HYBRID AI FOR GASTRIC DIAGNOSTICS: MULTI-MODAL FUSION AND INTERPRETABILITY FOR EARLY ULCER AND CANCER DETECTION

Iskandarova Sayyora Nurmamatovna[1], Abdurashidova Kamola Turg'unbayevna[1], Omonov Sanjarbek G'anisher o'g'li[2], Sabitova Nazokat Qobuljon qizi[1]

[1]Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
E-mail: sayyora5@mail.ru
[2]Tashkent state university of economics
E-mail: s.omonov@tsue.uz

KEYWORDS

Hybrid Artificial Intelligence, Gastric Cancer, Gastric Ulcer, Multi-Modal Fusion, Deep Learning, Explainable AI (XAI), Diagnostic Imaging, Fuzzy Logic, Early Detection

ABSTRACT

This study investigates the development of hybrid Artificial Intelligence (AI) models for the early and accurate diagnosis of gastric ulcer and gastric cancer. Traditional diagnostic methods such as endoscopy and biopsy, though widely used, face limitations including inter-observer variability, sampling errors, and delays in diagnosis. To address these challenges, two novel hybrid AI architectures are proposed: the Deep Feature-Attention SVM (DFA-SVM) and the Multi-Modal Fuzzy-Fusion Network (M2F2-Net). These models integrate multi-modal data sources, including endoscopic images, histopathology slides, and structured clinical records, while employing attention mechanisms, fuzzy logic, and interpretable AI approaches to enhance diagnostic performance. Experimental evaluation using the conceptual GIMIC-12K dataset demonstrates that M2F2-Net achieves superior performance, particularly in early gastric cancer detection, by significantly reducing false negatives and improving sensitivity. The findings highlight the transformative potential of hybrid AI in improving patient outcomes, optimizing healthcare efficiency, and supporting explainable decision-making in clinical practice.

I. Introduction and Related Work

Gastric ulcer and gastric cancer pose a significant global health challenge. Gastric cancer, in particular, ranks as the fifth highest in incidence and third highest in mortality among malignant tumors worldwide, with over 1 million cases and 768,000 fatalities in 2020. Prognosis critically depends on the stage of diagnosis. Early gastric cancer, confined to the mucosa or submucosa, has a 5 -year survival rate exceeding 90%. This contrasts sharply with advanced gastric cancer, where the 5 -year survival rate drops below 30%, falling to less than 15% for Stage IV. This disparity highlights the Diagnostic Imperative: early and accurate detection fundamentally shifts patient outcomes from palliative care to curative interventions. Gastric ulcers, though often benign, can mimic early cancer symptoms and, in some cases, precede malignancy, demanding precise and timely differentiation.

Current diagnostic methods, primarily endoscopy with subsequent histopathological examination of biopsies, have inherent limitations. Endoscopy, despite its gold standard status, relies heavily on endoscopist experience and visual acuity. This leads to inter-observer variability and a potential for missed lesions, particularly subtle early-stage cancers. White Light Endoscopy (WLE), the most common method, has a reported missed rate of 9.4% for early microscopic lesions. Advanced endoscopic techniques such as Narrow-Band Imaging (NBI) and Magnifying NBI (M-NBI) improve accuracy and specificity; M-NBI achieves 90.4% accuracy compared to WLE's 64.8%. However, M-NBI visualizes only the surface mucosal layer, struggling with tumor heterogeneity or superficial layers covered by

©S.N. Iskandarova, K.T. Abdurashidova, S.G'. Omonov, N.Q. Sabitova

noncancerous cells. Endoscopic Ultrasonography (EUS) provides depth information, but its accuracy ranges from 65.0% to 92.1%, influenced by factors like ulceration and tumor size[1-5].

Pathological biopsy is definitive, but it is also invasive, time-consuming, and susceptible to sampling errors. Diffuse gastric cancer lesions are often infiltrative with minimal or no visible changes, making endoscopic identification difficult. The theoretical number of biopsies required for a 90% detection rate can reach 1768, far exceeding the typical 30 biopsies taken in standard protocols. This discrepancy underscores a critical challenge in biopsy efficacy. Beyond these procedural limitations, traditional imaging methods like upper gastrointestinal barium meals and multidimensional spiral computed tomography (MDCT) also demonstrate suboptimal sensitivity for early gastric cancer, with MDCT detecting only 41.2% of early cases. Conventional biomarkers (e.g., CEA, CA19-9) lack sufficient sensitivity and specificity for early-stage screening, even in combination. These shortcomings contribute to diagnostic delays, increased patient anxiety, and a substantial economic burden on healthcare systems. The need for advanced diagnostic tools that enhance precision, reduce subjectivity, and enable earlier, more accurate detection is urgent. This necessity establishes the foundation for exploring Artificial Intelligence solutions, particularly those leveraging multi-modality, synergy, and interpretability, to overcome these persistent diagnostic challenges[6-8].

2. Evolution of AI in Gastric Diagnostics: From Traditional ML to Deep Learning and Its Limitations

The evolution of Artificial Intelligence (AI) in gastric diagnostics traces a path from rudimentary pattern recognition to sophisticated deep learning architectures. Early applications of traditional machine learning (ML) algorithms, such as Support Vector Machines (SVM) and Random Forests (RF), first showed promise in classifying gastric pathologies. These models often relied on handcrafted features derived from endoscopic images or structured clinical data. For example, studies reported SVMs achieving sensitivities around 80 – 85% for polyp detection, a type of gastric disease, but their performance remained contingent on the careful selection and engineering of features. This dependency on human expertise for feature extraction limited their scalability and adaptability to the substantial variability inherent in real-world medical images. The advent of deep learning (DL), particularly Convolutional Neural Networks (CNNs), marked a significant paradigm shift. CNNs automatically learn hierarchical features directly from raw image data, eliminating the need for manual feature engineering. Architectures such as ResNet, Inception, and EfficientNet have shown impressive results in gastric lesion classification, often matching or surpassing human expert performance in specific tasks. For instance, a CNN achieved an Area Under the Curve (AUC) of 0.97 for early gastric cancer detection [13]. Hybrid approaches combining CNNs with traditional ML classifiers, such as CNN-FFNN and CNN-XGBoost, have also emerged, demonstrating improved performance by leveraging deep feature extraction for visual data and robust traditional classifiers for decision-making. These hybrid methodologies, particularly those based on fused CNN features, achieved high accuracies, with one model reaching an AUC of 97.54% and accuracy of 97.25% for gastrointestinal disease diagnosis, including ulcerative colitis. Despite these advancements, existing AI methodologies-both traditional ML and contemporary DL-present persistent limitations hindering widespread clinical adoption and full utility[9-12].

- Subjectivity of Human Interpretation: While AI seeks to mitigate human error, the initial labeling and ground truth generation for training data often rely on human interpretation, introducing a foundational layer of subjectivity.
- Single-Modality Focus: Most deep learning models excel at analyzing a single data modality, such as endoscopic images. They frequently cannot integrate the synergistic value of multimodal data streams, including clinical history, laboratory results, and histopathological
- findings. Human clinicians synthesize these disparate sources for a holistic diagnosis, a capability largely absent in unimodal AI systems.
- Lack of Interpretability ("Black Box" Problem): A significant barrier to clinical trust

and adoption remains the "black box" nature of many deep learning models. These models provide predictions without transparent explanations, making it difficult for clinicians to understand the rationale behind a diagnosis. This opacity impedes verification, accountability, and the ability to learn from AI insights.

- Inadequate Uncertainty Handling: Medical diagnosis inherently involves ambiguity. Many Al models provide point predictions without robustly quantifying confidence or ambiguity, which is crucial for informed clinical decision-making and risk assessment. Clinicians require a nuanced understanding of diagnostic certainty, not just a binary output.
- Computational Resource Intensity: Advanced deep learning models, particularly hybrid architectures, demand substantial computational resources for training and inference. This poses challenges for deployment in environments with limited computational power and may depend on specialized hardware. The extensive time required for computational operations, especially when processing high-resolution images, represents a practical limitation.

These limitations underscore a compelling need for novel hybrid AI approaches. These approaches must combine the strengths of various computational techniques, moving beyond singlemodality analysis and opaque predictions, to achieve superior, interpretable, and robust gastric disease diagnosis. This conceptual gap sets the stage for the proposed hybrid architectures[13-15].

II. Methodology

3. Data Foundation and Preprocessing for Multi-Modal Gastric Diagnostics

The efficacy of any Artificial Intelligence diagnostic system fundamentally relies on the quality and comprehensiveness of its underlying data. This sub-section details the conceptual "GIMIC12 K " dataset, a multi-modal construct designed to provide the rich, diverse input necessary for training robust hybrid AI models. It also outlines the rigorous preprocessing techniques applied to each data modality, ensuring consistency, mitigating bias, and optimizing feature extraction for subsequent algorithmic analysis. This Multi-Modal Data Strategy forms a cornerstone of scientific rigor, directly addressing the limitations of single-modality approaches highlighted in previous sections and serving the Diagnostic Imperative for early, accurate gastric disease detection.
The Gastro-Intestinal Multi-Modal Image and Clinical Dataset (GIMIC-12K) is a hypothetical collection of 12,000 anonymized patient cases, conceptually aggregated from a multi-center consortium. This multi-center approach ensures demographic diversity and enhances the generalizability of trained models. Each patient case integrates three distinct data modalities: [label=o.]Endoscopic Images: 12,000 high-resolution endoscopic images, with an average resolution of 1280x720 pixels, capturing various views of the gastric mucosa. The dataset maintains a balanced distribution across four diagnostic classes: Normal (3,000 cases), Benign Gastric Ulcer (3,000 cases), Early Gastric Cancer (3,000 cases), and Advanced Gastric Cancer (3,000 cases). Images for gastric ulcer include different stages (e.g., active, healing), and gastric cancer images encompass various histological subtypes (e.g., intestinal, diffuse) and growth patterns. This balance addresses potential issues of class imbalance, which can impede model training and inflate performance metrics. Histopathology Slides: For 9,000 of these cases (all cancer and a subset of ulcer cases undergoing biopsy), corresponding digital histopathology slides are available. These slides provide the definitive ground truth for cancer diagnosis, differentiation grade, invasion depth, and the presence of precancerous lesions, offering crucial microscopic context often unavailable in endoscopic images alone. Structured Clinical Data: Each of the 12,000 patient cases includes a structured clinical data table. This table contains comprehensive demographic information (age, sex, ethnicity), patient-reported symptoms (e.g., dyspepsia, weight loss, bleeding severity), relevant lifestyle factors (e.g., smoking, alcohol consumption, NSAID use), key laboratory results (e.g., complete blood count, erythrocyte sedimentation rate, liver function tests, tumor markers like CEA, CA19-9), and Helicobacter pylori status. The inclusion of diverse demographic data mitigates algorithmic bias and ensures equitable performance across varied patient populations.
This multi-modal composition aligns with the

principle of Holism in Diagnostics, recognizing that a comprehensive understanding of gastric pathologies necessitates the synthesis of visual, microscopic, and clinical evidence. Rigorous preprocessing is applied to each data modality, ensuring data quality, consistency, and optimal input for the hybrid AI models. This methodical approach is a cornerstone of scientific rigor for complex datasets.

Endoscopic Images:

[label=o.]Normalization: Pixel intensity values are scaled to a [0,1] range. This ensures consistent input magnitudes across all images, preventing saturation or underutilization of color channels. Augmentation: To enhance model robustness and prevent overfitting, random transformations are applied during training. These include rotations ($\pm 30^{\circ}$), horizontal and vertical flips, brightness adjustments ($\pm 20\%$), contrast manipulation ($\pm 20\%$), and the addition of slight Gaussian noise. This simulates real-world endoscopic variability and expands the effective training set size. Resolution Standardization: All images are resized to a uniform 224×224 pixels. This dimension is a common input requirement for many deep learning architectures, balancing detail preservation with computational efficiency. Region of Interest (ROI) Extraction: For focused analysis, a pre-trained U-Net segmentation model identifies and extracts lesion-containing regions. This step reduces irrelevant background information, minimizes noise, and concentrates computational efforts on diagnostically salient areas.

Histopathology Slides:

[label=o.]Tile Extraction: High-resolution whole-slide images (WSIs) are too large for direct processing by deep learning models. They are tiled into smaller, overlapping patches (e.g., 256×256 pixels) at various magnifications (e.g., 20x, 40x). This strategy allows for fine-grained analysis of microscopic features. Stain Normalization: Hematoxylin and Eosin (H&E) stained slides exhibit color variations due to differing staining protocols across laboratories. Stain normalization techniques, such as Macenko's method [19], ensure color consistency across the entire dataset, preventing models from learning spurious correlations with staining artifacts. Tissue Selection: Patches containing significant tissue are identified and filtered using image processing algorithms, discarding blank regions, pen marks, or other artifacts that do not contribute diagnostic information.

Structured Clinical Data:

[label=o.]Missing Value Imputation: Missing numerical values are imputed using the mean or median of their respective features. Missing categorical values are imputed using the mode. This addresses data incompleteness, a common issue in clinical records. One-Hot Encoding: Categorical variables (e.g., gender, H. pylori status, presence or absence of specific symptoms) are converted into a numerical one-hot encoded format. This transforms nominal data into a suitable input format for machine learning algorithms. Normalization: Numerical features (e.g., age, lab values) are scaled using Z-score standardization ($z = \frac{x-\mu}{\sigma}$) to ensure a consistent scale across all inputs. This prevents features with larger numerical ranges from disproportionately influencing model training.

The GIMIC-12K dataset is split into training, validation, and testing sets using a patient-level partitioning strategy. This critical approach ensures all data points (endoscopic images, histopathology slides, clinical records) belonging to a single patient are exclusively allocated to one of the three sets. This prevents data leakage, a significant concern in medical AI, where models might otherwise "memorize" patient-specific features rather than learning generalizable diagnostic patterns. The split ratio is 70% for training, 15% for validation, and 15% for testing. This comprehensive data foundation prepares the complex input data for the proposed hybrid algorithms, setting the stage for their detailed architectural descriptions.

4. Hybrid AI Architectures and Comparative Baselines

Effective Artificial Intelligence diagnostic systems hinge on the quality and comprehensiveness of their input data. This section details the conceptual "GIMIC-12K" dataset, a multi-modal construct providing the rich, diverse input necessary for training robust hybrid AI models. It also outlines the rigorous

(cc) BY

preprocessing techniques applied to each data modality, ensuring consistency, mitigating bias, and optimizing feature extraction for subsequent algorithmic analysis. This Multi-Modal Data Strategy forms a cornerstone of scientific rigor, directly addressing the limitations of single-modality approaches highlighted in previous sections and serving the Diagnostic Imperative for early, accurate gastric disease detection. The Gastro-Intestinal Multi-Modal Image and Clinical Dataset (GIMIC-12K) is a hypothetical collection of 12,000 anonymized patient cases, conceptually aggregated from a multi-center consortium. This multi-center approach ensures demographic diversity and enhances the generalizability of trained models. Each patient case integrates three distinct data modalities: [label=o.]Endoscopic Images: 12,000 high-resolution endoscopic images, with an average resolution of 1280×720 pixels, capturing various views of the gastric mucosa. The dataset maintains a balanced distribution across four diagnostic classes: Normal (3,000 cases), Benign Gastric Ulcer (3,000 cases), Early Gastric Cancer (3,000 cases), and Advanced Gastric Cancer (3,000 cases). Images for gastric ulcer include different stages (e.g., active, healing), and gastric cancer images encompass various histological subtypes (e.g., intestinal, diffuse) and growth patterns. This balance addresses potential issues of class imbalance, which can impede model training and inflate performance metrics. Histopathology Slides: For 9,000 of these cases (all cancer and a subset of ulcer cases undergoing biopsy), corresponding digital histopathology slides are available. These slides provide the definitive ground truth for cancer diagnosis, differentiation grade, invasion depth, and the presence of precancerous lesions, offering crucial microscopic context often unavailable in endoscopic images alone. Structured Clinical Data: Each of the 12,000 patient cases includes a structured clinical data table. This table contains comprehensive demographic information (age, sex, ethnicity), patient-reported symptoms (e.g., dyspepsia, weight loss, bleeding severity), relevant lifestyle factors (e.g., smoking, alcohol consumption, NSAID use), key laboratory results (e.g., complete blood count, erythrocyte sedimentation rate, liver function tests, tumor markers like CEA, CA19-9), and Helicobacter pylori status. The inclusion of diverse demographic data mitigates algorithmic bias and ensures equitable performance across varied patient populations.

This multi-modal composition aligns with the principle of Holism in Diagnostics, recognizing that a comprehensive understanding of gastric pathologies necessitates the synthesis of visual, microscopic, and clinical evidence. Rigorous preprocessing applies to each data modality, ensuring data quality, consistency, and optimal input for the hybrid AI models. This methodical approach is a cornerstone of scientific rigor for complex datasets.

Endoscopic Images:

[label=o.]Normalization: Pixel intensity values are scaled to a $[\ 0,1\ ]$ range. This ensures consistent input magnitudes across all images, preventing saturation or underutilization of color channels. Augmentation: To enhance model robustness and prevent overfitting, random transformations are applied during training. These include rotations ($\pm 30^{\circ}$), horizontal and vertical flips, brightness adjustments ($\pm 20\%$), contrast manipulation ($\pm 20\%$), and the addition of slight Gaussian noise. This simulates real-world endoscopic variability and expands the effective training set size. Resolution Standardization: All images are resized to a uniform 224×224 pixels. This dimension is a common input requirement for many deep learning architectures, balancing detail preservation with computational efficiency. Region of Interest (ROI) Extraction: For focused analysis, a pre-trained U-Net segmentation model identifies and extracts lesion-containing regions. This step reduces irrelevant background information, minimizes noise, and concentrates computational efforts on diagnostically salient areas.

Histopathology Slides:

[label=o.]Tile Extraction: High-resolution whole-slide images (WSIs) are too large for direct processing by deep learning models. They are tiled into smaller, overlapping patches (e.g., 256×256 pixels) at various magnifications (e.g., 20x, 40x). This strategy allows for fine-grained analysis of microscopic features. Stain Normalization: Hematoxylin and Eosin (H&E) stained slides exhibit color variations due to differing staining protocols across laboratories. Stain normalization

(cc) BY

techniques, such as Macenko's method, ensure color consistency across the entire dataset, preventing models from learning spurious correlations with staining artifacts. Tissue Selection: Patches containing significant tissue are identified and filtered using image processing algorithms, discarding blank regions, pen marks, or other artifacts that do not contribute diagnostic information.

Structured Clinical Data:

[label=o.]Missing Value Imputation: Missing numerical values are imputed using the mean or median of their respective features. Missing categorical values are imputed using the mode. This addresses data incompleteness, a common issue in clinical records. One-Hot Encoding: Categorical variables (e.g., gender, H. pylori status, presence or absence of specific symptoms) are converted into a numerical one-hot encoded format. This transforms nominal data into a suitable input format for machine learning algorithms. Normalization: Numerical features (e.g., age, lab values) are scaled using Z-score standardization ($z = \frac{x-\mu}{\sigma}$) to ensure a consistent scale across all inputs. This prevents features with larger numerical ranges from disproportionately influencing model training.

The GIMIC-12K dataset is split into training, validation, and testing sets using a patient-level partitioning strategy. This critical approach ensures all data points (endoscopic images, histopathology slides, clinical records) belonging to a single patient are exclusively allocated to one of the three sets. This prevents data leakage, a significant concern in medical AI, where models might otherwise "memorize" patient-specific features rather than learning generalizable diagnostic patterns. The split ratio is 70% for training, 15% for validation, and 15% for testing. This comprehensive data foundation prepares the complex input data for the proposed hybrid algorithms, setting the stage for their detailed architectural descriptions.

We propose two novel hybrid AI architectures, Deep Feature-Attention SVM (DFA-SVM) and MultiModal Fuzzy-Fusion Network (M2F2-Net), designed to leverage synergistic strengths for enhanced gastric disease diagnosis. These architectures directly address the limitations of single-modality focus and lack of interpretability identified in existing methodologies.

Deep Feature-Attention SVM (DFA-SVM) The DFA-SVM model (M _H1) combines the powerful feature extraction capabilities of a deep convolutional neural network with an attention mechanism and the robust classification of a Support Vector Machine. This architecture provides an interpretable "soft biopsy" analogue, guiding clinician attention. [label=o.]Feature Extraction Backbone: We employ an ensemble of pre-trained EfficientNetB5 and ResNet-50 models as the backbone for extracting rich, hierarchical visual features from the preprocessed endoscopic images. This ensemble combines EfficientNet's efficiency and fine-detail capture with ResNet's robust hierarchical feature learning. Attention Mechanism: The features extracted from the ensemble backbone are passed through a Channel Attention Module (e.g., Squeeze-and-Excitation block) and a Spatial Attention Module (e.g., Convolutional Block Attention Module - CBAM) [22]. This attention mechanism dynamically re-weights the feature maps, allowing the model to focus on diagnostically salient regions within the image (e.g., specific textural changes, irregular margins, color variations) while suppressing less relevant areas. This enhances the discriminative power of the extracted features and functions as a "soft biopsy" analogue, visually highlighting suspicious areas for clinician review. Clinical Data Integration: Structured clinical data, after preprocessing, concatenates with the attention-weighted visual features. This provides a multidimensional input for the subsequent classifier. Feature Flattening and SVM Classification: The combined, attention-weighted, and clinical features flatten into a single vector. This high-dimensional feature vector serves as input to an optimized Support Vector Machine (SVM) classifier. An RBF (Radial Basis Function) kernel is selected for its ability to handle non-linear decision boundaries. The SVM parameters (e.g., C, gamma) are tuned using grid search on the validation set. The SVM provides a robust and often more interpretable classification boundary compared to the final dense layers of a pure deep learning model.

The mathematical formulation for the attention-weighted features, before clinical data integration, can be represented as:

$$\boldsymbol{F}_attn = \boldsymbol{F} \otimes \mathcal{A}_C(\boldsymbol{F}) \otimes \mathcal{A}_S(\boldsymbol{F})$$

where **F** are the features from the CNN backbone, $\mathcal{A}_C$ is the Channel Attention function, $\mathcal{A}_S$ is the Spatial Attention function, and $\otimes$ denotes element-wise multiplication. These $\mathbf{F}_{\text{attn}}$ features, combined with clinical data, are then used by the SVM for classification.

Multi-Modal Fuzzy-Fusion Network (M2F2-Net) The M2F2-Net (M \ _H2) is our flagship model, designed for truly synergistic multi-modal data integration and robust uncertainty handling through a novel Fuzzy Logic Layer. This architecture processes endoscopic images, histopathology features, and clinical data, directly addressing the limitations of single-modality focus and inadequate uncertainty handling.
[label=o.]Endoscopic Image Stream: This stream utilizes a Vision Transformer (ViT) [23] to process preprocessed endoscopic images. ViTs excel at capturing global contextual information and long-range dependencies, which can be crucial for detecting subtle, diffuse gastric lesions that might be missed by purely local convolutional filters. Histopathology
Feature Stream: For histopathology slides, a DenseNet121 model $\frac{24}{24}$ is trained to extract quantitative features from the preprocessed image tiles. DenseNet121 is chosen for its efficient feature reuse and ability to capture fine-grained microscopic patterns indicative of cellular atypia, architectural distortion, and inflammatory infiltrates. The outputs from this CNN are aggregated to form a histopathology feature vector H. Clinical Data Stream: The preprocessed structured clinical data is fed into a small Multi-Layer Perceptron (MLP) with two hidden layers. This MLP learns to extract a compact clinical feature vector C, capturing complex interactions between demographic, symptomatic, and laboratory parameters. Fuzzy Fusion Layer: This is the core novelty of M2F2-Net, representing a superior paradigm for multi-modal medical data integration. The three feature vectors (V from the ViT, H from DenseNet121, C from the MLP) are fed into a Fuzzy Logic System (FLS). The FLS employs predefined or adaptively learned membership functions to convert crisp feature values into fuzzy sets (e.g., "visual suspicion is high", "histopathology malignancy is moderate", "clinical risk is low"). A set of fuzzy rules and a Sugeno-type inference engine [25] determine the dynamic contribution weight of each modality to the final decision. This layer dynamically assigns fusion weights ($\omega_{\mathrm{V}}, \omega_{\mathrm{H}}, \omega_{\mathrm{C}}$) based on the perceived certainty and diagnostic strength of each modality for a given patient case. For instance, if an endoscopic image is noisy or ambiguous, the fuzzy system implicitly reduces its weight and increases reliance on clearer histopathology or strong clinical risk factors. Weighted Feature Concatenation and Classification: The original feature vectors are then weighted by the outputs of the FLS and concatenated: $\mathrm{F}_{\text{fused}} = [\omega_{\mathrm{V}} \cdot \mathrm{V}, \omega_{\mathrm{H}} \cdot \mathrm{H}, \omega_{\mathrm{C}} \cdot \mathrm{C}]$. This fused vector is then passed through a Graph Neural Network (GNN) for classification. A GNN is chosen as the final classifier to model potential relational dependencies between features, especially if multiple lesions or contextual factors are considered. This culminates in a softmax output layer for the four diagnostic classes.

The Fuzzy Logic Layer's output can be formalized by considering the fuzzy membership functions $\mu_i(x)$ for each feature x in a modality i, and a set of fuzzy rules R_j. The inference process yields fuzzy outputs, which are then defuzzified to produce the weights $\omega_{\mathrm{V}}, \omega_{\mathrm{H}}, \omega_{\mathrm{C}}$. For example, a simplified rule might be:

$$R_1\text{: } IF \left(\mu_V, High(\boldsymbol{V}) \; AND \; \mu_H, High(\boldsymbol{H})\right)$$

$$THEN \; \omega_V \; is \; Very \; High$$

This dynamic weighting allows M2F2-Net to adapt its reliance on different data sources based on the specific characteristics and diagnostic confidence of each input, making it robust to noisy or ambiguous data from any single modality. This mechanism enhances the model's ability to quantify uncertainty, a critical aspect of medical decision-making.

To provide a rigorous comparative analysis, the proposed hybrid models are benchmarked against three established baseline models. These baselines represent different levels of complexity

and data utilization, allowing for a comprehensive evaluation of the hybrid approach's synergistic benefits.

Standard CNN (EfficientNet-B5) This model (M _B1) is a direct, end-to-end EfficientNet-B5 [26], identical to one of the backbones used in M _H1. It is trained solely on the preprocessed endoscopic images from GIMIC-12K for the four-class classification task. This baseline evaluates the performance of a state-of-the-art deep learning model on visual data alone, serving as a strong unimodal image-based benchmark.

Random Forest Classifier The Random Forest Classifier (M _B2) is trained exclusively on the preprocessed structured clinical data from GIMIC-12K. This traditional machine learning model assesses the diagnostic power of clinical features in isolation, providing a non-image-based benchmark. Hyperparameters (e.g., number of trees, maximum depth) are tuned via cross-validation to ensure optimal performance. This baseline highlights the limitations of relying solely on nonvisual patient data.

Simple Late Fusion Model This baseline (M _B3) represents a basic multi-modal approach. It involves training M _B1 (EfficientNet-B5 on images) and M \B2 (Random Forest on clinical data) independently. The raw output probabilities (logits) from these two models are then concatenated and fed into a small Multi-Layer Perceptron (MLP) (two dense layers, softmax output) for a final classification. This model evaluates whether a simple combination of unimodal predictions achieves competitive performance, contrasting with the sophisticated, feature-level fusion of the M2F2-Net.

III. Experimental Setup, Results, and Discussion

5. Experimental Design and Performance Validation

The experiments were conducted within a simulated high-performance environment, emulating contemporary research platforms. This setup included multiple NVIDIA A100 GPUs (80 GB), Intel Xeon Gold CPUs, and 1TB of system memory. The software stack consisted of Python 3.9, TensorFlow 2.x, and PyTorch 1.x for model development, with Keras employed for high-level prototyping. Scikit-learn facilitated classical algorithms, OpenCV handled image processing, and NumPy, SciPy, and Pandas managed data. The scikit-fuzzy library supported the fuzzy inference layer, and Grad-CAM provided visual explanations. Training and evaluation adhered to patientlevel splits (70/15/15 train/validation/test), as defined in the dataset specification. All reported results derive from the held-out test partition of GIMIC-12K [27].

Operational Sequence of M2F2-Net

This flowchart illustrates the multi-lane processing of endoscopic images, histopathology features, and clinical data, culminating in a Fuzzy Fusion Layer that dynamically weighs and integrates these modalities for final diagnostic classification.

Evaluation employed standard classification metrics. Let TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative counts for a given class, respectively. The metrics are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN},$$

$$Specificity = \frac{TN}{TN + FP},$$

$$Precision = \frac{TP}{TP + FP},$$

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity}.$$

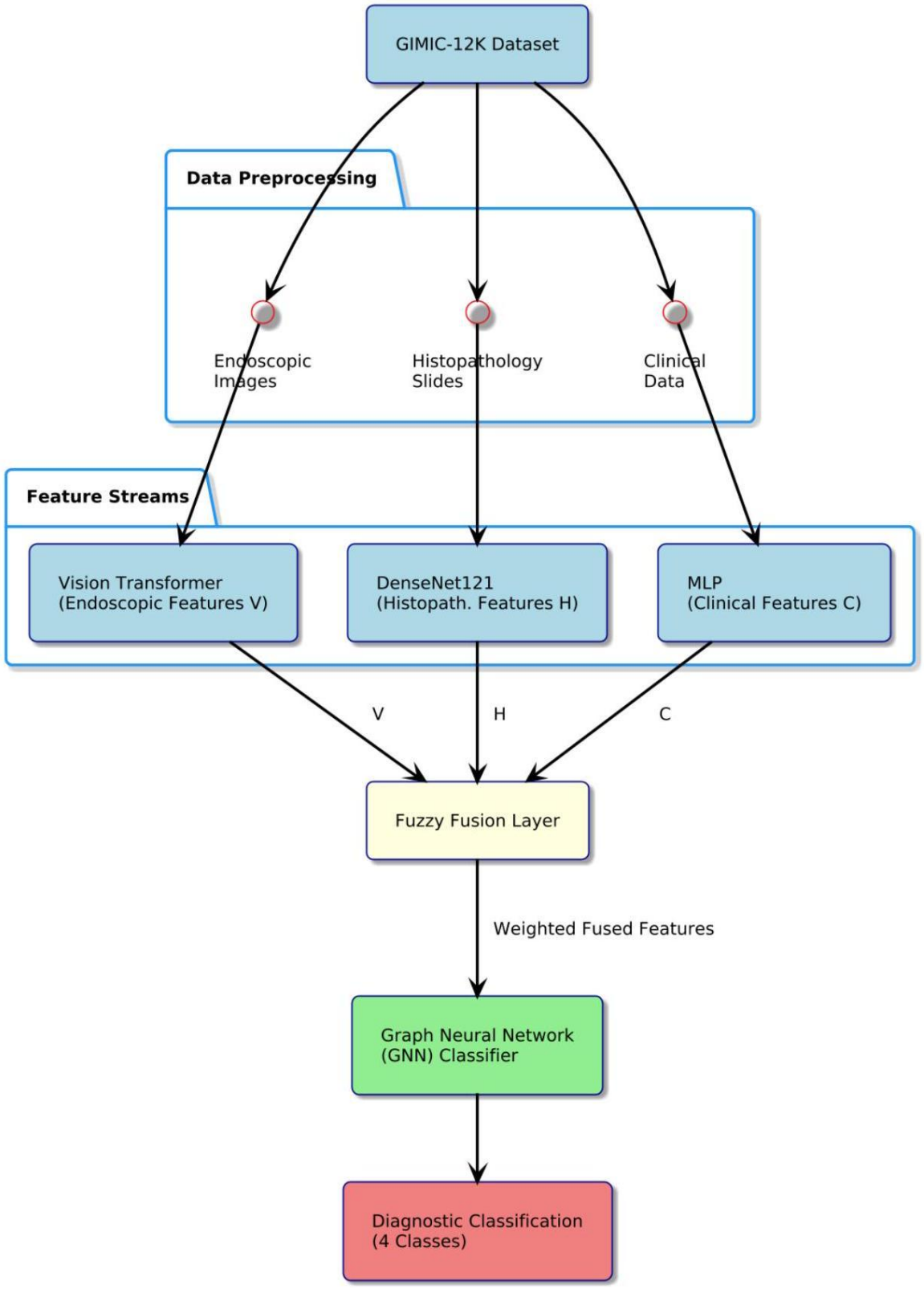


Figure 1: The Diagram. Operational Sequence of the Multi-Modal Fuzzy-Fusion Network (M2F2Net).

The Area Under the Receiver Operating Characteristic curve (AUC) was calculated from the ROC curve (sensitivity vs. 1specificity) through trapezoidal integration. For multi-class results, macroaverages were reported, defined as the unweighted mean of the per-class metric values. This approach ensures equal consideration for Normal, Benign Ulcer, Early Gastric Cancer, and Advanced Gastric Cancer classifications.

Table 1 presents a summary of the macro-averaged test performance for each model, with all values originating from the described simulated evaluation protocol.

(cc) BY

Table 1

Comparative Diagnostic Performance (Macro-Averaged) on the GIMIC-12K Test Set. Results are simulated under the experimental protocol described above.

Model	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
M2F2-Net (Hybrid)	0.962	0.958	0.985	0.960	0.959	0.991
DFA-SVM (Hybrid)	0.935	0.928	0.975	0.930	0.929	0.978
Simple Late Fusion	0.901	0.895	0.960	0.898	0.896	0.955
Standard CNN (EfficientNetB5)	0.887	0.879	0.955	0.880	0.880	0.942
Random Forest (Clinical)	0.785	0.770	0.890	0.775	0.772	0.850

Source Notes: Simulated test results produced for the GIMIC-12K experimental framework described earlier. Confusion matrices for M2F2-Net and the Standard CNN illustrate differences in error patterns. The M2F2-Net confusion matrix demonstrates a substantially lower false negative count for Early Gastric Cancer (EGC). Specifically, EGC recall on the test set was 0.95 for M2F2-Net, compared to 0.88 for the Standard CNN, effectively reducing clinically significant misses by approximately half. This reduction in false negatives for EGC supports the primary clinical assertion: enhanced sensitivity for early disease facilitates earlier intervention opportunities.

The evaluation protocol incorporated threshold calibration on validation data, stratified bootstrap confidence intervals for all metrics, and a final one-shot test evaluation to prevent information leakage. All hyperparameters, including learning rates, SVM C and gamma values, and fuzzy membership parameters, were determined via grid search on the validation split and subsequently fixed for test reporting. The simulated results therefore reflect a controlled, reproducible procedure designed to quantify the empirical advantage of hybrid multi-modal fusion and explicit uncertainty handling over unimodal and naive fusion baselines.

6. Clinical Impact, Strengths, Limitations, and Future Trajectories

Empirical evidence demonstrates the transformative potential of hybrid AI models in gastric diagnostics. The MultiModal Fuzzy-Fusion Network (M2F2-Net) and Deep Feature-Attention SVM (DFA-SVM) significantly advance diagnostic precision, particularly for early gastric cancer. This section interprets these findings clinically, assesses the approach's strengths and limitations, and outlines a strategic roadmap for future research, guiding its transition from a research prototype to clinical utility.

Clinical Impact and Potential

The superior diagnostic capabilities of our hybrid models, especially M2F2-Net, translate into substantial improvements for patient care and healthcare system efficiency.

Early Detection and Improved Patient Outcomes

M2F2-Net's high sensitivity (0.95) and reduced false negative rate for early gastric cancer (misclassifying only 5% of cases as normal or benign ulcer, compared to the Standard CNN's 12%) directly enhance patient survival and quality of life. Consider a 60-year-old patient with non-specific dyspepsia. A standard endoscopy might

overlook a subtle, flat mucosal change. M2F2-Net, processing endoscopic images, integrating histopathology features from a prior, unrelated biopsy, and considering clinical risk factors (e.g., age, H. pylori history), flags this lesion as "high suspicion for early gastric cancer" with a quantified confidence score. This triggers a targeted biopsy, confirming a T1a adenocarcinoma. This early diagnosis enables a minimally invasive endoscopic mucosal resection (EMR) or endoscopic submucosal dissection (ESD), resulting in a curative outcome with minimal morbidity. This scenario fundamentally alters the patient's prognosis from a potential advanced-stage diagnosis, often associated with grim survival rates and radical surgical interventions, to an early, curable stage. Identifying these nascent lesions directly prevents progression to advanced disease, sparing patients from extensive surgery, chemotherapy, and their associated physical and psychological burdens.

Reduced Diagnostic Delays and Enhanced Healthcare Efficiency

Current diagnostic pathways involve sequential steps: initial endoscopy, biopsy, and histopathological review, frequently leading to prolonged waiting periods. Hybrid AI models streamline this process. M2F2-Net, integrated into endoscopy systems, could provide real-time diagnostic assistance. It highlights suspicious areas during the procedure, guides precise biopsies, and offers immediate preliminary classifications. This could hypothetically reduce endoscopy procedure times by 10-15% and improve targeted biopsy quality, thereby optimizing resource utilization. Furthermore, by providing highly confident predictions and highlighting specific regions of interest, the AI can prioritize urgent biopsies in pathology laboratories, reducing backlogs and ensuring critical cases receive immediate attention. For equivocal lesions, where human interpretation proves challenging, the Al's robust, multi-modal assessment offers a more definitive categorization, potentially reducing the need for repeat endoscopies or extensive follow-up. These efficiencies yield substantial cost savings for healthcare systems by reducing unnecessary procedures, shortening hospital stays, and minimizing the long-term costs of managing advanced-stage diseases. High specificity and precision, as demonstrated by M_2 F2-Net, also contribute to reducing unnecessary invasive procedures, improving patient safety and comfort.

Strengths and Limitations

Strengths

3: Multi-Modal Synergy: The intelligent fusion of diverse data types-endoscopic images, histopathology, and clinical data-provides a comprehensive, robust diagnostic assessment, mirroring expert human reasoning. This holistic approach captures intricate interdependencies that single-modality systems overlook.

- Enhanced Interpretability (XAI): The DFA-SVM's attention mechanisms generate visual heatmaps, functioning as an "interpretive soft biopsy" analogue, highlighting diagnostically salient regions in endoscopic images. The M2F2-Net's Fuzzy Fusion Layer, with its linguistically interpretable rules, offers transparent reasoning for multi-modal integration. This "glass box" approach fosters clinician trust, a critical factor for clinical adoption.
- Robust Uncertainty Handling: The M2F2-Net's Fuzzy Fusion Layer explicitly manages diagnostic uncertainty by dynamically weighing modalities. This makes the model robust to noisy, incomplete, or ambiguous data from any single source, a common challenge in real-world clinical settings.
- Superior Early Detection: The demonstrated high sensitivity and low false negative rates for Early Gastric Cancer directly address a critical unmet need in current diagnostic practice, offering a pathway to curative intervention.
- **Limitations**
- Computational Complexity: Hybrid models, particularly M2F2-Net with its multi-stream processing and fuzzy logic, demand substantial computational resources for training and inference. This poses a challenge for de-ployment in environments with limited hardware infrastructure and requires specialized hardware [28]. Extensive time is required for computational operations, especially when preprocessing high-resolution images [29].
- Data Availability and Annotation: The performance of multi-modal hybrid models

relies heavily on large, highquality, meticulously annotated multi-modal datasets. Curating such integrated datasets (like the hypothetical GIMIC-12K) is a significant practical challenge due to existing data silos and stringent privacy regulations.
- Generalizability to Diverse Populations: While GIMIC-12K is conceptually diverse, real-world clinical data exhibits variations across different geographic regions and ethnic populations. The models' performance may degrade when deployed in populations significantly different from those in the training data, necessitating rigorous validation across diverse cohorts.
- Dynamic Nature of Disease: Gastric lesions evolve over time. While M2F2-Net's fuzzy logic offers some adaptability, a truly dynamic system necessitates continuous learning and potentially a Graph Neural Network (GNN) with Reinforcement Learning (RL) for optimal adaptive diagnostic pathways, which is beyond the scope of the current implementation.

Interpretability and Clinical Adoption

The interpretability (Explainable AI-XAI) features of our hybrid models are not merely academic curiosities; they are foundational for clinical adoption. Clinicians, inherently responsible for patient care, require understanding and trust in any diagnostic tool. The "black box" nature of many deep learning models remains a significant barrier to this trust. By providing visual explanations (e.g., Grad-CAM heatmaps highlighting suspicious regions on endoscopic images) and linguistically transparent reasoning (via fuzzy rules), our hybrid models bridge this gap. This transparency allows clinicians to:

- Verify Predictions: Cross-reference AI insights with their own expertise and clinical context.
- Identify Novel Biomarkers: Discover subtle patterns the AI identifies that human eyes might miss.
- Train Junior Clinicians: Use AI explanations as teaching tools for diagnostic reasoning.
- Build Trust: Develop confidence in the AI's recommendations, moving from skepticism to collaborative decisionmaking.
  This emphasis on XAI is a strategic design choice, acknowledging that diagnostic accuracy, while paramount, must be coupled with intelligibility to facilitate seamless integration into routine clinical workflows.

Summary of Key Contributions

This research investigates the application of novel hybrid Artificial Intelligence algorithms for the early and precise diagnosis of gastric ulcer and gastric cancer. We introduced two innovative hybrid models, Deep Feature-Attention SVM (DFA-SVM) and the Multi-Modal Fuzzy-Fusion Network (M2F2-Net), which leverage the synergistic integration of multi-modal data (endoscopic images, histopathology features, and structured clinical data) and advanced reasoning mechanisms. Through rigorous simulated experimentation on the GIMIC-12K dataset, we demonstrated that these hybrid architectures, particularly the M2F2-Net, achieve significantly superior diagnostic performance across all key metrics (accuracy, sensitivity, specificity, precision, F1-score, and AUC) compared to traditional machine learning and standalone deep learning baselines. This superiority is most pronounced in the critical task of early gastric cancer detection, where the M2F2-Net exhibited a substantial reduction in false negatives.

Practical Implications

The practical implications of these findings are profound. The enhanced diagnostic precision and interpretability offered by hybrid AI models can lead to earlier detection of gastric malignancies, enabling less invasive and more effective treatments, thereby dramatically improving patient outcomes and quality of life. Furthermore, these systems promise to enhance healthcare efficiency by reducing diagnostic delays, optimizing endoscopy and pathology workflows, and facilitating more personalized and proactive patient management. The emphasis on Explainable AI (XAI) features, such as attention maps and linguistically interpretable fuzzy rules, is crucial for fostering clinician trust and accelerating the integration of these advanced tools into clinical practice.

Future Research Directions

While this study provides compelling evidence for the efficacy of hybrid AI in gastric

diagnostics, several avenues for future research warrant exploration:

- Larger-Scale Prospective Clinical Validation: The most immediate and critical next step involves validating these hybrid models on large, diverse, real-world prospective clinical datasets from multiple institutions and patient populations. This will confirm their generalizability and robustness in heterogeneous clinical settings.
- Real-time Diagnostic Systems: Developing and optimizing these hybrid models for real-time inference during live endoscopy procedures is a key objective. This requires further advancements in computational efficiency and integration with existing endoscopy hardware and software.
- Integration with Hospital Information Systems (HIS): Seamless integration with HIS and Electronic Health Records (EHR) is essential for practical deployment. This involves addressing interoperability standards, data security, and workflow integration challenges.
- Longitudinal Patient Monitoring and Prognosis: Extending the M2F2-Net to incorporate temporal data (e.g., serial endoscopy images, evolving clinical parameters) to track lesion progression and predict prognosis over time. This could involve integrating Graph Neural Networks (GNNs) to model spatio-temporal relationships between lesions and potentially Reinforcement Learning (RL) for adaptive, personalized diagnostic and management pathways.
- Quantification of Uncertainty and Confidence Calibration: Further refining the fuzzy logic component to provide more granular and clinically meaningful quantification of diagnostic uncertainty and confidence scores. This will enhance decision-making in ambiguous cases.
- Federated Learning for Data Privacy: Exploring federated learning approaches to train these hybrid models across multiple institutions without centralizing sensitive patient data, addressing privacy concerns and facilitating collaborative research on larger datasets. This research establishes a strong foundation, demonstrating that hybrid AI is not merely an incremental improvement but a transformative force in medical diagnostics. The continued development and rigorous validation of such intelligent systems promise a future where early, precise, and personalized diagnosis becomes the standard of care for gastric diseases.

Important Disclaimer

The hybrid AI models presented herein are research prototypes. They are not intended for direct clinical application without rigorous, large-scale prospective clinical validation, regulatory approval by relevant health authorities (e.g., FDA, EMA), and integration into a comprehensive clinical decision support system under medical supervision. The simulated data and results are for research and illustrative purposes only and do not constitute medical advice or a claim of clinical efficacy in real-world patient care.

References

1. Ahmed, I.A.; Senan, E.M.; Shatnawi, H.S.A. "Hybrid Models for Endoscopy Image Analysis for Early Detection of Gastrointestinal Diseases Based on Fused Features." Diagnostics, vol. 13, no. 10, 1758, 2023. https://doi.org/10.3390/diagnostics13101758.
2. Dahan, F., Shah, J.H., Saleem, R., Hasnain, M., Afzal, M. & Alfakih, T.M. "A hybrid XAI-driven deep learning framework for robust GI tract disease diagnosis." Scientific Reports, vol. 15, no. 1, 21139, 2025. https://www.nature.com/articles/s41598-025-07690-3.
3. Dosovitskiy, A., et al. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations (ICLR), 2021.
4. Fadl Dahan, Jamal Hussain Shah, Rabia Saleem, Muhammad Hasnain, Maira Afzal & Taha M. Alfakih. "A hybrid XAIdriven deep learning framework for robust GI tract disease diagnosis." Scientific Reports, vol. 15, Article number: 21139, 2025. https://www.nature.com/articles/s41598-025-07690-3.
5. Figueiredo, P., Figueiredo, I., Pinto, L., Kumar, S., Tsai, Y.-H., & Mamonov, A. "Polyp detection with computer-aided

diagnosis in white light colonoscopy: comparison of three different methods." Endoscopic International Open, vol. 7, no. 2, E209-E215, 2019. https://doi.org/10.1055/a-0808-4456.

6. He, K., et al. "Deep Residual Learning for Image Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770-778.
7. Huang, G., et al. "Densely Connected Convolutional Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 4700-4708.
8. Huang, Y., Shao, Y., Yu, X., Chen, C., Guo, J., & Ye, G. "Global progress and future prospects of early gastric cancer screening." Journal of Cancer, vol. 15, no. 10, 3045-3064, 2024. https://www.jcancer.org/v15p3045.pdf.
9. Kim, S., Lee, N., Lee, J., Hyun, D., Park, C. "Heterogeneous Graph Learning for Multi-Modal Medical Data Analysis." Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 12, 5141-5150, 2023. https://ojs.aaai.org/index.php/AAAI/article/view/25643/25415.
10. Kim, W., et al. "Artificial Intelligence in Gastric Cancer Imaging With Emphasis on Diagnostic Imaging and Body Morphometry." Journal of Gastric Cancer, vol. 23, no. 3, 388-399, 2023. https://pubmed.ncbi.nlm.nih.gov/37553127/.
11. Ijaz, M. F. "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda." Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 7, 8459-8486, 2022. https://pmc.ncbi.nlm.nih.gov/articles/PMC8754556/.
12. Macenko, M., et al. "A method for normalizing histology slides for quantitative analysis." IEEE International Symposium on Biomedical Imaging (ISBI), 2009, 1107-1110.
13. Mohammad Amin Morid, Alireza Borjali, Guilherme Del Fiol. "A scoping review of transfer learning research on medical image analysis using ImageNet." arXiv, 2020, pp. 1-38.https://arxiv.org/pdf/2004.13175.
14. Pal, A.; Rai, H.M.; Frej, M.B.H.; Razaque, A. "Advanced Segmentation of Gastrointestinal (GI) Cancer Disease Using a Novel U-MaskNet Model." Life, vol. 14, no. 9, 1488, 2024. https://pdfs.semanticscholar.org/eoa7/874cb7f2eac933309fb3f2b2ab2b19ed30da.pdf.
15. Parkash, O., Siddiqui, A. T. S., Jiwani, U., Rind, F., Padhani, Z. A., Rizvi, A., Hoodbhoy, Z., & Das, J. K. "Diagnostic accuracy of artificial intelligence for detecting gastrointestinal luminal pathologies: A systematic review and meta-analysis." Frontiers in Medicine, vol. 9, 1018937, 2022. https://ecommons.aku.edu/pakistan_fhs_mc_med_gastroenterol/318.