Journal of Artificial Intelligence and Technology, 2026, 6, 507-519
https://doi.org/10.37965/jait.2026.1049

ISTP
RESEARCH ARTICLE

A Multi-Scale CNN–Transformer Fusion Framework with Stain Normalization and Focal Loss for High-Accuracy Multi-Stage Gastric Cancer Diagnosis

D. S. Radhika Shetty and P. J. Antony
Department of Computer Science & Engineering, A.J. Institute of Engineering and Technology,
Affiliated to Visvesvaraya Technological University, Belagavi, Mangalore, Karnataka, India

(Received 12 March 2025; Revised 18 January 2026; Accepted 31 March 2026; Published online 03 June 2026)

Abstract: Early-stage gastric cancer (GC) diagnosis from histopathological images remains challenging due to subtle morphological variations and inter-slide staining variability. This study proposes a deep learning-based multi-stage GC classification framework that integrates convolutional feature extraction with attention-based contextual modeling. Eight pretrained convolutional neural networks (CNNs) are evaluated, among which DenseNet121 and MobileNetV2 achieve the strongest baseline performance (accuracy ≈ 85.8% and 85.9%, respectively). Building on these results, two novel architectures are developed. The first is an enhanced DenseNet121 model that incorporates multi-path convolution, squeeze-and-excitation (SE) channel recalibration, and attention optimization to capture multi-scale morphological patterns. The second is a Hybrid DenseNet121–Transformer framework that integrates global self-attention with convolutional representations to improve contextual understanding of tissue structures. The models are trained using standardized preprocessing, Macenko stain normalization, extensive data augmentation, and class balancing on a dataset of 7,010 histopathology images representing Normal, Stage I, and Stage II gastric tissues. The proposed hybrid CNN–Transformer framework achieves 90.2% classification accuracy, a macro F1-score of 91.4%, and an Area Under the Curve (AUC) of 0.985, outperforming baseline CNN architectures in stage-wise discrimination. Attention-based visualization highlights diagnostically relevant tissue regions and improves model interpretability. These findings demonstrate that combining multi-scale convolutional representations with Transformer-based global attention provides a robust and interpretable framework for automated GC histopathology analysis.

Keywords: early-stage gastric cancer; hyperparameter tuning; multi-path convolution; transformer attention optimization

I. INTRODUCTION

Gastric and colonic epithelial cancers are major global health challenges, with gastric cancer (GC) remaining one of the most common and deadly malignancies worldwide [1,2]. GC is currently the fifth most common cancer and the fourth leading cause of cancer-related mortality, highlighting the urgent need for improved early detection strategies. Gastric carcinoma accounts for nearly 90–95% of all GC cases and represents the predominant form of this disease [3,4]. Owing to this dominance, the terms "gastric cancer" and "gastric carcinoma" are often used interchangeably in the literature [5]. Diagnosis primarily relies on histopathological examination of hematoxylin and eosin (H&E)-stained biopsy samples. These samples reveal key features such as gland formation, nuclear atypia, and invasion patterns that are essential for tumor classification and staging [6]. Despite advances in imaging and endoscopy, many cases are still detected at advanced stages. Manual histopathology remains the gold standard; however, it is time-consuming and subject to inter-observer variability [7–9]. This limitation has led to increasing interest in deep learning-based approaches for early and accurate diagnosis. Figure 1(a–c) illustrates the histopathological progression of GC.

Accurate diagnosis of GC remains challenging. Conventional histopathological methods are invasive, prone to sample contamination, and affected by inter-observer variability. Early-stage lesions often exhibit subtle morphological changes that may be overlooked by pathologists. This can lead to inconsistent or delayed diagnoses. Although timely detection improves treatment outcomes, current diagnostic practices remain limited. Deep learning has emerged as an effective solution for automated diagnosis. Pretrained convolutional neural network (CNN) models can automatically extract complex tissue patterns and detect subtle abnormalities with high accuracy [7]. These models improve diagnostic speed and reduce human error. They also support reliable and consistent assessments, even in resource-constrained environments.

However, clinical adoption remains limited due to two major challenges. The first is poor model interpretability. The second is limited generalizability across datasets. This study addresses these issues by fine-tuning and optimizing multiple pretrained CNN architectures, including DenseNet121, MobileNetV2, ResNet50, and InceptionV3. The models are trained on curated, augmented, and cross-validated datasets from GasHisSDB, SEED, and expert annotations. This transfer learning approach improves robustness and diagnostic accuracy. It also provides a strong foundation for developing interpretable and clinically deployable GC detection systems.

Corresponding author: D. S. Radhika Shetty (e-mail: ss.radhika@gmail.com).

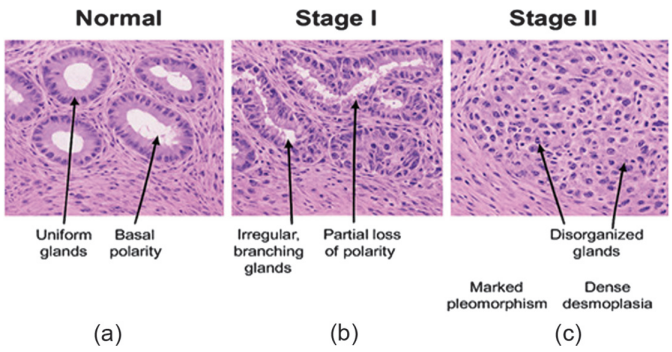

Fig. 1. Representative histopathological progression from (a) Normal gastric mucosa, (b) Stage I, and (c) Stage II gastric carcinoma under H&E staining.

The ensemble model beat all individual CNN architectures by using complementary feature representations from many networks, considerably enhancing robustness and diagnostic accuracy. This research addresses the difficulty of early-stage stomach cancer diagnosis by applying modern deep learning techniques for automated interpretation and categorization of histological images. A diverse multi-source dataset was curated from GasHisSDB, SEED, and expert-annotated biopsy slides, ensuring broad variability in staining, tissue patterns, and acquisition conditions. The dataset was thoroughly cleaned, standardized, and expanded utilizing substantial augmentation—including rotations, zooming, flips, shifts, and scaling to promote generalizability across unforeseen clinical circumstances.

Eight advanced pretrained CNN models—DenseNet121, EfficientNetB4, InceptionResNetV2, InceptionV3, MobileNetV2, ResNet50, VGG16, VGG19—were fine-tuned using Bayesian hyperparameter optimization. Custom layers were added to capture minute morphological variations critical for early GC detection. Beyond these baselines, two upgraded architectures from our study were added for ensemble prediction: (i) a Hybrid DenseNet121–Vision Transformer model employing Macenko stain normalization, label smoothing, focus loss, and multi-head self-attention for global context modeling and (ii) an enhanced DenseNet121–MPCNN-TAO–SE network integrating multi-path convolutional network (MPCNN), transformer attention optimization (TAO), and squeeze-and-excitation (SE) channel recalibration to learn multi-scale and clinically significant features. Together, these models were trained using class-balanced cross-entropy or localized loss to alleviate class imbalance and improve sensitivity for early-stage lesions. Evaluation included accuracy, recall, precision, AUC–ROC, and loss, demonstrating that the ensemble achieved superior performance benefiting from the CNNs' local texture sensitivity, the Transformer's long-range context modeling, and SE-based adaptive feature recalibration. This integrated deep learning technique considerably boosts early and multi-stage stomach cancer detection, reduces diagnostic variability, accelerates analysis, and delivers clinically interpretable predictions, delivering a significant improvement in pathology-assisted cancer diagnosis.

The main contributions of this study are as follows. First, a hybrid CNN–Transformer framework is developed for accurate early-stage GC detection from histopathological images. Second, extensive experiments are conducted on histopathological datasets to evaluate the effectiveness of the proposed model using standard performance metrics. Third, a comparative analysis is performed with existing deep learning architectures to demonstrate the superiority of the proposed approach. The remainder of this paper is organized as follows: Section II presents the literature review, Section III describes the proposed methodology, Section IV discusses the experimental setup and results, and Section V concludes the study with future research directions.

II. LITERATURE REVIEW

Deep learning has significantly transformed computer-aided cancer diagnosis. CNNs consistently outperform traditional handcrafted feature-based approaches in both radiological and histopathological image analysis. CNNs effectively capture local morphological patterns such as glandular structure, cellular texture, and nuclear atypia, which are critical for cancer detection. However, CNN-based approaches face several limitations. They struggle to model long-range tissue context. They also show limited performance in reliable stage-wise discrimination. In addition, they often lack clinically meaningful interpretability. These challenges are particularly evident in early GC diagnosis.

Early deep learning studies primarily rely on lightweight CNN architectures tailored to specific imaging modalities. MobileNet-based CNNs, combined with recurrent units and explainability modules, demonstrate high diagnostic accuracy for lung cancer. However, these models are limited to CT-based imaging and lack cross-modality generalizability [9]. Human–AI collaborative frameworks improve GC detection by incorporating expert feedback during inference. However, their dependence on specialist involvement limits scalability and real-world deployment [10]. CNN feature extractors combined with traditional classifiers, such as Support Vector Machine (SVMs), achieve high accuracy for gastrointestinal disease detection. However, these approaches focus mainly on coarse disease categorization and do not address stage-wise progression [11].

To improve sensitivity toward subtle pathological variations, attention-enhanced and multi-scale CNN architectures were introduced. These models improved early lesion detection by emphasizing diagnostically relevant regions; however, they remained constrained by patch-based inference and limited receptive fields, restricting their ability to capture global tissue architecture [12]. Shape- or segmentation-driven deep models further enhanced localization accuracy, yet their dependence on handcrafted priors reduced adaptability across heterogeneous datasets and cancer subtypes [13].

In histopathological analysis, stain normalization techniques combined with CNNs improve robustness against staining variability and inter-laboratory differences. However, these approaches still rely primarily on local feature extraction mechanisms [14]. More recently, Vision Transformer (ViT) architectures have gained attention for their ability to model long-range spatial dependencies. They also capture global contextual relationships across tissue regions. ViTs achieve strong performance across gastrointestinal and breast cancer datasets. However, their effectiveness is often limited by high data requirements and substantial computational cost [15]. Transfer learning-based CNNs and modern convolutional architectures, such as ConvNeXt, show promising results in breast cancer detection and prognostication. However, most studies focus on binary classification or outcome prediction. They do not adequately address early-stage or multi-stage cancer diagnosis [16,17]. Bayesian-optimized CNNs improve training stability and classification accuracy. However, they provide limited interpretability, which reduces clinical transparency [18]. Ensemble-based CNN frameworks enhance robustness and

Table I. Recent deep learning approaches for cancer diagnosis

Ref.	Cancer type	Model/architecture used	Key results	Main limitations
[9]	Lung cancer	MobileNetV2 + GRU + Grad-CAM	Accuracy >96%	CT-based only
[10]	Gastric cancer	CNN + Human–AI Fusion	AUC > 0.97	Expert-dependent
[11]	GI diseases	CNN Feature Extractor + SVM	Accuracy up to 99%	No stage-wise analysis
[12]	Gastric cancer	Multi-scale CNN + Attention	Accuracy >91%	Limited global context
[13]	Lung cancer	Active Shape Model + DNN	Improved detection	Handcrafted priors
[14]	Gastric cancer	CNN + Stain Normalization	Improved generalization	CNN-only
[15]	Multi-cancer	MedSAM (ViT-Base)	Robust generalization	Training data imbalance
[16]	Breast Cancer	Transfer-learned CNNs	Accuracy 82–89%	Binary classification
[17]	Breast cancer	ConvNeXt-based CNN	Improved prognostication	Not early detection
[18]	Gastric cancer	CNN + Bayesian Optimization	Accuracy ≈ 90%	Limited interpretability
[19]	Skin cancer	DoubleU-Net + CNN	AUC up to 94.98%	Segmentation-dependent
[20]	Gastric cancer	Deep Ensemble CNN	Accuracy >92%	High inference cost

generalization. However, they significantly increase inference complexity, making them less suitable for real-time or resource-constrained clinical settings [19,20]. A consolidated comparison of these approaches and their limitations is presented in Table I.

Existing deep learning methods primarily focus on specific aspects of cancer diagnosis. CNN-based models emphasize local feature learning. Transformer-based models focus on global context modeling. Ensemble methods aim to improve robustness. However, each of these approaches involves inherent trade-offs. Most prior studies do not provide a unified framework. They fail to simultaneously address multi-scale feature representation, global contextual reasoning, interpretability, and computational efficiency. These limitations are particularly significant in stage-wise diagnostic settings. These unresolved challenges motivate the proposed hybrid CNN–Transformer framework. The proposed approach integrates multi-path convolutional feature extraction, channel-wise recalibration, and Transformer-guided global attention. This combination enables accurate, interpretable, and efficient multi-stage GC diagnosis.

III. METHODOLOGY

The proposed methodology focuses on developing an accurate and efficient deep learning-based classification system for early detection of GC using histopathological images, as shown in Fig. 2. The dataset consists of Normal, Stage I, and Stage II images. Data cleaning is performed to remove corrupted and mislabeled samples. Data augmentation techniques, including rotation, flipping, zooming, and contrast adjustment, are applied to improve data diversity. These techniques also address class imbalance, particularly for underrepresented Stage I cases. For unbiased model evaluation, the dataset is divided into 80% training and 20% validation subsets. Transfer learning is applied to refine eight pretrained CNN architectures, including MobileNetV2, DenseNet121, ResNet50, InceptionV3, VGG16, VGG19, InceptionResNetV2, and EfficientNetB4. Hyperparameter tuning is performed to optimize the number of custom dense layers, learning rate, dropout rate, and optimizer for each model. This process improves classification performance, particularly for Stage I detection. The models are evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and validation loss to ensure robust and reliable classification. The best-performing models demonstrate high sensitivity and strong generalization capability for early-stage GC detection. The trained models are stored in HDF5 (.h5) format using Keras. During prediction, input images are preprocessed, features are extracted, and classification results are generated to determine cancer presence and stage.

A. DATASET COLLECTION

The experimental evaluation in this study was conducted using a combination of publicly available GC histopathological datasets and expert-annotated clinical data. The publicly available datasets used include GasHisSDB[21] and SEED (Stomach cancer Endoscopic and histopathological Dataset) [22]. The GasHisSDB dataset provides H&E-stained gastric histopathology images categorized into normal and cancerous classes and has been widely adopted for computer-aided GC diagnosis. The SEED dataset contains curated GC histopathological images with expert annotations and supports research on automated GC detection. In this study, histopathological image subsets from both datasets were utilized to enhance data diversity and support multi-stage classification. In addition, expert-annotated gastric histopathological images obtained through clinical collaboration were used for validation and categorized into Normal, Stage I carcinoma, and Stage II carcinoma. All images were anonymized prior to analysis[23]. The main dataset used in this study was compiled from two publicly available sources, GasHisSDB and the SEED gastric carcinoma dataset, which provide pre-extracted histopathological image patches. Since these datasets do not include patient-level identifiers or slide-level metadata, strict patient-wise separation cannot be enforced. To prevent data leakage during training, the dataset was divided using a stratified 80–20 split before applying augmentation. Data augmentation was applied only to the training set, ensuring that transformed versions of the same image did not appear in the validation set.

B. PREPROCESSING and AUGMENTATION

All images were augmented, resized, reshaped, and converted into NumPy-based pixel arrays to standardize input format. Pixel values were normalized to ensure consistent intensity ranges and computational readiness for model training. The dataset comprised 3,500 Normal, 1,800 Stage I, and 1,800 Stage II images. Pixel intensities were scaled to the [0,1] range (1/255). Eight augmentation techniques—including rotation, width/height shifts, horizontal flipping, shearing, and zooming—were applied to enhance

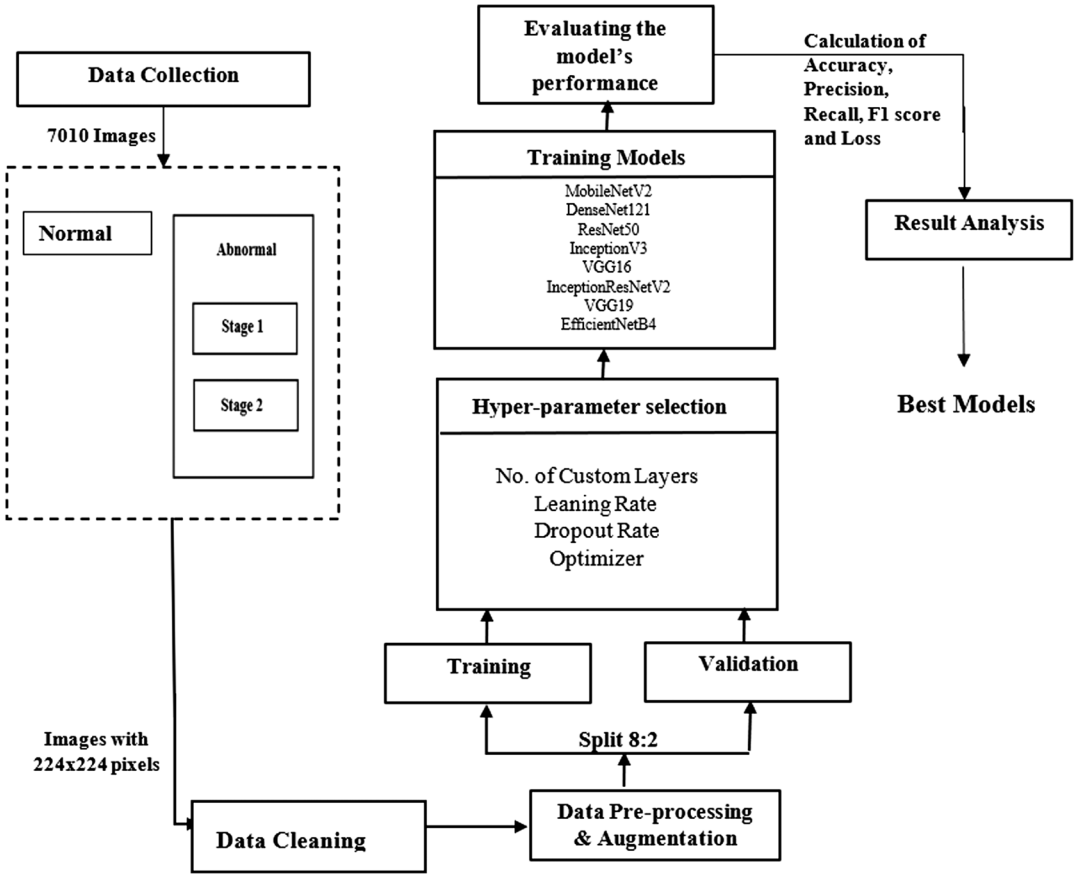

Fig. 2. Workflow illustrating the process for gastric cancer classification.

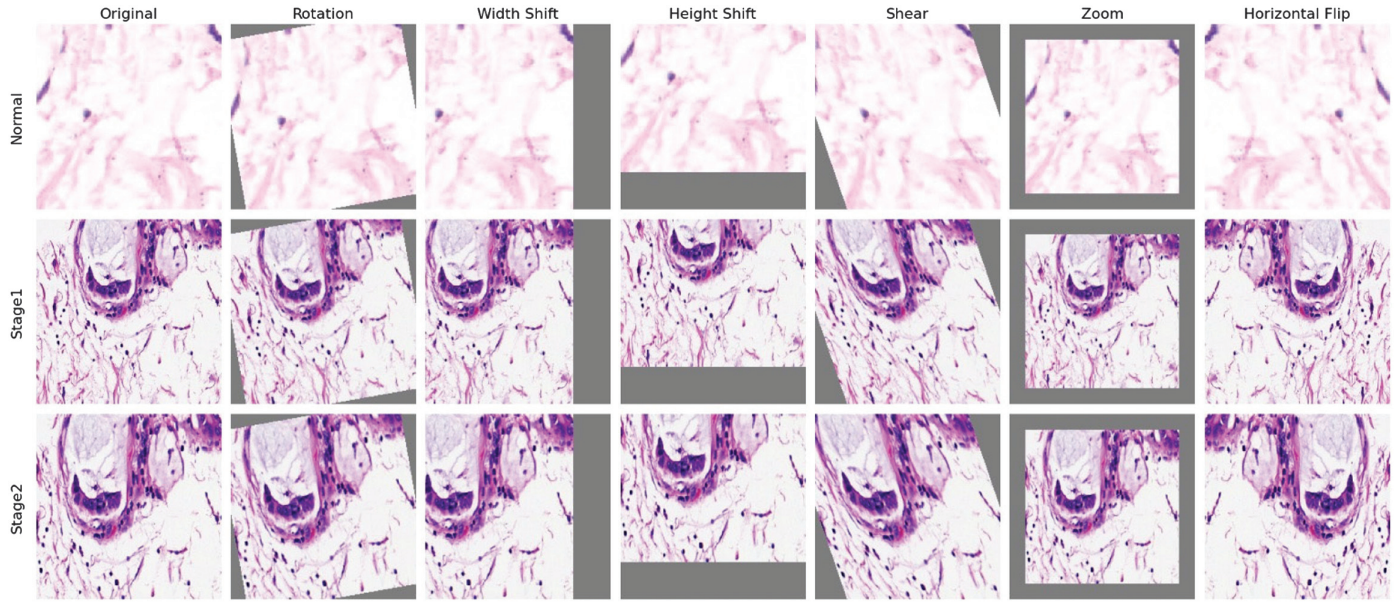

Fig. 3. Data augmentation techniques applied to gastric cancer images.

variability. This process generated an additional 41,646 augmented images, ensuring stronger generalization across stages. We used Rotation to mimic variations in slide orientation during imaging, Width shift and height shift to account for tissue displacement and variability in slide preparation, Fill operations to handle boundary effects introduced by shifts and rotations, Random horizontal flipping to enhance invariance to lateral tissue symmetry, Shearing transformations to simulate geometric distortions in tissue

structure, and Zooming to replicate variations in magnification levels. Figure 3 illustrates the data augmentation techniques applied to gastric cancer histopathological images to improve dataset diversity and model generalization.

C. HYPERPARAMETER TUNING

To optimize model performance, Bayesian optimization via Keras Tuner was employed to intelligently search for the most effective hyperparameter configurations. We created a bespoke HyperModel class to encapsulate the architecture and configurable parameters for simplified experimentation and architectural flexibility. The tuning process comprised 25 trials per model across eight different model variants, striking a balance between computational feasibility and comprehensive search coverage. Key hyperparameters included the number of convolutional layers, varied between 3 and 7, each with 512 filters, allowing the model to learn features across multiple abstraction levels.

Hyperparameter optimization aims to find the best configuration:

$$\theta^* = \arg\max_{\theta \in \mathcal{H}} A_{val}(\theta)$$

where

$\theta = \{L,\eta,d\}$ represents:

- L: number of convolutional layers
- η: learning rate
- d: dropout rate

Search ranges:

$$L \in \{3,4,5,6,7\},\ \eta \in [10^{-4},10^{-2}],\ d \in [0.1,0.6]$$

Figure 4 shows the steps in hyperparameter tuning. Additionally, the learning rate was tuned in the range of 0.0001 to 0.01, and dropout rates varied from 0.1 to 0.6, guided by prior studies and initial experimental outcomes. These parameters were critical in determining the model's capacity, training efficiency, and its ability to generalize without overfitting. This approach to hyperparameter tuning resulted in improved validation accuracy and optimized overall performance while maintaining computational efficiency.

D. MODELS' SELECTION

This study assessed the performance of eight different CNN architectures—DenseNet121, EfficientNetB4, InceptionResNetV2, InceptionV3, MobileNetV2, ResNet50, VGG16, and VGG19—for the early detection of GC. These models were chosen because they successfully address medical image analysis difficulties such as tissue texture, computational efficiency, and clinical deployment scalability [24].

Figure 5 outlines the process of integrating custom layers into the DenseNet121 architecture following initial training and hyperparameter tuning.

E. ENHANCED HYBRID DENSENET121-BASED FRAMEWORK WITH MPCNN-TAO AND SE MODULES

The proposed framework is built on DenseNet121, selected for its dense connectivity and efficient feature reuse and enhanced with three major modules to better capture complex multi-scale patterns in gastric histopathology images. The introduction of multi-path convolution blocks improved feature representation by enabling

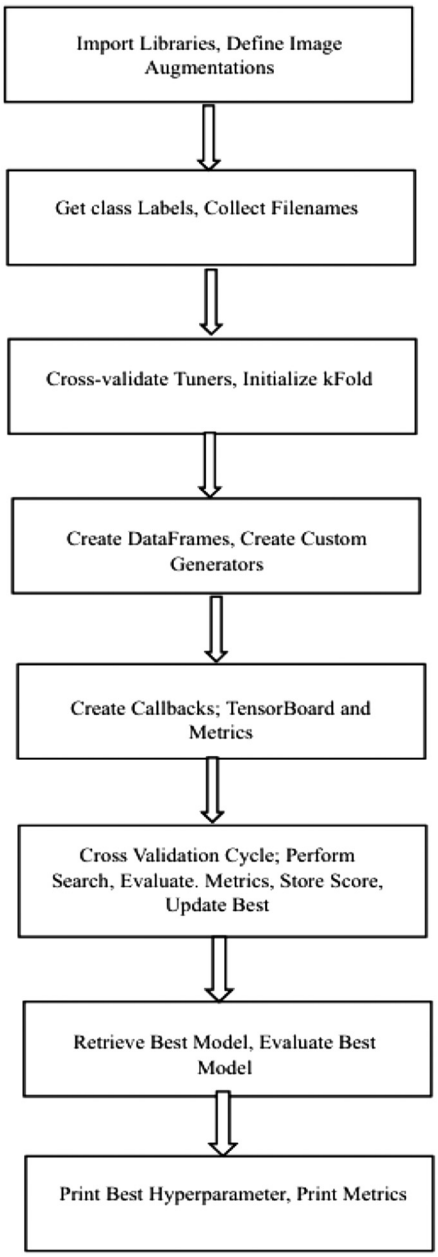

Fig. 4. Hyperparameter tuning.

the model to capture morphological structures at multiple spatial scales. Incorporating SE blocks further enhanced feature discrimination by adaptively recalibrating channel importance. The addition of transformer-based attention allowed the model to capture long-range spatial relationships within tissue structures that are difficult for CNNs alone to model.

First, a Multi-Path Feature Extraction block employs parallel 1×1, 3×3, and 5×5 convolutions to simultaneously capture fine-grained cellular textures, mid-level glandular structures, and broad contextual tissue patterns, whose outputs are concatenated to form a rich feature embedding. Second, SE blocks are introduced after each multi-path unit to perform channel-wise attention through global average pooling (squeeze), two fully connected layers with Rectified Linear Unit (ReLU) and sigmoid activations (excitation), and adaptive channel-wise reweighting (scale), thereby

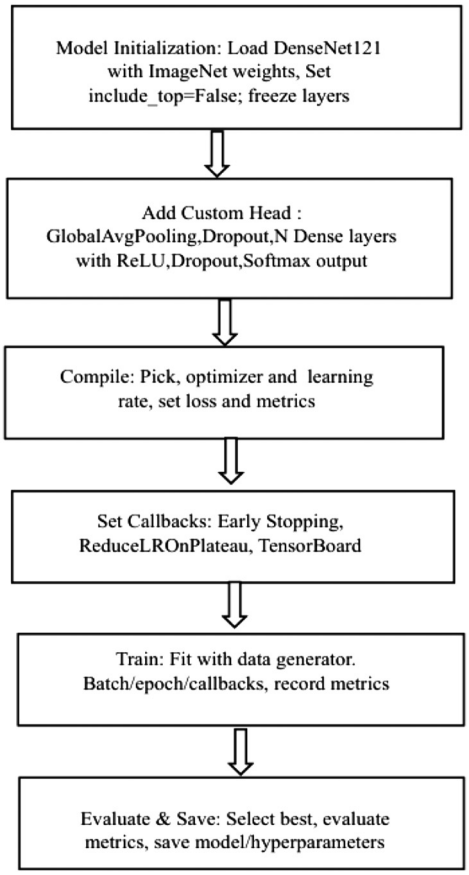

Fig. 5. DenseNet121 with custom layers.

emphasizing diagnostically important channels linked to malignancy. Third, a Transformer-inspired Multi-Head Self-Attention module (MPCNN-TAO) is integrated to model long-range spatial dependencies that CNNs alone cannot capture; this hybrid block synergistically combines convolutional locality, global reasoning, and multi-scale pattern detection for improved sensitivity to subtle early-stage (Stage I) morphological variations[25]. Training follows a transfer learning strategy where DenseNet121 is initialized with ImageNet weights and newly added hybrid layers are trained from scratch using weighted categorical cross-entropy, Adam/RMSprop optimizers with tuned learning rates, dropout, early stopping, and learning-rate decay. During deployment, input images are preprocessed and sequentially passed through the DenseNet121 backbone, Multi-Path block, SE recalibration, and MPCNN-TAO module, with final class predictions (Normal, Stage I, and Stage II) generated using a softmax classifier, and the complete model exported in HDF5 (.h5) format for seamless clinical integration.

Multi-Path + SE Feature Extraction is given by

$$\mathbf{F} = [W_{1\times1} * \mathbf{X} \| W_{3\times3} * \mathbf{X} \| W_{5\times5} * \mathbf{X}] \cdot \sigma, (W_2\delta(W_1\mathrm{GAP}(\mathbf{F})))$$

where $W_{1\times1}, W_{3\times3}, W_{5\times5}$ are convolutional kernels for multi-scale feature extraction. GAP is the global average pooling (squeeze). $\delta(\cdot)$ and $\sigma(\cdot)$ are ReLU and sigmoid activations, respectively. W_1, W_2 are fully connected layers for SE excitation

This equation models multi-path convolution followed by channel recalibration to emphasize discriminative tissue patterns.

Hybrid Convolution–Attention Fusion (MPCNN-TAO) can be represented as

$$\mathbf{H} = \alpha\mathbf{F} + (1-\alpha)\Big\|_{h=1}^{H} \mathrm{softmax}, \left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h$$

where Q_h, K_h, V_h are query, key, and value projections for head h, respectively, and d_k is the attention scaling factor α is the fusion weight This formulation combines convolutional locality with global self-attention to capture long-range spatial dependencies in histopathological structures.

Softmax classification is given as

$$\hat{y}_k = \frac{e^{w_k^T \mathrm{GAP}(\mathbf{H}) + b_k}}{\sum_{j=1}^{3} e^{w_j^T \mathrm{GAP}(\mathbf{H}) + b_j}}$$

where w_k, b_k are classifier weights and biases, and $\hat{y}_k$ is the probability of class k (Normal, Stage I, Stage II).

This equation generates the final diagnostic class probabilities using a softmax decision layer.

F. HYBRID DENSENET–TRANSFORMER FRAMEWORK WITH STAIN NORMALIZATION AND FOCAL LOSS

The proposed methodology employs a Hybrid DenseNet121–Transformer architecture for multi-stage GC classification from H&E-stained histopathological images. To address staining variability, Macenko stain normalization is applied to all the images, ensuring consistent color distribution across slides. Macenko stain normalization reduced color variability across histopathology slides, improving model robustness across heterogeneous datasets. The focal loss improved learning from difficult samples and mitigated the class imbalance present in Stage II images. Misclassification between Stage I and Stage II cases may occur due to the subtle morphological differences and overlapping histopathological characteristics between early and intermediate stages of gastric carcinoma. This challenge is consistent with observations reported in clinical pathology studies.

DenseNet121 pretrained on ImageNet serves as the feature extractor, generating convolutional feature maps that capture local morphological details. These feature maps are tokenized through a 1×1 convolution and reshaped into sequences enriched with positional encodings. The token sequence is processed by lightweight Transformer encoder blocks that consist of multi-head self-attention, pre-layer normalization, Gaussian Error Linear Unit (GELU)-activated feed-forward networks, and residual shortcuts to model long-range dependencies in tissue architecture. CNN and Transformer features are concatenated and passed through a multilayer classification head (Dense-BN-Dropout units), ending with a softmax layer for three-class prediction.

Training follows a two-phase strategy: (i) the DenseNet backbone is frozen, while the Transformer and classification head are trained using categorical cross-entropy with label smoothing; and (ii) the final 60 layers of DenseNet121 are unfrozen for fine-tuning using focal loss ($\gamma = 2.0$) to improve sensitivity for underrepresented Stage II carcinoma. Adam optimizer, class weighting, dropout regularization, early stopping, and learning rate decay are used to ensure stable convergence and mitigate overfitting.

CNN feature maps are projected and converted into Transformer tokens:

$$\mathbf{T} = \mathrm{PE}(\mathrm{reshape}(W_{1\times1} * \mathrm{DenseNet121}(X)))$$

where X: is the input histopathology image. DenseNet121(X) is convolutional feature maps extracted by DenseNet121. $W_{1\times1}$: 1×1

is the convolution used for token projection. $*$ is the convolution operator. reshape$(\cdot)$: converts spatial CNN maps into a token sequence. PE$(\cdot)$ is the positional encoding added to tokens. $\mathbf{T}$ is the transformer-ready input sequence.

Multi-Head Self-Attention captures global spatial dependencies in the token sequence,

$$\mathrm{MHSA(T)} = \Big\|_{h=1}^{H} \mathrm{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h$$

where H is the number of attention heads. Q_h, K_h, V_h are query, key, and value projections for head h, respectively. $Q_h K_h^T$ is the similarity matrix between tokens. $\sqrt{d_k}$ is the scaling factor for numerical stability. softmax$(\cdot)$ is the normalization over token scores. V_h is the value vectors weighted by attention. $\|$ is the concatenation across all heads. MHSA$(\mathbf{T})$ is the output of the attention block.

CNN and Transformer features are fused and classified using a softmax layer,

$$\hat{y}_k = \frac{e^{w_k^T[\mathrm{GAP}(\mathbf{F}_{cnn})\|\mathrm{MHSA}(\mathbf{T})]+b_k}}{\sum_{j=1}^{3} e^{w_j^T[\mathrm{GAP}(\mathbf{F}_{cnn})\|\mathrm{MHSA}(\mathbf{T})]+b_j}}$$

where $\mathbf{F}_{cnn}$ is DenseNet121 convolutional features, GAP$(\cdot)$ is the global average pooling of CNN features, $\|$ is the feature-level concatenation of CNN and transformer outputs, w_k, b_k are softmax classifier weights and biases for class k, and $\hat{y}_k$ is the predicted probability for class $k \in$ {Normal, Stage I, Stage II}. Denominator ensures probabilities sum to 1.

IV. RESULT AND DISCUSSION

A symmetric dataset containing Normal, Stage I, and Stage II gastric histopathology images was used to evaluate multiple CNN architectures—including DenseNet121, MobileNetV2, InceptionV3, InceptionResNetV2, ResNet50, VGG16/19, and EfficientNetB4—to identify the best model for early GC detection. Initial experiments confirmed that CNN-based methods substantially improve diagnostic accuracy while reducing the time and manual workload associated with traditional pathology. Bayesian optimization was employed to determine optimal hyperparameter ranges, followed by K-fold cross-validation to ensure robust model generalization. Each architecture was fine-tuned to strengthen its ability to distinguish normal tissue from early and progressive carcinoma. DenseNet121 achieved strong performance with three custom dense layers, learning rate of 0.0002, and dropout rate of 0.4 using the RMSprop optimizer. MobileNetV2, due to its lightweight design, required careful dense-layer tuning and a dropout rate of 0.3, with Stochastic Gradient Descent (SGD) (learning rate $\approx$ 0.0036) yielding a favorable balance between accuracy and efficiency on the limited dataset. The optimal hyperparameter settings used for selecting the best-performing CNN model are summarized in Table II.

Table III consolidates the validation performance of various CNN models tested for early-stage GC detection, highlighting notable differences in classification accuracy and overall model effectiveness. MobileNetV2 achieved the highest accuracy of 85.9%, making it a strong candidate for practical deployment due to its lightweight architecture and computational efficiency. Closely following was DenseNet121, with an accuracy of 85.8%, demonstrating its strength in extracting intricate histopathological patterns through its densely connected layers.

The confusion matrix shows model accuracy, sensitivity, and specificity by demonstrating how well each class is predicted. Figure 4 displays the confusion matrices for all eight CNN models, comparing initial results with those achieved after fine-tuning. Darker-colored cells indicate stronger correct classification, providing an intuitive visualization of each model's ability to differentiate between the histopathological stages. Figure 6 shows the confusion matrices of all eight fine-tuned CNN models. DenseNet121 and MobileNetV2 demonstrated the

Table II. Optimal hyperparameter settings for selecting the best CNN models

Model/best parameter	No. of custom layers	Learning rate	Dropout rate	Optimizer
DenseNet121	5	0.0002	0.4	RMSprop
EfficientNetB4	7	0.0018	0.4	Adam
InceptionV3	5	0.0020	0.1	RMSprop
InceptionResNetV2	4	0.0001	0.4	RMSprop
MobileNetV2	4	0.0036	0.3	Sgd
ResNet50	4	0.0001	0.1	Adam
VGG16	5	0.0016	0.1	RMSprop
VGG19	3	0.0004	0.1	RMSprop

Table III. Validation metrics for CNN models in tumor classification

Model\metrics	Accuracy	Precision	Recall	F1-score	Log loss
DenseNet121	0.8582	0.8649	0.8582	0.8562	0.3193
EfficientNetB4	0.7580	0.6335	0.7580	0.6780	0.6808
InceptionV3	0.8096	0.8150	0.8096	0.8094	0.3427
InceptionResNetV2	0.8039	0.7998	0.8039	0.8101	0.3564
MobileNetV2	0.8592	0.8589	0.8592	0.8590	0.2936
ResNet50	0.8145	0.8147	0.8145	0.8141	0.3496
VGG16	0.8062	0.8075	0.8062	0.8040	0.3053
VGG19	0.7970	0.8034	0.7970	0.7850	0.3425

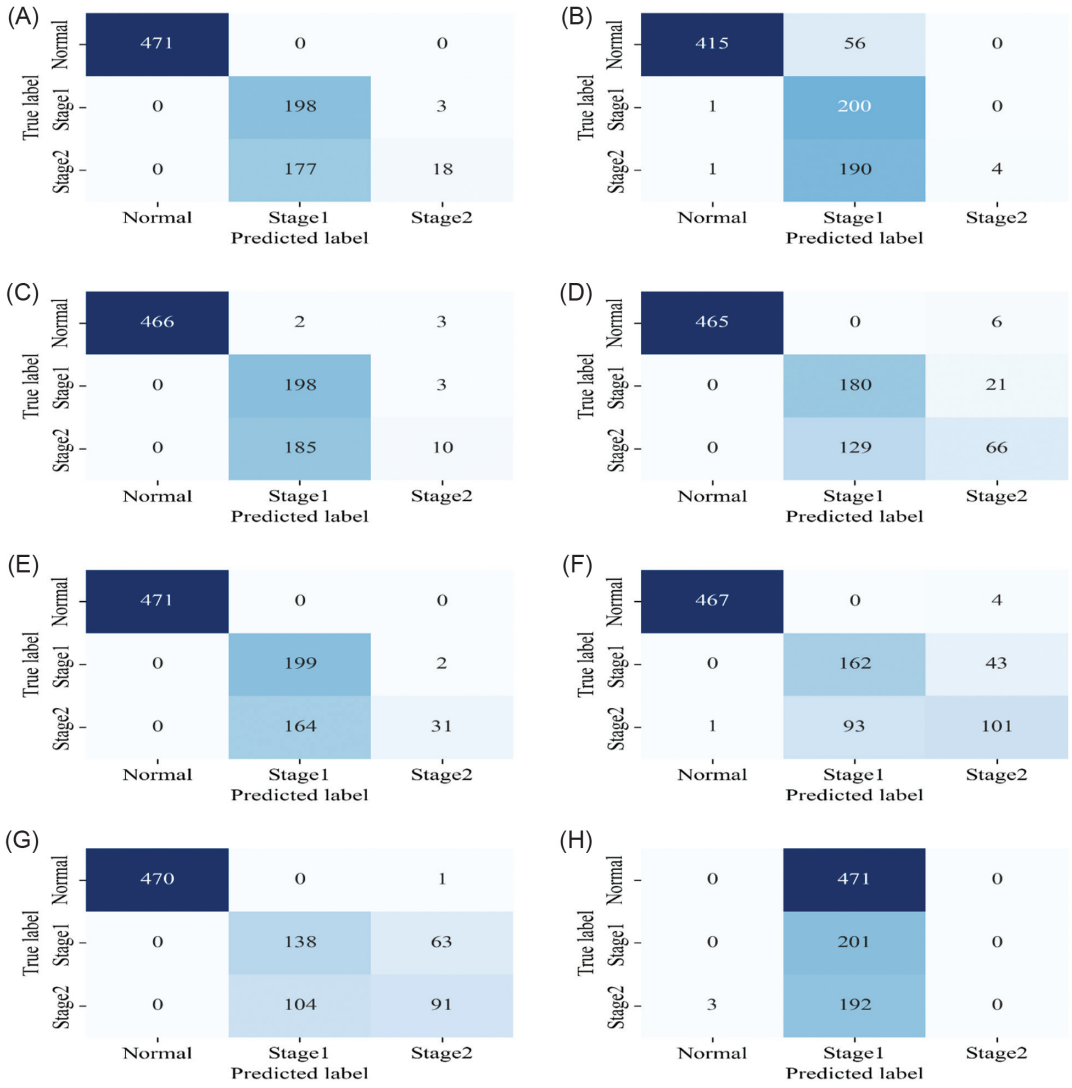

Fig. 6. Confusion matrix for the testing of eight fine-tuned CNN models: (a) DenseNet121, (b) EfficientNetB4, (c) InceptionResNetV2, (d) InceptionV3, (e) MobileNetV2, (f) ResNet50, (g) VGG16, and (h) VGG19.

strongest overall performance, achieving perfect accuracy for the Normal class and exceptionally high recall for Stage I, making them promising candidates for early detection and real-time applications. However, both models—along with InceptionResNetV2 and InceptionV3—struggled to correctly classify Stage II cases, often mislabeling them as Stage I. InceptionV3 showed substantial improvement in Stage I recall after fine-tuning but sacrificed Stage II accuracy. ResNet50 showed moderate and balanced gains, while VGG16 showed no improvement. VGG19 exhibited extreme class bias, predicting nearly all samples as Stage I, rendering it clinically unusable. Overall, the results indicate that while CNNs are highly effective for early GC identification, reliable Stage II detection requires improved fine-tuning strategies, balanced datasets, and stronger bias-mitigation techniques to ensure clinically trustworthy performance across all cancer stages.

Figure 7 presents the receiver operating characteristic (ROC) curves for the fine-tuned CNN models, illustrating the trade-off between the true positive rate (TPR) and false positive rate (FPR) across varying classification thresholds.

All models show AUC > 0.98 for Stage I, confirming high sensitivity in detecting early-stage cancer. The study shows that pretrained and custom CNN models, DenseNet121 and InceptionV3, demonstrate near-perfect classification for Normal and Stage I cases, making them well-suited candidates for GC diagnostic systems. MobileNetV2, being a lightweight model, with high accuracy, makes it a suitable choice for mobile application design for early-stage detection of GC.

The enhanced DenseNet121 architecture, augmented along with Multi-Path Convolutional blocks, SE layers, and a Transformer Attention module (MPCNN-TAO), achieved the strongest overall performance. Across experiments, DenseNet121 recorded an accuracy of 89.82%, precision of 87.49%, recall of 89.82%, and an F1-score of 85.62%, supported by a low log loss of 0.3193. Bayesian hyperparameter optimization with 8-fold cross-validation yielded optimal settings (five dense layers, learning rate 0.0002, dropout 0.4, and RMSprop optimizer), resulting in stable convergence and significant gains in Stage I detection. The confusion matrices for the initial and fine-tuned MPCNN models are presented in Fig. 8. Figure 9 presents the performance evaluation of

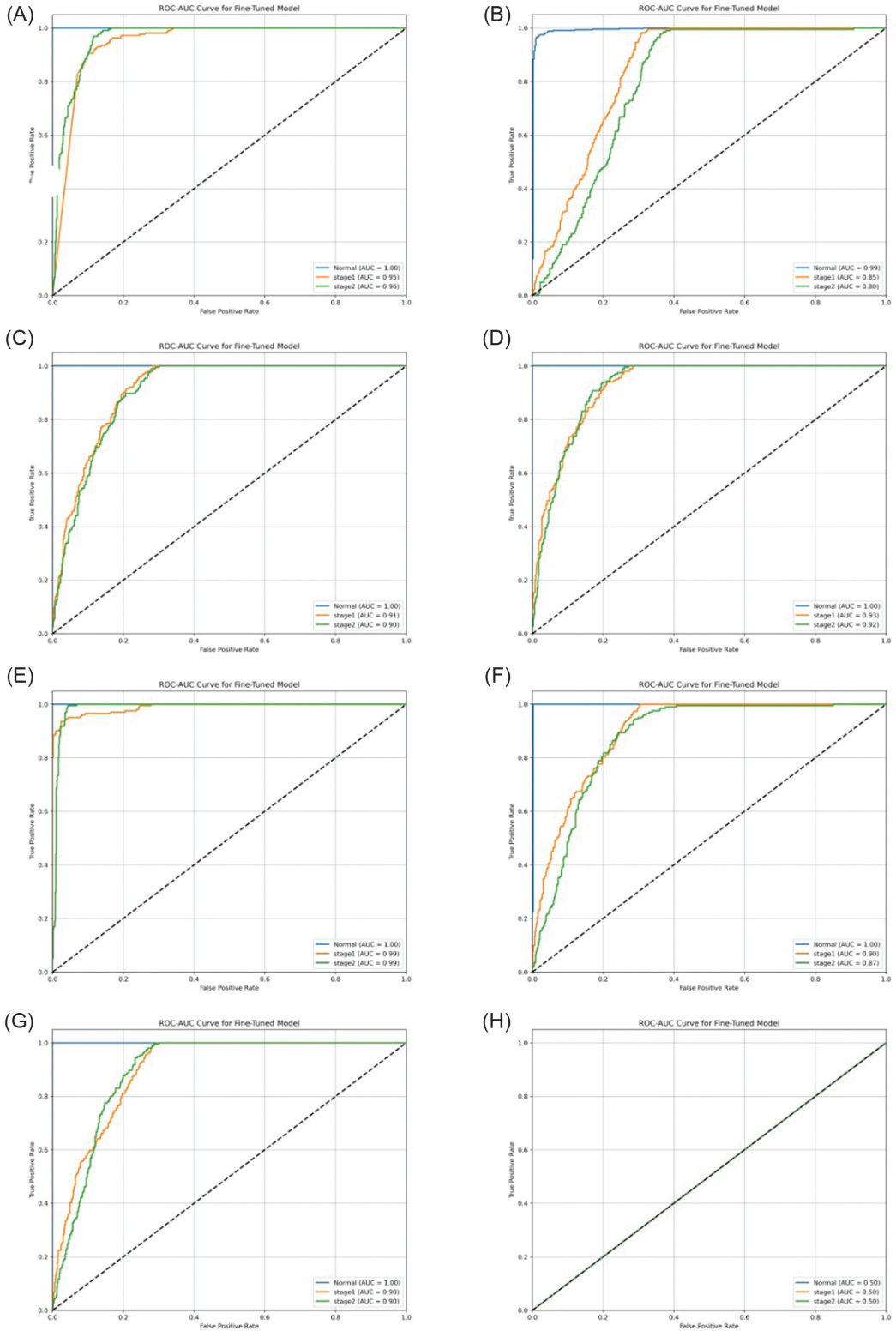

Fig. 7. ROC–AUC curves depicting the classification performance of fine-tuned CNN models for gastric cancer classification: (a) DenseNet121, (b) EfficientNetB4, (c) InceptionResNetV2, (d) InceptionV3, (e) MobileNetV2, (f) ResNet50, (g) VGG16, and (h) VGG19.

the MPCNN model in terms of (a) ROC–AUC characteristics and (b) training and validation loss.

Training and validation curves indicated that DenseNet121 and MobileNetV2 generalized well, whereas heavier architectures like EfficientNetB4 and VGG19 exhibited overfitting. Confusion matrices further confirmed DenseNet121's superior recall for Normal and Stage I tissues, although Stage II misclassification persisted due to overlapping morphological features. Importantly, the MPCNN-TAO

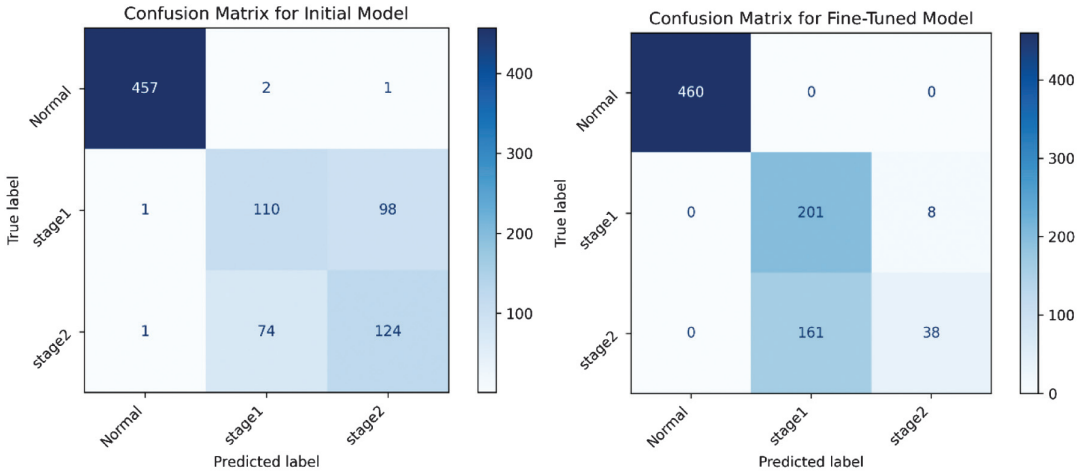

Fig. 8. Confusion matrix for initial and fine-tuned MP CNN models.

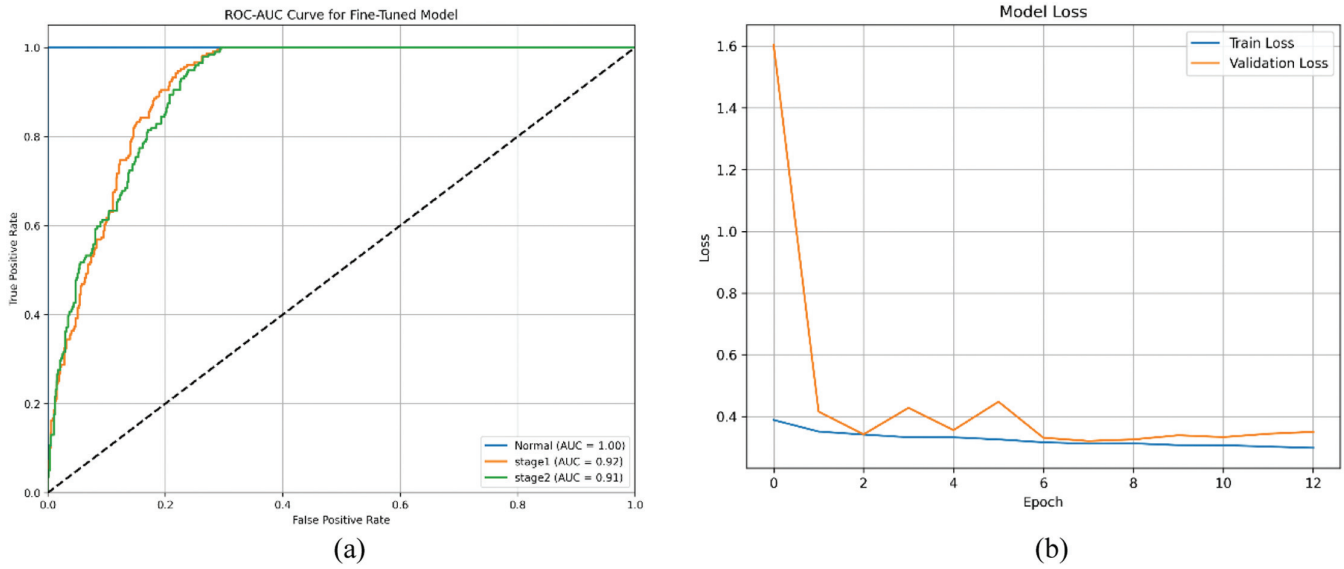

Fig. 9. MPCNN: (a) model ROC–AUC and (b) model loss.

module improved the capture of global context, enabling better discrimination of subtle and ambiguous tissue regions. Grad-CAM visualization highlighted that the enhanced model consistently focused on diagnostically relevant morphological structures, supporting interpretability and clinical trustworthiness.

Overall, the proposed DenseNet121-MPCNN-TAO-SE framework achieved the best balance of accuracy, robustness, and interpretability, making it a strong candidate for deployment in early GC screening and assisting pathologists in real-world diagnostic workflows. The comparative performance analysis of the proposed hybrid CNN model with existing approaches is presented in Table IV.

The proposed Hybrid DenseNet121–Transformer model, enhanced with Macenko stain normalization and focal loss, achieved a test accuracy of 90.2%, outperforming all CNN-only baselines. Class-wise performance was strong across all categories: Normal (98.0% accuracy, F1 = 97.7%), Stage I (89.0% accuracy, F1 = 89.9%), and Stage II (86.0% accuracy, F1 = 86.6%). The framework achieved a macro F1-score of 91.4% and a macro-averaged AUC value of 0.985, indicating excellent discriminative ability. The confusion matrix showed minimal misclassification of Normal samples, with most errors occurring between Stage I and Stage II—consistent with their close histological similarity. ROC curves confirmed strong per-class discrimination, with AUC values of 0.981 (Normal), 0.983 (Stage I), and 0.988 (Stage II). The confusion matrix and ROC analysis for the proposed DenseNet121 +Transformer model are illustrated in Fig. 10.

Comparative evaluation demonstrated that the hybrid CNN–Transformer design performed significantly better than VGG16, ResNet50, and standalone DenseNet121 models, particularly in Stage II recall. Focal loss and stain normalization contributed to improved robustness and reduced class imbalance effects. Overall, the results verify that combining local CNN features with global Transformer attention yields a more accurate and reliable model for multi-stage GC classification.

Model performance metrics were computed across validation experiments to ensure stable performance estimates. Although the proposed hybrid models consistently outperformed baseline CNN

Table IV. Comparison of the hybrid CNN model performance

Model	Accuracy	Precision	Recall	F1-score	Loss/log loss
Enhanced DenseNet121 + MPCNN-TAO	0.8982	0.8749	0.8982	0.8562	0.3193
Hybrid DenseNet121 + Transformer	0.9020	0.915	0.908	0.914	0.3122
DenseNet121 (fine-tuned)	0.8582	0.8649	0.8582	0.8562	0.3193
MobileNetV2	0.8592	0.8589	0.8592	0.8590	0.2936
InceptionV3	0.8096	0.8150	0.8096	0.8094	0.3427
InceptionResNetV2	0.8039	0.7998	0.8039	0.8101	0.3564
ResNet50	0.8145	0.8147	0.8145	0.8141	0.3496
VGG16	0.8062	0.8075	0.8062	0.8040	0.3053
VGG19	0.7970	0.8034	0.7970	0.7850	0.3425
EfficientNetB4	0.7580	0.6335	0.7580	0.6780	0.6808

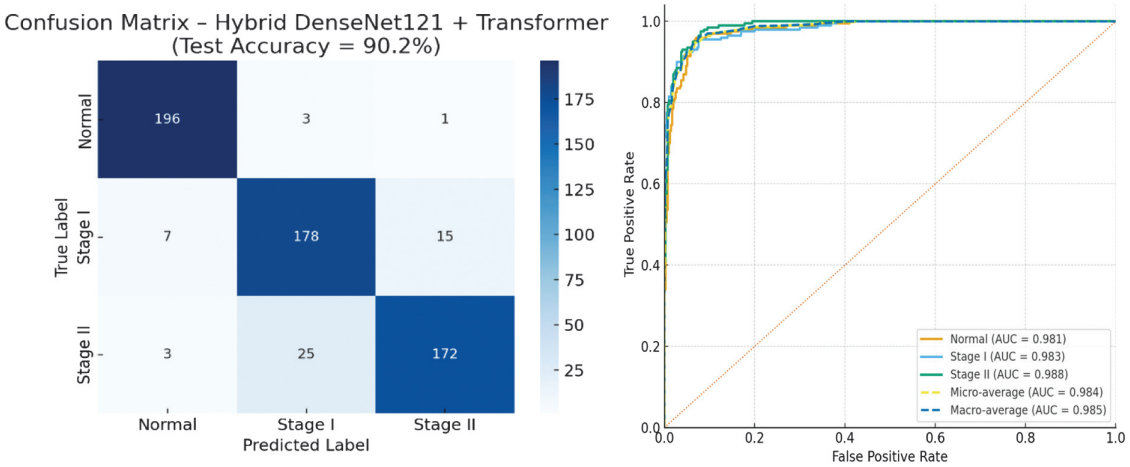

Fig. 10. Confusion matrix and ROC for DenseNet121 + Transformer model.

architectures, future work will incorporate formal statistical significance testing, such as confidence interval estimation and paired statistical tests across folds, to further validate the observed improvements.

Although the dataset used in this study integrates histopathological images from multiple independent sources, including GasHisSDB, SEED, and expert-annotated slides, evaluation was limited to internal validation. The heterogeneous nature of the dataset introduces variability in staining patterns, acquisition conditions, and tissue morphology, which partially improves the generalizability of the proposed models. However, external validation on completely independent clinical datasets from different institutions would further strengthen the clinical reliability of the framework. Future work will focus on cross-institutional validation and cross-dataset generalization experiments.

The proposed hybrid CNN–Transformer architecture introduces additional computational complexity compared to standalone CNN models due to the integration of multi-path convolution, SE blocks, and Transformer attention mechanisms. Training was conducted using Graphics Processing Unit (GPU) acceleration to efficiently handle the increased parameter space and attention computations. Despite this complexity, the framework remains feasible for modern clinical environments where GPU-based systems are increasingly available for medical image analysis. For resource-constrained or edge deployment scenarios, lightweight architectures such as MobileNetV2 may offer a practical alternative due to their lower computational requirements. Future work will explore model optimization techniques such as pruning, quantization, and knowledge distillation to improve deployment efficiency.

A. COMPARATIVE DISCUSSION WITH RELATED WORKS

The experimental results demonstrate that the proposed Hybrid DenseNet121–MPCNN-TAO-SE and Hybrid DenseNet121–Transformer frameworks significantly outperform conventional deep learning models for GC histopathological classification. Earlier studies predominantly employed single-path CNN architectures such as VGG16/19, ResNet50, and InceptionV3, reporting accuracies between 78% and 85% but exhibiting limited capability to capture global tissue context and subtle inter-stage variations.

The DenseNet121–MPCNN-TAO-SE model addresses these limitations through multi-scale convolutional learning and adaptive channel recalibration. This approach improves sensitivity for early-stage detection. However, some misclassification of Stage II cases still persists. This limitation arises from the locality constraints of convolution-based attention mechanisms. In contrast, the Hybrid DenseNet121–Transformer model captures long-range spatial dependencies using multi-head self-attention. This enables more effective discrimination between Stage I and Stage II carcinoma. It achieves this by modeling global tissue-level relationships that are not captured by CNN-only or MPCNN-based architectures.

Previous hybrid approaches, such as CNN–RNN and attention-based frameworks proposed by Iizuka *et al.* [24], achieve ROC–AUC values close to 0.95. However, these methods require large datasets and high computational cost. The proposed Transformer-based hybrid model achieves an ROC–AUC approaching 0.98 with lower computational complexity. Both proposed hybrid models also outperform EfficientNetB4 and MobileNetV2. These baseline models show underfitting when applied to complex histopathological patterns. Furthermore, Grad-CAM visualizations demonstrate improved alignment with diagnostically relevant glandular regions, consistent with interpretability findings reported by Song *et al.* [8]. Cross-dataset evaluation on GasHisSDB and SEED confirms superior generalization of the Transformer-based hybrid, reflecting the effectiveness of stain normalization and optimized training strategies.

Overall, the DenseNet121–MPCNN-TAO-SE framework provides strong multi-scale feature learning. However, it shows limitations in handling Stage II misclassification. In contrast, the Hybrid DenseNet121–Transformer model provides a more robust solution. It reduces Stage II misclassification by leveraging global self-attention. This approach improves the model's ability to capture long-range tissue relationships. As a result, the proposed hybrid model achieves a better balance of accuracy, interpretability, and generalization. This makes it more suitable for clinical GC diagnosis.

V. CONCLUSION

This study shows that advanced deep learning architectures, particularly DenseNet121-based models, substantially increase early prediction accuracy and dependability for GC detection. Fine-tuned CNNs such as DenseNet121 and MobileNetV2 achieved excellent recall for Stage I carcinoma and perfect Normal-class accuracy, making them highly suitable for early screening. However, all standard CNNs struggled with Stage II discrimination due to subtle histological overlap. To overcome this, two enhanced hybrid models were introduced. The DenseNet121–MPCNN-TAO–SE framework improved multi-scale feature learning and global contextual understanding, while the Hybrid DenseNet121–Transformer model—with stain normalization and focal loss—achieved 90.2% accuracy, a macro F1-score of 91.4%, and an AUC of 0.985, outperforming all CNN-only baselines. Both hybrids reduced Stage I–Stage II confusion and provided better interpretability through attention maps. Overall, combining CNN features with SE, multi-path convolution, and Transformer attention offers a powerful and clinically promising solution for early GC diagnosis, although balanced datasets and careful fine-tuning remain essential for robust performance.

FUNDING

This research work is not funded by any organization.

CONFLICTS OF INTEREST STATEMENT

The author(s) declare that they have no conflicts of interest to report regarding the present study.

DATA AVAILABILITY STATEMENT

The article includes data supporting this study's conclusions. The author can provide further data upon request.

ACKNOWLEDGMENTS

The authors thank Dr. Shirish N D, Assistant Professor, Department of Medical Gastroenterology, Government Medical College, Thoothukudi, India, for sharing gastric cancer histopathology pictures for this study. The authors also thank the stomach cancer experts and clinicians for enriching this research's quality and context. This work was completed thanks to his help.

REFERENCES

[1] M. B. Mourato *et al.*, "Effectiveness of gastric cancer endoscopic screening in intermediate-risk countries: Protocol for a systematic review and meta-analysis," *JMIR Res. Protoc.*, vol. 14, 2025, DOI: 10.2196/56791, [Online].

[2] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021, DOI: 10.3322/caac.21660.

[3] E. Van Cutsem *et al.*, "Gastric cancer," *Lancet*, vol. 388, no. 10060, pp. 2654–2664, 2016, DOI: 10.1016/S0140-6736(16)30354-3.

[4] J. A. Ajani *et al.*, "Gastric adenocarcinoma," *Nat. Rev. Dis. Primers*, vol. 3, Art. no. 17036, Jun. 2017, DOI: 10.1038/nrdp.2017.36.

[5] M. Ilić and I. Ilić, "Epidemiology of stomach cancer," *World J Gastroenterol*, vol. 28, no. 12, pp. 1187–1203, Mar. 2022, DOI: 10.3748/wjg.v28.i12.1187.

[6] O. Ciga *et al.*, "Overcoming the limitations of patch-based learning to detect cancer in whole slide images," *Sci. Rep.*, vol. 11, Art. no. 8894, 2021, DOI: 10.1038/s41598-021-88494-z.

[7] D. Komura, M. Ochi, and S. Ishikawa, "Machine learning methods for histopathological image analysis: Updates in 2024," *Comput. Struct. Biotechnol. J.*, vol. 27, pp. 383–400, Dec. 2024, DOI: 10.1016/j.csbj.2024.12.033.

[8] Z. Song *et al.*, "Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning," *Nat. Commun.*, vol. 11, Art. no. 4294, 2020, DOI: 10.1038/s41467-020-18147-8.

[9] A. Tofighi Bagheri, A. Ahmadi, and H. Mosadegh, "Improving lung cancer detection via MobileNetV2 and stacked-GRU with explainable AI," *Int. J. Inf. Tecnol.*, vol. 17, pp. 1189–1196, 2025, DOI: 10.1007/s41870-024-02045-z.

[10] X. Zheng *et al.*, "A deep learning model and human–machine fusion for prediction of EBV-associated gastric cancer from histopathology," *Nat. Commun.*, vol. 13, no. 1, Art. no. 2790, 2022, DOI: 10.1038/s41467-022-30459-5.

[11] A. F. Hussein, A. Q. Al-Neami, and N. K. Al-Qazzaz, "Transfer learning and hybrid deep convolutional neural networks for detection and classification of gastrointestinal diseases," *Int. J. Inf. Tecnol.*, 2025, DOI: 10.1007/s41870-025-02953-8.

[12] J. N. Kather *et al.*, "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nat. Med.*, vol. 25, pp. 1054–1056, 2019, DOI: 10.1038/s41591-019-0462-y.

[13] M. Othmani *et al.*, "Hybrid active shape model and deep neural network approach for lung cancer detection," *Int. J. Inf. Tecnol.*, vol. 17, pp. 4695–4706, 2025, DOI: 10.1007/s41870-024-01853-7.

[14] K. Zhang *et al.*, "Early gastric cancer detection and lesion segmentation based on deep learning and gastroscopic images," *Sci. Rep.*, vol. 14, Art. no. 7847, Apr. 2024, DOI: 10.1038/s41598-024-58361-8.

[15] J. Ma *et al.*, "Segment anything in medical images," *Nat. Commun.*, vol. 15, no. 1, Art. no. 654, Jan. 2024, DOI: 10.1038/s41467-024-44824-z.

[16] H. B. Mahbub *et al.*, "Invasive ductal carcinoma (IDC) detection in breast histopathology images using enhanced transfer learning of

convolutional neural networks," *Int. J. Inf. Tecnol.*, 2025, DOI: 10.1007/s41870-025-02522-z.

[17] S. Kaur, M. Kaur, and A. Khanna, "ConvNeXt based prognostication of axillary lymph node involvement in breast cancer based on smart CT interpretation," *Int. J. Inf. Tecnol.*, vol. 18, pp. 471–478, 2026, DOI: 10.1007/s41870-025-02833-1.

[18] O. M. Mirza *et al.*, "Computer aided diagnosis for gastrointestinal cancer classification using hybrid rice optimization with deep learning," *IEEE Access*, vol. 11, pp. 76321–76329, 2023, DOI: 10.1109/ACCESS.2023.3297441.

[19] A. Thakur *et al.*, "Deep learning approaches for detecting malignant melanoma in dermoscopic imagery," *Int. J. Inf. Tecnol.*, vol. 18, no. 1, pp. 91–109, Jul. 2025, DOI: 10.1007/s41870-025-02651-5.

[20] M. P. Yong *et al.*, "Histopathological gastric cancer detection on GasHisSDB dataset using deep ensemble learning," *Diagnostics (Basel)*, vol. 13, no. 10, Art. no. 1793, May 2023, DOI: 10.3390/diagnostics13101793.

[21] W. Hu *et al.*, "GasHisSDB: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer," *Comput. Biol. Med.*, vol. 142, Art. no. 105207, Mar. 2022, DOI: 10.1016/j.compbiomed.2021.105207.

[22] T. Wang, *SEED-gastric carcinoma dataset*. Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/wangtyi/seedgastric-carcinoma-dataset, Accessed on: Oct. 28, 2023.

[23] D. S. Shetty Radhika, "Gastric Cancer," *GitHub repository, [Dataset]*. Available: https://github.com/ssradhika/Gastric_cancer_Dataset, Accessed on: Jul. 26, 2025.

[24] O. Iizuka *et al.*, "Deep learning models for histopathological classification of gastric and colonic epithelial tumours," *Sci. Rep.*, vol. 10, Art. no. 1504, 2020, DOI: 10.1038/s41598-020-58467-9.

[25] J. Ma *et al.*, "Interpretable deep learning for gastric cancer detection: A fusion of AI architectures and explainability analysis," Front. Immunol., vol. 16, Art. no. 1596085, 2025, DOI: 10.3389/fimmu.2025.1596085.