Ensemble of adapted CNN methods for classifying colon histopathological images (#70538)

First submission

Guidance from your Editor

Please submit by **3 Mar 2022** for the benefit of the authors (and your $200 publishing discount) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files
Download and review all files from the materials page.

1 Latex file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- ✓ Clear, unambiguous, professional English language used throughout.
- ✓ Intro & background to show context. Literature well referenced & relevant.
- ✓ Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- ✓ Figures are relevant, high quality, well labelled & described.
- ✓ Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- ✓ Original primary research within Scope of the journal.
- ✓ Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- ✓ Rigorous investigation performed to a high technical & ethical standard.
- ✓ Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- ⓘ Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- ✓ All underlying data have been provided; they are robust, statistically sound, & controlled.
- ✓ Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip

Example

Support criticisms with evidence from the text or from other sources

Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.

Give specific suggestions on how to improve the manuscript

Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).

Comment on language and grammar issues

The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.

Organize by importance of the issues, and number your points

1. *Your most important issue*
2. *The next most important item*
3. *…*
4. *The least important points*

Please provide constructive criticism, and avoid personal opinions

I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC

Comment on strengths (as well as weaknesses) of the manuscript

I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.

Ensemble of adapted CNN methods for classifying colon histopathological images

Dheeb Albashish Corresp. 1

[1] Computer Science Department/ Prince Abdullah bin Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University, Alsalt, Jordan

Corresponding Author: Dheeb Albashish
Email address: bashish@bau.edu.jo

Deep convolutional neural networks (CNN) manifest the potential for computer-aided diagnosis systems (CADs) by learning features directly from images rather than using traditional feature extraction methods. Nevertheless, due to the limited sample sizes and heterogeneity in tumor presentation in medical images, CNN models suffer from training issues, including training from scratch, which leads to overfitting. Alternatively, a pre-trained neural network's transfer learning (TL) is used to derive tumor knowledge from medical image datasets using CNN that were designed for non-medical activations, alleviating the need for large datasets. This study proposes a new set of TL methods based on DenseNet121, MobileNetV2, InceptionV3, and VGG16 models for colon cancer multiclass classification of histopathology images. The pretrained models are adapted based on a block-wise fine-tuning policy, in which we join a set of dense and dropout layers to these pretrained models to make them more specific to colon classification. To boost the performance of these adjusted models, an ensemble is obtained via product and majority voting aggregation methods. The proposed ensemble (called E-CNNs) is based on softmax individual classifiers of the fully connected layers (FCC). The proposed E-CNNs is validated on the publicly available benchmark colon histopathological image dataset. The proposed E-CNNs is compared with other deep learning models on the same dataset. The obtained results provide clear evidence that E-CNNs outperform all baseline deep learning models in terms of accuracy, sensitivity, and specificity. The E-CNNs can be used in various medical image applications, which is approved from the achieved results on the colon dataset.

Ensemble of Adapted CNN Methods for classifying Colon Histopathological Images

Dheeb Albashish[1]

[1]**Computer Science Department, Prince Abdullah bin Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University, Salt, Jordan. Email: bashish@bau.edu.jo**

Corresponding author:
Dheeb Albashish[1]

Email address: bashish@bau.edu.jo

ABSTRACT

Deep convolutional neural networks (CNN) manifest the potential for computer-aided diagnosis systems (CADs) by learning features directly from images rather than using traditional feature extraction methods. Nevertheless, due to the limited sample sizes and heterogeneity in tumor presentation in medical images, CNN models suffer from training issues, including training from scratch, which leads to overfitting. Alternatively, a pre-trained neural network's transfer learning (TL) is used to derive tumor knowledge from medical image datasets using CNN that were designed for non-medical activations, alleviating the need for large datasets. This study proposes a new set of TL methods based on DenseNet121, MobileNetV2, InceptionV3, and VGG16 models for colon cancer multiclass classification of histopathology images. The pretrained models are adapted based on a block-wise fine-tuning policy, in which we join a set of dense and dropout layers to these pretrained models to make them more specific to colon classification. To boost the performance of these adjusted models, an ensemble is obtained via product and majority voting aggregation methods. The proposed ensemble (called E-CNNs) is based on softmax individual classifiers of the fully connected layers (FCC). The proposed E-CNNs is validated on the publicly available benchmark colon histopathological image dataset. The proposed E-CNNs is compared with other deep learning models on the same dataset. The obtained results provide clear evidence that E-CNNs outperform all baseline deep learning models in terms of accuracy, sensitivity, and specificity. The E-CNNs can be used in various medical image applications, which is approved from the achieved results on the colon dataset.

INTRODUCTION

Colon cancer is the third most deadly disease in males and the second most hazardous in females. According to the World Cancer Research Fund International, over 1.8 million new cases were reported in 2018(Belciug and Gorunescu, 2020). In colon cancer diagnosis, the study of histopathological images under the microscope plays a significant role in the interpretation of specific biological activities. Among the microscopic inspection functions, classification of images (organs, tissues, etc.) is one of considerable important tasks. However, classifying medical images into a set of different classes is a very challenging issue due to low inter-class distance and high intra-class variability(Sahran et al., 2018), as illustrated in Figure 1. Some objects in medical images may be found in images belonging to different classes, and different objects may appear at different orientations and scales in a given class. During the manual assessment, physicians examine the Hematoxylin and Eosin (H&E) stained tissues under a microscope to analyze their histopathological attributes, such as cytoplasm, nuclei, gland, and lumen, as well as change in the benign structure of the tissues. It is worth noting that early categorization of colon samples as benign or malignant, or discriminating between different malignant grades is critical for selecting the best treatment protocol. Nevertheless, manually diagnosing colon H&E stained tissue under a microscope is time-consuming and tedious, as illustrated in Figure 1. In addition, the diagnostic performance depends on the experience and personal skills of a pathologist. It, also, suffers from inter-observer variability with around 75% diagnostic agreement across pathologists (Elmore et al., 2015). As a result, the treatment protocol might differ from one pathologist to another. These issues motivate development and research into the automation of diagnostic and prognosis procedures(Stoean et al., 2016).

In recent decades, various computer aided diagnosis systems (CADs) have been introduced to tackle the classification problems in cancer digital pathology diagnosis to achieve reproducible and rapid results. CADs assist in enhancing the classification performance and, at the same time, minimize the variability in interpretations. The faults produced by CADs/machine learning model have been announced to be less than those produced by a pathologist(Kumar et al., 2020). These models can also assist clinicians in detecting cancerous tissue in colon tissue images. As a result, researchers are trying to construct CADs to improve diagnostic effectiveness and raise inter-observer satisfaction(Tang et al., 2009). Numerous conventional CADs for identifying colon cancer using histological images had been introduced by number of researchers in the past years(Stoean et al., 2016; Kalkan et al., 2012; Li et al., 2019). Most of the conventional CADs focus on discriminating between benign and malignant tissues. Furthermore, they focus on conventional machine learning and image processing techniques. In this regards, they emphasize on some complex tasks such as extracting features from medical images and require extensive preprocessing. The complex nature of these tasks in machine learning techniques degrades the results of the CADs regarding accuracy and efficiency(Ahmad et al., 2021). Conversely, recent advances in machine learning technologies make this task more accurate and cost-effective than traditional models.

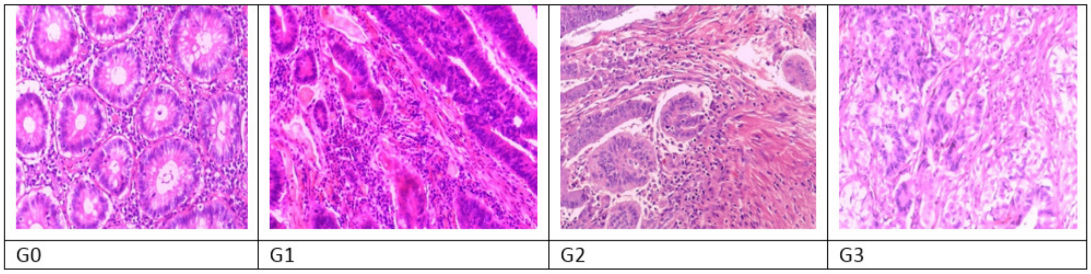


Figure 1. Colon histopathology images from the benchmark dataset(Stoean et al., 2016) with 40× magnification factor, from left to right: normal(G0), cancer grade 1(G1), cancer grade2(G2), and cancer grade 3(G3).

In the last few years, deep learning techniques have become a prevalent and leading tool in the field of machine learning for colon histopathlogical image classification. Recently, one of the most successful deep learning techniques is the deep convolutional neural networks (CNN) (Khan et al., 2020) that consists of series of convolutional and pooling layers. These are followed by fully connected layers (FCC) and softmax layers. The FCC and the softmax represent the neural networks classifiers. CNN has the ability to extract the features from images by parameter tuning of the convolutional and the pooling layers. Thus, it achieves great success in many fields especially in medical image classifications such as skin disease(Harangi, 2018), breast(Deniz et al., 2018) and colon cancer classification(Ghosh et al., 2021). CNN is categorized into two approaches: either training from scratch or pre-trained models (e.g., DenseNet (Huang et al., 2017), MobileNet(Sandler et al., 2018), and InceptionV3(Szegedy et al., 2016)). The most effective approach in medical image classification is the pretrained models due to the limited number of training samples(Saini and Susan, 2020).

CNN has been used in the domain of colon histopathlogical image classification. For example, Stefan Postavaru (Postavaru et al., 2017)utilized a CNN approach for the automated diagnosis of a set of colorectal cancer histopathological slides. They utilized CNN with 5 convolutional layers and reported accuracy of 91.4%. Ruxandra Stoean (Stoean, 2020) extended the work (Postavaru et al., 2017) and presented a modality method to tune the convolutional of the deep CNN. She introduced two Evolutionary algorithms for CNN parametrization. She conducted the experiments on colorectal cancer(Stoean et al., 2016) and reported the highest accuracy of 92%. It was obtained from these studies that the CNN models exceeded the handcrafted features.

While the CNN achieves high performance especially on large dataset size, it struggles to make such performance on small dataset size(Deniz et al., 2018; Mahbod et al., 2020), and simply results in overfitting issue(Zhao et al., 2017). To overcome this issue, the concept of transfer learning technique of pretrained CNN models is exploited for classification of colon histopathlogical images. In practical, the transfer learning technique of the pretrained models exports knowledge from previously CNN that has been trained on the large dataset to the new task with small dataset (target dataset). There are two approaches to transfer learning of pretrained models in medical image classification: feature extraction and fine-tuning (Benhammou et al., 2020). The former method extracts features from any convolutional or pooling layers and removes the last fully connected and softmax layers. While in the latter, the pretrained CNN models are adjusted for specific tasks. It is important to remember that the number of neurons in the final FC layer corresponds to the number of classes in the target dataset (i.e., the number of colon types). Following this replacement, the whole pre-trained model is retrained (Mahbod et al., 2020; Benhammou et al., 2020; Zhi et al., 2017) or the last FC layers are retrained (Benhammou et al., 2020). Various pretrained models (e.g., DenseNet (Huang et al., 2017), MobileNet(Sandler et al., 2018), VGG16(Simonyan and Zisserman, 2014), and InceptionV3(Szegedy et al., 2016)) have been introduced in recent years. Each pretrained model is constructed based on several convolution layers and filter sizes to extract specific features from the input image. However, transferring the begotten experience from the source (ImageNet) to our target (colon images) leads to losing some powerful features of histopathological image analysis(Boumaraf et al., 2021). For example, CNN pretrained AlexNet and GoogleNet models were used on the colon histopathological images classification (Popa, 2021). However, they achieved poor standard deviation results. Besides, using these pretrained models on the colon dataset needs a specific fine-tuning approach to achieve acceptable results.

To accommodate the pretrained CNN models to the colon image classification, we design a new set of transfer learning models (DenseNet (Huang et al., 2017), MobileNet(Sandler et al., 2018), VGG16(Simonyan and Zisserman, 2014), and InceptionV3(Szegedy et al., 2016) to refine the pre-trained models on the colon histopathological image tasks. Our transfer learning methods are based on a block-wise fine-tuning policy. We make the last set of residual blocks of the deep network models more domain-specific to our target colon dataset by adding dense layers and dropout layers while freezing the remaining initial blocks in the deep pretrained model. The adaptability of the proposed method is further extended by fine-tuning the neural network's hyper-parameters to improve the model generalization ability. Besides, a single pretrained model has a limited capacity to extract complete discriminating features, resulting in an inadequate representation of the colon histopathology performance (Yang et al., 2019). As a result, this study proposes an ensemble of pretrained CNN models architectures (E-CNN) to identify the representation of colon pathological images from various viewpoints for more effective classification tasks.

In this research, the following contributions are made:

- Investigation of the influence of the standard TL approaches (DenseNet, MobileNet, VGG16, and InceptionV3) on the colon cancer classification task.
- Design a new set of transfer learning methods based on a block-wise fine-tuning approach to learn the powerful features of the colon histopathology images. The new design includes adding a set of dense and dropout layers while freezing the remainder of the initial layers in the pretrained models (DenseNet, MobileNet, VGG16, and InceptionV3) to make them more specific for the colon domain requirements.
- Define and optimize a set of hyper-parameters for the new set of pretrained CNN models to classify colon histopathological images.
- An ensemble (E-CNN) is proposed to extract complementary features in colon histopathology images by using an ensemble of all the introduced transfer learning methods (base classifiers). The proposed E-CNN merges the decisions of all base classifiers via majority voting and product rules.

The remainder of this research is organized as follows. Section goes over the related works. Our proposed methodology is presented in detail in Section . Section presents and discusses the experimental results. Section (conclusion) brings this study to a close by outlining some research trends and viewpoints.

LITERATURE REVIEW

Deep learning pretrained models have made incredible progress in various kinds of medical image processing, specifically histopathological images, as they can automatically extract abstract and complex

features from the input images. Recently, CNN models based on deep learning design are dominant techniques in the CADs of cancer histopathological image classification (Kumar et al., 2020; Mahbod et al., 2020; Albashish et al., 2021). CNN learn high- and mid-level abstraction, which is obtained from input RGB images. Thus, developing CADs using deep learning and image processing routines can assist pathologists in classifying colon cancer histopathological images with better diagnostic performance and less computational time. Numerous CADs for identifying colorectal cancer using histological images had been introduced by a number of researchers in past years. These CADs vary from conventional machine learning algorithms of the deep CNN. In this study, we present the related work of the colorectal cancer classification relying on colorectal cancer dataset (Stoean et al., 2016) as real-world test cases.
The authors in (Postavaru et al., 2017) designed a CNN model for colon cancer classification based on colorectal histopathological slides belonging to a healthy case and three different cancer grades(1, 2, and 3). They used an input image with the size of 25xx256x3. They created five convolutional neural networks, followed by the RELU activation function. In the introduced CNN, various kernel sizes were utilized in each Conv. Layer. Besides, they utilized batch normalization and only two fully connected layers. They reported 91% accuracy in multiclass classification for the colon dataset in (Stoean et al., 2016). However, in the proposed approach, only the size of the kernels is considered, while other parameters, like learning rate and epoch size, were not taken into account.

The author in (Stoean, 2020) extended the previous study (Postavaru et al., 2017) by applying an Evolutionary algorithm (EA) in the CNN architecture to automate two tasks. She used genetic algorithm (GA) from EA methods to select the most import setting using SVM classifier with the mean square error (MSE). EA was first conducted for tuning the CNN hyper-parameters for the convolutional layers. She determined the number of kernels in CNN and their size. Afterward, a second EA was used to support SVM in parameters ranking to determine the variable importance within the hyper-parameterization of CNN. The proposed approach achieved 92% colorectal cancer grading accuracy on the dataset in (Stoean et al., 2016). However, using EV does not guarantee any diversity among the obtained hyper-parameters(solutions) (Bhargava, 2013).Thus, choosing the kernel size and depth of CNN may not ensure high accuracy.

The authors in (Popa, 2021) proposed a new framework for the colon multiclass classification task. They employed CNN pretrained AlexNet and GoogleNet models followed by softmax activation layers to handle the 4-class classification task. The best-reported accuracies on (Stoean et al., 2016) dataset ranged between 85% and 89%. However, the standard deviation of these results was around 4%. This means the results were not stable. AlexNet was also used in (Lichtblau and Stoean, 2019) as a feature extractor for the colon dataset. Then, an ensemble of five classifiers was built. The obtained results for this ensemble achieved around 87% accuracy.

In (Ohata et al., 2021), the authors use CNN to extract features of colorectal histological images. They employed various pretrained models, i.e., VGG16 and Inception, to extract deep features from the input images. Then, they employed ensemble learning by utilizing five classifiers (SVM, Bayes, KNN, MLP, and Random Forest) to classify the input images. They reported 92.083% accuracy on the colon histological images dataset in (Kather et al., 2016). A research study in (Rachapudi and Lavanya Devi, 2021) proposed light weighted CNN architecture. RGB-colored images of colorectal cancer histology dataset (Kather et al., 2016) belonging to eight different classes were used to train this CNN model. It consists of 16 convolutional layers, five dropout layers, five max-pooling layers, and one fully connected layer. This architecture exhibited high performance in term of incorrect classification compared to existing CNN models. Using ensemble learning model achieved around 77% accuracy (error of 22%).

In another study for colon classification but on a different benchmark dataset, the authors in (Malik et al., 2019) have proved that the transferred learning from a pre-trained deep CNN model using InceptionV3 on a colon dataset with fine-tuning provides efficient results. Their methodology was mainly constructed based on InceptionV3. Then, the authors modified the last FCC layers to become harmonious with the number of the classes in the colon classification task. Moreover, the adaptive CNN implementation was proposed to improve the performance of CNN architecture for the colon cancer detection task. The study achieved around 87% accuracy for the multiclass classification task.

In another study (Dif and Elberrichi, 2020a), a framework was proposed for the colon histopathological

Table 1. Summary of the major classification studies on colon cancer

Authors in	Dataset used	CNN architecture	Accuracy	Using pretrained either feature extraction/ fine- tuning
(Stoean, 2020)	colorectal in (Stoean et al., 2016)	CNN model from scratch	92%	Fine-tune: only kernel size and number of kernels in CNN using EA method
(Popa, 2021)	colorectal in (Stoean et al., 2016)	AlexNet and GoogleNet	89%	feature extractor
(Postavaru et al., 2017)	colorectal in (Stoean et al., 2016)	CNN model from scratch	91%	The number of filters and the kernel size
(Lichtblau and Stoean, 2019)	colorectal in (Stoean et al., 2016)	AlexNet	87%	Feature extractor with ensemble learning
(Ohata et al., 2021)	colorectal in (Kather et al., 2016)	Set of pretrained models (VGG16, Inception, Resent)	92.083%	Feature extraction extraction
(Rachapudi and Lavanya Devi, 2021)	colorectal in (Kather et al., 2016)	CNN architecture	77%	Fine-tune CNN model
(Dif and Elberrichi, 2020a)	colorectal in (Kather et al., 2016)	Pretrained Resnet121	94%	Feature extraction
(Boruz and Stoean, 2018)	colorectal in (Stoean et al., 2016)	Contour low-level image features	92.6%	

image classification task. The authors employed a CNN based on transferred learning from Resnet121 generating a set of models followed by a dynamic model selection using the particle swarm optimization (PSO) metaheuristic. The selected models were then combined by a majority vote and achieved 94.52% accuracy on the colon histopathological dataset (Kather et al., 2016). In the same context, the authors in (Dif and Elberrichi, 2020b) explored the efficiency of reusing pre-trained models on histopathological images dataset instead of ImageNet based models for transfer learning. For this target, a fine-tuning method was presented to share the knowledge among different histopathological CNN models. The basic model was created by training InceptionV3 from scratch on one dataset while transfer learning and fine-tuning were performed using another dataset. However, this transfer learning-based strategy offered poor results on the colon histopathological images due to the limited number of the training dataset.

The conventional machine learning techniques had been utilized also for the colon dataset (Stoean et al., 2016) to achieve accepted results. For example, the study in (Boruz and Stoean, 2018) conducted on the 4-class colon cancer classification task on the dataset in (Stoean et al., 2016). Contour low-level image features from grayscale transformed images were used to train the SVM classifier. Despite its simplicity, the study displayed a comparable performance to some computationally expensive approaches. The authors reported accuracy averages between 84.1% and 92.6% for the different classes. However, transforming the input images to grayscale leads to losing some information and degrades the classification results. Besides, using thresholding needs fine-tuning, which is a complex task.

In (Khadilkar, 2021), the authors extracted morphological features from the colon dataset. Mainly, they extracted Harris corner and Gabor Wavelet features. These features were then used to feed the Neural Network classifier. They applied their framework on the colon dataset (Stoean et al., 2016) to discriminate between benign vs. malignant cases. However, they ignored the multiclass classification task, which is more complex task in this domain.

Overall, the earlier studies, summarized in Table 1, revealed a notable trend in using deep CNN to classify colon cancer histopathological images. It was used to provide much higher performance than the conventional machine learning models. Nevertheless, training CNN models are not that trivial as they need considerable memory resources and computation and are usually hampered by over-fitting problems. Besides, they require a large amount of training dataset. In this regard, the recent studies (Ahmad et al., 2021; Boumaraf et al., 2021) have demonstrated that sufficient fine-tuned pretrained CNN models performance is much more reliable than the one trained from scratch, or in the worst cases the same. Besides, using ensemble learning of pretrained models show effective results in various applications of image classification tasks. Therefore, this research fills the gap in the previous studies for colon histopathological images classification by introducing a set of transfer learning models based on Dense. Then, reap the benefits of the ensemble learning to fuse their decision.

METHODOLOGY

This study constructs an ensemble of the pretrained models with fine-tuning for the colon diagnosis based on histopathological images. Mainly, four pretrained models (DenseNet121 MobileNetV2, InceptionV3, and VGG16) are fine-tuned, and then their predicted probabilities are fused to produce a final decision for a test/image. The pretrained models utilize transfer learning to mitigate these models' weights to handle a similar classification task. Ensemble learning of pretrained models attains superior performance for histopathological image classification.

Transfer Learning (TL) and pretrained Deep Learning Models for medical image

Transferring knowledge from one expert to another is known as transfer learning. In deep learning techniques, this approach is utilized where the CNN is trained on the base dataset (source domain), which has a large number of samples (e.g., ImageNet). Then, the weights of the convolutional layers are transferred to the new small dataset (target domain). Using pretrained models for classification tasks can be divided into two main scenarios: freezing the layers of the pretrained model and fine-tuning the models. In the former scenario: the convolutional layers of a deep CNN model are frozen, and the last fully connected layers are omitted. In this way, the convolutional layers act as feature extractions. Then these features are passed to a specific classifier (e.g., KNN, SVM). While in the latter case, the layers are fine-tuned, and some hyper-parameters are adjusted to handle a new task. Besides, the top layer (fully connected layer) is adjusted for the target domain. In this study, for example, we configure the number of neurons in this layer to (4) in accordance with the number of classes in the colon dataset. TL aims to boost the target field's accuracy (i.e., colon histopathological) by taking full advantage of the source field (i.e., ImageNet). Therefore, in this study, we transfer the weights of the set of four powerful pretrained CNN models (DenseNet (Huang et al., 2017), MobileNet(Sandler et al., 2018), VGG16(Simonyan and Zisserman, 2014), and InceptionV3(Szegedy et al., 2016)) with fine-tuning to increase the diagnosis performance of the colon histopathological image classification. The pretrained Deep CNN models and the proposed ensemble learning are presented in the subsequent section.

Pretrained DenseNet121

Dense CNN(DenseNet) was offered by Huang et al.(Huang et al., 2017). The architecture of DenseNet was improved based on the ResNet model. The prominent architecture of DenseNet is based on connecting the model using dense connection instead of direct connection within all the hidden layers of the CNN(Alzubaidi et al., 2021). The crucial benefits of such an architecture are that the extracted features/features map is shared with the model. The number of training parameters is low compared with other CNN models similar to CNN models because of the direct synchronization of the features to all following layers. Thus, the DenseNet reutilizes the features and makes their structure more efficient. As a result, the performance of the DenseNet is increased (Ahmad et al., 2021; Ghosh et al., 2021). The main components of the DenseNet are: the primary composition layer, followed by the ReLU activation function, and dense blocks. The final layer is a set of FC layers (Talo, 2019).

Pretrained MobileNetV2

MobileNet(Sandler et al., 2018) is a lightweight CNN model based on inverted residuals and a linear bottleneck, which form shortcut connections between the thin layers. It is designed to handle limited hardware resources because it is a low-latency model, and a small low power. The main advantage of the MobileNet is the tradeoff between various factors such as latency, accuracy, and resolution(Krishnamurthy et al., 2021). In MobileNet, depth separable convolutional (DSC) and point-wise convolutional kernels are used to produce feature maps. Predominantly, DSC is a factorization approach, which replaces the standard convolution with a faster one. In MobileNet, DSC first uses depth-wise kennels 2-D filters to filter the spatial dimensions of the input image. The size of the depth-wise filter is Dk x Dk x1, where Dk is the size of the filter, which is much less than the size of the input images. Then, it is followed by a point-wise convolutional filter that mainly applied to filter the depth dimension of the input images. The size of the depth filter is1x1xn, where n is the number of kernels. They separate each DSC from point-wise convolutional using batch normalization and ReLU function. Therefore, DSC is called (separable). Finally, the last FCC is connected with the Softmax layer to produce the final output/ classification result. Using depth-wise convolutional can reduce the complexity by around 22.7%. This means the DSC takes only approximately 22% of the computation required by the standard convolutional. Based on this reduction, MobileNet is becoming seven times faster than the traditional convolutional. Thus, it becomes more desirable when the hardware is limited (Srinivasu et al., 2021).

Pretrained InceptionV3

Google teams in (Szegedy et al., 2016) introduced the InceptionV3 CNN. The architecture of InceptionV3 was updated based on the inceptionV1 model, as illustrated in Figure 2. It mainly addressed some issues in the previous inceptionV1 such as auxiliary classifiers by add batch normalization and representation bottleneck by adding kernel factorization(Mishra et al., 2020). The architecture of the inceptionV3 includes multiple various types of kernels (i.e., kernel size) in the same level. This structure aims to solve the issue of extreme variation in the location of the salient regions in the input images under consideration (Mishra et al., 2020). The inceptionV3 (Szegedy et al., 2016) utilizes a small filter size (1x7 and 1x5) rather than a large filter (7x7 and 5x5). In addition, a bottleneck of 1x1 convolution is utilized. Therefore, better feature representation.

The architecture of inceptionV3(Szegedy et al., 2016) is demonstrated in Figure 2. It starts with input data (image), and then mapped parallel computations will be shaped into three different convolutional layers with 3x3 or 5x5 filter size. The output of these layers is aggregated into a single layer, which represents the output layer (e.g., ensemble technique). Using parallel layers with each other will save a lot of memory and increase the model's capacity without increasing its depth.

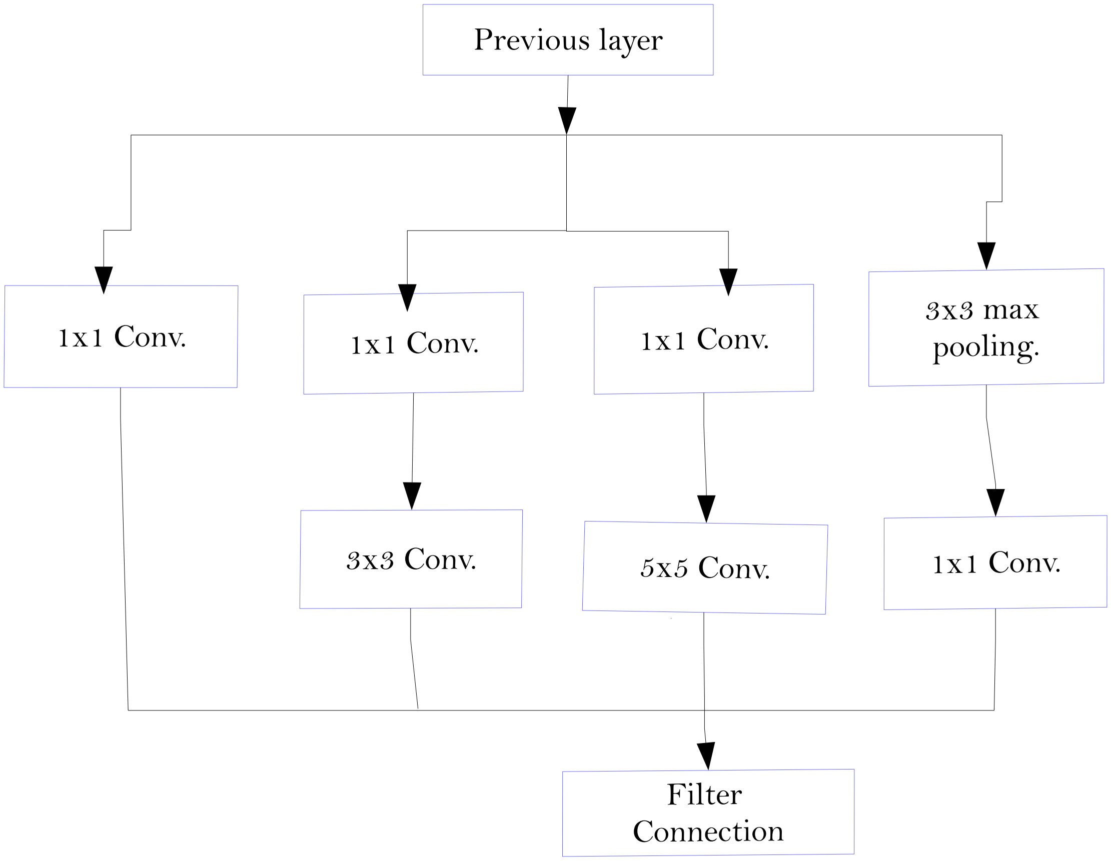


Figure 2. The inception model from (Talo, 2019)

Pretrained VGG16

VGG16 was presented by Simonyan et al. (Simonyan and Zisserman, 2014) as a deeper convolutional neural network model. The basic design of this model is to replace the large kernels with smaller kernels,

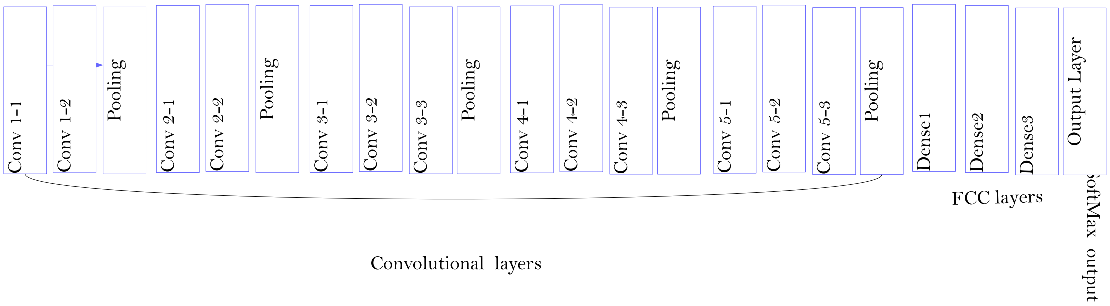


Figure 3. The VGG16 model (Simonyan and Zisserman, 2014)

Table 2. Summary of deep architectures used in this work..

Architecture	No. of Conv layers	No. of FCC layers	No. of training parameters	Minimum image size	Number of extracted features	Top 5 error on ImageNet
DenseNet121	120	1	7 million	221x221	1024	7.71%
InceptionV3	42	1	22 million	299x299	2048	3.08%
VGG16	13	3	134 million	227x227	4096	7.30%
MobileNet	53	3	3.4 million	224x224	1024	-%

and extending the depth of the CNN model(Alzubaidi et al., 2021). Thus, the VGG16 becomes potentially more reliable in carrying out different classification tasks. Figure 3 shows the basic VGG16 (Simonyan and Zisserman, 2014) architecture. It consists of five blocks with 41 layers, where 16 layers have learnable weights; 13 convolutional layers and 3 FCC layers from the learnable layers(Khan et al., 2020). The first two blocks include two convolutional layers, while the last three blocks consist of three convolutional layers. The convolutional layers use small kernels with size of 3x3 and padding 1. These convolutional layers are separated using the max-pooling layers that use 2x2 filter size with padding 1. The output of the last convolutional layer is 4096, which makes the number of neurons in the FCC 4096 neurons. As illustrated in Table 2 , VGG16 uses around 134 million parameters, which raises the complexity of VGG16 relating to other pretrained models (Tripathi and Singh, 2020).

The proposed Deep CNN Ensemble Based on softmax

The previous pretrained models extract various features from the training images to discriminate between different types of the classes/ cancer in the colon images datasets. However, each pretrained model is constructed based on various convolutions layers and filter size to extract different features. Thus, no pretrained model can be more general to extract all discriminating features from the input training images. Besides, using initial weights in pretrained models affect the classification performance since the CNN pretrained models are nonlinear designs. These pretrained models learn complicated associations from training data with assist of backpropagation and stochastic optimization (Ahmad et al., 2021). These problems can be alleviated by employing an ensemble method that involves training several deep CNN architectures and combining their predictions.

This study introduces the ensemble for four CNN pretrained models (DenseNet (Huang et al., 2017), MobileNet(Sandler et al., 2018), VGG16(Simonyan and Zisserman, 2014), and InceptionV3(Szegedy et al., 2016)) (E-CNNs) for the automated classification of colon H&E histopathological images after adapting them. Figure 4 and 5 show the main phases of the proposed E-CNNS. In E-CNNS, the above pretrained models represent the individual classifiers. Each individual is adapted to the current task by introducing block-wise fine tuning. Figure 4 illustrates the main steps for the design of the block-wise fine-tuning technique. First, we use the benchmark colon dataset in (Stoean et al., 2016). Then, we perform some processes on this dataset to prepare it for the pretrained model. After splitting the dataset into training and testing, the four pretrained models are loaded. Then, the last fully-connected and softmax

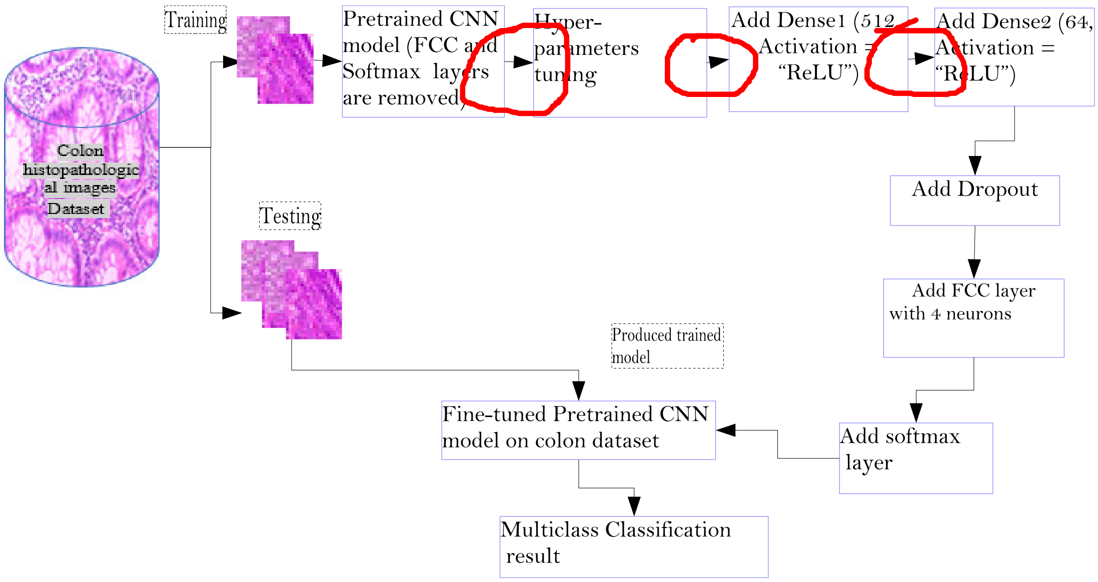

Training
Colon histopathological images Dataset
Testing
Pretrained CNN model (FCC and Softmax layers are removed)
Hyper-parameters tuning
Add Dense1 (512 Activation = "ReLU")
Add Dense2 (64, Activation = "ReLU")
Add Dropout
Add FCC layer with 4 neurons
Add softmax layer
Produced trained model
Fine-tuned Pretrained CNN model on colon dataset
Multiclass Classification result

Figure 4. Block diagram of the proposed Block-wise fine-tuning for each pretrained model from (DenseNet121, MobileNetV2, InceptionV3, and VGG16

layers are omitted from the loaded pretrained CNN models. These layers were originally designed to output 1000 classes from the ImageNet dataset. To strengthen the vital data-particular feature learning from each individual pretrained model, we, then, add two dense layers with a varying number of hidden neurons. These dense layers are followed by the ReLU nonlinear activation function, which allows us to learn complex relationships among the data (Ahmad et al., 2021; Garbin et al., 2020). Next, a 0.3 dropout layer is added to address the long training time and overfitting issues in classification tasks (Deniz et al., 2018; Boumaraf et al., 2021). The last fully connected layer (FCC) with the softmax layer represents the primary base classifier in the proposed ensemble. The FCC is simply a feed-forward neural network, which is fed by flattened input from the last pooling layer of the pretrained model. In this study, the number of neurons in FCC is set to four, based on the number of classes in this study, instead of 1000 classes of ImageNet. At the end of each model, a softmax layer (activation layer) is inserted on top of each model to train the obtained features and produce the classification output based on max probability. Algorithm 1 shows the main steps of the block-wise fine-tuning technique for each individual model in the proposed E-CNNs.

To boost the performances of the four proposed individual models, the ensemble learning is introduced in this study. Figures 4, and 5 show the proposed E-CNNs with its individual models. The four adaptive models are trained on the training dataset. Then evaluated on the tested dataset. The output probabilities of the four pretrained models are connected to produce 16-D feature vector (i.e., each individual with its softmax produce four probabilities based on the number of classes in colon images). Then, various combination methods (majority voting, and product rule) are employed to produce a final decision for the test image. Algorithm 2 shows the main steps in the proposed E-CNNs with majority voting and product rule.

Based on Algorithms 1, 2, and Figures 4,5, the following points are taken into account: First, the CNN model is adapted to handle the heterogeneity in the colon histopathological images using the Block-wise fine-tuning technique for each of the pretrained models. It extracts additional abstract features from the image that aid in increasing intra-class discrimination. Second, ensemble learning is employed to improve the performance of the four adaptive pretrained models. As a result, the final decision regarding the test images will be more precise.

Resources used

All the experiments are implemented using TensorFlow, Keras API, and utilized python programming in Google Colaboratory, or "CoLab." In the CoLab, we utilize Tesla GPU to run our experiment after

Algorithm 1 Building and training the adaptive pretrained models [Block-wise fine-tuning for each pretrained model].

```
input:Training data(T), N samples: T = [$x_1$, $x_2$,. . ., $x_N$], with Category: y = [$y_1$, $y_2$,. . ., $y_N$], pretrained CNN models( M), M=[ DenseNet121, MobileNetV2, InceptionV3, and VGG16 models].
for each $I$ in $M$ do
    Remove the last FCC and softmax layers
    Add Dense1 layer with number of neurons equal to 512 and activation function=”ReLU”
    Add Dense2 layer with number of neurons equal to 64 and activation function=”ReLU”
    Add dropout layer
    Add FCC layers with number of neurons equal to 4( based on number classes in the colon dataset)
    Add softmax layer ( for output probabilities)
    Initialize the Hyper-parameters values, as listed in Table 3
    Build the final model ($adaptI$)
    Train the $adapI$ on $T$
    Append $adapI$ into $adapM$
end for
Output: Adaptive models ($adaptM$), adaptM=[ adapt_DenseNet121, adapt_MobileNetV2, adapt_InceptionV3, and adapt_VGG16 models]
```

Algorithm 2 Ensemble of adaptive models and evaluating the ensemble model on test colon histopathological images.

```
Input: Adaptive models ($adaptM$), $adaptM$=[ adapt_DenseNet121, adapt_MobileNetV2, adapt_InceptionV3, and adapt_VGG16 models], Test images set( D), with z samples: R = [$x_1$, $x_2$, $x_3$,. . ., $x_z$], with Category: y = [$y_1$, $y_2$,. . ., $y_z$]
for $j$ in $D$ do
    for each individual $I$ in $adaptM$ do
        Evaluate the performance of I using the test data $j$.
        $P[j,I]$=probabilities of each class for the test image $j$ when using the individual $I$.
        $V[j,I]$=prediction for the test image $j$ when using the individual $I$.
    end for
    Compute the ensemble final prediction for test image $j$ based on majority voting and $V[j,:]$ (ECNN(majority_voting))
    Compute the ensemble final prediction for test image $j$ based on the product rule(ECNN(product_rule)) and $p[j,:]$
    Output: class prediction for $D$ using (ECNN(majority_voting)) and (ECNN(product_rule))
end for
```

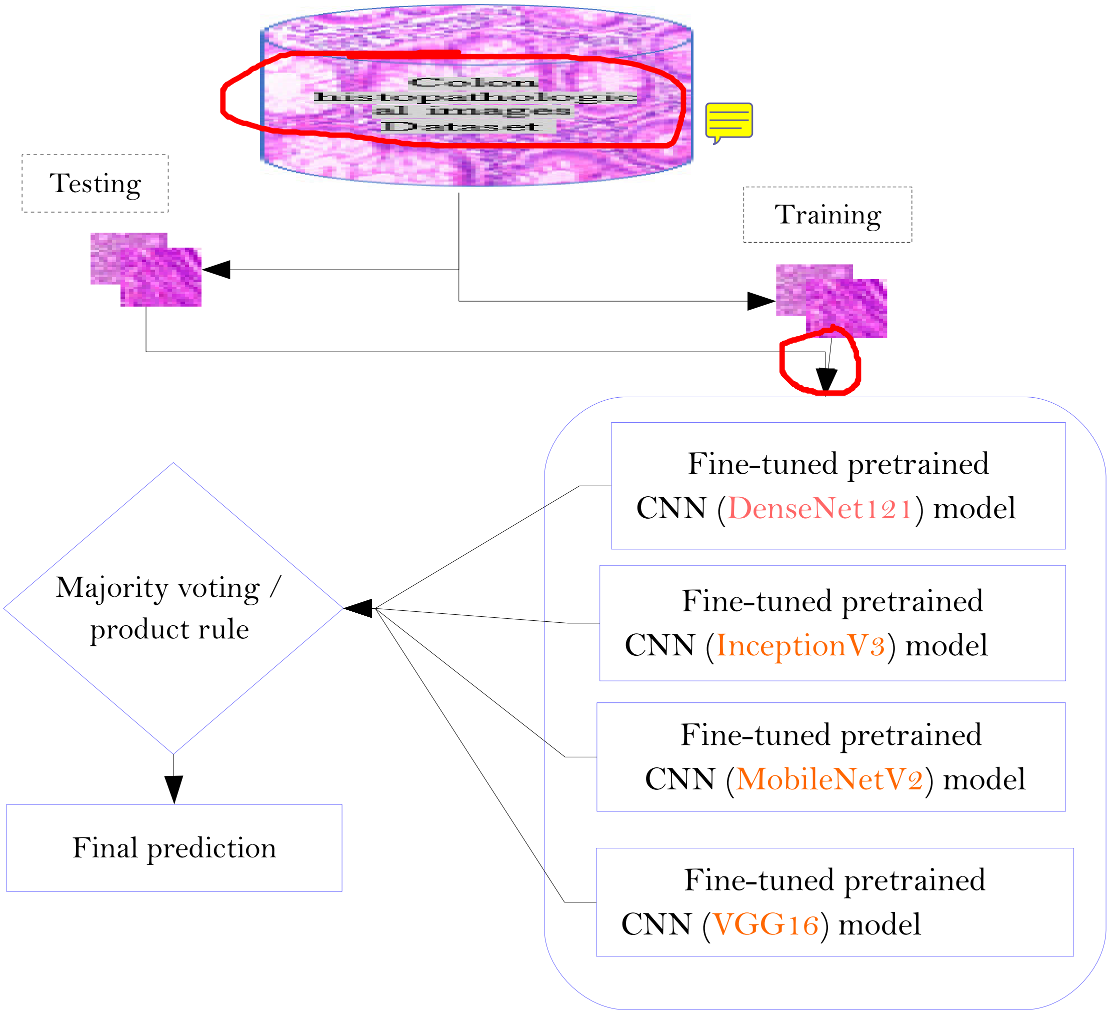

Colon histopathological images Dataset
Testing
Training
Fine-tuned pretrained CNN (DenseNet121) model
Fine-tuned pretrained CNN (InceptionV3) model
Fine-tuned pretrained CNN (MobileNetV2) model
Fine-tuned pretrained CNN (VGG16) model
Majority voting / product rule
Final prediction

Figure 5. The proposed E-CNNS with the four adaptive pretrained models (DenseNet121 MobileNetV2 InceptionV3, and VGG16)

loading the dataset into the Google drive.

EXPERIMENTS RESULTS AND DISCUSSION

This section outlines the experiments and evaluation results from the (E-CNN) and its individual models presented in this research. This section also entails a synopsis of the training and test datasets. The results using the proposed E-CNN, with majority voting and product rule, other standard pretrained models, and state-of-the-art colon cancer classification methods are also presented in this section. Comparisons between the proposed E-CNN and other CNN models from scratch are presented in this section.

Dataset

For experimental evaluation of the proposed E-CNNs model for colon diagnosis from histopathological images, we have employed the benchmark colon H&E from (Stoean et al., 2016).The histopathological images were collected from the Hospital of Craiova, Romania. The benchmark dataset consist of 357 histopathological H&E of normal grade (grade 0) and for cancer grades (grades 1, 2, and 3), with 10x magnification. They have a similar 800x600 pixels resolution. The images' distribution for the classes is as follows: Grade 0: 62 images, grade 1: 96 images, grade 2: 99 images, and grade 3: 100 images. All images are RGB color 8-bit depth with JPEG format. Figure 1 shows some samples from the images and how they are close to each other in the structure, which discriminates between various complicated grades.

Table 3. Hyperparameters used in the proposed individual transfer learning models and an ensemble model.

Hyperparameters	Value
Image size	224 × 224
Optimizer	0.005
Maximum Habitat probability	SGD with momentum
Learning rate	1e-6
Batch size	16
Number of epochs	10
Dropout	0.3
Loss function	Cross Entropy

Experimental Setting

As the proposed E-CNN aims to assist in diagnosing colon cancer based on the histopathological images, the benchmark dataset in (Stoean et al., 2016) is considered during the experiments' work. The dataset was divided into 80% training and 20% testing. In E-CNN, the Hyperparameters, as illustrated in Table 3, were fine-tuned with the same setting for all the proposed transfer learning models. The training and testing images were resized to 224 × 224 for comfort with the proposed transform learning models. The batch size was chosen as 16; the minimum learning rate was specified as min_lr=0.000001. The learning rate was determined to be small enough to slow down learning in the modelsPopa (2021); Kaur and Gandhi (2020). The number of epochs was selected as 10. These models were trained by stochastic gradient descent(SGD) with momentum. all the proposed TL models emploed the cross entrotpy (CE)as the loss function. The cross-entropy is mainly utilized to estimate the distance between the prediction likelihood vector(E) and the one-hot-encoded ground truth label(T)(Boumaraf et al., 2021) probability vector(The following equation depicts the CE Eq.1:

$$CE(E,T) = -\sum_{t=1} T_i \log E_i \tag{1}$$

where CE is used to tell how well the output E matches the ground truth T. Furthermore, the dropout layer was added to all the proposed TL models to avoid over-fitting affair during training. As a result, it drops the activation randomly during the training phase and avoiding units from over co-adapting (Boumaraf et al., 2021). In this study, dropout was set to 0.3 to randomly drop out the units with a probability of 0.3, which is typical when introducing the dropout in deep learning models.

Evaluation Criteria

In this work, multiclass (four-class) classification tasks have been carried out using the base classifiers and their ensembles on the benchmark colon dataset(Stoean et al., 2016). The obtained results have been evaluated using accuracy, sensitivity, specificity. All of these metrics are counted based on the confusion matrix, which includes the true negative (TN) and true positive (TP) values. TN and TP symbolize the acceptably classified benign and malignant samples, respectively. The false negative (FN), and false positive (FP) denote the wrong classified malignant and benign samples. These metrics are designed as follows:

- The average classification accuracy: The correctly categorized TP and TN numbers combined with the criterion parameter, are generally referred to as accuracy. A technique's classification accuracy is measured in Equation 2 as follows:

$$Acc = \frac{1}{M}\sum_{j=1}^{M}\frac{TP+TN}{TP+TN+FP+FN}, \tag{2}$$

where M is the number of independent runs of the proposed ECNN with its individual.

- Average sensitivity: Sensitivity is also called recall. It represents the proportion of positive samples, which are efficiently determined as described in Eq. 3:

$$Sensitivity = \frac{1}{M}\sum_{j=1}^{M}\frac{TP}{TP+FN} * 100\%, \tag{3}$$

The sensitivity value is between [0, 1] scale. One shows the ideal classification, while zero shows the worst classification possible. Multiplication by 100 is applied on the sensitivity to obtain the required percentage.

- Average Specificity: Specificity represents an evaluation metric that is provided for negative samples within a classification approach. In particular, it attempts to measure the negative samples' proportion, which is efficiently classified. Specificity is computed as Eq. 4:

$$Specificity = \frac{1}{M}\sum_{j=1}^{M}\frac{TN}{TN+FP} * 100\% \tag{4}$$

Results and discussion

This subsection presents the experimental results obtained from the proposed E-CNN and its individuals. These results are compared to the classification accuracy results using standard pretrained models (e.g., DenseNet, MobileNet, VGG16, and InceptionV3). After that, the performance of the standard pretrained models was compared to the adaptive pretrained models' performance to evaluate the influence of block-wise fine-tuning policy. The proposed E-CNN was also compared with the state-of-the-art CNN models for colon cancer classification such as (Postavaru et al., 2017; Stoean, 2020; Popa, 2021). In the end, to assess the significance of the proposed E-CNN, statistical test methods were used to verify whether there is a statistically significant difference between the performance of the E-CNN and the performance of the state-of-the-art CNN models.

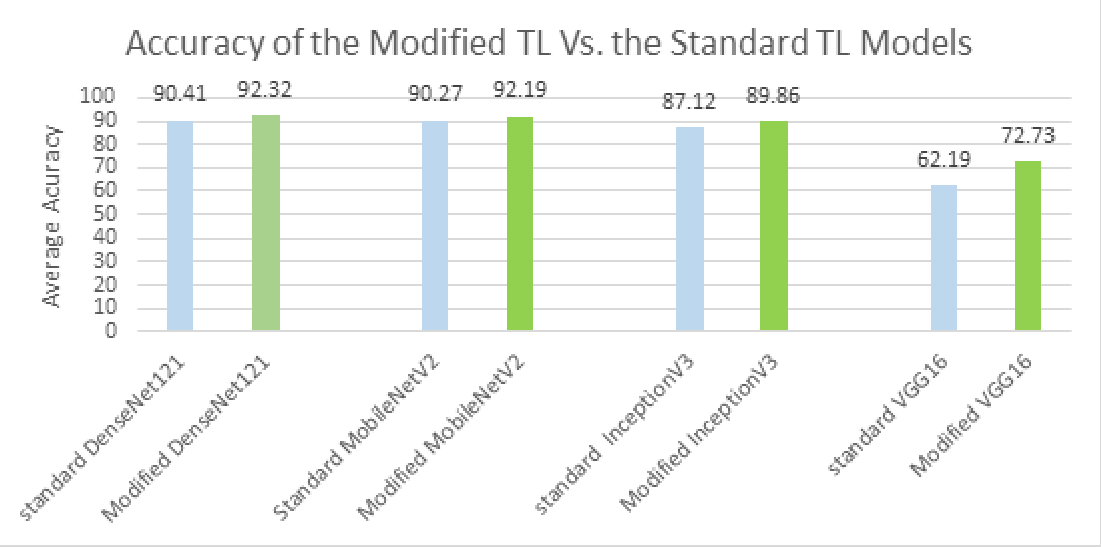


Figure 6. A comparison of modified TL models with standard TL models (original) in terms of average classification accuracy.

The experimental results of this study were built based on the average runs. Mainly, ten separate experiments were used to obtain the average value. The experiments were carried out on the benchmark colon histopathological images dataset of benign, grade1, grade2, and grade3. The classification tasks were accomplished using individual classifiers of new transfer learning set (Modified DenseNet121, Modified

Table 4. Evaluation results for the proposed E-CNN and its individuals (modified TL models) on colon histopathlogical images dataset

Modified pretrained Models	Accuracy	Sensitivity	Specificity
Modified DenseNet121	92.32±2.8	92.99±2.8	100±0.0
Modified MobileNetV2	92.19±3.8	90.75±2.0	100±0.0
Modified InceptionV3	89.86±2.2	95.0±1.5	100±0.0
Modified VGG16	72.73±3.9	73.0±3.6	87.43±12.4
Proposed E-CNN(product)	**95.20±1.64**	**95.62±1.50**	**100±0.0**
Proposed E-CNN (Majority voting)	94.52±1.73	95.0±1.58	100±0.0

Table 5. Evaluation results for the standard TL models on colon histopathlogical images dataset

Standard model without modification	Accuracy	Sensitivity	Specificity
DenseNet121	90.41±3.1	91.25±2.9	100±0
MobileNetV2	90.27±2.9	88.25±1.9	99.23±2.3
InceptionV3	87.12±2.0	92.75±2.0	100±0
VGG16	62.19±7.0	63.21±7.3	100±9.9

MobileNetV2, Modified InceptionV3, and Modified VGG16). The softmax of the fully connected layers of these transfer learning set is used as the classification algorithm. Then, the ensemble (E-CNN) was obtained via product and majority voting aggregation methods. To illustrate the proposed E-CNN performance, the average accuracy, sensitivity, and specificity over the ten runs are used for evaluation on the testing dataset. Besides, their stand deviation (STD) for each base classifier and the E-CNN were also used to estimate the effectiveness of the proposed E-CNN. The experimental results of the proposed E-CNN and its individuals (i.e., base classifiers) are shown in Tables 4,5,and 6, and in Figures 6, 7, and 8 , respectively.

Table 4 shows the performance of the proposed individual TL models and their ensemble (E-CNN) using product rule and majority voting techniques. According to this table, the proposed DenseNet121 transfer learning model achieved approximately 92% test accuracy, which was the highest rate among the other individuals (MobileNetV2, InceptionV3, and VGG16). This is due to the modified DenseNet121 architecture's custom design, which aids in extracting discriminating features from the input colon histopathological images, and which can distinguish between different classes in this domain. Among the four proposed individual pretrained models, the modified VGG16 obtained the lowest accuracy of 79% for the multiclass classification task. This could be the explanation for VGG16's limited number of layers (16), which prevents it from extracting abstract characteristics during the training phase. By analyzing and comparing each individual's results using the modified architecture as depicted in Figure 4(mythology), one can observe that all the proposed modified TL models outperform with superior improvement. For example, the modified DenseNet121 and MobileNetV2 attained around 92% accuracy. Furthermore, the STD is minimum for the DenseNet121, demonstrating its stability and ability to get optimal results regardless of the randomness utilized.

The above proposed modified TL models used the softmax classifier for the fully connected layers to classify the colon histopathlogical image. Margining the decision of these models, in this study, then fusing in the proposed Ensemble E-CNN, where their output probabilities were combined using two aggregation methods: majority voting and product rule. The proposed E-CNN(majority voting) and

Table 6. Summary of the major classification studies on colon cancer

Authors in	Dataset used	CNN architecture	Accuracy	T-test/p-value
(Stoean, 2020)	colorectal in (Stoean et al., 2016)	CNN model from scratch	92%	P<0.0001
(Popa, 2021)	colorectal in (Stoean et al., 2016)	AlexNet	89.53%	P<0.0001
(Popa, 2021)	colorectal in (Stoean et al., 2016)	GoogleNet	85.62%	P<0.0001
(Postavaru et al., 2017)	colorectal in (Stoean et al., 2016)	CNN model from scratch	91%	P<0.0001
Proposed E-CNN (product rule)	colorectal in (Stoean et al., 2016)	**Modified TL models with ensemble learning (using product rule)**	**95.20%**	
Proposed E-CNN (Majority voting)	colorectal in (Stoean et al., 2016)	Modified TL models with ensemble learning (using majority voting)	94.52%	

E-CNN(product rule) achieved accuracy of 94.5% and 95.2%, respectively. These accuracy values were higher compared to the individual models. For example, the result of the E-CNN(product rule) showed a percentage increase of 3.2% compared to the modified DenseNet121 model. This result reveals the significance of the product rule in the proposed E-CNN for colon image classification because it is based on the independent event(Albashish et al., 2016).

As a classification task, the proposed TL models have a loss function (error function). It quantifies the cost of a particular set of network parameters based on how often they generate output in comparison to the ground truth labels in the training set. The TL models employ SGD to determine the optimal set of parameters for minimizing the loss error. Figure 7 and8 depicts the proposed TL models' accuracies and loss curves for the training and testing datasets over ten epochs. Figure 7 shows that the proposed DenseNet and Inception models achieved good accuracy for the training and test datasets over various epochs, while the MobileNet and VGG15 models performed adequately. One possible explanation is that the proposed models are stable during the training phase, allowing them to converge to the best effect. The Densenet121 loss curve indicates that the training loss dropped significantly much faster than the VGG16 and that the testing accuracy improved much faster. In more detail, the VGG16 loss function was linearly reduced, whereas the DenseNet loss function was dramatically reduced. This is consistent with Densenet121's classification performance in Table 4, where it outperformed all other proposed models. Furthermore, one can see that all of the proposed TL models, except the VGG16, achieved high testing accuracies. These models improve the generalization performance simultaneously.

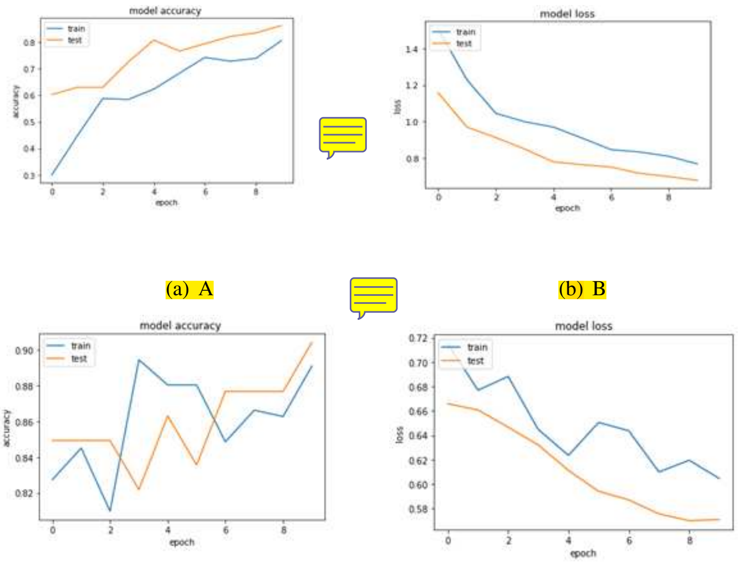
To show the adequacy of the proposed E-CNN, sensitivity was computed. Table 4 confirmations that the E-CNN had higher sensitivity values than the individual model. It worths noting that the sensitivity performance values were also in accordance with the accuracy values, thereby emphasizing the consistency of the E-CNN results. E-CNN(product rule) was able to yield better sensitivity value (95.6%). Among all the proposed transfer learning models, InceptionV3 yielded overall maximum sensitivity performance. Besides, the specificity measure shows that the E-CNN and its individual are able to detect negative samples which are correctly classified for each class.

To demonstrate the importance of the modified TL models in colon histological image classification, the proposed TL models' results were compared to the standard TL models' results (original models without adding any extra layers). The average classification results of the traditional TL models versus the modified TL models are shown in Tables 4, 5 and Figure 6. For the four proposed pretrained models, it was clear that modified TL models performed significantly better than standard models. As visually evaluated, the margin difference was statistically significant.

The modified DenseNet121 model had a relatively high average accuracy rate of 92.3%, which was significantly more prominent than the accuracy rate of 90 percent. The major difference between the two of 2.2 was significant. In more detail, the average accuracy rate difference between the modified VGG16 and the standard VGG16 was more than 10%, which was large and statistically significant. This astounding level of performance of the modified models could be attributed to the adaptation layers' ability to find the most abstract features, which aid the fully connected layers and softmax classifier in discriminating between various grades in colon histopathological. As a result, it reduces the problem of inter-class classification.

To further demonstrate the efficacy of the proposed E-CNNs, we also compare the obtained results on the colon histopathlogical images benchmark dataset with the most recent related works(Stoean, 2020; Popa, 2021). Table 6 contains the comparison between the proposed E-CNNs and the recent state-of-the-art studies. From the tabular results, one see that the proposed E-CNNs achieved higher results comparing either to pretrained models, such as in (Stoean et al., 2016), or to constructing CNN from scratch. One of the main reasons for these superior results is the adaptation of the transfer learning models with the appropriate layers; additionally, using ensemble learning demonstrates the ability of the proposed E-CNNs to increase discriminations between various classes in the histopathological colon images dataset. In comparison to the recent study by Stoean et al. (Stoean, 2020), our ensemble E-CNNs has shown better performance. They constructed CNN from scratch and then used Evolutionary Algorithms (EA) to fin-tune its parameters. Their classification accuracy on the colon histopathlogical images dataset was 92%. We find that the proposed method's superior due to expanding deeper architecture and utilizing ensemble learning.

Moreover, the obtained classification accuracies were compared with the pretrained models GoogleNet and AlexNet in (Popa, 2021). The proposed method exceeded the pretrained models. The classification accuracy of GoogleNet and AlexNet on colon histopathological images was 85.62% and 89.53%, respectively. The average accuracy rate difference between the proposed method and these pretrained models was more than 10% and 6%, respectively, which was large and statistically significant. Two critical observations are to be made here: first, adapting pretrained models to a specific task increases performance. Second, using pretrained models as a feature extraction without the softmax classifier may degrade the classification accuracy in the colon histopathological image dataset.



(a) A

(b) B

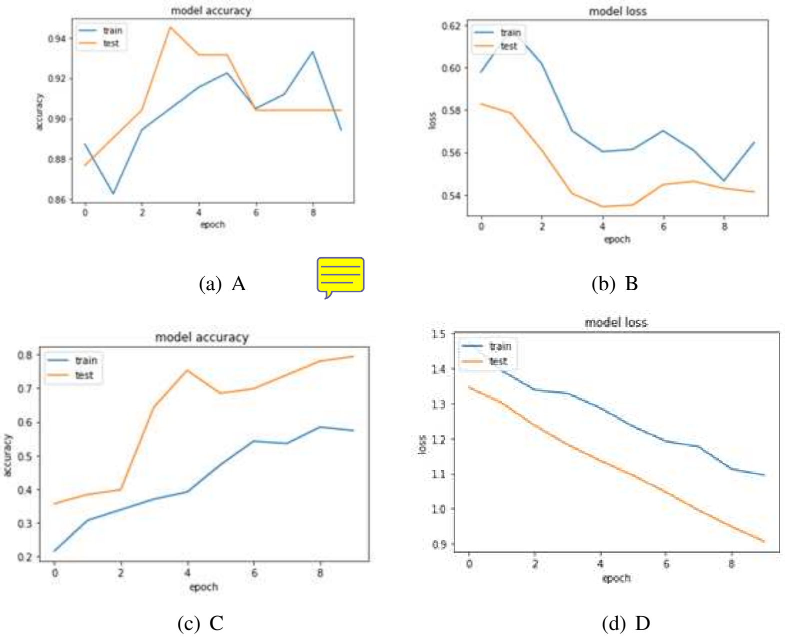
(c) C

(d) D

Figure 7. Learning curves for (a) Modified DenseNet: training and test accuracy, and (b) Modified DenseNet: training and test loss (c) Modified Inception: training and test accuracy (D) Modified Inception: training and test loss using the colon histopathological image benchmark dataset.

Discussion

According to the above experimental results, it is clear that the proposed E-CNNs and adapted TL predictive models outperform other state-of-the-art models and standard TL models in the colon histopathological image classification task. The experimental results indicate that adapting the TL models for medical image classification tasks improves classification tasks. The results in Tables 4 and 5 demonstrate the critical importance of the introduced adapted TL models (DenseNet121, MobileNetV2, InceptionV3, and VGG16) in comparison to conventional TL methods. For example, the adaptive DenseNet model outperformed the standard DenseNet model. These findings show that tailoring the TL models to a specific



(a) A

(b) B

(c) C

(d) D

Figure 8. Learning curves for (a) Modified MobileNetV2: training and test accuracy, and (b) Modified MobileNetV2: training and test loss (c) Modified VGG16: training and test accuracy (D) Modified VGG16: training and test loss using the colon histopathological image benchmark dataset.

classification task can boost performance. It has also been experimentally verified that using TL models in medical image classification results in superior performance when compared to training CNN from scratch (as in previous works by (Postavaru et al., 2017) and (Stoean, 2020)). One reason for this finding is that training a CNN from scratch would necessitate a large number of training samples. Moreover, it must be confirmed that the large number of parameters of the CNN are trained effectively and with a high degree of generalization to obtain acceptable results. Thus, the limitation of the number of training samples causes overfitting in classification tasks. Furthermore, based on the results, it was found that the selection of appropriate hyperparameters in TL models plays a vital role in the proper learning and performance of the TL model.

In this study, two ensemble learning (E-CNN (Majority voting), E-CNN (product rule)) models had been designed to further boost the colon histopathological image classification performance. In the proposed ensemble learning models, the adaptive TL models were used as base classifiers. Through the experimental results, one can find that ensemble learning outperformed using individual classifiers. Furthermore, using product rules in the ensemble allows the probabilities of independent events to be fused, ultimately improving performance. This finding is in line with the results in Table 4, where the proposed E-CNN (product) outperformed the proposed E-CNN (majority voting).

Furthermore, the T-test was computed to compare the proposed E-CNN (product) with the most related works. The reached statistics were carried out using the proposed E-CNN (product), which relied on the accuracy of the results specific to the colon dataset. By handling a T-test with a 95% spectrum of significance (alpha =0.05) on the collected p-values and the classification accuracy, the corresponding difference statistics are shown in Table 6. As shown in Table 6, the proposed E-CNN (product) outperforms most of the related works on the colon histopathological image dataset, where the majority of the p-values of$<$ 0.0001. For example, comparing the proposed E-CCNN with CNN from scratch in (Stoean, 2020), E-CNN is significantly better with a p-value $<$0.001. These findings show that using the E-CNN (product) is effective for handling medical image classification tasks.

In summary, it has been demonstrated that the use of the proposed TL models assists in the colon histopathological image classification task, which can be used in the medical domain. Besides, using ensemble learning for the machine learning classification tasks can improve the classification results.

CONCLUSION AND FUTURE WORK

Deep learning plays a key role in diagnosing colon cancer by grading captured images from colon histopathological images. In this study, we introduced a new set of transfer learning-based methods to help classify colon cancer from histopathological images, which can be used to discriminate between different classes in this domain. To solve this classification task, the pre-trained CNN models DenseNet121, MobileNetV2, InceptionV3, and VGG16 were used as backbone models. We introduced the TL technique based on a block-wise fine-tuning process to transfer learned experience to colon histopathological images. To accomplish this, we added new dense and drop-out layers to the pretrained models, followed by new FCC with softmax layers to handle the four-class classification task. The adaptability of the proposed models has been enhanced further by the utilized ensemble learning. Two deep ensemble learning methods (E-CNN(product) and E-CNN (Majority voting),)have been proposed. The adapted pretrained models were used as individual classifiers in these proposed ensembles. Next, their output probabilities were fused using the majority voting and the product rule. The acquired results revealed the efficiency of the suggested E-CNNs and their individuals.

We achieved accuracy results of 95.20% and 94.52% for the proposed E-CNN (product) and E-CNN (majority voting), respectively. The proposed E-CNNs and its individual performances were evaluated and compared against the standard(without adaption)pretrained models(DenseNet121, MobileNetV2, InceptionV3, and VGG16 models) as well as state-of-the-art pretrained models and CNN from scratch on colon histopathological images. On all evaluation metrics and the colon histopathological images benchmark dataset, the proposed E-CNNs considerably outperformed the standard pretrained and state-of-the-art CNN from scratch models. The results indicate that the adaptation of pretrained models for TL is a viable option for dealing with the limited number of samples in any new classification task. As a result, the findings indicate that E-CNNs are being used in diagnostic pathology to assist pathologists in making final decisions and accurately diagnosing colon cancer.

Future research could be considered to introduce a new strategy to select the best hyperparameters for the adaptive pretrained models—we recommend wrapper methods for this task.

REFERENCES

Ahmad, F., Farooq, A., and Ghani, M. U. (2021). Deep ensemble model for classification of novel coronavirus in chest x-ray images. *Computational intelligence and neuroscience*, 2021.

Albashish, D., Al-Sayyed, R., Abdullah, A., Ryalat, M. H., and Almansour, N. A. (2021). Deep cnn model based on vgg16 for breast cancer classification. In *2021 International Conference on Information Technology (ICIT)*, pages 805–810. IEEE.

Albashish, D., Sahran, S., Abdullah, A., AbdShukor, N., and Hayati Md Pauz, S. (2016). Ensemble learning of tissue components for prostate histopathology image grading. *International Journal on Advanced Science, Engineering and Information Technology*, 6(6):1132–1138.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74.

Belciug, S. and Gorunescu, F. (2020). Data mining-based intelligent decision support systems. In *Intelligent Decision Support Systems—A Journey to Smarter Healthcare*, pages 103–258. Springer.

Benhammou, Y., Achchab, B., Herrera, F., and Tabik, S. (2020). Breakhis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. *Neurocomputing*, 375:9–24.

Bhargava, S. (2013). A note on evolutionary algorithms and its applications. *Adults Learning Mathematics*, 8(1):31–45.

Boruz, D. B. M. and Stoean, C. (2018). On supporting cancer grading based on histological slides using a limited number of features. *Annals of the University of Craiova-Mathematics and Computer Science Series*, 45(1):156–165.

Boumaraf, S., Liu, X., Zheng, Z., Ma, X., and Ferkous, C. (2021). A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images. *Biomedical Signal Processing and Control*, 63:102192.

Deniz, E., Şengür, A., Kadiroğlu, Z., Guo, Y., Bajaj, V., and Budak, Ü. (2018). Transfer learning based histopathologic image classification for breast cancer detection. *Health information science and systems*, 6(1):1–7.

Dif, N. and Elberrichi, Z. (2020a). A new deep learning model selection method for colorectal cancer classification. *International Journal of Swarm Intelligence Research (IJSIR)*, 11(3):72–88.

Dif, N. and Elberrichi, Z. (2020b). A new intra fine-tuning method between histopathological datasets in deep learning. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, 11(2):16–40.

Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J., et al. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11):1122–1132.

Garbin, C., Zhu, X., and Marques, O. (2020). Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, pages 1–39.

Ghosh, S., Bandyopadhyay, A., Sahay, S., Ghosh, R., Kundu, I., and Santosh, K. (2021). Colorectal histology tumor detection using ensemble deep neural network. *Engineering Applications of Artificial Intelligence*, 100:104202.

Harangi, B. (2018). Skin lesion classification with ensembles of deep convolutional neural networks. *Journal of biomedical informatics*, 86:25–32.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Kalkan, H., Nap, M., Duin, R. P., and Loog, M. (2012). Automated classification of local patches in colon histopathology. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 61–64. IEEE.

Kather, J. N., Weis, C.-A., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., Marx, A., and Zöllner, F. G. (2016). Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1):1–11.

Kaur, T. and Gandhi, T. K. (2020). Deep convolutional neural networks with transfer learning for automated brain image classification. *Machine Vision and Applications*, 31(3):1–16.

Khadilkar, S. P. (2021). Colon cancer detection using hybrid features and genetically optimized neural network classifier. *International Journal of Image and Graphics*, page 2250024.

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516.

Krishnamurthy, S., Srinivasan, K., Qaisar, S. M., Vincent, P., and Chang, C.-Y. (2021). Evaluating deep neural network architectures with transfer learning for pneumonitis diagnosis. *Computational and Mathematical Methods in Medicine*, 2021.

Kumar, A., Singh, S. K., Saxena, S., Lakshmanan, K., Sangaiah, A. K., Chauhan, H., Shrivastava, S., and Singh, R. K. (2020). Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. *Information Sciences*, 508:405–421.

Li, M., Wu, L., Wiliem, A., Zhao, K., Zhang, T., and Lovell, B. (2019). Deep instance-level hard negative mining model for histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 514–522. Springer.

Lichtblau, D. and Stoean, C. (2019). Cancer diagnosis through a tandem of classifiers for digitized histopathological slides. *PloS one*, 14(1):e0209274.

Mahbod, A., Schaefer, G., Wang, C., Dorffner, G., Ecker, R., and Ellinger, I. (2020). Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. *Computer methods and programs in biomedicine*, 193:105475.

Malik, J., Kiranyaz, S., Kunhoth, S., Ince, T., Al-Maadeed, S., Hamila, R., and Gabbouj, M. (2019). Colorectal cancer diagnosis from histology images: A comparative study. *arXiv preprint arXiv:1903.11210*.

Mishra, A. K., Das, S. K., Roy, P., and Bandyopadhyay, S. (2020). Identifying covid19 from chest ct images: a deep convolutional neural networks based approach. *Journal of Healthcare Engineering*, 2020.

Ohata, E. F., das Chagas, J. V. S., Bezerra, G. M., Hassan, M. M., de Albuquerque, V. H. C., and Reboucas Filho, P. P. (2021). A novel transfer learning approach for the classification of histological images of colorectal cancer. *The Journal of Supercomputing*, pages 1–26.

Popa, L. (2021). A statistical framework for evaluating convolutional neural networks. application to colon cancer. *Annals of the University of Craiova-Mathematics and Computer Science Series*, 1(48):159–166.

Postavaru, S., Stoean, R., Stoean, C., and Caparros, G. J. (2017). Adaptation of deep convolutional neural networks for cancer grading from histopathological images. In *International Work-Conference on*

Artificial Neural Networks, pages 38–49. Springer.

Rachapudi, V. and Lavanya Devi, G. (2021). Improved convolutional neural network based histopathological image classification. *Evolutionary Intelligence*, 14(3):1337–1343.

Sahran, S., Albashish, D., Abdullah, A., Shukor, N. A., and Pauzi, S. H. M. (2018). Absolute cosine-based svm-rfe feature selection method for prostate histopathological grading. *Artificial intelligence in medicine*, 87:78–90.

Saini, M. and Susan, S. (2020). Deep transfer with minority data augmentation for imbalanced breast cancer dataset. *Applied Soft Computing*, 97:106759.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srinivasu, P. N., SivaSai, J. G., Ijaz, M. F., Bhoi, A. K., Kim, W., and Kang, J. J. (2021). Classification of skin disease using deep learning neural networks with mobilenet v2 and lstm. *Sensors*, 21(8):2852.

Stoean, C., Stoean, R., Sandita, A., Ciobanu, D., Mesina, C., and Gruia, C. L. (2016). Svm-based cancer grading from histopathological images using morphological and topological features of glands and nuclei. In *Intelligent interactive multimedia systems and services 2016*, pages 145–155. Springer.

Stoean, R. (2020). Analysis on the potential of an ea–surrogate modelling tandem for deep learning parametrization: an example for cancer classification from medical images. *Neural Computing and Applications*, 32(2):313–322.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Talo, M. (2019). Automated classification of histopathology images using transfer learning. *Artificial Intelligence in Medicine*, 101:101743.

Tang, J., Rangayyan, R. M., Xu, J., El Naqa, I., and Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *IEEE transactions on information technology in biomedicine*, 13(2):236–251.

Tripathi, S. and Singh, S. K. (2020). Ensembling handcrafted features with deep features: an analytical study for classification of routine colon cancer histopathological nuclei images. *Multimedia Tools and Applications*, 79(47):34931–34954.

Yang, Z., Ran, L., Zhang, S., Xia, Y., and Zhang, Y. (2019). Ems-net: ensemble of multiscale convolutional neural networks for classification of breast cancer histology images. *Neurocomputing*, 366:46–53.

Zhao, B., Huang, B., and Zhong, Y. (2017). Transfer learning with fully pretrained deep convolution networks for land-use classification. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1436–1440.

Zhi, W., Yueng, H. W. F., Chen, Z., Zandavi, S. M., Lu, Z., and Chung, Y. Y. (2017). Using transfer learning with convolutional neural networks to diagnose breast cancer from histopathological images. In *International Conference on Neural Information Processing*, pages 669–676. Springer.