BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 13, No. 1, March 2025
Research Article

# Ensemble-Based Deep Transfer Learning for Robust Gastrointestinal Endoscopy Image Classification

Sehmus Aslan

***Abstract*— Gastrointestinal (GI) diseases remain a significant global health challenge, particularly in low-income settings where diagnostic resources are often scarce. Endoscopic examination is essential for detecting and monitoring these diseases, yet the manual analysis of the resulting images is time-consuming, prone to observer variability, and demanding of clinical expertise. Recent advances in computer-aided diagnosis (CAD) using deep convolutional neural networks (CNNs) have shown promise in automating endoscopic image classification, but limited annotated data and the subtlety of GI findings continue to pose challenges. To address these constraints, this study proposes a two-level stacking ensemble framework that combines three pre-trained CNN architectures—ResNet50, DenseNet201, and MobileNetV3Large—with four classical machine-learning meta-classifiers (Logistic Regression, Random Forest, Support Vector Machine, and K-Nearest Neighbors). The KvasirV2 dataset, comprising 8,000 GI endoscopic images across eight classes, was used to train and evaluate the models. Results indicate that the stacking ensemble achieved a top accuracy of 94.33%, surpassing individual CNN baselines by 1–2%. Notably, this multi-level ensemble approach demonstrated improved diagnostic consistency for challenging classes like early-stage esophagitis and normal Z-line, suggesting that synergizing diverse CNN feature extractors can mitigate the limitations of single-network methods. These findings underscore the potential of ensemble-based transfer learning to enhance clinical decision support, reduce observer variability, and facilitate earlier, more accurate detection of GI diseases.**

*Index Terms*—: **Ensemble Learning, Transfer learning, Gastrointestinal Endoscopy, Deep Convolutional Neural Networks (CNNs), Computer-Aided Diagnosis (CAD)**

## I. Introduction

GASTROINTESTINAL(GI) DISEASES pose a major global health concern, ranking as the seventh leading cause of death in low-income countries in 2021, according to the World Health Organization (WHO) [1]. The diversity of gastrointestinal disorders, ranging from mild inflammatory conditions to life-threatening cancers, underscores the need for accurate and timely diagnosis to prevent complications and improve patient outcomes.

A primary method for diagnosing GI diseases is endoscopic examination, which involves the use of a flexible tube with an attached camera to visualize the GI tract in real time. While this technique facilitates direct observation of the esophagus, stomach, and intestines, and permits biopsy or treatment during the procedure, it also places substantial demands on clinical resources. Gastroenterologists spend considerable time interpreting large numbers of images or videos, which may result in increased workload and heightened fatigue. Moreover, visual assessment is inherently susceptible to inter- and intra-observer variability, as clinicians may reach different conclusions depending on their expertise, training, or even geographic practice environment.

In response to these diagnostic challenges, computer-aided diagnosis (CAD) systems have gained momentum as valuable clinical support tools. By harnessing the power of artificial intelligence (AI), these systems can potentially standardize diagnostic criteria, detect abnormalities more consistently, and expedite the process of disease identification. Specifically, deep convolutional neural networks (CNNs) have shown promise for classifying endoscopic images, as they autonomously extract relevant features from raw data, thereby reducing the need for manual feature engineering. However, one persistent obstacle in developing robust deep learning models in medicine is the limited availability of annotated data, given that privacy regulations often constrain the sharing of medical images among institutions.

To address this limitation, transfer learning has emerged as a strategic approach. Rather than training a deep neural network from scratch, researchers leverage pre-trained models from large-scale datasets—such as ImageNet—and adapt them to medical image classification tasks [2], [3]. This process not only reduces computational burden but also allows the model to inherit feature representations from millions of natural images, improving performance on smaller and more specialized medical datasets. Overall, the confluence of increased GI disease prevalence, the growing volume of endoscopic data, and advancements in deep learning highlights an urgent need for integrating AI-driven diagnostic tools into clinical practice. By doing so, healthcare systems can potentially detect pathologies earlier, reduce clinician workload, and enhance patient care. A summary of the studies in the literature is as follows:

**Sehmus Aslan**, is with Department of Business Administration of Mardin Artuklu University, Mardin, Turkey,(e-mail: sehmusaslan@artuklu.edu.tr).
https://orcid.org/0000-0003-1886-3421

Manuscript received Jan 31, 2025; accepted Feb 20, 2025.
DOI: 10.17694/bajece.1630294

Gebreslassie et al. [4] compared DenseNet121 and ResNet50 on a subset of 2,000 images from the Kvasir v2 dataset, employing a split ratio of 0.6, 0.3, and 0.1 for training, testing, and validation, respectively. Their findings indicated that ResNet50 attained the highest accuracy of 87.8%, thereby demonstrating the potential utility of transfer learning for gastrointestinal (GI) endoscopic image classification. Poudel et al. [5] employed ResNet50 with scaled-dilation convolutions to classify 4,000 endoscopic images into eight categories, achieving an accuracy of 88%. Their methodology incorporated a batch size of 16, a learning rate of 0.001, and stochastic gradient descent, illustrating the importance of meticulous hyperparameter selection to mitigate overfitting in convolutional neural networks (CNNs). Lonseko et al. [6] adopted an attention-guided CNN incorporating spatial attention and encoder-decoder layers on the Kvasir dataset, achieving 93.19% accuracy and an F1 score of 92.8%. They addressed data imbalance via augmentation strategies, underscoring the relevance of data diversity in model training. Musha et al. [7] investigated 16 pre-trained models, including MobileNetV2, on 2,000 Kvasir v2 images focused on dyed lifted polyps and resection margins. MobileNetV2 performed best, reaching an accuracy of 82.25% under a learning rate of 0.001 with Adam. Auzine et al. [8] explored InceptionV3, InceptionResNetV2, and VGG16 on 9,852 images from the Endoscopic Artifact Detection and Kvasir v2 datasets, reporting 77.65% accuracy with InceptionV3. In a similar vein, Gupta et al. [9] proposed a hybrid architecture combining EfficientNetB7 and ResNet50 to classify 8,000 Kvasir v2 images, achieving 88.19% accuracy. These studies collectively highlighted the continued success of transfer learning in GI endoscopic tasks and the inherent challenges of avoiding overfitting.

Mukhtorov et al. [10] examined DenseNet201, MobileNetV2, ResNet18, ResNet152, and VGG16 on 8,000 wireless endoscopic images, identifying overfitting in ResNet152, with a training accuracy of 98.28% versus a validation accuracy of 93.46%. Gunasekaran et al. [2] reported analogous issues using an ensemble of DenseNet201, InceptionV3, and ResNet50 obtaining 95% accuracy but with diminished generalization on validation data. Demirbaş et al. [11] developed a Spatial-Attention ConvMixer (SAC) architecture, surpassing Vanilla ViT, Swin Transformer, and the baseline ConvMixer on the Kvasir dataset, with a final accuracy of 93.37%. This study demonstrated the efficacy of spatial attention mechanisms in enhancing classification performance. In parallel, Ayan [12] investigated the classification of gastrointestinal diseases using thirteen different CNN models and two different ViT architectures. The authors observed that while ViT models reached accuracies of 91.25% and 90.50%, a well-optimized DenseNet201 variant, leveraging optimized transfer learning parameters, recorded an accuracy of 93.13%, a recall of 93.17%, a precision of 93.13%, and an F1 score of 93.11%, thereby outperforming both ViT models. Similarly, Huo et al. [13] proposed Self-Peripheral-Attention (SPA), inspired by human peripheral vision, to improve classification and segmentation on Kvasir and Kvasir-SEG datasets, attaining an accuracy of 92.7%.

If the literature is examined, it can be seen that despite the demonstrated success of transfer learning and deep CNNs in classifying GI endoscopic images, several notable gaps remain unaddressed. First, most existing studies rely on either a single deep network or straightforward ensemble averaging, without systematically exploiting more advanced multi-level ensemble frameworks. As a result, valuable complementary features learned by different architectures may not be fully leveraged, especially for visually subtle classes such as early-stage esophagitis and the Z-line. Second, while overfitting and limited annotated data are frequently acknowledged challenges in GI image analysis, there is comparatively little research into robust strategies—beyond basic augmentation—for mitigating these issues across diverse endoscopic conditions. Finally, few works provide a detailed examination of how stacking ensembles with classical machine-learning meta-classifiers can improve diagnostic consistency and reduce the variance inherent in individual CNNs. Addressing these gaps could lead to more precise classification performance, particularly in clinically challenging contexts where subtle tissue changes are critical for early diagnosis.

This study contributes to the gastrointestinal (GI) endoscopy classification literature in several key ways. First, it proposes a two-level stacking ensemble approach, systematically combining multiple state-of-the-art CNN architectures (ResNet50, DenseNet201, MobileNetV3Large) with classical meta-classifiers (Logistic Regression, Random Forest, SVM, KNN). By moving beyond single-network solutions and basic ensemble averaging, the method fully exploits complementary learned features, which is particularly important for handling the subtle visual distinctions in GI images such as early-stage esophagitis and normal Z-line. Second, the detailed comparison of base CNNs against multiple stacking ensembles offers new insights into how meta-classifiers can improve diagnostic consistency, bridging an existing gap in the literature on advanced ensemble frameworks for GI endoscopic image analysis. By presenting robust evidence that such multi-level ensembles outperform individual CNNs, the study sets a foundation for future research aimed at achieving more accurate and clinically relevant GI disease detection systems.

## II. MATERIALS AND METHODS

### A. *Dataset*

This study employs the Kvasir-V2 dataset [14], a collection of gastroenterological endoscopic images gathered by a Norwegian healthcare organization, designed to facilitate research in medical image analysis. The Kvasir-v2 dataset is a comprehensive collection of 8,000 gastrointestinal tract endoscopic images, released in 2017 through the MediaEval Medical Multimedia Challenge. The dataset contains eight balanced classes with 1,000 images each, all annotated and verified by certified endoscopists. These classes are divided into three main categories: anatomical landmarks (pylorus, z-line, and cecum), pathological findings (esophagitis, ulcerative colitis, and polyps), and medical procedures (dyed lifted polyps and dyed resection margins). The classes and image examples used for the application are presented in Fig. 1. The images vary in resolution from 720×576 pixels to 1920×1072 pixels, with

some containing annotations in the leftmost quarter and green boxes indicating endoscope location. Each image varies in capture angle, brightness, zoom level, resolution, and centerpoint, making it a challenging dataset for deep learning applications. Despite its relatively small size compared to standard deep learning datasets, Kvasir-v2 has become a crucial benchmark dataset for evaluating machine learning approaches in gastrointestinal image analysis, particularly for testing classification accuracy, developing computer-aided diagnosis tools, and assessing model generalization capabilities. The dataset's standardized format and public availability through Kaggle [15] make it particularly valuable for research in automated gastrointestinal disease detection and medical image classification systems. By applying a random sampling strategy, each category was partitioned into training and test sets at a 70:30 ratio, yielding 5,600 images for training and 2,400 images for testing.

To enhance the robustness and generalization capability of the model, a data augmentation pipeline was implemented, applying a series of random transformations to the input images during training. Each image undergoes random rotations within the range $[-15°, 15°]$ and discrete 90-degree increments, along with adjustments to brightness, contrast, saturation, and hue to introduce variability in illumination and color. Small translations are simulated by padding the image by 20 pixels and cropping it back to its original dimensions, while additive Gaussian noise with a mean of 0.0 and a standard deviation of 0.1 is introduced to mimic real-world imperfections. These augmentations are applied dynamically during training, ensuring that the model is exposed to a diverse range of variations, thereby improving its ability to generalize and reducing the risk of overfitting. Originally, there were 5,600 training instances; after applying these augmentations at a rate of 5×, the total number of augmented samples increases to 28,000. The test dataset remains unchanged at 2,400 samples. All images were resized at 224×224x3 pixels..

### *B. Transfer Learning*

Transfer learning is highly beneficial in medical imaging tasks primarily because acquiring large, well-annotated datasets in clinical settings can be challenging due to patient privacy concerns, labeling costs, and the specialized expertise needed for annotation [16]. By leveraging models pre-trained on extensive and diverse non-medical image datasets—such as ImageNet—researchers can repurpose learned features (e.g., edges, textures) and adapt them to medical contexts. This process not only saves significant computation time and resources but also mitigates the risk of overfitting when working with relatively small medical datasets [17]. According to Litjens et al. [18], transfer learning facilitates faster convergence and can enhance classification or detection accuracy in a wide range of medical imaging applications, from lesion identification to organ segmentation. Numerous CNN architectures have been introduced in the literature. In this work, three pre-trained CNN models ( ResNet50, DenseNet201 and MobileNetV3Large) are employed to categorize endoscopic images into eight distinct classes.

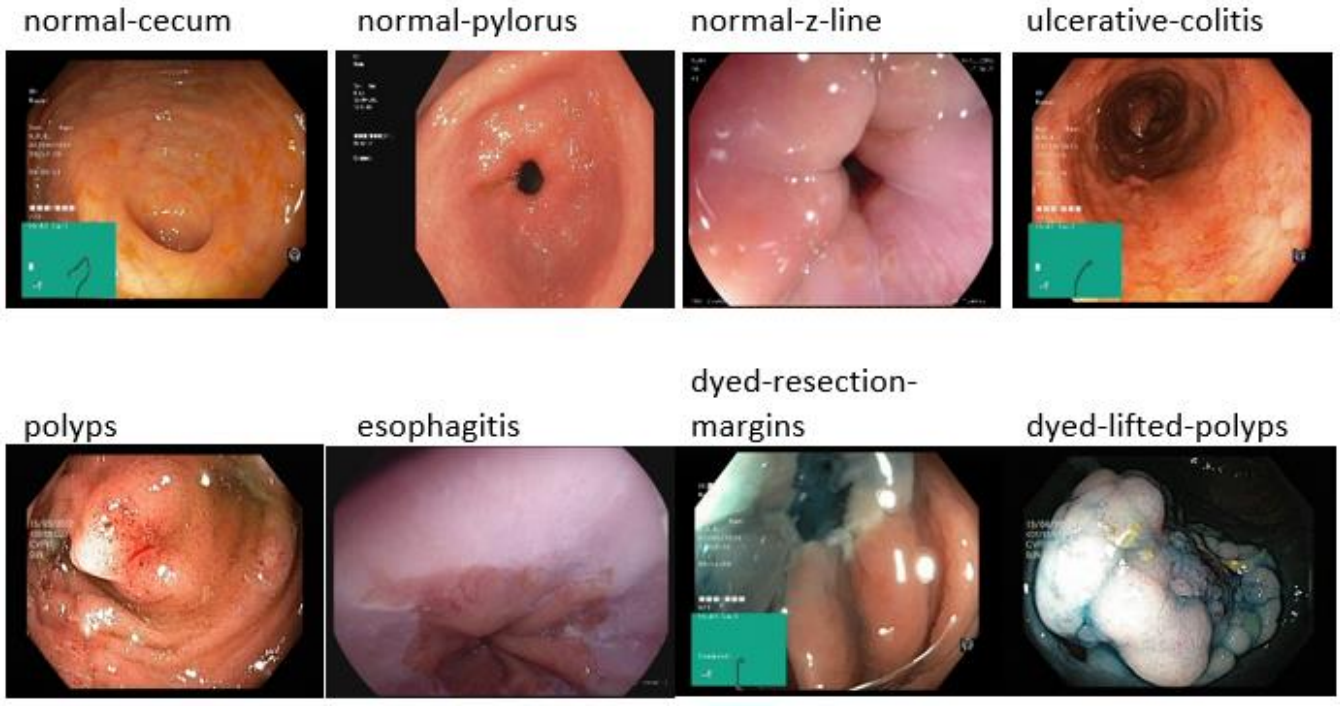


Fig.1. KvasirV2 classes

*1. ResNet50*

ResNet50 is a 50-layer convolutional neural network introduced by He et al. [19] as part of the ResNet (Residual Network) family, which was primarily designed to overcome the vanishing gradient problem in deeper neural networks. By incorporating residual blocks with identity connections (or skip connections), the architecture allows gradients to flow more effectively, facilitating the training of much deeper networks than earlier CNN models like VGGNet or AlexNet. Owing to its balance between depth and computational efficiency, ResNet50 has become a widely adopted backbone for various computer vision tasks, including image classification, object detection, and segmentation [20]. In medical image analysis, the model's pre-trained weights on large-scale datasets (e.g., ImageNet) have proven highly beneficial when performing transfer learning with limited labeled data, enhancing accuracy and accelerating convergence [18].

*2. DenseNet201*

DenseNet201 is a 201-layer convolutional neural network that is part of the Densely Connected Convolutional Network (DenseNet) family introduced by Huang et al. [20]. Unlike traditional CNNs, DenseNet layers are directly connected to every other layer in a feed-forward manner, allowing each layer to access the gradients from all preceding layers. This dense connectivity pattern mitigates the vanishing gradient problem and promotes feature reuse, enabling the construction of deeper and more efficient networks without a substantial increase in computational complexity [20]. DenseNet201 has demonstrated strong performance on large-scale image classification tasks such as ImageNet. In medical imaging, transferring pre-trained

DenseNet201 weights to specialized domains has shown to accelerate model convergence and enhance diagnostic accuracy, especially when the available datasets are relatively small [18]. Its depth and ability to capture complex hierarchical features make DenseNet201 particularly suitable for tasks like disease classification, segmentation, and detection, where subtle variations in medical images are critical for accurate predictions.

*3. MobileNetV3Large*

MobileNetV3Large is a lightweight convolutional neural network introduced as part of the MobileNetV3 family by Howard et al. [21]. It was designed through a combination of neural architecture search (NAS) and platform-aware model optimization (NetAdapt), balancing high accuracy with reduced computational complexity. Key features include the use of squeeze-and-excitation (SE) modules, novel activation functions such as the h-swish, and inverted residual blocks that improve both efficiency and representational power [21]. While originally optimized for resource-constrained devices (e.g., smartphones), MobileNetV3Large has also shown promise in medical imaging contexts, particularly when transferring pre-trained weights to smaller medical datasets for tasks like classification and segmentation [18]. Its efficient design enables faster inference and lower resource usage, which are critical factors for real-time, point-of-care diagnostics.

### *C. Proposed Stacking Ensemble Model*

The proposed framework leverages transfer learning to construct a robust and scalable deep learning model for multi-class classification. Transfer learning is employed to utilize the feature extraction capabilities of pre-trained convolutional neural networks (CNNs), which have been trained on the large-scale ImageNet dataset. This approach not only reduces the computational cost of training from scratch but also enhances the model's ability to generalize to new datasets, particularly when labeled data is limited.

Three state-of-the-art CNN architectures are explored as backbone models: ResNet50, DenseNet201, and MobileNetV3Large. These architectures are chosen for their proven performance in various computer vision tasks, with each offering unique advantages:

1. *ResNet50:* Known for its residual learning framework, which mitigates the vanishing gradient problem and enables the training of very deep networks.
2. *DenseNet201:* Utilizes dense connections between layers, promoting feature reuse and improving parameter efficiency.
3. *MobileNetV3Large:* Designed for efficiency, this architecture is optimized for mobile and edge devices, offering a balance between accuracy and computational cost.

Each backbone model is initialized with pre-trained weights from ImageNet and configured to exclude the fully connected classification head. This allows the model to retain only the feature extraction layers, which are then adapted to the specific task at hand. Global average pooling is applied to the output feature maps to reduce spatial dimensions and produce a fixed-size feature vector. This is followed by a Flatten layer to convert the pooled features into a one-dimensional vector, which is then passed to a Dense layer with eight output units and a softmax activation function. This final layer enables multi-class classification into eight distinct categories.

All layers of the base models are set to be trainable, allowing for fine-tuning of the pre-trained weights during training. This ensures that the model can adapt to the specific characteristics of the target dataset while retaining the generalizable features learned from ImageNet. The models are optimized using Stochastic Gradient Descent (SGD) with a learning rate of 0.01, momentum of 0.9, and Nesterov acceleration. The loss function is defined as categorical cross-entropy, which is well-suited for multi-class classification tasks. Model performance is evaluated using accuracy as the primary metric.

To further enhance classification performance and robustness, a stacking ensemble approach is employed. Stacking combines the predictions of multiple base models (in this case, the three deep learning models) using a meta-classifier, which learns to optimally weigh and combine the predictions. This approach leverages the strengths of diverse models, reducing the risk of overfitting and improving generalization.

The predictions from the three deep learning models are concatenated to form a combined feature representation, which serves as input to the meta-classifier. Four distinct meta-

classifiers are implemented and evaluated, each chosen for its unique characteristics and suitability for the task:

1 *Logistic Regression (LR):* A linear classifier with L2 regularization (C=10) and the newton-cg solver, configured for one-vs-rest multi-class classification. Logistic regression is chosen for its interpretability and efficiency in handling linearly separable data.
2 *Random Forest (RF):* An ensemble of 300 decision trees with Gini impurity as the splitting criterion, a maximum depth of 20, and balanced class weights to handle class imbalance. Random forests are robust to overfitting and capable of capturing complex, non-linear relationships in the data.
3 *Support Vector Machine (SVM):* A linear SVM with a regularization parameter (C=0.1) and a maximum of 100 iterations, configured to output probability estimates. SVMs are known for their ability to find optimal decision boundaries in high-dimensional spaces.
4 *k-Nearest Neighbors (k-NN):* A non-parametric classifier with three neighbors, Manhattan distance as the metric, and uniform weighting. k-NN is simple yet effective, particularly for datasets with well-defined clusters.

Each meta-classifier is preceded by a StandardScaler to normalize the input features, ensuring consistent scaling across the concatenated predictions. This preprocessing step is critical for algorithms like SVM and k-NN, which are sensitive to the scale of input features.

TABLE I
HYPERPARAMETERS FOR DEEP LEARNING MODELS

| Hyperparameter | Value/Configuration |
|---|---|
| Backbone Architectures | ResNet50, DenseNet201, MobileNetV3Large |
| Pre-trained Weights | ImageNet |
| Include Top | False (exclude fully connected layers) |
| Pooling | Global Average Pooling |
| Classifier Activation | Softmax |
| Trainable Layers | All layers trainable |
| Optimizer | Stochastic Gradient Descent (SGD) |
| Learning Rate | 0.01 |
| Momentum | 0.9 |
| Nesterov Acceleration | Enabled |
| Loss Function | Categorical Crossentropy |
| Metrics | Accuracy |
| Epochs | 30 |
| Batch Size | 32 |
| Early Stopping | Patience = 10 (monitor validation loss) |
| Learning Rate Scheduler | ReduceLROnPlateau (factor = 0.2, patience = 5, min_lr = 1e-5) |
| Model Checkpoint | Save best weights based on validation accuracy |

The meta-classifiers are trained on the combined predictions from the deep learning models and evaluated using standard classification metrics, including accuracy, precision, recall and F1 score. These metrics provide a comprehensive assessment of model performance. The hyperparameters for the deep learning models are provided in Table 1, while those for the meta-classifiers are detailed in Table 2.

TABLE II
HYPERPARAMETERS FOR META-CLASSIFIERS

| Meta-Classifiers | Hyperparameter | Value/Configuration |
|---|---|---|
| Logistic Regression | Regularization (C) | 10 |
| | Solver | newton-cg |
| | Max Iterations | 100 |
| | Multi-Class Strategy | One-vs-Rest (OvR) |
| Random Forest | Number of Estimators | 300 |
| | Criterion | Gini Impurity |
| | Max Depth | 20 |
| | Max Features | sqrt |
| | Min Samples Split | 2 |
| | Min Samples Leaf | 1 |
| | Bootstrap | True |
| | Class Weight | Balanced |
| Support Vector Machine (SVM) | Regularization (C) | 0.1 |
| | Kernel | Linear |
| | Max Iterations | 100 |
| | Probability Estimates | Enabled |
| k-Nearest Neighbors (k-NN) | Number of Neighbors (k) | 3 |
| | Distance Metric | Manhattan (p=1) |
| | Weights | Uniform |

## III. RESULTS

### A. Evaluation Metrics

The proposed ensemble model is evaluated using key performance metrics, including *accuracy*, *precision*, *recall*, and *F1 score*. In multiclass classification, where the number of classes exceeds two, the predictions generated by the model can be either correct or incorrect for each class. To evaluate the model's performance, the predictions are analyzed based on the following classification states for each class:

- *True Positive (TP):* The model correctly predicts the class of interest.
- *True Negative (TN):* The model correctly identifies instances that do not belong to the class of interest.
- *False Positive (FP):* The model incorrectly predicts an instance as belonging to the class of interest.
- *False Negative (FN):* The model fails to identify an instance that belongs to the class of interest.

For multiclass classification, these metrics are typically computed using a one-vs-rest approach, where each class is evaluated against the rest of the classes. The formulas for the evaluation metrics are defined as follows:

$$Accuracy = \frac{\sum_{i=1}^{C}(\mathrm{TP}_i + \mathrm{TN}_i)}{\sum_{i=1}^{C}(\mathrm{TP}_i + \mathrm{TN}_i + \mathrm{FP}_i + \mathrm{FN}_i)} \quad (1)$$

$$Precision = \frac{\sum_{i=1}^{C} \mathrm{TP}_i}{\sum_{i=1}^{C} (\mathrm{TP}_i + \mathrm{FP}_i)} \quad (2)$$

$$Recall = \frac{\sum_{i=1}^{C} \mathrm{TP}_i}{\sum_{i=1}^{C} (\mathrm{TP}_i + \mathrm{FN}_i)} \quad (3)$$

$$F1\ Score = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (4)$$

Here, $C$ represents the total number of classes, and the metrics are aggregated across all classes. *Accuracy* (1) is a metric that measures the overall correctness of the model's predictions across all classes. It is calculated as the ratio of the sum of true positives (TP) and true negatives (TN) for all classes to the total number of instances, including true positives, true negatives, false positives (FP), and false negatives (FN). *Precision* (2) quantifies the proportion of predicted instances for a class that are actually correct. *Recall* (3) measures the proportion of actual instances of a class that the model correctly identifies. The *F1 score* (4) provides a balanced measure by computing the harmonic mean of precision and recall, ensuring that both metrics are equally weighted in the evaluation.

### B. *Experimental Setup*

The experiment was conducted in Google Colab Pro with Python 3 Google Compute Engine backend (GPU-A100) with 40 GB GPU RAM. The deep learning models are trained on the KvasirV2 dataset for a maximum of 30 epochs, with early stopping implemented to prevent overfitting. Training is monitored using validation loss, and the learning rate is dynamically adjusted using the *ReduceLROnPlateau* callback, which reduces the learning rate by a factor of 0.2 if the validation loss does not improve for five consecutive epochs. The best model weights are saved based on validation accuracy using the *ModelCheckpoint* callback.

The meta-classifiers are trained on the concatenated predictions from the deep learning models, ensuring that they learn to effectively combine the strengths of each base model. The performance of the meta-classifiers is evaluated on a held-out test set derived from the KvasirV2 dataset, with results visualized using confusion matrices and summarized using classification metrics.

### C. *Test Results*

Fig. 2, 3, 4, 5, 6 and 7 present the accuracy and loss curves of the base models. The curves demonstrate that each architecture (ResNet, DenseNet, MobileNet) learns the dataset effectively, reaching high overall performance. The differences among models primarily manifest in how smoothly the validation accuracy evolves and how tightly the validation loss tracks the training loss. While the current results already achieve strong classification performance, the remaining gap between training and validation highlights a potential avenue for fine-tuning regularization or data augmentation strategies to bolster robustness further.

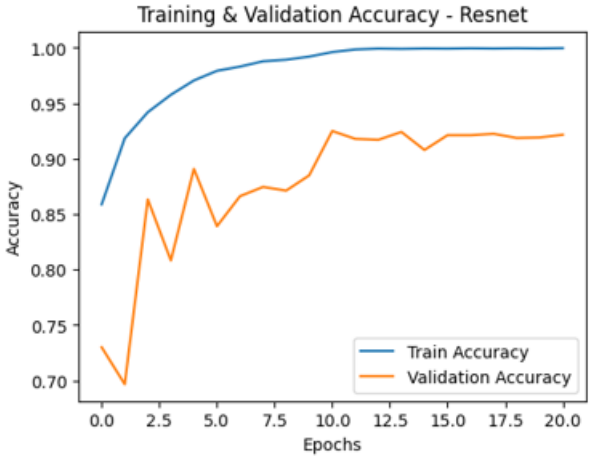


Fig.2. ResNet50 accuracy curve.


Fig.3. ResNet50 loss curve.

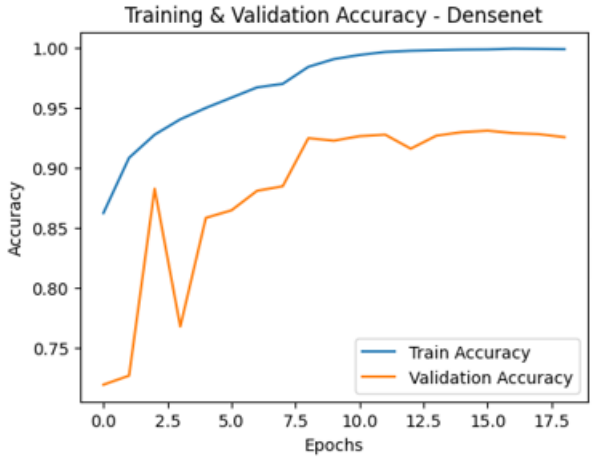


Fig.4 DenseNet201 accuracy curve.

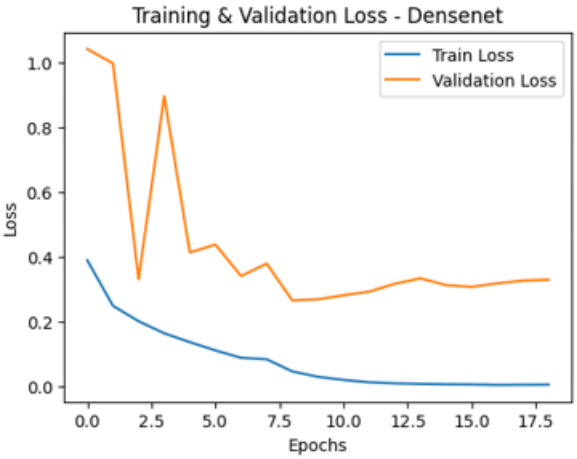


Fig.5 DenseNet201 loss curve.

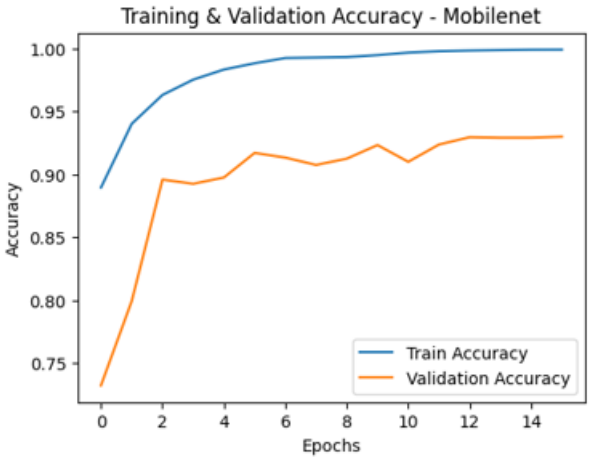


Fig.6 MobileNetV3Large accuracy curve.

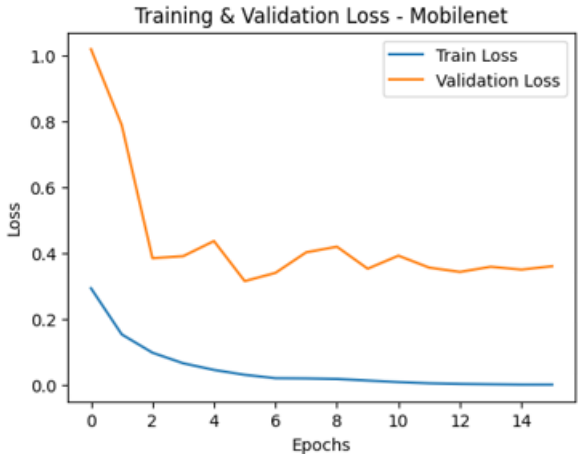


Fig.7 MobileNetV3Large loss curve.

Table III provides a comparison of the models' performances. Among the three base CNNs, MobileNetV3Large achieves the highest accuracy (93.00%), slightly outperforming ResNet50 (92.17%) and DenseNet201 (92.54%). This finding suggests that MobileNetV3Large, which balances depth and efficiency through its inverted residual blocks and attention mechanisms, adapts effectively to the Kvasir dataset. Logistic Regression, Random Forest, and SVM ensembles each hover around 94.3% accuracy, while KNN trails only slightly at 94.17%. Precision, recall, and F1 scores follow a similarly tight range. These minimal differences may reflect the relatively uniform effectiveness of combining three high-performing CNN backbones; once robust feature representations are available, multiple classical classifiers can exploit them effectively.

All stacking ensembles outperform their single CNN counterparts, with the best results reaching 94.33% accuracy (Stacking Ensemble with Logistic Regression or Random Forest). In other words, ensembling the probability outputs from ResNet50, DenseNet201, and MobileNetV3Large typically yields a 1–2% improvement in accuracy, precision, recall, and F1 score. From a healthcare perspective, slight performance gains can be critical, as more reliable diagnoses translate to fewer missed pathologies and better patient outcomes. This is particularly true in endoscopic procedures, where subtle changes can be indicative of early disease progression.

TABLE III
PERFORMANCE EVALUATION OF DEEP LEARNING MODELS ON KVASIR-V2 DATASET.

| Models | ACC(%) | Precision(%) | Recall(%) | $F_1\ Score$(%) |
|---|---|---|---|---|
| ResNet50 | 92.17 | 92.06 | 92.07 | 92.03 |
| DenseNet201 | 92.54 | 92.46 | 92.46 | 92.44 |
| MobileNetV3Large | 93.00 | 92.99 | 92.94 | 92.95 |
| Stacking Ensemble with SVM | 94.29 | 94.29 | 94.23 | 94.22 |
| Stacking Ensemble with LR | 94.33 | 94.33 | 94.27 | 94.25 |
| Stacking Ensemble with RF | 94.33 | 94.36 | 94.27 | 94.27 |
| Stacking Ensemble with KNN | 94.17 | 94.17 | 94.10 | 94.10 |

Fig. 8, 9, 10 and 11 show the confusion matrices of the stacking ensemble models. These matrices reveal that the stacking ensemble models can identify all classes with high performance, except for the Z-line and esophagitis classes. In many endoscopic images, the visual distinctions between a normal Z-line and mild esophagitis are very subtle, often manifesting as slight color changes or faint lesions. As a result, even advanced CNNs may struggle to reliably differentiate these two classes. The similarity between a normal Z-line and early-stage esophagitis—characterized by minor discoloration or subtle shifts in tissue texture—makes these distinctions less prominent compared to other classes (e.g., polyps or ulcers). Occasional misclassifications persist, indicating the presence of a few particularly challenging cases. Nevertheless, nearly all classes achieve high recall and precision, underscoring the overall robustness of the ensemble approach.

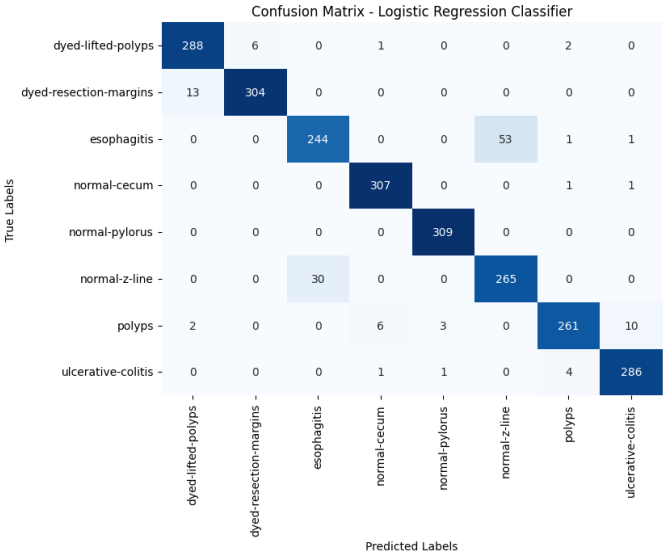


Fig.8 Confusion matrix of LR meta-classifier.

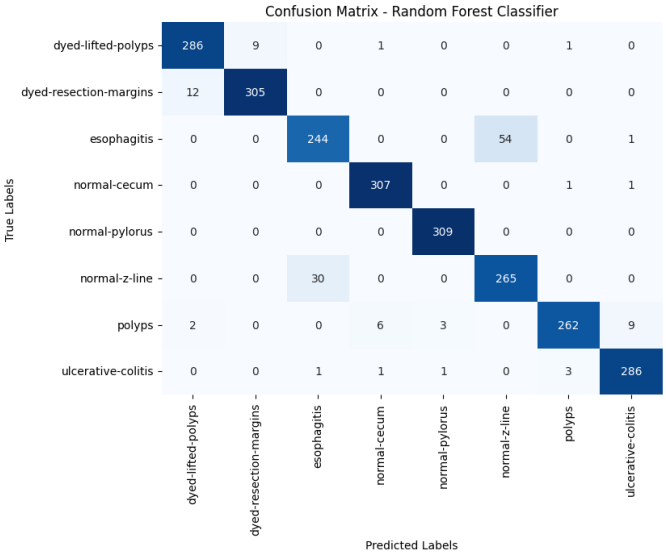


Fig.9 Confusion matrix of RF meta-classifier.

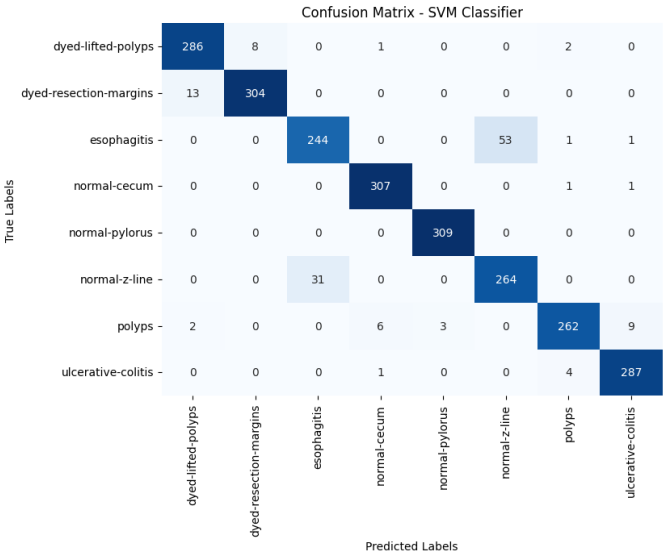


Fig.10 Confusion matrix of SVM meta-classifier.

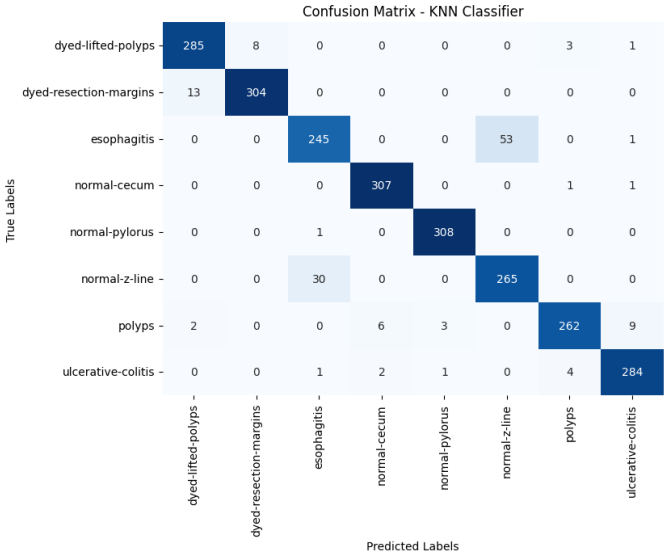


Fig.11 Confusion matrix of KNN meta-classifier.

Table IV compares performance metrics (Accuracy, Precision, Recall, F1-Score) across studies using the Kvasir dataset. Yogapriya et al. [22] achieved the highest accuracy (96.33%) and F1-score (96.50%), setting a strong benchmark. This study ranks second with competitive metrics (Accuracy: 94.33%, F1-Score: 94.27%), showing balanced performance across precision and recall. Other studies like Losenko et al. [6], Demirbaş et al. [11], and Huo et al. [13] also report strong results, while Gupta et al. [9] lags with metrics below 90%. Some studies, such as Mukhtorov et al. [10] and Gunasekaran et al. [2], only report accuracy, limiting comprehensive comparison. Overall, this study demonstrates robust performance, though Yogapriya et al. [22] remains the top performer. The findings highlight the promise of ensemble-based deep learning strategies and underscore the field-wide progress toward robust, clinically relevant models for gastrointestinal disease detection and classification.

TABLE IV
COMPARISON OF PROPOSED MODEL WITH OTHER RECENT MODELS

| Study | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Yogapriya et al [22] | 96.33 | 96.50 | 96.37 | 96.50 |
| Losenko et al. [6] | 93.19 | 92.8 | 92.7 | 92.8 |
| Gupta et al. [9] | 89.3 | 89 | 89.3 | 88.6 |
| Mukhtorov et al. [10] | 93.46 | - | - | - |
| Gunasekaran et al. [2] | 95.00 | - | - | - |
| Huo et al. [13] | 92.87 | 93.01 | 92.87 | 92.88 |
| Demirbaş et al. [11] | 93.37 | 93.66 | 93.37 | 93.42 |
| Ayan [12] | 93.13 | 93.17 | 93.13 | 93.11 |
| This study | 94.33 | 94.36 | 94.27 | 94.27 |

## IV. Discussions

The findings of this study reinforce the value of leveraging transfer learning and ensemble techniques for robust endoscopic image classification. As evidenced by the strong performance of individual CNN backbones (ResNet50, DenseNet201, MobileNetV3Large), pre-trained models offer a reliable starting point when working with relatively small yet challenging medical datasets such as Kvasir v2. The slight variations in baseline performance among these networks likely stem from differences in architecture design—ranging from ResNet's residual connections to DenseNet's dense connectivity and MobileNetV3's efficient inverted residual blocks—each of which provides unique advantages for feature extraction in endoscopic images.

Despite these variations, the proposed stacking framework demonstrates consistent improvements across accuracy, precision, recall, and F1 score. Such gains underscore the ensemble's ability to reconcile the complementary strengths of different CNNs. By uniting multiple feature representations at the meta-classifier level, the method mitigates the variance inherent in individual models and achieves a more robust overall performance. From a clinical perspective, even marginally higher metrics (1–2% above single-model baselines) can be particularly valuable in reducing missed pathologies, given the high-stakes nature of GI disease diagnosis.

The confusion matrices, however, highlight a recurrent challenge in distinguishing subtle classes like early-stage esophagitis versus a normal Z-line. This difficulty points to the inherent complexity of GI endoscopy images, where slight color shifts or minor morphological differences can be easily overlooked. Addressing this gap may require additional strategies, such as more targeted data augmentation, higher-resolution inputs, or region-of-interest (ROI) detection methods that emphasize the gastroesophageal junction. Likewise, incorporating advanced attention mechanisms or domain adaptation techniques may further refine the model's ability to capture faint textural changes indicative of mild esophagitis.

In a broader sense, the results align with existing literature that showcases the benefits of transfer learning in medical imaging, particularly when annotated data are scarce [16], [18]. Pre-trained weights allow the network to capitalize on foundational visual features, reducing the risk of overfitting and expediting convergence. Ensemble approaches, in turn, harness these strengths in a synergistic manner, as documented by related research that reports similar performance boosts when combining models [2].

Overall, this study's findings emphasize two key takeaways: first, that combining multiple CNN architectures through a stacking ensemble is effective in boosting classification metrics on the Kvasir dataset; and second, that additional focus on nuanced, easily confounded classes remains a priority for future work. By refining the proposed framework with enhanced data handling, attention modules, and specialized augmentation, researchers and clinicians can continue to push the boundaries of AI-driven GI diagnostics, ultimately contributing to earlier and more accurate detection of critical gastrointestinal conditions.

## V. Conclusion

This work demonstrates that ensemble-based transfer learning can significantly improve the classification of gastrointestinal endoscopic images, addressing both the scarcity of annotated data and the inherent complexity of subtle GI conditions. By combining ResNet50, DenseNet201, and MobileNetV3Large as base architectures and employing a second-level meta-classifier, we achieved higher accuracy, precision, recall, and F1 scores compared to single-network models. These findings underscore the synergy that arises when leveraging diverse CNN features in a stacking framework.

Moreover, the results highlight the practical benefits of enhanced diagnostic accuracy for conditions such as esophagitis and normal Z-line, where visual differences are often minimal. Although occasional misclassifications occur in these subtle classes, the overall performance points to the promise of refined augmentations, region-of-interest approaches, and advanced attention mechanisms in bridging the remaining performance gap.

From a clinical standpoint, the observed improvements in detection rates and classification reliability translate into potentially earlier interventions and reduced workload for gastroenterologists. Future research directions may focus on integrating larger, multi-center datasets, exploring novel attention modules, and automating the identification of key anatomical landmarks. By continuing to refine ensemble strategies and transfer learning pipelines, the field can move closer to real-time, AI-driven diagnostic support that is both efficient and clinically robust.

## References


[1] WHO, "The top 10 causes of death." Accessed: Jan. 30, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] H. Gunasekaran, K. Ramalakshmi, D. K. Swaminathan, A. J, and M. Mazzara, "GIT-Net: An Ensemble Deep Learning-Based GI Tract Classification of Endoscopic Images," *Bioengineering*, vol. 10, no. 7, p. 809, Jul. 2023, doi: 10.3390/bioengineering10070809.

[3] S. Mohapatra, J. Nayak, M. Mishra, G. K. Pati, B. Naik, and T. Swarnkar, "Wavelet Transform and Deep Convolutional Neural Network-Based Smart Healthcare System for Gastrointestinal Disease Detection," *Interdiscip. Sci. Comput. Life Sci.*, vol. 13, no. 2, pp. 212–228, Jun. 2021, doi: 10.1007/s12539-021-00417-8.

[4] A. KahsayGebreslassie, YaecobGirmayGezahegn, M. T. Hagos, AchimIbenthal, and Pooja, "Automated Gastrointestinal Disease Recognition for Endoscopic Images," in *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India: IEEE, Oct. 2019, pp. 312–316. doi: 10.1109/ICCCIS48478.2019.8974458.

[5] S. Poudel, Y. J. Kim, D. M. Vo, and S.-W. Lee, "Colorectal Disease Classification Using Efficiently Scaled Dilation in Convolutional Neural Network," *IEEE Access*, vol. 8, pp. 99227–99238, 2020, doi: 10.1109/ACCESS.2020.2996770.

[6] Z. M. Lonseko *et al.*, "Gastrointestinal Disease Classification in Endoscopic Images Using Attention-Guided Convolutional Neural Networks," *Appl. Sci.*, vol. 11, no. 23, p. 11136, Nov. 2021, doi: 10.3390/app112311136.

[7] A. Musha, R. Hasnat, A. A. Mamun, and T. Ghosh, "Deep Learning-Based Comparative Study to Detect Polyp Removal in Endoscopic Images," in *2022 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Pune, India: IEEE, Mar. 2022, pp. 1–5. doi: 10.1109/ESCI53509.2022.9758254.

[8] M. M. Auzine, P. Bissoonauth-Daiboo, M. H.-M. Khan, S. Baichoo, X. Gao, and N. G. Sahib, "Classification of artefacts in endoscopic images using deep neural network," in *2022 3rd International Conference on*

*Next Generation Computing Applications (NextComp)*, Flic-en-Flac, Mauritius: IEEE, Oct. 2022, pp. 1–5. doi: 10.1109/NextComp55567.2022.9932202.

[9] D. Gupta, G. Anand, P. Kirar, and P. Meel, "Classification of Endoscopic Images and Identification of Gastrointestinal diseases," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Faridabad, India: IEEE, May 2022, pp. 231–235. doi: 10.1109/COM-IT-CON54601.2022.9850571.

[10] D. Mukhtorov, M. Rakhmonova, S. Muksimova, and Y.-I. Cho, "Endoscopic Image Classification Based on Explainable Deep Learning," *Sensors*, vol. 23, no. 6, p. 3176, Mar. 2023, doi: 10.3390/s23063176.

[11] A. A. Demirbaş, H. Üzen, and H. Fırat, "Spatial-attention ConvMixer architecture for classification and detection of gastrointestinal diseases using the Kvasir dataset," *Health Inf. Sci. Syst.*, vol. 12, no. 1, p. 32, Apr. 2024, doi: 10.1007/s13755-024-00290-x.

[12] E. Ayan, "Classification of Gastrointestinal Diseases in Endoscopic Images: Comparative Analysis of Convolutional Neural Networks and Vision Transformers," *Iğdır Üniversitesi Fen Bilim. Enstitüsü Derg.*, vol. 14, no. 3, pp. 988–999, Sep. 2024, doi: 10.21597/jist.1501787.

[13] X. Huo, S. Tian, Y. Yang, L. Yu, W. Zhang, and A. Li, "SPA: Self-Peripheral-Attention for central–peripheral interactions in endoscopic image classification and segmentation," *Expert Syst. Appl.*, vol. 245, p. 123053, Jul. 2024, doi: 10.1016/j.eswa.2023.123053.

[14] K. Pogorelov *et al.*, "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection." Association for Computing Machinery, Haziran 2017. doi: 10.1145/3193289.

[15] Kaggle, "Kvasir v2." Accessed: Jan. 30, 2025. [Online]. Available: https://www.kaggle.com/datasets/plhalvorsen/kvasir-v2-a-gastrointestinal-tract-dataset

[16] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, doi: 10.1109/TMI.2016.2535302.

[17] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[18] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

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

[22] J. Yogapriya, V. Chandran, M. G. Sumithra, P. Anitha, P. Jenopaul, and C. Suresh Gnana Dhas, "Gastrointestinal Tract Disease Classification from Wireless Endoscopy Images Using Pretrained Deep Learning Model," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–12, Sep. 2021, doi: 10.1155/2021/5940433.

## BIOGRAPHIES

**Şehmus Aslan** received his undergraduate degree from İhsan Doğramacı Bilkent University, Department of Industrial Engineering in 2008. He completed his master's degree at Dicle University, Institute of Social Sciences, Department of Business Administration in 2015. In 2018, he began his doctoral studies at Hasan Kalyoncu University, Institute of Social Sciences, Department of Business Administration, and earned his Ph.D. in 2020. He is currently working as an Assistant Professor at Mardin Artuklu University, Faculty of Economics and Administrative Sciences, Department of Business Administration, Production Management Division. His research areas include operations research, combinatorial optimization, evolutionary algorithms and machine learning.