Modified CNN Approach for Enhanced Gastrointestinal Tract Classification

Dimple Sethi, Prashant Kumar, Milind Kashyap, Pratham Agarwal, Pranshu Saini, and Aashi Jadoun
School of Computer Science and Engineering, Bennett University, Greater Noida, India
Email: dimple.sethi@bennett.edu.in, prashant.kumar@bennett.edu.in, {e22cseu1317, e22cseu1308, e22cseu1300, e22cseu0450}@bennett.edu.in

***Abstract*—Gastrointestinal (GI) diseases affect a significant portion of the global population, making accurate diagnosis and mitigation of their impact on patient health essential. Traditionally, diagnosing GI diseases involves manual analysis of endoscopic images by medical professionals—a process that is not only time-consuming but also prone to variability in interpretation among different clinicians. To address these challenges, our research investigates the application of deep learning techniques for the automated classification of GI diseases using the KVASIR dataset. This study evaluates the performance of various machine learning approaches, including Random Forest, Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), and K-Nearest Neighbors (KNN), in distinguishing multiple GI disease types. Our findings reveal that the Random Forest model, leveraging selected global features, achieves the highest accuracy and F1 score among the tested methods. However, while CNN-based models initially underperformed without transfer learning, a modified CNN architecture demonstrated superior accuracy in GI disease classification, highlighting the potential of advanced deep-learning models in this domain. The results emphasize the efficiency of specialized CNNs in analyzing GI endoscopic images, offering improved diagnostic precision and significantly reducing manual review time. Future research will focus on further enhancing CNN architectures, incorporating transfer learning techniques, and optimizing feature extraction processes to refine classification accuracy. These advancements provide a robust foundation for developing automated diagnostic systems for GI diseases, promising to improve clinical workflows and patient outcomes.**

Keywords— Convolutional Neural Networks, Gastrointestinal Tract, Image Classification, Deep Learning, Machine Learning, Automated Diagnosis, Endoscopic Imaging.

1. INTRODUCTION

GASTROINTESTINAL diseases refer to a vast spectrum of disorders with variable presentations and manifestations; some of the terminology is, however, a little bit indefinite-or actions for all concerned; anything between the esophagus and intestines, which are chief among the sites of disease and those other organs involved in nutrient absorption. Diseased states comprise infections and inflammations, among others-anomalies of form, like ulcers and polyps; more complex diseases may also be present in the mix such as colorectal cancer, Crohn's disease, colitis, and GERD. The many symptoms of GI disorders range from something that may seem minor on the surface to more serious, even life-threatening, concerns such as bleeding, perforations, or malignancies if left unattended. Actually, of the approximately 10 million deaths, 9.4% are due to colorectum, 7.7% from the stomach, and 5.5% from oesophagal. Other GI tract diseases also cause death. This can be seen in an estimated 9.9 million deaths from cancer in other GI tract diseases out of 19.3 million new cases diagnosed worldwide in either gender, according to Global Cancer Statistics 2020, a study conducted by Sung et al. [1]. According to the report from the Pan American Health Organization, published in 2021 [2], digestive diseases like peptic ulcer, appendicitis, gastritis and duodenitis, and inflammatory bowel disease led to 375,170 deaths in 2019. But there are also clinical studies which further reveal that early diagnosis of diseases in the GI tract helps in preventing mortality [3,4]. Tubular structure is diagnosed with small polyps, adenomas, or lesions that develop, thus establishing an early diagnosis.

Therefore, a quick and correct diagnosis of GI disorders is paramount, as a high standard of management and a means to improve the quality of life of patients depends heavily on early detection. Such medical diagnostic and analysis techniques include computer, artificial intelligence (AI), and machine learning (ML), among many emergent technologies, all of which have a promise to radically transform diagnostic methods and improve their efficacy, precision, and accessibility.

1.1. Motivation

- GI diseases are becoming more critical and thus the early and accurate detection of these diseases is becoming increasingly important. Colorectal cancer accounts for a sizeable number of deaths around the globe.
- Conventional endoscopic methods of examination possess drawbacks, including being slow and dependent on clinicians' experience. The latter is especially true with subtle manifestations of the disease that are mistaken for early stages and with clinicians not thoroughly checking patients.
- Consequently, there is a great need for automated diagnostic tools that assist healthcare providers in better identifying GI anomalies with a higher degree of accuracy, consistency, and efficiency. This work provides a solution to this based on the implementation of machine learning models for multi-class classification of GI diseases using the KVASIR dataset, which will serve as a strong base for future research in automated GI disease detection.

1.2. Contribution

The present work aims to provide a comprehensive analysis of various machine learning models on the KVASIR dataset-a well-curated collection of endoscopic images depicting a range of gastrointestinal disorders and anomalies. Major Contributions Include:

- **Baseline Classification Performance:** This study considers baseline performance through a basic Convolutional Neural Network (CNN) and other classical machine learning classifiers, such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN) and Random Forests. These trials give a baseline that is comparable for the future.
- **Examination of Feature Extraction Methods:** We discuss the effect of extreme dimensionality reduction on computational costs and show that if global features are chosen, the expense is shortened dimensionality.
- **Perspectives for Future Research:** Our research is the first to signal both opportunities and constraints inherent in different machine-learning techniques regarding computer-aided diagnosis of gastrointestinal illnesses.

1.3. Paper Outline

This paper is organized as follows: Section 2 provides a literature review on the use of the approach by other authors. Section 3 describes the methodology, which may include but is not limited to the description of the dataset, its processing, feature extraction, and model architectures for classification in the experiments. Section 4 illustrates the results and performance metrics for each model, followed by a discussion and comparison of the results. Section 5 concludes the paper, summing up the main takeaways. Section 6 discusses further scope for development, providing directions leading to additional research and development concerning automated GI disease detection, including advances in deep learning techniques, a steady merger of real-time video analysis with large-scale advanced datasets, and efforts to modernize model precision.

2. Literature Review

For over three decades, computer-aided diagnosis in gastroenterology has received heavy research interest [Table I]. Early efforts were largely confined to specific regions of the GI tract, especially for polyp detection [1], [2], [3]. The initial approaches relied heavily on manual feature extraction through image processing techniques combined with classification using statistical methods [4]. However, recent development has swung towards deep learning, specifically Convolutional Neural Networks (CNNs), which automatically extract spatial and high-level features. Recent approaches include transfer learning, hybrid models, and fine-tuning architectures from pre-trained CNN models with significant accuracy in the tasks of classifying GI diseases [5].

The work of Pogorelov et al. in 2017, introducing the Kvasir dataset, is considered a foundational milestone in the development of research on the detection of GI diseases. It started as an eight-class dataset with 500 samples, then later became known as KvasirV2, and it evolved into a more useful scope. The first multi-class detection experiment was conducted by Pogorelov et al. using a three-layer CNN, which achieved 95% accuracy [6]. The evolution of the dataset goes further with Borgli et al. [7], who released HyperKvasir, an improvement of KvasirV2, which included better-quality images. Although it has greatly improved, HyperKvasir still has issues because of its imbalanced sample distribution that limits its applicability in some research scenarios.

Transfer learning has been highly used for classification purposes using the KvasirV2 dataset. For example, Dheir and Abu-Naser [8] used pre-trained CNN models including VGG, ResNet, and MobileNet to classify endoscopic images with the accuracy of 98.30% using VGG. Hmoud Al-Adhaileh et al. [14] reported that AlexNet performed best and also obtained a maximum of 97% accuracy. Again, Yogapriya et al. [9] validated that VGG16 performed very well as they were able to obtain a validation accuracy of 96.33%.

New classifiers have been proposed to enhance the performance of transfer learning. Öztürk and Özkaya [10] introduced LSTM-based classifiers integrated with pre-trained AlexNet, GoogleNet, and ResNet models with an accuracy of 97.90%. They improved this to 98.05% by including a residual LSTM layer in the CNN architecture [11]. However, Dutta et al. [12] have approached the problem of model size by the development of the Tiny Darknet model, which achieves an MCC of 75.80% on the HyperKvasir dataset.

Increasingly, multi-feature fusion techniques have also been used in order to enhance classification performance. Khan et al. [13], [14] have thus used Bayesian optimization along with distance-canonical correlation fusion methods, and achieved accuracies of 98.02% and 97.20% on KvasirV2, respectively. In addition, Ramamurthy et al. [15] integrated features extracted by EfficientNetB0 using a custom CNN architecture called Effimix, attaining an accuracy of 97.99%. The capsule networks have emerged as another promising alternative too, wherein Afriyie et al. [16] applied the denoising capsule networks (Dn-CapsNets) on the dataset of KvasirV2 and attained an accuracy of 94.16%. Wang et al. [17] further combined CNNs with capsule networks in a two-stage classification framework, achieving strong performance on KvasirV2 (94.83%) and HyperKvasir (85.99%).

The literature puts forth strong emphasis on an association between polyps and lesions and GI diseases in diagnosis. However, existing studies often lack generalized computer-aided diagnosis systems that can concurrently detect and classify various findings within the GI tract. This work addresses this gap by applying a multi-class image dataset in developing an advanced machine learning pipeline with clinical utility, offering timely and accurate diagnoses.

There is a significant limitation in the existing state of the literature, which relies on relatively diverse and limited datasets, usually to a single dataset with imbalanced class distribution. This study tries to alleviate these limitations by applying both KvasirV2 and HyperKvasir datasets to enhance robustness and generalizability. Many classical methods separate the feature extraction process from the classification process, which heavily limits their ability to classify together

pathological and non-lesional tissues in the same run. To this end, a hybrid deep-learning-based stacking ensemble approach is proposed here, one that integrates predictions from multiple CNN architectures with secondary classifiers in pursuit of superior performance and better diagnostic solutions in gastroenterology.This approach utilizes the advantages of many CNN models and integrates them through meta-learning approaches, guaranteeing reliable and precise categorization. This method has the potential to greatly improve automated diagnostic processes in gastroenterology.

3. METHODOLOGY

To design and implement a comprehensive machine-learning pipeline that can be utilized for differentiating among various gastrointestinal diseases through the use of a multi-class image dataset . With this regard, this paper is aimed to build robust and scalable models that can effectively distinguish a range of disease classes while important diagnoses occur in very early time with an accurate manner, which is the most important factor for medical interventions within a needed time. This research paper follows four phases: dataset preparation, data preprocessing, architecture design of each model, and classification. Each stage was clearly delineated and executed to optimize performance, reliability, and clinical validity of the final model, hence ensuring its suitability for real-world diagnostic contexts. This thorough design emphasizes its potential for easy incorporation into clinical workflows.

3.1. Dataset

Vestre Viken Health Trust (VV) in Norway used endoscopic machines to gather the data provided in this dataset. VV is made up of 4 hospitals and provides health care services to 470.000 people. This study uses the **Kvasir dataset**, including an overall of **8000 images across 8 classes** , where each class represents a different type of gastrointestinal disease or condition, which has 1000 images per class. The stabilised distribution ensures every class is represented well-the critical factors for achieving accurate multi-class classification. The dataset contains images with various resolutions from 720x576 to 1920x1072 pixels and is structured so that they are divided into different subfolders categorized by subject. It is divided into three major anatomical landmarks and three clinically significant findings. Besides, there are two types of pictures related to endoscopic polyp removal. In sorting and annotating the dataset, medical doctors, being experienced endoscopists, are used.

The collection of images is classified into three anatomical landmarks and three clinically significant findings. Additionally, it contains two categories of images related to endoscopic polyp removal. Sorting and annotation of the dataset are performed by medical doctors (experienced endoscopists).

1. Anatomical Landmarks

An anatomical landmark is an easily visualized feature in the GI tract by the endoscope. They are necessities for navigation and as a reference point for describing the location of any specific finding. These landmarks include:

Z-line: It is a border site between oesophagus and stomach. In Endoscopy, it presents as an optical border at the site where white color of oesophageal mucosa meets the red gastric mucosa. [Fig. 1] **Pylorus**: The pylorus is the area surrounding the opening from the stomach into the first part of the small bowel (duodenum). The opening contains circumferential muscles controlling the movement of food from the stomach. The recognition of pylorus is necessary in instrumentation to the duodenum within endoscopic procedures one of the most difficult manoeuvres of gastroscopy. [Fig. 2] **Cecum**: The cecum is the most proximal part of the large bowel. Reaching cecum is the proof for a complete colonoscopy and completion rate has shown to be a valid quality indicator for colonoscopy. Therefore, recognition and documentation of the cecum is important. [Fig. 3]

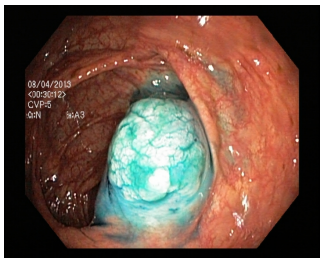

Fig. 1: Class 0: Dyed and Lifted Polyps

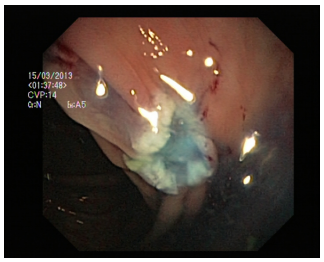

Fig. 2: Class 1: Dyed Resection Margins

Fig. 3: Class 2: Esophagitis

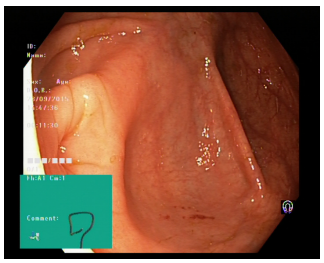

Fig. 4: Class 3: Cecum

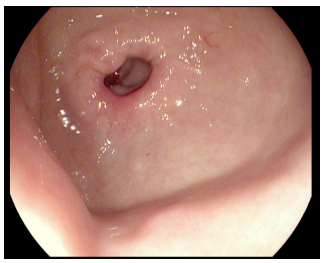

Fig. 5: Class 4: Pylorus

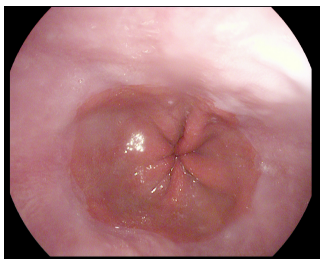

Fig. 6: Class 5: Z-line

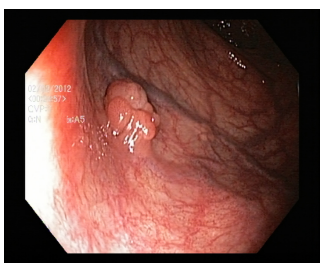

Fig. 7: Class 6: Polyps

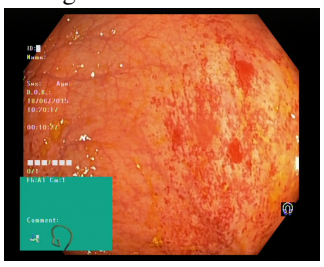

Fig. 8: Class 7: Ulcerative Colitis

TABLE I: Summary of Past Research on GI Disease Classification

Author(s)	Dataset	Methodology	Results	Remarks
Pogorelov et al. [15]	Kvasir	Three-layer CNN	95.00%	Introduced the Kvasir dataset, later updated to KvasirV2 with double samples.
Borgli et al. [16]	HyperKvasir	Pre-trained CNNs	90.20%	HyperKvasir is under-researched due to imbalanced sample distribution.
Dheir & Abu-Naser [17]	KvasirV2	Transfer learning (VGG, ResNet, MobileNet)	98.30%	Highlights VGG's success in GI tract classification.
Hmoud Al-Adhaileh et al. [18]	KvasirV2	AlexNet with frozen convolution bases	97.00%	AlexNet performed best among tested models.
Yogapriya et al. [19]	KvasirV2	VGG16, ResNet-18, GoogLeNet	VGG16: 96.33%	Further supports VGG16's performance.
Öztürk & Özkaya [20]	KvasirV2	LSTM-based classifier on AlexNet, GoogleNet, ResNet	97.90%	Proposes innovative LSTM-based classifiers.
Öztürk & Özkaya [21]	KvasirV2	Residual LSTM layer CNN	98.05%	Improved performance using residual LSTM.
Dutta et al. [22]	HyperKvasir	Tiny Darknet model	75.80%	High speed with a lightweight model.
Ramamurthy et al. [23]	HyperKvasir	Multi-feature fusion with Efimix	97.99%	Innovative fusion approach for feature extraction.
Khan et al. [24]	KvasirV2	Bayesian optimization	98.02%	Optimized hyperparameters for high accuracy.
Afriyie et al. [25]	KvasirV2	Dn-CapsNets	94.16%	Capsule networks applied to GI disease classification.

2. Pathological Findings In this picture, a pathological finding means an abnormal feature observed in the gastrointestinal tract. It can be observed through endoscopy as a damage or alteration of the normal mucosa . These include: **Esophagitis**: This is an inflammation of the oesophagus that is observed as a break in the oesophagal mucosa approximately on the Z-line. The grade of inflammation is identified through the extent of the mucosal breaks and the fraction of the circumference involved.[Fig. 4] **Polyps**:- Polyps are lesions within the bowel which can be seen as mucosal outgrowths. The polyps are either flat, elevated, or pedunculated, and may be differentiated from normal mucosa on account of colour and surface patterns. Most bowel polyps are harmless although some of them have a potential to grow into a cancer. [Fig. 5] **Ulcerative Colitis -** Ulcerative colitis is an inflammatory chronic disease involving the large bowel. The disease could significantly impact quality of life, and diagnosis is primarily colonoscopic findings. [Fig. 6]

3. Polyp Removal Polyps in the large bowel are precursors to cancer, and therefore they should be removed during endoscopy if possible. **Dyed and Lifted Polyps**: The light blue polyp margins are visible against the darker normal mucosa. Usefulness of adjunctive information for reporting, in an automatized way, may include success of lift and eventual presence of non-lifted areas, which could be suggestive of malignancy. [Fig. 7]

Dyed Resection Margins: The Resection margins are important in helping to decide whether the polyp has indeed fully been removed or not. The residual tissue of a polyp can continue to grow and even, in worst cases development into malignancy. [Fig. 8]

3.1.1. Training and Validation Dataset: The dataset which contains 8,000 high-quality images of endoscopy across eight classes with 1,000 images in each class has been split into two subsets: training set and the validation set. [Table II] This has been divided in an 80:20 ratio to allocate 6,400 images towards training set and 1,600 images towards the validation set. Each class in the validation set consists of 200 images so that under the disease categories, every class remains fairly represented.

TABLE II: Number of images in each training and validation dataset.

Category	Training Dataset	Validation Dataset
Z-line	6400	1600
Pylorus	6400	1600
Cecum	6400	1600
Esophagitis	6400	1600
Polyps	6400	1600
Ulcerative Colitis	6400	1600
Dyed and Lifted Polyps	6400	1600
Dyed Resection Margins	6400	1600

This split had the rationale of having enough size in the training set to enable the model to learn hard and complex patterns and features efficiently. Simultaneously, a proper validation set was needed to judge the performance of the model reliably so that it doesn't overfit. With such a distribution, the model is trained on diverse examples while tested on unseen data assessing its generalization capability. This well-thought split, besides allowing the model to learn at an accuracy level at training also turned out to be a strong construction when put to test its generalization capability on unseen data. The balanced and systemic way ensured that the validation metrics reflected well the model's potential for real application in practice use cases.

3.2. Preprocessing

Several preprocessing techniques were applied to the images to improve the robustness and generalization of the model. These techniques ensured that the model effectively learned the data while being robust to common variations occurring in medical imaging. The preprocessing process entailed two components: data augmentation and feature extraction.

3.2.1. Data Augmentation: **Rescaling:** All pixel intensity values were normalized to the range [0, 1] by dividing each pixel value by 255. This kept the convergence time shorter for the model during training due to uniformity in the input data and also reduced computation.

Random Rotation: Images were rotated randomly within a range of 30 degrees in all directions to simulate different viewing angles as encountered during endoscopic tests.

Width and Height Shifts: Random shifts along both width and height, up to 20%, were applied to introduce positional variations in the captured images.

Shear Transformation: A shearing range of 0.2 was used to stretch or tilt the images slightly, introducing common geometric distortions encountered in real-world imaging applications.

Zooming: Random zooms of up to 20% were applied to focus on more localized regions of the images.

Horizontal Flipping: Mirrored versions of the images were added to the training set by flipping the images horizontally.

Fill Mode: When transformations extended regions outside the original image boundaries, the ”nearest” fill mode was applied. This method used the nearest pixel values to fill missing areas, ensuring no structural discontinuity within the images.

These combinations collectively improved the diversity and variability within the training set, making the model more robust against variations in angles, positions, and focus, as manifested in real clinical applications.

3.2.2. Feature Extraction: Feature extraction was applied to convert the preprocessed images into structured formats suitable for input into traditional algorithms like SVM, Random Forest, KNN, and Logistic Model Tree.

Preprocessing: Each image in a batch was converted into an array using `image.img_to_array()` and normalized to prepare for feature extraction.

Storage: Extracted features were appended to a list of features, and corresponding labels were added to a labels list.

Batch: Images were processed in batches to ensure that all samples in the generator were utilized.

Labeling: Labels were encoded into a categorical format using `np.argmax()`.

Flattening: Extracted features were reshaped into a 2D array with appropriate dimensions, compatible with algorithms like SVM, KNN, Random Forest, and Logistic Model Tree. The flattened features represented the most important characteristics of the images in a compact, structured form, enabling effective classification.

For deep learning models, raw extracted features were directly passed as input because these models exploit the hierarchical spatial relationships within the data to learn.

This preprocessing and feature extraction stage facilitated the training and testing of different classification models, ensuring their robustness and accuracy.

3.3. Network Architecture

The network architectures proposed in this work are adapted to bring out the best of multiple learning and deep learning models to be used in multi-class classification of gastrointestinal images. Each of the models will be elaborated on for their architectures, training approaches, salient features, and how they contribute to robust classification performance.

3.3.1. Modified Convolutional Neural Network (CNN): The CNN was designed solely with an aim to detect the automatic feature extraction and classification of the images into the eight classes of gastrointestinal diseases. This CNN integrates the convolutional, pooling, and fully connected layers[Fig. 9], whose main purpose is to extract meaningful patterns and reduce computational complexity while still achieving a high accuracy of 0.75

Structure of the CNN : Input Layer : The CNN inputs the images resized to 224×224×3 dimensions, with 3 representing the RGB channels. **Convolutional Layers : First Convolutional Block:** 32 filters of 3×33×3 with ReLU activation will extract low-level features like edges and textures. Max pooling (2×2) reduces the spatial dimensions; this ensures that they will not overfit. **The second Convolutional Block:** This will take 64 filters of size 3×3 to fine-tune the features to be oriented at the pattern-making site. It is then followed by a max-pooling layer. **Third Convolutional Block:** This has 128 filters of size 3×3 which extract high-level features. It captures complex patterns representing the fine grain differences between the eight classes. **Flatten Layer:** The output from the convolutional layers is flattened to a one dimensional vector, to bridge the convolutional blocks with the fully connected layers. **Fully Connected Layers:** A dense layer with 128 units and ReLU activation will enable the network to learn a complex, non-linear pattern from the extracted features. The output layer will be chosen to have 8 units using softmax activation, which gives probability distributions over the classes. **Compilation and Optimization:** The model uses the **Adam optimizer** to balance convergence speed and accuracy.**Categorical Cross-Entropy** is employed as the loss function for multi-class classification.Performance metrics include accuracy, tracked on both training and validation datasets.Early stopping (patience of 10 epochs) prevents overfitting, while learning rate reduction (by a factor of 0.1 after 5 epochs without improvement) accelerates convergence. This CNN architecture yielded a maximum classification accuracy of 0.75 and happened to be the most optimal model for the research at hand. Therefore, better performance was accomplished with any other model for achieving appropriate classification of diseases within the gastrointestinal system.

3.3.2. Support Vector Machine (SVM): The SVM was trained by a linear kernel over the flattened image features.

SVMs are good for high-dimensional data, and a linear kernel is especially suitable for structured output given by the feature extraction procedure. Training: The flattened feature vector went into the model, and it was trained to differentiate between classes with hyperplanes. Output: The SVM was only moderately accurate but was always limited to its handcrafted feature extraction. Furthermore, SVM is computationally inexpensive for smaller datasets, and therefore can be used for preliminary experiments. SVM is well-suited for binary classification tasks and can be easily generalized to multi-class problems using the one-vs-one or one-vs-rest strategy.

3.3.3. Random Forest (RF): A Random Forest classifier was implemented using only two global features extracted from the dataset. RF is an ensemble learning approach in which several decision trees are combined to enhance classification quality. Hyperparameters: There are 100 trees, with a "max features" value set as 2. Feature Handling: As the RF comes shown with minimal input features of two mere high-quality features, it showed high competitive accuracy and therefore implies its robustness, by virtue of which it would operate on minimal input dimensionality. Its ability to manage non-linear relationships amid features has added flexibility toward making predictions by the model. Random Forests have relatively fewer chances of overfitting compared to individual decision trees, with high generalization capability.

3.3.4. K-Nearest Neighbors (KNN): This exercise employed KNN as a baseline model because it possesses good interpretability and is simple. The algorithm classified flattened features by computing the majority class among the 5 nearest neighbors (k=5). Advantages: KNN is non-parametric and does not rely on training weights, making it efficient for small-scale datasets. Weaknesses: The performance of the model was limited by the curse of dimensionality inherent in high-dimensional feature spaces. The simplicity of KNN also made it suitable as a reference for benchmarking other, more complex models. However, KNN's performance can degrade as the dataset size grows, particularly when dealing with noisy data or irrelevant features.

3.3.5. Logistic Model Tree (LMT): The Logistic Model Tree integrates a decision tree structure with logistic regression models fitted at the leaf nodes. This hybrid model used decision trees for interpretability and logistic regression for the classification. Construction of Tree: A maximum of 6 global features was used for the decision tree. Leaf Nodes: Logistic Regression on the data subsets corresponding to the leaf nodes as classification became granular. Performance: LMT achieved balanced accuracy and interpretability and turned out to be suitable for low-to-medium complexity datasets.

The partnership of traditional machine learning models with a deep learning-based Modified CNN provided a rich validation of the dataset. Here, the CNN established itself as the state-of-the-art model, with its ability to automatically extract hierarchical features causing it to have better classification accuracy.

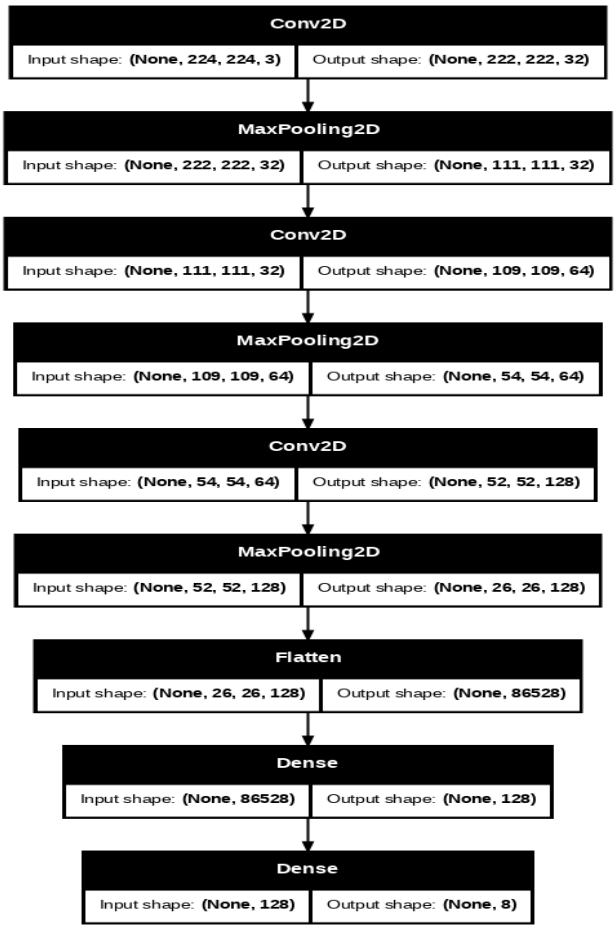

Fig. 9: Modified CNN Architecture

3.4. Classification

After rigorous training and testing, the **updated CNN** was put to performance comparison with other classification models, namely **SVM, Random Forest, KNN, and Logistic Model Tree**. Each of these models attempted to classify input images of the gastrointestinal tract into one of eight exclusive classes[Fig.1] corresponding to different ailments or conditions, which were: **Z-Line, Pylorus, Cecum, Esophagitis, Polyps, Ulcerative Colitis, Dyed and Lifted Polyps, Dyed Resection Margin**. Each of these classes portrayed a landmark anatomical/pathological condition of the gastrointestinal tract.

Performance Metrics: The classification performance of each model is given below: Accuracy: The ratio of correctly classified instances to total instances. F1 Score: This provides the harmonic mean of precision and recall, balancing performance in imbalanced datasets. For an imbalanced dataset Recall: It measures the sensitivity of the model for each class that identifies true positive instances. **Performance Highlights:** At all evaluation metrics, CNN outperformed the other models with highest accuracy at 0.750 [Fig. 14]. This is owing to the fact that it can extract, in an automatic way, hierarchical and high-level features, designed for complex multi-class-classification tasks. SVM [III] achieved a reasonable performance, but its flattening features prevented it from taking the spatial and contextual information behind the images fully.

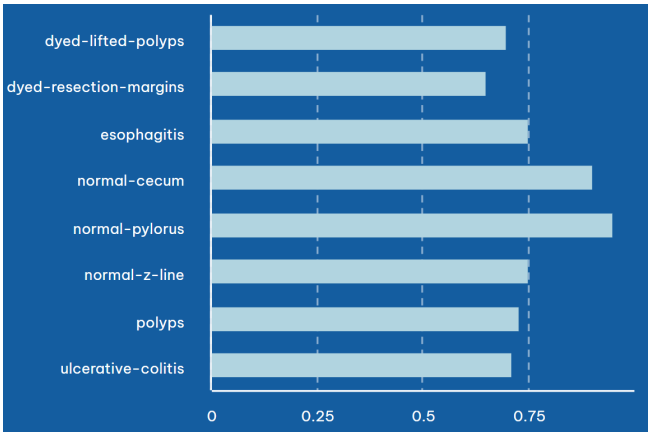

Fig. 10: Class-vise Accuracy of CNN model on validation dataset

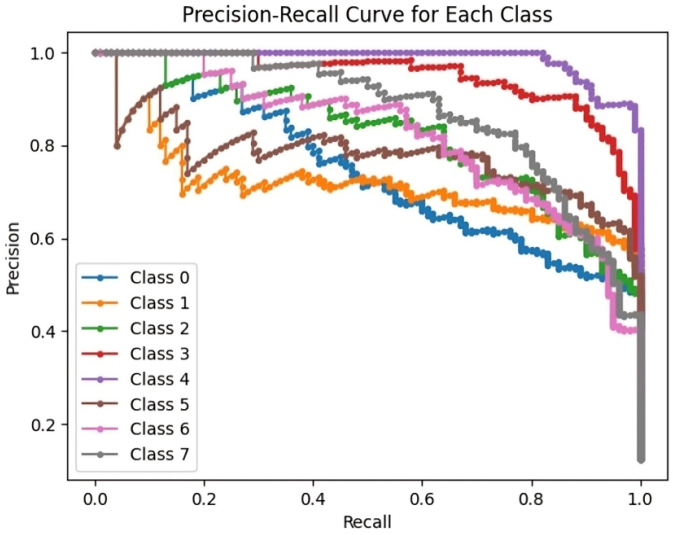

Fig. 11: Class-vise Precision-Recall curve of CNN model on validation dataset

Random Forest [IV] only used two global features, surprisingly being highly efficient but lacking depth that dealing with complex image-based data needs. KNN [VI] succeeded regarding the measure of recall but failed when it came to good accuracy levels due to the high dimension of the feature space. Logistic Model [V] Tree caught the balance between being interpretative and performing well. Thus, scenarios where model explainability should be crucial will find this one very useful.

The above comparative analysis shows that the Modified CNNs are significantly effective in solving the multi-class image classification problem. Having a strong performance for discrimination between different classes, the scalability of CNNs is found to be suitable to solve the complex tasks related to medical imaging. Their ability to learn hierarchical features makes them exceptionally well-suited for real-world applications, especially in the area of computer-aided diagnosis where accuracy and efficiency are very important, particularly for detecting gastrointestinal disease.[Fig. 17]

4. Results

The results showed that the revised CNN achieved the highest accuracy of 0.75, outperforming the other models, as shown in VII. The high-ranking performance of CNN suggests its ability to discover hierarchical, complex features is an important factor in facilitating the learning and distinction of the different disease classes.

TABLE III: SVM Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.50	0.51	0.50	200
1	0.58	0.56	0.56	200
2	0.54	0.56	0.55	200
3	0.56	0.68	0.62	200
4	0.72	0.78	0.75	200
5	0.56	0.53	0.55	200
6	0.41	0.39	0.40	200
7	0.57	0.47	0.52	200
Accuracy	-	-	0.56	1600
Macro Avg	0.56	0.56	0.56	1600
Weighted Avg	0.56	0.56	0.56	1600

TABLE IV: RF Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.43	0.33	0.37	200
1	0.52	0.60	0.56	200
2	0.66	0.59	0.63	200
3	0.60	0.79	0.68	200
4	0.70	0.78	0.73	200
5	0.59	0.60	0.59	200
6	0.50	0.35	0.42	200
7	0.62	0.65	0.64	200
Accuracy	-	-	0.59	1600
Macro Avg	0.58	0.59	0.58	1600
Weighted Avg	0.58	0.59	0.58	1600

TABLE V: LMT Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.29	0.28	0.29	200
1	0.42	0.38	0.40	200
2	0.43	0.36	0.39	200
3	0.45	0.43	0.44	200
4	0.44	0.49	0.47	200
5	0.42	0.45	0.43	200
6	0.23	0.26	0.24	200
7	0.41	0.41	0.41	200
Accuracy	-	-	0.38	1600
Macro Avg	0.39	0.38	0.38	1600
Weighted Avg	0.39	0.38	0.38	1600

TABLE VI: KNN Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.50	0.36	0.42	200
1	0.58	0.51	0.54	200
2	0.58	0.46	0.51	200
3	0.48	0.81	0.60	200
4	0.38	0.80	0.52	200
5	0.39	0.27	0.32	200
6	0.36	0.26	0.30	200
7	0.57	0.26	0.36	200
Accuracy	-	-	0.47	1600
Macro Avg	0.48	0.46	0.45	1600
Weighted Avg	0.48	0.47	0.45	1600

Making, CNN the final model for this classification task.CNN is the recommended model for use in clinical settings since this examination validates its usefulness for gastrointestinal illness detection.

TABLE VII: **Model Performance Comparison**

Model	Accuracy	F1-Score	Recall
Modified CNN	0.750	-	-
SVM	0.558	0.550	0.558
Random Forest (2 GF)	0.586	0.577	0.586
KNN	0.465	0.4464	0.465
Logistic Model Tree (6 GF)	0.384	0.383	0.383

4.1. Comparative Analysis

The performances of the five tested models—Modified CNN, SVM, Random Forest, KNN, and Logistic Model Tree using six global features—are compared intensively with three important performance metrics: accuracy, F1-score, and recall. These metrics provide a panoramic view of the capacity of the models to differentiate between the eight different classes of gastrointestinal (GI) diseases. The comparison results, as shown in Table VII, further elaborate on the relative merits and demerits of each of these models while executing the complex multi-class classification task for GI disease detection.

4.1.1. Modified Convolutional Neural Network (CNN): The CNN model seemed to be the strongest and most accurate classifier, with an accuracy of 0.750 [Fig. 10], thus significantly outperforming all other models. This could be because of the hierarchical feature extraction ability of the CNN that is able to learn the complex spatial patterns and relationships that exist in an endoscopic image [Fig. 11]. Leverage the ability of convolutional, pooling, and dense layers to differentiate between the eight classes of GI diseases based on a CNN [Fig. 12]. Hence, this result presents the efficacy of CNN in handling those complex datasets for which traditional machine learning models falter.[Fig. 13]

4.1.2. Support Vector Machine (SVM): The SVM model achieved an accuracy of 0.558 [Fig. 14], an F1-score of 0.550 [Fig. 15], and a recall of 0.558 [Fig. 16], making it the second-best performer. While the SVM utilized a linear kernel to manage high-dimensional feature spaces, its reliance on manually extracted features limited its ability to handle the complexities of the dataset. Despite its reasonable performance, the SVM struggled to differentiate subtle variations between the disease classes.

4.1.3. Random Forest (Two Global Features): Random Forest achieved an accuracy of 0.586 [Fig. 14] and an F1-score of 0.577 [Fig. 15] on recall 0.586 [Fig. 16], considering only two global features, whereas it outperforms the SVM by great margins in both accuracy and recall. However, its performance is rather much lower than that of the CNN, showing that a richer feature representation is needed for the higher accuracy in this multi-class problem.

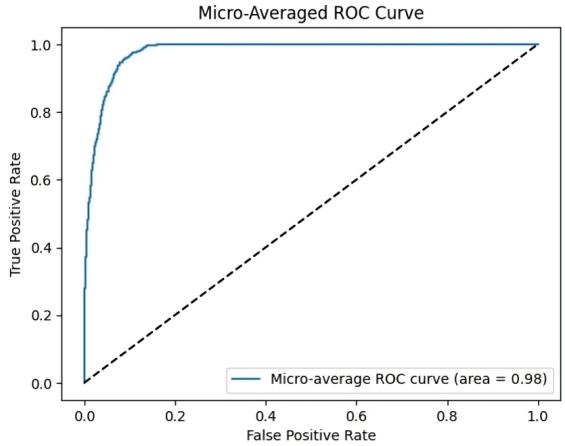

Fig. 12: Micro-Averaged ROC Curve of Modified CNN

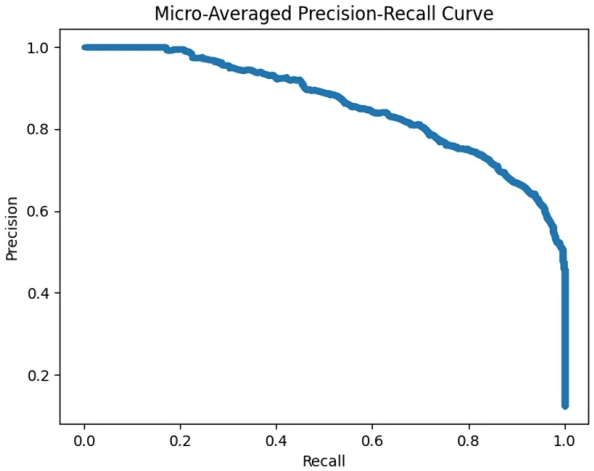

Fig. 13: Micro-Averaged Precision-Recall Curve of Modified CNN

4.1.4. K-Nearest Neighbor (KNN): The KNN classifier performed poorly, with an accuracy of 0.465 [Fig. 14], an F1-score of 0.446 [Fig. 15], and a recall of 0.465 [Fig. 16].

Its reliance on feature similarity without the ability to learn hierarchical patterns limited its generalization capability for the complex dataset. The results demonstrated the inefficiency of KNN for high-dimensional medical imaging tasks.

4.1.5. Logistic Model Tree (Six Global Features): The worst outcome was by the Logistic Model Tree model; it had the lowest performance result out of the evaluated models with an accuracy of 0.384 [Fig. 14], F1-score of 0.383 [Fig. 15], and recall of 0.383 [Fig. 16]. This model was a decision tree combined with logistic regression. Still, the model was incapable of capturing high-level patterns and also the critical interactions within the dataset, thus degrading classification performance. This clearly shows how traditional hybrid approaches fail with regard to complex data, such as in medical image problems. In conclusion, the very good performance of the CNN model poses an argument in support of employing this model in clinical-based real-world applications. The precision and robustness of diagnostic accuracy of the models are highly significant. Future work will be on trying to integrate CNN with ensemble techniques or transfer learning to improve diagnostic accuracy and reliability.[Fig. 17]

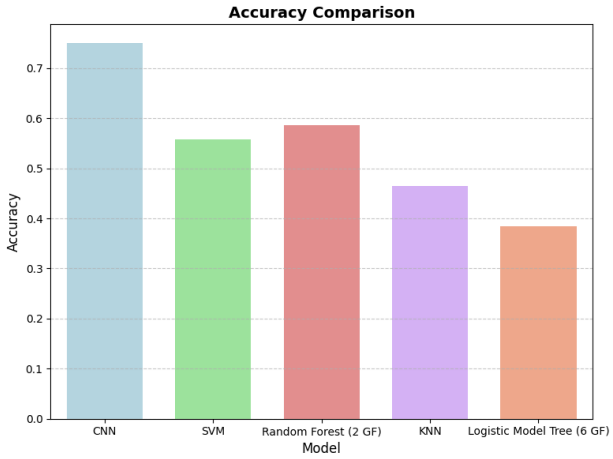

Fig. 14: Accuracy Comparion of CNN, SVM, Random Forest, KNN, Logistic Model Tree

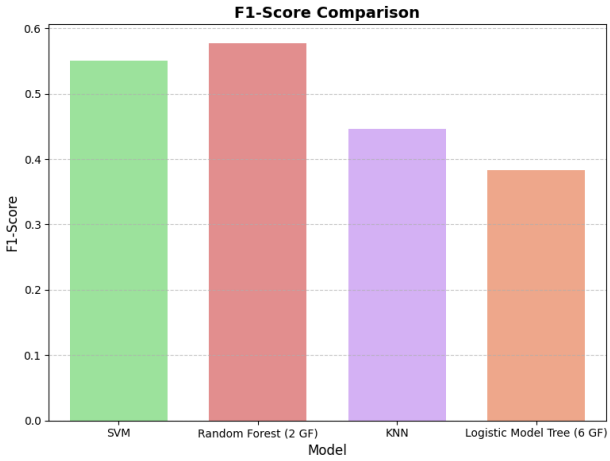

Fig. 15: F1-Score Comparion of SVM, Random Forest, KNN, Logistic Model Tree

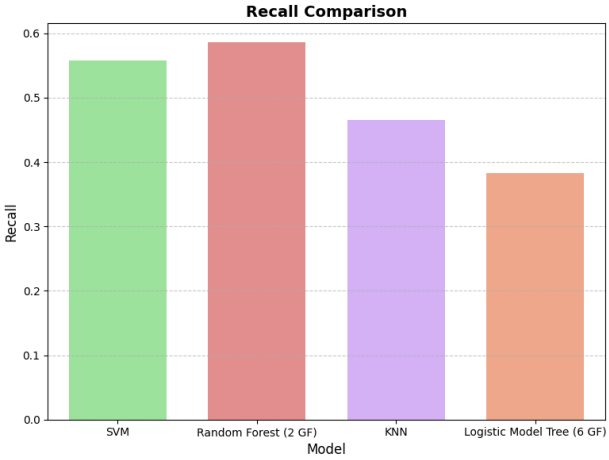

Fig. 16: Recall Comparion of SVM, Random Forest, KNN, Logistic Model Tree

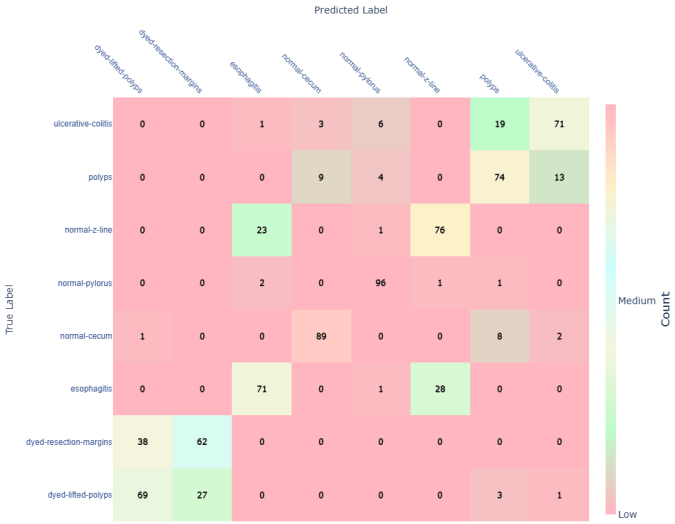

Fig. 17: Confusion matrix of testing dataset (100 images of each class) of CNN model

5. Discussion and Conclusion

5.1. Conclusion

This work was entirely successful in building a machine-learning pipeline that focuses on a variant of the CNN model to distinguish gastrointestinal diseases from the Kvasir dataset. Splendid differentiating capabilities between eight classes of GI conditions were achieved, with deep learning techniques significantly enhancing diagnostic accuracy within this domain. The findings underline the utility of the CNN model in enriching clinical workflows by reducing reliance on manual assessment and demonstrating the growing relevance of AI-driven approaches in medical diagnostics.[Fig. 18] The promising results obtained in this study highlight the potential for innovative breakthroughs in such technologies, enabling

early and accurate detection of GI diseases, ultimately translating into improved patient outcomes and reduced delays in diagnosis. Future work involves exploring advanced deep learning techniques, transfer learning, and ensemble modeling to further improve model performance. Additionally, incorporating real-time video analysis features and enriching the dataset with diverse and larger samples may enhance the adaptability and robustness of such systems in varied clinical settings.

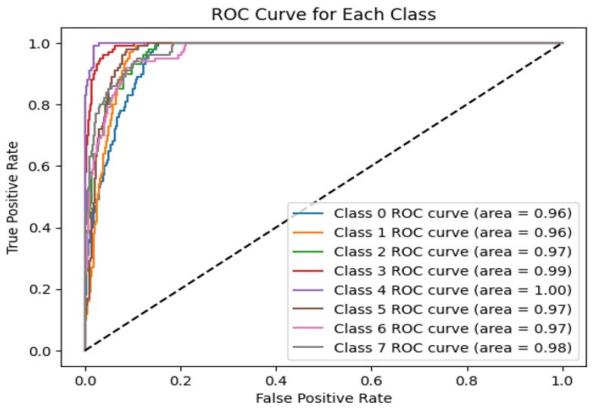

Fig. 18: ROC curve for each class of CNN model

5.2. Discussion

This work demonstrates the feasibility of a modified CNN architecture for the classification of GI diseases from endoscopic images, yielding valuable findings. The CNN model significantly outperformed other classifiers, including Support Vector Machine (SVM), Random Forest, K-Nearest Neighbor (KNN), and Logistic Model Tree. This highlights CNN's superiority in multi-class classification tasks due to its inherent feature extraction capability, which captures hierarchical patterns effectively. Such capabilities make CNN well-suited for complex tasks like GI disease detection.

Extensive preprocessing techniques, including normalization, rotation, and augmentation, played a crucial role in improving the model's generalizability and robustness. These techniques helped the model perform consistently despite the variability and noise inherent in real medical images, a compelling feature for clinical applications. The traditional machine learning models struggled to achieve similar performance due to their reliance on manual feature extraction, which could not adequately represent the high-dimensional and complex features of GI images.

Even though traditional models base their predictions on manual feature extraction, the CNN's automatic feature learning, combined with careful architectural tuning, proved to be a decisive advantage. This work validates CNNs as efficient and reliable tools for the early and accurate detection of GI diseases, underlining their potential to enhance diagnostic performance in clinical workflows.

In conclusion, the limitations of traditional machine learning models stem from their inability to capture the inherent complexity and hierarchical patterns of GI tract features. CNN models, with their automatic feature learning and robust architecture, demonstrated superior performance and the potential to transform GI disease detection. This study confirms CNNs as a promising tool for improving healthcare diagnostic efficiency and sets the stage for future advancements in AI-driven medical diagnostics.

6. Future Scope

Future directions toward automation of the detection of GI diseases in their development may be very critical in many ways toward further improvement of diagnostic accuracy, robustness, and practicality. Below are some key areas for future research:

1) **Latest Techniques in Deep Learning:** Transfer learning and ensemble models can lead to more advanced classification performances by leveraging pre-trained networks and combining the strengths of various architectures. Proper training techniques could enable accurate predictions even with minimal data, significantly enhancing model adaptability and performance.
2) **Real-Time Video Analysis:** Real-time classification of abnormalities through video-supported endoscopy could greatly enhance clinical workflows. Providing on-the-fly feedback to physicians during procedures could significantly improve the practical usefulness of automated detection systems.
3) **Augmented Data:** Expanding datasets to include a greater number and diversity of GI images—spanning various demographics and disease presentations—could improve model generalization. This would reduce biases and ensure applicability across different patient populations, making models more robust and universally applicable.
4) **Patient-Level Data Integration:** Incorporating patient-level data, such as demographics, clinical history, and lifestyle factors, into diagnostic systems could lead to more specific and personalized predictions. Models adapted to individual patient requirements would enhance both diagnostic accuracy and the relevance of predictions.
5) **Engagement of Practicing Clinicians:** Actively involving clinicians in the design and testing of these systems would ensure that AI-centered solutions are tailored to the tools and workflows used in real-world clinical settings. This collaboration could bridge the gap between technological innovation and practical application, making these systems more user-friendly and effective.

From these explorations, future research can push the frontiers of GI disease detection systems toward becoming highly accurate, adaptive, and personalized diagnostic tools. Such advancements will not only improve patient outcomes but also set new standards for medical diagnostics in gastroenterology, ultimately bridging the gap between AI innovation and clinical application.

REFERENCES

[1] Monika Thapliyal, Neelu Jyothi Ahuja, Achyut Shankar, Xiaochun Cheng, and Manoj Kumar. A differentiated learning environment in domain model for learning disabled learners. *Journal of Computing in Higher Education*, 34(1):60–82, 2022.

[2] Supriya Raheja, Shreya Kasturia, Xiaochun Cheng, and Manoj Kumar. Machine learning-based diffusion model for prediction of coronavirus-19 outbreak. *Neural Computing and Applications*, 35(19):13755–13774, 2023.

[3] Supriya Raheja, Shreya Kasturia, Xiaochun Cheng, and Manoj Kumar. Machine learning-based diffusion model for prediction of coronavirus-19 outbreak. *Neural Computing and Applications*, 35(19):13755–13774, 2023.

[4] Mehul Patel, Shraddha Gulati, Sarah O'Neil, Nicholas Wilson, Sophie Williams, Shonette Charles-Nurse, Guy Chung-Faye, Rajaventhan Srirajaskanthan, Amyn Haji, and Bu'Hussain Hayee. P59 artificial intelligence increases adenoma detection even in 'high-detector'colonoscopy: early evidence for human: machine interaction, 2021.

[5] Mehul Patel, Shraddha Gulati, Sarah O'Neil, Nicholas Wilson, Sophie Williams, Shonette Charles-Nurse, Guy Chung-Faye, Rajaventhan Srirajaskanthan, Amyn Haji, and Bu'Hussain Hayee. P59 artificial intelligence increases adenoma detection even in 'high-detector'colonoscopy: early evidence for human: machine interaction, 2021.

[6] Pu Wang, Tyler M Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*, 68(10):1813–1819, 2019.

[7] Mustain Billah and Sajjad Waheed. Gastrointestinal polyp detection in endoscopic images using an improved feature extraction method. *Biomedical engineering letters*, 8:69–75, 2018.

[8] Wei-Lun Chao, Hanisha Manickavasagan, and Somashekar G Krishna. Application of artificial intelligence in the detection and differentiation of colon polyps: a technical review for physicians. *Diagnostics*, 9(3):99, 2019.

[9] Linjie Guo, Hui Gong, Qiushi Wang, Qiongying Zhang, Huan Tong, Jing Li, Xiang Lei, Xue Xiao, Chuanhui Li, Jinsun Jiang, et al. Detection of multiple lesions of gastrointestinal tract for endoscopy using artificial intelligence model: a pilot study. *Surgical Endoscopy*, 35:6532–6538, 2021.

[10] Said Charfi, Mohamed El Ansari, and Ilangko Balasingham. Computer-aided diagnosis system for ulcer detection in wireless capsule endoscopy images. *IET Image Processing*, 13(6):1023–1030, 2019.

[11] Xinle Wang, Haiyang Qian, Edward J Ciaccio, Suzanne K Lewis, Govind Bhagat, Peter H Green, Shenghao Xu, Liang Huang, Rongke Gao, and Yu Liu. Celiac disease diagnosis from videocapsule endoscopy images with residual learning and deep feature extraction. *Computer Methods and Programs in Biomedicine*, 187:105236, 2020.

[12] Francesco Renna, Miguel Martins, Alexandre Neto, António Cunha, Diogo Libânio, Mário Dinis-Ribeiro, and Miguel Coimbra. Artificial intelligence for upper gastrointestinal endoscopy: a roadmap from technology development to clinical practice. *Diagnostics*, 12(5):1278, 2022.

[13] Michael Liedlgruber and Andreas Uhl. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: a review. *IEEE reviews in biomedical engineering*, 4:73–88, 2011.

[14] Javeria Naz, Muhammad Sharif, Mussarat Yasmin, Mudassar Raza, and Muhammad A Khan. Detection and classification of gastrointestinal diseases using machine learning. *Current Medical Imaging*, 17(4):479–490, 2021.

[15] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017.

[16] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020.

[17] Ibtesam M Dheir and Samy S Abu-Naser. Classification of anomalies in gastrointestinal tract using deep learning. 2022.

[18] Mosleh Hmoud Al-Adhaileh, Ebrahim Mohammed Senan, Fawaz Waselallah Alsaade, Theyazn H H Aldhyani, Nizar Alsharif, Ahmed Abdullah Alqarni, M Irfan Uddin, Mohammed Y Alzahrani, Elham D Alzain, and Mukti E Jadhav. Deep learning algorithms for detection and classification of gastrointestinal diseases. *Complexity*, 2021(1):6170416, 2021.

[19] J Yogapriya, Venkatesan Chandran, MG Sumithra, P Anitha, P Jenopaul, and C Suresh Gnana Dhas. Gastrointestinal tract disease classification from wireless endoscopy images using pretrained deep learning model. *Computational and mathematical methods in medicine*, 2021(1):5940433, 2021.

[20] Şaban Öztürk and Umut Özkaya. Gastrointestinal tract classification using improved lstm based cnn. *Multimedia Tools and Applications*, 79(39):28825–28840, 2020.

[21] Şaban Öztürk and Umut Özkaya. Residual lstm layered cnn for classification of gastrointestinal tract diseases. *Journal of Biomedical Informatics*, 113:103638, 2021.

[22] Amartya Dutta, Rajat Kanti Bhattacharjee, and Ferdous Ahmed Barbhuiya. Efficient detection of lesions during endoscopy. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, pages 315–322. Springer, 2021.

[23] Karthik Ramamurthy, Timothy Thomas George, Yash Shah, and Parasa Sasidhar. A novel multi-feature fusion method for classification of gastrointestinal diseases using endoscopy images. *Diagnostics*, 12(10):2316, 2022.

[24] Muhammad Attique Khan, Naveera Sahar, Wazir Zada Khan, Majed Alhaisoni, Usman Tariq, Muhammad H Zayyan, Ye Jin Kim, and Byoungchol Chang. Gestronet: a framework of saliency estimation and optimal deep learning features based gastrointestinal diseases detection and classification. *Diagnostics*, 12(11):2718, 2022.

[25] Yaw Afriyie, Benjamin A. Weyori, and Alex A. Opoku. Gastrointestinal tract disease recognition based on denoising capsule network. *Cogent Engineering*, 9(1):2142072, 2022.