

COMBINING SALIENCY WITH PREDICTION FOR
ENDOSCOPIC DIAGNOSIS

by

Mahmoud Tarek Mahmoud Rashad Abdelaziz Rezk

A Thesis presented to the Faculty of the
American University of Sharjah
College of Engineering
In Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in
Electrical Engineering

Sharjah, United Arab Emirates

May 2020

Declaration of Authorship

I declare that this thesis is my own work and, to the best of my knowledge and belief, it does not contain material published or written by a third party, except where permission has been obtained and/or appropriately cited through full and accurate referencing.

Signature: Mahmoud Rezk

Date: 15/5/2020

The Author controls copyright for this report.
Material should not be reused without the consent of the author. Due
acknowledgement should be made where appropriate.

© Year 2020

Mahmoud Tarek Rezk

ALL RIGHTS RESERVE

Approval Signatures

We, the undersigned, approve the Master's Thesis of Mahmoud Tarek Mahmoud
Rashad Abdelaziz Rezk.

Thesis Title: Combining Saliency with Prediction for Endoscopic Diagnosis.

Date of Defense: 13/05/2020

Name, Title and Affiliation

Signature

Dr. Usman Tariq
Assistant Professor, Department of Electrical
Engineering
Thesis Advisor

Dr. Abhinav Dhall
Lecturer, Monash University
Assistant Professor, Indian Institute of Technology Ropar
Thesis Co-Advisor

Dr. Hasan Al-Nashash
Professor, Department of Electrical Engineering
Thesis Co-Advisor

Dr. Hasan Mir
Professor, Department of Electrical Engineering
Thesis Committee Member

Dr. Salam Dhou
Assistant Professor, Department of Computer Science and
Engineering
Thesis Committee Member

Dr. Nasser Qaddoumi
Head
Department of Electrical Engineering

Dr. Lotfi Romdhane
Associate Dean for Graduate Studies and Research
College of Engineering

Dr. Sirin Tekinay
Dean
College of Engineering

Dr. Mohamed El-Tarhuni
Vice Provost for Graduate Studies
Office of Graduate Studies

Acknowledgements

I would like to first thank Allah for the chance to complete my masters in AUS. Then I would like to thank my parents for their continuous motivation support throughout my journey.

I would like to thank my advisors Dr. Usman Tariq, Dr. Abhinav Dhall and Dr. Hasan Al-Nashash for providing knowledge, guidance, support, and motivation throughout my research stages. I am deeply beholden for their great assistance, worthy discussions, and suggestions.

I would like to thank the professors of the Electrical Engineering Department who taught me the master level courses that allowed me to gain the necessary skills and knowledge to perform this research. I really appreciate their dignified advice and motivation.

Finally, I would like to thank the American University of Sharjah for the Graduate Teaching Assistant opportunity.

Abstract

Healthcare sector has advanced tremendously in the past few years. With the advancement in technology, many image diagnostic techniques have been introduced to help doctors in identifying diseases and abnormalities inside the human body. However, the increase in population and access to affordable healthcare have increased the patient population significantly, which requires a bigger infrastructure in medical diagnostics. The demand and supply imbalance of expert doctors in the field had led to the increase in healthcare bills. As a solution to the scarcity problem, one of the advancements that has been introduced to this sector is automated diagnostics using artificial intelligence (AI). The automated systems are made to help doctors in two ways. Firstly, they decrease the time required by the doctor to diagnose the patient and, secondly, they act as a second layer of diagnostic verification. This thesis aims to automate the classification of endoscopic images to eight disease and non-disease classes using a deep network architecture that would detect the salient region and classify the images accordingly. This thesis further studies the effect of jointly performing both tasks on the overall quality of attention masks and the classification results. The automated system is achieved by concatenating a U-net architecture to a dense-net architecture to jointly predict the salient medical masks and classify them to their respective classes. Furthermore, the automated system proved that medical image masks can be achieved by transfer learning the knowledge learned from natural images. Additionally, jointly predicting the masks and reusing the masks for classification demonstrated that the joint behavior would increase the classification accuracy.

Keywords: *Autoencoders, Convolutional Neural Networks, Deep learning, Endoscopy, Medical Diagnosis*

Table of Contents

Abstract	5
List of Figures	8
List of Tables	9
List of Abbreviations	11
Chapter 1. Introduction	12
1.1 Thesis Objectives	13
1.2 Research Contribution.....	13
1.3 Thesis Organization.....	13
Chapter 2. Background and Literature Review.....	15
2.1 Deep Neural Networks	15
2.2 Convolutional Neural Networks.....	15
2.3 Autoencoder	17
2.4 Related Work in Medical Field	17
2.4.1 Lesion segmentation.	17
2.4.2 Lesion classification.....	19
2.4.3 GI tract diseases and detection.....	20
Chapter 3. Methodology	24
3.1 Saliency Detection.....	24
3.1.1 Saliency datasets.	24
3.1.2 Saliency architectures.	25
3.1.3 Input preprocessing.....	27
3.2 GI Classification.....	29
3.2.1 GI classification dataset.	29
3.2.2 Basic classification network.	30
3.2.3 Transfer learning.....	32
3.2.4 Proposed architecture for medical image classification.....	32

3.3	Cost Functions and Evaluation Metric	33
3.4	The Procedure Summary	35
Chapter 4. Results and Discussion.....		37
4.1	Saliency Prediction.....	37
4.1.1	Selection of U-net architecture.	37
4.1.2	Saliency mask from input combinations.....	40
4.2	Combined Saliency predictions with labels classification	44
4.2.1	Concatenated RGB input.	45
4.2.3	Concatenated HS input.	51
4.2.4	Multiplied HS input.	51
4.3	Five-Fold Cross Validation	52
Chapter 5. Conclusion and Future Work		54
References.....		56
Appendix.....		61
Vita.....		62

List of Figures

Figure 1: Classification CNN example [4]	16
Figure 2: Auto-encoder	17
Figure 3: Salient image example, right is ground truth [46].....	25
Figure 4: U-net architecture [23]	25
Figure 5: U-net Dense architecture	26
Figure 6: U-net Skip architecture.....	27
Figure 7: Dense-net architecture [53]	31
Figure 8: Proposed final architecture for joint saliency and prediction.....	32

List of Tables

Table 1: Target GI Tract endoscopy images for classification	29
Table 2: Training: DUTS Training validation: DUTS Test metric: MAE & F1	38
Table 3: Training: DUTS Train validation: DUTSOMRON metric: MAE & F1	38
Table 4: Visual output of selected architectures	39
Table 5: Comparison between Liu et al. [55] and current results	39
Table 6: Comparison between U-net with and without max pooling and with strided convolution	40
Table 7: Comparison between different color space inputs on DUTSOMRON	41
Table 8: Mask predicted before and after zoom + crop	41
Table 9: Input image, and mask outputs for RGB, HS and Value	42
Table 10: Different iteration executed	45
Table 11: Results obtained from concatenated RGB model	45
Table 12: Saliency masks labeling illustration	46
Table 13: Saliency mask enumeration results for RGB concatenated network	47
Table 14: Enumerated saliency and classification results for concatenated RGB framework	48
Table 15: Saliency and classification results for multiplied RGB framework	49
Table 16: Examples of distortions caused by multiplications	50
Table 17: Saliency and classification results for concatenated HS framework	51
Table 18: Saliency and classification results for multiplied HS framework	52
Table 19: Five-fold cross validation results	53
Table 20: Training: DUTS Training validation: DUTS Test metric: validation loss	61
Table 21: Training: DUTS Training validation: DUTSOMRON metric: validation loss	61

Table 22: Training: DUTS Training + Test | validation: DUTSOMRON | metric:
validation loss61

List of Abbreviations

AE	Autoencoder
CNN	Convolutional Neural Network
FC	Fully Connected
GI	Gastrointestinal
MAE	Mean Absolute Error
MSE	Mean Square Error

Chapter 1. Introduction

Healthcare sector has developed rapidly in the previous years with the tremendous advancements in technology. The advancements in technology have brought several diagnostic tools and methods for the doctors to help them in assessing the health state of an individual. As a result of the improvements in modern medicine, the average life expectancy of humans has increased in the past century from 47 to 76 years [1]. Additionally, healthcare has become more affordable for the ever-increasing population. Consequently, there has been a disproportionate increase in patient population relative to medical experts. Therefore, increments in the number of experts are needed in the medical field [2]. Artificial Intelligence (AI) can help experts in their diagnostic tasks making it more efficient and reducing their workload [3]. AI have proven its significant contributions in different sectors of the medical field such as disease diagnosis, drug interaction, radiology, and medicine creation.

Continuous researches are made to automate the process of finding various types of diseases and abnormalities in different parts of the body, such as detecting disease from endoscopic images or videos. Endoscopy is the process of detecting diseases and abnormalities in the Gastrointestinal (GI) tract by inserting a camera ending tube inside the body to observe the internal organs and tissues. An example of the significant abnormalities detected by endoscopy is polyps. Polyps are abnormal tissue growth in the GI tract that could develop into cancerous cells if not detected through endoscopy and treated. A challenge associated with endoscopy is the vast number of images or long videos that a specialist must go through to detect and mark the abnormalities investigated in these images. This process is very time consuming and is prone to human vision errors. Hence, AI can help to assist or automate the process of detecting and classifying these diseases. This would contribute in reducing the errors and time needed by the physicians for classification of these diseases. Therefore, the purpose of this research is to program an automated endoscopic image classifier that would be used in identifying the salient part of the GI tract image and classifying it to its corresponding label.

1.1 Thesis Objectives

Motivated by previous contributions done in the field of biomedical image processing, the main aim of this work is to build an automated classifier of GI endoscopic images; additionally, the classifier is further modified to detect salient regions. The purpose of the image classifier is to provide a diagnostic tool that could assist in the medical analysis by highlighting the salient regions in GI tract. Such augmentation would help provide accurate results to the medical examiner while drawing his attention to salient regions which might, also, speed up the process of visual classification. Furthermore, this paper will examine the effect of jointly predicting both saliency and classification on the overall classification results and attention maps predictions.

1.2 Research Contribution

The contribution of this thesis are two-folds: First, we develop a U-net architecture that was built, trained, and tuned to predict saliency map in natural images. The trained model was utilized to predict saliency maps from medical images. From the experimental results, we concluded that saliency networks trained on natural images can also perform well for medical images that look drastically different.

Second, we study the difference in performance between the usage of a classification network, and the usage of a classification network that is merged with a saliency network. To elaborate more, we study the effect of having a network that would predict its own mask from the medical input and utilizes this mask along with the input through further classification layers to predict the class label. This contrasts with, utilizing only the original input in classification layers to predict the class label. After several experiments we concluded that combining saliency prediction and classification network improves the overall classification performance.

1.3 Thesis Organization

The rest of the thesis is organized as follows: Chapter 2 discusses various architectures in deep learning; neural networks, convolutional neural networks, autoencoders and convolutional autoencoders. Additionally, the chapter mentions the contributions of deep learning in the field of medical diagnostics. Subsequently, a brief introduction of endoscopy followed by the contributions of deep learning in the GI tract diagnostics is covered. Chapter 3 illustrates the different datasets, architectures and the

evaluation metrics that will be used in both stages, saliency detection and image classification. Chapter 4 discusses the results achieved from predicting saliency maps of medical images using networks trained on natural image. Furthermore, this chapter discusses the results obtained from jointly detecting saliency maps and predicting class labels from different inputs and architectures variations. Finally, Chapter 5 includes the concluding remarks and the future work to follow.

Chapter 2. Background and Literature Review

This chapter introduces the concept of deep neural networks, convolutional neural networks and autoencoders. Subsequently, the chapter proceeds to the contributions of deep learning in the field of medical diagnostics. Then it advances to discuss different gastrointestinal diseases and the algorithms proposed to detect them.

2.1 Deep Neural Networks

Artificial Neural Networks are inspired from the structure of neurons in the brain. Originally these neural networks were proposed as a set of algorithms that imitate the human brain in recognizing patterns. Patterns in real world data, such as images, sound, and time series, are translated into numerical vectors and then recognized using neural networks. Neural networks are also known for their capability of adapting to changing inputs, where the network can generate satisfying results without changing the output criteria. They are made of layers of interconnected perceptron known as nodes. A perceptron is a simple model of how a biological neuron operates. Each multi-layered perceptron has an input layer, output layer, and intermediate/hidden layer(s). The target of the network is to maximize a performance criterion during learning, and this is done by propagating the errors back through the hidden layer to match the inputs to the desired outputs by changing the weights of the network. The higher the number of levels in the network the deeper it is; hence, the name deep learning represents architectures composed of numerous hidden layers. There are different architectures in deep networks. One more famous in image recognition and processing is known as Convolution Neural Networks (CNNs).

2.2 Convolutional Neural Networks

CNNs are very popular in the field of deep learning and machine learning due to their capability of achieving satisfying results in different applications such as: image recognition, segmentation, saliency detection and many more. From its name it can deduced that the operation of the CNN relies mainly on convolution. An example of a CNN classification network is shown in Figure 1 [4]. The network consists of three main components similar to the feed-forward neural network, which are the input layer/s, hidden layer/s, and the output layer/s. The difference is that the hidden layers constitutes of convolutional layer, pooling layer, sometimes a fully connected layer and

occasionally skipped connections. On the contrary, traditional feed forward neural network consists of fully connected layers, only.

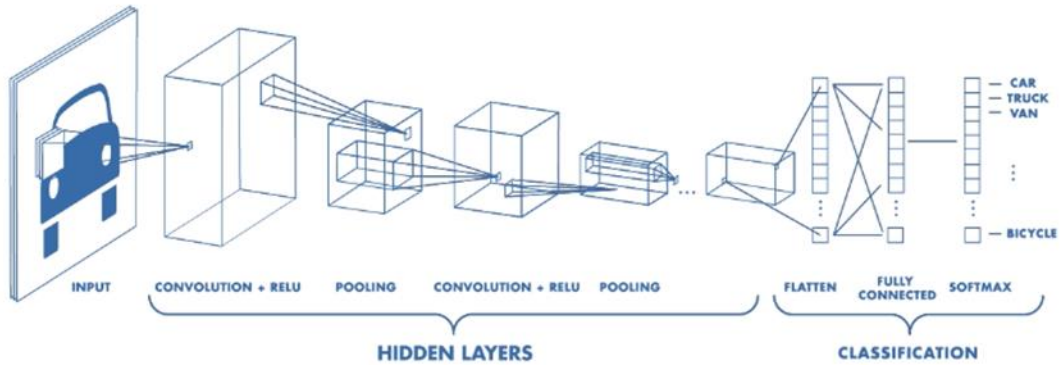


Figure 1: Classification CNN example [4]

Furthermore, each layer in the CNNs is characterized by multiple parameters. For instance, the convolution layer has a kernel/filter size, kernel numbers, and strides. The filter size is used to determine the part of the input that will undergo convolution, and the filter number governs the number of learnt filters that will help to achieve the goal of the network. Finally, the strides determine the number of units that filter window would slide while convolving it [5].

A pooling layer may be used to aggregate the filter responses. Hence it facilitates the extraction of the most important features that would help in later layers. Max pooling is a common pooling that is used heavily in several papers, and it extracts the maximum value on a local neighborhood in feature maps. However, there are other pooling categories such as average and L2-norm pooling.

The fully connected layers are often used at the end of a CNN, especially when the function of the network is classification as seen in Figure 1. In these layers each neuron in one hidden layer is connected to all neurons in the subsequent hidden layer. The output size of the final fully connected layer in a network is usually equal to the number of classes that the network is set to classify. However, if the problem pertains to segmentation or saliency, the CNN may end with a convolutional layer, and the output may be the segmentation labels or saliency maps (to be described later).

2.3 Autoencoder

Autoencoders are self-supervised learning algorithms whose main aim is to retrieve an output value that is similar, to the input with the least distortions [6]. Autoencoders (AE) are made mainly from an encoder (reduction) and a decoder (reconstruction) section. Autoencoders may be used for various kind of data; images are one such example. Its mechanism focuses on extracting the most important features of the image in the encoding process and then reconstruct an output image in the decoding process from the learned important features as shown in Figure 2.

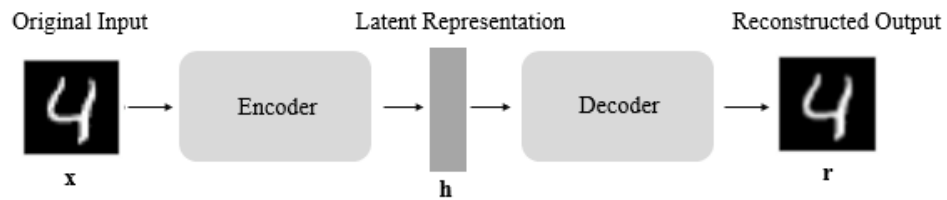


Figure 2: Auto-encoder

The main aim is to teach the neural network the important features that would help in reconstructing the image. There are different types of AE such as denoising AE, convolutional AE, etc. The denoising AE is used mainly to reconstruct an output image from a noisy input image. Additionally, the convolutional AE is made mainly from convolution layers similar to a CNN [7]. The encoder is constructed from convolutional layers along with pooling layers. On the contrary, the decoder is constructed from transposed convolution and up-sampling to counter the encoder effect (i.e. the decoder would retrieve the image back).

2.4 Related Work in Medical Field

This section starts with different contributions of machine learning in medical images various applications such as lesion segmentation and classification. We also discuss selected examples. Then, the procedure of the endoscopy is discussed, followed by the different GI tract diseases and their detection methods.

2.4.1 Lesion segmentation. A lesion is an abnormal appearance in regions of a tissue or an organ due to disease or injury, such as ulcer, abscess, tumor, or a wound. Lesion segmentation is the process of identifying the set of voxels that makes up the interior or the contour of a lesion. To elaborate, the process delineates the boundaries

of a lesion. Lesion segmentation of neuronal structure and prostate volume are some examples discussed in [8-13].

One of the most popular architectures for biomedical segmentation is the U-net architectures made by Ronneberger et al. [8]. The architecture is a Full Convolutional Neural Network (FCNN) and has similar characteristics of an AE in feature extraction, where the output of the network is a segmented image. The network has equal down-sampling using convolution and max pooling and then up sampling using transposed convolution. Furthermore, satisfactory results are obtained with this architecture while having a very few training images. For instance, Ronneberger et al. [8] applied the architecture on Electron Microscope (EM) segmentation challenge dataset which was composed of 30 images of size 512×512 pixels of microscopic cells. The images were augmented, and weighted loss was added to tackle the touching cells. Furthermore, the model's performance was evaluated using warping error, rand error (the recurrence with which a couple of segmentations disagree over two pixels belonging to different or same object) and pixel error (the ratio of the number of disagreeing pixels over the total number of pixels). Additionally, the architecture was also applied on ISBI (International Symposium on Biomedical Imaging) cell neuronal structure tracking challenge 2014 dataset, in which the algorithm achieved an IOU(intersection over union: the overlap area of the predicted bounding box and the ground truth box over the union of these boxes) of 92%. Due to its significant performance, a similar architecture is considered in the work of this paper.

A tuned version of U-net, V-net is introduced by Milletari et al. in [9] for segmenting 3D Prostate images using 3D fully convolutional architecture. Furthermore, the architecture is applied on 50 MRI prostate scans from PROMISE2012 challenge dataset and is used to extract a volume segmented image as an output of the network. The performance of segmentation was evaluated using dice coefficient, and Hausdorff distance. Additionally, a comparison of the performance of the V-net with other competitors is provided in the paper. In this thesis, localizing the endoscope using anatomical landmarks will help us detect, classify, and highlight the lesions.

As for brain lesions, Kamnits et al. [14] proposes a three-dimensional CNN for segmenting traumatic brain injuries, brain tumors and ischemic stroke. Moreover, the network consists of a dual pathway for multiple scale processing of the input image. One pathway is fed with normally sized images to capture the detailed appearance of

the structure of the lesion, and the second pathway is devoured with down sampled images to be able to capture the location of the lesion within the brain. Hence, the network was able to integrate both regional and contextual information. The outputs of the parallel convolutional layers are fed to a fully connected conditional Markov Random Field model to segment the tumor areas and decline the effect of noise in the input images. The proposed architecture by [14] outperformed the benchmark results on the dataset obtained from the Brain Tumor Segmentation Challenge (BRATS).

A combination between supervised and unsupervised deep representation learning approach was tackled by Baur et al. [15] for segmenting white matter lesion in the brain. The framework works on modeling abnormality with an unsupervised abnormality detection deep learning algorithm. Additionally, the detected abnormality targets are used as labels for training a supervised segmentation model. The model is first trained on reconstructing images of a healthy brain. Then it is used to detect any abnormalities in unlabeled brain data. Finally, a U-net is trained for jointly segmenting labeled brain data with their predefined ground truths along with the predicted ground truths from the unlabeled data. The experiments were conducted on the public datasets such as MICCAI 2008 challenge dataset and a private dataset retrieved from the University of North Carolina [16].

2.4.2 Lesion classification. A lesion is an abnormal appearance in regions of a tissue or an organ due to disease or injury, such as ulcer, abscess, tumor, or a wound. Lesion detection is the process of detecting key objects in classifying and labelling abnormalities in tissue's or organ's image. Some examples of different lesion detection in medicine, such as skin lesion detection, lung nodule classification, and Angiodysplasia detection can be found in [17-25].

Kawahara and Harmana in [17] classified 10- skin lesion classes with an accuracy of 81.8%. The team had 1300 captured images labelled to 10 different classes. Their network utilized transfer learning by using a pretrained Alex-Net that was trained on natural images, and then it was applied on the skin images. The network, also, used a multi-stream approach in which different resolutions were concatenated. More specifically, they extracted features from 227×227 and 339×399 -pixel images and then concatenated both feature vectors to obtain the accuracy result stated earlier.

Similar to Kawahara, Shen et al. [19] used a multiscale approach for classifying lung nodule as benign or malignant. In their approach, they had 3 CNNs working

simultaneously. Each CNN were input 3D-CT scanned images of size 32, 64 and 96 pixels of length, width, and height, respectively. Moreover, feature vectors were extracted by the 3 CNNs and then concatenated to one vector which was then processed using random forest technique. The multiscale tackle was done due to the varying size of the nodule, thus making the model flexible in detecting the different sizes and to accurately classify the nodule. Additionally, the simulation was done on LIDC-IDRI datasets which contains 1010 CT scanned lung images. The dataset was modified by having 5 modules rated from 1 to 5, where 1 being least malignant and 5 being the most malignant, to fit the binary classification; images that had a rating of 3 were deleted and images having a rating less than 3 were classified as benign and the rest were classified as malignant. The model achieved an average accuracy of 86.84%.

Another type of lesion classification is breast cancer mammographic images classification which was tackled by Ting et al. [26]. They proposed a modified convolutional neural network for classifying medical images into malignant, benign, and healthy patients. The CNN consists of an input layer, followed by 28 convolutional layers, and an output dense layer for classification. The data assessed in this study consisted of 221 patient subjects that were collected from the Mammographic Image Analysis society. Furthermore, data Augmentation was applied as an image preprocessing step to overcome the overfitting problems of the network. The framework succeeded to achieve an average accuracy of 90.5%.

In brain lesion classification, Dou et al. [22] focuses on detecting cerebral microbleeds (CMB) in the brain from magnetic resonance images. Utilizing 3D convolutional neural networks, the author proposes a multistage framework to firstly screen the candidates with highest probability of CMB and secondly distinguish between CMBs and their mocks using a discrimination model. The author focuses on building an architecture that could be able to achieve a high sensitivity and decrease the prediction time per subject. The architecture was applied on 320 volumetric MR scans and managed to achieve 98.29% sensitivity and 1.0725 minutes pers subject. After focusing on different problems in medical applications now the focus will be on GI tract diseases and their detection techniques.

2.4.3 GI tract diseases and detection. GI tract consists of a group of organs that are part of the digestive system. They are responsible for: digesting and absorbing minerals from foods and liquids, and excreting wastes as feces. According to the

national institute of diabetes and digestive and kidney diseases, 60 to 70 million Americans are affected by gastrointestinal diseases [27]. This subsection will discuss the endoscopy procedure and then will explore some well-known diseases and how are they detected.

An Endoscope is a tool that is used in a procedure called an endoscopy. Endoscope is mainly an illuminated optical tubular which is used to examine deep body parts by capturing images of internal organs. The tube is inserted from the mouth and the images are recorded throughout its trip inside the body. A doctor is then able to diagnose the patient without referring to a surgery [28]. In this thesis, deep learning algorithms are used to classify, and label images extracted from an endoscopy dataset.

Colorectal Polyp is a clump of cells that forms on the lining of the colon, which has a chance to develop to a colon cancer [29]. A CNN architecture was made by Komeda et al. in [23] to classify if the colorectal Polyp is adenomatous or non-adenomatous. Adenomatous means that the polyp is highly prone to turn cancerous [30]. Therefore, it is important to classify if it is adenomatous or not to avoid unnecessary resection operations. The CNN architecture is made of two layers, including convolutional and pooling, followed by a flattening layer, then a SoftMax function to present the output as a probability. The measure of success was 10-fold cross validation which resulted in an accuracy of 75.1%.

Hookworm is a parasite which grows inside the intestines of humans [31]. The worm threatens human health as it could cause maternal and child morbidity. Two deep convolutional networks based on CNN and Inception models has been adapted by [24] to detect and classify the hookworm from, 440k wireless capsule, endoscopy images. The first network is known to be edge detection network and the second one, inception based, is used mainly for classification. Inception module is used mainly to tackle the changing dimensions, sizes, and shapes of the worm. Furthermore, this paper utilizes two edge pooling sides which transfers features from edge extraction model to the low level and the high-level parts of the classification network. The concatenation of the feature and tubular maps regions is done to enhance the performance of classification. Accuracy, sensitivity, specificity, and ROC were used to evaluate the performance of the network.

Angiodysplasia is an asymptomatic disease that could lead to anemia and or gastrointestinal bleeding [32]. A comparison between four different networks of

detecting the disease was conducted by A. A. Shvets in [25]. The four adopted networks are U-net, Teraus-Net 11, and Teraus-Net 16 (like U-net but uses pretrained VGG (11 or 16) as encoder) and AlbuNet-34 (uses pretrained ResNet34 as encoder). The four networks were applied on a dataset in the form of 1200, 576×576 -pixel images split equally between training and validation. Additionally, the images had the salient points as white pixels in binary masks for disease localization. IOU, Dice co-efficient, and inference time were used, as evaluation metrics, to compare between the four networks. Moreover, AlbuNet-34 outperformed the rest of the networks achieving an IOU of 75.35, Dice of 84.98 and inference time of 21 MS.

Another approach for lesion detection and classification of several diseases was proposed by Zhu and his team in [33]. They had 180 endoscopy images of bleeding, ulcer, oncoids, polyps, umbilication and anabrosis, that were gathered by the department of gastroenterology in Anhui Medical University. The RGB images were resized from 390×390 pixels to 32×32 pixels to be fed to the network. Furthermore, the network consisted of a LeNet-5 famous architecture for feature extraction and followed by SVM for predicting patch labels. The model used accuracy, true positive rate, and true negative rate to evaluate the performance of the network. However, no numeric results were provided.

Our research focuses on three main diseases in the GI tract, which are Polyps (discussed earlier), Esophagitis, and Ulcerative Colitis. Esophagitis is the inflammation in the esophagus (the tube between the mouth and stomach) that might harm the tissues, while the Ulcerative Colitis is the inflammation in the lining of large intestine and rectum and can cause ulcers in the digestive tract. A modified Inception V3 model made by Andrea et al. [34] classifies these three diseases along with other non-disease classes that are found in the Kvasir dataset (used in current thesis). The paper utilizes transfer learning and data-augmentation to improve the performance of the classification. Additionally, they transfer weights learned from ImageNet dataset and augment the images in the dataset to increase the amount of data that the network will train on. The evaluation measurements of the paper were accuracy and F1 measure. Results of the training set has achieved approximately 91.5 percentile for both measurements.

Another approach for the same classification problem (Kvasir dataset classification) was tackled by Chaturika et al. [35]. Their framework composes of three different popular CNN models: VGG-16, ReNet-18 and DenseNet-201. The three

architectures are pretrained on ImageNet and are connected in a parallel fashion. Moreover, the features extracted from each architecture is passed through a corresponding global average pooling layer; the output from the three global pooling layers are appended and fed to a truncated singular value decomposition layer that acts as a noise filter. The output of the single decomposition layer is fed to a small neural network composed of a hidden layer and an output layer of 8 classes. The proposed framework managed to achieve an accuracy of 97.28% and F1 score of 0.9721,

The same problem was further tackled by Syed et al. [36] utilizing the concept of feature engineering. They developed their work based on features that were supported with the dataset such as: JCD, Tamura, Color Layout Edge Histogram, Auto Color Correlogram and PHOG. In addition to the provided features the authors compute the texture of the image. Moreover, they feed the features to logistic regression models and combine the votes by all models to contribute in the final decision of the class. The proposed model achieves a 94% accuracy and an F1 score of 0.76.

Another framework that also utilizes the baseline features is adopted by Agrawal et al. [37]. Their framework incorporates the features extracted from the fully connected layer of VGG-Net, the output features of the Inception-V3 and the baseline features. The set of features are combined and fed to a support vector machine (SVM) algorithm, linear kernel, for classification. Applying the combined set of features, they managed to achieve 96.1% accuracy and 0.847 F1 score. Our research interest stems out of [34]'s problem statement; however, as will be discussed soon in the methodology section, this paper discusses a different method of classification. The main aim of this work will be to jointly predict saliency map and classify the medical input images to their corresponding labels.

Chapter 3. Methodology

Our work is divided into two phases, which is saliency image detection followed by medical image classification. A U-net architecture will be selected from different variant to predict salient images from natural images. The transferred weights learned from training on natural images will be used to predict attention maps for medical images. The attention maps are used as ground truth label for a merged network that will jointly perform salient map prediction and class label classification for the medical dataset. This chapter discusses the problem of saliency in computer vision, then introduces the datasets used to train the proposed networks. Moreover, the architectures that are tested for saliency prediction will be discussed. Then the metrics used to evaluate the performance are explained. For the second phase, the dataset that will be used to train the proposed network for classification will be discussed. The chapter proceeds into the proposed adjustments to the saliency network, that are added, to output the labels for the images.

3.1 Saliency Detection

It is a process that trains the network to detect and highlight the most important object in an image or a video from a human visual system perspective [38]. Methods of saliency detections can be branched into two categories: bottom-up and top-down models. Bottom-up, named after its mechanism, focuses on making deductions based on low-level vision features inspired from human visual system such as, compactness prior [39], background prior [40] and contrast prior [41]. On the other hand, top-down means that the detection is based on previously known information in the image and is task driven. Therefore, it focuses on utilizing supervised learning using labeled images that is planned to be used throughout this research trained on two saliency datasets.

3.1.1 Saliency datasets. The two datasets that are used for training and validating the architectures are DUTS [42] and DUTSOMRON [43]. The DUTS dataset is known for its large-scale data, in which 10,553 training images and 5019 validation are included. The images are collected from, the well-known, ImageNet DET [44] and SUN [45] datasets. Additionally, the images are accompanied with their pixel-level ground truth labels that are manually annotated by 50 subjects. An example of the image and its pixel-level ground truth label is shown Figure 2.



Figure 3: Salient image example, right is ground truth [46]

The DUTOMRON dataset is mainly used in evaluating saliency detection networks for its characteristic of being one of the largest evaluation datasets with 5168 images along with their pixel level ground truths. The images are known for their relatively challenging complexity in saliency prediction compared to other datasets such as MIT [46] and NUSEF [47]. Additionally, the dataset can also help evaluate the network ability in predicting multiple salient objects in one image.

3.1.2 Saliency architectures. The well-known, U-net architecture has been utilized in many image segmentation applications such as nodule segmentation in low dose CT scans of chest [48], nuclei segmentation in microscopy images [49], liver segmentation in abdominal CT [50] and many more. This section will discuss the original U-net architecture along with its different variants that were built for this study.

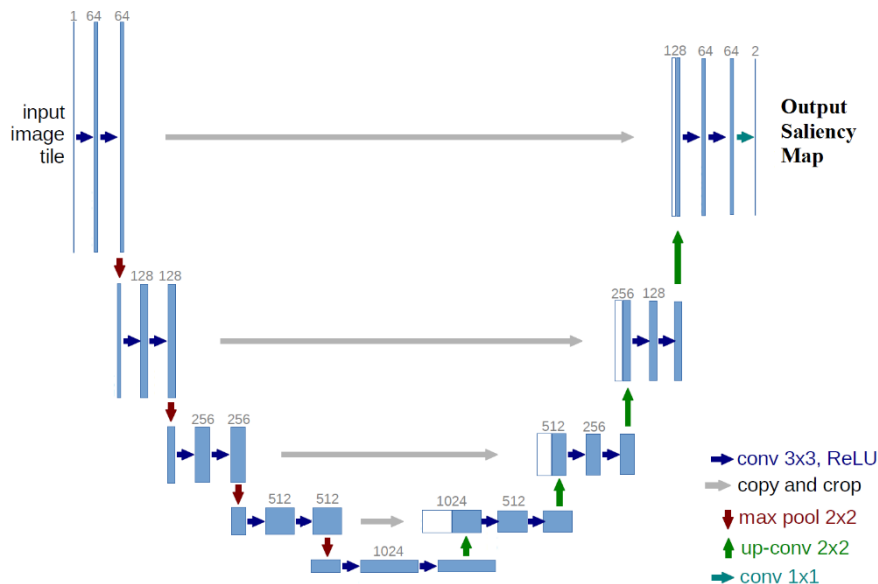


Figure 4: U-net architecture [23]

The original U-net architecture consists of contractive path and an expansive path, left half, and right half in the Figure 4, respectively. The contractive path consists

of layers of two 3×3 unpadded convolutions. These layers are followed by a rectified linear unit and a 2×2 max pooling operation with a stride of 2 for down sampling by half, which also means doubling the feature channels number.

On the other hand, the expansive part starts from the first 2×2 up-sampling and a 2×2 convolution layer, that would half the feature channels and up-sample by a factor of 2. The convolution layer that precedes the up sampling is also concatenated with corresponding feature map from the contractive path, followed by two 3×3 convolution, in which each is followed by a ReLU. The skip connections are presumed to improve the results of the network as it would transfer feature maps (information) from the contractive path to the expansive path before being lost in down sampling.

We additionally propose two extra U-net architectures, U-net dense and U-net skip, which has some slight additions relative to the original architecture.

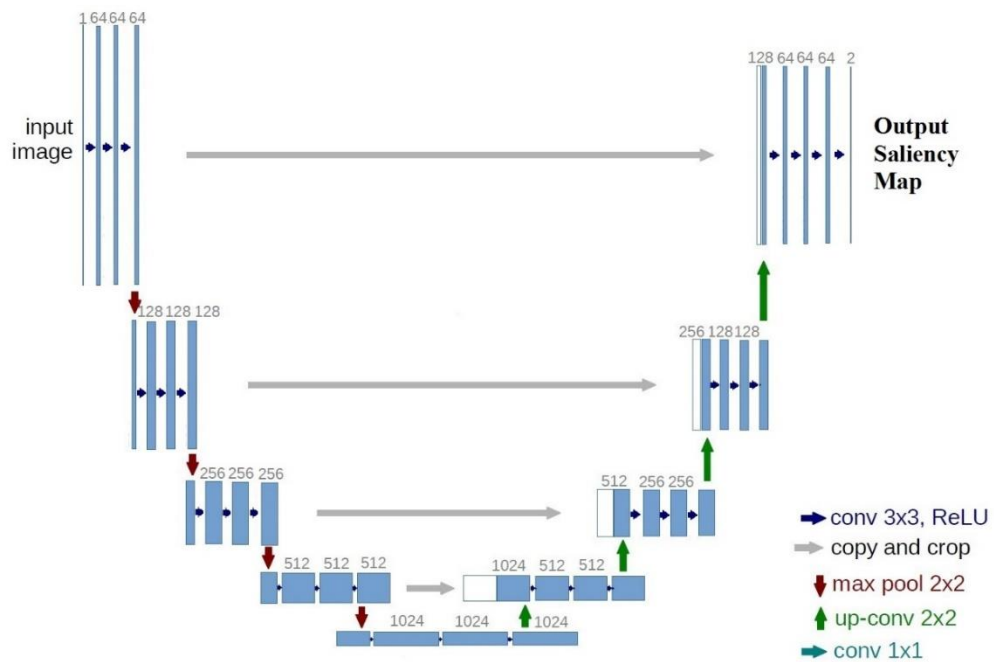


Figure 5: U-net Dense architecture

In contrast to the original architecture, the U-net dense has an extra 3×3 convolutional layer in all layers as shown in the Figure 5. Dense U-net has also been constructed as to study and observe the relative performance of denser and deeper networks in saliency detection applications.

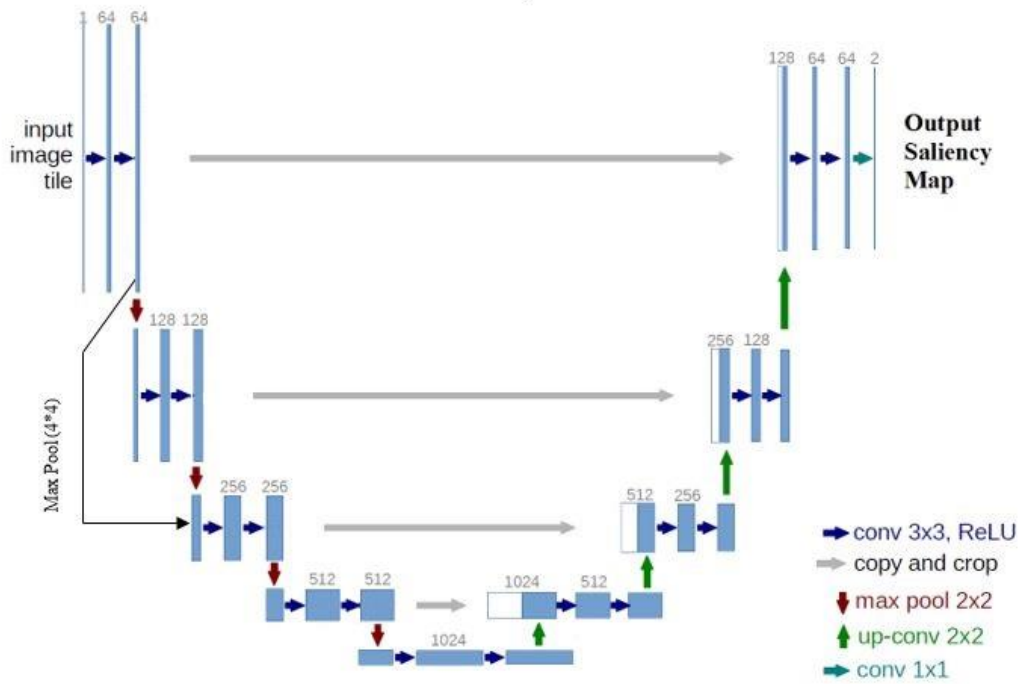


Figure 6: U-net Skip architecture

U-net Skip defining feature would be the supplemental skip connection in the contractive part that down sample by a factor of 4 using a max pooling 4×4 and stride 2. The result is concatenated to a layer in the contractive region by skipping some connections in the same region as shown in Figure 6. This is to check the effect of transferring information from earlier layers to deeper layers in the down sampling side on the overall performance.

The normal network and the U-net skip outperformed the U-net dense architecture and hence they were considered for more tuning. Since both had very close results and the skip connection did not contribute much in the results the rest of the experimentations was applied on the normal U-net. The normal U-net was further modified to improve its saliency prediction results by replacing the max pooling layers with stride convolutional layers. The results did show a bit of improvement, hence the modified U-net architecture with no max pooling was used further.

3.1.3 Input preprocessing. After choosing the architecture from the above experimentations, the focus is shifted to experimenting if inputting a different color space of the same image would improve the result of predicting the saliency masks. The different color spaces that are experimented are the RGB and Hue, Saturation, Value

(HSV) color spaces. The difference between the two was retrieved from [51] and is demonstrated below.

RGB images are the ones which constitutes of three main channels: red, green, and blue. Each color component represents 8 bits of a 24-bit image, which results that each channel has a range of values from 0 to 255. Each color that the eye can see is basically a weighted combination of these three primary colors which is represented numerically.

HSV on the other hand, represents the image in a different perspective. It tries to imitate how the humans view colors: where hue represents the color, saturation represents the shade, and the value represents the brightness. The ranges of these components are different from the RGB color space, as hue ranges from 0 to 360 degrees, while saturation and value ranges from 0 to 255 [51]. The Hue, saturation and value of an image can be retrieved from the RGB channel components using the following calculations:

$$R' = R/255, G' = G/255, B' = B/255 \quad (1)$$

$$Cmax = \max(R', G', B'), Cmin = \min(R', G', B') \quad (2)$$

$$\gamma = Cmax - Cmin \quad (3)$$

$$H = \begin{cases} 60^\circ * \left(\frac{G' - B'}{\gamma} * \text{mod } 6 \right), & Cmax = R' \\ 60^\circ * \left(\frac{B' - R'}{\gamma} + 2 \right), & Cmax = G' \\ 60^\circ * \left(\frac{R' - G'}{\gamma} + 4 \right), & Cmax = B' \\ 0^\circ, & \gamma = 0 \end{cases} \quad (4)$$

$$S = \begin{cases} 0 & Cmax = 0 \\ \frac{\gamma}{Cmax} & Cmax \neq 0 \end{cases} \quad (5)$$

$$V = Cmax \quad (6)$$



R', G', B' are the normalized values of the original R, G, B points. C_{max} and C_{min} contains the maximum and the minimum value of the normalized points, respectively. γ is the difference between the maximum value and the minimum value amongst the normalized R', G', B' values. The H, S, V represents the hue saturation and value points being calculated from the R, G, B values. The input combinations that were assessed are RGB, Hue and Saturation only (HS), and Value alone. The performances from the combination of architecture and input to the network were evaluated based on the perceptual sense of the mask and defined evaluation metrics that will be discussed shortly.







3.2 GI Classification

We now proceed to the discussion of the dataset that is used to train the network for classification, followed by the demonstration of the suggested architecture. It is notable to mention that the same evaluation metrics used in the saliency detection stage will be used to evaluate the network performance in classification. Finally, the section is concluded with a summary of the whole process.

3.2.1 GI classification dataset. The training dataset, Kvasir dataset [52], consists of 8000 images that are classified to 8 different classes. The classes show anatomical landmarks, pathological findings, and different dyed margins in GI tract. The images and their labels are collected and presented by GI endoscopists annotations. An image of each class along with a brief description is shown in Table 1.

Table 1: Target GI Tract endoscopy images for classification

<i>LABEL</i>	Sample	Description
<i>Anatomical Landmarks:</i>		<i>Features present consistently in a tissue that helps in indicating its structure or position</i>
Z-line		Dark thin bands across a striated muscle fiber
Pylorus		Muscular opening from the stomach into the intestine

Cecum		A pouch connected to the junction of the small and large intestines.
<i>Pathological Finding:</i>		<i>Abnormal alteration in tissues caused by a disease</i>
Esophagitis		Inflamation in the esophagus
Polyp		Abnormal tissue growth that sometimes looks like a small pump, and can be flat in cases that are hard to detect
Ulcerative Colitis		Inflamotory disease of the colon/bowel
<i>Others:</i>		
Dyed-lifted-polyps		The polyp is dyed by a saline solution to give a clearer vision for the margins, making it easier to detect
Dyed-resection-margins		Surgical margins of tissues that will be removed for examination (biopsy) are dyed

3.2.2 Basic classification network. The architecture that is considered as the baseline for classification is the Dense-net network [53]. The Dense-net was chosen due its interesting advantages such as: strong feature propagation and re-usage which reduces the number of trainable parameters. In addition, this architecture helps in reducing the possibility of the vanishing gradient problem occurrence.

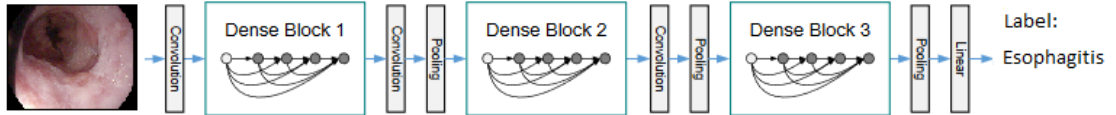


Figure 7: Dense-net architecture [53]

The strong feature propagation and re-usage is established by the connectivity pattern of the network. Each layer of a dense block is connected directly to all the subsequent layers in the same dense block, as observed in Figure 7. To elaborate more, let us assume that the dense block consists of L layers and the output of each layer ℓ is x_ℓ . Additionally, the non-linear transformation, H , applied to a layer output (which is the input to the next layer ℓ) is denoted by:

$$X_\ell = H_\ell(x_\ell - 1) \quad (7)$$

Normally without skip connection the equation of the information movement from one layer from the previous layer is presented by equation (7). However, in the Dense-net architecture, the ℓ^{th} layer receives the information concatenated from all the previous layers which can be denoted as:

$$X_\ell = H_\ell ([x_0, x_1, x_2, \dots, x_{\ell-1}]) \quad (8)$$

Each dense block that adapts the aforementioned skip connection feature shares the same size of feature maps throughout the whole block in order to be able to conduct the concatenation task. Moreover, the feature maps outputs are down sampled using the transition layers found in the middle which is made up mainly of a 1×1 convolutional layer followed by a 2×2 average pooling layer. Except the last dense block which is followed by a global average pooling layer and a SoftMax classifier to predict the label. The interconnection of layers within the dense block allows each layer to access the gradients from the input signal and the loss function, leading to an improved implicit supervision and reduces the chances of the loss of gradient within the training process.

The classification results score of this architecture was chosen as a baseline score for comparison with the proposed architecture. Our composed architecture merges the U-net architecture that is pretrained on saliency prediction of the DUTS dataset and the Dense-net architecture. The concept of reusing the weights of a previously trained architecture is known as transfer learning.

3.2.3 Transfer learning. The use of pre-trained networks, on different image datasets, to serve the purpose of a different task that the original one was trained to fulfill is a type of transfer learning. This method helps in improving the learning of the new task, as knowledge is transferred from the pretrained task [54]. Different transfer learning strategies that were discussed in papers [17] and [32] in the literature are identified: 1) fine-tuning a pretrained network on its new target purpose and 2) using the pretrained network as a feature extractor technique. Moreover, in this paper transfer learning will be utilized after training the network on detecting saliency map. The architecture utilized for saliency detection with its learned weights will be transferred and modified. Additionally, the network is retrained and restructured for class prediction along with saliency map detection, as seen in the proposed architecture in Figure 8.

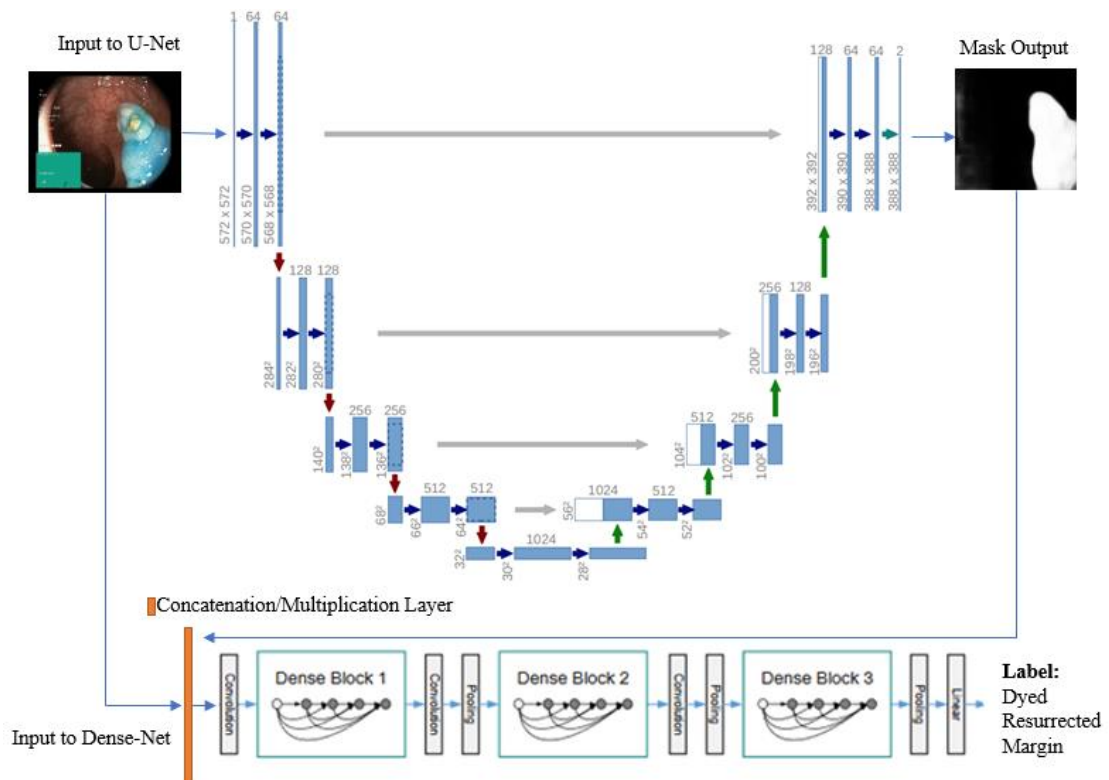


Figure 8: Proposed final architecture for joint saliency and prediction

3.2.4 Proposed architecture for medical image classification. This section will discuss the proposed architecture that will be utilized for jointly predicting the attention maps and the classification labels of the medical images.

The network shown in Figure 7 is a result merging the U-net architecture and Dense-net architectures. As seen by the figure, the network will have two outputs which is the output mask after the U-net architecture and the output label at the end of the network. We further study the effect of concatenating, or multiplying, the saliency mask of the medical image that is a product of the U-net architecture with the original input image to check if these actions would actually improve the performance of classification or will it deteriorate it. Concatenating the input with the mask, is like adding an extra layer of information to the network to base its decision upon. While multiplying the input with the mask is like focusing the framework to predict the label based on a selected product of the original input that is highlighted by the mask. The summary of the whole procedure will be discussed with the hypothesis that is applied in current research.

3.3 Cost Functions and Evaluation Metric

Cost functions are used to quantify the error between the predicted output and the actual output. These errors are used to guide the network to find the set of weights for their neurons in the neural network layers. Since the model objective is to perform two different tasks, each one was performed using a certain cost function. The saliency prediction was guided using the mean square error (MSE) per pixel, and the label classification was guided by the categorical cross entropy (CE) function. A combined weighted loss function, consisting of MSE and CE was then used to update the weights of the overall architecture

MSE is the measure of squared mean difference between the original label pixel, Y_i , and the predicted label pixel, \hat{Y}_i , of all pixels, N , and is defined by the following equation:

$$MSE = \frac{1}{N} * \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \quad (9)$$

CE function is calculated using the probability of each class that is fed to the function. The probability of each class is calculated using SoftMax function shown by equation 10. The ‘ t ’ represents the output label and C is the total number of classes, and the function works by dividing the exponential of a label, e^{t_i} , by the sum of all the exponentials.

$$f(t)_i = \frac{e^{t_i}}{\sum_j^c e^{t_j}} \quad (10)$$

The results from the SoftMax function is fed to the categorical cross entropy to calculate the network loss over the total number of classes, C :

$$CE = - \sum_i^c t_i \log (f(t)_i) \quad (11)$$

Hence the overall weighted loss function used can be illustrated in the following equation, where w_1 and w_2 resembles the weight assigned to each cost function.

$$\text{Overall Loss Function} = w_1 * MSE + w_2 * CE \quad (12)$$

The variation of these weights and their effect on the overall performance will be discussed in the results section. Additionally, the performance of these networks is then evaluated using three main evaluation metrics that are the Mean Absolute Error, F-measure and Accuracy.

Mean Absolute Error is a measure of the sum of the absolute difference between two variables: it computes the difference across each image pair (predicted and expected), sum up the values and divide by the number of pixels. Its equation is given below:

$$MAE = \frac{\sum_n |Predicted - Expected|}{n} \quad (13)$$

F-measure is a method of measuring the robustness of classification. It is also known as the harmonic mean of precision and recall. Precision is the fraction of relevant instances (true positives) among the retrieved instances (true positives+ false positives). Additionally, recall is known as the fraction of fraction of relevant instance (true positive) over the relevant instances (true positive + false negative). The β in F-measure is a weight distributor, to choose, which one from the precision or recall would have more effect on the overall score. For instance, if β is 2, the weight of recall will be higher than precision on the overall score, hence more emphasis will be focused on the false negative. On the other hand, if β is less than 1 the weight of precision will be higher than recall, therefore less influence of false negative. However, β was set to one, to have a balanced weight distribution to both precision and recall, in the results attained

(discussed later). The equations of precision, recall and the general F-measure are shown respectively below:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (14)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (15)$$

$$F_{\beta} \text{ measure} = (1 + \beta^2) * \frac{Precision * Recall}{\beta^2 * Precision + Recall} \quad (16)$$

Accuracy is a measure of the number of correct predictions to the total number of input samples. This metric is used mainly the dataset is balanced; i.e. equal number of images distributed among all labels. It can be presented using the following equation:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (17)$$

Stage 1, choosing the best architecture and input variation, was evaluated using the MAE and F1 measure guided by [55]. Stage 2, jointly training the network on saliency prediction and label classification, was evaluated using the MAE to show the change in the mask prediction behavior for the mask output. As for the label output it was evaluated using Accuracy and F1 score. A summary of the whole procedure will be illustrated to connect the steps and ideas of the project.

3.4 The Procedure Summary

To begin with, saliency prediction is implemented using different U-net architecture variations and different input types. The U-net architectures were trained on DUTS and DUTSOMRON. We then check the performance of each architecture relative to the paper discussed by Liu et al. [55]. The best performing network in terms of the MAE, F1 score on the DUTS and DUTSOMRON datasets was then selected. We then used the pre-trained network to generate the saliency masks labels for the medical image dataset KVASIR, to evaluate the ability of network in highlighting the salient regions in the medical image dataset subjectively. This was done using the perceptual of the human eye, by checking if the masked area corresponds to the key features of a certain label. This presents an interesting result that the network trained on natural

images also gives reasonable results in medical image dataset. Different input variations, RGB, HS and Value only, were used to predict corresponding masks. Moreover, the best predicted masks were then used as the ground truth labels for the second stage.

The second stage merges the best U-net architecture of the first stage, along with Dense-net architecture as shown in Figure 7. Additionally, the effect of multiplying or concatenating the input image with the masks is examined. The network is guided using a weighted combined loss function, where the effect of changing the weights on the overall score and the predicted mask was also studied. We can ensure that saliency layer output does not deviate much by adding a term corresponding to the MSE between the predicted saliency mask and the ground truth in the loss function at the last layer. Additionally, this work's classification results are compared with the results obtained from the Dense-net to see if jointly predicting saliency and classes helps in enhancing the classification. The results from the different iterations were tabulated and the best ones of each iteration were collected and applied to a five-fold cross validation to ensure that the results are not biased, and the network is not over fitting.

Chapter 4. Results and Discussion

The following sections encompasses the results gathered through the different stages of the thesis. Moreover, section 4.1 will discuss the saliency prediction and section 4.2 will discuss the results obtained from training the combined network.

4.1 Saliency Prediction

This section will discuss the results obtained in each stage in finding the best U-net architecture, then will discuss the different input combinations and their corresponding medical mask output.

4.1.1 Selection of U-net architecture. The following section encompasses results gathered from the training of the saliency detection of objects. Moreover, the outputs of the three architectures, discussed in the methodology section, are collected after running through different learning rates, training datasets and validation datasets. Furthermore, the results are compared with results retrieved from a paper by Liu et al. [55]. Our aim is to get results that are close to the benchmark results and could predict masks that are reasonable and acceptable.

The three proposed architectures (U-net, U-net skip, and U-net dense) for saliency detection are trained on DUTS training dataset and tested on two different datasets: DUTS test and DUTOMRON. Moreover, the three architectures are trained over 60 epochs and with three different learning rates. These three learning rates are chosen after running the network on 7 logarithmic learning rates (i.e. $1e-1$, $1e-2\dots$) from $1e-1$ to $1e-7$. The results from these 7 learning rates showed acceptable results between $1e-4$ and $1e-6$, hence they are demonstrated in this paper. Additionally, the evaluation metrics that were used are mean absolute error and F_1 measure.

The yellow and green highlighting are used as an indication for the outperforming learning rates and the best performing architecture score for the convenience of the reader. Yellow highlight is used to highlight the row of best learning rate. Green highlight with bold font is used to highlight the architecture's name and value of best metric score.

Table 2: Training: DUTS Training | validation: DUTS Test | metric: MAE & F1

Training: DUTS Train Validation: DUTS Test Metric: MAE & F1						
Learning rate	U-net		U-net skip		U-net dense	
	MAE	F ₁	MAE	F1	MAE	F1
1e-4	0.1615	0.5943	0.1617	0.5865	0.4759	0.5527
1e-5	0.1172	0.6568	0.1282	0.6550	0.1334	0.6244
1e-6	0.4963	0.0000463	0.2534	0.5207	0.3103	0.4541

Table 2 shows results that are obtained after the network has been trained on DUTS train and validated on DUTS test. In that event, U-net architecture outperformed in both the MAE and the F1 measure scores, where the scores achieved were 0.1172 and 0.6568, respectively.

Table 3: Training: DUTS Train | validation: DUTSOMRON | metric: MAE & F1













Training: DUTS Train Validation: DUTSOMRON Metric: MAE & F1						
Learning rate	U-net		U-net skip		U-net dense	
	MAE	F ₁	MAE	F1	MAE	F1
1e-4	0.1268	0.6732	0.1193	0.6864	0.1451	0.6716
1e-5	0.1779	0.5852	0.1685	0.6197	0.1348	0.6118
1e-6	0.2728	0.5005	0.2557	0.5220	0.3029	0.4816

Table 3 shows results that are obtained after the network has been trained on DUTS train and validated on DUTSOMRON. The results show that the U-net skip architecture outperformed in both the MAE and the F1 measure scores, where the scores achieved were 0.1193 and 0.6864, respectively.

Additionally, to be able to visualize the network output 8 images were extracted from the DUTS training test dataset and were excluded from training and validation of any of the networks. This was done to ensure that the networks would have a fair comparison between them, and that none of the networks is biased toward a single network. Below are the visual results of the best performing and the poorest performing networks from all iterations. The best performing network was extracted from Table 2 and was found to be the U-net architecture with a learning rate of 1e-4. On the other hand, the worst performing network was found to be the U-net architecture with a learning rate of 1e-6. Noticeably, the networks that had the best scores shows images that are very close to the expected output images. On the other hand, the network with

the poorest results gave unacceptable predictions that resembles distorted or noisy images with no visual knowledge that can be deduced from them. The information collected from the tables and the images were able to provide some important deductions that helped in the thesis’s later work.

Table 4: Visual output of selected architectures

Expected								
Best Prediction								
Poorest Prediction								

From the above results it is noticeable that the proposed U-net skip and U-net have very close results in both metrics. Additionally, it is safe to deduce that the optimum learning rate lies between $1e-4$ and $1e-5$ as the best performance network varied between these two learning rates. Results obtained from these simulations are compared with the results extracted from the U-net architecture built by [54]. The author’s architecture has the same architecture the original U-net utilized in this research but with different hyper parameters. The numerical comparison is demonstrated in 5 Table:

Table 5: Comparison between Liu et al. [55] and current results

	DUTs test		DUTSOMRON	
	MAE	F measure	MAE	F measure
Liu et al. [55]	0.060	0.819	0.073	0.761
Thesis result	0.1172	0.6568	0.1193	0.6864

As seen from the results of the first trial, they were a bit far off from the results obtained by the author as shown in Table 5. The deduction from the above results, in different tables is the adding skip connections might improve the results but not significantly hence the other U-net architectures were excluded from current work.

The original U-net architecture was further modified by changing the max pooling with strided convolution layers to allow more flow of information of the network and additional hyperparameter tuning were applied such as adaptive learning rate and increasing dropout layers. All of these has contributed in improving the performance of the U-net architecture which is shown in Table 6:

Table 6: Comparison between U-net with and without max pooling and with strided convolution

	DUTs test		DUTSOMRON	
	MAE	F measure	MAE	F measure
Liu et al. [55]	0.060	0.819	0.073	0.761
Previous U-net	0.1172	0.6568	0.1193	0.6864
U-Net (with strided convolution)	0.098	0.7258	0.086	0.732

Table 6 shows that the results have improved, yet it did not reach the paper results. However, in the paper they were using image sizes of 256×256 while in this research 128×128 images are utilized instead due to hardware limitation. The decrease in the size of the image means that less information is supplied to the network which might lead to the degradation in the results. Hence, the modified U-net was set to be utilized for further practical experimentation without extra hyper parameter modifications.

4.1.2 Saliency mask from input combinations. After reconstructing the U-net architecture with stride convolution, replacing the max pooling layers, different color space variations of the input were experimented to check if they affect the classification evaluation metrics and the predicted medical masks. The images were first resized to 128×128 and different input variations were experimented, such as the RGB, HS and Value inputs on the DUTSOMRON dataset only, and their results are illustrated in table 7. The RGB input had shown that it has a better evaluation scores when compared to HS and Value. Additionally, the HS mask showed an average behavior between both value and RGB. However, as the scores' differences are not major, the medical masks were predicted using these three inputs.

Table 7: Comparison between different color space inputs on DUTSOMRON

Input Type	MAE	F measure
U-Net RGB	0.086	0.732
U-Net HS	0.0921	0.7198
U-Net Value	0.098	0.7111

A comparison between the results of the RGB, HS and value masks was conducted in terms of which one had the best perceptual sense visually to be chosen as the ground truth for the medical images' saliency map label for stage 2. A sample image will be shown from each class label and will be compared with the original image if it does cover the important marks of the image. The images initially showed inaccurate masks as the black spaces in the images had an effect of finding the salient region using the network. A sample is shown in table 8.

Table 8: Mask predicted before and after zoom + crop


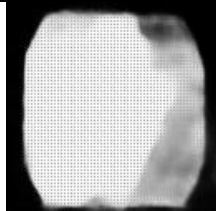





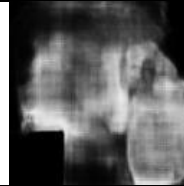



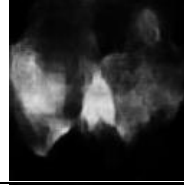


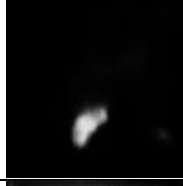
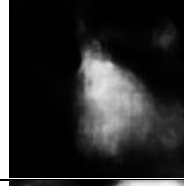


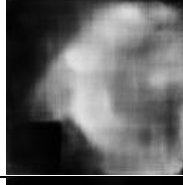
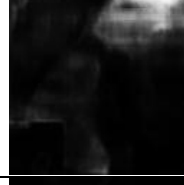

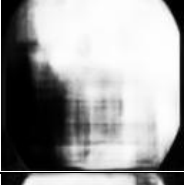
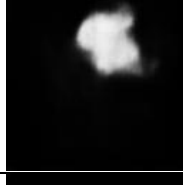








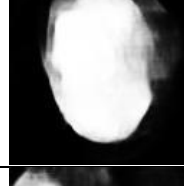




Label	Image	Mask
Original Image		
Zoomed and Cropped Image		

Table 8 shows the image of esophagitis and its predicted saliency mask before and after zooming in and cropping by a factor of 1.275. The masks, after cropping, have a more focused attention maps highlighting certain parts of the image instead of being distorted by the black spaces in the image.

Furthermore, the difference between the predicted attention maps of different color spaces were assessed. The original input medical image with its corresponding mask that is a result of the RGB, HS and Value inputs are illustrated in the Table 9.

Table 9: Input image, and mask outputs for RGB, HS and Value

Label Name	Original Input Image	RGB output	HS Output	Value
Dyed Lifted polyps				
Dyed Resection margins				
Esophagitis				
Normal Cecum				
Normal Pylorus				
Normal Z-Line				
Polyps				
Ulcerative Colitis				

In order to recap the explanation of each label please refer to Table 1. The dyed lifted polyp saliency mark is the pigmented cells in blue showed in the image. We can

see from the images above that masks predicted from RGB and HS inputs show a decent behavior in extracting it out of the image, while the value fails drastically. Additionally, the masks from RGB input gets attracted by the trap of the bottom left window and merely highlights it which should not be the case and should be ignored by the architecture.

Dyed Resection Margin on the other hand shows the pigmented surgical margins that should be removed for examination. The RGB input shows the inverse of the salient region by highlighting the clear region instead of highlighting the pigmented region. The HS input, on the other hand highlights the area of the darkest pigment present. While the Value highlights the dark hole along with the sides of the hole, which does not show anything that could be of importance to the examiner.

As for esophagitis, it is known to be the inflammation of the esophagitis and it is very hard to capture by any of the three networks, as the color pigment or the variations in the pixels is not significant. As for the RGB input it highlights a semi-circular shape as shown, while the HS highlights only a dot of the highly pigmented hole of esophagitis. Value, also, highlights the darker area along with an unrelated region.

Cecum is just a connecting junction between small and large intestine. As seen in the above image, it is having a rutted shape and would be challenging for the network to predict as well. We can observe from the Table 9 that HS input somehow manages to mimic the image, while the RGB and Value inputs fails to highlight anything that is useful.

Pylorus is just an opening from the stomach to the intestine and can be represented as a hole. The HS and Value succeeds in highlighting these holes, while the RGB fails to highlight what is important in this task.

Z-line is the dark thin bands across a muscular fiber which can be observed in the image above darks lines. The networks do not precisely highlight the lines; however, the HS and Value succeeds in highlighting the region in which the z-line is present while the RGB input shows a poor performance.

Polyps are abnormal tissue growth and sometimes look as a small pump in the image and this pump would sometimes have the same color as the surrounding healthy

tissues, hence it is hard to predict using saliency networks. As seen above the three-color inputs are having trouble in finding the polyp which is deep inside the cecum lookalike tissues. The value only might be highlighting the region in which a surgeon would search for the polyps. However, the three variations had shown their weakness in finding the polyp inside the image for other examples.

Ulcerative colitis (UC) is an inflammatory disease in the colon and is the shrinking of the size of the colon hole that is observed in the image above. The masks corresponding to the labels above show that output saliency masks due to HS inputs and Value inputs are a bit precise about their prediction as there is a bit of green pigmented element present in the element which could attract the attention of the viewer and hence supports the theory of attention maps.

Overall, the saliency masks due to HS inputs outperformed the RGB even though the evaluation score for the RGB input is higher than HS and value as shown in Table 7. The RGB and value masks showed an overall poor performance, which made it reasonable to use the masks predicted from the HS input to be used as the medical mask labels for stage 2.

4.2 Combined Saliency predictions with labels classification

In this section, the results obtained from concatenating and multiplying the input layer with the generated mask from the U-net architecture will be discussed. Different tables will be used to illustrate the numerical results. Then the behavior of the masks with the different variations will be demonstrated. As mentioned earlier, the images fed to the U-net architecture for saliency were of size 128×128 , however due to the increased complexity of the network and the computational limitation, for this section the images were resized to 64×64 . The weight assigned for each cost function was changed in a sense that the summation of both weights should equate to one, and the change in step size was 0.1.

$$\textit{Constraint: } w_1 + w_2 = 1 \quad w_1, w_2 = [0, 0.1, 0.2, \dots, 1] \quad (18)$$

Even though RGB inputs did not perform well initially in the previous phase while predicting the medical masks, trials are still set on RGB and HS images, while Value images were excluded. Therefore, the search narrowed to 4 combinations which are shown in Table 10.

Table 10: Different iteration executed

Network:	Input	Operation in middle layers with input
1	RGB	Concatenation
2	RGB	Multiplication
3	HS	Concatenation
4	HS	Multiplication

The dataset was randomly shuffled and then kept constant for all the different operations, to ensure fair comparison between all networks. The results of these networks were assessed upon two main criteria: first, their ability of outperforming the results obtained from Dense-net, second, if their masks' prediction ability improved, maintained same quality, or deteriorated. The best combination of weights for each network was extracted with limited testing and then applied on five-fold cross validation to ensure that the results are not biased and performs well on different selections of the data.

4.2.1 Concatenated RGB input. The evaluation results obtained from the network of RGB input and that is concatenated with masks on different set of weights are shown in Table 11.

Table 11: Results obtained from concatenated RGB model

U-NET	Dense-net	MAE	Acc	F1m
1	0	0.1392	0.1288	0
0.9	0.1	0.1446	0.875	0.8765
0.8	0.2	0.1567	0.8875	0.8859
0.7	0.3	0.1725	0.8975	0.899
0.6	0.4	0.1957	0.8938	0.8934
0.5	0.5	0.2038	0.8888	0.8841
0.4	0.6	0.2105	0.9012	0.8983
0.3	0.7	0.2596	0.8925	0.8961
0.2	0.8	0.2925	0.8852	0.8829
0.1	0.9	0.3163	0.8738	0.8757
0	1	0.4319	0.88	0.881
Untrained joined network		0.219	12.5	12.49
Dense-net results		-	0.8662	0.8685

The overall view of the results shows that concatenating an extra mask that is generated while training would improve the results of the network. The concatenated

mask acts as extra information that is supplied to the network in order to make its final decisions. Additionally, the MAE enumerates the change in mask predicted from the mask label present. However, the quantity of the number does not show if the masks improved or did it worsen, hence the masks were assessed visually for each label and was given scores accordingly, scores varied between -2 and 1 integers. Furthermore, (-2) represented an equally highlighted mask without any detectable discrepancy, (-1) represents that the masks prediction quality has worsen, (0) stayed the same, and (1) improved. A glimpse of each score label is shown in Table 12.

Table 12: Saliency masks labeling illustration

Label Name	Original Input Image	Ground Truth	worsen (-1)	Same quality (0)	Improved (1)
Dyed Lifted Polyps					
Dyed Resection Margins					
Esophagitis					
Normal Cecum					-
Normal Pylorus					-
Normal Z-Line					-
Polyps				-	

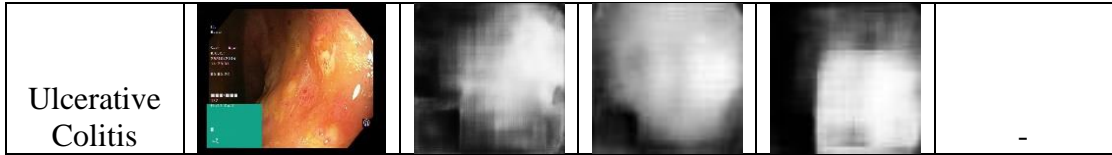


Table 12 extracts the variations in the ground truth from all the trials that has been implemented through this thesis. It is noticeable that some labels did not improve from the original ground truth, only four labels have shown improvement in some of the trials, such as Dyed Lifted Polyps (DLP), Dyed Resection Margin (DRM), Esophagitis (Esoph), and the polyps. The networks improved in the sense of capturing the key features of a certain label by increasing its highlighting intensity on these features while giving the rest of the regions less importance. As seen by DLP, esophagitis and polyps for example, it captures more susceptible regions of DLP, esophagitis and polyps featured cells. As for DRM, it gives more intensity for the dyed region which is the key feature of the label and gives less label to the non-dyed part as opposed to the original ground truth highlighting both. The rest of the labels showed same or lower quality mask than the ground truth. The deteriorated masks were labeled based on their lack or decrease of the important key features highlighted as seen throughout all the labels. In most of the cases it shows the disability of the network in figuring out what is important in the network. The process of labeling the output masks was performed to enumerate the effect of changing the weight on the overall visual output. An example of the enumeration process is shown in Table 13.

Table 13: Saliency mask enumeration results for RGB concatenated network

U-NET	Dense-Net	DLP	DRM	Esoph	Norm. cecum	Norm. pylorus	z-line	polyps	UC
1	0	-	-	-	-	-	-	-	-
0.9	0.1	0	0	0	0	0	0	1	0
0.8	0.2	-1	-1	0	-1	0	0	0	0
0.7	0.3	-1	-1	1	0	0	0	-1	0
0.6	0.4	-1	-1	0	0	0	0	-1	-1
0.5	0.5	-1	-1	1	0	-1	0	-1	-1
0.4	0.6	-1	-1	-1	0	0	0	-1	0
0.3	0.7	-1	-1	-1	0	-1	0	-1	-1
0.2	0.8	-1	-1	-1	0	-1	-1	-1	-1
0.1	0.9	-2	-2	-2	-2	-2	-2	-2	-2
0	1	-2	-2	-2	-2	-2	-2	-2	-2

The first iteration in terms of giving full attention to U-net was excluded from the calculations, as its classifications results were highly unacceptable. Moreover, the Table illustrates that increasing the weight of Dense-net and decreasing the weights of the U-net results in reducing the performance of the masks. This is observed by the increase in the negative label numbers, showing an inverse proportionality between classification and saliency prediction. As more weight percentage is shifted toward the classification loss function, the masks start to disappear as shown by the -2 labels. The summation of all the labels with respect to weights was made to show the total improvement in the masks and could be evaluated as the more positive the value the better it is. The summation is attached to Table 12 forming the final Table format that will be used to discuss the further results in this thesis. There is a tradeoff between the performance of the saliency mask prediction and the performance of the classification as shown in Table 14.

Table 14: Enumerated saliency and classification results for concatenated RGB framework

U-NET weight	Dense-net weight	MAE	Acc	F1m	Mask Score
1	0	0.1392	0.1288	0	-
0.9	0.1	0.1446	0.875	0.8765	1
0.8	0.2	0.1567	0.8875	0.8859	-3
0.7	0.3	0.1725	0.8975	0.899	-3
0.6	0.4	0.1957	0.8938	0.8934	-3
0.5	0.5	0.2038	0.8888	0.8841	-4
0.4	0.6	0.2105	0.9012	0.8983	-4
0.3	0.7	0.2596	0.8925	0.8961	-6
0.2	0.8	0.2925	0.8852	0.8829	-7
0.1	0.9	0.3163	0.8738	0.8757	-16
0	1	0.4319	0.88	0.881	-16
Untrained joined network		0.219	12.5	12.49	-
Dense-net results		-	0.8662	0.8685	-

When higher portion of the weights was applied for the U-net it had better mask prediction yet lower classification score. However, the scores are close to each other in term of classification accuracy with a little bit of fluctuation. This behavior shows that the extra layer of masks being concatenated has improved the performance of the

network over the Dense-net architecture alone. However, the behavior of the mask in terms of improvement or depreciation does not seem to affect the classification results. In other words, extra information provided would lead in improving classification despite its condition. The untrained joined network’s result was added to test the behavior of the just using the pretrained U-net and pretrained dense-net on the overall performance without synching between both networks. The classification accuracy and F1 score shows that the network without training is unable to perform the joint task hence it was applied for further training to do so. The highlighted weights were to act as a middle ground between classification and saliency prediction as it has high classification value, with close importance between loss functions’ weights. Further examination using five-fold cross validation are conducted on chosen weights to compare with the other results obtained from different settings.

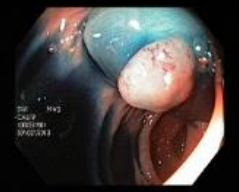




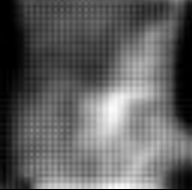
4.2.2 Multiplied RGB input. In this subsection the effect of multiplying the RGB inputs with predicting masks architecture is examined. This will study the behavior of focusing the inputs on what is classified as important by the saliency masks. Different weight variation on loss functions are applied and studied, like the previous section. Table 15 will be used as the source of information to discuss the effect.

Table 15: Saliency and classification results for multiplied RGB framework

U-NET weight	Dense-net weight	MAE	Acc	F1m	Mask Score
1	0	0.1443	0.1288	0	-
0.9	0.1	0.1582	0.8588	0.8555	-1
0.8	0.2	0.1865	0.8688	0.8671	-3
0.7	0.3	0.2177	0.87	0.8721	-3
0.6	0.4	0.2486	0.8862	0.8848	-4
0.5	0.5	0.1146	0.8925	0.8934	-16
0.4	0.6	0.1146	0.8762	0.875	-8
0.3	0.7	0.3124	0.88	0.8814	-8
0.2	0.8	0.3155	0.8888	0.8864	-16
0.1	0.9	0.3156	0.8862	0.887	-16
0	1	0.4796	0.8888	0.8905	-16
Untrained joined network		0.206	16.43	15.06	-
Dense-net Results		-	0.8662	0.8685	-

The multiplication effect clearly shows the mask effect on the classification results. As the mask turns clear and empty, the classification results increase as shown in Table 15. This is similar to multiplying the RGB input by 1 and allowing it to pass on all the information it carries. However, there may be minor variations that may explain the increase in performance. Additionally, as the Dense-net's cost function gets a higher portion of the weight the mask starts fading away and showing less information which reflects the behavior of concentrating the weights for classification. Hence, this proves that both classification and saliency prediction cost functions are contradicting cost functions and they do not support each other. The predicted masks from this framework also contains some white noise in the background which can be considered as a rebelling behavior of the classification cost function to allow more information to pass for classification. An example can be shown in Table 16.

Table 16: Examples of distortions caused by multiplications

Label	Image	Ground Truth	Predicted Mask
DLP			
DRM			

The predicted masks in Table 16 shows that the networks highlight the important information with higher intensity of white color and contains some distortion noise. This proves that the multiplication layer acts as a bottleneck layer of the information that could pass through, in which the framework tends to expand in order to achieve higher classification results. However, the extra guidance provided had shown that the network with extra layers and guided information outperforms the normal Dense-net benchmark. The regions which has a unified color with no salient information, denoted with -16 mask score, appears to act as a weighting factor which helps in improving the classification results. The highlighted results are then used for further investigation using five-fold cross validation.

4.2.3 Concatenated HS input. The following section shows the results of jointly predicting saliency labels and classification using HS input. Similar to RGB concatenation, the results in this section supports the hypothesis that adding an extra layer of information will be valuable for the classification purpose. The results of training the whole framework on saliency prediction and classification are shown in Table 17. The network succeeded in achieving results higher than the Dense net in most weights. It can be observed that, similar to the previous networks, as the weights are shifted toward classification, the performance of the masks worsens. Additionally, the masks output in this configuration happens to worsen faster in terms with reduction in weight. Once the classification weights are higher than saliency cost function, the saliency masks become very poor. However, the masks information seems to be contributing significantly while concatenation as it achieves very similar classification score even when masks perform relatively with -2 score and -16 score. However, the overall performance of the network in classification has improved as proven by the results illustrated.

Table 17: Saliency and classification results for concatenated HS framework

U-NET weight	Dense-net weight	MAE	Acc	F1m	Mask Score
1	0	0.1543	11.66	0	-
0.9	0.1	0.1537	87.62	87.71	0
0.8	0.2	0.1653	87.5	87.43	-3
0.7	0.3	0.1827	88.38	88.05	-3
0.6	0.4	0.2046	89.75	89.72	-2
0.5	0.5	0.2399	88.12	88.05	-3
0.4	0.6	0.3067	89.12	89.15	-16
0.3	0.7	0.307	87.62	87.91	-16
0.2	0.8	0.5662	88	87.95	-16
0.1	0.9	0.3012	89.38	89.37	-16
0	1	0.3732	88.25	88.3	-16
Untrained joined network		0.209	18.775	17.12	-
Dense-net Results		-	0.8662	0.8685	-

4.2.4 Multiplied HS input. This section shows the results obtained from using the same U-net weights used in the previous section for HS saliency prediction. The difference is that the merging layer between the networks is a multiplication layer. The results obtained are shown in Table 18.

Table 18: Saliency and classification results for multiplied HS framework

U-NET weight	Dense-net weight	MAE	Acc	F1m	Mask Score
1	0	0.1496	12.88	0	0
0.9	0.1	0.1728	0.8617	0.8625	-1
0.8	0.2	0.1792	0.8588	0.8594	-2
0.7	0.3	0.2097	0.8562	0.8573	-5
0.6	0.4	0.2159	0.8875	0.885	-6
0.5	0.5	0.2569	0.87	0.8716	-5
0.4	0.6	0.2735	0.855	0.8579	-7
0.3	0.7	0.2828	0.8788	0.876	-8
0.2	0.8	0.2962	0.87	0.8672	-7
0.1	0.9	0.7142	0.8938	0.894	-16
0	1	0.6955	0.8888	0.8892	-16
Untrained joined network		0.203	16.9	15.89	-
Dense-net Results		-	0.8662	0.8685	-

The results mimics the behavior of multiplication RGB mentioned in section 4.2.2 in terms of the noise present in the masks, and that increasing the weights toward classification would actually cause the network to have a more noisy, closer to a clear, mask to allow more information flow. This variation of network however did maintain much better classification results than the Dense-net architecture.

4.3 Five-Fold Cross Validation

A technique to avoid overfitting and biasing of the data is five-fold cross validation. It ensures that the model is trained and tested on different variations of the overall dataset every time. Five-fold cross validation means that the data will be split into 5 equal segments, and it will be trained on 4/5 of this data and will be tested on the remaining 1/5. Training and testing will occur 5 times in a way that each segment will have a chance to be used as a testing set. The average and the standard deviation of this mechanism is recorded and used to compare between different results. In this research it will be applied on the Dense-net framework and to the combined architectures with the selected best as weights highlighted in section 4.2. The results for five-fold cross validation are illustrated in Table 19. The results show that adding an extra network that predicts a saliency mask improves the overall results of the network. The RGB input models outperform the HS input models in both multiplication and concatenation

operation. The value for classification accuracy has decreased as tests are now done on the entire dataset, which may be more challenging than the subset we used earlier.

Table 19: Five-fold cross validation results

Label	MAE	Acc	F1m
RGB – CAT	0.196	86.25	86.18
RGB - MUL	0.2168	85.675	85.57
HS - CAT	0.2097	85.325	85.323
HS - MUL	0.2108	84.6125	84.608
Dense-net	-	83.025	82.86978

Additionally, The RGB concatenation model had the highest classification accuracy and F1 measure which shows that adding in more information the network would lead to better classification results. The MAE results from the past tables can be safely associated with the degradation in the overall performance of the attention map prediction. The least performance was achieved by the HS multiplication model which showed a poor performance in terms of mask prediction as well due to the shared weights.

Chapter 5. Conclusion and Future Work

This thesis addressed two main contributions which are: predicting medical masks from saliency architecture trained on natural images, and the improvement of classification accuracy when a network jointly predicts salient masks with classification labels. As for the first contribution, four different U-net variants were introduced: U-net original, U-net Skip, U-net Dense, U-net with strided convolution. The first three mentioned architectures were examined on DUTS and DUTOMRON natural image datasets to check their performance using accuracy and F1 scores. The U-net original showed very close results to the U-net Skip architecture. Hence, the original U-net was further modified and introduced as the fourth variant which is U-net strided convolution which showed that replacing max-pooling layer with strided convolution would improve the results of predicting saliency masks for natural images.

The U-net with strided convolution trained on natural images was utilized to predict the saliency masks of the medical images. Different input variation of the same inputs, RGB, HS and Value, were analyzed to check the effect of changing the color space on the quality of the predicted masks visually. The medical masks quality was analyzed subjectively. The HS input showed the best quality of masks in term of capturing the proper attention regions from the medical inputs. The masks of the best predicted quality were saved as ground truths for the next step.

As for the second contribution: four different variations of the framework were explored. U-net architecture was merged with a Dense-net architecture for jointly predicting saliency masks and classifying input images to their corresponding labels. The different variation were as follows: RGB input concatenated with predicted mask and fed to Dense-net, RGB input multiplied with predicted mask and fed to Dense-net, HS input concatenated with predicted mask and fed to Dense-net and HS input multiplied with predicted mask and fed to Dense-net. Different weights for the cost function of these architectures were applied to check the behavior of assigning different weights on classification accuracy and on masks prediction on a subset of images. The best set of weights which had a decent improvement in classification accuracy and a visually acceptable mask were chosen then for each of the frameworks for final comparison using 5-fold cross validation. It is fair to conclude that changing the weights

toward the classification loss function would increase the classification accuracy yet worsen the mask's quality showing that both loss functions are competing.

The networks which had multiplication layers in the middle illustrated that as the weights of the cost functions shift towards classification, the mask becomes more speckle like, which shows that the network is trying to get further data from the input image. Networks constituting concatenation in their middle layer, showed the classification would improve no matter the overall quality of the mask. Therefore, we can conclude that adding masks by multiplication or concatenation had illustrated an overall performance enhancement compared to classifying input images only using the classification network (Dense-net).

For future work, different classification architectures other than the Dense-net should be tried to validate the theory that providing extra mask information improves overall classification results even further. Additionally, the masks could be assigned a non-linear learnable weight factor, as a function for mask intensity values, which may improve the results further. Moreover, the work could be expanded to study other types of medical images, other than GI tract images, to achieve similar improvement in classification with addition of saliency inputs.

References

- [1] "20 Types of Physicians in High Demand." Medical Blog | St. George's University. <https://www.sgu.edu/blog/medical/types-of-physicians-in-demand/> (accessed 10-8-2019, 2019).
- [2] Dyson and Tauren. "Demand for doctors in U.S. jumped by 7 percent in 2017." United Press International. https://www.upi.com/Health_News/2018/12/05/Demand-for-doctors-in-US-jumped-by-7-percent-in-2017/5191544015835/ (accessed 10-08-2019, 2019).
- [3] D. S. Kermany *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, 2019/08/17 2018, doi: 10.1016/j.cell.2018.02.010.
- [4] S. Patel and J. Pingel. "Introduction to Deep Learning: What Are Convolutional Neural Networks?" MathWorks. <https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html> (accessed 26/8/2019, 2019).
- [5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7-12 June 2015 2015, pp. 3431-3440, doi: 10.1109/CVPR.2015.7298965.
- [6] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006.
- [7] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network," *IEEE Transactions on Big Data*, vol. 3, pp. 180-189, 2017, doi: 10.1109/TBDATA.2017.2717439.
- [8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234-241
- [9] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 25-28 Oct. 2016 2016, pp. 565-571.
- [10] G. Litjens *et al.*, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, p. 26286, 2016.
- [11] T. Brosch, L. Y. W. Tang, Y. Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, "Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1229-1239, 2016, doi: 10.1109/TMI.2016.2528821.
- [12] C. Wang *et al.*, "A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015: IEEE, pp. 2415-2418.
- [13] M. Ghafoorian *et al.*, "Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation," in *2016 IEEE*

- 13th International Symposium on Biomedical Imaging (ISBI)*, 13-16 April 2016 2016, pp. 1414-1417, doi: 10.1109/ISBI.2016.7493532.
- [14] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61-78, 2017, doi: <https://doi.org/10.1016/j.media.2016.10.004>.
- [15] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Fusing Unsupervised and Supervised Deep Learning for White Matter Lesion Segmentation," in *2nd International Conference on Medical Imaging with Deep Learning*, 2019, pp. 63-72.
- [16] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993-2024, 2014.
- [17] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 13-16 April 2016 2016, pp. 1397-1400, doi: 10.1109/ISBI.2016.7493528.
- [18] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, "Automated detection of pulmonary nodules in PET/CT images: Ensemble false-positive reduction using a convolutional neural network technique," *Medical Physics*, vol. 43, no. 6 Part1, pp. 2821-2827, 2019/08/17 2016, doi: 10.1118/1.4948498.
- [19] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale Convolutional Neural Networks for Lung Nodule Classification," in *International Conference on Information Processing in Medical Imaging*, 2015, pp. 588-599.
- [20] J. M. Wolterink, T. Leiner, B. D. de Vos, R. W. van Hamersvelt, M. A. Viergever, and I. Išgum, "Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks," *Medical Image Analysis*, vol. 34, pp. 123-136, 2016.
- [21] M. J. J. P. v. Grinsven, B. v. Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1273-1284, 2016, doi: 10.1109/TMI.2016.2526689.
- [22] Q. Dou *et al.*, "Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1182-1195, 2016, doi: 10.1109/TMI.2016.2528129.
- [23] Y. Komeda *et al.*, "Computer-Aided Diagnosis Based on Convolutional Neural Network System for Colorectal Polyp Classification: Preliminary Experience," *Oncology*, vol. 93(suppl 1), no. Suppl. 1, pp. 30-34, 2017, doi: 10.1159/000481227.
- [24] J. He, X. Wu, Y. Jiang, Q. Peng, and R. Jain, "Hookworm Detection in Wireless Capsule Endoscopy Images With Deep Learning," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2379-2392, 2018, doi: 10.1109/TIP.2018.2801119.
- [25] A. A. Shvets, V. I. Iglovikov, A. Rakhlin, and A. A. Kalinin, "Angiodysplasia Detection and Localization Using Deep Convolutional Neural Networks," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 17-20 Dec. 2018 2018, pp. 612-617, doi: 10.1109/ICMLA.2018.00098.

- [26] F. F. Ting, Y. J. Tan, and K. S. Sim, "Convolutional neural network improvement for breast cancer classification," *Expert Systems with Applications*, vol. 120, pp. 103-115, 2019, doi: <https://doi.org/10.1016/j.eswa.2018.11.008>.
- [27] "Digestive Diseases Statistics for the United States." National Institute of Diabetes and Digestive and Kidney Diseases. <https://www.niddk.nih.gov/health-information/health-statistics/digestive-diseases> (accessed 24/8/2019, 2019).
- [28] Webmd. "Digestive Diseases and Endoscopy." webmd. <https://www.webmd.com/digestive-disorders/digestive-diseases-endoscopy#> (accessed 13/05/19, 2019).
- [29] M. Clinic. "Colon Polyps." Mayo Clinic. <https://www.mayoclinic.org/diseases-conditions/colon-polyps/symptoms-causes/syc-20352875>" (accessed 5/8/2019, 2019).
- [30] A. C. Society. "Understanding Your Pathology Report: Colon Polyps (Sessile or Traditional Serrated Adenomas)." American Cancer Society. <https://www.cancer.org/treatment/understanding-your-diagnosis/tests/understanding-your-pathology-report/colon-pathology/colon-polyps-sessile-or-traditional-serrated-adenomas.html> (accessed 5/8/2019, 2019).
- [31] C. f. D. C. a. Prevention. "Parasites - Hookworm." U.S. Department of Health & Human Services. https://www.cdc.gov/parasites/hookworm/gen_info/faqs.html (accessed 8/5/2019, 2019).
- [32] J. Regula, E. Wronska, and J. Pachlewski, "Vascular lesions of the gastrointestinal tract," *Best Practice & Research Clinical Gastroenterology*, vol. 22, no. 2, pp. 313-328, 2008, doi: <https://doi.org/10.1016/j.bpg.2007.10.026>.
- [33] R. Zhu, R. Zhang, and D. Xue, "Lesion detection of endoscopy images based on convolutional neural network features," in *2015 8th International Congress on Image and Signal Processing (CISP)*, 14-16 Oct. 2015 2015, pp. 372-376, doi: 10.1109/CISP.2015.7407907.
- [34] A. Asperti and C. Mastronardo, "The effectiveness of data augmentation for detection of gastrointestinal diseases from endoscopical images," in *the 5th International Conference on Bioimaging*, 2017, pp. 588-599.
- [35] C. Gamage, I. Wijesinghe, C. Chitraranjan, and I. Perera, "GI-Net: anomalies classification in gastrointestinal tract through endoscopic imagery with deep learning," in *2019 Moratuwa Engineering Research Conference (MERCon)*, 2019: IEEE, pp. 66-71.
- [36] S. Nadeem, M. Tahir, S. Sadiq, A. Naqvi, and M. Memon, "Ensemble of Texture and Deep Learning Features for Finding Abnormalities in the Gastro-Intestinal Tract," in *International Conference on Computational Collective Intelligence*, 2018, pp. 469-478.
- [37] T. Agrawal, R. Gupta, S. Sahu, and C. Y. Espy-Wilson, "SCL-UMD at the Medico Task-MediaEval 2017: Transfer Learning based Classification of Medical Images," in *MediaEval*, 2017, pp. 512-515.
- [38] R. Cong, J. Lei, H. Fu, M. Cheng, W. Lin, and Q. Huang, "Review of Visual Saliency Detection With Comprehensive Information," *IEEE Transactions on*

- Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2941-2959, 2019.
- [39] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, and D. Hu, "Salient region detection via integrating diffusion-based compactness and local contrast," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3308-3320, 2015.
- [40] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2814-2821.
- [41] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569-582, 2014.
- [42] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 136-145.
- [43] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166-3173.
- [44] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015, doi: 10.1007/s11263-015-0816-y.
- [45] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485-3492.
- [46] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, 29 Sept.-2 Oct. 2009 2009, pp. 2106-2113, doi: 10.1109/ICCV.2009.5459462.
- [47] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *European Conference on Computer Vision*, 2010: Springer, pp. 30-43.
- [48] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*: Springer, 2018, pp. 3-11.
- [49] C. McQuin *et al.*, "CellProfiler 3.0: Next-generation image processing for biology," *PLoS biology*, vol. 16, no. 7, pp. 150-163, 2018.
- [50] X. Li, H. Chen, X. Qi, Q. Dou, C. Fu, and P. Heng, "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663-2674, 2018, doi: 10.1109/TMI.2018.2845918.
- [51] S. N. Gowda and C. Yuan, "ColorNet: Investigating the importance of color spaces for image classification," in *Asian Conference on Computer Vision*, 2018: Springer, pp. 581-596.
- [52] K. Pogorelov *et al.*, "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection," in *the Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 164-169
- [53] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and*

- Pattern Recognition (CVPR)*, 21-26 July 2017 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.
- [54] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017, doi: <https://doi.org/10.1016/j.media.2017.07.005>.
- [55] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3089-3098.

Appendix

Table 20: Training: DUTS Training | validation: DUTS Test | metric: validation loss

Training: DUTS Training Validation: DUTS Test Metric: validation loss			
Learning rate	U-net	U-net skip	U-net dense
1e-4	0.0862	0.0870	0.2367
1e-5	0.0810	0.0835	0.0773
1e-6	0.2472	0.1168	0.1329

Table 21: Training: DUTS Training | validation: DUTSOMRON | metric: validation loss

Training: DUTS Training Validation: DUTSOMRON Metric: validation loss			
Learning rate	U-net	U-net skip	U-net dense
1e-4	0.0830	0.0811	0.0742
1e-5	0.0857	0.0828	0.0766
1e-6	0.1170	0.1201	0.1300

Table 22: Training: DUTS Training + Test | validation: DUTSOMRON | metric: validation loss

Training: DUTS Training + Test Validation: DUTSOMRON Metric: validation loss			
Learning rate	U-net	U-net skip	U-net dense
1e-4	0.0653	0.0624	0.0656
1e-5	0.0760	0.0711	0.0703
1e-6	0.1184	0.1026	0.2462

Vita

Mahmoud Rezk was born in 1996, in Dammam, Saudi Arabia. He received his primary and secondary education in Dammam, Saudi Arabia. He received his B.Sc. degree in Electrical Engineering from the American University of Sharjah in 2018.

In September 2018, he joined the Electrical Engineering master's program in the American University of Sharjah as a graduate teaching assistant. His research interests are machine learning, deep learning, computer vision and big data analytics.