Please cite the Published Version

Data Access Statement: Data will be made available on request.
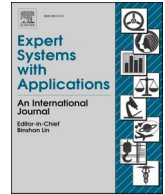
Detection of various gastrointestinal tract diseases through a deep learning method with ensemble ELM and explainable AI

Md. Faysal Ahamed [a,1], Md. Nahiduzzaman [a,2], Md. Rabiul Islam [b,3], Mansura Naznine [b,4], Mohamed Arselene Ayari [c,5], Amith Khandakar [c,6], Julfikar Haider [d,7,*]

[a] *Department of Electrical & Computer Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh*
[b] *Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh*
[c] *Department of Electrical Engineering, Qatar University, Doha 2713, Qatar*
[d] *Department of Engineering, Manchester Metropolitan University, Chester Street, Manchester M1 5GD, UK*

ARTICLE INFO

ABSTRACT

The rising prevalence of gastrointestinal (GI) tract disorders worldwide highlights the urgent need for precise diagnosis, as these diseases greatly affect human life and contribute to high mortality rates. Fast identification, accurate classification, and efficient treatment approaches are essential for addressing this critical health issue. Common side effects include abdominal pain, bloating, and discomfort, which can be chronic and debilitating. Nausea and vomiting are also frequent, leading to difficulties in maintaining adequate nutrition and hydration. The current study intends to develop a deep learning (DL)-based approach that automatically classifies GI tract diseases. For the first time, a GastroVision dataset with 8000 images of 27 different GI diseases was utilized in this work to design a computer-aided diagnosis (CAD) system. This study presents a novel lightweight feature extractor with a compact size and minimum number of layers named Parallel Depthwise Separable Convolutional Neural Network (PD-CNN) and a Pearson Correlation Coefficient (PCC) as the feature selector. Furthermore, a robust classifier named the Ensemble Extreme Learning Machine (EELM), combined with pseudo inverse ELM (ELM) and L1 Regularized ELM (RELM), has been proposed to identify diseases more precisely. A hybrid pre-processing technique, including scaling, normalization, and image enhancement techniques such as erosion, CLAHE, sharpening, and Gaussian filtering, are employed to enhance image representation and improve classification performance. The proposed approach consists of twenty-four layers and only 0.815 million parameters with a 9.79 MB model size. The proposed PD-CNN-PCC-EELM extracts essential features, reduces computational overhead, and achieves excellent classification performance on multiclass GI images. The PD-CNN-PCC-EELM achieved the highest precision, recall, f1, accuracy, ROC-AUC, and AUC-PR values of 88.12 ± 0.332 %, 87.75 ± 0.348 %, 87.12 ± 0.324 %, 87.75 %, 98.89 %, and 92 %, respectively, while maintaining a minimum testing time of 0.000001 s. A comparative study utilizes 10-fold cross-validation, ablation study and various state-of-the-art (SOTA) transfer learning (TL) models as feature extractors. Then, the PCC and EELM are integrated with TL to generate predictions, notably in terms of performance and real-time processing capability; the proposed model significantly outperforms the other models. Moreover, various explainable AI (XAI) methods, such as SHAP (Shapley Additive Explanations), heatmap, guided heatmap, Grad-Cam (Gradient-weighted Class Activation Mapping), guided Grad-CAM, and guided Saliency mapping, have been employed to explore the interpretability and decision-making capability of the proposed model. Therefore, the model provides practical intelligence for increasing confidence in diagnosing GI diseases in real-world scenarios.

* Corresponding author.
E-mail address: j.haider@mmu.ac.uk (J. Haider).
[1] 0000-0002-7014-3205
[2] 0000-0003-4126-0389
[3] 0000-0003-1989-4385
[4] 0009-0007-0296-9981
[5] 0000-0002-8663-886X
[6] 0000-0001-7068-9112
[7] 0000-0001-7010-8285

1. Introduction

The gastrointestinal (GI) system, which includes the organs associated with digestion and food absorption, is essential for sustaining good health. This intricate system is susceptible to multiple disorders that can significantly impact its daily functioning. GI diseases such as polyps, esophageal disorders, colon cancer, and ulcerative colitis affect organs such as the stomach, intestines, liver, and pancreas. Medical imaging technology has made significant strides toward automatic diagnosis of these diseases in the last 20 years. Early identification and accurate diagnosis are essential for successful treatment of many diseases, but a large number of healthcare experts are needed, which is costly, prone to error, and time consuming. Moreover, rural areas often struggle to fulfill the need for more skilled medical professionals (Sung et al., 2021). Addressing these issues necessitates technological solutions that can automatically and accurately detect and assess GI diseases.

Digestive diseases significantly increase mortality rates, indicating an alarming phenomenon in public health recently. Colorectal cancer is one of the most common GI illnesses. Since 2015, almost 132,000 new cases of colorectal cancer have been reported in the USA, affecting 1.6 million individuals with bowel infections. Approximately 200,000 new cases occur each year (Khan, Sarfraz, et al., 2020). In 2017, 135,430 cases of various GI diseases were documented in the USA (Khan, Khan, et al., 2020). Additionally, 18 % of adults in Brazil, 11 % in China, 20 % in the EU-5, 12 % in Russia, and 21 % in the US were diagnosed with GI diseases (Sharif et al., 2021). In 2017, a worldwide survey reported 765,000 fatalities from stomach diseases, with colon cancer being responsible for 525,000 deaths (Khan, Kadry, et al., 2020). Worldwide, there were approximately 4.8 million new cases of GI malignancies and 3.4 million deaths related to these illnesses in 2018 (Arnold et al., 2020). Approximately 3.6 million children are affected by stomach infections each year (Khan, Sarfraz, et al., 2020). Esophageal cancer is the seventh most prevalent cancer worldwide, whereas stomach cancer is the third leading cause of cancer-related fatalities globally.

Accurate and early diagnosis of GI diseases plays a significant role in reducing the mortality rate. Endoscopy, which includes esophagogastroduodenoscopy (EGD) and colonoscopy, is one of the best and most effective ways to examine the upper and lower intestines for potential health problems. Furthermore, capsule endoscopy, endoscopic ultrasonography (EUS), CT scan, magnetic resonance imaging (MRI), and positron emission tomography (PET) scan are other important techniques for the thorough diagnosis of GI disorders. These techniques support medical personnel in observing the GI tract, evaluating organ health, and detecting abnormalities quickly, which helps in timely intervention and enhances the possibility of successful treatment outcomes (Arnold et al., 2020).

The complicated structure of small bowls makes push gastroscopy instruments unsuitable for the identification and analysis of GI infections such as polyps, ulcers, and bleeding. The small bowel, or small intestine, is a long, coiled tube where most of the digestion and absorption of nutrients occurs. Its intricate structure and length can make it challenging to navigate certain endoscopic instruments, making it difficult to identify and analyze gastrointestinal issues within this part of the digestive tract. Standard endoscopy may fail to detect numerous lesions because of the presence of secretions. During colon cleansing operations intended for cancer or precursor lesion diagnosis, a significant number of polyps remain undiscovered, with rates ranging from 21.4 % to 26.8 %, which poses significant challenges (Kim et al., 2017a). Furthermore, as polyp growth may exhibit similarities among multiple categories, accurate diagnosis can be challenging. In 2000, a new technology called Wireless Capsule Endoscopy (WCE) was introduced to resolve these issues to a certain extent (Iddan et al., 2000). During WCE, a medical professional visually inspects the interior of the GI tract to identify any diseases. The patient swallowed a capsule with a wireless camera, light-emitting diodes, radio frequency emitter, and a battery throughout this procedure. The system autonomously navigates through the GI tract, and the camera captures thousands of images. The images are stored on recorders and then transmitted to a computer with specialized software that compiles them to form a video. The gastroenterologists evaluated those images and tracked the lesion. However, the primary issue of this process is the longer time needed to classify many types of GI diseases. Over 50,000 pictures are generated during a WCE scan. Physicians need an average of two hours to analyze the images, and the risk of incorrect detection (25 % overall) is very high (Fan et al., 2018).

Previous research has shown substantial progress in developing various artificial intelligence (AI) models for classifying the GI tract. These models utilize a variety of methodologies, such as rule-based reasoning and neural networks (NNs) (Aruna et al., 2007; Awais & Awan, 2011; Saraiva et al., 2016). Although significant progress has been made, certain challenges need to be addressed. Previous research has mostly concentrated on developing image diagnosis methods to precisely classify precursor lesions associated with GI disorders (Iakovidis & Koulaouzidis, 2014; Lee et al., 2019; Li & Meng, 2009; Noya et al., 2017; Pan et al., 2011; Ye & Prince, 2016). Traditional approaches in these studies included improving contrast, removing noise, and segmenting regions. Several studies have also delved into classifying diseases within the GI tract (Gunasekaran et al., 2023; Noor et al., 2023; Nouman Noor et al., 2023). However, a significant drawback of these approaches is their focus on a limited number of diseases and a limited number of samples, approximately 1650 to 4854, and lack of models' interpretability. Additionally, certain researchers have utilized transfer learning (TL)-based methods that involve a significant number of parameters and layers. However, these methods require substantial processing time because of extracting irrelevant features, which creates significant challenges for real-time applications.

The aim of this research is to mitigate the conventional problems of previous studies by introducing a comprehensive disease classification framework. The primary contributions of this research are outlined as follows:

- For the first time, a DL model is employed to classify a large number of GI tract diseases (27 classes) which contains a large number of upper, lower, and combined GI samples.
- A hybrid preprocessing method (CLAHE, erosion, sharpening, and Gaussian filtering) was introduced to enhance image quality on the multi-class GastroVision dataset.
- A novel lightweight DL model called Parallel Depth-wise separable CNN (PD-CNN) is proposed to perform feature extraction with distinct characteristics and minimal parameters, resulting in a significant decrease in model size, parameters, layers, and testing time.
- The Pearson correlation coefficient (PCC) has been utilized to reduce irrelevant features by assessing the linear connection between the features and target classes, which improves the effectiveness of the proposed PD-CNN model.
- A novel Ensemble Extreme Learning Machine (EELM) classifier, which is a combination of pseudo inverse-ELM (ELM) and L1-regularized ELM (RELM), is designed to accelerate performance in classifying GI tract diseases.
- This study evaluated the classification performance, parameters, layers, and sizes of the proposed PD-CNN model with different established transfer learning (TL) models to demonstrate the'superiority of the proposed model.
- The'interpretability of the framework is highlighted by the use of various explainable AI (XAI) techniques, including Shapley Additive exPlanations (SHAP), heatmap, guided heatmap, Grad-CAM, guided Grad-CAM and guided Saliency mapping, which demonstrate significant insights into the model's decision-making capability.

Section 2 provides a detailed summary of the previous relevant research. Section 3 outlines the suggested methodology, consisting of a proposed framework and dataset, and section 4 demonstrates the

detailed model architecture. Section 5 provides an elaborate overview of complete classification outcomes, accompanied by interpretability of the proposed model using XAI. Section 6 presents the key conclusions.

2. Related works

Multiple approaches have been explored in the field of medical diagnosis and decision support for GI diseases (Fujii-Lau et al., 2023; Gupta et al., 2022; Johannes et al., 2008; Jun et al., 2022; Kusano et al., 2024; Nass et al., 2022; Parasa et al., 2023; Parsa et al., 2018). Researchers have utilized several ML and DL approaches to detect and analyze various GI diseases, including cancer and ulcer, from endoscopic images [13–15]. Additionally, ongoing research is being conducted in the field of multi-class classification but mostly encompasses a smaller number of classes (Gunasekaran et al., 2023; Noor et al., 2023; Nouman Noor et al., 2023; Rustam et al., 2021; Sivari et al., 2023).

A medical diagnosis decision support model for gastrointestinal cancer was presented by Saraiva et al. (Saraiva et al., 2016) using a mix of rule-based and case-based reasoning. Subsequently, Aruna et al. (Aruna et al., 2007) presented an NN model for GI diagnosis that used radial basis functions and backpropagation. The fuzzy inputs used in this model were derived from patient interviews. Furthermore, Awais et al. (Awais & Awan, 2011) presented a unique model for myocardial infarction detection that was based on the defense mechanism of the human digestive system. Additionally, a polyp detection method for WCE images was presented by Li et al. (Li & Meng, 2012). Using a support vector machine (SVM), classifier features were extracted through the integration of wavelet transform and a uniform local binary pattern. All these studies were carried out on comparable tract areas, yet very few of these studies specifically addressed multiclass classification.

Researchers have also used different aspects of the human body, such as cancer, blood, polyps, ulcer lesions, dyed-lifted-polyps, and ileocecal, to detect problems in endoscopic images (Iakovidis & Koulaouzidis, 2014; Lee et al., 2019; Li & Meng, 2009; Noya et al., 2017; Pan et al., 2011; Ye & Prince, 2016). Musha et al. (Musha et al., 2023) suggested a method utilizing a chromaticity moment color feature and uniform local binary pattern for bleeding region detection in endoscopy images. Similarly, Pan et al. (Pan et al., 2011) used a probabilistic NN for bleeding detection. On the other hand, Noya et al. (Noya et al., 2017) applied a boosted decision tree (DT) classifier using a combination of color-based, texture, statistical and morphological features for detecting angiodysplasia lesions. Li et al. (Li & Meng, 2009) introduced a texture extraction process curvelet-based local binary pattern for the detection of ulcer regions in capsule endoscopy images. Using multilayer perceptron NN and SVM, they classified ulcer regions. Most of these studies have emphasized extracting many features overachieving a balanced real-time benchmark. Morphological operations and statistical analysis were conducted to produce the data.

Yeh et al. (Yeh et al., 2014) proposed a method for detecting ulcers and bleeding in images acquired using WCE. Color features have been used to evaluate the condition of the small intestine, and various feature selection approaches and classifiers have been utilized, with the DT demonstrating the highest accuracy in detecting bleeding. In a separate study, Lee et al. (Lee et al., 2019) evaluated TL models such as ResNet50, Inceptionv3, and VGG16 to classify stomach endoscopic images and distinguish between normal and benign ulcers. The results showed high accuracy, with ResNet50 consistently performing better than the other methods. Yuan et al. (Yuan et al., 2015) generated a computer-aided technique for detecting ulcers. The approach effectively identified ulcer regions through a multi-level super-pixel representation. The method utilized Locality-Constrained Linear Coding (LLC) and saliency max-pooling to save visual features. Furthermore, Pogorelov et al. (Pogorelov et al., 2017) created Kvasir, a multi-class image dataset designed for computer-aided identification of GI diseases, with annotated images from the GI tract. The collection contains anatomical landmarks, pathological observations (such as esophagitis, polyps, and ulcerative colitis), and images associated with endoscopic polyp removal. Jain et al. (Jain et al., 2021) developed WCENet, a deep CNN model for detecting and locating anomalies in WCE images. Their proposed model worked in two stages: an initial stage utilizing an attention-based CNN to categorize images into certain groups (polyp, vascular, inflammatory, or normal), followed by a phase that combines Grad-CAM++ with a customized SegNet for locating anomalies in images.

As mentioned before, very little work has been done recently on the multi-class classification of GI tract diseases. Nouman et al. (Nouman Noor et al., 2023) proposed a method for classifying GI tract diseases using the Kvasir (Pogorelov et al., 2017) and Hyper-Kvasir (Borgli et al., 2020) datasets, which contain 4854 images of five different classes. The process included contrast optimization using a genetic algorithm (GA), utilizing MobileNetV2 for feature extraction, and applying the machine learning classifier SoftMax. Similarly, Noor et al. (Noor et al., 2023) suggested a computer-aided diagnosis system for GI diseases with a lightweight MobileNetV2 feature extractor and SoftMax classifier. They also have integrated attention mechanisms and a cosine similarity-based feature selection technique to reduce the number of features to improve the effectiveness of the classification. Using 810 key features, the framework achieved high accuracy (97.68 %) in classifying GI tract images into 5 different classes of the Kvasir dataset (Pogorelov et al., 2017). Gunasekaran et al. (Gunasekaran et al., 2023) presented GIT-NET, an ensemble model that uses the pretrained models DenseNet201, InceptionV3, and ResNet50 to classify GI diseases accurately. The Kvasir v2 (Pogorelov et al., 2017) dataset, which has 8000 photos from 8 classes, was utilized in this study. With an accuracy of 95 %, the proposed weighted average ensemble method outperforms individual models. Sivari et al. (Sivari et al., 2023) also utilized the Kvasir v2 and Hyper-Kvasir datasets to develop a DL-based hybrid stacking ensemble model for the detection and classification of the GI tract from endoscopic images. The models were trained using a two-level stacking architecture. The second level includes logistic regression, linear SVM, multi-layer perceptron, and k-nearest neighbor algorithms. Rustam et al. (Rustam et al., 2021) introduced a bleedy image recognizer (BIR) combining MobileNet and a custom-built CNN model for automatic analysis of WCE images. With a dataset of 1650 images, BIR achieved impressive performance, demonstrating a high accuracy of 99.3 %. Lan et al. (Lan & Ye, 2021) carried out a study that introduced a hybrid unsupervised DL technique to summarize videos within a weakly supervised cross-modal embedding framework. They used networks such as long short-term memory (LSTM) and autoencoder to help healthcare professionals analyze WCE videos in detail. In (Alhajlah et al., 2023), a method that integrates Mask R-CNN, fine-tuned ResNet models, and an Enhanced Ant Colony Optimization algorithm was proposed. The ResNet-50 and ResNet-152 models achieved an impressive classification accuracy of 96.43 %. Another study (Mohapatra et al., 2023) utilized discrete wavelength transformation (WT) and CNN approaches to categorize polyp and esophagitis classes. The method achieved an impressive accuracy rate of 96.65 %.

Several researchers achieved a higher level of accuracy ranging from 93 % to 98 % in their research. However, it is important to note that these researchers utilized datasets with a limited number of classes (ranging from 5 to 8) and a small number of image samples (ranging from 4000 to 8000) to demonstrate the performance of their proposed models. Furthermore, most researchers have employed TL- and DL-based models, including GIT-NET, MobileNetV2, ResNet-50, ResNet-152, and a hybrid stacking ensemble model. Nevertheless, a major problem arises regarding the computational requirements (parameter, size, layer) that lead to longer processing times, which makes it difficult to use the classification model effectively. Furthermore, certain advanced techniques require high-resolution images to achieve precise classification, which is crucial for real-world applications, especially in embedded systems (Alhajlah et al., 2023; Aruna et al., 2007; Iakovidis & Koulaouzidis, 2014; Khan, Sarfraz, et al., 2020; Kim et al., 2017b; Lee et al., 2019; Li & Meng, 2009; Noor et al., 2023; Nouman Noor et al.,
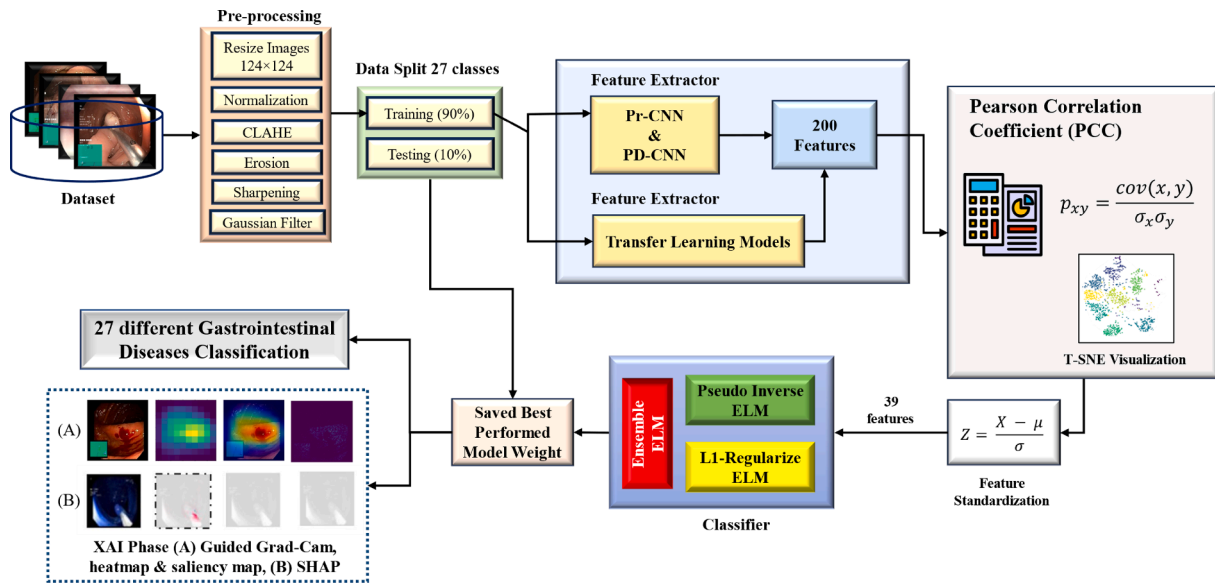
Fig. 1. Proposed working framework for multi-class classification of GI diseases.

Table 1
Dataset distribution in twenty-seven classes and data split into training and testing sets.

Testing phase	GI Position	labeled areas	Disease Types	Class No.	Training	Testing
GI tract Diseases (27-classes)	Upper GI	Normal findings	Normal stomach	20	872	97
			Normal esophagus	18	126	14
		Anatomical Landmarks	Gastroesophageal_junction_normal z-line	15	297	33
			Duodenal bulb	8	185	20
			Pylorus	21	354	39
		Pathological Findings	Barrett's esophagus	2	86	9
			Esophagitis	13	96	11
			Gastric polyps	14	59	6
			Ulcer	26	5	1
			Esophageal varices	12	6	1
	Lower GI	Normal Findings	Normal mucosa and vascular pattern in the large bowel	19	1320	147
		Anatomical Landmarks	Cecum	4	102	11
			Colon diverticula	5	26	3
			Ileocecal valve	16	180	20
			Retroflex rectum	24	60	7
			Small bowel_terminal ileum	25	761	85
		Pathological Findings	Angioectasia	1	15	2
			Mucosal inflammation large bowel	17	26	3
			Colon polyps	6	738	82
			Colorectal cancer	7	125	14
		Therapeutic interventions	Dyed-lifted-polyps	9	127	14
			Dyed-resection-margins	10	221	25
			Resected polyps	22	83	9
			Resected margins	23	23	2
	Upper & Lower GI	Pathological findings	Blood in lumen	3	154	17
			Erythema	11	14	1
		Therapeutic interventions	Accessory tools	0	1139	127
	Total				**7200**	**800**

2023; Noya et al., 2017; Pan et al., 2011; Saraiva et al., 2016). To accelerate the widespread use of the GI disease classification model, it is crucial to improve existing models by reducing parameter counts, size, and layers; minimizing processing times; and boosting classification accuracy. In addition, certain studies have demonstrated the application of XAI techniques such as heatmaps and Grad-CAM. However, most of the SOTA research has not focused much on assessing the impact of individual features. This study acknowledges these challenges and suggests a novel and efficient solution.

3. Methodology

3.1. Proposed framework

A novel methodology utilizing the DL approach has been developed to address the complicated issues linked to identifying a large number of GI disorders. The procedural phases involved in this research are illustrated in Fig. 1. Initially, the annotated dataset was divided into twenty-seven unique disease categories at a 90:10 ratio for the training and testing sets. Several data preprocessing steps were executed to enhance the model's learning ability. These procedures included implementing data normalization, CLAHE, erosion, sharpening, Gaussian filtering, and resizing images from the training set to 124 × 124 pixels. Next, a
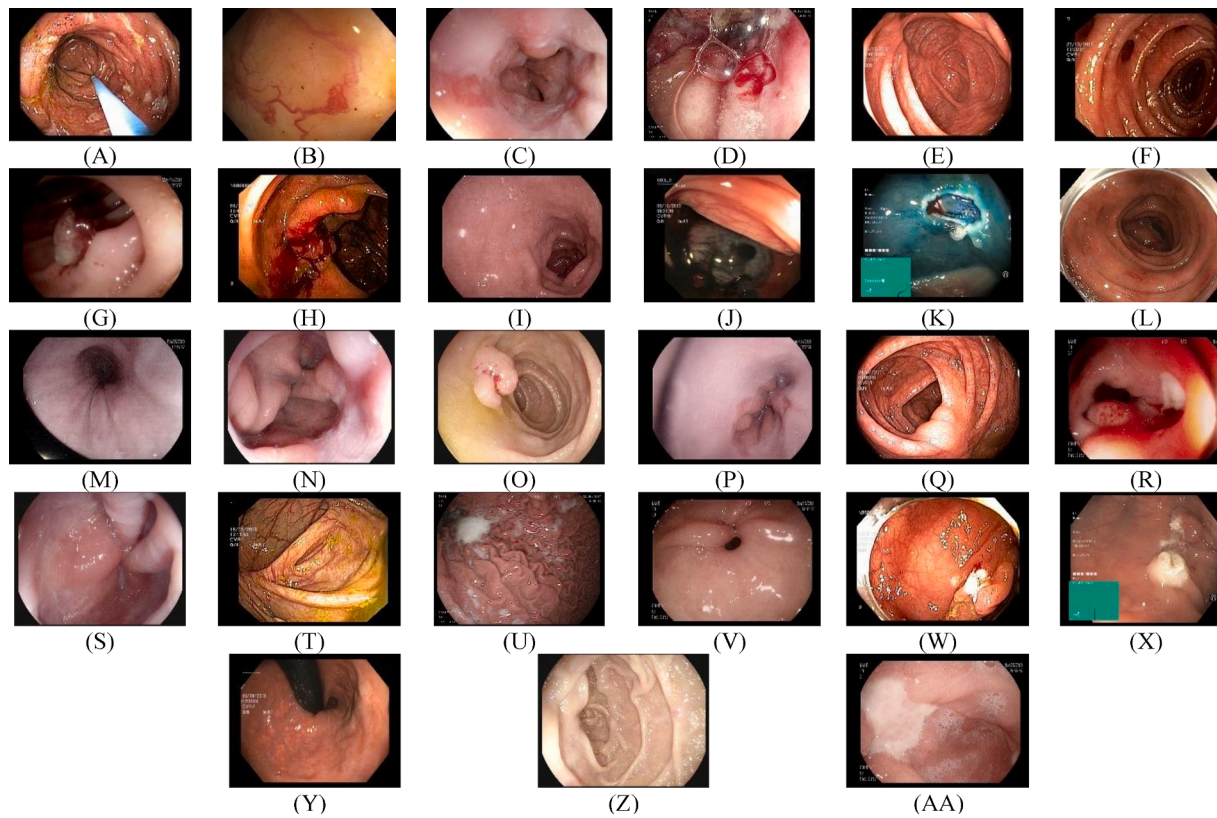
Fig. 2. GastroVision dataset include (A) Accessory tools, (B) Angiectasia, (C) Barrett's esophagus, (D) Blood in lumen, (E) Cecum, (F) Colon diverticula, (G) Colon polyps, (H) Colorectal cancer, (I) Erythema, (J) Dyed-lifted-polyps, (K) Dyed-resection-margins, (L) Erythema, (M) Esophageal varices, (N) Esophagitis, (O) Gastric polyps, (P) Gastroesophageal_junction_normal z-line, (Q) Ileocecal valve, (R) Mucosal inflammation large bowel, (S) Normal esophagus, (T) Normal mucosa and vascular pattern in the large bowel, (U) Normal stomach, (V) Pylorus, (W) Resected polyps, (X) Resection margins, (Y) Retroflex rectum, (Z) Small bowel_terminal ileum, and (AA) Ulcer multi-classes.

lightweight novel Parallel Depthwise separable CNN (PD-CNN) architecture was constructed in conjunction with a parallel CNN (Pr-CNN) in which traditional convolutional layers were used. This architecture was utilized to test multiple TL models simultaneously and extract 200 features. Thirty-nine significant features were identified by eradicating 161 irrelevant features via the PCC algorithm and t-distributed stochastic neighbor embedding (t-SNE) visualization of the feature distribution. Subsequently, Z score normalization was applied to achieve standardization. An ensemble extreme learning machine (ELM) classifier was designed to improve the model's classification performance from the PCC features. This classifier incorporated the ELM and RELM approaches. The model weights that yielded the most accurate predictions were maintained after a comparative analysis. Moreover, by employing various XAI techniques, the decision-making capabilities of the proposed models were graphically presented.

3.2. Dataset description

A comprehensive multi-center dataset developed for GI endoscopy applications, denoted GastroVision, was used in this study (Jha et al., 2024). The objective of GastroVision is to aid in the advancement and assessment of AI-driven algorithms utilized for the identification and categorization of gastrointestinal disorders. With a total of 27 unique classes representing different GI tract diseases, the dataset contains a broad variety of anatomical landmarks, clinical abnormalities, normal results, and instances of polyp removal. A combined group of images representing both normal variations and GI pathology is available in addition to the upper GI and lower GI categories, which are based on the digestive tract. The data collection process for GastroVision involved a collaborative effort between Bærum Hospital in Norway and Karolinska

University Hospital in Sweden. Skilled GI endoscopists meticulously conducted endoscopic procedures, capturing high-resolution images of diverse GI tract regions, including the esophagus, stomach, small intestine, colon, rectum, and terminal ileum. Following data acquisition, expert GI endoscopists meticulously annotated and verified the images. The dataset comprised a comprehensive collection of 8,000 high-resolution endoscopic images from various GI regions on the human body. Moreover, the dataset is thoughtfully distributed across different areas of the GI tract, ensuring comprehensive coverage of GI pathology and normal variations. Comprehensive details of the dataset, along with sample images of the model training and testing sets, are provided in Table 1 and Fig. 2. Notably, the KvasirV2 dataset was also utilized with the proposed model to justify the comparative performances (Pogorelov et al., 2017).

3.3. GI tract diseases

Understanding the various disease classes is crucial for accurate diagnosis and treatment. Table 2 presents a brief description of 27 GI tract diseases (Jha et al., 2024).

3.4. Data preprocessing

The image processing step is vital for optimizing the performance of deep learning models. GastroVision contains images of varying sizes, with the following distributions: 2,647 samples at 576×720 pixels, 3,890 samples at 576×768 pixels, 976 samples at 1024×1280 pixels, 8 samples at 1048×1232 pixels, 67 samples at 1064×1350 pixels, 347 samples at 1072×1920 pixels, and 65 samples at 1080×1350 pixels. Each image was scaled to a consistent resolution of 124×124 pixels,

Table 2
Definitions of GI tract diseases.

GI tract Diseases	Description
Normal Stomach	A healthy stomach with no disease or abnormalities, showing a smooth mucosal lining.
Normal Esophagus	A healthy esophagus that appears pink and soft without any signs of inflammation or lesions
Gastroesophageal Junction Normal Z-Line	The area where the esophagus meets the stomach, marked by a normal Z-line indicating no abnormal tissue growth.
Duodenal Bulb	The first part of the small intestine just beyond the stomach, appearing healthy and free of ulcers or inflammation.
Pylorus	The opening from the stomach into the duodenum, functioning normally without obstruction or thickening
Barrett's Esophagus	A condition where the lining of the esophagus changes, becoming similar to the lining of the intestine, often due to acid reflux.
Esophagitis	Inflammation of the esophagus, usually caused by acid reflux, infections, or medications, resulting in pain and difficulty swallowing
Gastric Polyps	Small growths on the lining of the stomach, which can be benign or precancerous, requiring monitoring or removal.
Ulcer	A sore on the lining of the stomach or duodenum, often caused by *Helicobacter pylori* infection or the use of NSAIDs, leading to pain and bleeding.
Esophageal Varices	Swollen veins in the esophagus, usually due to liver disease, posing a risk of bleeding.
Normal Mucosa and Vascular Pattern in the Large Bowel	Healthy large bowel tissue with no signs of disease, showing a normal vascular pattern.
Cecum	he beginning of the large intestine, appearing healthy and free of inflammation or polyps.
T Colon Diverticula	Small pouches that can form in the colon wall, which can become inflamed or infected, causing diverticulitis.
Ileocecal Valve	The valve between the small intestine and large intestine, functioning normally without signs of disease or obstruction.
Retroflex Rectum	A technique used to view the rectum from a different angle during endoscopy, showing normal tissue.
Small Bowel Terminal Ileum	The last part of the small intestine, appearing healthy without signs of Crohn's disease or other conditions.
Angioectasia	Abnormal blood vessels in the GI tract, which can cause bleeding and anemia.
Mucosal Inflammation Large Bowel	Inflammation of the lining of the large intestine, often due to conditions like ulcerative colitis or Crohn's disease.
Colon Polyps	Growths on the lining of the colon, which can be benign, precancerous, or cancerous, requiring removal and monitoring.
Colorectal Cancer	Cancerous growths in the colon or rectum, often detected through screening methods like colonoscopy.
Dyed-Lifted Polyps	Polyps that have been lifted using a dye during endoscopy to aid in removal and visualization.
Dyed-Resection Margins	Margins of tissue that have been dyed during polyp removal to ensure complete resection.
Resected Polyps	Polyps that have been surgically removed from the colon or rectum.
Resected Margins:	The edges of tissue that have been removed along with a polyp, are checked to ensure no cancerous cells remain.
Blood in Lumen	Presence of blood in the lumen of the GI tract, indicating bleeding from a lesion, ulcer, or varices.
Erythema	Redness of the mucosa, often a sign of inflammation or irritation.
Accessory Tools	Instruments used during endoscopic procedures to aid in diagnosis and treatment, such as biopsy forceps, snares, and injection needles.

and z score normalization procedures were performed to ensure consistency and optimal data representation for model training. Additionally, using a slightly smaller size helps avoid issues with padding or borders that might arise during convolution operations within the network (Hesse et al., 2023). Furthermore, pretrained models might have been trained with non-standard input sizes, and to maintain compatibility and benefit from transfer learning, a 124×124 input size was used. A random splitting technique was used to divide the dataset, allocating 90 % of the images for training and 10 % for testing. This method was crucial due to the limited number of images in certain classes, ensuring that a sufficient number of training samples remained to effectively train the model across all classes.

Sequential preprocessing steps significantly enhance the robustness and performance of the model across diverse datasets (Li et al., 2022). It includes a series of image enhancement techniques, such as scaling, normalization, erosion, CLAHE, sharpening, and Gaussian filters. Erosion reduces noise and sharpens image boundaries, while CLAHE enhances contrast, particularly in regions with varying illumination. Sharpening techniques emphasize edge detection and overall image clarity, while Gaussian filters effectively reduce noise and improve image characteristics. The sequential integration of these preprocessing methods ensures that the model receives refined and standardized input data, leading to improved performance and accuracy in classification tasks. Examples demonstrating the enhancement of the original images through each preprocessing step are illustrated in Fig. 3. No augmentation methods were used to address the dataset's class imbalance, influenced by the limited number of samples in classes such as "Mucosal inflammation large bowel", "Resected margins", "Colon diverticula", "Ulcer", "Esophageal varices", "Angioectasia", and "Erythema," which contained only 29, 25, 29, 6, 7, 17, and 15 samples, respectively, compared to classes with a larger number of images. Balancing the dataset across these classes presented significant challenges due to their uneven distribution (Xu et al., 2023).

4. Architecture

The dataset is ready for training after successfully performing the preprocessing stage. In the current research era, the most challenging term is to build a robust, lightweight model that can provide the best performance while maintaining minimum testing time, parameters, layers, and sizes. Next, a novel lightweight PD-CNN feature extractor was proposed, and for fundamental feature selection, the PCC was concurrently utilized. Furthermore, a novel classifier, the EELM, has been proposed to improve the model's decision-making capability.

4.1. Feature extraction

The primary objective of this research was to construct a custom CNN design that can extract critical features while simultaneously reducing both parameters and network depth. The configuration of layers in a CNN is of the utmost importance. An overabundance of parameters and layers may impede the model's ability to differentiate distinctive attributes, thus imposing a performance constraint. Conversely, an overabundance of parameters and layers increases the demand for computational resources and lengthens processing times due to the chance of overfitting. Therefore, it is crucial to achieve an ideal equilibrium to guarantee precise feature extraction and feasible implementation.

Fig. 4 presents the proposed PD-CNN model architecture adept at managing layer complexities and adhering to parameter constraints to fulfill its objective effectively. The model incorporates convolution layers (CLs) and fully connected layers (FCs) to find an optimal balance. Instead of relying on a single CL model, five parallel CL models were employed to identify the essential features. This challenge was addressed by concurrently running the initial parallelly connected five depth-wise separable CLs instead of employing five consecutive CLs,
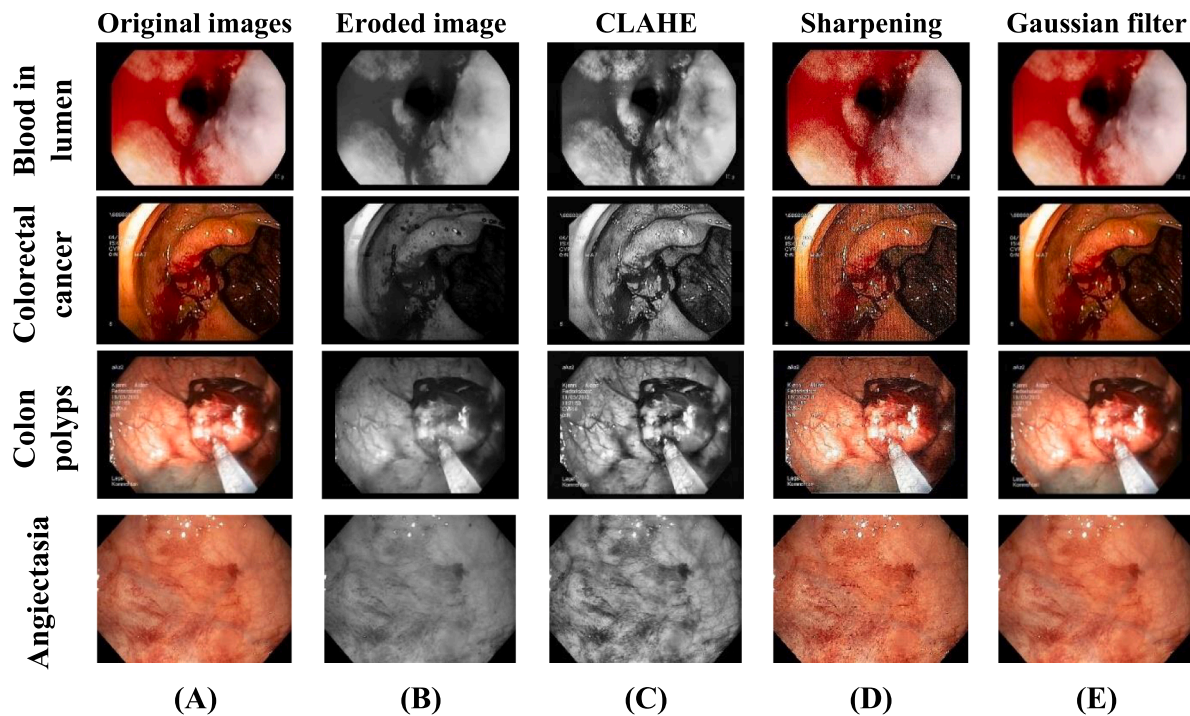
Fig. 3. Image pre-processing include: (A) original images, (B) eroded images, (C) CLAHE images, (D) sharpened images, and (E) Gaussian filtered images.
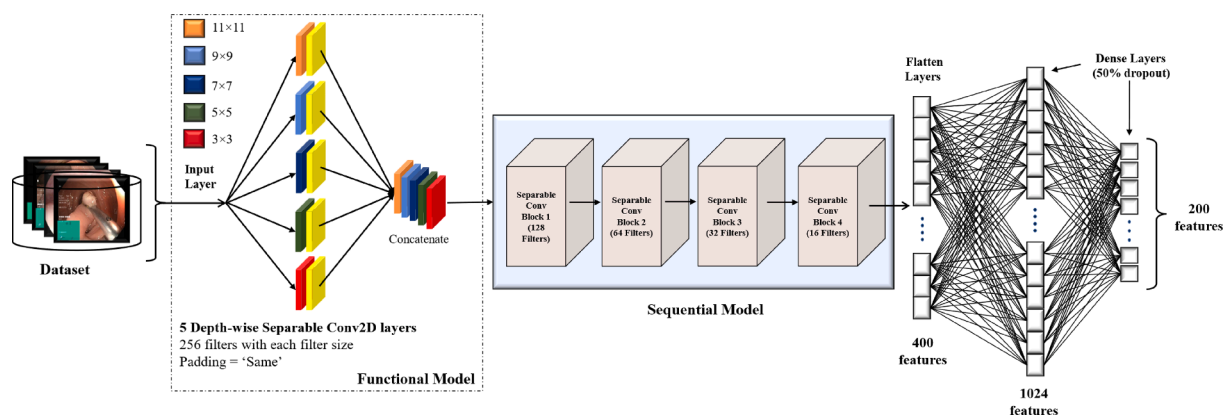

Fig. 4. Proposed PD-CNN-EELM architecture.

which would amplify network depth and complexity (Kaiser et al., 2017). Their selection was determined through a systematic trial-and-error process.

The model begins with an input layer that accommodates images of variable dimensions. Subsequently, a series of parallel CLs with varying kernel sizes (11×11, 9×9, 7×7, 5×5, and 3×3) are applied to capture spatial hierarchies and diverse patterns within the input data. For this research, the kernel size selection method suggested by Krizhevsky et al. was implemented (Krizhevsky et al., 2017). This method entails the utilization of 11×11 kernel sizes, which have been found to yield satisfactory classification performance. Acknowledging the importance of the diverse proportions of kernels, we undertook a comprehensive examination and synthesis of various kernels to identify critical characteristics and enhance the efficacy of classification. This methodology recognizes the distinct feature maps produced by various kernels. To optimize the results obtained from extracting critical data from the frame features of GI images, it is crucial to ensure that the initial five CLs have a consistent buffer size. The feature maps acquired from these concurrent CLs must be error-free and seamlessly integrated

into a sequential CL to preserve the integrity of the classification process.

The concatenated feature maps are then fed into subsequent layers comprising standard separable convolutions, each followed by batch normalization and rectified linear unit (ReLU) activation functions. This hierarchical architecture enables the model to progressively distil and abstract high-level features from the input data while mitigating the risk of overfitting through regularization techniques. An updated feature map with fewer channels is produced by applying a 1×1 convolutional kernel separately to each channel during the pointwise convolution process. This emphasizes the pivotal significance of Depthwise Separable Convolution (DSC), as it substantially reduces the computational complexity. During the concluding stage, three CLs were incorporated in addition to implementing BN and MP using a 2×2 kernel. The respective filter values for these CLs were 128, 64, 32, and 16; each filter utilized 3×3 kernels and VALID padding. Integrating the BN enhances the model's efficiency by recalibrating each layer's input mean and standard deviation, improving the execution speed and stability. Further details of the convolution block are presented in Fig. 5.

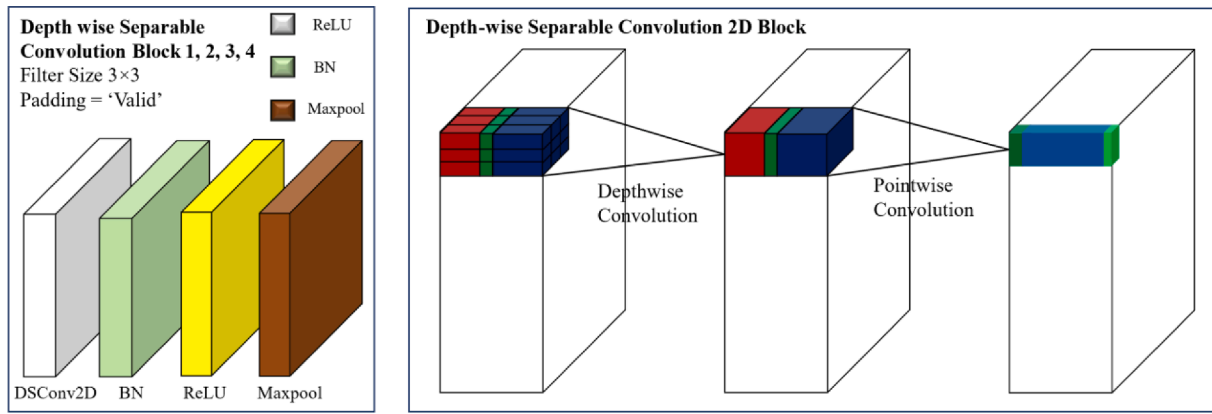The ReLU activation function was used for each CL. In addition to

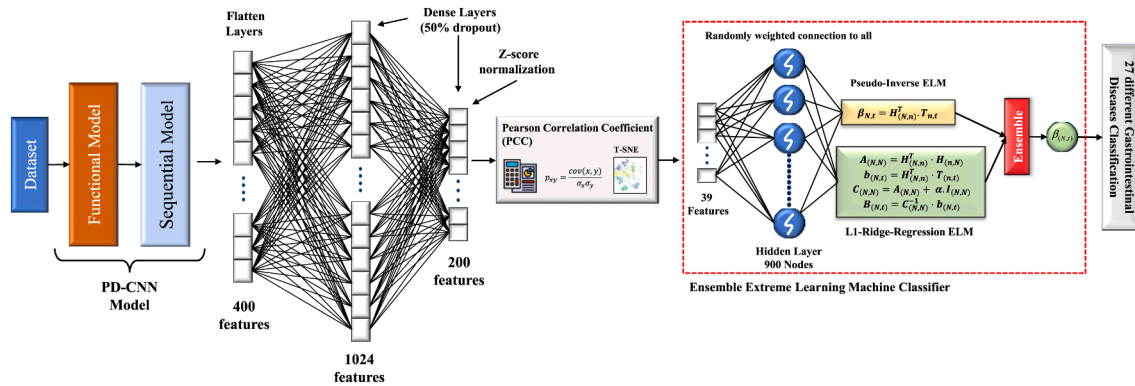**Fig. 5.** Detailed overview of the convolution block.

Fig. 6. Integration of EELM architecture with PCC from PD-CNN model's last layer.

Table 3
Summary of proposed PD-CNN model.

Layer Type	Output Shape	Parameters
Model input layer	(None, 124, 124, 3)	0
Functional model	(None, 124, 124, 1280)	5975
Separable conv2d layer	(None, 122, 122, 128)	175,488
Batch normalization	(None, 122, 122, 128)	512
Activation	(None, 122, 122, 128)	0
Max pooling	(None, 61, 61, 128)	0
Separable conv2d layer	(None, 59, 59, 64)	9408
Batch normalization	(None, 59, 59, 64)	256
Activation	(None, 59, 59, 64)	0
Max pooling	(None, 29, 29, 64)	0
Separable conv2d layer	(None, 27, 27, 32)	2656
Batch normalization	(None, 27, 27, 32)	128
Activation	(None, 27, 27, 32)	0
Max pooling	(None, 13, 13, 32)	0
Last convolution layer	(None, 11, 11, 16)	816
Batch normalization	(None, 11, 11, 16)	64
Activation	(None, 11, 11, 16)	0
Max pooling	(None, 5, 5, 16)	0
Dropout	(None, 5, 5, 16)	0
Flatten	(None, 400)	0
Dense	(None, 1024)	410,624
Batch normalization	(None, 1024)	4096
Dropout	(None, 1024)	0
Dense Last	(None, 200)	205,000
Total Parameters:	815,023	
Trainable Parameters:	812,495	
Non-Trainable Parameters:	2,528	

generalization and accelerate convergence. The final FC layer retrieved 200 features. After that, the PCC algorithm is employed to achieve significant features (39 features) from the final layer of 200 features. Furthermore, the EELM classifier is integrated to determine the final classification performance among 27 classes (Fig. 6). Table 3 provides a comprehensive overview of the model summary.

The z score normalization was applied to the 1-dimensional feature vector obtained from feature extraction, specifically to the 200 extracted features $X_i = \{x_1, x_2, \cdots, x_{200}\}$ (Singh & Singh, 2020). Z-score normalization transforms each extracted feature X_i into its z-score Z_i using the following formula:

$$Z_i = \frac{X_i - \mu_i}{\sigma_i} \tag{1}$$

where μ_i is the mean of feature X_i across the dataset and σ_i is the standard deviation of feature X_i across the dataset. This normalization standardizes the data by ensuring that each feature has a mean of zero and a standard deviation of one, facilitating easier interpretation of feature importance and enabling comparisons across different features. It also helps in stabilizing the learning process of machine learning models, leading to improved performance and robustness.

4.2. Ensemble Extreme learning Machine (EELM)

The pseudo-inverse Extreme Learning Machine (ELM) is a well-recognized approach for single-hidden layer feedforward neural networks (SLFNs) (Ding et al., 2014). In contrast to traditional NN training methods, the ELM utilizes a unique strategy of randomizing and fixing the parameters that connect the input layer to the hidden layer by using the pseudo-inverse technique. This allows for the exclusive training of
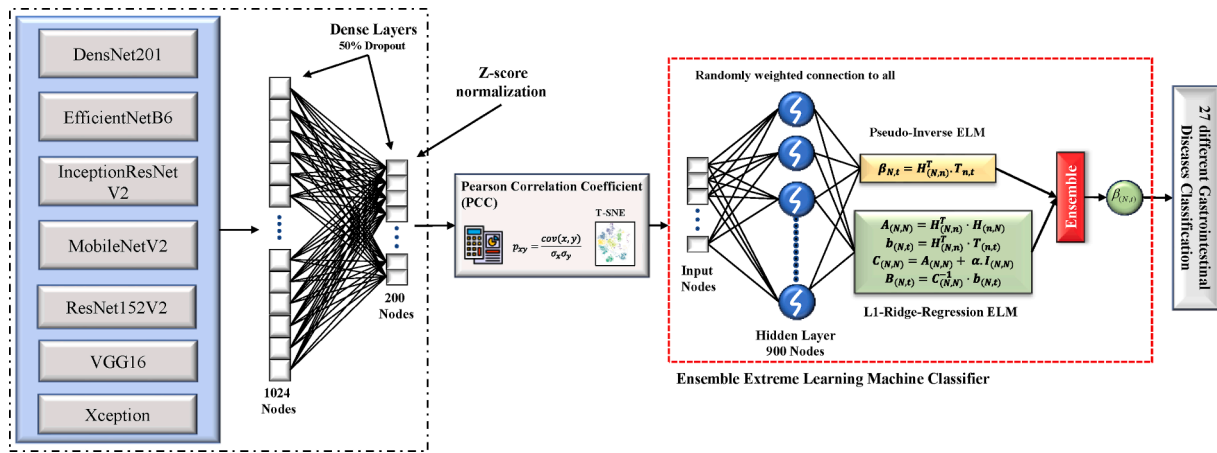
two FC layers, dropout was used to mitigate overfitting and enhance the efficiency of the training process. In every training cycle, random deactivation was applied to 50 % of all nodes, which helped enhance

Fig. 7. The modified TL architecture with PCC and EELM to classify GI diseases.

Table 4
Hyper-parameter settings for the experimental approach.

Name	Parameters
Learning Rate	0.001
Batch Size	32
epochs	200
Optimizer	Adam
Activation Function	ReLU
Loss Function	Sparse categorical cross-entropy

the parameters that link the hidden layer to the output layer. Randomized initialization accelerates training processes and enhances generalization capacities. A useful feature selection and regularization procedure is introduced into the ELM framework by including L1 regularization, often known as Lasso regularization (Shi et al., 2022). L1 regularization encourages sparsity in feature representations by adding a penalty term to the loss function, which effectively pushes many feature weights toward zero. By incorporating L1 regularization into ELM (RELM), the ability to distinguish features is improved, and the chance of overfitting is reduced, resulting in a stronger performance of

the model in generalizing.

A novel ensemble method that combines ELM and RELM classifiers has been proposed. Ensemble learning is a method that integrates many classifiers to enhance accuracy and resilience. However, the proposed methodology distinguishes itself by including ELM and L1 Regularized ELM models in this strategy. The ensemble operation combines predictions from individual ELM and RELM classifiers, each trained on separate subsets or with various initializations. This integration combines the ELM's efficiency with the RELM's feature selection, improving the classification performance. The ELM has shown remarkable competence in handling large-scale multi-class classification tasks, outperforming current machine learning models. However, this work enhances the complexity by substituting the pseudoinverse method and L1 Regularized methodology with an ensemble approach. This augmentation greatly enhances the model's ability to learn and control features, improving its potential for generalization and obtaining unmatched accuracy compared to each approach. In the classifier's design, 39 nodes are in the input layer after employing the PCC, and a staggering 900 nodes are in the hidden layer. In addition, the EELM algorithm produces twenty-seven nodes crucial for categorizing different samples from GI tract images. The ELM, RELM and proposed EELM are described

Table 5
Comparative performance among the baseline models with proposed model.

Method	Precision	Recall	F1-score	Accuracy	ROC-AUC	Testing Time
Pr-CNN-ELM	82.8 ± 0.476	81.8 ± 0.321	80.7 ± 0.224	81.8	97.87	0.00078
Pr-CNN-RELM	83.1 ± 0.471	82.6 ± 0.322	81.3 ± 0.248	82.6	98.36	0.00094
Pr-CNN-EELM	83.1 ± 0.314	82.1 ± 0.342	80.9 ± 0.315	82.1	98.32	0.00143
Pr-CNN-PCC-ELM	82.6 ± 0.489	83.0 ± 0.378	81.9 ± 0.274	83.0	98.03	0.00006
Pr-CNN-PCC-RELM	81.2 ± 0.541	81.0 ± 0.374	79.7 ± 0.233	81.0	97.97	0.00008
Pr-CNN- PCC-EELM	81.7 ± 0.521	82.1 ± 0.368	81.1 ± 0.285	82.12	98.23	0.00007
PD-CNN-ELM	87.59 ± 0.333	87.62 ± 0.343	87 ± 0.321	87.62	98.88	0.01562
PD-CNN-RELM	87.31 ± 0.33	87.25 ± 0.338	86.54 ± 0.316	87.25	98.50	0.0156
PD-CNN-EELM	**88.11 ± 0.3**	**87.75 ± 0.316**	**87.12 ± 0.285**	**87.75**	**98.92**	0.00001
PD-CNN-PCC-ELM	87.59 ± 0.334	87.62 ± 0.339	87.04 ± 0.318	87.62	98.91	0.00006
PD-CNN- PCC-RELM	87.31 ± 0.329	87.25 ± 0.3387	86.54 ± 0.314	87.25	98.43	0.00004
PD-CNN-PCC-EELM (proposed)	**88.12 ± 0.332**	**87.75 ± 0.348**	**87.12 ± 0.324**	**87.75**	98.89	**0.000001**

*Bold values indicate the best results.

Table 6

Class-wise performance of PD-CNN model without PCC on test set.

GI Disease classes	Precision			Recall			F1-score			Accuracy (%)			ROC-AUC (%)		
	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM
0	0.95	0.94	**0.95**	0.99	1	**1**	0.97	0.97	**0.97**	87.62	87.25	**87.75**	98.88	98.50	**98.92**
1	1	0	**1**	0.5	0	**0.5**	0.67	0	**0.67**						
2	0.73	0.7	**0.73**	0.89	0.78	**0.89**	0.8	0.74	**0.8**						
3	0.94	1	**1**	0.88	0.88	**0.88**	0.91	0.94	**0.94**						
4	**1**	0.9	0.91	0.91	0.82	**0.91**	**0.95**	0.86	0.91						
5	1	1	**1**	**1**	0.67	0.67	**1**	0.8	0.8						
6	0.74	0.74	**0.76**	0.88	0.9	**0.91**	0.8	0.81	**0.83**						
7	0.83	0.83	**0.91**	0.71	0.71	**0.71**	0.77	0.77	**0.8**						
8	**0.77**	0.74	0.74	0.85	0.85	**0.85**	**0.81**	0.79	0.79						
9	0.92	**0.92**	0.91	0.79	**0.79**	0.71	0.85	**0.85**	0.8						
10	**0.92**	0.88	0.85	0.92	0.92	**0.92**	**0.92**	0.9	0.88						
11	0	0	0	0	0	0	0	0	0						
12	0	0	0	0	0	0	0	0	0						
13	1	0.83	**1**	0.45	0.45	**0.45**	0.62	0.59	**0.62**						
14	0.67	0.67	**0.67**	0.33	0.33	**0.33**	0.44	0.44	**0.44**						
15	0.82	0.79	**0.82**	0.94	0.94	**0.94**	0.87	0.86	**0.87**						
16	0.85	**0.92**	0.91	0.55	**0.55**	0.5	0.67	**0.69**	0.65						
17	1	1	**1**	0.33	0.33	**0.33**	0.5	0.5	**0.5**						
18	0.86	0.86	**0.86**	0.86	0.86	**0.86**	0.86	0.86	**0.86**						
19	0.88	0.88	**0.88**	0.91	0.91	**0.91**	0.89	0.9	**0.9**						
20	0.9	**0.9**	0.89	0.94	0.94	**0.95**	0.92	0.92	**0.92**						
21	0.83	0.83	**0.83**	0.87	0.87	**0.87**	0.85	0.85	**0.85**						
22	1	1	**1**	**0.67**	0.33	0.44	**0.8**	0.5	0.62						
23	0	1	**1**	0	0.5	**0.5**	0	0.67	**0.67**						
24	0.78	0.88	**0.88**	1	1	**1**	0.88	0.93	**0.93**						
25	0.95	0.96	**0.96**	0.84	0.84	**0.84**	0.89	0.89	**0.89**						
26	0	0	0	0	0	0	0	0	0						
Average (μ) ± SD (σ) (%)	87.59 ± 0.333	87.31 ± 0.33	**88.11 ± 0.3**	87.62 ± 0.343	87.25 ± 0.338	**87.75 ± 0.316**	87 ± 0.321	86.54 ± 0.316	**87.12 ± 0.285**						

*Bold values indicate the best results.

in Algorithm 1.

The explanation of the EELM algorithm is given below:

Algorithm 1: Proposed EELM Classifier algorithm.

1. Feature sample is 'S', and output is 'O'.

$$S_{(n,m)} = \begin{bmatrix} s_{(1,1)} & s_{(1,2)} & \cdots & s_{(1,m)} \\ s_{(2,1)} & s_{(2,2)} & \cdots & s_{(2,m)} \\ s_{(3,1)} & s_{(3,2)} & \cdots & s_{(3,m)} \\ \vdots & \vdots & \ddots & \vdots \\ s_{(1,1)} & s_{(n,2)} & \cdots & s_{(n,m)} \end{bmatrix} O_{(n,t)} = \begin{bmatrix} o_{(1,1)} & o_{(1,2)} & \cdots & o_{(1,t)} \\ o_{(2,1)} & o_{(2,2)} & \cdots & o_{(2,t)} \\ o_{(3,1)} & o_{(3,2)} & \cdots & o_{(3,t)} \\ \vdots & \vdots & \ddots & \vdots \\ o_{(n,1)} & o_{(n,2)} & \cdots & o_{(n,t)} \end{bmatrix}$$

2. Input weight and bias metrics is presented as $W_{m,N}$, and $B_{(1,N)}$.

$$W_{(m,N)} = \begin{bmatrix} w_{(1,1)} & w_{(1,2)} & \cdots & w_{(1,N)} \\ w_{(2,1)} & w_{(2,2)} & \cdots & w_{(2,N)} \\ w_{(3,1)} & w_{(3,2)} & \cdots & w_{(3,N)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{(m,1)} & w_{(m,2)} & \cdots & w_{(m,N)} \end{bmatrix} B_{(1,N)} = \begin{bmatrix} b_{(1,1)} & b_{(1,2)} & \cdots & b_{(1,N)} \end{bmatrix}$$

3. Hidden layer $H_{(n,N)}$ is used to generate the output.

$$H_{(n,N)} = \begin{bmatrix} h_{(1,1)} & h_{(1,2)} & \cdots & h_{(1,N)} \\ h_{(2,1)} & h_{(2,2)} & \cdots & h_{(2,N)} \\ h_{(3,1)} & h_{(3,2)} & \cdots & h_{(3,N)} \\ \vdots & \vdots & \ddots & \vdots \\ h_{(n,1)} & h_{(n,2)} & \cdots & h_{(n,N)} \end{bmatrix} H_{(n,N)} = G(S_{(n,m)} \cdot W_{(m,N)} + B_{(1,N)}) \text{Where, } 'G'$$

denotes activation function

4. In ELM, output weight metric is presented as $\beta_{(N,t)}, \bullet \beta_{(N,t)} = H^{\dagger}_{(N,N)} \bullet T_{(n,t)}$

5. In RELM, output weight metric is presented as $\beta'_{(N,t)}$, where the ELM equations are replaced as follows:

A(N,N)=H(N,n)T•H(n,N)b(N,t)=H(N,n)T•T(n,t)C(N,N)=A(N,N)+ α•I(N, N)β'(N,t)=C(N,N)-1•b(N,t)Here, 'α' is called regularization parameter

6. The following formula denotes the proposed Ensemble operation:

$$E(N,t) = \beta(N,t) + \beta'(N,t) 2E_{(N,t)} = \frac{H^{\dagger}_{(N,N)} \cdot T_{(n,t)} + C^{-1}_{(N,N)} \cdot b_{(N,t)}}{2}$$

7. The generated prediction, $E_{(N,t)}$.

4.3. Transfer learning (TL)

The ability to diagnose GI diseases across many classes can be greatly improved by using transfer learning models such as DenseNet201 (Zhao et al., 2021), EfficientNetB6 (Tan & Le, 2019), InceptionResNetV2 (Bhatia et al., 2019), MobileNetV2 (Sandler et al., 2018), ResNet152V2

(He et al., 2016), VGG16 (Simonyan & Zisserman, 2014), and Xception (Chollet, 2017). These models extract large numbers of features from images due to their extensive pre-training on large datasets. Fine-tuning them on limited data for a specific task enables the effective capture of intricate patterns and subtle details associated with GI diseases. The pre-trained models were trained using more than 14 million classifications from 1,000 categories (ImageNet). We integrated the training of TL models with the PCC to reduce unnecessary features and the EELM classifier to attain accurate classification outcomes and compared the PD-CNN model to TL approaches in terms of classification results and computational resources, as there is no previous research on this dataset. This comparison encompasses performance metrics, model parameters, layer, sizes, and the duration of testing. After initializing the TL models, two FC layers were added with 1024 and 200 nodes each to improve the detection of GI diseases. PCC was employed simultaneously to reduce 200 features to only 39. Fig. 7 shows a comprehensive illustration of the TL models with the PCC and EELM classifier.

DenseNet is a CNN architecture that utilizes dense connectivity, allowing each layer to receive input from all preceding layers (Zhao et al., 2021). This promotes effective information transmission and improves overall performance. The variants of DenseNet, including DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264, differ in the number of layers they contain, with DenseNet-121 having 121 layers and DenseNet-201 having 201 layers. The EfficientNetB6 architecture employs compound scaling, which involves scaling the network's depth, width, and resolution (Tan & Le, 2019). This model is trained using the ImageNet dataset and comprises 87 million parameters. It has been used for several kinds of TL applications, such as semantic segmentation, object detection, and image categorization. The InceptionResNetV2 (Bhatia et al., 2019) model combines the Inception and ResNet architectures, utilizing inception modules and residual connections to extract features effectively. MobileNetV2 (Sandler et al., 2018) is another CNN architecture introduced by Sandler et al., which is based on an inverted residual structure. It employs lightweight depth-
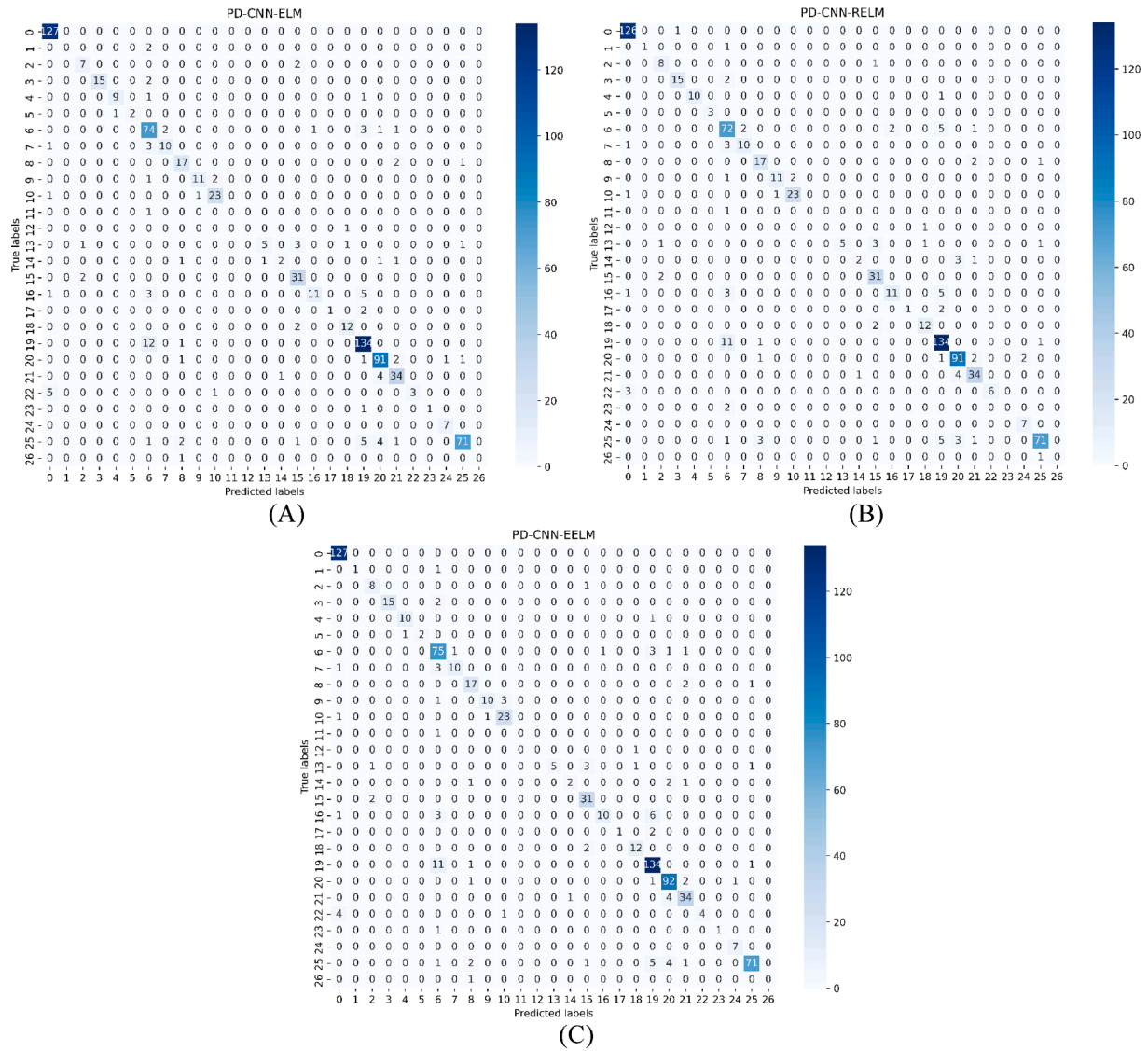
Fig. 8. Confusion metrics of (A) PD-CNN-ELM, (B) PD-CNN-RELM, and (C) PD-CNN-EELM models for GI tract disease classification.

Fig. 9. Performance of the PD-CNN model without PCC based on (A) ROC-AUC and (A) AUC-PR.

wise convolutions and bottleneck layers to achieve better performance while being computationally efficient, so this architecture is particularly efficient for mobile devices. In 2017, He et al. (He et al., 2016) presented

ResNet152V2, an NN model that utilizes residual learning by incorporating shortcut connections across layers to enhance learning efficiency. ResNet152V2 contains 60 million parameters. The Visual Geometry

Table 7
Class-wise performance using PD-CNN-PCC on test set.

GI Disease classes	Precision			Recall			F1-score			Accuracy (%)			ROC-AUC (%)		
	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM
0	**0.95**	0.94	0.94	0.99	0.99	**0.99**	0.97	0.97	**0.97**	87.62	87.25	**87.75**	**98.91**	98.43	98.89
1	0	**1**	0	0	**0.5**	0	0	**0.67**	0						
2	0.73	0.67	**0.73**	0.89	0.89	**0.89**	0.8	0.76	**0.8**						
3	0.88	0.83	**0.94**	0.88	0.88	**0.88**	0.88	0.86	**0.91**						
4	1	0.91	**1**	0.91	0.91	**0.91**	0.95	0.91	**0.95**						
5	1	0.67	**1**	0.67	0.67	**1**	0.8	0.67	**1**						
6	0.76	**0.76**	0.75	0.9	0.87	**0.9**	**0.83**	0.81	0.82						
7	**0.91**	0.83	0.83	0.71	0.71	**0.71**	0.8	0.77	0.77						
8	0.74	0.74	**0.74**	0.85	0.85	**0.85**	0.79	0.79	**0.79**						
9	1	1	**1**	0.71	0.71	**0.71**	0.83	0.83	**0.83**						
10	0.89	0.89	**0.89**	0.96	0.96	**0.96**	0.92	0.92	**0.92**						
11	0	0	0	0	0	0	0	0	0						
12	0	0	0	0	0	0	0	0	0						
13	0.83	**1**	0.83	0.45	0.45	**0.45**	0.59	**0.62**	0.59						
14	0.67	0.67	**0.67**	0.33	0.33	**0.33**	0.44	0.44	**0.44**						
15	0.84	0.84	**0.84**	0.94	0.94	**0.94**	0.89	0.89	**0.89**						
16	**0.92**	0.85	0.85	0.55	0.55	**0.55**	**0.69**	0.67	0.67						
17	1	1	**1**	0.33	0.33	**0.33**	0.5	0.5	**0.5**						
18	0.87	0.87	**0.87**	0.93	0.93	**0.93**	0.9	0.9	**0.9**						
19	0.87	0.89	**0.9**	0.93	0.93	**0.93**	0.9	0.91	**0.91**						
20	0.91	0.91	**0.91**	0.93	0.94	**0.94**	0.92	0.92	**0.92**						
21	0.83	0.81	**0.83**	0.87	0.87	**0.87**	0.85	0.84	**0.85**						
22	1	1	**1**	0.56	**0.56**	0.44	0.71	**0.71**	0.62						
23	1	0	**1**	0.5	0	**0.5**	0.67	0	**0.67**						
24	0.88	0.88	**0.88**	1	1	1	0.93	0.93	**0.93**						
25	0.95	**0.96**	0.95	**0.82**	0.81	0.81	0.88	**0.88**	0.87						
26	0	0	0	0	0	0	0	0	0						
Average (μ)	87.59	87.31	**88.12**	87.62	87.25	**87.75**	87.04	86.54	**87.12**						
± SD (σ)	±	±	±	±	±	±	±	±	±						
(%)	0.334	0.329	**0.332**	0.339	0.3387	**0.348**	0.318	0.314	**0.324**						

*Bold values indicate the best results.

Group (VGG) is characterized by multiple CLs and filters (Simonyan & Zisserman, 2014). After each CL, feature extraction is enhanced with a Max Pooling (MP) layer and a Rectified Linear Unit (ReLU) function. Google developed Xception, an architecture based on the Inception framework, in 2016 (Chollet, 2017). The system employs pointwise convolutions and DSCs to filter each channel of the input feature map separately. This approach preserves precision while significantly decreasing memory consumption and processing requirements. Xception is commonly utilized for various computer vision tasks due to its high efficiency, especially in situations with limited computational resources.

4.4. Feature selection with PCC

In the current era of ML, where data are paramount, it is critical to emphasize the importance of identifying pertinent features. Pattern recognition systems are based on features, measured parts of things that help identify patterns. However, of the many characteristics that may be provided, only a specific subset is significantly relevant to the final output. The extensive feature space of ML methods causes many problems, such as slow learning and complicated computations. Therefore, finding the best group of features, which can be achieved by carefully choosing which features to use, becomes important for overcoming these problems. The Pearson Correlation Coefficient (PCC)-based method stands out among the many feature selection procedures as a potential way to isolate essential characteristics from various possibilities (Benesty et al., 2009). Using the PCC, this method aims to reduce complexity and improve efficiency and processing speed by selecting the most relevant feature subset from those recovered by CNNs. By computing correlation values across all features, finding pairs with correlations more robust than certain limits is easier. This reduces the number of features that are not needed and improves the feature space. Additionally, the correlation coefficient is calculated by dividing the product of the standard deviations of two variables by the covariance

between them. This ensures that its output remains within the interval of −1 to 1. Standardization approaches, such as the ordinary score equation for a sample, focus on data pretreatment, which is essential for achieving effective machine learning outcomes. This thorough procedure of selecting features and standardizing them emphasizes the crucial significance of precise data pretreatment techniques in fully harnessing the capabilities of machine learning algorithms. Algorithm 2 presents the working steps of the PCC.

Algorithm 2: Feature selection utilizing PCC

1. **BEGIN**
2. Define data,$X = [x_1, x_2, ..., x_n], and Y = [y_1, y_2, ..., y_n]$
3. CorrMat = features.corr()
a) Calculate the mean of each dataset:
$$\mu_X = \left(\frac{1}{n}\right) * \sum_i^n x_i, \text{ and } \mu_y = \left(\frac{1}{n}\right) * \sum_i^n y_i$$
b) Calculate the standardized values for each data point:
$$z_i = x_i - \mu X w_i = y_i - \mu_Y$$
c) Calculate the covariance:$cov_{X,Y} = \left(\frac{1}{n}\right) * \sum_i^n z_i * w_i$
d) Calculate the standard deviation:
$$\sigma_X = \sqrt{\left[\left(\frac{1}{n}\right) * \sum_i^n (z_i)^2\right]}, \text{ and } \sigma_X = \sqrt{\left[\left(\frac{1}{n}\right) * \sum_i^n (w_i)^2\right]}$$
e) Calculate PCC, corr():$PCC = \frac{cov_{X,Y}}{(\sigma_X * \sigma_Y)}$
4. for i ≤ CorrMat.col:
5.   for j ≤ i:
6.   If CorrMat.iloc[i, j] > threshold:
7.   ColName = CorrMat.col[i]
8.   CorrCol.add(ColName)
9. Dropped.features(CorrCol)
10. **END**

4.5. Explainable artificial intelligence (XAI)

In the realm of DL, XAI refers to understanding and clarifying the decision-making process of a deep neural network (Tjoa & Guan, 2020).

Fig. 10. Confusion metrics of (A) PD-CNN-PCC-ELM, (B) PD-CNN-PCC-RELM, and (C) PD-CNN-PCC-EELM models for GI tract disease classification.
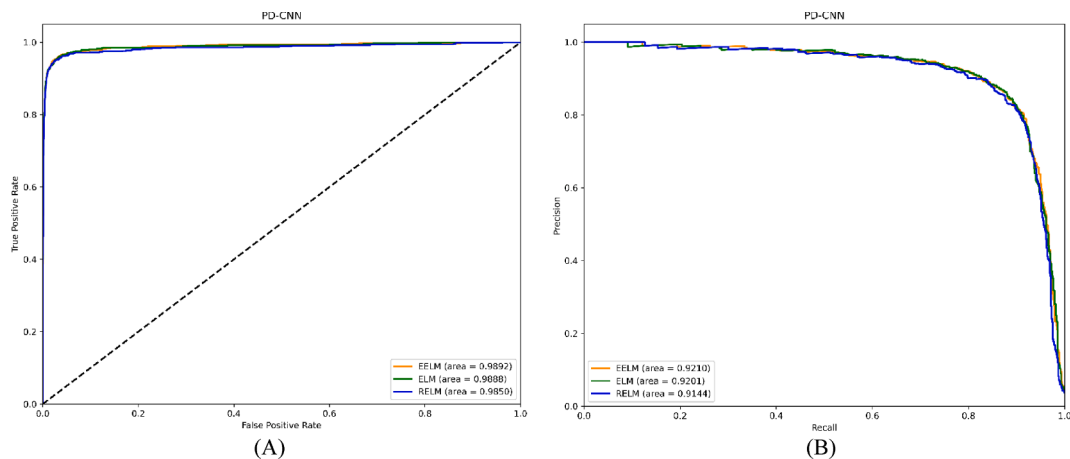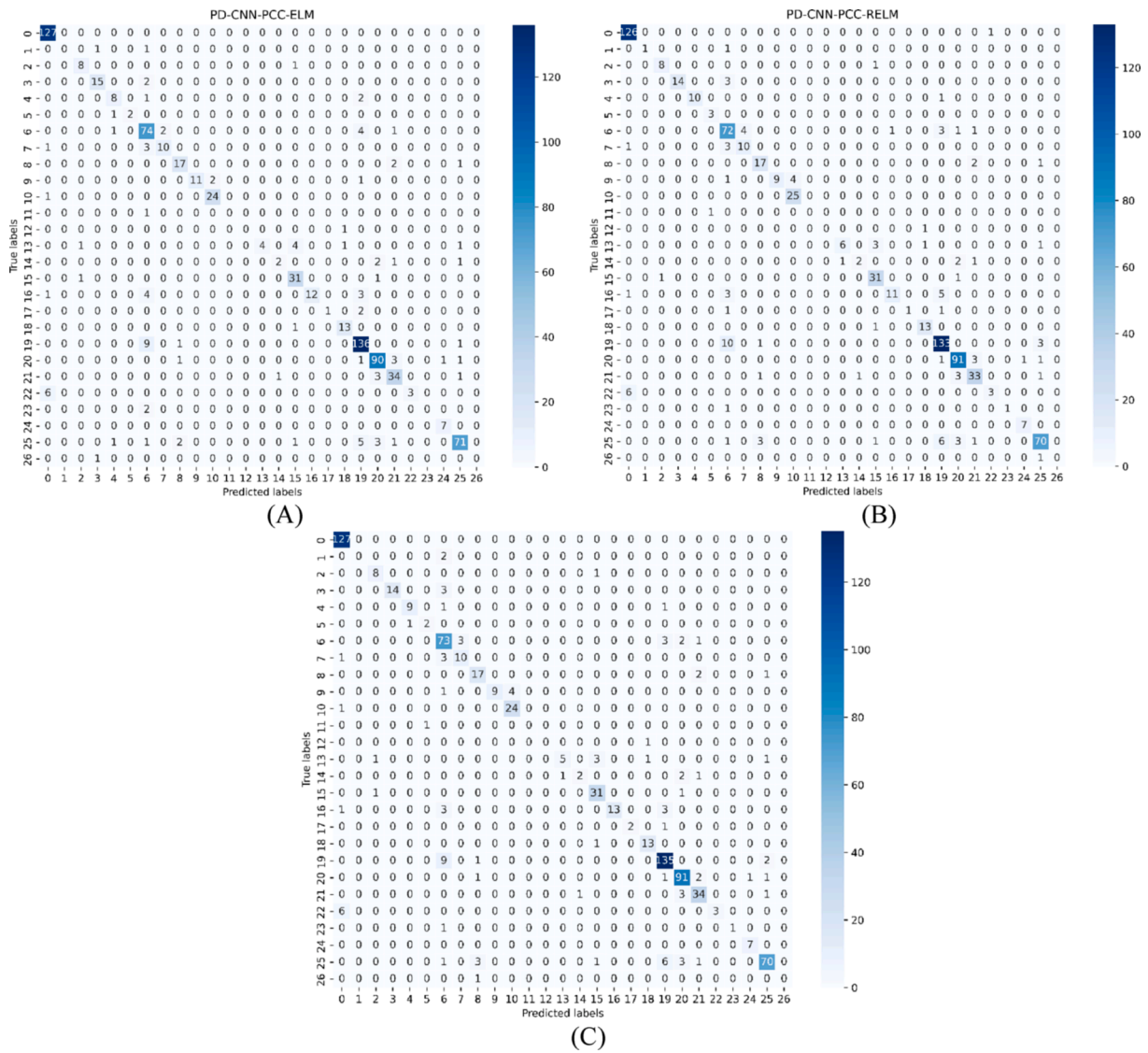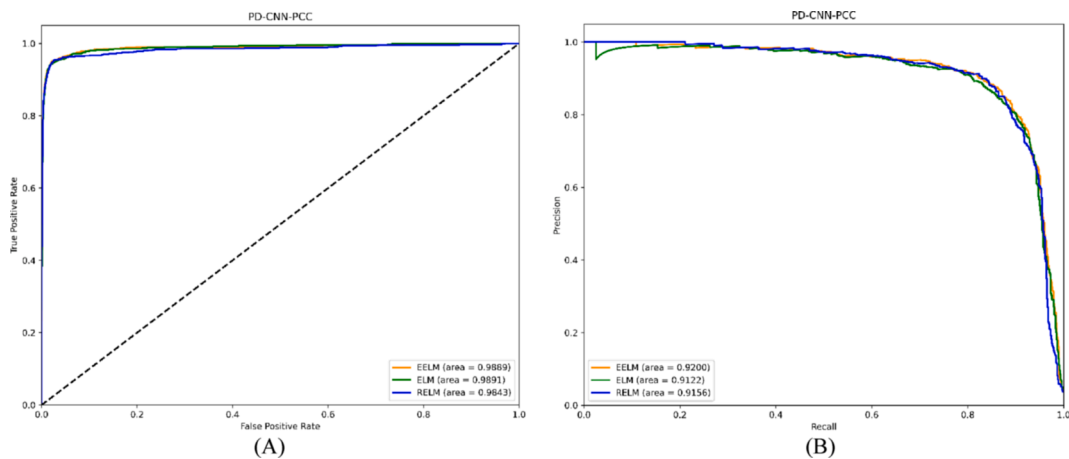

Fig. 11. Performance of the PD-CNN model with respect to the PCC based on (A) ROC-AUC and (A) AUC-PR.

Table 8

Twenty-seven class classification performances of the TL models and proposed PD-CNN-PCC model.

Models	Precision			Recall			F1-score			Accuracy (%)			Testing Time (Seconds)		
	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM
DenseNet201-PCC	0.831	0.808	0.829	0.825	0.815	0.825	0.805	0.792	0.804	82.5	81.5	82.5	0.00008	0.00007	0.00008
EfficientNetB6-PCC	0.121	0.0896	0.11	0.208	0.207	0.207	0.128	0.123	0.1259	20.87	20.75	20.75	0.00007	0.00006	0.00007
InceptionResNetV2-PCC	0.71	0.76	0.753	0.738	0.751	0.761	0.709	0.725	0.734	73.8	75.1	76.1	0.00009	0.00009	0.00009
MobileNetV2-PCC	0.727	0.758	0.751	0.752	0.772	0.778	0.717	0.738	0.745	75.25	77.25	77.87	0.00009	0.00009	0.00009
ResNet152V2-PCC	0.834	0.827	0.832	0.837	0.831	0.838	0.823	0.823	0.826	83.75	83.12	83.87	0.00007	0.00007	0.00007
VGG16-PCC	0.81	0.812	0.838	0.832	0.827	0.848	0.81	0.81	0.828	83.25	82.75	84.87	0.00005	0.00032	0.0141
Xception-PCC	0.669	0.674	0.654	0.686	0.72	0.715	0.648	0.677	0.672	68.62	72	71.5	0.00009	0.00008	0.00023
PR-CNN	0.828	0.831	0.831	0.818	0.826	0.821	0.807	0.813	0.809	0.818	0.826	82.1	0.00078	0.00094	0.00143
PR-CNN- PCC	0.826	0.812	0.817	0.83	0.81	0.821	0.819	0.797	0.811	83	81	82.12	0.00006	0.00008	0.00007
PD-CNN	0.8759	0.8731	0.8811	0.8762	0.8725	0.8775	0.87	0.8654	0.8712	87.62	87.25	87.75	0.01562	0.0156	0.00001
PD-CNN-PCC	0.8759	0.8731	**0.8812**	0.8762	0.8725	**0.8775**	0.8704	0.8654	**0.8712**	87.62	87.25	87.75	0.00006	0.00004	0.000001

*Bold values indicate the best results.

This is essential due to the model's complexity and difficulty in understanding. XAI was used in this study to diagnose GI illnesses across 27 different categories to verify its accuracy. SHAP, heatmap, guided heatmap, Grad-CAM, guided Grad-CAM, and saliency mapping were employed to address the "black box" character of the DL models, which can hinder their usefulness. The goal was to enhance the transparency and interpretability of the proposed PD-CNN model by utilizing XAI. By combining the PD-CNN model with XAI for disease classification, endoscopists can make more accurate and confident decisions when diagnosing diseases more efficiently. This approach will help healthcare professionals confirm the model's predictions, identify errors, and reduce biases and missing data by providing accurate diagnoses. This advancement creates new opportunities for better disease management techniques and more efficient therapies for GI issues.

4.5.1. Shapley Additive explanations (SHAP)

This study utilized Shapley values to determine the importance of individual pixels, which exhibited a clear pattern. Red pixels enhance accurate class recognition, but blue pixels hinder it by reducing the likelihood of successful categorization (Bhandari et al., 2022). The Shapley values were computed using Eq. (1).

$$\varnothing_k = \sum_{M \subseteq N \setminus k} \frac{M!(A - |M| - 1)!}{A!} [f_x(M \cup k) - f_x(M)] \qquad (2)$$

f_x indicates the impact on the output resulting from the Shapley values of a specific feature, k. The subset M includes all features coming from feature N, excluding feature k. $\frac{M!(A-|M|-1)!}{A!}$ represents the weighted factor of the subset M permutations. Eq. (2) gives the predicted result, denoted by the sign $f_x(M)$.

$$f_x(M) = P[f(x)|x_M] \qquad (3)$$

The SHAP method involves the substitution of every initial identifiable (x_k) with a binary value (b'_k) that denotes the presence or absence of x_k, as illustrated in Eq. (3).

$$l(b') = \varnothing_0 + \sum_{k=1}^{A} \varnothing_k b'_k \qquad (4)$$

The contribution of the feature is represented by $\varnothing_k b'_k$ in the proposed framework $f(x)$, where $l(b')$ is the substitute model for the framework. The bias is indicated by $\varnothing_0$. A crucial component that helps comprehend the fundamental workings of the model is the contribution of feature k to the result and the function of $\varnothing_k$.

4.5.2. Heatmap visualization

A heatmap was generated to display the regions of the original image that had the greatest impact on the ultimate classification outcome. This heatmap is generated by computing the gradient of the last convolutional layer's output class score with respect to the feature maps (Jin et al., 2023).

$$Heatmap(x) = \Sigma_i \frac{\partial Score_c}{\partial Feature_i} \qquad (5)$$

Here, $Heatmap(x)$ represents the heatmap for a given input image x, $Score_c$ represents the class score, and $Feature_i$ corresponds to the ith feature map.

4.5.3. Guided heatmap visualization

The heatmap from the previous phase was refined using guided visualization techniques. Gradients are calculated via guided backpropagation and then scaled by the ReLU activation of the corresponding feature map (Jin et al., 2023).
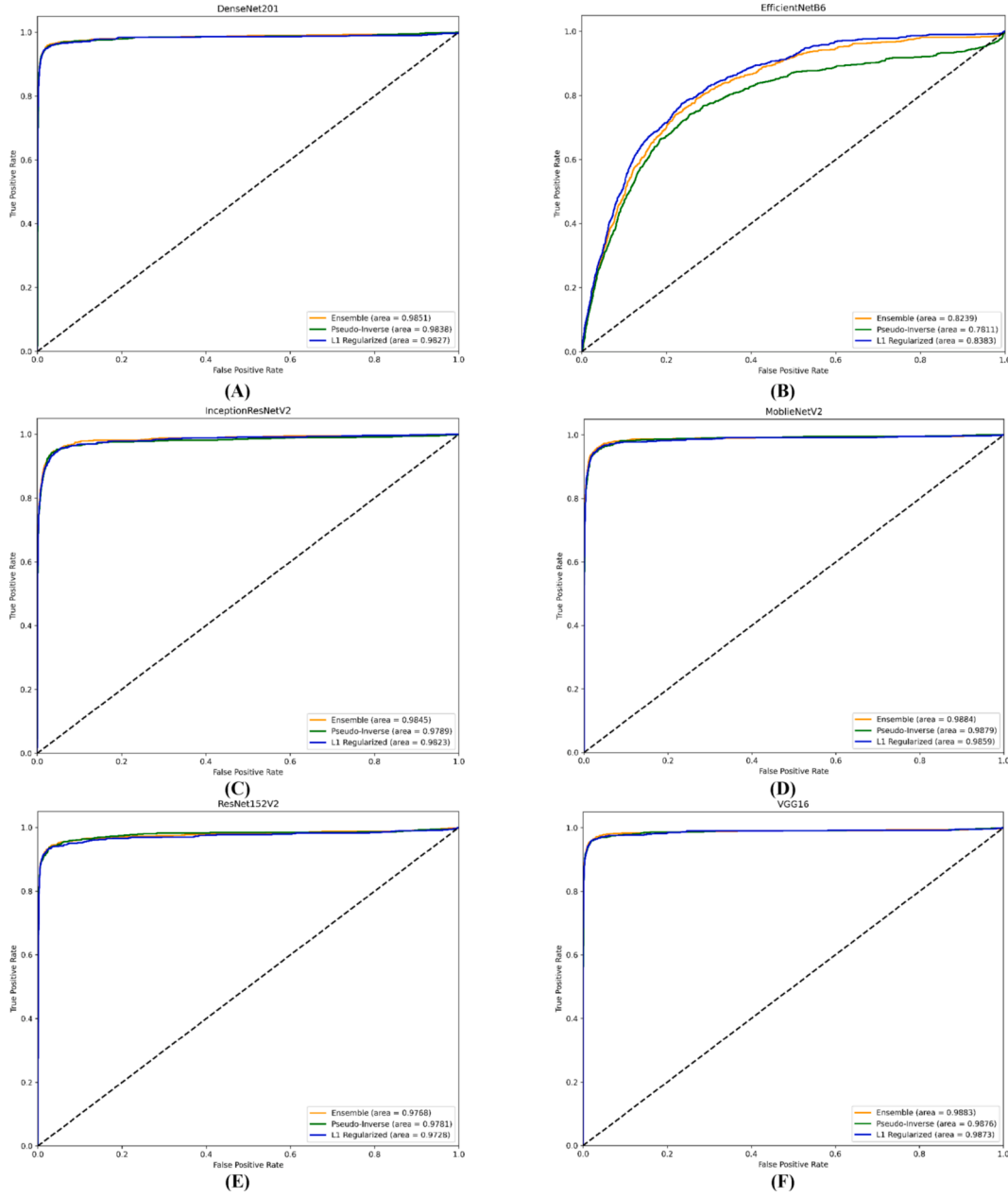
Fig. 12. ROC-AUC curves for (A) DensNet201, (B) EfficientNetB6, (C) InceptionResNetV2, (D) MobileNetV2, (E) ResNet152V2, (F) VGG16, and (G) Xception with PCC and ELM (Pseudo-Inverse), RELM (L1-Regularized), EELM (Ensemble) classifier on test-set.

4.5.4. Gradient-weighted class Activation mapping (Grad-CAM)

Grad-CAM merges class discrimination with location. It creates a heatmap after computing the weights for each feature map based on the gradient of the class score compared to the feature maps (Selvaraju et al., 2017).

$$\alpha_k^c = \frac{1}{z}\Sigma_i\Sigma_j\frac{\partial y^C}{\partial A_{ij}^k} \tag{6}$$

$$L_{GradCAM}^c = ReLU(\Sigma_k\alpha_k^c.A^k) \tag{7}$$

Here, y^C indicate score for class c with respect to feature map A^k, α_k^c indicates calculated weight for every neuron, $\frac{1}{z}\Sigma_i\Sigma_j$ defines a global average pooling over the width (i) and height (j).

4.5.5. Guided Grad-CAM

Guided Grad-CAM visualization is an interpretability technique that merges guided backpropagation and Grad-CAM principles. This technique enhances the heatmap produced by Grad-CAM by including guided backpropagation gradients and highlighting the most significant features in the ultimate classification determination (Chen et al., 2020).

$$GuidedGrad-CAM^c(x,y) = ReLU(\Sigma_k\alpha_c^k.A^k(x,y)).^*G^c(x,y) \tag{8}$$

Here, c indicates the class of interest, (x,y) is the co-ordinates of a pixel in the input image, α_c^k indicates important weights for each feature map, A^k is the activation of feature map k at pixel (x,y), and $G^c(x,y)$ indicates the refined gradient map obtained through guided backpropagation for class c.
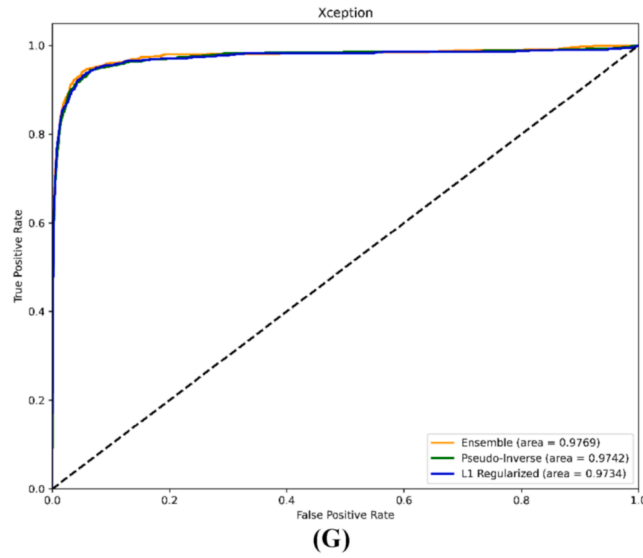
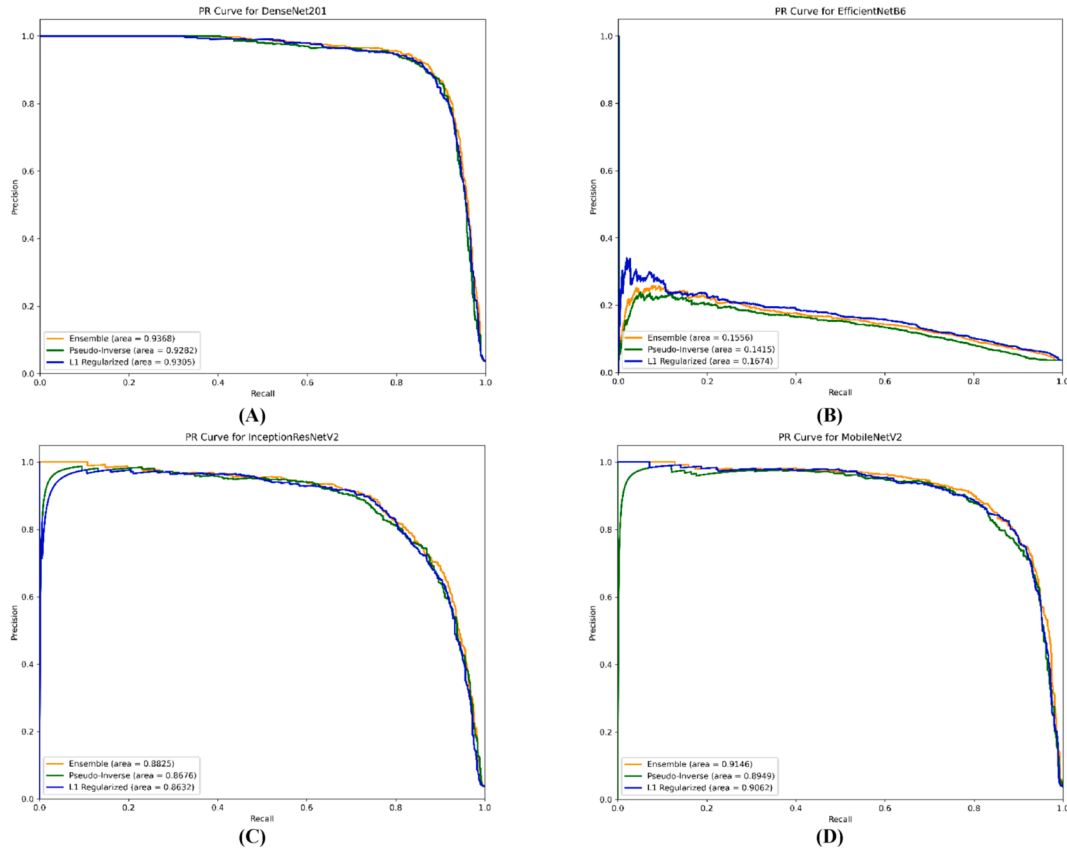**Fig. 12.** (*continued*).

Fig. 13. AUC-PR curves for (A) DensNet201, (B) EfficientNetB6, (C) InceptionResNetV2, (D) MobileNetV2, (E) ResNet152V2, (F) VGG16, and (G) Xception with PCC and ELM (Pseudo-Inverse), RELM (L1-Regularized), EELM (Ensemble) classifier on test-set.

4.5.6. Guided saliency mapping

Saliency mapping assesses the spatial support of a class. It facilitates interpretability in neural networks by presenting an image that emphasizes the region of interest. The saliency map is generated using backpropagation. This method enhances comprehension of the model's judgments by locating pixels that have minimal influence on the score and calculating the derivative of the class score regarding the image (Yang & Berdine, 2023).

First, the distance of each pixel to the remaining pixels in the same frame is calculated:

$$SALS(I_k) = \sum_{i=1}^{N} |I_k - I_i| \tag{9}$$

I_i is the value of pixel *i*. The following equation is an expanded form of Eq. (9)

$$SALS(I_k) = |I_k - I_1| + |I_k - I_2| + \cdots + |I_k - I_N| \tag{10}$$
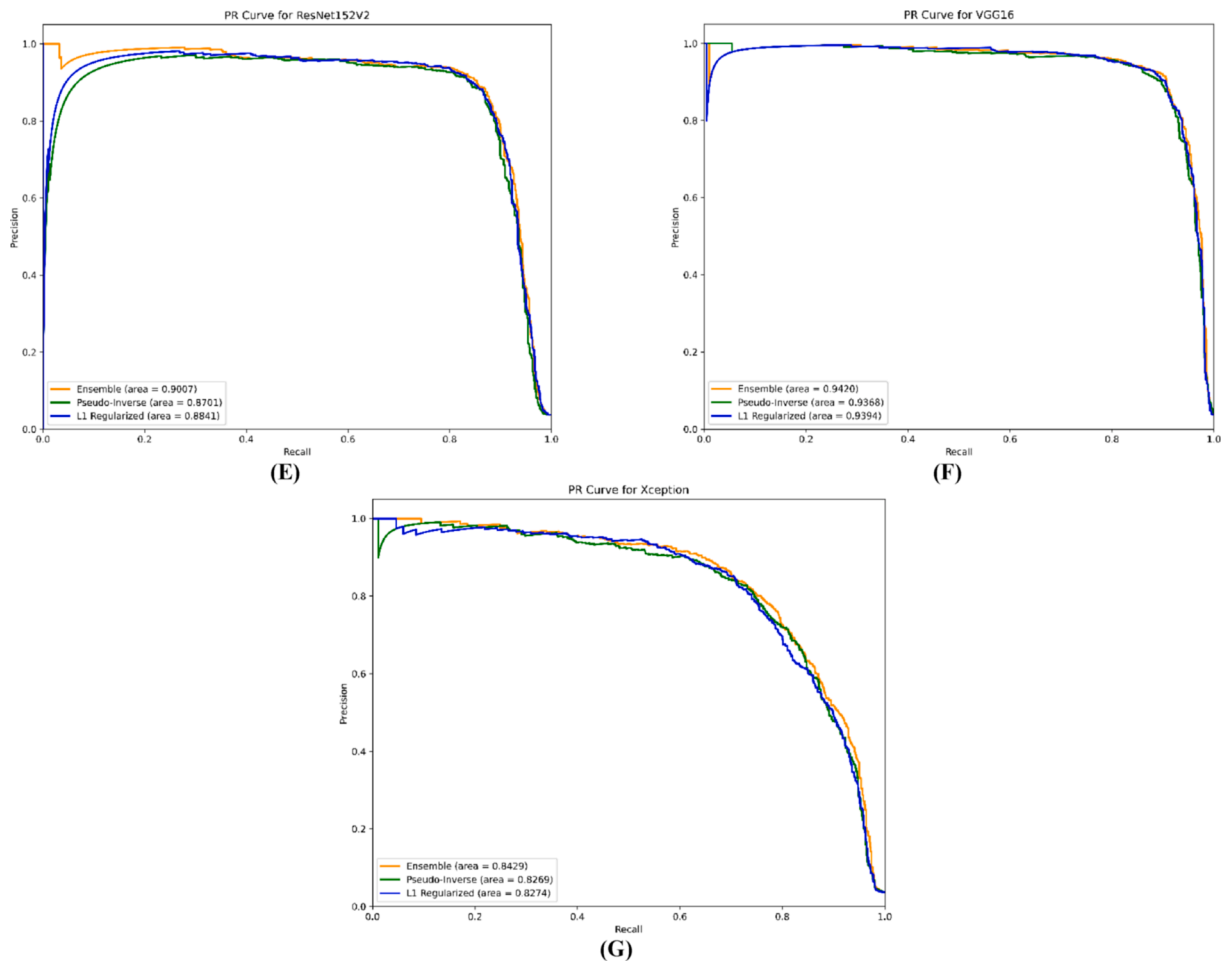
(E)

(F)

(G)

Fig. 13. (*continued*).

where N represents all the pixels in the present frame. After that, formula 8 is refined. The values that have the same I are combined.

$$SALS(I_k) = \Sigma F_n \times |I_k - I_N| \tag{11}$$

where F_n is the frequency of I_N. The value of n is in the range [0, 255].

4.6. Hyperparameter settings and classification matrices

The optimal hyperparameters were selected based on a trial-and-error approach during the experimental work. Once the best parameters were identified, both the proposed and transfer learning models were trained under similar conditions. Table 4 provides a detailed overview of the training hyperparameters.

To assess the classification effectiveness of the proposed model, various metrics, such as accuracy, precision, recall, F1-score, and area under the curve (AUC), were computed (Powers, 2020). The cross-entropy formula evaluates the correspondence between the integer-based actual class label and the probability distribution generated by the model. It calculates the difference between the real and predicted labels to reduce cross-entropy loss to the greatest extent possible. The sparse categorical cross-entropy loss is commonly utilized in deep learning scenarios such as image classification, particularly when dealing with numerous classes (Chaithanya et al., 2021).

5. Results and discussions

This section presents a quantitative performance analysis of the proposed PD-CNN-PCC-EELM framework, evaluated using the

GastroVision test set. Prior to this, the results of an ablation study involving the following baseline models—Pr-CNN, Pr-CNN-PCC, PD-CNN, and PD-CNN-PCC—each paired with ELM, RELM, and EELM classifiers—have been presented, followed by classwise performance metrics for the PD-CNN and PD-CNN-PCC models. In addition, the performances of the SOTA-TL models were compared against those of the best-performing frameworks in terms of the classification accuracy, number of parameters, number of layers, model size, and computational cost. To determine the optimal PCC threshold, a comprehensive analysis. For qualitative assessment of model interpretability, XAI visualizations are presented. Furthermore, the versatility of the model's performance was validated using K-fold cross-validation, and a comparative performance analysis was performed using the KvasirV2 dataset.

5.1. Ablation study

Table 5 presents the results of an ablation study comparing the proposed PD-CNN-PCC-EELM model with several baseline models, focusing on the impact of the PCC for feature selection, the Parallel CNN (Pr-CNN) as a feature extractor and the ELM, RELM, and EELM as classifiers.

In terms of Precision, the proposed PD-CNN-PCC-EELM model achieved the highest value of 88.12 ± 0.332. This represents a notable improvement of 6.03 % over the Pr-CNN-ELM model (82.8 ± 0.476) and a 0.601 % improvement over the PD-CNN-ELM model (87.59 ± 0.333). For recall, the PD-CNN-PCC-EELM model also achieved a score of 87.75 ± 0.348, indicating a 6.78 % enhancement compared to that of the Pr-CNN-ELM model (81.8 ± 0.321) and a slight improvement over that

Table 9
Results of the proposed PD-CNN-PCC-Extreme Learning Machine (Ensemble-ELM with RELM) model with different PCC values.

| PCC Threshold	Features on PD-CNN		Accuracy			Precision			Recall			F1-score			Testing Time (Seconds)		
	W/O PCC	With PCC	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM
0.79	200	36	0.85	0.8487	0.8569	0.8558	0.8552	0.8614	0.85	0.8487	0.8562	0.8459	0.8436	0.8509	0.0157	0.0157	0.00005
0.78	200	**39**	0.8762	0.8725	**0.8775**	0.8759	0.8731	**0.8812**	0.8762	0.8725	**0.8775**	0.8704	0.8654	**0.8712**	0.00006	0.00004	**0.000001**
0.77	200	46	0.8537	0.8537	0.8531	0.856	0.8578	0.8568	0.8537	0.8537	0.8537	0.8593	0.8489	0.8479	0.00007	0.00005	0.00006
0.76	200	53	0.8412	0.855	0.8609	0.8477	0.8585	0.8654	0.8512	0.855	0.8612	0.8356	0.8505	0.8569	0.00009	0.00009	0.00009
0.75	200	60	0.85	0.8562	0.8538	0.8572	0.8601	0.8632	0.85	0.8562	0.8575	0.8453	0.8518	0.8536	0.0156	0.00005	0.007825
0.74	200	72	0.845	0.8475	0.8499	0.8497	0.8539	0.842	0.845	0.8475	0.8475	0.839	0.8423	0.8417	0.0008	0.0006	0.0007

*Bold values indicate the best results.
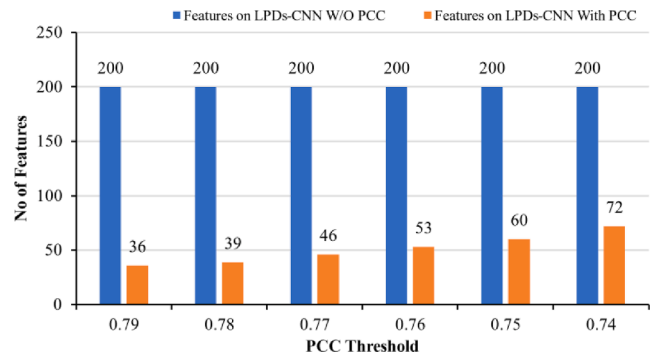
Fig. 14. Number of features with various threshold values.
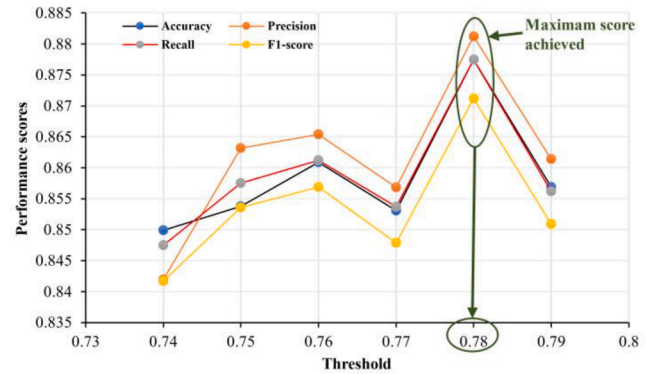

Fig. 15. Performance scores for various threshold values for the PD-CNN-PCC-EELM model.

of the PD-CNN-ELM model (87.62 ± 0.343). Similarly, the F1-score for the PD-CNN-PCC-EELM model was the highest at 87.12 ± 0.324, reflecting a 7.36 % improvement over that of the Pr-CNN-ELM model (80.7 ± 0.224) and a marginal increase over that of the PD-CNN-ELM model (87.0 ± 0.321).

The PD-CNN-PCC-EELM model achieved the highest accuracy at 87.75 %, marking a 6.78 % enhancement over that of the Pr-CNN-ELM model (81.8) and a slight improvement over that of the PD-CNN-ELM model (87.62). Additionally, the proposed model demonstrated an excellent ROC-AUC of 98.89, indicating superior classification performance. This represents a 1.02 % improvement over the Pr-CNN-ELM model (97.87) and is on par with the PD-CNN-ELM model (98.88). Notably, the PD-CNN-PCC-EELM model exhibited the fastest testing time at 0.000001 s, a significant improvement over both the Pr-CNN-ELM model (0.00078 s) and the PD-CNN-ELM model (0.01562 s), underscoring its efficiency.

The incorporation of the PCC feature extractor improved the overall performance in terms of precision, recall, F1-score, and accuracy compared to models without PCC. For instance, comparing the Pr-CNN-PCC-EELM to the Pr-CNN-EELM, the precision improved from 83.1 ± 0.314 to 81.7 ± 0.521, and the F1-score increased from 80.9 ± 0.315 to 81.1 ± 0.285, although the accuracy slightly decreased from 82.1 to 82.12. Similarly, comparing the PD-CNN-PCC-EELM to the PD-CNN-EELM, both models performed equally well in terms of the recall (87.75), with the precision slightly increasing from 88.11 ± 0.3 to 88.12 ± 0.332, and the F1-score remained consistent (87.12), while the ROC-AUC showed a minor decrease from 98.92 to 98.89.

In conclusion, the proposed PD-CNN-PCC-EELM model significantly outperformed the other models across all the key metrics, including the precision, recall, F1-score, accuracy, and testing time. The inclusion of the PCC enhances feature selection, contributing to superior performance metrics, particularly when combined with the PD-CNN architecture. These results confirmed the model's efficiency and effectiveness
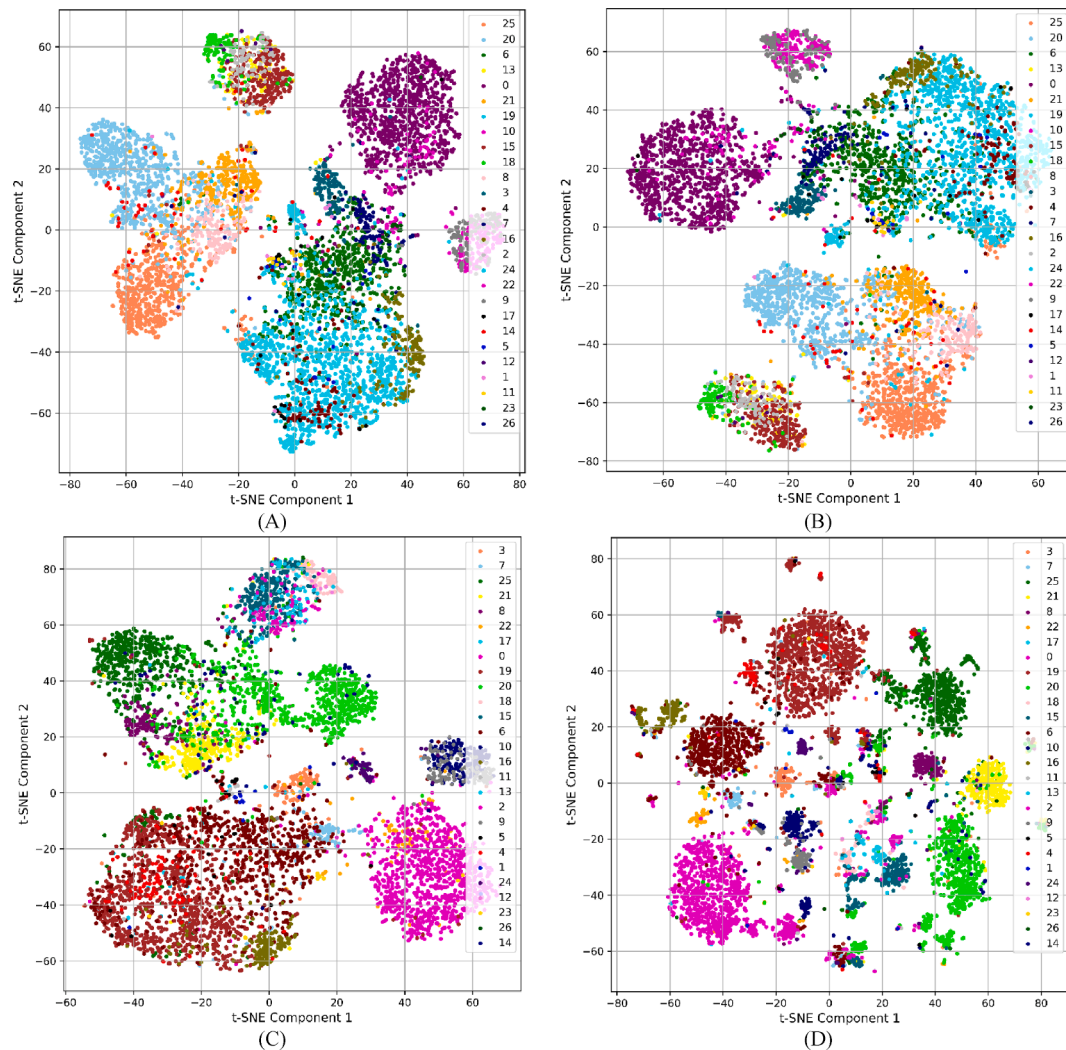
Fig. 16. T-SNE visualization in a 2D space of testing samples after training the models (A) PD-CNN, (B) PD-CNN-PCC, (C) Pr-CNN-PCC, (D) DensNet201-PCC, (E) EfficientNetB6-PCC, (F) InceptionResNetV2-PCC, (G) MobileNetV2-PCC, (H) ResNet152V2-PCC, (I) VGG16-PCC, and (J) Xception-PCC with EELM for twenty-seven test-set classification.

in processing and classifying data with high accuracy and speed, underscoring its robustness and suitability for practical applications.

5.2. Depth-wise Separable CNN without PCC

Table 6 presents the performance metrics, including precision, recall, F1-score, accuracy, and AUC, for various GI disease classes using the PD-CNN model without the PCC approach. Across the board, the EELM classifier consistently outperforms both the ELM and RELM classifiers within the PD-CNN framework. The EELM demonstrates superior performance across all disease classes. The confusion matrices among the ELM, RELM, and EELM classifiers are displayed in Fig. 8, providing essential insights into the classification outcomes. According to the analysis, EELM was identified as the most precise classifier. For instance, in disease class 7, the EELM achieved a precision of 0.91, a recall of 0.71, and an F1-score of 0.80, surpassing the performances of the ELM and RELM. Notably, in disease class 24, EELM exhibited competitive improvements in precision, recall, and F1-scores, with values of 0.88, 1.00, and 0.93, respectively, outscoring both ELM and RELM. For disease class 25, ELM and RELM exhibited competitive precision, recall, and F1-scores of 0.96, 0.84, and 0.89, respectively. Additionally, the model could not identify classes 11, 12, and 26 due to a lack of training data,

and only one sample was available for testing, as shown in Table 1.

In comparison to ELM (average precision of 87.59 %, recall of 87.62 %, and F1-score of 87 %), and RELM (average precision of 87.31 %, recall of 87.25 %, and F1-score of 87.12 %), the EELM demonstrated considerable improvements, with the highest average precision of 88.11 %, recall of 87.75 %, and F1-score of 87.12 %. Among the examined GI diseases, the EELM achieves an estimated 0.594 % improvement in precision, 0.15 % in recall, and 0.14 % in F1-score compared to the ELM. This significant improvement highlights how EELM effectively addresses the categorization issues presented by certain classes.

With respect to accuracy, the EELM achieved a competitive score of 87.75 %, which was 0.15 % greater than that of the ELM (87.62 %) and 0.57 % greater than that of the RELM (87.25 %). The reason for the lower EELM scores for some classes is that combining multiple base learners (ELM and RELM) improved the overall predictive performance (precision, recall, F1 score); however, in specific scenarios, insufficient data diversity among the base learners prevented the ensemble from harnessing distinct perspectives and patterns to greatly enhance the predictive accuracy. Another performance measure with the AUC showed that the EELM had a better area coverage of 98.92 % (Fig. 9A). This represents an increase of approximately 0.2 % in comparison to ELM (98.88 %) and RELM (98.50 %). Fig. 9B presents the AUC-PR
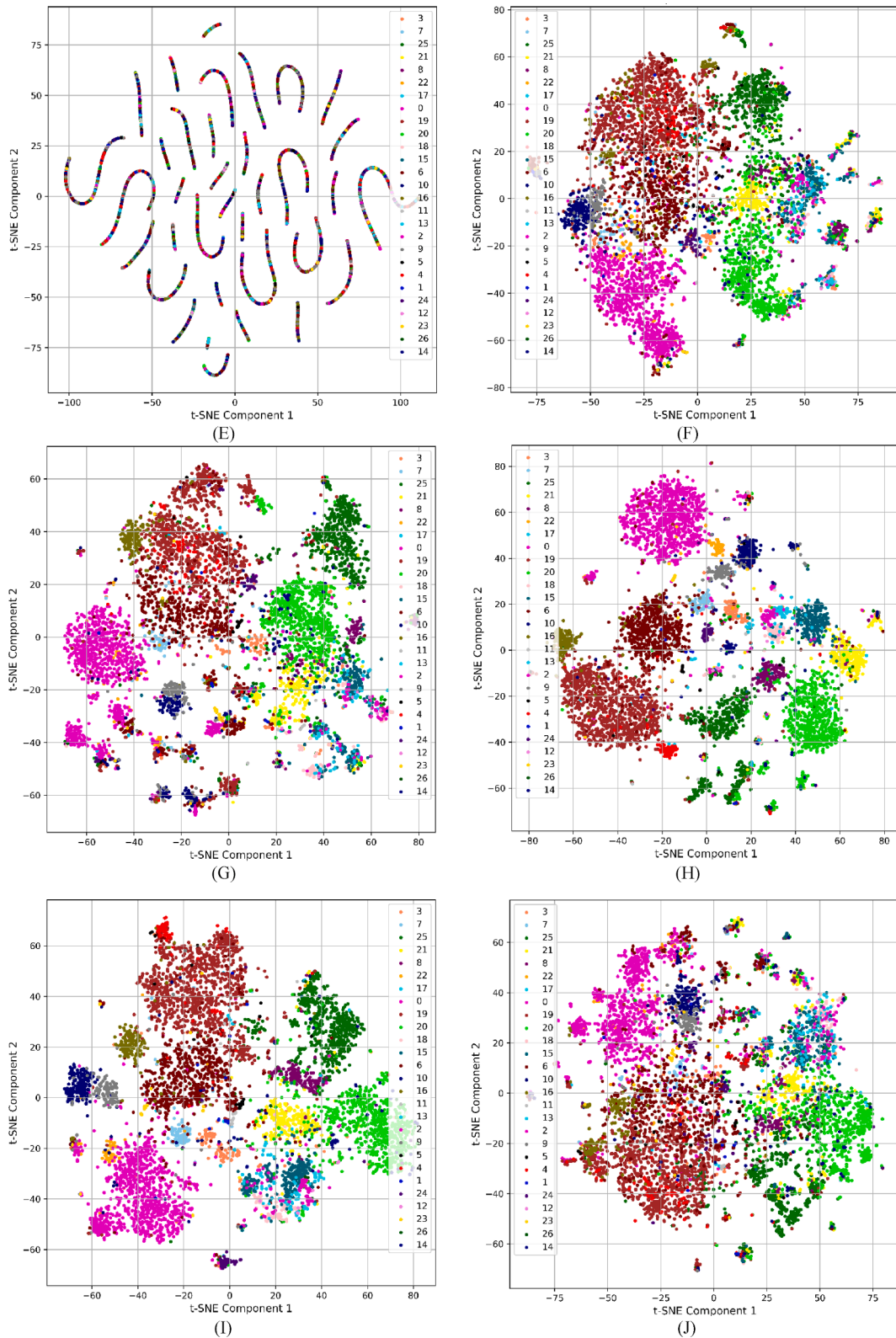
Fig. 16. (*continued*).

curves, where the EELM achieved a maximum area coverage of 92.10 %. It is evident from the performance study that using the EELM classifier improves average precision, recall, and F1-scores, underscoring the classifier's potential application in a variety of contexts, particularly in the fields of predictive modeling and medical diagnostics.

5.3. Depth-wise Separable CNN with PCC

The performance of the proposed feature extractor, PD-CNN, after incorporating the PCC with three different classifiers, ELM, RELM, and EELM, is shown in Table 7. Fig. 10 illustrates the confusion matrices among the EELM, RELM, and ELM classifiers. These matrices offer
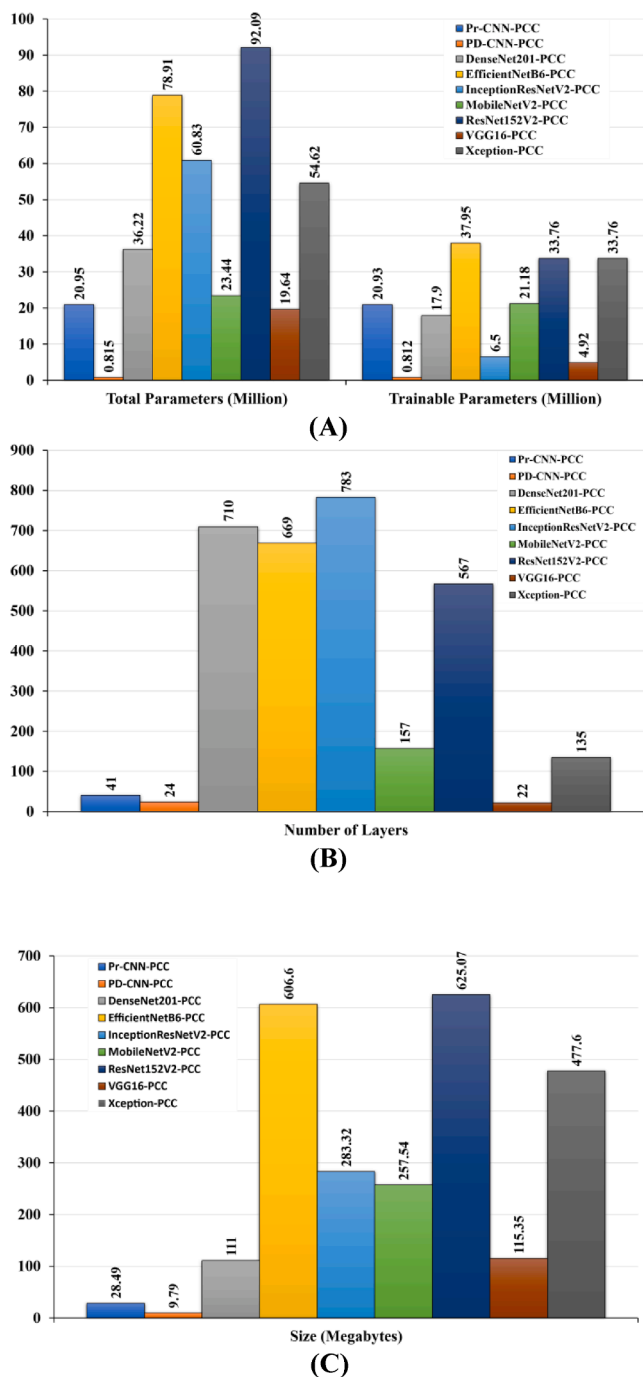
Fig. 17. Computational resource (A) parameters, (B) number of layers and (C) size comparisons among the proposed PD-CNN-EELM with TL models.

crucial insights into the outcomes of the classification process. The novel PD-CNN model extracted an excess of 200 features. Unnecessary and insignificant features were subsequently eliminated to reduce the predictive complexity of the classification. Upon completion of the feature extraction process, the PCC algorithm was implemented to eliminate 161 redundant features, resulting in the retention of only 39 of the most salient features. The classification was then performed using the PD-CNN-PCC-EELM framework and displayed by employing a test set.

The EELM classifier frequently outperforms the ELM and RELM classifiers in terms of the maximum number of evaluated metrics. The EELM consistently outperforms the other methods. In class 5, the EELM achieved perfect precision, recall, and F1 scores (1.00), but without the PCC, it performed inadequately (Tables 6 and 7).

The performance enhancement of EELM compared to that of ELM and RELM is substantial. The EELM obtained the highest average accuracies of 87.75 %, 0.15 % greater than the ELM's accuracy of 87.62 %, and 0.57 % greater than the RELM's accuracy of 87.25 %.

The EELM classifier accurately identifies and classifies GI disorders within the PD-CNN-PCC model, demonstrating its outstanding capabilities. Reducing superfluous features is a crucial benefit of the PCC, substantially improving the PD-CNN model's performance. Compared to the PD-CNN-EELM model, the PD-CNN-PCC-EELM model achieves competitive ROC-AUC (98.89 %) and AUC-PR (98.89 %) scores (Fig. 11).

In summary, the proposed configuration emerges as the optimal selection, amalgamating the robust feature extraction capabilities inherent to the PCC. This synergy yields exceptional performance, rendering PD-CNN-PCC-EELM the most efficacious solution for precise and dependable GI disease classification tasks.

5.4. Performance comparison with TL models

The PCC was incorporated with the TL models for comparative study, as shown in Table 8. Compared to regular CL, DSC layers performed better in the suggested architecture.

Among the models, the Pr-CNN, Pr-CNN-PCC, PD-CNN, and proposed PD-CNN-PCC models demonstrate substantial improvements. For every metric, the PD-CNN-PCC model outperformed the Pr-CNN-PCC model. Similarly, PD-CNN-PCC achieved a recall of 0.863, 4.87 % higher than that of Pr-CNN-PCC (0.821). An analysis of the F1-score revealed that it increased by 5.26 %, from 0.811 for the Pr-CNN-PCC to 0.856 for the PD-CNN-PCC. The computational time was seven times faster than that of the Pr-CNN-PCC technique.

The PD-CNN-PCC-EELM outperformed EfficientNetB6-PCC in terms of accuracy, with an impressive 0.866 score, which is well above 0.11 and represents an enormous improvement of 87.29 %. The recall metric also showed significant improvement, reaching 0.863, which was better than that of ResNet152V2, the second-best transfer model, which scored 0.838. The recall improved by an impressive 2.9 %. The F1-score, which stands at 0.856 and surpasses the transfer models' range of 0.672 to 0.828, further demonstrates the superiority of the PD-CNN-PCC model. The model performance increased by 21.5 % with Xception and 85.3 % with EfficientNetB6. Overall, the accuracy of the PD-CNN-PCC model exceeded that of the Xception and VGG16 models, reaching an impressive score of 86.13 %. There is a noticeable improvement of 1.66 % compared to the best performing VGG16. Compared to traditional TL approaches, the PD-CNN approach—which involves integrated PCC thresholding—has produced significant detection capabilities ranging from 2.9 to 87.29 % across major standards. This highlights the critical need for domain-specific model design and training.

The PD-CNN-PCC-EELM model was 1,410 times faster than the VGG16-PCC model, which took 0.0141 s to test. It is 120 times faster than Pr-CNN, 14 times faster than PD-CNN, and 7–8 times faster than other TL models, with processing durations ranging from 0.00007 to 0.00009 s. The substantial decrease in testing time demonstrates the efficiency advantages of the PD-CNN-PCC method. The computational load is reduced while testing by condensing the CNN model into a compact feature vector before implementing PCC thresholding. When real-time or low-latency forecasts are crucial, producing findings in 10 microseconds instead of multiple milliseconds can be very beneficial. The PD-CNN-PCC model demonstrates superior testing efficiency compared to the conventional deep TL methods.

The ROC-AUCs of DensNet201, EfficientNetB6, InceptionResNetV2, MobileNetV2, ResNet152V2, VGG16, and Xception with the PCC-EELM were 98.51 %, 82.39 %, 98.45 %, 98.84 %, 97.68 %, 98.83 %, and 97.69 %, respectively (Fig. 12). Furthermore, the AUC-PRs of these models reached 93.68 %, 15.56 %, 88.25 %, 91.46 %, 90.07 %, 94.20 %, and 84.29 %, respectively (Fig. 13).

Table 10

Comparative resource analysis among the trained models on test set.

Performance Criteria	Pr-CNN-PCC	PD-CNN-PCC	DenseNet201-PCC	EfficientNetB6-PCC	Inception ResNetV2-PCC	MobileNet V2-PCC	ResNet 152 V2-PCC	VGG16-PCC	Xception-PCC
Total Parameters (Million)	2.095	**0.815**	36.22	78.91	60.83	23.44	92.09	19.64	54.62
Trainable Parameters (Million)	2.093	**0.812**	17.9	37.95	6.5	21.18	33.76	4.92	33.76
Number of Layers	41	**24**	710	669	783	157	567	22	135
Size (Megabytes)	28.49	**9.79**	111	606.6	283.32	257.54	625.07	115.35	477.6
Testing Time (Seconds)-EELM	0.00005	**0.000001**	0.000098	0.000099	0.000085	0.000076	0.000086	0.0141	0.000069

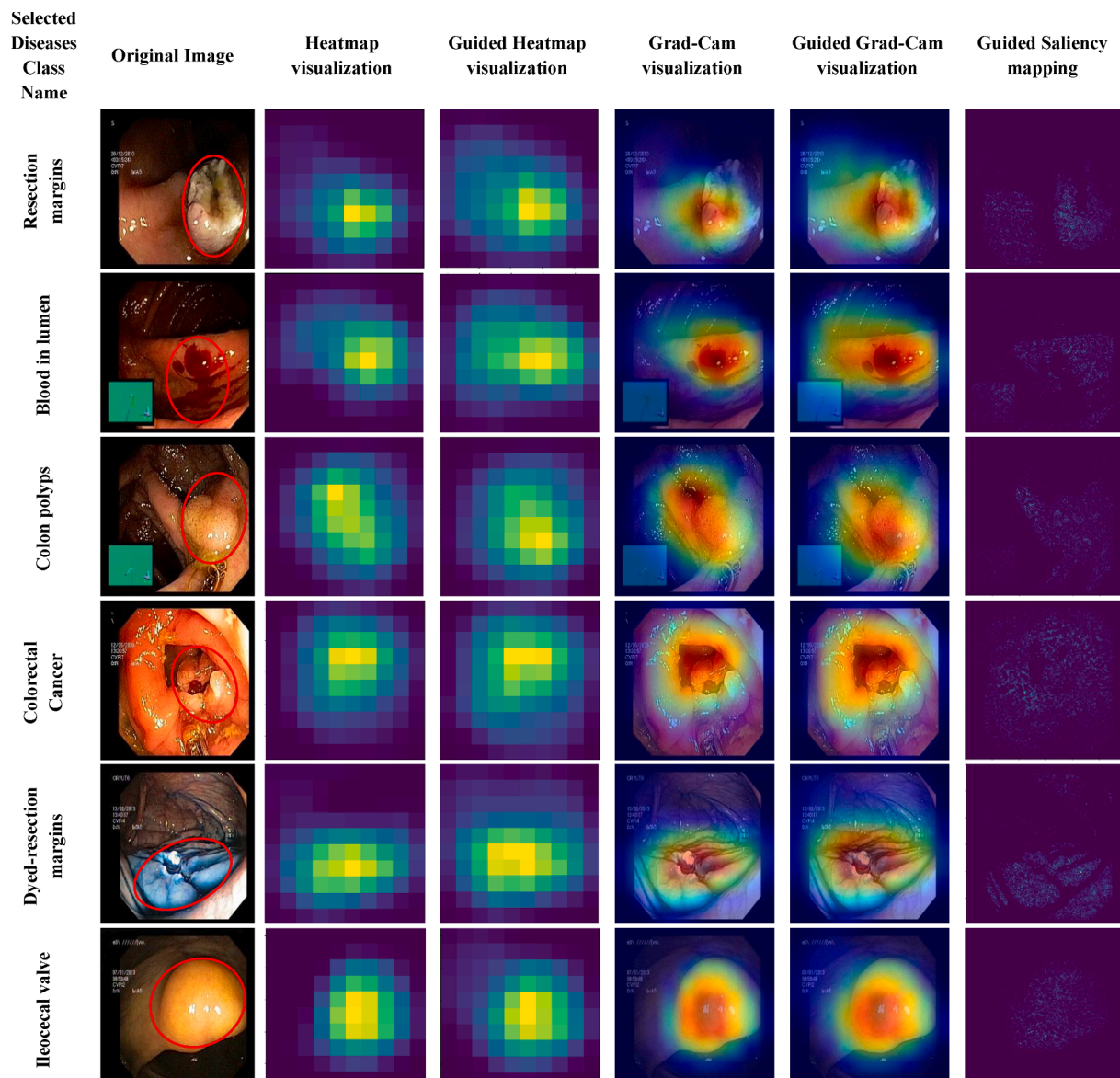*Bold values indicate the best score.

Fig. 18. Grad-CAM visualization demonstrating the most accurate prediction utilizing the proposed model, where the red circles on the original images indicate the region responsible for the specific diseases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.5. Optimization of PCC threshold

Once the feature extraction was completed, a total of 200 features were found; subsequently, the PCC was used to eliminate unnecessary features and select the most important ones, and the suggested EELM classifier was applied for classification. This led to the creation of the PD-CNN-PCC-EELM model. Table 9 displays the PCC values and their performance metrics for different schemes. The PCC threshold value was determined through trial and error.

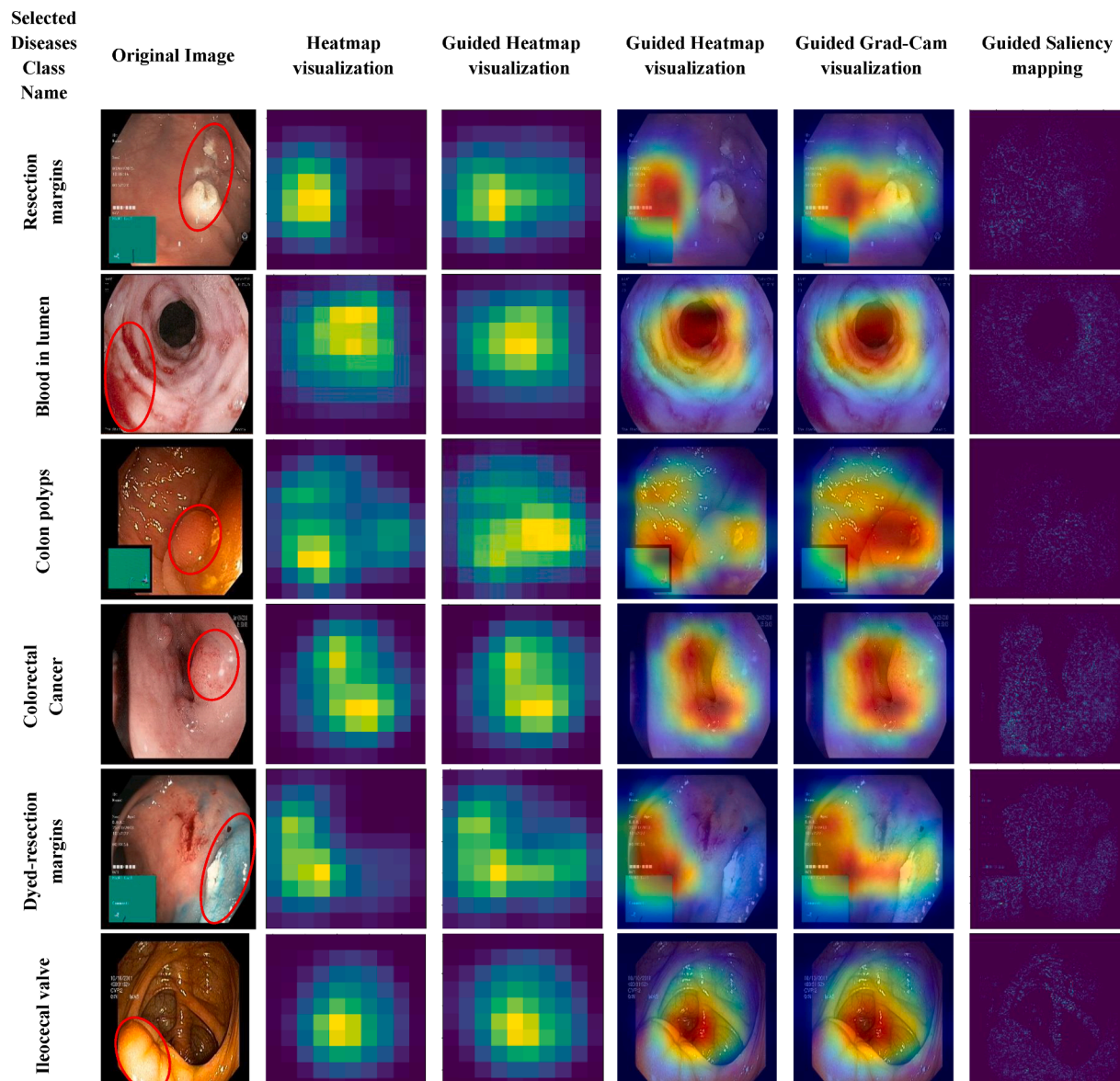The selection of significant features by employing the PCC is based

Fig. 19. Grad-CAM visualization demonstrates less accurate prediction utilizing the proposed model, where the red circles on the original images indicate the region responsible for the specific diseases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

on empirical methods known to optimize classification performance. The determination of significant features depends on their correlations and a specified threshold value in the PCC (refer to Algorithm 2, step 6). To achieve optimal performance, different threshold levels were systematically varied, and model performance metrics were assessed accordingly. It is evident that an increase in the number of features results from the lower threshold values, whereas a decrease in the number of features occurs from higher thresholds. However, the performance assessments are heavily impacted by these differences in feature values. Most importantly, when combined with the suggested EELM classifier, a PCC value of 0.78 produced the best results out of all the tested PCC values, outperforming both the lower and higher thresholds. Moreover, a threshold value of 0.78 for incrementing or decrementing diminishes the detection capabilities. A greater value ($\geq$0.79) will decrease the number of relevant aspects, resulting in less accurate findings with reductions of 1.017 % in recall and accuracy, 0.863 % in precision, and 1.104 % in the F1-score. Correspondingly, if the PCC value falls below the threshold ($\leq$0.77), it results in a reduction in discriminant features, thereby yielding unsatisfactory outcomes. Upon conducting time–cost analyses, the optimal processing efficiency was observed at a threshold

value of 0.78. With the PD-CNN-PCC model integrated into the EELM framework, the test score was 0.00001 s, which is five times faster than that achieved with thresholds greater than 0.79 and six times faster than that achieved with thresholds less than or equal to 0.77. Notably, both the ELM and RELM classifier models exhibited superior performance compared to the other threshold levels. Consequently, based on this comprehensive examination, the threshold value of 0.78 emerged as the optimal choice for the PCC in this research endeavor. Figs. 14 and 15 present the visualization of the feature numbers and performances on various threshold levels.

5.6. Data distribution with PCC

T-SNE is a method used to decrease nonlinear dimensions and display complicated information (Arora et al., 2018). Data visualization is achieved by transforming data from higher dimensions into two or three dimensions, highlighting the proximity of close spots and the uniqueness of distant ones. The t-SNE method works in two steps. It first creates a probability distribution for pairs of high-dimensional objects, assigning greater probabilities to comparable pairings and lower
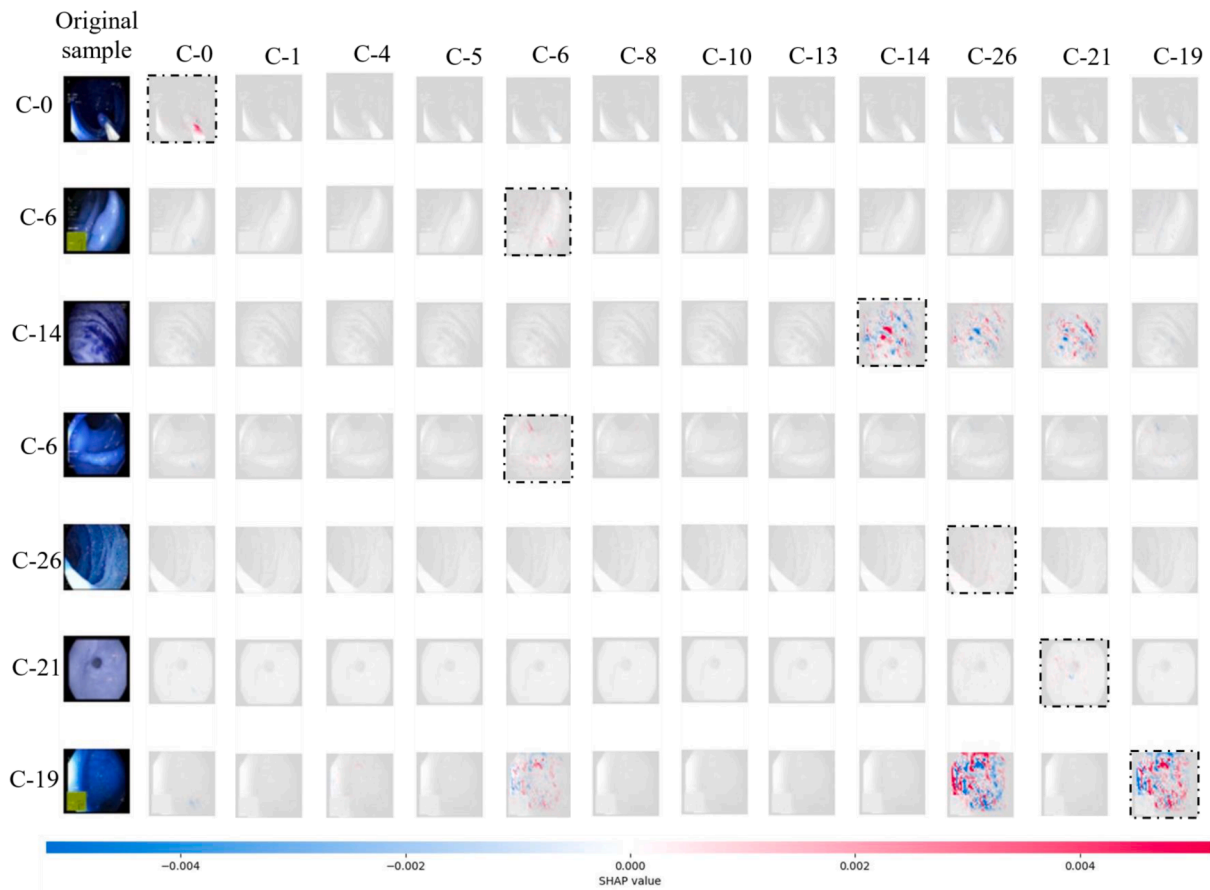
Fig. 20. SHapley Additive exPlanations (SHAP) accurately predicted images for the proposed model (C indicates the selected class).
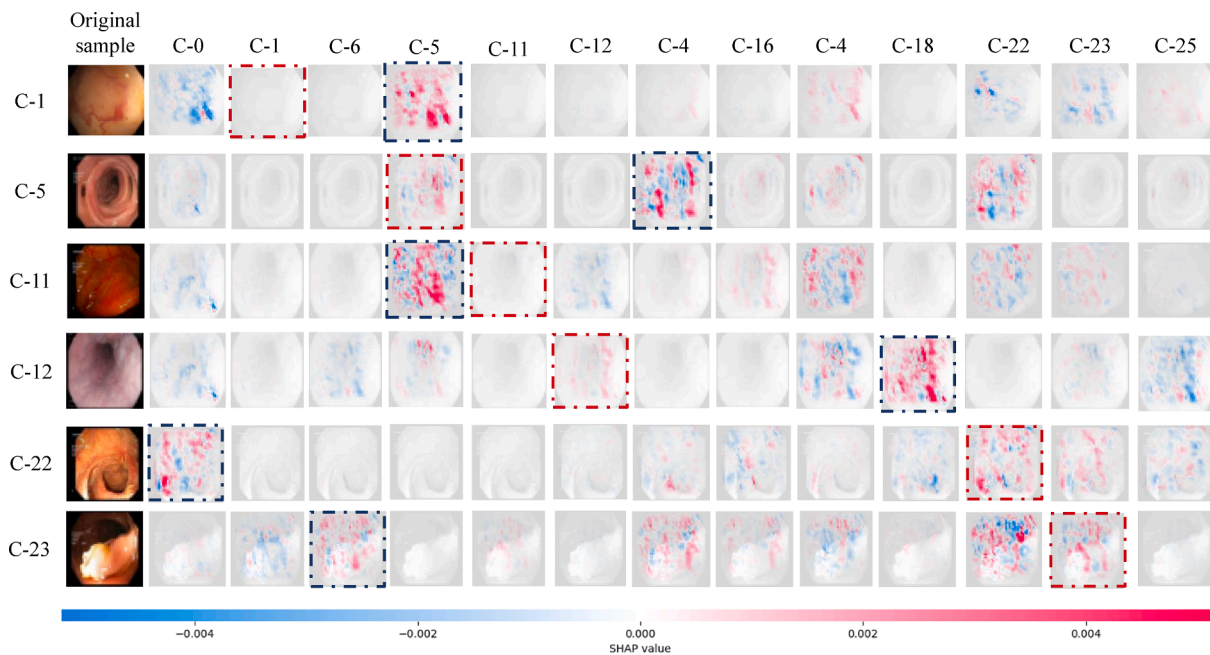

Fig. 21. SHapley Additive exPlanations (SHAP) less accurate predicted images for the proposed model (C indicates the selected class).

probabilities to dissimilar pairs. Subsequently, it establishes a corresponding probability distribution in lower dimensions, minimizing the Kullback–Leibler divergence (KL divergence) between the two distributions concerning the positions of the points on the map. Evaluating

the model's effectiveness involves visualizing its learned insights. The widely used t-SNE method aids in representing learning within the embedded space of a trained model. This analysis indicates that 27 class classifications exhibit reduced 2D representations for the testing

Table 11
PD-CNN-PCC performances on K-fold cross validation where K=10.

Fold Number	Precision			Recall			F1-score			Accuracy			ROC-AUC			Testing Time (Seconds)		
	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM
Fold 0	86.86	87.18	87.67	86.75	87.13	87.38	86.8	87.15	87.52	86.95	87.13	87.38	98.72	98.69	98.8	0.003989	0.003989	0.000003
Fold 1	88.13	86.53	87.49	87.85	86.75	87.38	87.99	86.64	87.43	87.87	86.75	87.38	98.59	98.38	98.59	0.003989	0.003883	0.0000001
Fold 2	87.1	87.99	88.07	87.82	87.75	87.75	87.46	87.87	87.91	88	87.75	88.75	98.61	98.54	98.79	0.004236	0.004984	0.000001
Fold 3	86.31	86.72	87.46	86.63	86.87	87.5	86.47	86.79	87.48	86.63	86.87	87.5	98.65	98.73	98.87	0.004898	0.005003	0.000002
Fold 4	86.87	87.59	88.11	86.87	87.5	87.13	86.87	87.54	87.62	87.37	88.5	87.13	98.48	98.99	98.9	0.00399	0.01034	0.0000002
Fold 5	86.73	86.94	87.73	86.87	86.87	87.5	86.8	86.9	87.61	86.87	86.87	87.5	98.53	98.71	98.71	0.005984	0.004986	0.000001
Fold 6	87.42	85.92	87.96	87.25	86.5	87.7	87.33	86.21	87.83	87.25	86.5	87.87	98.31	98.55	98.66	0.004942	0.003989	0.000002
Fold 7	87.23	87.68	88.19	87.13	86.88	87.25	87.18	87.28	87.72	87.82	86.88	87.85	98.57	98.65	98.86	0.004948	0.003948	0.0000001
Fold 8	87.49	87.54	88.5	87.38	87.5	87.5	87.43	87.52	88	87.38	87.5	88.5	98.71	98.57	98.79	0.004882	0.004986	0.000001
Fold 9	88.11	86.82	88.25	88	86.73	87.13	88.05	86.77	87.69	88	87.63	87.13	98.73	98.43	98.88	0.004987	0.003989	0.000001
Average (μ) ±	87.225	87.091	**87.943**	87.255	87.048	**87.422**	87.238	87.067	**87.681**	87.414	87.238	**87.699**	98.59	98.624	**98.785**	0.0046883	0.0050097	**0.00000114 ± 0**
SD (σ) (%)	± 0.5838	± 0.6266	± **0.3445**	± 0.4946	± 0.4069	± **0.2126**	± 0.5198	± 0.4984	± **0.1874**	± 0.4961	± 0.6028	± **0.5506**	± 0.1288	± 0.1733	± **0.1026**	± 0.0006	± 0.0019	

*Bold values indicate the best results.

datasets. In Fig. 16, the embedded space segregates sample points at the testing phase among the multiclass classification datasets.

For a comprehensive analysis, a two-dimensional t-SNE embedding was utilized on PD-CNN-PCC, as shown in Fig. 16A. Significantly distinct categories, such as "Normal esophagus (class 18)" and "Small bowel_terminal ileum (class 25)", demonstrated a reduced occurrence of misclassification in the t-SNE embedding. The high F1 scores of 0.90 and 0.87 for the "Normal esophagus (class 18)" and "Small bowel_terminal ileum (class 25)" classes are likely attributable to these distinct separations. Conversely, certain classes, such as "Ileocecal valve" (class 16), "Erythema" (class 11), "Esophageal varices" (class 12) and "Angioectasia" (class 1), overlap, making them susceptible to misclassification due to the absence of well-defined boundaries and limited data diversity with other classes. Additionally, t-SNE visualizations with the PCC were generated for the remaining TL models.

5.7. Computational time and resource comparison

Fig. 17 illustrates the comparison of resources among the models. The PD-CNN-PCC model stands out among the evaluated models across various performance measures. Compared with the other models, the PD-CNN-PCC has a notable reduction in architectural complexity, consisting of a mere 0.815 million total parameters. On the other hand, Pr-CNN-PCC required a 1.57 times greater number of parameters. With only 0.812 million trainable parameters, it is the most parameter-efficient model compared to Pr-CNN by 1.58 % and other TL models by 5.06 to 45.74 % (Table 10). Furthermore, compared to DensNet201-PCC models, which have 710 layers and are more difficult to analyze, the proposed model's 24-layer design features a streamlined architecture and a 28.58 % decrease in the number of layers. The compactness of PD-CNN-PCC is evident in its size, which is only 9.79 megabytes, positioning it as one of the smallest models in comparison. The model's compact size and fewer layers enhance its efficiency for deployment on embedded devices, enabling real-time processing. PD-CNN-PCC is approximately 5 times faster than Pr-CNN-PCC, 10 times faster than EfficientNetB6-PCC, and nearly 140 times faster than VGG16-PCC when considering testing time with EELM. In summary, this computational time analysis demonstrates the remarkable speed and efficiency of PD-CNN-PCC compared to the other models. These characteristics enable faster processing and better outcomes on hardware devices. It should be noted that no embedded system was used to test the model in real-world applications. Future work will focus on employing the proposed model for clinical application. Additionally, general-purpose computing systems can provide significant utility due to the model's low computational overhead.

5.8. Interpretability with XAI

The proposed framework enhances transparency and interpretability in decision-making processes by employing multiple XAI approaches. For reliability, widely used XAI methods such as heatmap, Grad-CAM, Guided Heatmap, Guided Grad-CAM, Guided Saliency Mapping, and SHAP were utilized because they are effective for predicting image data. Specifically, the heatmap, Grad-CAM, Guided Heatmap, Guided Grad-CAM, and Guided Saliency Mapping provide feature-capturing explanations, while SHAP offers class-wise feature representation (Abusitta et al., 2024). This comprehensive quantification ensures the trustworthiness of the proposed model.

Several test samples of images were randomly selected to generate the XAI predictions, as depicted in Fig. 18. For instance, in the first row, the sample image belongs to the resection margin class. The heatmap visualizes the affected area crucial for class identification, whereas the guided heatmap highlights this region more precisely. Remarkably, the Guided Grad-CAM integrates the original image as a background, revealing specific features relevant to the class prediction. The sixth column shows that the model predominantly focuses on the center of the resection margin regions, while the last column's guided saliency map
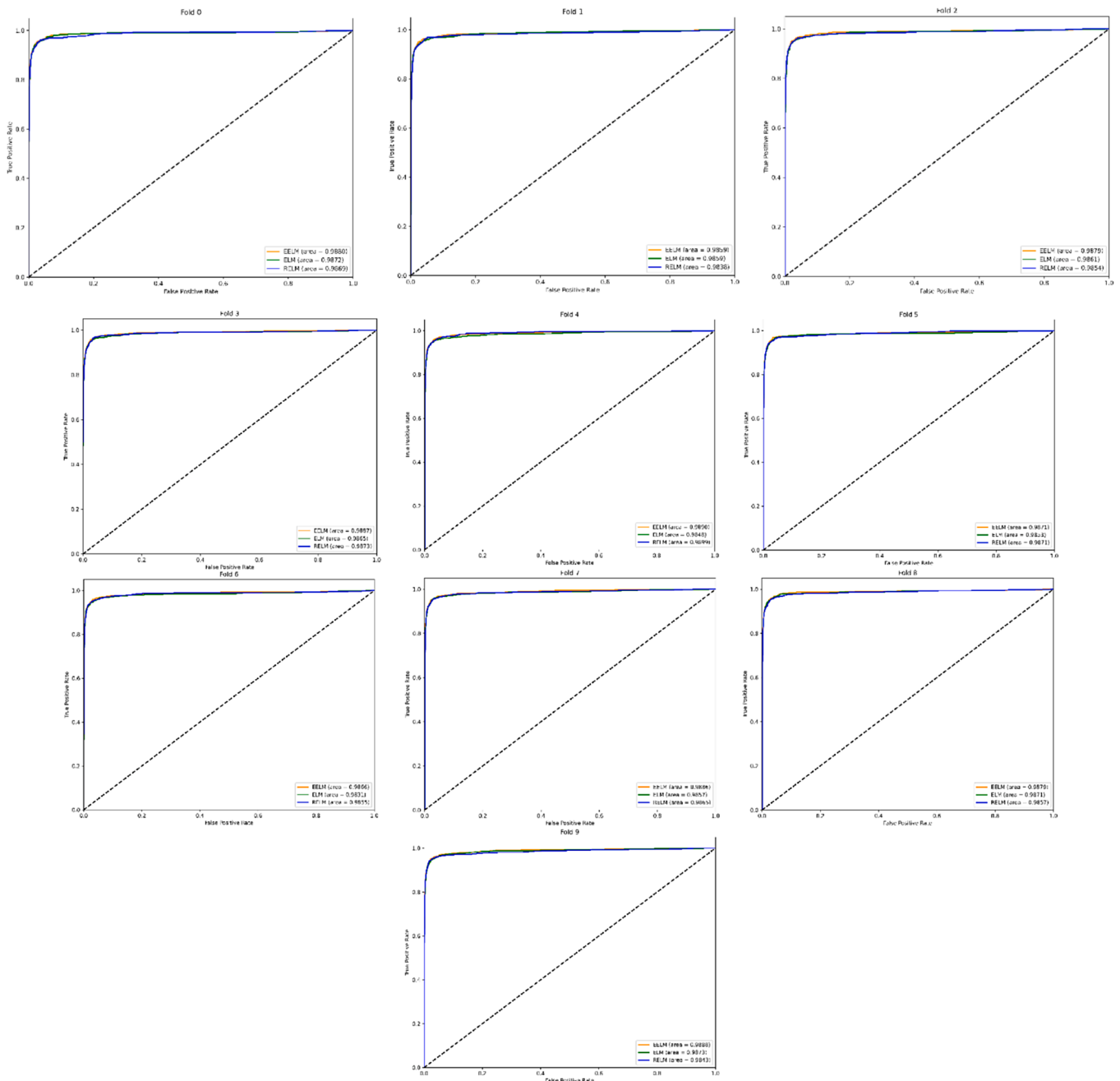
Fig. 22. ROC-AUC performances of the PDCNN-PCC model on the K-fold scheme.

indicates the affected pixels with greater precision, eliminating irrelevant background noise. Similarly, in the third row, a sample image of a colon polyp is shown. The visualizations demonstrate that the guided heatmap identifies the polyp area more accurately than does the heatmap, and the Guided Grad-CAM focuses on the polyp with greater precision. Additionally, the brighter pixels in the guided saliency map accurately highlight the affected pixels. Furthermore, Fig. 19 presents some examples of less accurate predictions for similar classes of data, as shown in Fig. 18. The primary reason for these misclassifications is the complexity and similarity of specific data features across different classes.

In the SHAP explanation, random samples were selected to generate predictions and test the interpretability performance of the proposed method. An exhaustive analysis of several GI features led to the generation of Shapley values, resulting in pixelated visualizations. The analysis demonstrated a clear pattern: red pixels effectively indicated

specific classes, while blue pixels indicated a lower probability of belonging to the target class. To obtain the SHAP results, a set of test samples was randomly selected for prediction. Fig. 20 displays the SHAP results using faint gray backgrounds combined with the original images. The red pixels in the SHAP explanation images in the top row represent the presence of Accessory tools (C-0). Conversely, the lack of blue pixels and the reduced number of red pixels accurately removed other class groupings. The second row shows a clear pattern where red pixels in the SHAP explanation images represent the Colon polyps (C-6) class, with an excess of red pixels correctly indicating class membership. Subsequently, blue pixels in the SHAP explanation graphics for other classes indicated a lower probability. The red pixels in the third row of the SHAP explanation images suggested strong evidence of Gastric polyps (C-16) class GI disease. Significantly, C-19 exhibited a very competitive XAI prediction compared to C-26 and C-19, with red pixels depicted in both classes. However, a significant presence of blue pixels in the C-26

Table 12
Class-wise performance of PD-CNN-PCC on the KvasirV2 test set.

GI Disease classes	Precision			Recall			F1-score			Test Samples	Accuracy (%)			Testing Time (Second)		
	ELM	RELM	EELM	ELM	RELM	EELM	ELM	RELM	EELM		ELM	RELM	EELM	ELM	RELM	EELM
dyed-lifted-polyps (0)	0.97	0.96	0.98	0.97	0.96	0.97	0.97	0.96	0.97	100	97.87	97.37	98.01	0.00000023	0.00000041	**0.00000005**
dyed-resection-margins (1)	0.96	0.95	0.97	0.97	0.96	0.98	0.96	0.95	0.97	104						
esophagitis (2)	0.98	0.97	0.98	0.96	0.96	0.98	0.97	0.96	0.98	97						
normal-cecum (3)	0.98	0.98	0.98	0.99	0.99	0.99	0.98	0.98	0.98	88						
normal-pylorus (4)	0.99	0.99	0.99	1	1	1	0.99	0.99	0.99	103						
normal-z-line (5)	0.95	0.97	0.97	0.95	0.97	0.98	0.95	0.97	0.97	109						
Polyps (6)	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	86						
ulcerative-colitis (7)	0.99	0.98	0.99	0.97	0.97	0.98	0.98	0.97	0.98	113 / 100						
Average (μ) ± SD (σ) (%)	97.5 ± 0.01414	97.25 ± 0.01281	**98 ± 0.00755**	97.37 ± 0.01597	97.37 ± 0.01505	**98.25 ± 0.00886**	97.25 ± 0.01281	97 ± 0.01309	**97.75 ± 0.007071**							

*Bold values indicate the best results.

projection resulted in an increased false projection for C-26. Additionally, rows 4, 5, 6, and 7 correctly recognized disease classifications by emphasizing red pixels in specific areas.

Finally, SHAP visualizations were generated for specific classes with less accurate model predictions, as shown in Fig. 21. Due to fewer training samples, the model struggled to learn specific class features perfectly. However, in the 2nd, 3rd, 5th, and 6th rows (C-5, 11, 22, 23), the model showed competitive feature extraction visibility, with red pixels found in accurate classes as well. These visual SHAP explanations validated the model's outcomes, offering doctors a deeper understanding of specific disease categories.

5.9. K-fold cross-validation performances

Table 11 presents the comparative performance metrics of the proposed PD-CNN-PCC architecture using three classifiers, ELM, RELM, and EELM, evaluated on the GastroVision dataset through a 10-fold cross-validation scheme. Across all folds, the EELM consistently outperforms the ELM and RELM. Specifically, the EELM achieved the highest average Precision (87.943 ± 0.3445), Recall (87.422 ± 0.2126), F1-score (87.681 ± 0.1874), Accuracy (87.699 ± 0.5506), and ROC-AUC (98.785 ± 0.1026). Beyond superior classification metrics, the EELM model exhibited a notable improvement in computational efficiency. The average testing time for the EELM was significantly lower (0.00000114 s) than that for the ELM (0.0046883 s) and RELM (0.0050097 s). This reduction in testing time underscores the efficiency of the EELM model in rapidly processing data, which is critical for practical applications. Furthermore, the average performance metrics across the 10 folds were consistent with the class-wise average performance metrics. The ROC-AUC curves for each fold are presented in Fig. 22. This consistency affirms the robustness and reliability of the proposed model across various dataset configurations.

5.10. Validation on KvasirV2 dataset

Table 12 presents the class-wise performance of the proposed PD-CNN-PCC-EELM model on the KvasirV2 test set and compares it with that of the ELM and RELM classifiers. The proposed model was trained on the KvasirV2 dataset in a similar way as the GastroVision dataset. The EELM model trained on the KvasirV2 dataset achieved an accuracy of 98.01 %, while the ELM and RELM models achieved accuracies of 97.87 % and 97.37 %, respectively. Additionally, the EELM model demonstrated superior computational efficiency, with an average testing time of 0.00000005 s, which is significantly faster than that of the ELM (0.0000023 s) and RELM (0.0000041 s). Similar improvement trends were also found for other performance parameters, such as precision, recall and F1-score. Furthermore, the ROC-AUC and AUC-PR curves, shown in Fig. 23, indicate that the EELM model outperforms the RELM model in both metrics and demonstrates competitive performance with the ELM model. These results further validated the effectiveness and efficiency of the proposed PD-CNN-PCC-EELM model on a new dataset.

5.11. Discussion, Limitations, and Future work

While numerous studies have focused on GI disease diagnosis, identification, and segmentation, more research on multi-class classification encompassing a broad spectrum of GI diseases is needed. This discussion presents a thorough review of the experimental results achieved using the proposed method, which effectively categorizes 27 types of GI tract issues for the first time. The proposed approach comprises four main steps: dataset preprocessing, feature extraction, feature selection, and classification with interpretability. The preprocessing step ensures the use of refined and standardized input data, enhancing the performance and accuracy of classification tasks. The PD-CNN extracted 200 features, but the feature selection stage retained only 39 features using the PCC. The EELM outperforms the other classifiers in terms of
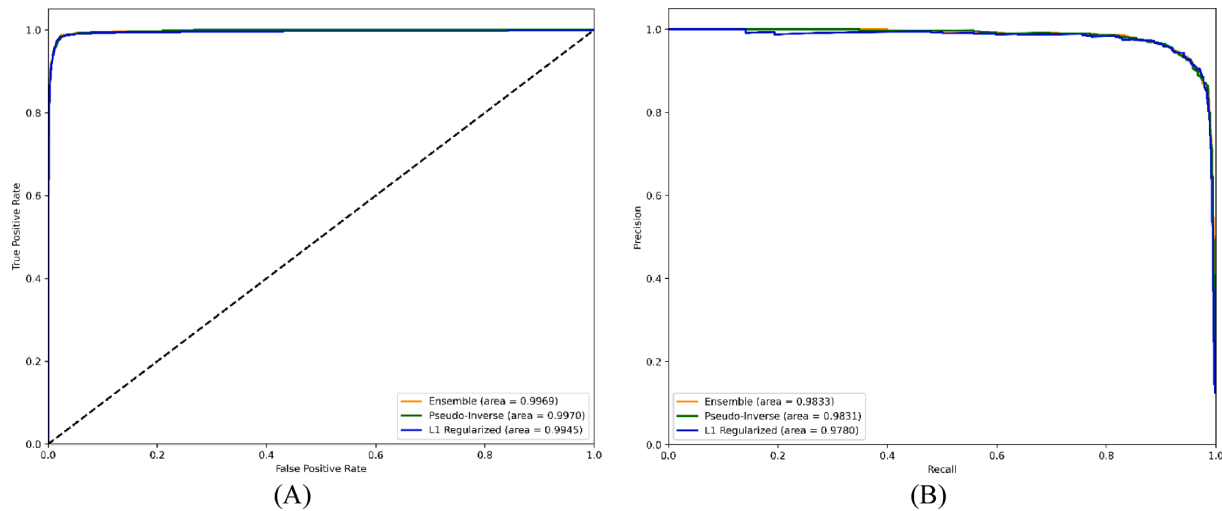
Fig. 23. Performance of the PD-CNN-PCC-ELM (Pseudo-Inverse)/RELM (L1-Regularized)/EELM (Ensemble) models on the KvasirV2 test set: (A) ROC-AUC and (A) AUC-PR performances.

precision, recall, F1 score, accuracy, ROC-AUC and AUC-PR performance.

Table 13 compares previous classification models with the proposed PD-CNN-PCC-EELM model. Noor et al. (Noor et al., 2023) show that MobileNetV2, with 3.4 million parameters and 154 layers, achieved an accuracy of 97.68 % for five classes. Nouman et al. (Nouman Noor et al., 2023) also employed MobileNetV2 for the 5-class classification task, utilizing 3.4 million parameters and 1210 extracted features. Gunasekaran et al. (Gunasekaran et al., 2023) obtained a 95 % accuracy rate for eight classes by employing the ensemble TL model, which comprised 66.94 million features. Öztürk et al. (Öztürk & Özkaya, 2021) combined a ResNet50 TL model with a residual LSTM classifier to achieve 98.05 % accuracy in an eight-class classification task using the Kvasir dataset. Other researchers (Khan et al., 2024; Lonseko et al., 2021; Ramzan et al., 2023; Thomas Abraham et al., 2023; Yogapriya et al., 2021) have utilized TL-based models for classifying gastrointestinal disorders. As previously stated, all these models contained many parameters, ranging from 5.3 to 138 million. Most of these models prioritize extracting a large number of features (943–1000), which leads to higher computational requirements for both training and inference. Large feature sets in models can result in slower inference times, making them unsuitable for practical use.

On the other hand, the PD-CNN-PCC-EELM model achieved a comparable accuracy of 86.13 % across 27 classes with only 0.815 million parameters, 24 layers, and 39 features. The proposed model, which is a lightweight NN, attained notable accuracy for all 27 classes within a testing time of 0.00001 s. The design of the model is optimized by simultaneously running the first five CLs to improve feature extraction. The model succeeds in terms of classification performance and computational requirements compared to the SOTA-TL models, as shown in Tables 8 and 10. This approach reduces the number of parameters, layers, and testing time while maintaining acceptable accuracy. The use of SHAP, heatmap, Grad-CAM, guided Grad-CAM and guided saliency map has improved the interpretability of the proposed model by showing that it focuses on relevant image regions to extract useful features.

A few datasets, such as Kvasir, HyperKvasir, Kvasir-Capsule and KID, provide multiple GI findings. However, Kvasir-Capsule and KID are video capsule endoscopy datasets that contain a minimum number of classes. Most previous studies have demonstrated their proposed models using these datasets, classifying 5 to 8 types of GI diseases. For the first time, the GastroVision dataset contained 27 classes (highest) and included more labeled classes of anatomical landmarks, pathological findings, and normal findings. Additionally, baseline results have been

established on this dataset for GI disease detection and classification of the upper, lower, and combined GI tract, offering valuable research resources for advancing GI endoscopy studies.

Although the suggested method yields outcomes comparable to those of a lightweight model, it has several drawbacks. The model's classification accuracy is inferior to that of other existing studies. The primary cause of this lower classification accuracy is that the current study worked on a large dataset with 27 different classes. Additionally, the GastroVision dataset has diverse types of images of different GI diseases. Another issue affecting the classification accuracy is the resizing of images during the preprocessing stage, which results in a significant loss of resolution. Moreover, some classes of the dataset contain very few sample images, such as ulcer (class 26), which has only 6 sample images, and esophageal varices (class 12), which has 7 sample images. Therefore, there is still potential for additional improvement in the model. Future endeavors of the authors will focus on improving the model's efficiency by balancing the GastroVision dataset. The researchers plan to gather and create a more evenly distributed dataset to enhance classification results.

For the KvasirV2 dataset, the proposed model achieved an accuracy of 98.01 %, maintaining a significantly low testing time of 0.00000005 s and utilizing only 39 essential features. This performance surpasses that of other SOTA works (Nouman Noor et al., 2023; Noor et al., 2023; Gunasekaran et al., 2023; Yogapriya et al., 2021; Lonseko et al., 2021; Khan et al., 2024).

Additionally, as a future avenue of exploration, implementing a real-world hardware-based design is envisioned to provide enhanced visualization capabilities for the proposed model, thus improving its practical utility.

6. Conclusion

This study presented a novel method for precisely categorizing gastrointestinal (GI) tract disorders by integrating the parallel Depthwise Separable CNN (PD-CNN) feature extractor and PCC feature selector with the Ensemble ELM (EELM) classifier. The proposed model, consisting of 24 layers and 0.815 million parameters, effectively categorizes twenty-seven types of different anatomical positions of GI diseases while decreasing the computational burden. The testing duration of the EELM model was only 0.0001 s after integrating the PCC, which reduced irrelevant features. The hybrid EELM classifier improves the classification performance by combining the ELM and RELM algorithms. The proposed approach has shown excellent classification performance, with precision, recall, f1, accuracy, ROC-AUC, and AUC-PR values of

Table 13

Results of previous studies compared with those of the proposed PD-CNN-PCC-EELM model.

Ref.	Dataset	Number of Sample Images	Number of Class	Feature Extractor	Parameters (Million)	Number of Layers	Model Size (MB)	Number of Features	Best Classifier	Testing Accuracy (%)	Testing Time (Seconds)	Real-time XAI
Nouman et al. (Nouman Noor et al., 2023)	KvasirV-2 and Hyper-Kvasir	4854	5	MobileNetV2	3.4	154	–	1210	Softmax	96.40	––	No
Noor et al. (Noor et al., 2023)	KvasirV-2 and Hyper-Kvasir	4854	5	MobileNetV2	3.4	154	–	810	Softmax	97.68	––	Yes (Grad-CAM)
Gunasekaran et al. (Gunasekaran et al., 2023)	KvasirV-2	8000	8	Ensemble Model (InceptionV3, DenseNet201, ResNet201)	66.94	––	–	––	––	95	––	No
Öztürk et al. (Öztürk & Özkaya, 2021)	Kvasir	6000	8	ResNet50	23.9	––	–	––	Residual LSTM	98.05	––	No
Yogapriya et al. (Yogapriya et al., 2021)	KvasirV-2	8000	8	VGG16	138	**16**	–	––	––	96.33	––	No
Hmoud Al-Adhaileh et al. (Hmoud Al-Adhaileh et al., 2021)	Kvasir	5000	5	AlexNet	62.3	25	–	1000	––	97.00	––	No
Lonseko et al. (Lonseko et al., 2021)	KvasirV-2	8000	8	Deep CNN based Attention	19.92	––	–	––	––	93.19	––	Yes (Heatmap)
Ramzan et al. (Ramzan et al., 2023)	Kvasir	4000	5	InceptionNetV3, GITNet	––	50	–	1000	QSVM	99.32	0.0016	Yes (Grad-CAM)
Thomas Abraham et al. (Thomas Abraham et al., 2023)	Kvasir	5000	5	Custom CNN with EfficientNetB0	5.3	––	–	––	––	98.01	––	Yes (Grad-CAM)
Khan et al. (Khan et al., 2024)	KvasirV-2	8000	8	Darknet 53, Xception	––	––	–	943	Ensemble-based Subspace KNN (ESKNN)	98.25	––	No
Proposed Model	**GastroVision** **Kvasir v2**	**8000** **8000**	**27** **8**	**PD-CNN**	**0.815**	**24**	**9.79**	**39**	**EELM**	**87.75** **98.01**	**0.000001** **0.00000005**	**Yes** (Heatmap, guided Heatmap Grad-CAM, Guided Grad-CAM, Salience Mapping and SHAP)

*Bold values indicate the best result.

88.12 ± 0.332 %, 87.75 ± 0.348 %, 87.12 ± 0.324 %, 87.75 %, 98.89 %, and 92 %, respectively. Additionally, the proposed model is compact, with a size of only 9.79 MB, which makes it suitable for practical use. Moreover, its low computational requirements (parameters, layer, size) would enable its deployment on cost-effective edge devices quite easily. Furthermore, combining real-time explainable (XAI) helps medical experts by offering a reliable explanation of the model's results in revealing the right type of GI disorder. In conclusion, the PD-CNN-PCC-EELM technique greatly enhances the accuracy of classifying 27 types of GI tract diseases and is easy to implement and evaluate in real-world scenarios. However, there is still room for further improvement in accuracy.

CRediT authorship contribution statement

Md. Faysal Ahamed: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Visualization. **Md. Nahiduzzaman:** Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Visualization. **Md. Rabiul Islam:** Methodology, Investigation, Data curation, Writing – original draft. **Mansura Naznine:** Conceptualization, Methodology, Investigation, Data curation, Writing – original draft. **Mohamed Arselene Ayari:** Formal analysis, Validation, Writing – review & editing, Supervision. **Amith Khandakar:** Formal analysis, Validation, Writing – review & editing, Supervision. **Julfikar Haider:** Formal analysis, Validation, Writing – review & editing, Visualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

References

Abusitta, A., Li, M. Q., & Fung, B. C. M. (2024). Survey on explainable AI: Techniques, challenges and open issues. *Expert Systems with Applications*, Article 124710.

Alhajlah, M., Noor, M. N., Nazir, M., Mahmood, A., Ashraf, I., & Karamat, T. (2023). Gastrointestinal diseases classification using deep transfer learning and features optimization. *CMC-Computers Materials & Continua, 75*(1), 2227–2245.

Arnold, M., Abnet, C. C., Neale, R. E., Vignat, J., Giovannucci, E. L., McGlynn, K. A., & Bray, F. (2020). Global burden of 5 major types of gastrointestinal cancer. *Gastroenterology, 159*(1), 335–349.e15. https://doi.org/10.1053/j. gastro.2020.02.068

Arora, S., Hu, W., & Kothari, P. K. (2018). An Analysis of the t-SNE Algorithm for Data Visualization. In S. Bubeck, V. Perchet, & P. Rigollet (Eds.), *Proceedings of the 31st Conference On Learning Theory* (Vol. 75, pp. 1455–1462). PMLR. https://proceedings. mlr.press/v75/arora18a.html.

Aruna, P., Puviarasan, N., & Palaniappan, B. (2007). Diagnosis of gastrointestinal disorders using DIAGNET. *Expert Systems with Applications, 32*(2), 329–335. https:// doi.org/10.1016/j.eswa.2005.11.039

Awais, M. M., & Awan, S. K. (2011). Gastro-intestinal tract inspired computational model for myocardial infarction diagnosis. *Expert Systems with Applications, 38*(5), 5633–5641. https://doi.org/10.1016/j.eswa.2010.10.072

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). *Pearson Correlation Coefficient BT - Noise Reduction in Speech Processing* (I. Cohen, Y. Huang, J. Chen, & J. Benesty (eds.); pp. 1–4). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5.

Bhandari, M., Shahi, T. B., Siku, B., & Neupane, A. (2022). Explanatory Classification of CXR Images into COVID-19, Pneumonia and Tuberculosis Using Deep Learning and

XAI. *Computers in Biology and Medicine, 150*(C). https://doi.org/10.1016/j. compbiomed.2022.106156

Bhatia, Y., Bajpayee, A., Raghuvanshi, D., & Mittal, H. (2019). Image captioning using Google's inception-resnet-v2 and recurrent neural network. *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 1–6.

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, S., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., & Nguyen, D. T. D. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data, 7*(1), 283.

Chaithanya, B. N., Swasthika Jain, T. J., Usha Ruby, A., & Parveen, A. (2021). An approach to categorize chest X-ray images using sparse categorical cross entropy. *Indonesian Journal of Electrical Engineering and Computer Science*, 1700–1710.

Chen, L., Chen, J., Hajimirsadeghi, H., & Mori, G. (2020). Adapting grad-cam for embedding networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2794–2803.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258.

Ding, S., Xu, X., & Nie, R. (2014). Extreme learning machine and its applications. *Neural Computing and Applications, 25*(3), 549–556. https://doi.org/10.1007/s00521-013-1522-8

Fan, S., Xu, L., Fan, Y., Wei, K., & Li, L. (2018). Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images. *Physics in Medicine & Biology, 63*(16), Article 165001.

Fujii-Lau, L. L., Thosani, N. C., Al-Haddad, M., Acoba, J., Wray, C. J., Zvavanjanja, R., … Qumseya, B. J. (2023). American Society for Gastrointestinal Endoscopy guideline on the role of endoscopy in the diagnosis of malignancy in biliary strictures of undetermined etiology: Summary and recommendations. *Gastrointestinal Endoscopy, 98*(5), 685–693. https://doi.org/10.1016/j.gie.2023.06.005

Gunasekaran, H., Ramalakshmi, K., Swaminathan, D. K., & Mazzara, M. (2023). GIT-Net: An ensemble deep learning-based GI tract classification of endoscopic images. *Bioengineering, 10*(7), 809.

Gupta, S., Seleq, S., Gimpaya, N., Khan, R., Scaffidi, M. A., & Grover, S. C. (2022). Interobserver reliability of the paris classification for superficial gastrointestinal tract neoplasms: A systematic review and meta-analysis. *Gastrointestinal Endoscopy, 95*(6, Supplement), AB96–AB97. https://doi.org/10.1016/j.gie.2022.04.260

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition* (pp. 770–778). http://image-net.org/challenges/LSVRC/2015/.

Hesse, R., Schaub-Meyer, S., & Roth, S. (2023). Content-adaptive downsampling in convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4544–4553.

Hmoud Al-Adhaileh, M., Mohammed Senan, E., Alsaade, F. W., Aldhyani, T. H. H., Alsharif, N., Abdullah Alqarni, A., Uddin, M. I., Alzahrani, M. Y., Alzain, E. D., & Jadhav, M. E. (2021). Deep learning algorithms for detection and classification of gastrointestinal diseases. *Complexity, 2021*, 6170416. https://doi.org/10.1155/2021/6170416

Iakovidis, D. K., & Koulaouzidis, A. (2014). Automatic lesion detection in capsule endoscopy based on color saliency: Closer to an essential adjunct for reviewing software. *Gastrointestinal Endoscopy, 80*(5), 877–883. https://doi.org/10.1016/j. gie.2014.06.026

Iddan, G., Meron, G., Glukhovsky, A., & Swain, P. (2000). Wireless capsule endoscopy. *Nature, 405*(6785), 417.

Jain, S., Seal, A., Ojha, A., Yazidi, A., Bures, J., Tacheci, I., & Krejcar, O. (2021). A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images. *Computers in Biology and Medicine, 137*, Article 104789. https://doi.org/10.1016/j.compbiomed.2021.104789

Jha, D., Sharma, V., Dasu, N., Tomar, N. K., Hicks, S., Bhuyan, M. K., Das, P. K., Riegler, M. A., Halvorsen, P., Bagci, U., & de Lange, T. (2024). In *GastroVision: A Multi-class Endoscopy Image Dataset for Computer Aided Gastrointestinal Disease Detection BT - Machine Learning for Multimodal Healthcare Data* (pp. 125–140). Springer Nature Switzerland.

Jin, W., Li, X., Fatehi, M., & Hamarneh, G. (2023). Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical Image Analysis, 84*, Article 102684. https://doi.org/10.1016/j.media.2022.102684

Johannes, R. S., Tabak, Y. P., Sun, X., Wolf, A. T., & Saltzman, J. R. (2008). Development and validation of a simple risk classification rule for patients with acute lower gastrointestinal bleeding. *Gastrointestinal Endoscopy, 67*(5), AB83. https://doi.org/10.1016/j.gie.2008.03.046

Jun, H., Song, H. J., Boo, S.-J., & Kim, H. U. (2022). Upper gastrointestinal involvement of Behcet's disease. *Gastrointestinal Endoscopy, 95*(6, Supplement), AB466–AB467. https://doi.org/10.1016/j.gie.2022.04.1170

Kaiser, L., Gomez, A. N., & Chollet, F. (2017). Depthwise separable convolutions for neural machine translation. *ArXiv Preprint.* ArXiv:1706.03059.

Khan, M. A., Kadry, S., Alhaisoni, M., Nam, Y., Zhang, Y., Rajinikanth, V., & Sarfraz, M. S. (2020). Computer-aided gastrointestinal diseases analysis from wireless capsule endoscopy: A framework of best features selection. *IEEE Access, 8*, 132850–132859.

Khan, M. A., Khan, M. A., Ahmed, F., Mittal, M., Goyal, L. M., Hemanth, D. J., & Satapathy, S. C. (2020). Gastrointestinal diseases segmentation and classification based on duo-deep architectures. *Pattern Recognition Letters, 131*, 193–204.

Khan, M. A., Sarfraz, M. S., Alhaisoni, M., Albesher, A. A., Wang, S., & Ashraf, I. (2020). StomachNet: Optimal deep learning features fusion for stomach abnormalities classification. *IEEE Access, 8*, 197969–197981. https://doi.org/10.1109/ACCESS.2020.3034217

Khan, Z. F., Ramzan, M., Raza, M., Khan, M. A., Iqbal, K., Kim, T., & Cha, J. H. (2024). Deep convolutional neural networks for accurate classification of gastrointestinal

tract syndromes. *Computers, Materials and Continua, 78*(1), 1207–1225. https://doi.org/10.32604/cmc.2023.045491

Kim, N. H., Jung, Y. S., Jeong, W. S., Yang, H.-J., Park, S.-K., Choi, K., & Park, D. I. (2017). Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intestinal Research, 15*(3), 411–418.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM, 60*(6), 84–90. https://doi.org/10.1145/3065386

Kusano, C., Gotoda, T., Ishikawa, H., Suzuki, S., Ikehara, H., & Matsuyama, Y. (2024). Gastric cancer detection rates using gastrointestinal endoscopy with serological risk stratification: A randomized controlled trial. *Gastrointestinal Endoscopy*. https://doi.org/10.1016/j.gie.2024.01.022

Lan, L., & Ye, C. (2021). Recurrent generative adversarial networks for unsupervised WCE video summarization. *Knowledge-Based Systems, 222*, Article 106971. https://doi.org/10.1016/j.knosys.2021.106971

Lee, J. H., Kim, Y. J., Kim, Y. W., Park, S., Choi, Y.i., Kim, Y. J., Park, D. K., Kim, K. G., & Chung, J. W. (2019). Spotting malignancies from gastric endoscopic images using deep learning. *Surgical Endoscopy, 33*(11), 3790–3797. https://doi.org/10.1007/S00464-019-06677-2/METRICS

Li, B., & Meng, M.-Q.-H. (2009). Texture analysis for ulcer detection in capsule endoscopy images. *Image and Vision Computing, 27*(9), 1336–1342. https://doi.org/10.1016/j.imavis.2008.12.003

Li, B., & Meng, M.-Q.-H. (2012). Automatic polyp detection for wireless capsule endoscopy images. *Expert Systems with Applications, 39*(12), 10952–10958. https://doi.org/10.1016/j.eswa.2012.03.029

Li, Y., Su, Y., Guo, M., Han, X., Liu, J., Vishwasrao, H. D., Li, X., Christensen, R., Sengupta, T., & Moyle, M. W. (2022). Incorporating the image formation process into deep learning improves network performance. *Nature Methods, 19*(11), 1427–1437.

Lonseko, Z. M., Adjei, P. E., Du, W., Luo, C., Hu, D., Zhu, L., Gan, T., & Rao, N. (2021). Gastrointestinal Disease Classification in Endoscopic Images Using Attention-Guided Convolutional Neural Networks. In *Applied Sciences* (Vol. 11, Issue 23). https://doi.org/10.3390/app112311136.

Mohapatra, S., Kumar Pati, G., Mishra, M., & Swarnkar, T. (2023). Gastrointestinal abnormality detection and classification using empirical wavelet transform and deep convolutional neural network from endoscopic images. *Ain Shams Engineering Journal, 14*(4), Article 101942. https://doi.org/10.1016/j.asej.2022.101942

Musha, A., Hasnat, R., Mamun, A. A., Ping, E. P., & Ghosh, T. (2023). Computer-aided bleeding detection algorithms for capsule endoscopy: A systematic review. *Sensors, 23*(16), 7170.

Nass, K. J., Zwager, L. W., Van Der Vlugt, M., Dekker, E., Bossuyt, P. M., Ravindran, S., Thomas-Gibson, S., & Fockens, P. (2022). A Novel classification for adverse events in gastrointestinal endoscopy: The agree classification. *Gastrointestinal Endoscopy, 95*(6, Supplement), AB67. https://doi.org/10.1016/j.gie.2022.04.189

Noor, M. N., Nazir, M., Ashraf, I., Almujally, N. A., Aslam, M., & Fizzah Jilani, S. (2023). GastroNet: A robust attention-based deep learning and cosine similarity feature selection framework for gastrointestinal disease classification from endoscopic images. CAAI Transactions on Intelligence Technology.

Nouman Noor, M., Nazir, M., Khan, S. A., Song, O.-Y., & Ashraf, I. (2023). Efficient gastrointestinal disease classification using pretrained deep convolutional neural network. *Electronics, 12*(7), 1557.

Noya, F., Álvarez-González, M. A., & Benítez, R. (2017). Automated angiodysplasia detection from wireless capsule endoscopy. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3158–3161. https://doi.org/10.1109/EMBC.2017.8037527.

Öztürk, Ş., & Özkaya, U. (2021). Residual LSTM layered CNN for classification of gastrointestinal tract diseases. *Journal of Biomedical Informatics, 113*, Article 103638. https://doi.org/10.1016/j.jbi.2020.103638

Pan, G., Yan, G., Qiu, X., & Cui, J. (2011). Bleeding detection in wireless capsule endoscopy based on probabilistic neural network. *Journal of Medical Systems, 35*(6), 1477–1484. https://doi.org/10.1007/s10916-009-9424-0

Parasa, S., Repici, A., Berzin, T., Leggett, C., Gross, S. A., & Sharma, P. (2023). Framework and metrics for the clinical use and implementation of artificial intelligence algorithms into endoscopy practice: Recommendations from the American Society for Gastrointestinal Endoscopy Artificial Intelligence Task Force, 815–824.e1 *Gastrointestinal Endoscopy, 97*(5). https://doi.org/10.1016/j.gie.2022.10.016.

Parsa, N., Haito-Chavez, Y., Brewer Gutierrez, O. I., Paiji, C., Inoue, H., Beard, K. W., Draganov, P. V, Ujiki, M., Rahden, B. H. A., Desai, P. N., Pioche, M., Hayee, B., Haji, A., Saxena, P., Reavis, K., Onimaru, M., Balassone, V., Nakamura, J., Hata, Y., … Khashab, M. A. (2018). Sa1907 Classification and grading of adverse events related to peroral endoscopic myotomy (POEM): A comparison between the american society of gastrointestinal endoscopy lexicon and the Clavien-Dindo classification.

Gastrointestinal Endoscopy, *87*(6, Supplement), AB244–AB245. https://doi.org/10.1016/j.gie.2018.04.429.

Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.-T., Lux, M., Schmidt, P. T., Riegler, M., & Halvorsen, P. (2017). KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, 164–169. https://doi.org/10.1145/3083187.3083212.

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint*. ArXiv:2010.16061.

Ramzan, M., Raza, M., Sharif, M. I., Azam, F., Kim, J., & Kadry, S. (2023). Gastrointestinal tract disorders classification using ensemble of InceptionNet and proposed GITNet based deep feature with ant colony optimization. *PLoS One1, 18*(10), e0292601.

Rustam, F., Siddique, M. A., Siddiqui, H. U. R., Ullah, S., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Wireless capsule endoscopy bleeding images classification using CNN based model. *IEEE Access, 9*, 33675–33688.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Saraiva, R., Perkusich, M., Silva, L., Almeida, H., Siebra, C., & Perkusich, A. (2016). Early diagnosis of gastrointestinal cancer by using case-based and rule-based reasoning. *Expert Systems with Applications, 61*, 192–202. https://doi.org/10.1016/j.eswa.2016.05.026

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Sharif, M., Attique Khan, M., Rashid, M., Yasmin, M., Afza, F., & Tanik, U. J. (2021). Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images. *Journal of Experimental & Theoretical Artificial Intelligence, 33*(4), 577–599.

Shi, X., Kang, Q., An, J., & Zhou, M. (2022). Novel L1 Regularized Extreme Learning Machine for Soft-Sensing of an Industrial Process. *IEEE Transactions on Industrial Informatics, 18*(2), 1009–1017. https://doi.org/10.1109/TII.2021.3065377

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint*. ArXiv:1409.1556.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing, 97*, Article 105524. https://doi.org/10.1016/j.asoc.2019.105524

Sivari, E., Bostanci, E., Guzel, M. S., Acici, K., Asuroglu, T., & Ercelebi Ayyildiz, T. (2023). A new approach for gastrointestinal tract findings detection and classification: Deep learning-based hybrid stacking ensemble models. *Diagnostics, 13*(4), 720.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249. https://doi.org/10.3322/caac.21660

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, 6105–6114.

Thomas Abraham, J. V., Muralidhar, A., Sathyarajasekaran, K., & Ilakiyaselvan, N. (2023). A deep-learning approach for identifying and classifying digestive diseases. *Symmetry, 15*(2). https://doi.org/10.3390/sym15020379

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813.

Xu, M., Yoon, S., Fuentes, A., & Park, D. S. (2023). A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition, 137*, Article 109347.

Yang, S., & Berdine, G. (2023). Interpretable artificial intelligence (AI)–saliency maps. *The Southwest Respiratory and Critical Care Chronicles, 11*(48), 31–37.

Ye, C., & Prince, J. L. (2016). A Bayesian approach to fiber orientation estimation guided by volumetric tract segmentation. *Computerized Medical Imaging and Graphics, 54*, 35–47.

Yeh, J.-Y., Wu, T.-H., & Tsai, W.-J. (2014). Bleeding and ulcer detection using wireless capsule endoscopy images. *Journal of Software Engineering and Applications, 7*(05), 422.

Yogapriya, J., Chandran, V., Sumithra, M. G., Anitha, P., Jenopaul, P., Dhas, S. G., & C.. (2021). Gastrointestinal tract disease classification from wireless endoscopy images using pretrained deep learning model. *Computational and Mathematical Methods in Medicine, 2021*. https://doi.org/10.1155/2021/5940433

Yuan, Y., Wang, J., Li, B., & Meng, M.-Q.-H. (2015). Saliency based ulcer detection for wireless capsule endoscopy diagnosis. *IEEE Transactions on Medical Imaging, 34*(10), 2046–2057. https://doi.org/10.1109/TMI.2015.2418534

Zhao, C., Shuai, R., Ma, L., Liu, W., Hu, D., & Wu, M. (2021). Dermoscopy image classification based on StyleGAN and DenseNet201. *IEEE Access, 9*, 8659–8679.