Journal of Artificial Intelligence and Technology, (Ahead of Print)
https://doi.org/10.37965/jait.2026.1383

ISTP
RESEARCH ARTICLE

MoVi-HER2: A MobileNetV3-ViT Fusion Network for HER2 Status Prediction in Gastroesophageal Adenocarcinoma from Tissue Microarray Images

Untari Novia Wisesty,[1,2] Kurniawan Nur Ramadhani,[1,2] Gia Septiana Wulandari,[1,3] Aaz Muhammad Hafidz Azis,[1,4] Febryanti Sthevanie,[1,2] Isman Kurniawan,[1,2] Lidya Ningsih,[1,4] Eva Nurmala,[6] and Rita Rismala[1,5]

[1]School of Computing, Telkom University, Bandung, Indonesia
[2]Center of Excellence of Artificial Intelligence for Learning and Optimization, Telkom University, Bandung, Indonesia
[3]CoE Social Wellness and Data Analytics, Telkom University, Bandung, Indonesia
[4]CoE Inspiring Digital Transformation for Social Innovation, Telkom University, Bandung, Indonesia
[5]Center of Excellence of Advanced ICT Infrastructure and Services, Telkom University, Bandung, Indonesia
[6]Directorate of Human Resources, Telkom University, Bandung, Indonesia

(Received 01 May 2026; Revised 24 July 2026; Accepted 12 August 2026; Published online 29 August 2026)

Abstract: Accurate assessment of HER2 status is critical for therapeutic decision-making in gastroesophageal adenocarcinoma (GEA), yet manual evaluation of HER2 immunohistochemistry (IHC) remains labor-intensive and prone to interobserver variability. This research proposes MobileNetV3-ViT Fusion Network (MoVi-HER2), a deep learning model designed to automatically predict HER2 status and IHC scores from tissue microarray (TMA) images by integrating local and global feature representations. The model fuses MobileNetV3 as a convolutional backbone for fine-grained texture extraction with Vision Transformer (ViT) for contextual spatial reasoning, combined through concatenation and addition strategies followed by dense refinement layers for classification. Performance was evaluated on a public TMA dataset of GEA stained for HER2 expression, using macro F1-score, weighted F1-score, and balanced accuracy. MoVi-HER2 achieved superior performance across all metrics, obtaining a macro F1-score of 0.8531 and balanced accuracy of 0.9794 for HER2 status prediction, and 0.8119 and 0.9370 for IHC score prediction, surpassing MobileNetV3, ViT, Xception, EfficientNetV2, TinyViT, and EfficientFormerV2. Notably, these results were achieved without complex preprocessing or augmentation, indicating strong generalization and stain invariance. Fusion attention visualizations further confirmed that the model focuses on membrane-rich regions consistent with clinical HER2 scoring criteria, demonstrating its potential as an accurate, robust, and interpretable tool for AI-assisted digital pathology in GEA diagnosis.

Keywords: fusion network; gastroesophageal adenocarcinoma; HER2 status; MobileNetV3; ViT

I. INTRODUCTION

Gastroesophageal adenocarcinoma (GEA) is a malignancy arising at the gastroesophageal junction (GEJ), originating from glandular epithelial cells. Its incidence has risen significantly in industrialized nations, driven by risk factors including central obesity, high-fat diet, smoking, and chronic gastroesophageal reflux leading to Barrett's esophagus. GEA predominantly affects men aged 60 years and older [1–4]. Clinically, GEA presents with nonspecific early symptoms, resulting in frequent late-stage diagnosis. It ranks sixth among cancer-related deaths globally, with approximately 500,000 new cases and 400,000 deaths annually, and a five-year survival rate below 25% [5–8].

HER2 overexpression or amplification occurs in 10–30% of GEA cases, particularly at the GEJ and in intestinal-type histology. As a tyrosine kinase receptor regulating cell proliferation and survival, HER2 amplification activates oncogenic pathways such as PI3K/AKT and MAPK, promoting tumor aggressiveness and metastasis [9–12]. The ToGA trial established trastuzumab as the first targeted therapy for HER2-positive advanced GEA, significantly improving overall survival in patients with immunohistochemistry (IHC) scores of 2+ or 3+ [13].

Despite HER2 testing becoming standard practice, its implementation remains challenged by tumor heterogeneity, variability in tissue fixation, and significant interobserver inconsistency in IHC scoring, particularly in distinguishing scores of 1+, 2+, and 3+. In equivocal IHC 2+ cases, confirmatory in situ hybridization (ISH) is required, yet access to this technology remains unequal across healthcare facilities [14–17].

Deep learning has emerged as a noninvasive approach for automated HER2 prediction from histopathology images, with methods ranging from convolutional neural networks (CNNs) to attention-based multiple instance learning frameworks [18–20]. However, existing approaches often face a trade-off between capturing fine-grained local morphology and modeling global tissue

Corresponding author: Untari Novia Wisesty (e-mail: untarinw@telkomuniversity.ac.id).

context, as they typically rely on a single architectural paradigm. To address these limitations, the proposed framework integrates two complementary architectures. MobileNetV3 is a lightweight CNN that employs depthwise separable convolutions and squeeze-and-excitation (SE) modules to efficiently capture fine-grained local features such as texture, edges, and cellular morphology, making it well suited for resource-constrained settings. The Vision Transformer (ViT), by contrast, partitions an image into fixed-size patches and employs multi-head self-attention to represent long-range spatial relationships, allowing it to perceive global tissue organization and macrostructural context that convolutional models frequently miss. By combining these two architectures, MoVi-HER2 leverages the local efficiency of MobileNetV3 and the global contextual reasoning of ViT to overcome the local–global trade-off inherent in prior single-architecture approaches.

Building on this motivation, this research proposes MoVi-HER2, a fusion network for automated HER2 status and IHC score prediction from GEA tissue microarray (TMA) images. MobileNetV3 serves as a lightweight local feature extractor capturing texture, edges, and cellular morphology, while ViT models global spatial relationships via self-attention. Their integration yields a model that is accurate, generalizable across staining and scanner variations, and interpretable through fusion attention visualizations, supporting clinical transparency in HER2 classification decisions.

The key contributions of this study are threefold. First, we present a deep learning model for automated HER2 status and IHC score prediction from GEA TMA images, extending HER2 prediction research beyond its predominant focus on breast cancer. Second, we propose a novel MobileNetV3-ViT fusion architecture that integrates local morphological features with global spatial context for improved prediction accuracy and robustness. Third, the model provides enhanced interpretability via fusion attention visualizations, enabling pathologists to verify tissue regions contributing to classification decisions and promoting clinical trust.

II. RELATED WORK

Research on automated HER2 status prediction in gastric and gastroesophageal cancers has employed diverse approaches, broadly categorized into CNN-based methods on H&E slides and IHC-based or attention-based frameworks. For H&E-based prediction, Zhang *et al.* proposed CLAM_Sim, an attention-based MIL model with contrastive learning for gastric adenocarcinoma, achieving an AUC of 0.847 on surgical samples and 0.833 for trastuzumab response prediction; however, performance dropped on biopsy samples (AUC 0.723), indicating sensitivity to tissue quality [18]. Liao *et al.* introduced HER2Net, a semi-quantitative pixel-level model trained on 531 internal whole slide images (WSIs) and validated on 102 multi-center WSIs, achieving an AUROC of 0.9769 by quantifying HER2 high-expression regions rather than relying on binary tile classification, thereby better mimicking the clinical interpretation process [21].

For IHC-based and attention-based frameworks, Han *et al.* developed a two-stage pipeline combining a RepVGG-based tile-level classifier with a WSI-level predictor (SVM/MLP), reaching 94% accuracy without requiring H&E matching, although validation was limited to a single-center dataset of 183 WSIs [19]. Pisula *et al.* presented an attention-based MIL model for IHC slide evaluation, achieving balanced accuracies of 0.8249 for IHC scoring and 0.9429 for HER2 status with interpretability through attention maps; nevertheless, challenges remained in handling HER2 expression heterogeneity, data imbalance for IHC 2+ cases, and substantial computational infrastructure requirements [20].

Despite these advances, several gaps remain that this study aims to address. First, most existing studies focus on gastric cancer broadly rather than GEA specifically, which exhibits distinct biological and morphological characteristics. Second, CNN- and MIL-based approaches often struggle to balance local feature precision with global spatial understanding, as they typically rely on a single architectural paradigm. Third, few studies provide interpretable predictions that align with the membrane-based criteria used in clinical HER2 assessment, limiting their transparency and clinical trust. To address these gaps, this study proposes MoVi-HER2, a fusion network combining the local feature extraction of MobileNetV3 with the global contextual modeling of ViT, specifically targeting HER2 status and IHC score prediction from GEA TMA images.

III. PROPOSED METHOD

This section describes the proposed model, MoVi-HER2, a MobileNetV3-ViT fusion network, to predict HER2 status and IHC score in GEA. The proposed model integrates the lightweight CNN MobileNetV3 with the transformer-based architecture ViT to effectively extract and fuse both local and global features from TMA images. The proposed fusion network aims to improve classification accuracy and interpretability in HER2 assessment, addressing the limitations of traditional CNN-based methods that often fail to capture long-range spatial dependencies within histopathological images.

A. DATA SPECIFICATION

This research employs the HER2 overexpression in GEA public dataset [20], which provides digitized TMA cores of GEA cases stained with IHC for HER2 protein expression. Each TMA core was manually annotated and scored by expert pathologists following established clinical HER2 scoring criteria, with four IHC score labels (0, 1, 2, and 3) corresponding to negative (0 and 1) and positive (3) HER2 status, while IHC score 2 requires further fluorescence ISH analysis. These annotations serve as the ground truth for supervised learning and model evaluation. Images are provided in high-resolution WSI format or as extracted patches from TMA cores.

The dataset is provided by its original source with a predefined division into training and test sets, ensuring a representative distribution of HER2 expression classes. Notably, the test set was collected later than the training data and originates from a different experimental batch, introducing natural variability in staining intensity, slide preparation, and imaging conditions, providing a realistic and challenging evaluation scenario for assessing model generalization. In this study, the predefined test set was reserved exclusively for final evaluation, while the training set was further partitioned into 80% training and 20% validation subsets. The validation set was used for hyperparameter tuning and model selection while keeping the test set fully unseen during development to prevent overfitting and ensure reliable performance evaluation. The resulting data splits are presented in Table I.

B. DATA AUGMENTATION AND PREPROCESSING

The original dataset exhibited strong class imbalance in both tasks: IHC 0 dominated with 1,093 samples versus only 33, 59, and 93 samples for IHC 1, 2, and 3, respectively, while HER2-negative

Table I. Data distribution for IHC score and HER2 status prediction tasks

	Before augmentation			After augmentation		
Class	**Train**	**Validation**	**Test**	**Train**	**Validation**	**Test**
IHC score prediction						
IHC 0	1093	274	284	1093	274	284
IHC 1	33	9	-	231	9	-
IHC 2	59	14	15	236	14	15
IHC 3	93	23	3	279	23	3
Subtotal	1278	320	302	1839	320	302
HER2 status prediction						
HER2-negative	1140	286	291	1140	286	291
HER2-positive	138	34	16	138	34	16
Subtotal	1278	320	307	1278	320	307

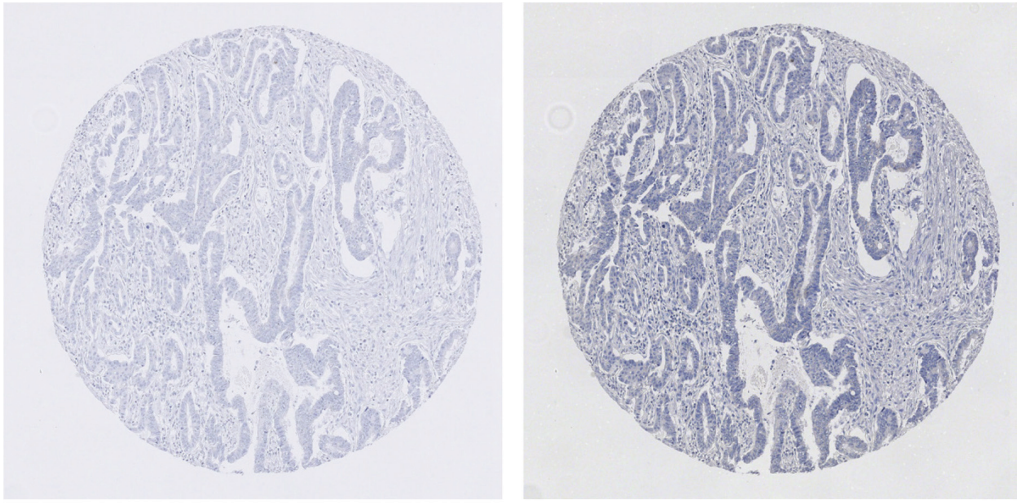
samples (1,140) far outnumbered HER2-positive samples (138). To address this, augmentation transformations were applied to the minority classes, grouped into geometric transformations (random rotation, flipping, and shear) and color transformations (brightness and contrast adjustment, color jitter, and Gaussian noise), selected to increase sample diversity while preserving essential histopathological characteristics. Post-augmentation, IHC 1, 2, and 3 expanded to 231, 236, and 279 samples, and the HER2-positive class to 690 samples (Table I).

To systematically evaluate the effect of preprocessing on model performance, a complete preprocessing pipeline was developed and decomposed into four stages that were later assessed individually in an ablation study: (1) resizing each RGB image to a maximum dimension of 1024×1024 pixels while preserving aspect ratio; (2) isolating tissue regions by converting images to HSV color space with thresholds HSV_LOWER = (0, 5, 5) and HSV_UPPER = (180, 255, 250), followed by morphological opening and closing using a 3×3 elliptical kernel, retaining only images with tissue coverage $\geq 0.3\%$ and color quality ≥ 0.4; (3) applying CLAHE independently to each RGB channel (clip limit = 2.0, tile grid size = 8×8) [22,23] to minimize inter-slide color variability; and (4) replacing background pixels outside the tissue mask with a uniform light-gray value [240,240,240] to reduce background bias. The visual results of the preprocessing pipeline are illustrated in Fig. 1.

C. MOBILENetV3-VIT FUSION NETWORK

The proposed MoVi-HER2 unites the distinct advantages of CNNs and ViTs for accurate and interpretable prediction of HER2 status and IHC scores from high-resolution TMA images of GEA (Fig. 2). The architecture captures both fine-grained cellular textures and global structural dependencies, addressing the limitations of models relying solely on local or global representations. The network takes as input a preprocessed RGB TMA image of size $1024 \times 1024 \times 3$. Both backbone networks, MobileNetV3-Large and ViT-Base, are initialized with ImageNet pretrained weights, leveraging transfer learning for generalized low-level features (edges, textures, and contours) and mid-to-high-level semantic patterns proven transferable to medical imaging. A two-stage fine-tuning strategy is employed: in the first stage, the pretrained backbones are frozen while only the adapter and classification layers are trained; in the second stage, selected backbone layers are progressively unfrozen in a top-down manner to enable domain adaptation to TMA-specific patterns such as glandular architecture, cell clusters, and staining textures.

The MobileNetV3-Large backbone serves as the CNN component for capturing localized and fine-scale texture information through depthwise separable convolutions and SE modules, making it lightweight yet effective in modeling nuclear boundaries, cytoplasmic intensity, and intra-tissue variations [24–27]. A global average pooling layer aggregates the spatial feature maps into a 1280-dimensional vector, which is then refined through a dense layer (1280→512) to produce the CNN-based embedding. In parallel, the ViT-Base backbone models global contextual dependencies by dividing the input into 16×16 patches, linearly projecting them into patch embeddings augmented with positional encodings, and passing them through multi-head self-attention transformer blocks [28–31]. This encoder-only transformer design has been shown effective in attention-guided classification tasks where interpretability and computational efficiency are essential



(a) (b)

Fig. 1. Preprocessing result of TMA image: (a) original image and (b) image after preprocessing.

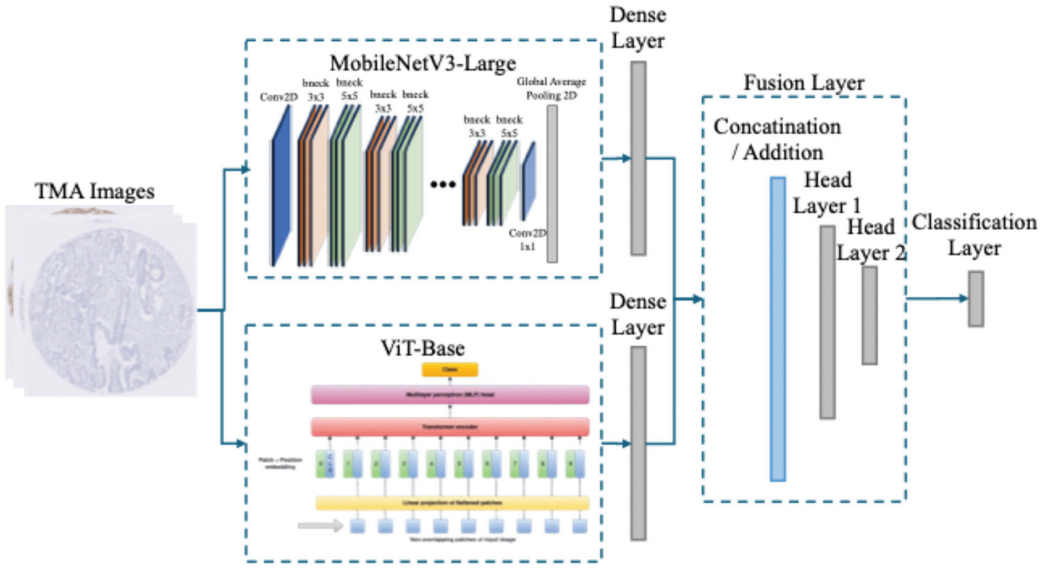


Fig. 2. The proposed MobileNetV3–ViT Fusion Network (MoVi-HER2) for HER2 status prediction of GEA.

[32], leveraging multi-head scaled dot-product attention to selectively focus on diagnostically relevant regions. The [CLS] token from the final encoder, representing long-range dependencies such as tissue organization, gland formation, and stromal alignment, is then projected through a dense layer (768→512) for dimensional alignment with the CNN branch. The use of attention mechanisms in this design supports clinical interpretability by enabling identification of tissue regions that contribute most to HER2 classification decisions, reinforcing transparency and trust in the model's predictions.

To exploit the complementary representations from both branches, a feature fusion module is introduced. Two fusion strategies were explored: concatenation, which combines both 512-dimensional embeddings into a 1024-dimensional vector to retain maximum feature diversity, and addition, which performs element-wise summation to produce a compact 512-dimensional vector with lower computational cost but assumes feature homogeneity. Layer Normalization (LayerNorm) is applied before fusion to ensure scale compatibility. The fused 1024-dimensional vector is then refined through two head layers: Head Layer 1 with a dense layer (1024→512), LayerNorm, GELU activation, and dropout (0.2), and Head Layer 2 with a dense layer (512→128) using the same configuration to produce a 128-dimensional discriminative embedding. Finally, the classification layer adapts to each task: a dense layer (128→4) with softmax for IHC score prediction (classes 0, 1, 2, 3), and a dense layer (128→2) with softmax for binary HER2 status prediction (negative and positive). This fusion design bridges CNN-based local pixel-level analysis with transformer-based image-wide relational learning, providing a robust, scalable, and interpretable framework for computational pathology.

IV. RESULTS AND DISCUSSION

This section presents the experimental results of the proposed MoVi-HER2 for HER2 status and IHC score prediction in GEA. The experiments were designed to evaluate (1) the effect of preprocessing on model performance, (2) the impact of data augmentation on class balance and generalization, and (3) the comparative performance against baseline architectures including standalone MobileNetV3, ViT, and conventional CNN and ViT models. To ensure fair comparison and stable convergence, a consistent hyperparameter configuration was applied across all experiments: dropout rate of 0.2, initial learning rate of 1×10^{-5}, weight decay of 1×10^{-5}, gradient clipping threshold of 1.0, and 50 training epochs. A Reduce-on-Plateau learning rate scheduler was employed in 'max' mode, monitoring the validation F1-score with patience = 3, reduction factor of 0.5, and a lower bound of 1×10^{-6} to maintain convergence stability and enable effective fine-tuning near local minima.

Three primary evaluation metrics were used: macro F1-score to measure uniform classification performance across all classes (emphasizing minority class), weighted F1-score to provide a class-frequency-adjusted estimate of overall performance, and balanced accuracy to define the average of sensitivity and specificity across all classes, to assess robustness under class imbalance and ensure equitable evaluation of the model's discriminative reliability for each HER2 expression level. All performance metrics reported in this section were computed exclusively on the test set, which was not used during training or model selection. The validation set was used only for hyperparameter tuning and selecting the best model configuration, while the test set remained completely unseen throughout development.

A. EFFECT OF PREPROCESSING AND AUGMENTATION

To analyze the impact of preprocessing and data augmentation on model performance, controlled experiments were conducted across three architectures (MobileNetV3, ViT, and the proposed MoVi-HER2) using three configurations: (1) resize only at 1024×1024 pixels, (2) all preprocessing steps (resizing, tissue masking, stain normalization, and background replacement), and (3) with augmentation (random rotation and flipping, brightness and contrast adjustment, color jittering, mild shear, Gaussian noise, and mild elastic deformation). This focused comparison was designed as an ablation-style analysis to isolate the contribution of the fusion mechanism by evaluating MoVi-HER2 against its two constituent

Table II. Preprocessing and data augmentation experimental results

Model	Configuration	IHC Macro F1	IHC Weighted F1	IHC Balanced Acc.	HER2 Macro F1	HER2 Weighted F1	HER2 Balanced Acc.
MobileNetV3	Resize only	0.7165	0.9250	0.9361	0.8334	0.9656	0.8630
MobileNetV3	All preprocessing	0.6299	0.8972	0.8840	0.7969	0.9486	0.9674
MobileNetV3	Augmentation	0.6840	0.8967	0.9433	0.6734	0.8945	0.9261
ViT	Resize only	0.7376	0.9121	0.9502	0.8441	0.9633	0.9777
ViT	All preprocessing	0.6679	0.8979	0.9222	0.7541	0.9378	0.9014
ViT	Augmentation	0.6257	0.9004	0.8919	0.7969	0.9486	0.9674
MoVi-HER2 (Addition)	Resize only	**0.8202**	**0.9487**	0.9620	**0.8531**	**0.9658**	**0.9794**
MoVi-HER2 (Addition)	All preprocessing	0.6125	0.8867	0.9005	0.7479	0.9355	0.8997
MoVi-HER2 (Addition)	Augmentation	0.7055	0.9020	0.9456	0.7969	0.9486	0.9674
MoVi-HER2 (Concatenation)	Resize only	0.7867	0.9303	**0.9699**	**0.8531**	**0.9658**	**0.9794**
MoVi-HER2 (Concatenation)	All Preprocessing	0.7541	0.9383	0.9364	0.7823	0.9455	0.9361
MoVi-HER2 (Concatenation)	Augmentation	0.7210	0.9070	0.9479	0.7469	0.9300	0.9536

backbones (MobileNetV3 and ViT) under identical preprocessing and augmentation conditions. For the IHC score task, the model was trained on three classes (IHC 0, 2, and 3) since the test set contained no IHC 1 samples. Results are summarized in Table II.

The findings indicate that the baseline configuration yielded the highest performance across most models. The MoVi-HER2 with addition fusion achieved the best overall results under baseline, with a macro F1-score of 0.8202, weighted F1-score of 0.9487, and balanced accuracy of 0.9620 for IHC score prediction; and macro F1-score of 0.8531, weighted F1-score of 0.9658, and balanced accuracy of 0.9794 for HER2 status prediction. The MoVi-HER2 with concatenation fusion showed a similar trend, achieving the highest balanced accuracy of 0.9699 for IHC score prediction (macro F1 0.7867, weighted F1 0.9303) and matching performance for HER2 status prediction (macro F1 0.8531, balanced accuracy 0.9794).

In contrast, both the full preprocessing pipeline and data augmentation generally led to performance degradation. Under full preprocessing, MoVi-HER2 (Addition) declined to a macro F1-score of 0.6125 for IHC and 0.7479 for HER2 status, while augmentation reduced its performance to 0.7055 (IHC) and 0.7969 (HER2). A similar degradation pattern was also observed for MoVi-HER2 (Concatenation) and the MobileNetV3 and ViT baselines. These observations suggest that conventional preprocessing and augmentation strategies, while beneficial in many computer vision tasks, may interact unfavorably with histopathological features critical to HER2 discrimination.

Across all configurations, the proposed MoVi-HER2 consistently outperformed both of its standalone constituent backbones (MobileNetV3 and ViT), confirming that the fusion of CNN-based local feature extraction and transformer-based global context modeling provides complementary strengths beyond what either component achieves individually. The consistently high balanced accuracy (>0.96) across fusion configurations further demonstrates reliable performance across both HER2 categories and IHC scores, while the relative robustness of MoVi-HER2 under augmentation indicates that the fusion representation effectively mitigates the negative effects of unnecessary transformations.

1). ABLATION STUDY ON PREPROCESSING COMPONENTS. To investigate which specific preprocessing components contribute to the performance degradation observed in the previous experiment, an ablation study was conducted on the proposed MoVi-HER2 model under three progressively complex preprocessing configurations: (1) resize only (baseline), (2) resize with CLAHE-based stain normalization, and (3) all preprocessing steps (resize, CLAHE, tissue masking, and background replacement). To eliminate confounding effects, no data augmentation was applied in this experiment. The results are summarized in Table III.

The findings reveal that performance progressively declined as more preprocessing steps were added. For MoVi-HER2 (Addition) on the IHC score task, the macro F1-score dropped from 0.8202 under resize only to 0.7228 with CLAHE, and further to 0.6125 under full preprocessing. A consistent pattern was observed for the HER2 status task, where the macro F1-score decreased from 0.8531 to 0.7962 and then to 0.7479 across the same configurations. The concatenation fusion variant exhibited a similar trend, with its IHC macro F1-score declining from 0.7867 to 0.6635 under CLAHE, and its HER2 macro F1-score from 0.8531 to 0.8192 and then to 0.7823 under full preprocessing.

The degradation introduced by CLAHE can be explained by the nature of HER2 IHC scoring, which relies on the absolute intensity and completeness of membrane staining, ranging from weak, incomplete staining (IHC 1+) to strong, complete circumferential staining (IHC 3+). CLAHE operates by redistributing pixel intensity histograms locally within small tile regions and equalizing contrast across them. While this enhances local contrast for general image analysis, in HER2 IHC it distorts the very intensity gradients that encode diagnostic information that can amplify weak background or stromal staining while compressing the intensity gap between moderate and strong membrane staining. As a result, the discriminative color and intensity cues that differentiate HER2 expression levels are partially lost, reducing the model's ability to distinguish between adjacent IHC categories. The further decline under full preprocessing indicates that tissue masking and background replacement introduce additional information loss, likely by discarding peritumoral context or creating artificial boundaries at the mask edges. Overall, the ablation confirms that the model performs best when learning directly from minimally processed images, supporting the use of resize-only preprocessing in subsequent experiments.

2). ABLATION STUDY ON DATA AUGMENTATION COMPONENTS. To address whether specific augmentation types are responsible for the observed performance degradation, a controlled

Table III. Ablation study on preprocessing components for MoVi-HER2

Model	Preprocessing	IHC Macro F1	IHC Weighted F1	IHC Balanced Acc.	HER2 Macro F1	HER2 Weighted F1	HER2 Balanced Acc.
MoVi-HER2 (Addition)	Resize only	**0.8202**	**0.9487**	0.962	**0.8531**	**0.9658**	**0.9794**
MoVi-HER2 (Addition)	Resize +CLAHE	0.7228	0.9142	0.9192	0.7962	0.952	0.9117
MoVi-HER2 (Addition)	All preprocessing	0.6125	0.8867	0.9005	0.7479	0.9355	0.8997
MoVi-HER2 (Concatenation)	Resize only	0.7867	0.9303	**0.9699**	0.8531	0.9658	0.9794
MoVi-HER2 (Concatenation)	Resize +CLAHE	0.6635	0.9201	0.8595	0.8192	0.9558	0.9725
MoVi-HER2 (Concatenation)	All preprocessing	0.7541	0.9383	0.9364	0.7823	0.9455	0.9361

ablation study was conducted by separating the augmentation operations into distinct groups: (1) without augmentation (baseline), (2) geometric augmentation (rotation, flipping, and shear), (3) color augmentation (brightness and contrast adjustment, color jitter, and Gaussian noise), and (4) all augmentations combined. All experiments used resize-only preprocessing to isolate the effect of augmentation and were performed on both fusion variants of MoVi-HER2. The results are summarized in Table IV.

The ablation reveals that the effect of augmentation is highly dependent on both the task and the fusion strategy. For the IHC score task, MoVi-HER2 (Addition) achieved its highest macro F1-score under color augmentation (0.7603), exceeding geometric augmentation (0.7358), although neither surpassed the resize-only baseline (0.8202). Geometric augmentation consistently produced weaker results for the IHC task, reducing the concatenation variant's macro F1-score to 0.6807. For the HER2 status task, both fusion variants achieved their best performance without augmentation, reaching a macro F1-score of 0.8531 and balanced accuracy of 0.9794. Among the augmentation settings, geometric augmentation was the least detrimental for the addition variant (macro F1 0.8120), whereas for the concatenation variant it caused the largest drop, reducing the macro F1-score to 0.6936.

A key observation is that no single augmentation strategy consistently improved performance over the resize-only baseline across both tasks. Geometric transformations tended to increase false positives in minority classes, particularly degrading the concatenation variant, while color-based perturbations had a milder and occasionally beneficial effect for IHC scoring. This task-dependent and fusion-dependent behavior explains why the combined augmentation pipeline used in the initial experiment led to overall degradation: the simultaneous application of geometric and color transformations introduced conflicting effects that obscured the subtle morphological and staining patterns critical for HER2 discrimination. These findings demonstrate that conventional augmentation strategies, though effective in general computer vision, must be applied selectively in histopathological analysis, and justify the resize-only configuration adopted for the final model.

B. EFFECT OF IHC SCORE CLASS CONFIGURATION

In the original dataset, the test set contains samples from only three IHC score categories (IHC 0, 2, and 3), as no IHC 1 samples were available for testing, while the training set includes all four categories (IHC 0, 1, 2, and 3). This discrepancy introduces a methodological consideration regarding whether the model should be trained on three classes to match the test set or on four classes to utilize all available training data, including the unseen IHC 1 category. To examine the effect of this class configuration on model performance, experiments were conducted on the proposed MoVi-HER2 under both settings, using resize-only preprocessing without augmentation to ensure a fair comparison. The results are summarized in Table V.

The findings reveal that the optimal class configuration depends on the fusion strategy. MoVi-HER2 (Addition) achieved its best performance under the three-class configuration, with a macro

Table IV. Ablation study on data augmentation components for MoVi-HER2

Model	Preprocessing	IHC Macro F1	IHC Weighted F1	IHC Balanced Acc.	HER2 Macro F1	HER2 Weighted F1	HER2 Balanced Acc.
MoVi-HER2 (Addition)	Without Augmentation	**0.8202**	**0.9487**	0.9620	**0.8531**	**0.9658**	**0.9794**
MoVi-HER2 (Addition)	Geometric Augmentation	0.7358	0.8977	0.9098	0.8120	0.9552	0.9430
MoVi-HER2 (Addition)	Color Augmentation	0.7603	0.9232	0.9239	0.8042	0.9527	0.9413
MoVi-HER2 (Addition)	All Augmentation	0.7055	0.9020	0.9456	0.7969	0.9486	0.9674
MoVi-HER2 (Concatenation)	Without Augmentation	0.7867	0.9303	**0.9699**	0.8531	0.9658	0.9794
MoVi-HER2 (Concatenation)	Geometric Augmentation	0.6807	0.9132	0.8560	0.6936	0.9055	0.9347
MoVi-HER2 (Concatenation)	Color Augmentation	0.6981	0.8966	0.9098	0.7966	0.9503	0.9395
MoVi-HER2 (Concatenation)	All Augmentation	0.7210	0.9070	0.9479	0.7469	0.9300	0.9536

F1-score of 0.8202, weighted F1-score of 0.9487, and balanced accuracy of 0.9620. However, its performance declined under the four-class configuration to a macro F1-score of 0.7913, weighted F1-score of 0.9026, and balanced accuracy of 0.9306. In contrast, MoVi-HER2 (Concatenation) exhibited the opposite trend, with its macro F1-score improving to 0.8119 (4-class), although its balanced accuracy decreased from 0.9699 to 0.9370.

This contrasting behavior reflects the different feature integration mechanisms of the two fusion strategies. The addition fusion compresses both branch representations through element-wise summation into a compact 512-dimensional vector, which may be more sensitive to the increased class complexity introduced by the additional IHC 1 category. The concatenation fusion, which preserves the full dimensional structure of both branches in a 1024-dimensional vector, appears to handle the additional class more effectively by retaining greater representational capacity. These observations indicate that training on all four IHC categories benefits the concatenation variant by enabling smoother decision boundaries between adjacent expression levels, while the addition variant achieves stronger discrimination when the class scope is restricted to those present in the testing set.

Considering both tasks, the concatenation fusion variant was selected for the final proposed model. Although both variants performed comparably on HER2 status prediction without augmentation (macro F1 0.8531, balanced accuracy 0.9794), the concatenation variant achieved higher performance on the more complex four-class IHC scoring task (macro F1 0.8119 vs. 0.7913 for the addition variant, as shown in Table V). By preserving the full dimensional structure of both branches in a 1024-dimensional fused vector, the concatenation variant retains greater representational capacity, enabling it to handle the increased class complexity more effectively and to provide more stable performance across both tasks.

C. PERFORMANCE COMPARISON ACROSS MODELS

To comprehensively evaluate the proposed MoVi-HER2, additional experiments were conducted to compare its performance against several established deep learning architectures spanning CNN-based, transformer-based, and hybrid lightweight designs: MobileNetV3, ViT, Xception, EfficientNetV2, TinyViT, and EfficientFormerV2 [29,33–35]. All models were trained on four IHC categories (0, 1, 2, and 3) for IHC score prediction and the corresponding binary task for HER2 status prediction. To isolate the effect of model architecture, no data augmentation was applied, and only image resizing to 1024×1024 pixels was used as preprocessing. The results are summarized in Table VI.

For the IHC score prediction task, the proposed MoVi-HER2 achieved the highest overall performance with a macro F1-score of 0.8119, weighted F1-score of 0.9180, and balanced accuracy of 0.9370, surpassing all comparative models. Among the other model, Xception showed the weakest overall performance with a macro F1-score of only 0.4980 and balanced accuracy of 0.6882, likely due to its lack of attention or channel-wise recalibration mechanisms, which limits its ability to adaptively emphasize discriminative features in histopathological textures. EfficientNetV2 (macro F1 0.7009, balanced accuracy 0.8904), TinyViT (0.7280, 0.9302), and EfficientFormerV2 (0.7017, 0.8819) achieved moderate performance, with their compact designs trading off some classification consistency for efficiency. Although ViT obtained the highest balanced accuracy (0.9558), its lower macro F1-score (0.7982) indicates less consistent classification across classes. MobileNetV3 yielded a competitive weighted F1-score (0.9366) but a weaker macro F1 (0.7597), reflecting its limited capacity to capture global structural context despite incorporating SE modules for local channel recalibration.

For the HER2 status prediction task, MoVi-HER2 again outperformed all comparative models across every metric, achieving a macro F1-score of 0.8531, weighted F1-score of 0.9658, and balanced accuracy of 0.9794. ViT followed closely (macro F1 0.8441, balanced accuracy 0.9777), validating the importance of global contextual understanding in HER2 recognition. Xception again recorded the lowest performance (macro F1 0.7415, balanced accuracy 0.8441), while TinyViT (0.7893, 0.9378) and EfficientFormerV2 (0.7861, 0.8544) performed moderately but inconsistently across metrics. In contrast, MobileNetV3 and EfficientNetV2 showed lower balanced accuracies (0.8630 and 0.9674, respectively) despite strong weighted F1-scores, indicating a tendency to overfit majority classes.

These quantitative comparisons demonstrate that MoVi-HER2 consistently outperforms both pure CNN-based architectures

Table V. Effect of IHC score class configuration on MoVi-HER2 performance

Model	3 class IHC score			4 class IHC score		
	Macro F1	Weighted F1	Balanced Acc.	Macro F1	Weighted F1	Balanced Acc.
MoVi-HER2 (Addition)	0.8202	0.9487	0.9620	0.7913	0.9026	0.9306
MoVi-HER2 (Concatenation)	0.7867	0.9303	0.9699	0.8119	0.9180	0.9370

Table VI. Performance comparison across models

Model	IHC Macro F1	IHC Weighted F1	IHC Balanced Acc.	HER2 Macro F1	HER2 Weighted F1	HER2 Balanced Acc.
Proposed model (MoVi-HER2)	**0.8119**	0.9180	0.9370	**0.8531**	**0.9658**	**0.9794**
MobileNetV3	0.7597	**0.9366**	0.9231	0.8334	0.9656	0.8630
ViT	0.7982	0.9154	**0.9558**	0.8441	0.9633	0.9777
Xception	0.498	0.9183	0.6882	0.7415	0.9384	0.8441
EfficientNetV2	0.7009	0.8969	0.8904	0.7969	0.9486	0.9674
TinyViT	0.7280	0.9063	0.9302	0.7893	0.9479	0.9378
EfficientFormerV2	0.7017	0.9264	0.8819	0.7861	0.9527	0.8544

(MobileNetV3, Xception, EfficientNetV2) and pure transformer-based or hybrid lightweight architectures (ViT, TinyViT, EfficientFormerV2) across all evaluation metrics. Notably, MoVi-HER2 achieved these results using only resized raw images, without complex preprocessing or augmentation. This highlights the model's intrinsic robustness and strong feature generalization. The balanced accuracy exceeding 0.97 for HER2 status prediction indicates that MoVi-HER2 maintains both high sensitivity and specificity, two properties essential for clinical decision support, particularly in identifying patients eligible for HER2-targeted therapies such as trastuzumab. The consistent superiority across both tasks confirms that the fusion of convolutional and transformer-based representations provides a powerful framework for histopathological biomarker analysis, maintaining high accuracy and balanced class performance under challenging real-world conditions. Beyond predictive accuracy, the practical deployability of the model also depends on its computational requirements, which are analyzed below.

In addition to predictive performance, the computational requirements of MoVi-HER2 were compared against all baseline architectures in terms of trainable parameters, model size, and inference time per image, as summarized in Table VII. As expected, MoVi-HER2 has a larger parameter count (93.3M) than the individual baselines, since it integrates both the MobileNetV3 (4.93M) and ViT (87.92M) backbones, with the overhead primarily attributable to the ViT component. However, this does not translate into proportionally higher inference latency: MoVi-HER2 achieved an inference time of 100.3 ms per image, which is faster than several baselines with fewer parameters, such as Xception (141.2 ms) and EfficientNetV2 (106.85 ms), supporting near-real-time inference. Overall, while MoVi-HER2 is not the most lightweight model in absolute terms, it achieves the highest predictive performance across all metrics with a competitive inference speed, representing a favorable accuracy and efficiency balance for HER2 assessment in digital pathology.

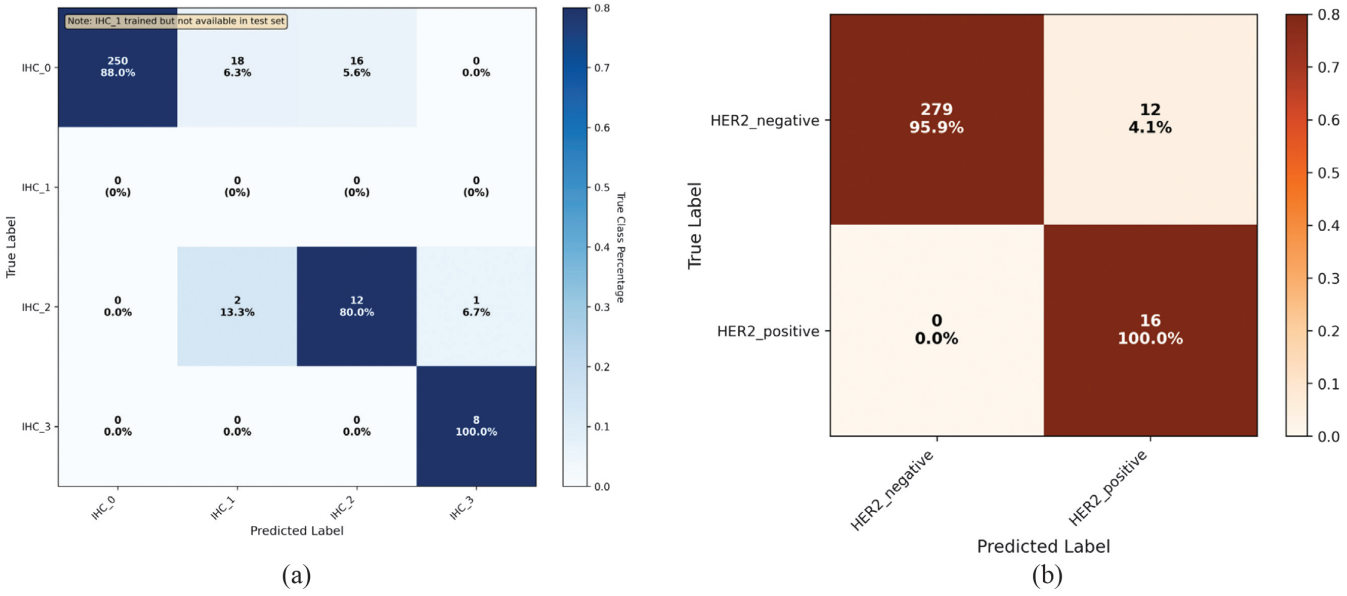
Table VII. Performance comparison across models

Model	Trainable params (M)	Model size (MB)	Inference time (ms)
MobileNetV3	4.93	19.8	13.66
ViT	87.92	351.67	73.35
Xception	38.98	156.68	141.2
EfficientNetV2	53.58	215.5	106.85
TinyViT	21.05	136.72	28.89
EfficientFormerV2	25.82	104.36	30.99
Proposed model (MoVi-HER2)	93.3	373.3	100.3

D. DISCUSSION

The quantitative results demonstrate that MoVi-HER2 consistently achieved the highest macro F1-score, weighted F1-score, and balanced accuracy across all evaluation scenarios, even under minimal preprocessing and without data augmentation. The consistently high balanced accuracy (>0.97) for HER2 status prediction indicates strong sensitivity for HER2-positive cases while preserving specificity for HER2-negative tissues, a clinically critical property, since misclassifying HER2-positive cases as negative could exclude patients from life-saving anti-HER2 therapies such as trastuzumab. Detailed per-class analysis revealed that MoVi-HER2 achieved 100% recall for the HER2-positive class (16/16) and 100% accuracy for the IHC 3 category (8/8), while correctly classifying 80% of IHC 2 samples (12/15), a clinically important class that typically requires confirmatory ISH testing (Fig. 3). As shown in Fig. 3a, the IHC 1 category was included during training but absent from the test set, since the dataset contained no IHC 1 samples for evaluation.

For HER2 status prediction (Fig. 3b), the model achieved perfect recall for the HER2-positive class (16/16), correctly identifying all positive cases, with a false-positive rate of only 4.1% (12 of 291 HER2-negative cases). However, the precision for the HER2-positive class was comparatively lower at 57% (16/28), as the small number of positive cases in the test set means that even a few false positives substantially reduce precision. Importantly, this



(a) (b)

Fig. 3. Confusion matrices of the proposed MoVi-HER2 (Concatenation) for (a) IHC score and (b) HER2 status prediction. Note: IHC 1 was included during training but absent from the test set.

precision–recall trade-off is clinically favorable in the context of HER2 screening: a false negative, in which a HER2-positive case is missed, is far more consequential than a false positive, as it could exclude an eligible patient from HER2-targeted therapy such as trastuzumab. Since false positives can be resolved through the standard confirmatory diagnostic workflow, the model's prioritization of sensitivity aligns with the clinical requirement to minimize missed HER2-positive cases.

A noteworthy property of MoVi-HER2 is its robustness under minimal preprocessing. As shown in Table II, applying full preprocessing or augmentation generally degraded performance, suggesting that aggressive normalization may inadvertently remove discriminative color and texture cues. This observation is further supported by the dedicated ablation studies (Tables III and IV), which confirmed that CLAHE-based stain normalization and combined augmentation each contributed to performance decline. The fusion of local feature normalization with global attention distribution appears to provide adaptive invariance to variations in hue, brightness, and texture statistics, a property rarely achieved in histopathological models, suggesting that MoVi-HER2 has the potential for deployment across varied imaging conditions, although validation on independent multi-center cohorts is required to confirm this capability.

To examine model interpretability, fusion attention visualizations were generated for representative samples (Fig. 4). Clinical HER2 assessment is based on the intensity and completeness of membrane staining, where strong, complete circumferential membrane staining indicates HER2 overexpression (IHC 3+), whereas weak, incomplete, or absent membrane staining corresponds to negative status. For HER2-negative samples (Figures 4a–4c), both fusion mechanisms produced diffuse, scattered activations distributed across the tissue without concentrating on any specific membranous pattern, consistent with the absence of the distinct membrane staining that characterizes HER2 overexpression. In contrast, HER2-positive samples (Fig 4d–4f) exhibited localized activations concentrated in the tumor epithelial regions exhibiting HER2 membrane staining, while stromal and luminal areas received minimal attention, directly aligning with the membrane-based diagnostic criteria for HER2 overexpression. Within these regions, concatenation fusion (Fig. 4e) showed broader contextual activation across glandular boundaries, while addition fusion (Fig. 4f) concentrated on specific high-intensity areas.

These visualizations confirm that MoVi-HER2 learns biologically plausible decision patterns aligned with established HER2 assessment criteria, rather than relying on spurious correlations. The transparency provided by fusion attention maps is essential for integration into digital pathology workflows, supporting both predictive accuracy and clinical trust, and establishing MoVi-HER2 as a robust, generalizable, and interpretable framework for HER2 prediction in GEA.

This study has several limitations that warrant consideration. First, the model was developed and evaluated on a single publicly available TMA dataset. Although the test set originated from a separate experimental batch, providing a degree of robustness assessment against inter-batch variability, external validation on independent multi-center cohorts to confirm the generalizability of model across different institutions, scanners, and staining protocols. Second, the dataset exhibits substantial class imbalance, particularly the scarcity of IHC 2+ and HER2-positive cases. While data

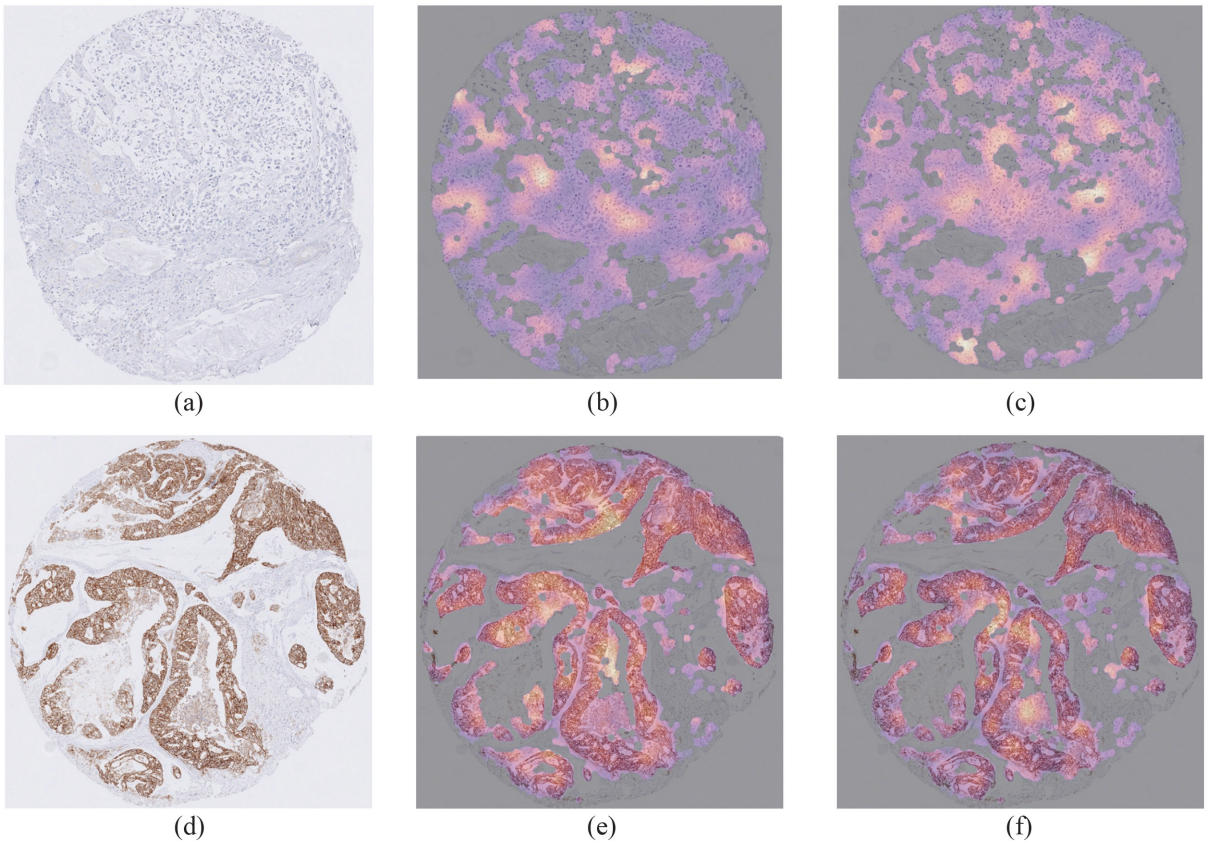


Fig. 4. MoVi-HER2 fusion visualization for HER2-negative (a–c) and HER2-positive (d–f), showing original images, concatenation fusion, and addition fusion (left to right).

augmentation was applied to the minority classes to mitigate this imbalance, our ablation experiments revealed that augmentation did not improve, and in some cases degraded, model performance, indicating that augmentation strategies are insufficient to fully resolve class imbalance in this histopathological context. Consequently, minority-class performance, especially for the clinically critical IHC 2+ category, remains an area for improvement.

To address these limitations, future work will pursue several directions. External multi-center validation will be conducted to assess generalizability across diverse clinical settings, and advanced imbalance-handling techniques such as cost-sensitive learning or specialized loss functions will be explored to improve minority-class performance. Integrating ordinal regression losses could further capture the inherently ordered nature of IHC scoring (0, 1+, 2+, 3+), while systematic hyperparameter optimization techniques such as the Tree-structured Parzen Estimator (TPE) may enhance the tuning of the ViT-Base component. Finally, extending the framework to incorporate multi-staining sources and multi-modal data, combining histopathological, genomic, and clinical information, represents a promising avenue toward more comprehensive and clinically translatable biomarker prediction in precision oncology.

V. CONCLUSION

This research presents MoVi-HER2, a novel MobileNetV3-ViT fusion network for predicting HER2 status and IHC scores in GEA from TMA images. The model integrates the distinct advantages of convolutional and transformer-based architectures, combining MobileNetV3's fine-grained local feature extraction with ViT's global contextual modeling. Across extensive experiments, MoVi-HER2 consistently outperformed all baselines, including MobileNetV3, ViT, EfficientNetV2, EfficientFormerV2, TinyViT, and Xception, in macro F1-score, weighted F1-score, and balanced accuracy, even under minimal preprocessing (resize-only), demonstrating strong generalization and robustness to staining variations. Dedicated ablation studies further revealed that aggressive preprocessing such as CLAHE-based stain normalization and combined data augmentation degraded performance, confirming that the model performs best when learning directly from minimally processed images. For HER2 status prediction, the model achieved a balanced accuracy of 0.9794 and macro F1-score of 0.8531, with complete recall of HER2-positive cases, ensuring clinical safety and reliability.

ACKNOWLEDGMENTS

This research was supported by Telkom University and Directorate of Research, Technology, and Community Service, Ministry of Higher Education, Science, and Technology of the Republic of Indonesia according to the contract number 125/C3/DT.05.00/PL/2025, 7925/LL4/PG/2025, and 114/LIT07/PPM-LIT/2025.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] S. Al-Haddad *et al.*, "Adenocarcinoma at the gastroesophageal junction," *Ann. N. Y. Acad. Sci.*, vol. 1325, no. 1, pp. 211–225, Sep. 2014, doi: 10.1111/nyas.12535.

[2] M. A. Shah *et al.*, "Improving outcomes in patients with oesophageal cancer," *Nat. Rev. Clin. Oncol.*, vol. 20, no. 6, pp. 390–407, Jun. 2023, doi: 10.1038/s41571-023-00757-y.

[3] W. P. M. Dijksterhuis *et al.*, "Gender differences in treatment allocation and survival of advanced gastroesophageal cancer: A population-based study," *JNCI: J. Natl. Cancer Inst.*, vol. 113, no. 11, pp. 1551–1560, Nov. 2021, doi: 10.1093/jnci/djab075.

[4] C. Giacometti, A. Gusella, and M. Cassaro, "Gastro-esophageal junction precancerosis: Histological diagnostic approach and pathogenetic insights," *Cancers (Basel).*, vol. 15, no. 24, p. 5725, Dec. 2023, doi: 10.3390/cancers15245725.

[5] Y. Jiang *et al.*, "Global trends in the burden of esophageal cancer, 1990−2019: results from the Global Burden of Disease Study 2019," J. Thorac. Dis., vol. 15, no. 2, pp. 348–364, Feb. 2023, doi: 10.21037/jtd-22-856.

[6] L. Sun *et al.*, "Global, regional, and national burden of esophageal cancer using the 2019 global burden of disease study," *Sci. Rep.*, vol. 15, no. 1, p. 3284, Jan. 2025, doi: 10.1038/s41598-025-86244-z.

[7] M. Elnaggar *et al.*, "Trends in esophageal cancer mortality in the United States (1999–2024): Disparities by sex, race/ethnicity, region, and urbanization," *Clin. Res. Hepatol. Gastroenterol.*, vol. 49, no. 9, p. 102691, Nov. 2025, doi: 10.1016/j.clinre.2025.102691.

[8] R. He *et al.*, "Analysis of CSF2RA expression and its functional and clinical significance in gastric cancer based on bioinformatics," *J. Biosci. Med. (Irvine).*, vol. 13, no. 09, pp. 405–425, 2025, doi: 10.4236/jbm.2025.139036.

[9] B. C. Grieb and R. Agarwal, "HER2-directed therapy in advanced gastric and gastroesophageal adenocarcinoma: Triumphs and troubles," *Curr. Treat. Options Oncol.*, vol. 22, no. 10, p. 88, Oct. 2021, doi: 10.1007/s11864-021-00884-7.

[10] A. Gordon *et al.*, "Targeting FGFR2 positive gastroesophageal cancer: Current and clinical developments," Onco. Targets. Ther., vol. 15, pp. 1183–1196, Oct. 2022, doi: 10.2147/OTT.S282718.

[11] J. F. Hechtman and A. D. Polydorides, "HER2/neu gene amplification and protein overexpression in gastric and gastroesophageal junction adenocarcinoma: A review of histopathology, diagnostic testing, and clinical implications," *Arch. Pathol. Lab. Med.*, vol. 136, no. 6, pp. 691–697, Jun. 2012, doi: 10.5858/arpa.2011-0168-RS.

[12] M. Alsina, M. Diez, and J. Tabernero, "Emerging biological drugs for the treatment of gastroesophageal adenocarcinoma," *Expert Opin. Emerg. Drugs*, vol. 26, no. 4, pp. 385–400, Oct. 2021, doi: 10.1080/14728214.2021.2010705.

[13] F. Wang *et al.*, "Trastuzumab plus docetaxel and capecitabine as a first-line treatment for HER2-positive advanced gastric or gastroesophageal junction cancer: A phase II, multicenter, open-label, single-arm study.," *Am. J. Cancer Res.*, vol. 10, no. 9, pp. 3037–3046, 2020.

[14] M. G. Kohale *et al.*, "Immunohistochemistry in pathology: A review," *J. Cell. Biotechnol.*, vol. 9, no. 2, pp. 131–138, Dec. 2023, doi: 10.3233/JCB-230110.

[15] D. Y. Mebratie and G. G. Dagnaw, "Review of immunohistochemistry techniques: Applications, current status, and future perspectives," *Semin. Diagn. Pathol.*, vol. 41, no. 3, pp. 154–160, May 2024, doi: 10.1053/j.semdp.2024.05.001.

[16] A. Hall *et al.*, "Incorporating molecular data into treatment decision making in gastroesophageal and pancreaticobiliary cancers: Timing and strategies," *Am. Soc. Clin. Oncol. Educ. Book*, vol. 44, no. 3, e433640, Jun. 2024, doi: 10.1200/EDBK_433640.

[17] N. Buza *et al.*, "Reproducibility of scoring criteria for HER2 immunohistochemistry in endometrial serous carcinoma: A multi-institutional interobserver agreement study," *Mod. Pathol.*, vol. 34, no. 6, pp. 1194–1202, Jun. 2021, doi: 10.1038/s41379-021-00746-5.

[18] Z. Wu *et al.*, "Deep learning-based prediction of HER2 status and trastuzumab treatment efficacy of gastric adenocarcinoma based on morphological features," *J. Transl. Med.*, vol. 23, no. 1, p. 13, Jan. 2025, doi: 10.1186/s12967-024-06034-5.

[19] Z. Han *et al.*, "A deep learning quantification algorithm for HER2 scoring of gastric cancer," *Front. Neurosci.*, vol. 16, p. 877229, May 2022, doi: 10.3389/fnins.2022.877229.

[20] J. I. Pisula *et al.*, "Predicting the HER2 status in oesophageal cancer from tissue microarrays using convolutional neural networks," *Br. J. Cancer*, vol. 128, no. 7, pp. 1369–1376, Mar. 2023, doi: 10.1038/s41416-023-02143-y.

[21] Y. Liao *et al.*, "Artificial intelligence for predicting her2 status of gastric cancer based on whole-slide histopathology images: A retrospective multicenter study," *Adv. Sci.*, vol. 12, no. 10, 2408451, Mar. 2025, doi: 10.1002/advs.202408451.

[22] M. Hayati *et al.*, "Impact of CLAHE-based image enhancement for diabetic retinopathy classification through deep learning," *Procedia Comput. Sci.*, vol. 216, pp. 57–66, 2023, doi: 10.1016/j.procs.2022.12.111.

[23] S. Okyere-Gyamfi *et al.*, "Contrast limited adaptive histogram equalization (CLAHE) and colour difference histogram (CDH) feature merging capsule network (CCFMCapsNet) for complex image recognition," *PLoS One*, vol. 20, no. 10, p. e0335393, Oct. 2025, doi: 10.1371/journal.pone.0335393.

[24] S. B. R. Prasad and B. S. Chandana, "Mobilenetv3: A deep learning technique for human face expressions identification," *Int. J. Inf. Technol.*, vol. 15, no. 6, pp. 3229–3243, Aug. 2023, doi: 10.1007/s41870-023-01380-x.

[25] B. Koonce, "MobileNetV3," in Convolutional Neural Networks with Swift for Tensorflow, Berkeley, CA: Apress, 2021, pp. 125–144. doi: 10.1007/978-1-4842-6168-2_11.

[26] A. Howard *et al.*, "Searching for MobileNetV3," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea: IEEE, Oct. 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

[27] S. Qian, C. Ning, and Y. Hu, "MobileNetV3 for Image Classification," in 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China: IEEE, Mar. 2021, pp. 490–497. doi: 10.1109/ICBAIE52039.2021.9389905.

[28] S. Khan *et al.*, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.

[29] X. Mao *et al.*, "Towards robust vision transformer," 2021. Accessed: Dec. 25, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:235211752

[30] X. Jiang, S. Wang, and Y. Zhang, "Vision transformer promotes cancer diagnosis: A comprehensive review," *Expert Syst. Appl.*, vol. 252, p. 124113, Oct. 2024, doi: 10.1016/j.eswa.2024.124113.

[31] I. Pacal, "MaxCerVixT: A novel lightweight vision transformer-based Approach for precise cervical cancer detection," *Knowl. Based. Syst.*, vol. 289, p. 111482, Apr. 2024, doi: 10.1016/j.knosys.2024.111482.

[32] P. K. Shukla *et al.*, "Encoder only attention-guided transformer framework for accurate and explainable social media fake profile detection," *Peer Peer Netw. Appl.*, vol. 18, no. 4, p. 232, Jul. 2025, doi: 10.1007/s12083-025-02047-z.

[33] R. S. Sandhya Devi, V. R. Vijay Kumar, and P. Sivakumar, "EfficientNetV2 model for plant disease classification and pest recognition," *Comput. Syst. Sci. Eng.*, vol. 45, no. 2, pp. 2249–2263, 2023, doi: 10.32604/csse.2023.032231.

[34] K. Wu *et al.*, "TinyViT: Fast pretraining distillation for small vision transformers," 2022, pp. 68–85. doi: 10.1007/978-3-031-19803-8_5.

[35] Y. Zhou *et al.*, "Boosting dental image classification performance with EfficientFormer via transfer learning," *Biomed. Signal Process. Control*, vol. 113, p. 109213, Mar. 2026, doi: 10.1016/j.bspc.2025.109213.