J. Inst. Eng. India Ser. B
https://doi.org/10.1007/s40031-024-01099-8

ORIGINAL CONTRIBUTION

Severity Grading of Ulcerative Colitis Using Endoscopy Images: An Ensembled Deep Learning and Transfer Learning Approach

Subhashree Mohapatra[1] · Pukhraj Singh Jeji[2] · Girish Kumar Pati[2] · Janmenjoy Nayak[3] · Manohar Mishra[4] · Tripti Swarnkar[5]

Received: 14 January 2024 / Accepted: 17 June 2024

Abstract Ulcerative colitis (UC) is a persistent condition necessitating prompt treatment to avert potential complications. Detecting UC severity aids treatment decisions. The Mayo-endoscopic subscore is a standard for UC severity grading (UCSG). Deep learning (DL) and transfer learning (TL) have enhanced severity grading, but ensemble learning's impact remains unexplored. This study designed DL-ensemble and TL-ensemble models for UCSG. Using the HyperKvasir dataset, we classified UCSG into two stages: initial and advanced. Three deep convolutional neural networks were trained from scratch for DL, and three pre-trained networks were trained for TL. UCSG was conducted using a majority voting ensemble scheme. A detailed comparative analysis evaluated individual networks. It is observed that TL models perform better than the DL models, and implementation of ensemble learning enhances the performance of both DL and TL models. Following a comprehensive assessment, it is observed that the TL-ensemble model has delivered the optimal outcome, boasting an accuracy of 90.58% and a MCC of 0.7624. This study highlights the efficacy of our methodology. TL-ensemble models, especially, excelled, providing valuable insights into automatic UCSG systems' potential enhancement. Ensemble learning offers promise for enhancing accuracy and reliability in UCSG, with implications for future research in this field.

Keywords Ulcerative colitis · Deep learning · Transfer learning · Severity · Ensemble learning · Endoscopic

Introduction

Colorectal cancer is a malignant condition occurring in the gastrointestinal tract and specifically in the large intestine [1]. Colorectal cancer stands as a significant apprehension in the present era, having secured the second position globally in terms of mortality rates [1]. Additionally, it ranks as

✉ Manohar Mishra
manohar2006mishra@gmail.com

Subhashree Mohapatra
subhashreemohapatra68@gmail.com

Pukhraj Singh Jeji
pukhrajsingh93@gmail.com

Girish Kumar Pati
pati.drgirishkumar@gmail.com

Janmenjoy Nayak
mailforjnayak@gmail.com

Tripti Swarnkar
swarnkar.tripti@gmail.com

[1] Department of Computer Science and Engineering, Siksha O Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

[2] Department of Gastroenterology, IMS and SUM Hospital, Siksha O Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

[3] Department of Computer Science, Maharaja Sriram Chandra Bhanja Deo University, Baripada, Odisha, India

[4] Department of Electrical and Electronics, Siksha O Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

[5] Department of Computer Application, National Institute of Technology (NIT), Raipur, India

Published online: 25 June 2024

the third most prevalent cancer in both males and females [2]. Figure 1 illustrates the global occurrences of colon cancer, a subtype of colorectal cancer [3]. A fatter diet and a less fiber diet is a major risk factors for colon cancer. UC is a long-term health condition characterized by recurrent sores and irritation, predominantly affecting the lower part of the digestive system. It is restricted to the mucosal and submucosal lining of the colorectal region. The course of the disease is long which easily relapses mainly showing the symptoms of diarrhea, stool associated with mucus and pus, abdominal cramps, fever as well as nausea. Studies show that if ulcerative colitis is left untreated for a long time, then it can lead to the initiation of cancer in the affected area, and it is one of the high-risk factors for colon cancer. Glandular cancer [2] of the colon is mostly observed in humans between the age of 60–79, but when colorectal carcinoma is diagnosed in a young mass then the initiation factor is majorly ulcerative colitis or polyps.

Endoscopy, particularly colonoscopy, serves as a crucial medical procedure in the identification and diagnosis of UC [4]. To evaluate the effect of therapeutics the severity of ulcerative colitis is graded. A popular grading technique to scale the severity of the disease is the Mayo score or also termed as Mayo Endoscopic Subscore (MES). The severity grade is calculated based on four different factors such as rectal bleeding, frequency of stool, endoscopic evidence, and expert assessment. The score of severity ranges from 0 to 3, where '0' represents the initial stage and 3 indicates the advanced stage of ulcerative colitis. The evaluation of MES from the endoscopic images requires the expertise of experienced gastroenterologists [5]. Observing the minute difference in the inflammation site of the disease needs the time of professionals and can be subjective in the medical scenario. In this case, if a system could be designed to score the severity of the disease, then that could be a very beneficial and great assistance to the doctors.

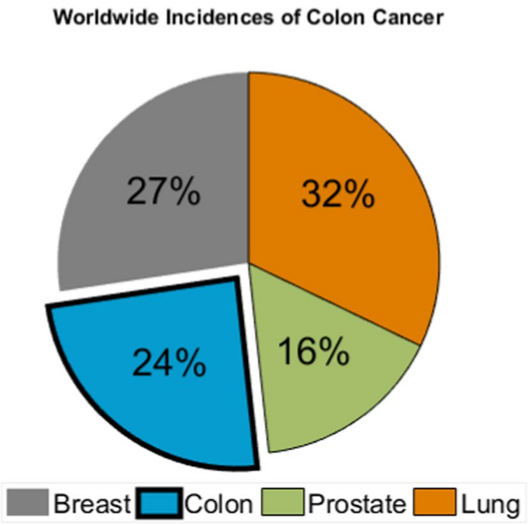


Fig. 1 Chart illustrating the global incidences of colon cancer

In recent times, artificial intelligence (AI) has emerged as a pivotal contributor in analyzing images within the medical domain [6]. Techniques within the realm of AI (such as machine learning and deep learning), encompassing various computational approaches, are becoming more widespread in systems designed to assist in computer-aided design processes. These technologies play an important role not only in detecting abnormalities from medical images but also in assisting healthcare professionals in delivering accurate diagnosis results [7]. Within deep learning (DL), the network is crafted and trained anew for the classification of given input images. The convolutional neural network (CNN) stands out as a DL method capable of autonomously learning features and categorizing images [8]. These network architectures have demonstrated superior performance compared to machine learning techniques, especially when applied to the analysis of medical images. A notable limitation of the DL technique is its dependency on large datasets to yield effective classification results. Additionally, the implementation of DL demands substantial computational power, constituting another significant drawback [9].

To address the previously noted limitations of deep learning, an alternative concept known as transfer learning (TL) has gained prominence, particularly in the context of classifying medical images. TL uses the parameters of a pre-trained network that already trained for a particular dataset, to classify any given new dataset. This concept is much helpful when the dataset size is less and computational power is low [10]. Again, to elevate the final output of the mentioned techniques another approach known as ensemble learning (EL) has garnered attention. Ensemble learning-based classification is an approach where the outcome of various networks is considered to give a final prediction result over the given dataset. In EL, there are several methods in which the result of different classifiers is combined to give a concluding prediction [11]. Even though DL and TL have gained much popularity in gastrointestinal disease detection and more specifically grading the severity of UC, utility of EL is unexplored. So, to evaluate each of the learning methodologies and to design an effective system to grade the UC severity according to the MES scale, authors are motivated to design different models based on these concepts. If an efficient system can be proposed to provide a proper severity score for the endoscopy images, then it can be of great assistance to gastroenterologists. This will in turn help the professional to deliver proper and timely treatment to patients affected by UC and avoid future complications.

The Prior State of Work

Numerous studies in the literature demonstrate the automatic severity grading of UC. Several authors have collected the data from different clinical centers to carry out the experimental study whereas others have used the standard data that are publicly available. Here, a summarized review is presented on the previous art for ulcerative colitis detection and grading.

Huang et al. [5] have designed a CAD system for the diagnosis of mucosal inflammation in the colon of the patients who suffer from ulcerative colitis. The model is developed using both deep learning (DL) and machine learning (ML). Its role involves extracting features and classifying the disease, respectively. The authors have gathered 856 images related to the gastrointestinal condition from 54 patients diagnosed with UC, sourced from a Taiwan based hospital. In this process, the images are inputted into a pre-trained DL model, which extracts features. Subsequently, these extracted features are employed to train three ML learners, namely the deep neural network (DNN), support vector machine, and k-nearest neighbor. Here, the final result is obtained by using ensemble learning through voting of the outcome generated by the machine learning classifiers. The system achieved a success rate of 94.5% for endoscopic evaluation of mucosal improvement and an 89.0% success rate for achieving full mucosal recovery. Stidham et al. [6] focused on comparing the ability of the DL model over experienced professionals in detecting the severity of ulcerative colitis. The dataset used consists of 16,514 colonoscopy images of 3082 patients having ulcerative colitis from a sole specialized referral center in the USA. A 159-layer deep CNN is trained to categorize the images into remission and moderate/severe disease. The designed model categorized the images into a proper group with 0.966 AUROC, 0.87 Positive Prediction Value (PPV), 83.0% sensitivity, 96.0% specificity, and 0.94 Negative Prediction Value (NPV). Here the deep network performed at par to the experienced doctors.

Bhambhvani and Zamora [4] proposed an approach to investigate the strength of a DNN to classify the UC image according to the MES severity scale. Here, 777 endoscopic images collected from 777 individual patients were used for the study. An experienced physician and a junior medical practitioner annotated the images with MES scores of 3, 2, or 1. The results indicate that ResNet-101 yielded the most favorable outcome, achieving an overall accuracy of 77.2%. Sutton et al. [12] carried out two binary classification models. In the first part of the work, UC and non-UC endoscopic images were classified. And in the second part of the work, only the UC images were classified into two groups that are inactive/mild and moderate/severe. In this investigation, the authors utilize the publicly accessible HyperKvasir dataset. The weights of the CNN models were initialized using ImageNet and the best hyperparameter were identified using the Grid Search technique using fivefold cross-validation. Within this study, the DenseNet121 pre-trained network achieved the highest classification accuracy at 87.50% and an AUC of 0.9.

Yao et al. [13] used CNN for the automatic grading of the UC severity and predict the image informativeness. The proposed classification learner used for still-image informativeness generated a high performance of 0.902 sensitivity and 0.870 specificities. For the high-resolution videos, the automated models rightly classified MESs in 78% of videos. Becker et al. [14] have proposed a fully automated model using DL for predicting MES from colonoscopy videos. The study utilized 1672 videos sourced from a multisite dataset derived from the etrolizumab phase 3 clinical program. The designed model achieved high accuracy and AUROC for different sub-score. Authors in [15] have suggested a deep CNN-based CAD system considering the GoogLeNet architecture. The training of the learner involved utilizing 26,304 colonoscopy-based imageries obtained from 841 patients diagnosed with UC. Here, the proposed DL approach demonstrated impressive AUROCs, achieving a score of 0.86 for Mayo 0 and 0.98 for Mayo 0–1. Authors in [16], have designed a CNN for classifying the UC severity grade using colonoscopy videos. Authors have experimented by finding the pattern of veins that is specified by the number of blood vessels in frames. Tejaswini et al. [17] have proposed an improved version of CNN model by adding significant image pre-processing to their previous work [16] to classify the UC according to severity. In this work, the authors have divided the categories of UC severity into subcategories, creating a more refined classification system.

As observed from the stated art of work in the field of classifying UC images according to the severity grading, the majority of study is carried out using DL or TL approach. However as per our knowledge no study has been done using the EL approach for severity grading of UC. Whereas, EL has been proved to be a better performer in various other classification problems [8, 18–22]. Again, most of the study is done using colonoscopy data from hospitals collected over some time. Very few works have been done considering the standard dataset available publicly is the HyperKvasir dataset [23]. This motivated the authors to grade the severity of UC using DL, TL, and EL approaches. Here a two-level comparison is done where in the first level DL approach is compared with TL and in the second level EL is implemented to observe the improvements in performance from DL and TL individually.

Motivation and Contribution

The literature study mentioned above offers an extensive analysis of recent methodologies proposed for the severity grading of ulcerative colitis (UC). The study emphasizes and discusses the major outcomes and research gaps. In response to the identified implications, a novel model is crafted, incorporating a DL module, a Transfer Learning (TL) module, and subsequently an Ensemble Learning (EL) module.

The main contribution of the study is highlighted below –

(a) Authors have designed a two-step experimental and comparative study comprising various modules using the proposed DL-ensemble approach and TL-ensemble approach for the classification of UC according to the severity.

(b) The initial phase incorporates an image pre-processing unit where the input imageries are resized in order to meet the training model's specifications, and augmentation techniques are applied to expand the training data size.

(c) In the first step, another module is designed using DL models where three individual CNN models are constructed and trained from scratch for the binary classification of UC. In addition to this one TL module is designed where three pre-trained CNN models named GoogleNet, ShuffleNet, and ResNet have been fine-tuned and trained for binary classification of UC.

(d) In this first step, a comparative analysis based on the classification performance is performed between DL and TL modules.

(e) In the second step of work, an EL module is designed using the majority voting ensemble method which is applied to the DL unit as well as the TL unit separately. Here a comparative analysis is carried out to show the performance improvement of both the DL and TL techniques by the implementation of EL.

Here in this work, authors hypothesize that TL generates better classification results than DL for which TL can be preferable for medical image analysis. Also, EL can help in increasing the performance of both DL and TL. Hence a step-by-step study is carried out to show the comparative analysis of all the techniques and give a conclusive outcome.

The organization of this paper is further enhanced through systematic structuring. In Sect. ”Materials and Methodology”, a detailed exploration of the dataset is presented accompanied by a discussion of the planned approach. Subsequently, Sect. ”Result Analysis” delves into the analysis of results, offering insights derived from the study. Critical discussions on the findings are then thoroughly examined in Sect. ”Critical Discussion”. Finally, Sect. ”Conclusion” encapsulates the concluding statements and introduces avenues for future research.

Materials and Methodology

In the study, authors have stated a methodology where two models have been designed named DL-ensemble and TL-ensemble models aiming to classify the endoscopy images of UC disease into two classes according to the severity grading as initial stage UC or advance stage UC. The entire experimentation and evaluation process are divided into different units, elaborated upon in this section. Figure 2 provides a visual representation of the methodology.

Dataset

Various datasets are available publicly which is very helpful for research based on gastrointestinal disease detection. In this work, one of such datasets that is the HyperKvasir dataset [23] is used to explore the ability of recent computational techniques. HyperKvasir is a large repository of endoscopy images which is a recent edition of the Kvasir dataset [24]. The images are collected over 8 years from 2008 to 2016 at Bærum Hospital. HyperKvasir dataset has endoscopy images belonging to both the upper GI and lower GI parts consisting of 110,079 images out of which 10,662 images are correctly labeled. These images belong to 23 different classes of findings. Figure 3 gives a clear picture of the whole HyperKvasir dataset. Given that the primary objective of this study is to assess the UCSG, the authors have exclusively selected endoscopy images corresponding to UC for analysis. The UC section consists of 851 images which are categorized according to Mayo Endoscopic Subscore (MES) and fit into 6 different classes. Table 1 gives an idea about the images specific to various classes of UC grading in the HyperKvasir dataset.

To fulfill the requirement of the study the six classes of UC images have been merged to carry out a binary classification task. Images from UC-grade-0-1 and UC-grade-1 are grouped into one class named “initial stage UC” whereas UC-grade-1-2, UC-grade-2, UC-grade-2-3, and UC-grade-3 are grouped into another class named “advanced stage UC”. The amalgamation of the classes is done with concern by experienced gastroenterologists. Based on expert opinion, grade-2 and grade-3 images are considered potentially confusing during classification, posing a potential impact on model performance. Hence, a binary classification of UC is adopted, consisting of 236 images in the initial stage and 615 images in the advanced stage. Among these images, 70% are allocated for training the model, 20% for validation, and the remaining 10% for testing the model.

Image Pre-Processing

N this study, the pre-processing section involves two steps. The initial step is image resizing, and the subsequent step

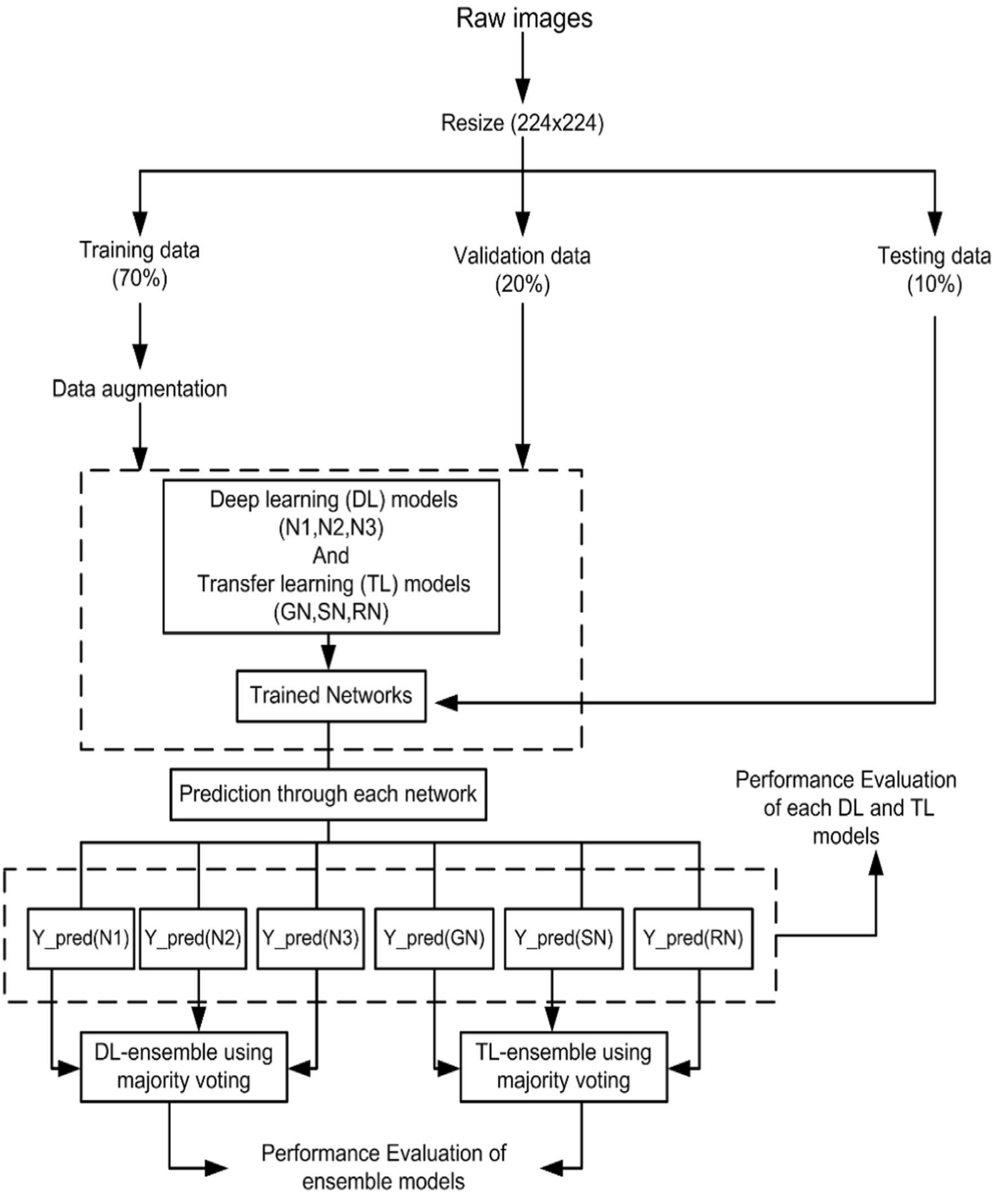

Raw images
Resize (224x224)
Training data (70%)
Validation data (20%)
Testing data (10%)
Data augmentation
Deep learning (DL) models (N1,N2,N3) And Transfer learning (TL) models (GN,SN,RN)
Trained Networks
Prediction through each network
Performance Evaluation of each DL and TL models
Y_pred(N1)
Y_pred(N2)
Y_pred(N3)
Y_pred(GN)
Y_pred(SN)
Y_pred(RN)
DL-ensemble using majority voting
TL-ensemble using majority voting
Performance Evaluation of ensemble models

Fig. 2 Workflow of proposed DL-ensemble and TL-ensemble system

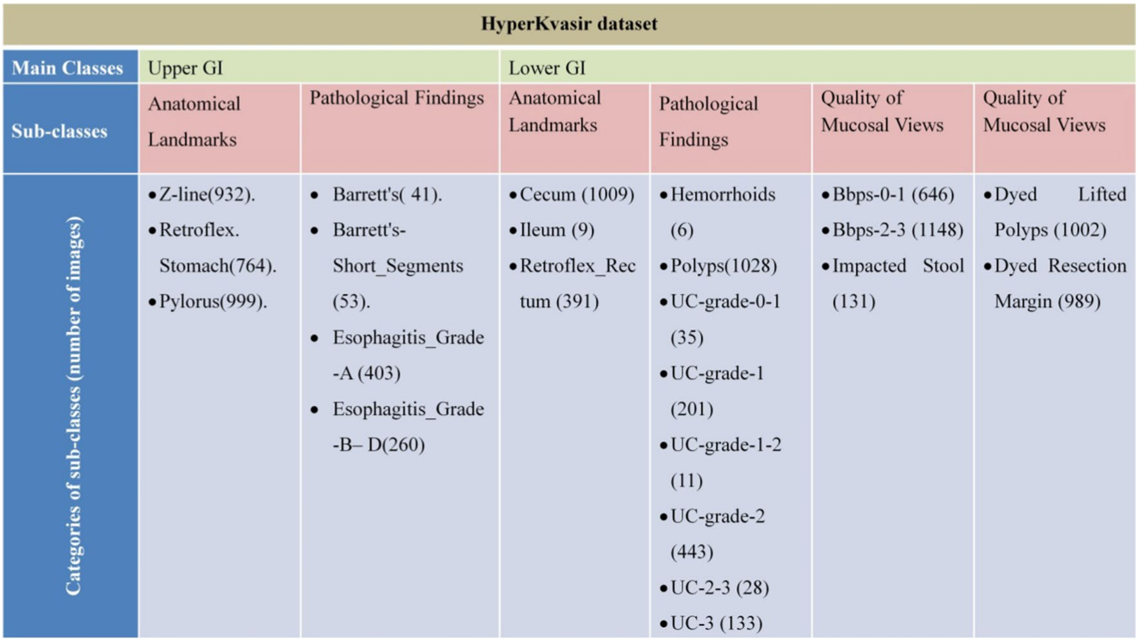
involves augmenting the training set of images. The UC images within the HyperKvasir dataset exhibit diverse resolutions ranging from 576 × 720 to 1072 × 1920. All original UC images are uniformly resized to 224 × 224 pixels. This specific dimension is chosen as the pre-trained models employed for transfer learning are trained on images with dimensions of 224 × 224 pixels. Consequently, the size remains constant during the training of DL models. In the subsequent phase, 70% of the images utilized for model training undergo augmentation to augment the dataset. DL models, particularly CNNs, yield improved outcomes when trained on a substantial dataset of images. Image augmentation is implemented by employing different approaches, including rotation, reflection, shearing, and scaling. After the augmentation is done, the initial stage UC consists of 1650 images, and the advanced stage UC has 4310 images.

Authors have chosen this particular sample size after several round of training using different size of data less than the finalized one. With less sample size the model outputs were not very effective. Table 2 gives an idea about the division of the dataset.

Deep Learning

DL is a subgroup of AI characterized by its deep-layered structure, which consists of multiple linear and non-linear processing modules. In recent times several DL methodologies have gained popularity in various domains of research. There are different DL algorithms such as autoencoder, Boltzman machines, deep belief networks, RNNs, and CNN. Out of these DL techniques, CNN has a wide area of application in the domain of image classification and more specifically medical image processing [25].

The overarching concept of CNN involves taking a raw image as input and, through several layers of processing; the model produces the ultimate class designation for the



HyperKvasir dataset						
Main Classes	Upper GI		Lower GI			
Sub-classes	Anatomical Landmarks	Pathological Findings	Anatomical Landmarks	Pathological Findings	Quality of Mucosal Views	Quality of Mucosal Views
Categories of sub-classes (number of images)	• Z-line(932). • Retroflex. Stomach(764). • Pylorus(999).	• Barrett's(41). • Barrett's-Short_Segments (53). • Esophagitis_Grade-A (403) • Esophagitis_Grade-B– D(260)	• Cecum (1009) • Ileum (9) • Retroflex_Rectum (391)	• Hemorrhoids (6) • Polyps(1028) • UC-grade-0-1 (35) • UC-grade-1 (201) • UC-grade-1-2 (11) • UC-grade-2 (443) • UC-2-3 (28) • UC-3 (133)	• Bbps-0-1 (646) • Bbps-2-3 (1148) • Impacted Stool (131)	• Dyed Lifted Polyps (1002) • Dyed Resection Margin (989)

Fig. 3 Comprehensive Overview of the HyperKvasir Dataset Contents

Table 1 The content of the complete ulcerative colitis images

Mayo grading	Findings	No. of images	Merged class name
UC-grade-0-1	It May belong to grade 0 or grade 1	35	Initial stage UC (236 images)
UC-grade-1	Few erosions, decrease in vascular patterns	201	
UC-grade-1-2	It May belong to grade 1 or grade 2	11	Advance stage UC (615 images)
UC-grade-2	Multiple confluent ulcers, multiple erosions, minimal vascular patterns	443	
UC-grade-2-3	It May belong to grade 2 or grade 3	28	
UC-grade-3	Multiple large ulcers, multiple erosions, bleeding with minimal pressure by the colonoscope	133	

Table 2 Division of the dataset

Grade of UC	#Images for training	#Images for training after augmentation	#Images for validation	#Images for testing
Initial stage UC (236 images)	165	1650	47	24
Advance stage UC (615 images)	431	4310	123	61

provided image. The convolution unit stands out as the most crucial component in the deep network. The convolutional unit facilitates the creation of a feature map for images through the application of a convolutional operation (CO). The CO represents a modest dot-product involving the input feature map at a precise spatial coordinate (x, y) and the 2D kernel given by Eq. (1).

$$CO_{(x,y)} = \sum_{i} a_i b_i \quad (1)$$

Here, a_i signifies the kernel weight and b_i signifies the input matrix from the preceding layer. The final output (FO) of the convolutional unit is achieved by tallying a bias 'B' into the result obtained by Eq. (1). The result is evaluated using Eq. (2).

$$FO_{(x,y)} = CO_{(x,y)} + B \quad (2)$$

In addition to the convolution unit, an activation function is incorporated to ensure a non-linear correlation in the resultant map. The most used activation function for a CNN network is the Rectified Linear Unit (ReLU), represented by Eq. (3) for a given input 'I'.

$$f(I) = \begin{cases} 0, ifI < 0 \\ I, ifI \geq 0 \end{cases} \quad (3)$$

This combination of convolutional units with ReLU includes subsequent max-pooling layers. The max-pooling layers aids to choose the most appropriate indigenous features that are generated by the convolutional unit. The entire network concludes with a fully connected layer and a softmax layer. The fully connected layer plays a role in aggregating the attributes cultured by the DNN, while the softmax layer provides the final prediction as a probability value. Figure 4 illustrates a generalized CNN architecture [26].

In this study, the authors aim to assess the efficacy of ensemble learning. To introduce an ensemble model, the authors have conceived and trained three distinct CNN models from the ground up. Two of the networks are self-designed networks and the third network that is designed is a ResNet-like architecture that is trained from scratch. The selection of a ResNet-like model is done based on the work proposed by the authors previously in [27]. In [27] authors have evaluated that out of three architectures taken for transfer learning ResNet has given the best performance. Hence, to evaluate a similar model the ResNet-like architecture is constructed and used as the third self-designed deep CNN model (N3). The details of self-designed models that are Network-1 (N1) and Network-2 (N2) are detailed in Tables 3 and 4, correspondingly.

Transfer Learning

Nowadays deep learning (DL) is recognized as an effective technique for the detection of disease conditions using colonoscopy images, moving ahead of the traditional machine learning methodologies. Among various DL algorithms, deep CNN is well known for the medical image classification task. The inbuilt modules for automated processing (feature extraction and classification) are the main strengths of CNN models, making them more preferable than manual techniques. Even though it has a major advantage, training a deep CNN model from scrap requires more computational power as well as more time for execution. Another requirement for a good-performing CNN model is that the sample size of the annotated dataset should be large which is quite difficult in the medical domain. Due to this shortfall of deep

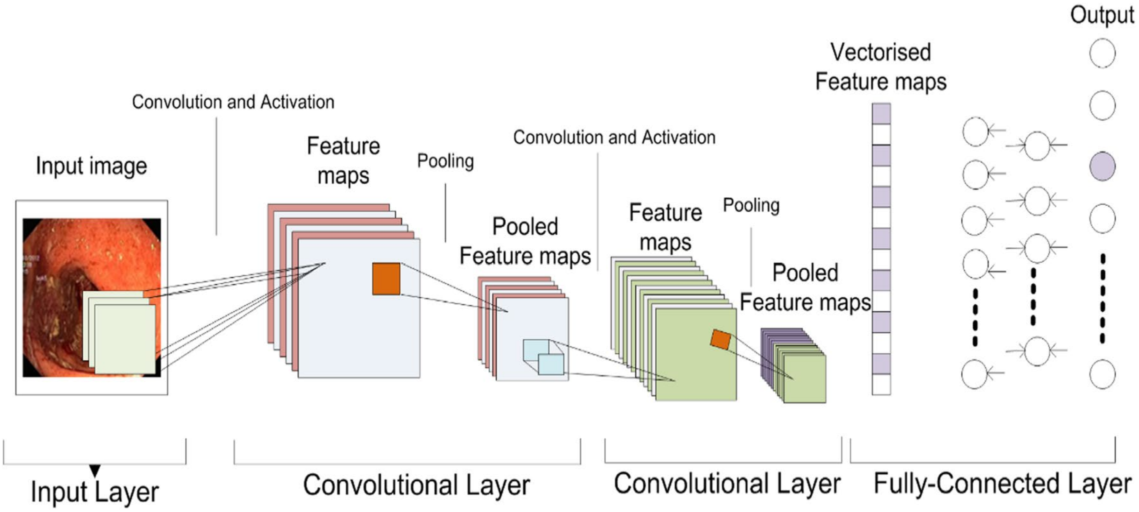


Fig. 4 Generalized CNN architecture

Table 3 Architectural overview of network-1

Layer	Filter size	Number of filters	Output: (width × height × channels)
Input layer	Input image size		112 × 112 × 3
Convolutional layer 1	3 × 3	32	112 × 112 × 32
Max-pooling layer 1	2 × 2	–	56 × 56 × 32
Convolutional layer 2	3 × 3	64	56 × 56 × 64
Max-pooling layer 2	2 × 2	–	28 × 28 × 64
Convolutional layer 3	3 × 3	128	28 × 28 × 128
Max-pooling layer 3	2 × 2	–	14 × 14 × 128
Convolutional layer 4	3 × 3	164	14 × 14 × 164
Max-pooling layer 4	2 × 2	–	7 × 7 × 164
Convolutional layer 5	3 × 3	–184	7 × 7 × 184
Max-pooling layer 5	2 × 2	–	3 × 3 × 184
Fully connected	–	–	1 × 1 × 1656
Softmax	–	–	1 × 2
Class output	–	–	Binary classes

Table 4 Architectural overview of network-2

Layer	Filter size	Number of filters	Output: (width × height × channels)
Input layer	Input image size		224 × 224 × 1
Convolutional layer 1	3 × 3	32	224 × 224 × 32
Convolutional layer2	3 × 3	32	224 × 224 × 32
Max-pooling layer 1	2 × 2		112 × 112 × 32
Convolutional layer 3	3 × 3	64	112 × 112 × 64
Convolutional layer 4	3 × 3	64	112 × 112 × 64
Max-pooling layer 2	2 × 2		56 × 56 × 64
Convolutional layer 5	3 × 3	128	56 × 56 × 128
Convolutional layer 6	3 × 3	128	56 × 56 × 128
Max-pooling layer 3	2 × 2		28 × 28 × 128
Convolutional layer 7	3 × 3	164	28 × 28 × 164
Convolutional layer 8	3 × 3	164	28 × 28 × 164
Max-pooling layer 4	2 × 2		14 × 14 × 164
Convolutional layer 9	3 × 3	184	14 × 14 × 184
Convolutional layer 10	3 × 3	184	14 × 14 × 184
Max-pooling layer 5	2 × 2		7 × 7 × 184
Convolutional layer 11	3 × 3	256	7 × 7 × 256
Convolutional layer 12	3 × 3	256	7 × 7 × 256
Max-pooling layer 6	2 × 2		3 × 3 × 256
Convolutional layer 13	3 × 3	512	3 × 3 × 512
Convolutional layer 14	3 × 3	512	3 × 3 × 512
Max-pooling layer 7	2 × 2		1 × 1 × 512
Two fully connected layers			1 × 1 × 512
Softmax			1 × 2
Class output			Binary classes

networks, another technique that is gaining attention is transfer learning technology [27].

TL is a method in which previously designed and trained deep CNN models are fine-tuned and used to classify their data of interest. The pre-trained models are initially trained on a huge image dataset of some other domain. Figure 5 represents the general working of transfer learning [28]. The name itself signifies the working process. In the process of TL, knowledge is transferred in the form of weights of the network. When training a deep CNN model from scratch, performance enhancement is achieved through iterative optimization of network weights, leading to increased evaluation time. This decrease is significant when the fine-tuning process involves transferring the optimized weights from pre-trained CNN models during training. In return, the transfer of pre-trained network weight adds to the improvement of the classification performance [29].

In this study which aims to grade the severity of the UC images, the authors have chosen three different pre-trained networks namely GoogleNet (GN) [30], ResNet-18 (RN) [31], and ShuffleNet (SN) [32]. As per the available literature, various research endeavors have explored the application of transfer learning in medical image analysis. And among the multiple pre-trained networks present, GoogleNet and ResNet have outperformed the other networks specifically for the disease detection task. On the other hand, the strength of ShuffleNet is not much explored for the task of medical image classification. Hence, authors have preferred to experiment with these networks. During the TL, the initial layer of the pre-trained deep networks is kept unchanged and the final classification layers are substituted with a new layer to fulfill the requirement of the study. For grading the UC severity, the last three layers of individual networks are substituted with the "fully connected layer", "softmax layer" and "classification output layer". The details of the three pre-trained networks are discussed in this section.

GoogleNet

GoogleNet [30], also known as Inception V1, is a widely embraced network for TL, not only in the realm of medical image classification but also in various research domains [33]. Comprising 27 layers, it includes the input layer, convolutional layer, inception layer, pooling layer, fully connected layer, and softmax layer. The input layer receives colored images with dimensions of 224 × 224 pixels. The input layer connects to two large filters of size 7 × 7, effectively reducing the dimensions of the input visuals. Following convolution, there is max pooling, another convolution, and subsequent max pooling with a 3 × 3 filter. The output from the max pooling layer serves as input to a two-layer inception module block, followed by a 3 × 3 max pool, a four-layer inception module, another max pool, and a two-layer inception module. This sequence is succeeded by a final max pooling layer and a 7 × 7 average pooling layer. The concluding layers of the original GoogleNet are replaced with a new fully connected layer, a softmax layer, and then a classification output layer, providing the ultimate prediction result for UC grading.

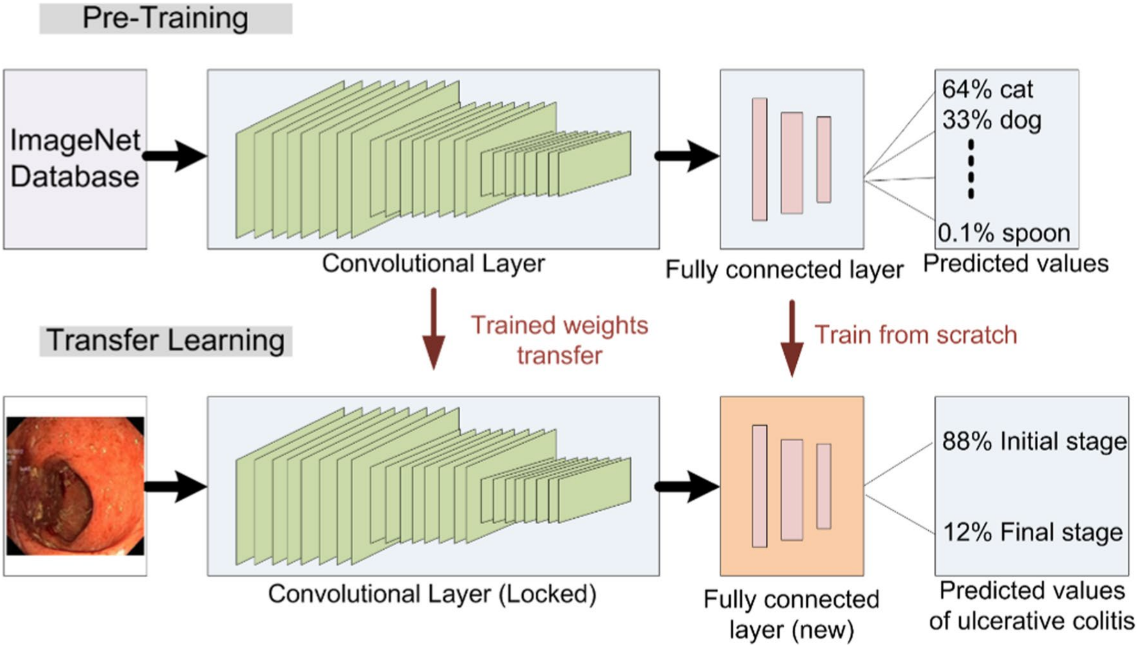


Fig. 5 Generalized transfer learning architecture

ResNet-18

ResNet-18 architecture also known as a stack of residual blocks, consists of 18 different layers of CNN units [34]. The network is pre-trained with the ImageNet database which comprises millions of images and can be efficiently used to classify any imagery dataset with around 1000 classes of samples. One of the issues of DNNs is the problem of vanishing gradient due to the execution of the backpropagation technique in the models. As the networks have several layers and implement a similar type of activation function, it leads to the loss of the gradient to zero. This results in ineffective training of the model. In their work [31], the authors introduced the skip connection concept in ResNet, aiming to alleviate the vanishing gradient problem. Skip connection is a typical unit in convolutional network design and this helps provide an alternate path for the gradient while executing the backpropagation algorithm. The additional routes are typically beneficial for convergence. By incorporating a skip connection, the output of the previous layer is not only fed into the immediate subsequent layer but also propagated to the subsequent layers in the network. This implementation ensures that features already captured in earlier layers are retained and not lost during the training process.

ShuffleNet

ShuffleNet is a pre-trained deep network architecture that is proposed by [32]. The main building blocks of ShuffleNet are the pointwise group convolution and channel shuffle which is helpful to achieve better classification accuracy. The importance of the channel shuffle is the shuffling of features in each of the channels that maximize the exchange of information among the channels and in turn, increase the efficiency of the network. ShuffleNet has a lightweight design and is less expensive than other networks.

Ensemble Learning (EL)

EL is a method that is designed around the 1990s to enhance weaker classifiers to stronger ones to obtain better classification output [35]. This approach involves initially training individual learning models, referred to as base learners, on a specific dataset. Subsequently, the results from the trained learners are amalgamated in diverse ways to assess the final outcome. Individual classifiers perform differently on a particular dataset, so it is difficult to say which model can give the best performance. Here EL aids in improving the classification performance by combining the result of several predicting models. The final output of the ensemble model does depend on the way the individual learning models are ensemble [36]. There are several methods by which EL can be implemented. The selection of methods for EL is contingent upon the outcomes generated by the underlying base learning models. The prediction value evaluated by the classifiers can be either a continuous value or some labelled class value depending on the classification task at hand. There are different techniques such as majority voting, max, average, posterior probability, or some sophisticated methods such as decision template which can be used as an ensemble method for continuous value results. But for the classifiers generating some labelled classes as result can use the majority voting technique for EL. In the case of majority voting, the maximum number of votes by the individual base models is considered as the final prediction class. Figure 6 shows the general idea of majority voting for EL. Assuming that labelled class results of the base learner model B_a are represented as a d-dimension binary vector as

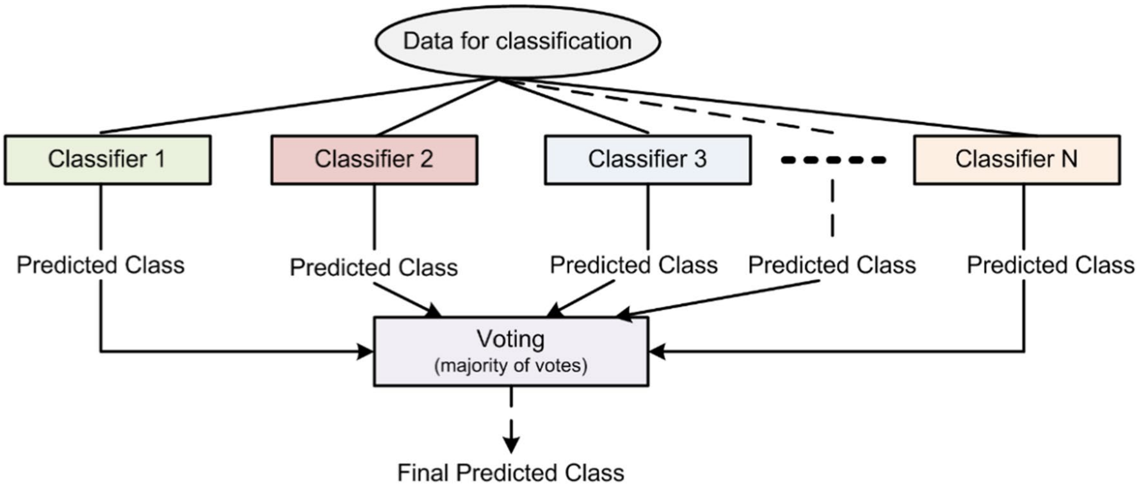


Fig. 6 Generalized architecture representing the majority voting scheme for ensemble learning

$$\left[b_{a,1}, \dots \dots, b_{a,d}\right] \in \{C1, C2\}^{d}, a = 1, \dots ., M \tag{4}$$

Here $b_{a,b}$=$C1$, if a base learner B_a label Y in X_b, and $C2$ otherwise. The majority voting method would generate an ensemble outcome for the class X_c, if the condition specified by Eq. (5) is satisfied:

$$\sum_{a=1}^{M} b_{a,c} = max_{b=1}^{d} \sum_{a=1}^{M} b_{a,b} \tag{5}$$

Performance Evaluation

In this work, six different deep CNN models are taken into consideration for grading the severity of UC. In this context, the performance of each network is assessed using various parameters. The evaluation metrics consist of classification accuracy (CA), recall (R), precision (P), F1-score (F), and specificity (S). As the dataset used for the experimentation is an unbalanced dataset so authors have considered another metric that is Matthew's Correlation Coefficient (MCC). As this is a binary class classification, the different parameters are evaluated using values which are true positive (T_p), true negative (T_n), false positive (F_p) also known as type-I error, and false negative (F_n) also known as type-II error. The confusion matrix structure is represented in Fig. 7 using which the different parameters are evaluated.

Classification accuracy represents how appropriately the model can predict the class label of given instances. Recall, also recognized as the true positive rate or sensitivity, denotes the percentage of predicted true values among all actual true values. In contrast, precision indicates the percentage of correctly predicted positive values among all correct predictions. The F1-score serves as the harmonic mean between precision and recall. Specificity, alternatively termed as the true negative rate or selectivity, signifies the percentage of predicted negative values among all actual negative values. MCC gives the correlation or strength of the statistical connection between two given variables. The result of MCC can range from [− 1, + 1], where − 1 represents no agreement between the true and predicted class output and + 1 represents the strongest agreement between the true and predicted result. All the parameters are evaluated using Eqs. (6–11).

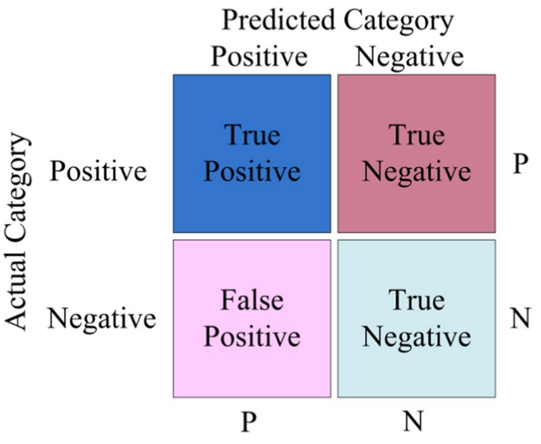


Fig. 7 General structure of the confusion matrix

$$CA = \frac{T_p + T_n}{T_p + F_p + T_n + F_n} \tag{6}$$

$$R = \frac{T_p}{T_p + F_n} \tag{7}$$

$$P = \frac{T_p}{T_p + F_p} \tag{8}$$

$$F = \frac{2 * R * P}{R + P} \tag{9}$$

$$S = \frac{T_n}{T_n + F_p} \tag{10}$$

$$MCC = \frac{\left[\left(T_p * T_n\right) - \left(F_p * F_n\right)\right]}{\sqrt{\left[\left(T_p + F_p\right)\left(T_p + F_n\right)\left(T_n + F_p\right)\left(T_n + F_n\right)\right]}} \tag{11}$$

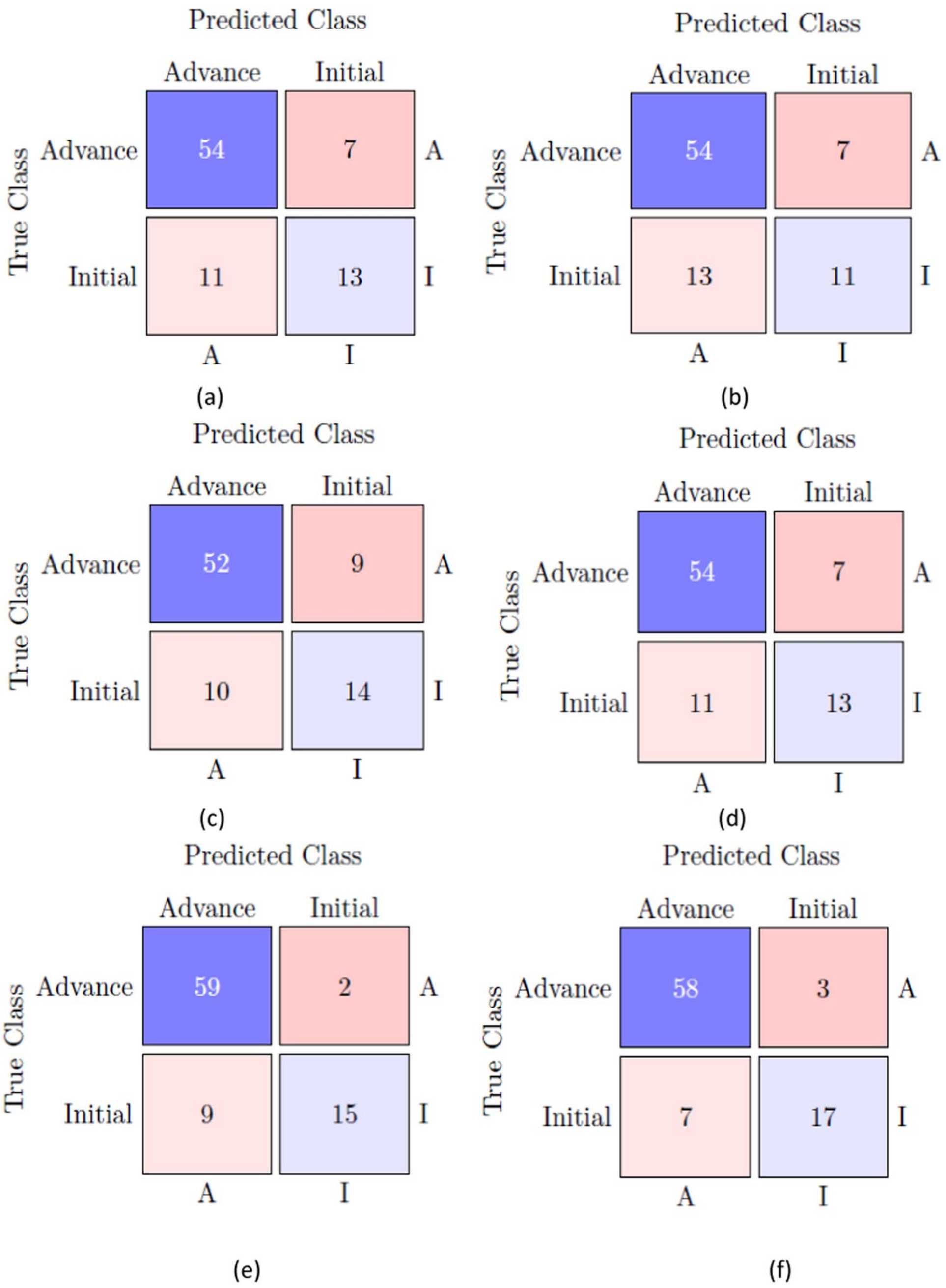
Result Analysis

In this study, all experimental assessments were performed using MATLAB 2022 on hardware consisting of an NVIDIA GeForce GTX 1650 GPU and an Intel Core-i7 processor running at 2.6 GHz. The tests are conducted on a system running Windows 10 with a total RAM capacity of 16 GB. The main motto of this evaluation is to grade the severity of UC using the HyperKvasir dataset considering the proposed DL-ensemble and TL-ensemble models. The dataset is partitioned into training, validation, and testing sets, maintaining a ratio of 7:2:1. The training set is employed for model training, the validation set is utilized to fine-tune the model parameters, and the testing set is reserved for the ultimate evaluation of the models. All the performance measures are calculated considering the testing set of data. Other parameters that are set during the training process are as follows: learning rate of 0.00010, 30 number of epochs, and Stochastic Gradient Descent with Momentum (SGDM) optimizer. The whole work is divided into two steps of evaluation.

In the first step, three deep CNN models and three pre-trained deep networks are trained and examined. Here a comparative investigation is done based on the task pertaining to the severity grading of UC, considering DL and TL techniques. In the second step of the assessment, the EL technique is applied for both the DL and TL models, and the improvement in the performance is analyzed. To demonstrate the effectiveness of the proposed model, a comparative study is conducted against state-of-the-art techniques that have already been developed for the severity grading of UC.

Results Based on DL and TL

The result discussed in this section corresponds to the performance evaluation of DL and TL methodologies for grading the severity of UC as an initial and advanced stage. As the outcome is predicted using EL, in the first stage three deep CNN models are trained from scratch and three pre-trained deep CNN models are fine-tuned and trained to get the individual network prediction of UC. The different parameters that are considered for evaluation of the networks are classification accuracy, recall, precision, specificity, F1-score, Matthews Correlation Coefficient, and the time



(a)

	Advance	Initial
Advance	54	7
Initial	11	13

(b)

	Advance	Initial
Advance	54	7
Initial	13	11

(c)

	Advance	Initial
Advance	52	9
Initial	10	14

(d)

	Advance	Initial
Advance	54	7
Initial	11	13

(e)

	Advance	Initial
Advance	59	2
Initial	9	15

(f)

	Advance	Initial
Advance	58	3
Initial	7	17

Fig. 8 Confusion matrix generated by **a** N1 **b** N2 **c** N3 **d** GN **e** SN and **f** RN for the test set of data

taken for training the network. The stated parameters are calculated by using the confusion matrices represented in Fig. 8 and the calculated values are revealed in different tables.

Table 5 shows the values of performance measures corresponding to the DL models. From the table, it is noted that N1 has the best performance where CA is 78.82% and MCC is 0.453. But the time taken by N1 for training is more than N3. N3 model which is a network resembling ResNet-18 architecture is trained in a minimum time of 27 min (mins) with a CA of 77.64% and MCC of 0.4416. Table 6 shows the values of performance measures corresponding to the TL models. Here it is observed that RN has the best results in grading the UC severity. RN gives a CA of 88.23% and MCC of 0.6995. RN also takes a minimum time of only 23 min 29 s (secs). Comparing the three pre-trained networks GN has the least CA of 78.82%, MCC of 0.453 and the time taken for training is 61 min 32 s. Comparing the results obtained by DL models and TL models shows that the pre-trained networks perform better than the networks trained from scratch. The results pertaining to the first step of evaluation are summarized below.

(a) Out of the three deep CNN models N1 generates the best result of 78.82% CA and 0.453 MCC. But if the training time is considered then N3 takes the least time of 27 min to train the network.
(b) Out of the three pre-trained models RN generated the best result of 88.23% CA, 0.6995 MCC and the network training time is also the least that is 23 min 29 s.
(c) Comparing the result of DL and TL models, all the pre-trained models have generated better results than the deep CNN models.
(d) Out of all the six models considered, RN has the best performance in terms of CA (88.23%), R (95.08%), P (89.23%), S (70.83%), F (92.06%), MCC (0.6995), and training time (23 min 29 s).
(e) The performance of RN in terms of CA is 11.94% more than the best performing deep CNN model which is N1. In terms of MCC also RN is 0.2465 more than the N1 network. Figure 9 illustrates the performance metrics for only the top-performing deep CNN network (N1) and the most effective pre-trained network (RN).

Table 5 Performance measures corresponding to the DL models

Network	CA	R	P	S	F	MCC	Time for training
N1	0.7882	0.8852	0.8307	0.5416	0.8570	0.453	40 min 20 s
N2	0.7647	0.8852	0.8059	0.4583	0.8436	0.3785	60 min 43 s
N3	0.7764	0.8524	0.8387	0.5833	0.8454	0.4416	27 min

Table 6 Performance measures corresponding to the TL models

Network	CA	R	P	S	F	MCC	Time for training
GN	0.7882	0.8852	0.8307	0.5416	0.8570	0.453	61 min 32 s
SN	0.8705	0.9672	0.8676	0.625	0.9146	0.6665	32 min 11 s
RN	0.8823	0.9508	0.8923	0.7083	0.9206	0.6995	23 min 29 s

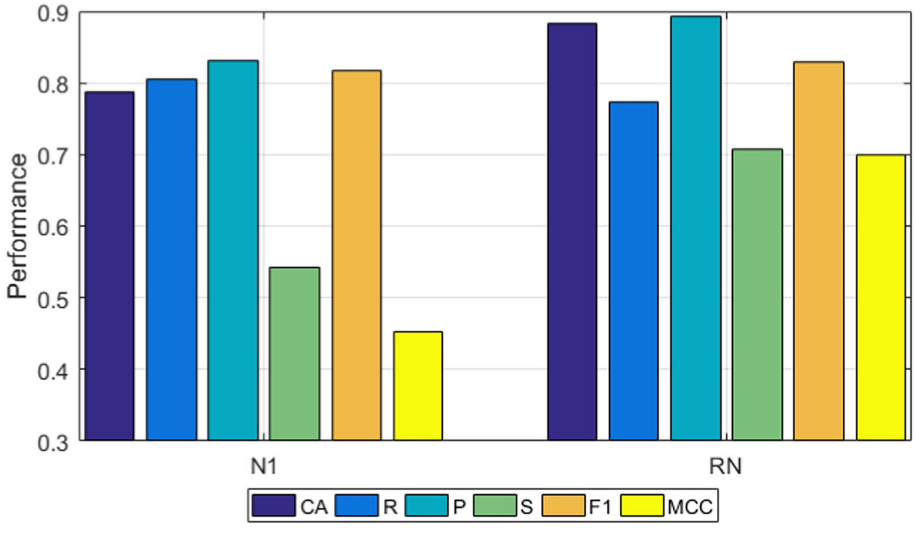


Fig. 9 Performance measures of only the best performing deep CNN network (N1) and best performing pre-trained network (RN)

(f) N3 model which is like RN network architecture takes very less time which is 27 min to train the model, near about the time taken by RN to train the network. But CA and MCC are much less than that RN. This justifies that any deep CNN model trained from scrap requires a large sample size of data to give a good result whereas, even with a small size of data, pre-trained model generates better outcome.

Figure 10 represents the performance graph generated during the training of the best performing DL model (N1) and the best performing TL model (RN). First part of each graph displays the accuracy values and the second part of the graph represents the loss values throughout the training of the network. From the graph it is observed that the training stability is achieved within 30 epochs, hence the quantity of 30 is chosen as the number of epochs for training the models.

Results Based on Ensemble Learning

In this work, the final prediction of the UC severity is done by implementing EL. Here the concept of majority voting is applied to the prediction outcome of both the DL (DL-ensemble) and TL (TL-ensemble) models. Table 7 illustrates the final improved result using majority voting EL. The result pertaining to the second step of the evaluation and a few important opinions from the table is summarized as mentioned.

(a) From Table 7 it is observed that combining the prediction outcome of the three individual deep CNN networks through majority voting has produced CA of 80%, R of 90.16%, P of 83.33%, S of 54.16%, F of 86.61% and MCC of 0.479. This has improved the CA by 1.49% as compared to the outcome of the N1 network.

(b) When the same majority voting scheme is implemented over the combined result of the three pre-trained models then it generates a CA of 90.58%, R of 98.36%, P of 89.55%, S of 70.83%, F of 93.74%, and MCC of 0.7624. This has elevated the CA by 2.66% when compared to the result of the RN network.

(c) Comparing the DL-ensemble and TL-ensemble model results it is observed TL-ensemble model generates a CA of 90.58% which is 13.23% more than the DL-ensemble model.

(d) The result justifies that the EL technique helps to improve the final prediction result of both DL as well as TL. However, the TL-ensemble model has generated the best result in the whole experimental evaluation. Figure 11 shows different performance measures with respect to the best performing DL model (N1), best performing TL model (RN), DL-ensemble model, and TL-ensemble model.

Critical Discussion

This section provides a comparative analysis of the proposed methodology with previous state-of-the-art approaches for grading the severity of ulcerative colitis (UC) using endoscopy images.

Results Compared to Prior State-of-the-Art

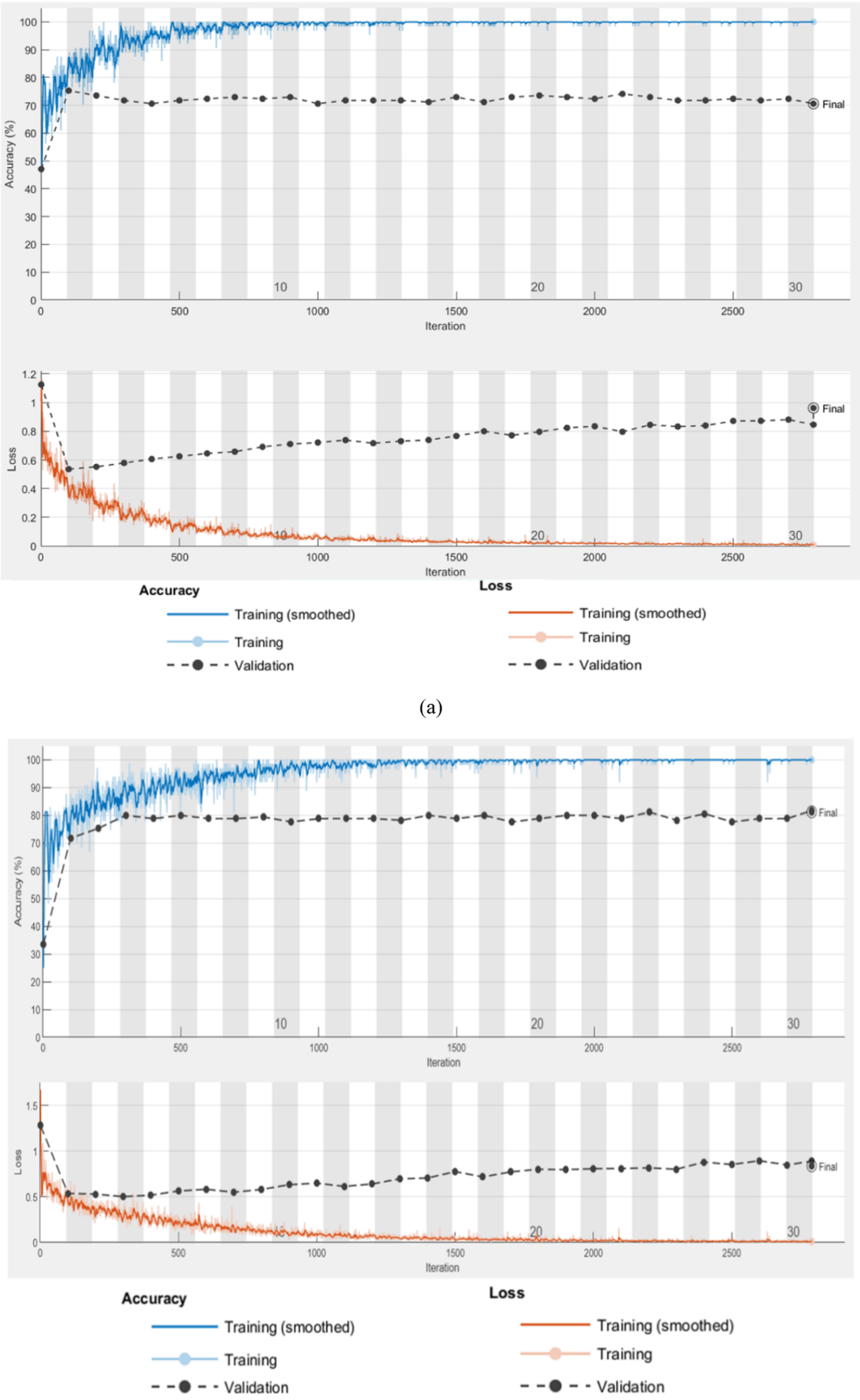
In this part of the study, a comparative analysis is carried out between the prior state-of-the-art and the proposed methodology. All the prior work is selected based on the task at hand which is grading the severity of UC. But the techniques, evaluation environment, and dataset used in the previous works are different from that considered in the proposed framework. So, to compare the result of the stated study and the earlier work a detailed description is mentioned in form of a matrix. Table 8 gives detailed information about each of the studies chosen for the comparative task. Each work is described particularly based on the objective of the research proposal, the dataset used for the experimental evaluation, the technique used, and the outcome of the learning. Some of the concluding remarks from the table are mentioned below.

(a) The objective of every work is to grade the severity of UC. Some of the works are a binary classification and a few are multiclass classification. In the proposed study authors have taken up a binary class classification to grade the UC as the initial stage or advance stage.

(b) Most of the work has either used a dataset that is directly collected from medical organizations and annotated by experienced reviewers or considered using the standard dataset that is publicly available the HyperKvasir dataset. But it is noted that none of the work has used the complete set of UC images from the HyperKvasir dataset. The proposed work has considered the whole UC data from HyperKvasir for the experimental work.

(c) All the prior studies mentioned have either used the DL technique or TL technique to classify the UC images. But none of the authors have worked with the EL. Here EL technique is used to show the improvement of both the DL as well as TL methodology.

(d) There are different parameters used by the authors for evaluating the grading models. But considering the classification accuracy, the proposed model has given the best result of 90.58% through the TL-ensemble model.



(a)

(b)

Fig. 10 Training accuracy and loss graph for **a** N1 and **b** RN network

Table 7 Performance measures corresponding to the DL-ensemble and TL-ensemble models

Network	CA	R	P	S	F	MCC
DL-ensemble	0.8	0.9016	0.8333	0.5416	0.8661	0.479
TL-ensemble	0.9058	0.9836	0.8955	0.7083	0.9374	0.7624

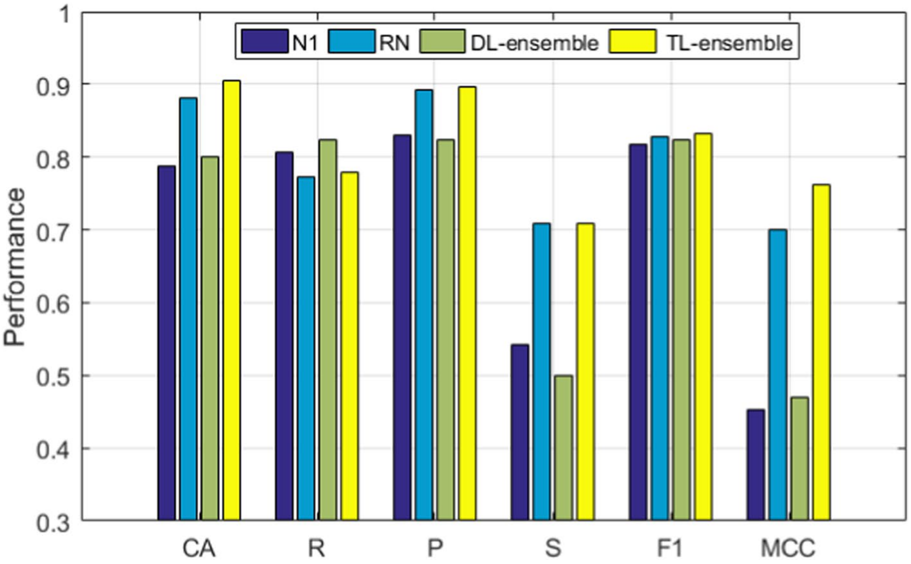


Fig. 11 Different performance measures with respect to best performing DL model, best performing TL model, DL-ensemble model, and TL-ensemble model

In this work, the researchers have suggested a DL-ensemble model as well as a TL-ensemble model to classify the UC images according to the severity grading as the initial stage or advance stage UC. It is very well justified that EL has helped to enhance the classification performance as compared to either DL or TL. Even though the work has achieved an efficient outcome, there is still room for enhancement. Here the experimental evaluation is carried out considering the publicly available dataset which has a small sample size, but it can be more effective when the model will be evaluated on real-time data and a bigger sample size data. Secondly, the model is designed considering two classes of severity. But as it is known that UC severity is graded in four classes, a system can be developed for multiclass classification of UC. And finally, if the models will be implemented in a real-time scenario, then it can be of great assistance to gastroenterologists in detecting the severity level of the disease by which proper treatment may be given to patients suffering from UC. In this regard, the above few statements may be considered as the limitation of the work and taken as a scope for exploration by researchers.

Conclusion

This work is taken up with the purpose to grade the severity of UC using endoscopy images. Even though the efficiency of techniques such as DL and TL has been examined in this domain previously but the strength of EL is not yet explored much. The previous studies are executed mostly taking privately collected data or publicly available HyperKvasir datasets (with a partial set of images). But in this study authors have utilized the complete set of UC images from the HyperKvasir dataset and grouped them into two classes of severity. UC with grades MES 0–1 is categorized as initial stage UC and MES 2–3 are categorized as advance stage UC. To observe the potential of EL to improve the overall performance, it is implemented along with DL as well as TL. A DL-ensemble model is designed where three different deep CNN models are trained from scratch. Then the final prediction is obtained by applying the majority voting ensemble scheme combining the individual prediction of the DL models. Likewise, a TL-ensemble model is designed by considering three pre-trained networks (GN, RN, and SN) which are trained using the weights obtained during the training of the ImageNet dataset. Here also the final classification result is generated by applying the majority voting ensemble scheme combining the individual prediction of the TL models. For evaluation purpose the performance analysis is done in two steps. Some major findings are summarized in this section.

(a) In the first step the performance of the DL and TL models are analyzed. In this phase, it is observed that TL models perform better than the DL models. The TL model which achieved the best result (RN) provided

Table 8 Comparative study of prior art-of-work with the proposed methodology

References	Objective	Dataset	Technique	Output
Bhambhvani et al. [4]	To carry out a multiclass classification task and grade the UC image as MES 1, MES 2, or MES 3	HyperKvasir dataset, out of 851 images 777 images are considered for the study	Trained four deep CNN models namely Resnext-101 pre-trained, Resnext-101 not pre-trained, Resnext-50 pre-trained, and Resnext-102 conv out of which Resnext-101 pre-trained has the best performance	Result of the best network: Overall accuracy is 77.2%, AUC of 0.96 for Grade-1 UC, AUC of 0.86 for Grade-2 UC, and AUC of 0.89 for Grade-1 UC
Stidham et al. [6]	Determining whether DL models can grade the UC severity as remission (MES 0 or 1) or moderate to severe (MES 2 or 3) equivalent to human reviewers	Endoscopic images from the University of Michigan database are collected having 16,514 images related to 3082 patients	Two human reviewers graded the UC images as per the MES grading which is finally decided by a third reviewer. A 159-layered deep CNN model is designed and trained for automatic grading of UC	Result of CNN model: AUROC of 0.966, PPV of 0.87, sensitivity of 83.0%, specificity of 96.0% and NPV of 0.94
Weighted kappa statistics agreement between deep model and deciding reviewer is 0.84				
Sutton et al. [12]	Two purposes of the study: (1) Binary classification to predict UC from non-UC images and (2) binary classification to grade the severity of UC as inactive/mild (MES 0–1) and moderate/severe (MES 2–3)	HyperKvasir dataset, out of 851 images 840 images are used for grading the severity of UC	Four different pre-trained CNN models are trained for evaluation of UC severity grading, model trained are Inception-V3, ResNet50, VGG19, and DenseNet121 out of which DenseNet121 is the best performing model	Result of best pre-trained model: Accuracy of 87.50%, and AUC of 0.90
Yao et al. [13]	Design a fully automated system for UCSG	For classifying still endoscopic images 16,000 still, images related to 3000 patients are collected and graded as MES 0, 1, 2, and 3 by specialists	For automatic grading of UC still images an Inception-V3 model, a 42-layered CNN model is trained using the pre-trained weights using ImageNet	Fully automated methods for high-resolution videos correctly predicted 78% of videos
Beckar et al. [14]	Designing, training, and deployment of a deep learning network for UCSG	1672 endoscopic videos are collected from e etrolizumab Phase II Eucalyptus and Phase III Hickory and Laurel clinical trials for training the model, evaluation of the model is done using the HyperKvasir dataset	The model that is trained is designed based on the ResNet50 architecture	The model achieved an AUROC of 0.84 (MES $\geq$ 1), AUROC of 0.85 (MES $\geq$ 2), and AUROC of 0.85 (MES $\geq$ 3)
Ozawa et al. [15]	Designing a computer-assisted evaluation of endoscopic disease activity in patients with UC	26,304 colonoscopy images related to 841patients suffering from UC are collected from Tada Tomohiro Institute of Gastroenterology and Proctology, Japan	A deep CNN network is designed based on GoogleNet architecture for the classification task	The model generated an AUROC of 0.86 for MES 0 and AUROC of 0.98 for MES 0–1

Table 8 (continued)

References	Objective	Dataset	Technique	Output
Mokter et al. [16]	Designed a model for classifying severity in UC from colonoscopy videos with help of vascular patterns	A total of 178 colonoscopy videos belonging to the class of normal, mild, moderate, and severe are collected	Three different CNN models are designed for classifying the patches extracted from colonoscopy videos. Model-1 is trained to classify the patches based on vascular patterns and then Model-2 and Model-3 are trained to classify the patches into four different classes	Patch level accuracy for each model: Model-1 gives an accuracy of 82.2%, Model-2 accuracy is 77.1% and Model-3 accuracy is 78.2%
Proposed model	Design a DL-ensemble and TL-ensemble framework for grading the severity as initial stage UC and advance stage UC	HyperKavsir dataset is used where the whole section of UC images is considered	Six different models, three models are designed and trained from scratch, three models are trained by taking the pre-trained weights. Then the result is ensembled using the majority voting scheme to get the final prediction	TL-ensemble model generated an accuracy of 90.58%, MCC of 0.7624 whereas the DL-ensemble model generated an accuracy of 80.0%, MCC of 0.4704

a classification accuracy of 88.23% which is 11.94% more than the best performing DL model (N1).

(b) In the second phase of evaluation, the efficiency of EL is analyzed in improving the performance of both DL and TL networks. In this phase, it is seen that the DL-ensemble model generates an accuracy of 80% which is 1.49% more than the best-performing DL network (N1).

(c) It is also seen that the TL-ensemble model gives an accuracy of 90.58% and enhanced the accuracy by 2.66% when compared to the result of the best-performing TL network (RN).

(d) But when DL-ensemble and TL-ensemble model is compared it is perceived that the accuracy generated by TL-ensemble is 13.23% more than the accuracy obtained by the DL-ensemble model. As a concluding remark it is stated that the TL-ensemble model has achieved the best performance with CA of 90.58%, R of 98.36%, P of 89.55%, S of 70.83%, F of 93.74%, and MCC of 0.7624.

Even though the study illustrates a good performance by using the concept of DL, TL, and EL but still the work could be boosted by considering a few of the following research options. Considering the DL section, the designed networks can be enhanced by tuning the hyperparameters using some optimization techniques and implementing some advanced options for the network parameters. In the TL section, three pre-trained models are considered here but there are various other efficient pre-trained networks that can be experimented upon. And finally coming to EL, in this study the majority voting scheme is implemented whereas in literature there are various other techniques available for combining the individual prediction outcome of the networks that can be worked on.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Conflict of interests The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval This study is retrospective and the HyperKvasir data used in this study is freely and publicly available on [23], so there is no need to obtain informed consent.

Consent to Participate The data studied in this research has been made freely available to the public on [23], and the current study was a retrospective study of the data, and did not require informed consent and publication.

References

1. L. Zhang, H. Gan, Secondary colon cancer in patients with ulcerative colitis: a systematic review and meta-analysis. J. Gastrointest. Oncol. **12**(6), 2882 (2021)
2. B. Mabika, Ulcerative colitis complicated by colon cancer in a young adult. Int. J. **4**(4), 110 (2021)
3. A.H. Hamza, H.A. Aglan, H.H. Ahmed (2017) Recent concepts in the pathogenesis and management of colorectal cancer. Recent Advanced in Colon Cancer
4. H. Bhambhvani, A. Zamora, Deep learning enabled assessment of endoscopic disease severity in patients with ulcerative colitis. Eur. J. Gastroenterol. Hepatol. **33**, 6485–6649 (2020)
5. T.Y. Huang, S.Q. Zhan, P.J. Chen, C.W. Yang, H.H.S. Lu, Accurate diagnosis of endoscopic mucosal healing in ulcerative colitis using deep learning and machine learning. J. Chin. Med. Assoc. **84**(7), 678–681 (2021)
6. R.W. Stidham, W. Liu, S. Bishu, M.D. Rice, P.D. Higgins, J. Zhu, B.K. Nallamothu, A.K. Waljee, Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. JAMA Netw. Open **2**(5), e193963–e193963 (2019)
7. W.S. Liew, T.B. Tang, C.H. Lin, C.K. Lu, Automatic colonic polyp detection using integration of modified deep residual convolutional neural network and ensemble learning approaches. Comput. Methods Programs Biomed. **206**, 106114 (2021)
8. R.A. Pratiwi, S. Nurmaini, D.P. Rini, M.N. Rachmatullah, A. Darmawahyuni, Deep ensemble learning for skin lesions classification with convolutional neural network. IAES Int. J. Artif. Intell. **10**(3), 563 (2021)
9. S. Mohapatra, T. Swarnkar, M. Mishra, D. Al-Dabass, R. Mascella, Deep learning in gastroenterology, in *Handbook of computational intelligence in biomedical engineering and healthcare*. (Elsevier, 2021), pp.121–149. https://doi.org/10.1016/B978-0-12-822260-7.00001-7
10. J. Yogapriya, V. Chandran, M.G. Sumithra, P. Anitha, P. Jenopaul, C. Suresh Gnana Dhas, Gastrointestinal tract disease classification from wireless endoscopy images using pretrained deep learning model. Comput. Math. Methods Med. **2021**(1), 5940433 (2021)
11. A. Das, S.K. Mohapatra, M.N. Mohanty, Design of deep ensemble classifier with fuzzy decision method for biomedical image classification. Appl. Soft Comput. **115**, 108178 (2022)
12. R.T. Sutton, O.R. Zaiane, R. Goebel, D.C. Baumgart, Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images. Sci. Rep. **12**(1), 1–10 (2022)
13. H. Yao, K. Najarian, J. Gryak, S. Bishu, M.D. Rice, A.K. Waljee, H.J. Wilkins, R.W. Stidham, Fully automated endoscopic disease activity assessment in ulcerative colitis. Gastrointest. Endosc. **93**(3), 728–736 (2021)
14. B.G. Becker, F. Arcadu, A. Thalhammer, C.G. Serna, O. Feehan, F. Drawnel, Y.S. Oh, M. Prunotto, Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. Ther. Adv. Gastrointest. Endosc. **14**, 263177452199062 (2021). https://doi.org/10.1177/2631774521990623
15. T. Ozawa, S. Ishihara, M. Fujishiro, H. Saito, Y. Kumagai, S. Shichijo, K. Aoyama, T. Tada, Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. Gastrointest. Endosc. **89**(2), 416–421 (2019)
16. Mokter, M.F., Oh, J., Tavanapong, W., Wong, J. and Groen, P.C.D., Classification of ulcerative colitis severity in colonoscopy videos using vascular pattern detection. In *International Workshop on Machine Learning in Medical Imaging*. Springer, Cham. pp. 552–562, (2020)
17. Tejaswini, S.V.L.L., Mittal, B., Oh, J., Tavanapong, W., Wong, J. and Groen, P.C.D., Enhanced approach for classification of ulcerative colitis severity in colonoscopy videos using CNN. In *International Symposium on Visual Computing* . Springer, Cham, pp. 25–37, (2019)
18. W.K. Moon, Y.W. Lee, H.H. Ke, S.H. Lee, C.S. Huang, R.F. Chang, Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput. Methods Programs Biomed. **190**, 105361 (2020)
19. A. Kumar, J. Kim, D. Lyndon, M. Fulham, D. Feng, An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J. Biomed. Health Inform. **21**(1), 31–40 (2016)
20. A. Manna, R. Kundu, D. Kaplun, A. Sinitca, R. Sarkar, A fuzzy rank-based ensemble of CNN models for classification of cervical cytology. Sci. Rep. **11**(1), 1–18 (2021)
21. Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. Javier Aguirre, A. María Vanegas, Breast cancer histopathology image classification using an ensemble of deep learning models. Sensors **20**(16), 4373 (2020)
22. D.T. Nguyen, M.B. Lee, T.D. Pham, G. Batchuluun, M. Arsalan, K.R. Park, Enhanced image-based endoscopic pathological site classification using an ensemble of deep learning models. Sensors **20**(21), 5982 (2020)
23. H. Borgli, V. Thambawita, P.H. Smedsrud, S. Hicks, D. Jha, S.L. Eskeland, K.R. Randel, K. Pogorelov, M. Lux, D.T.D. Nguyen, D. Johansen, HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data **7**(1), 1–14 (2020)
24. 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. and Riegler, M., Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. *In Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164–169, (2017)
25. S. Mohapatra, T. Swarnkar, J. Das, Deep convolutional neural network in medical image processing, in *Handbook of deep learning in biomedical engineering*. (Elsevier, 2021), pp.25–60. https://doi.org/10.1016/B978-0-12-823014-5.00006-5
26. A.S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on MRI. Z. Med. Phys. **29**(2), 102–127 (2019)
27. Mohapatra, S., Pati, G.K. and Swarnkar, T., Efficiency of transfer learning for abnormality detection using colonoscopy images: a critical analysis. *In 2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, IEEE, pp. 1–6, (2022)
28. J. Xu, K. Xue, K. Zhang, Current status and future trends of clinical diagnoses via image-based deep learning. Theranostics **9**(25), 7556 (2019)
29. Almanifi, O.R.A., Razman, M.A.M., Khairuddin, I.M., Abdullah, M.A. and Majeed, A.P.A., Automated gastrointestinal tract classification via deep learning and the ensemble method. In *2021 21st International Conference on Control, Automation and Systems (ICCAS)*, IEEE, pp. 602–606, (2021)
30. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, (2015)
31. He, K., Zhang, X., Ren, S. and Sun, J., Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 770–778, (2016)
32. Zhang, X., Zhou, X., Lin, M. and Sun, J., Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, (2018)

33. M. Hmoud Al-Adhaileh, E. Mohammed Senan, W. Alsaade, T.H. Aldhyani, N. Alsharif, A. Abdullah Alqarni, M.I. Uddin, M.Y. Alzahrani, E.D. Alzain, M.E. Jadhav, Deep learning algorithms for detection and classification of gastrointestinal diseases. Complexity **2021**, 6170416 (2021)
34. Kochgaven, C., Mishra, P. and Shitole, S., Detecting presence of COVID-19 with ResNet-18 using PyTorch. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pp. 1–6, IEEE, (2021)
35. L.D. Nguyen, R. Gao, D. Lin, Z. Lin, Biomedical image classification based on a feature concatenation and ensemble of deep CNNs. J. Amb. Intell. Human. Comput. **14**(11), 15455–15467 (2019). https://doi.org/10.1007/s12652-019-01276-4
36. K. Raza, Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule, in *U-Healthcare Monitoring Systems*. (Elsevier, 2019), pp.179–196. https://doi.org/10.1016/B978-0-12-815370-3.00008-6

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.