Multi-lane LBP-Gabor Capsule Network with K-means Routing for Medical Image Analysis

Patrick Kwabena Mensah[1], Anokye Acheampong Amponsah[2], Kwame Baffour Agyemang[3], Mighty Abra Ayidzoe[5], Faiza Umar Bawah[6], Adebayor Felix Adekoya[7], Benjamin Asubam Weyori[8], Mark Amo-Boateng[9]
Department of Computer Science & Informatics
University of Energy and Natural Resources
Sunyani, Ghana

Gabriel Kofi Armah[4]
Department of Business Computing
Faculty of Computing and Information Sciences (FCIS)
University of Technology and Applied Sciences, Navrongo, Ghana

***Abstract*—Medical images naturally occur in smaller quantities and are not balanced. Some medical domains such as radiomics involve the analysis of images to diagnose a patient's condition. Often, images of sick inaccessible parts of the body are taken for analysis by experts. However, medical experts are scarce, and the manual analysis of the images is time-consuming, costly, and prone to errors. Machine learning has been adopted to automate this task, but it is tedious, time-consuming, and requires experienced annotators to extract features. Deep learning alleviates this problem, but the threat of overfitting on smaller datasets and the existence of the "black box" still lingers. This paper proposes a capsule network that uses Local Binary Pattern (LBP), Gabor layers, and K-Means routing in an attempt to alleviate these drawbacks. Experimental results show that the model produces state-of-the-art accuracy for the three datasets (KVASIR, COVID-19, and ROCT), does not overfit on smaller and imbalanced datasets, and has reduced complexity due to fewer parameters. Layer activation maps, a cluster of features, predictions, and reconstruction of the input images, show that our model is interpretable and has the credibility and trust required to gain the confidence of practitioners for deployment in critical areas such as health.**

Keywords*—Convolutional neural networks; deep learning; Gabor filters; k-means routing; local binary pattern; power squash introduction

I. Introduction

Health is among the top critical areas that affect human life. For instance, 50,000 people die each year from pneumonia in the United States whereas colorectal polyps are projected to increase by 60% in 2030 which is likely to increase the number of causalities [1]. Images, videos, and text are the commonly generated and analyzed data used for the evaluation of most medical conditions. The analysis of these data requires the expertise of professionals which is rare and expensive in some regions and additionally susceptive to human errors [2], less effective [3], and falls below recommended levels in clinical procedures [4]. This calls for computer vision-assisted diagnosis. Machine learning-based methods such as support vector machines have been employed to assist in the effective diagnosis of medical diseases [5]. However, the performance of these methods was below the standard practices and the feature extraction procedure is time-consuming. To address these issues, deep learning models such as convolutional neural networks (CNNs) were adopted to improve feature extraction. Interestingly, CNNs achieved performance rivaling human experts. For example, a CNN model made up of 121 layers (termed CheXNet), was trained on 100,000 frontal view chest X-rays and performed far better than 4 radiologists [6].

Regardless of CNN's good performance, the research identified certain limitations such as being translationally invariant [7], requiring large datasets, being computationally expensive [8], and following certain criteria for effective feature selection [9]. In health, the availability of a large dataset is a major challenge coupled with the lack of unavailability of qualified annotators [8]. Therefore, to prevent CNNs from overfitting on these small datasets, data augmentation techniques are employed. These data augmentation techniques are time-consuming and laborious.

To address these challenges, Capsule Network (CapsNet) [7] was introduced, and unlike CNNs, they do not require large datasets making them suitable for health applications. Notwithstanding, CapsNets also have their limitations. They perform poorly on complex images and images with varied backgrounds, have complex routing processes, poor learning of lower-level description [10], and polarization.

The contributions of this paper, therefore, are a) architectural innovation: we propose a Local binary pattern (LBP) – Gabor Capsule Network to address the weak feature extraction problem and the inability of CapsNets to learn lower-level descriptions of a complex image, b) algorithmic innovation: we adopt K-means routing, power squash, and sigmoid functions to complement the feature extraction abilities of the LBP-Gabor layers, c) explainable artificial intelligence (XAI): we provide extensive visualizations of the outputs of our network in an attempt to "open" the "black box" in deep learning models for enhanced credibility and understandability.

The rest of the paper is organized as follows: Section 2 presents the related work leading to Section 3 where the proposed methods are outlined. The experiments and experimental results are presented in Section 4 and the work concluded in Section 5.

II. Related Work

The limitations of human-centered diagnosis led to the adoption of algorithms for predicting medical conditions found in domains such as "radiomics". Radiomics involves the use of data-characterization algorithms to extract features from radiographic images. Studies in the literature, such as Saif et al. [5] proposed a Capsule Network algorithm for the recognition of musculoskeletal conditions from radiographic images. The proposed model outperformed a 169-layer DenseNet in recognizing abnormality in musculoskeletal radiography. To address the inability of CNNs in encoding part-whole relationships, Mobiny et al. [11] proposed an efficient bi-directional long short-term recurrent capsule network for the recognition of apoptosis (cell death). The proposed model achieved competitive performance and outperformed CNNs especially when the number of training samples is small.

One of the deadliest medical conditions is brain tumors. Detection of the correct type of brain tumor at an early stage is vital to enable early treatment and reduce mortality in both children and adults. Consequently, there has been a surge of interest in developing efficient brain tumor detection algorithms. Afshar et al. [12] proposed a capsule network algorithm for the detection of a brain tumor on segmented images generated from the training images. The segmentation was done to avoid the negative effect of miscellaneous background objects on the model's performance. Afshar et al. [13] proposed a focus-oriented capsule network algorithm that takes coarse boundaries of brain tumor images as extra inputs to diagnose brain tumors. The proposed model achieved overall recognition accuracy of 90.89%.

Given the challenges encountered during human-centered diagnosis of other lung infections and COVID computed tomography (CT) scan and X-ray images, Afshar et al. [14] proposed a capsule network termed COVID-Caps. The proposed model achieved an accuracy of 95.7%, sensitivity of 90%, and specificity of 90% on small datasets of COVID-19. This study is more related to the works in [15, 16] and [17] where transfer learning and custom-built CNNs are designed to diagnose diseases such as COVID-19 and retinal diseases from Chest X-ray and retinal optical coherence tomography (ROCT) images respectively. However, we leverage on CapsNet's ability to avoid overfitting and identify the pose and deformation of objects and object parts to diagnose medical conditions from challenging medical images. Furthermore, the aforementioned works did extensive data preprocessing, augmentation, segmentation, and balancing of datasets (especially [15]) before fitting their models. We, however, used the raw datasets without augmentation and preprocessing to understand the model's performance on the natural data since it may not be feasible to perform augmentation or segmentation during a medical emergency. Although the work in [15] provided images of the regions recognized by the model, we provide elaborate visualizations of image regions that attract the attention of parts of our model, clusters of features at the class capsule layer to measure the performance of the routing algorithm, performance on imbalanced datasets in the form of Precision-Recall (PR) curves, and reconstruction of input images as a way to enhance model transparency and understandability.

III. Proposed Methods

In this section, we present the model modifications and the methodology adopted to achieve our objective of designing a capsule network with superior feature extraction capabilities compared to the original CapsNet. We avoid shallowness and at the same time strive to reduce the number of parameters by using layers that generate no or less trainable parameters.

A. K-Means Routing

We adopt the K-means routing in [18] with Sigmoid normalization, Power squash $v_j = \|v_j\|^n \frac{v_j}{\|v_j\|}$, and a modified logit updates procedure, instead of dynamic routing [7] in an attempt to minimize the problem of polarization [19] leading to improved performance on difficult medical images. Instead of using dot product and initializing b_{ij} with zero and adding the old logits to perform updates, our method respectfully uses the ℓ_2 distance measure, initializes b_{ij} as $b_j^{(0)} \leftarrow \sum_{i=1}^{n} \|W_{ij}u_i - v_j^{(0)}\|^2$ and does not add old logits to new logits during updates. Algorithm 1 shows the K-means routing procedure.

Algorithm 1 K-Means Routing for image classification [18]

1. **procedure** ROUTING (u_i, r)
2. **Initialize** $v_j^{(0)} \leftarrow \frac{1}{k}\sum_{i=1}^{k} W_{ij}u_i$
3. $b_j^{(0)} \leftarrow \sum_{i=1}^{n} \|W_{ij}u_i - v_j^{(0)}\|^2$
4. **for** r iterations **do**
5. $b_{ij} \leftarrow \sum_{i=1}^{n} \|W_{ij}u_i - v_j\|^2$
6. $c_{ij} \leftarrow Sigmoid(b_{ij})$
7. $v_j \leftarrow \sum_{i=1}^{n} c_{ij}W_{ij}u_i$
8. **return** $power_n(v_j)$

B. Feature Extraction with LBP and Gabor

Both LBP [20] and Gabor filters [21] have each been shown to be superior edge and texture feature extractors [22, 23] than convolutional layers [18, 24, 25].

Gabor filters belong to a special class of bandpass filters with frequency and orientation representation mimicking those of the mammalian cortex. They are made up of real and imaginary parts. It is the real part shown in equation 1 that is used to extract image features.

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma)=\exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x}{\lambda}+\psi\right) \quad (1)$$

where $x' = xcos\theta + ysin\theta$, $y = -xsin\theta + ycos\theta$ with (x, y) being the pixel position in the spatial domain. λ controls the width of the Gabor function strips, θ represents the orientation to the normal, ψ is the phase offset, γ is the spatial aspect ratio, and σ is the standard deviation of the Gaussian envelope. To extract features with Gabor filters, five frequencies f and eight orientations θ are adopted. These parameters are defined in equations 3 and 4.

$$=\frac{\pi}{2}\sqrt{2}^{-(n-1)} \quad (2)$$

$$\theta=\frac{\pi}{2}(m-1) \quad (3)$$

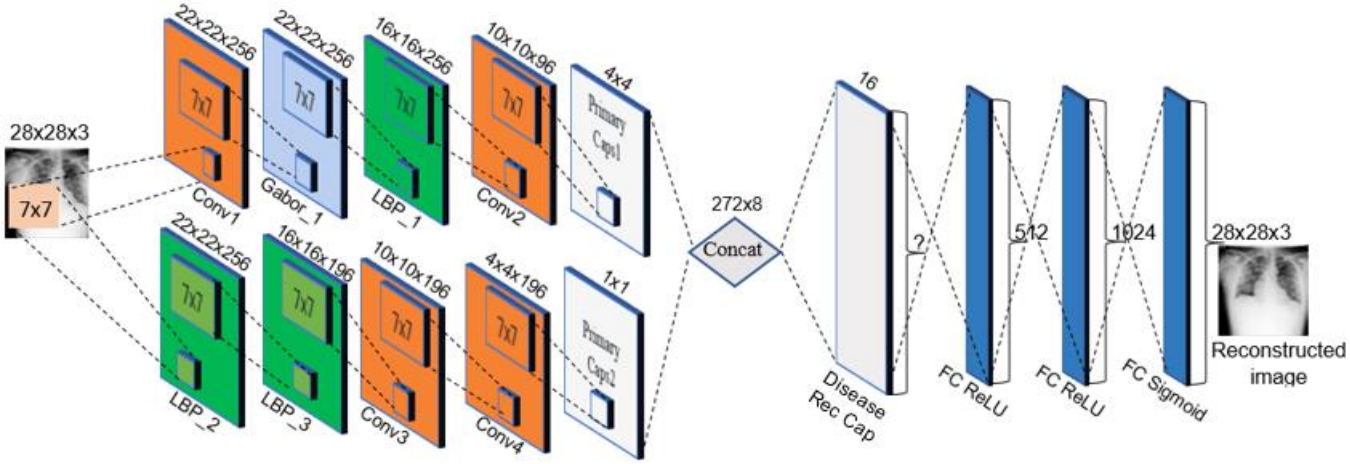


Fig. 1. The Gabor-LBP Capsule Network Architecture.

where $n = 1,2, \dots, 5$, $m = 1,2, \dots, 8$, and $\sigma = \frac{\pi}{f}$.

The Local Binary Pattern (LBP) [20] is a powerful feature extractor that adds no trainable parameters to a model when used to extract contrast and spatial patterns of an image. It accomplishes this by thresholding n neighbouring pixels and computing its equivalent binary number based on equation 4.

$$LBP(n,r) = \sum_{n=0}^{n-1} f(i_n - i_c) 2^n \quad (4)$$

where i_n= neighboring pixels' intensity, i_c= current pixels' intensity, n = number of selected neighboring pixels at radius r, and a sign function defined as $f = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$.

C. LBP-Gabor CapsNet Architecture

The proposed model is a combination of Conv-LBP-Gabor layers placed in a multi-lane fashion (see Fig. 1). The input images are resized to 28x28x3 and fed to both lanes simultaneously. The first lane (upper lane) has a conv1 layer made up of 256, 7x7 kernels with ReLU non-linear activation at a stride of 1 to produce 256, 22x22 feature maps. These feature maps serve as input to the Gabor_1 layer made up of 256, 7x7 kernels at a stride of 1 and valid padding to produce 256, 22x22 feature maps for subsequent layers. The feature maps are processed in this manner as they pass through each layer in lane one until they reach the Primary Caps 1 layer which is a convolutional capsule layer made up of 7x7 kernels with a stride of 2. It is a 16-component capsule each with 4x4 capsules in an 8-dimensional vector.

LBP_2 extracts the features directly from the input image to feed lane two (bottom lane). It is made up of 256, 7x7 kernels with stride 1 to produce 256, 22x22 feature maps. The features are refined as they pass through the rest of the layers to Primary Capsule 2 which has 3x3 kernels at a stride of 2. This too is a 16-component convolutional capsule each with a 1x1 capsule in an 8-dimensional vector.

The outputs of the two PCs are concatenated via axis 1 to produce a 272x8 dimensional tensor. It is the features of this tensor that are used for routing with the Disease recognition cap layer. The latter is 16-dimensional while the number of capsules is varied according to the number of classes in the dataset. We have used (?) to indicate that the number of capsules will vary from 8, 4, 4 for KVASIR, COVID-19, and ROCT datasets respectively. Reconstruction of the input image is carried out by the decoder. The quality of the reconstructed images (see Fig. A1 in Appendix A) depends on the performance of the classification.

IV. Experiments

In this section, we present the experiments conducted on each dataset as well as their respective results. Three publicly available datasets were used to evaluate the performance of the model's ability to generalize on unseen data.

A. Dataset Description

The Kvasir [24] is a dataset consisting of images from inside of the gastrointestinal (GI) tract. It consists of eight different classes made up of images from 720x576 to 1920x1072 pixels. The dataset can be used for multiclass classification [24] as the images can be categorized under three important anatomical landmarks. For a detailed description of this dataset, readers are encouraged to look at the work in [24]. This dataset is not balanced.

The COVID-19 dataset [16, 17] was collected by a team of doctors from 4 countries, and it is made up of chest X-ray images of COVID-19 positive cases plus some Normal and Viral Pneumonia images. Categories such as COVID, Lung_Opacity, Normal, and Viral_Pneumonia form the class in this dataset. This dataset is also imbalanced and details can be found in [16, 17].

The Retinal Optical coherence tomography (ROCT) dataset [15] contains high-resolution cross-sectional images of the retina. The dataset was collected from adult patients at the Shiley Eye Institute of the University of California San Diego, the California Retinal Research Foundation, Medical Center Ophthalmology Associates, the Shanghai First People's Hospital, and Beijing Tongren Eye Center [25]. It has four classes and is originally organized such that each test set has 250 images while the training set has 20,135 (i.e. approximately 95% to 5% train-test split). We, however, split all the three datasets into 80% training and 20% test. Additionally, we did not perform data augmentation to any of

our datasets as a means to measure the ability of the proposed model to decode the spatial orientation of the images. A summary of the datasets used in this study is provided in Table A1 in Appendix A.

B. Experimental Setup

We performed all the experiments using the following tools and software; Keras with TensorFlow backend, one 64-bit Windows machine with NVIDIA GeForce GTX 1060 Graphic Processing Unit (GPU), 8GB GPU memory, 16GB system memory, and CUDA 10.1 toolkit. Hyperparameters such as the number of epochs, batch size range, learning rate, learning rate decay, and early stopping were respectively set to 100, 50-100, 0.001, 0.9, and 15. We varied the number of routing iterations from 2 to 7 (see Section 4.5) to test the ability of the model to scale up. To calculate the loss, we adopted the margin loss from [7]. This loss is given by:

$$L_k = T_k \max(0, m^+ - \|v_k\|)^2 + \lambda(1 - T_k) \max(0, \|v_k\| - m^-)^2$$

where $T_k = \begin{cases} 1 & \text{if class k is active} \\ 0 & \text{otherwise} \end{cases}$, $\lambda = 0.5$, $m^+ = 0.9$, and $m^- = 0.1$

We adopted, customized, and modified the code from *https://github.com/XifengGuo/CapsNet-Keras* for this study.

C. Experimental Results

We present the experimental results in this Section and show that the model performed well when evaluated on the three datasets. To enhance confidence and reliability in the model's results, several evaluation methods were adopted and carefully conducted. Metrics such as the number of parameters, classification loss, and accuracy, the Area Under the Curve (AUC) for both the Receiver Operating Characteristic Curve (ROC) and Precision-Recall (PR) curves were used for the performance evaluation. Additionally, the model's robustness, ability to scale-up, fail-safe, extract only relevant features and the performance of the routing process were also evaluated. The traditional capsule network was also trained with the datasets and the results compared to our model based on the aforementioned performance metrics.

D. Accuracy

We used the multi-class confusion matrix to summarize the performance of the model on the datasets. This method includes powerful per-class metrics such as true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). The values in the principal diagonals of the confusion matrices are the TP values representing the level of correct identification of the true classes from the respective datasets. Few FNs as seen from Fig. A2 in Appendix A. indicates a good performance considering the field of application (i.e. health). In other words, the high TP values indicate good performance for a disease recognition model since it is not fatal for a healthy medical image (and by extension a healthy person) to be categorized as sick compared to when a sick person is classified as healthy.

From the confusion matrices, the accuracy of the model can be computed based on equation 5.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

It is worth noting that accuracy, even though very popular [26] at evaluating classification algorithms, is not appropriate for medical images since they tend to be small and highly imbalanced [27]. Despite its drawback, it can provide a snapshot of the entire system performance, especially when the datasets are balanced..

The performance of the model in terms of accuracy during training and validation can be monitored via the training and validation curves. These curves for the three datasets are depicted in Fig. 2, with (c) and (d) depicting that the model had some difficulty in extracting the relevant features from the COVID-19 dataset. This is indicated by the zig-zag nature of the curves. Consistently, the proposed GLC model outperformed the traditional capsule network on the respective datasets in terms of training and validation accuracy/loss. A comparison of the accuracies of the proposed model, the traditional CapsNet, and other models in the literature on the same datasets are shown in Table A2 in Appendix A. The 93.40% accuracy of [15] on the ROCT dataset was obtained on the original 95%-5% train-test split. However, we split the data into 80%-20% for training and testing respectively. Unavailable values in Table A2 in Appendix A are indicated by (?).

To further probe the superiority of the proposed model, we performed additional experiments to determine the accuracy of the model as it is subjected to architectural damages in what is known as ablation studies (see Section 4.6). Additionally, we performed more experiments to explore the effect of increasing the capacity of the model on accuracy by increasing the number of routing iterations from 2 to 7 (see Section 4.5).

E. Model's Ability to Scale

Dynamic routing has an inner loop [28] [18] which contributes to hindering the algorithm to scale on complex data and increases the threat of overfitting when the network capacity is increased through an increase in the number of routing iterations. To test the models on this score, we varied the number of routing iterations and the results of these experiments are depicted in Fig. A3 in Appendix A. It is observed that the proposed GLC maintains a marginal loss in accuracy for both KVASIR (Fig. A3 (a)) and COVID-19 (Fig. A3 (b)) as the number of routing iterations increases from 2 to 7. On the contrary, the traditional model begins to overfit after the third routing iteration (Fig. A3 (a)), probable because the number of classes is comparatively higher than the other datasets while at the same time the number of images in the dataset is relatively smaller. As the traditional model scales up, it becomes "hungrier" for data and tends to depend on the number of classes, consequently increasing the number of interrelationships to a level likely to cause overfitting.

We also observe from Fig. A3 (Appendix) that at 3 iterations, the traditional CapsNet achieved optimal performance as established in [7], however, this varies for the proposed model. For instance, GLC's accuracy for KVASIR and ROCT are highest at 2 and 4 routing iterations respectively.

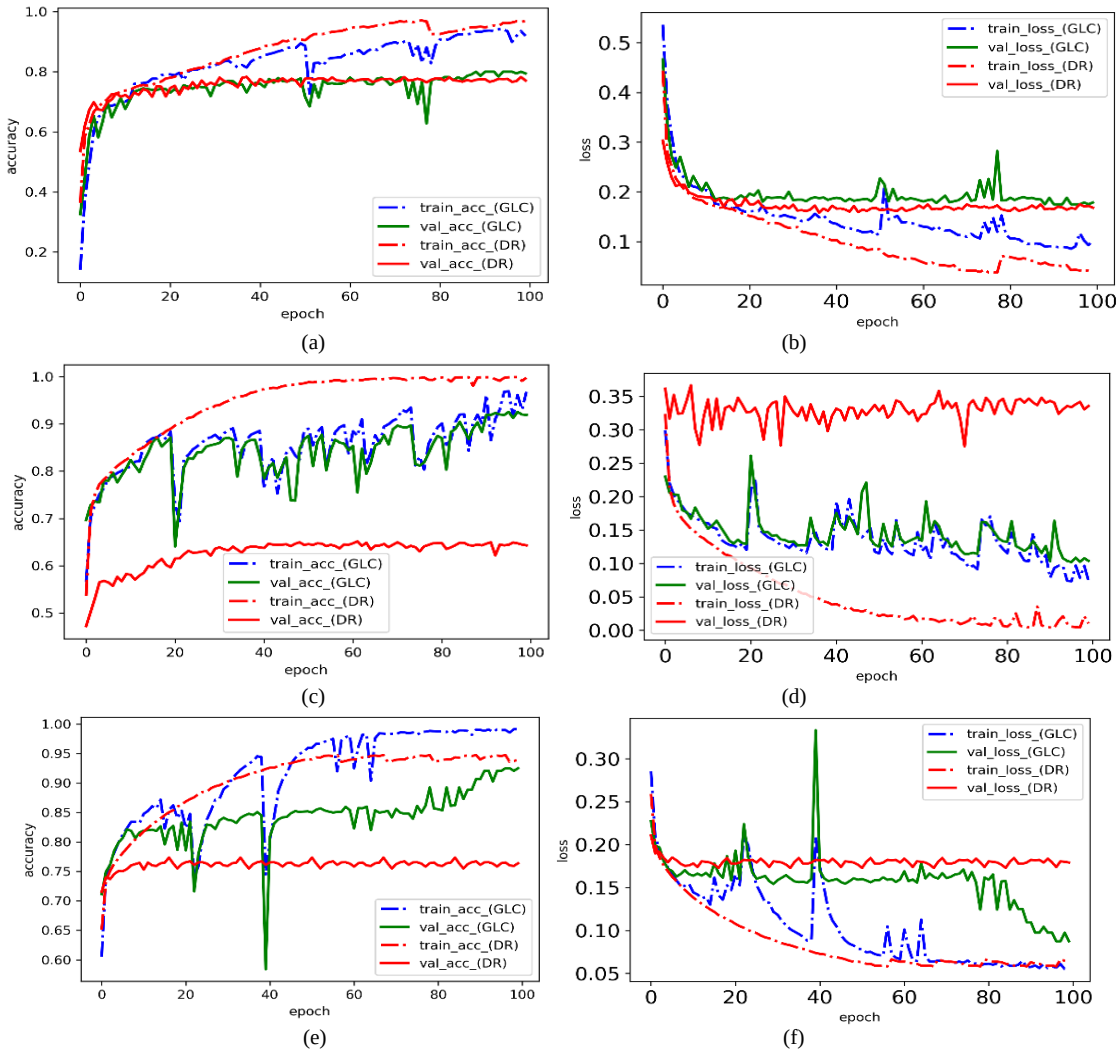


Fig. 2. Training and Validation Curves (a) Accuracy for KVASIR, (b) Loss for KVASIR (c) Accuracy for COVID-19, (d) Loss for COVID-19, (e) Accuracy for ROCT, and (f) Loss for ROCT Dataset.

F. Model's Robustness and Ability to Fail-Safe

Setting the number of routing iterations to 3, we performed additional experiments to determine parts and configurations of the model that made significant contributions to its high performance. We removed layers at a time and trained the network to measure the effect of their presence/absence in the network. Also, hyperparameters such as the squash and normalizer were varied and several pieces of training were carried out. This technique is called ablation study [29], and it can determine the ability of a network to fail-safe or undergo graceful degradation. Graceful degradation is a required property for critical applications. It is also a means to enhance confidence in the model since network components with the ability to stand in for failed parts can be identified and to also test for the robustness of the model to architectural changes.

From Table 1, Conv1 (row 1) and LBP2 (row 7) are very crucial in the network due to their positions as lower-layer (primary) feature extractors. Their removal causes a drop in accuracy across all the datasets. However, the removal of any of the rest of the conv layers causes a slight drop in accuracy, an indication that we could comfortably remove any one of them in situations where our objective is to reduce model parameters/size. Again, removing all the conv layers (row13) seems to have little effect on the performance compared to removing all LBP (row 11) and all LBP plus Gabor (row 12) layers. Rows 16 and 17 indicate the use of all layers in the network. We observe that the combination of the original squash and SoftMax underperformed relative to that of the Power squash and Sigmoid normalization consistent with what was reported in [18].

TABLE I. RESULTS OF ABLATION STUDY ON THE GLC MODEL

No	Layer(s) removed	Squash	Normalizer	Validation Accuracy (%)		
				KVASIR	COVID-19	ROCT
1	Conv1	Power	Sigmoid	79.11	90.02	90.13
2	Conv2	Power	Sigmoid	79.43	90.76	91.01
3	Conv3	Power	Sigmoid	79.51	90.62	90.35
4	Conv4	Power	Sigmoid	79.60	90.71	90.32
5	Gabor	Power	Sigmoid	77.07	80.23	81.56
6	LBP1	Power	Sigmoid	72.97	81.66	82.97
7	LBP2	Power	Sigmoid	70.94	79.52	80.50
8	LBP3	Power	Sigmoid	73.43	81.05	82.03
9	Gabor+LBP1	Power	Sigmoid	76.02	79.81	81.43
10	Conv1+Conv2	Power	Sigmoid	79.55	89.05	90.17
11	LBP1+LBP2+LBP3	Power	Sigmoid	66.91	70.12	75.57
12	Gabor+LBP1+…+LBP3	Power	Sigmoid	65.43	70.63	77.09
13	Conv1+…+Conv4	Power	Sigmoid	79.11	88.95	89.98
14	Lane 1 (top lane)	Power	Sigmoid	75.29	79.85	81.00
15	Lane 2 (bottom lane)	Power	Sigmoid	76.71	80.32	84.21
16	None	Original	SoftMax	78.54	88.70	89.01
17	None	Power	Sigmoid	80.91	91.96	91.30

G. Performance on Smaller and Imbalanced Datasets

Medical images are usually smaller and highly imbalanced [33]. Class imbalance, on the other hand, contributes to a problem called the "accuracy paradox" [31] which causes the larger classes to overshadow the smaller classes during accuracy computations. In other words, accuracy under these conditions is influenced or biased towards the class with the highest number of samples. Besides, the asymmetric misclassification costs and probability estimates of the classification are not taken into consideration during accuracy computations under class imbalance. The AUCs for the ROC and PR curves become handy when fitting a model with balanced and imbalanced classes respectively [36, 37]. The AUC is invariant to the a priori likelihoods of the classes as well as being independent of the decision threshold [34]. Large AUCs are preferred over their smaller counterparts.

Fig. 3. shows the ROC and PR curves for the GLC model. We observe that the ROC curves have relatively larger areas separating them from the diagonal. The impression is that the model performed very well in all the classes, however, the PR curves depict that the model did not perform equally well in all the classes. This is so because ROC tends to be overly optimistic with insufficient data [35] as well as when there is a large skew in the dataset class distribution [32]. A medical practitioner ultimately needs to see the PR curves of a model (not only accuracy) before taking critical decisions on a patient's condition. Compared to the ROC and PR curves of the DR model (shown in Fig. 3), the GLC model outperformed the traditional CapsNet model under class imbalanced conditions. The respective AUC values are; ROC -KVASIR (0.96), PR-KVASIR (0.71), ROC-COVID-19 (0.97), PR-COVID-19 (0.95), ROC-ROCT (0.93), and PR-ROCT (0.87).

On smaller datasets, CapsNets are known to outperform convolutional neural networks due to the ability of CapsNets to encode pose and orientation. This reason, plus our superior feature extractors explain why our model performed well on the KVASIR dataset (see Fig. 2(a)) without any data augmentation.

H. Prediction and Reconstruction

During prediction, the capsule outputs the class with the longest vector as the correct class. It is compared with the ground truth (GT) image to measure how well the trained model can classify an unseen image. This aspect of the model is very crucial for health applications since it quantifies the confidence the model has in its prediction. To introduce variability in the testing set, 1% of each dataset was reserved for prediction, and as such was not used as part of the training set. Sample prediction results on the unseen images are shown in Fig. A1 in Appendix A. The KVASIR dataset (Fig. A1 (a)) has eight classes, each of which is assigned a likelihood of being the correct class. The class with the highest probability is the predicted class. For both KVASIR and COVID-19 predictions, the model misclassified 0.5% of the unseen images (e.g. Fig. A1 (a) row 5 and Fig. A1 (b) row 4). We observe that the model imposed huge confidence (83%) in predicting class 2 of the KVASIR dataset as the correct class (Fig. A1 (a) row 3) while at the same time predicting class 1 with the confidence of 82% for the COVID dataset (Fig. A1 (b) row 5).

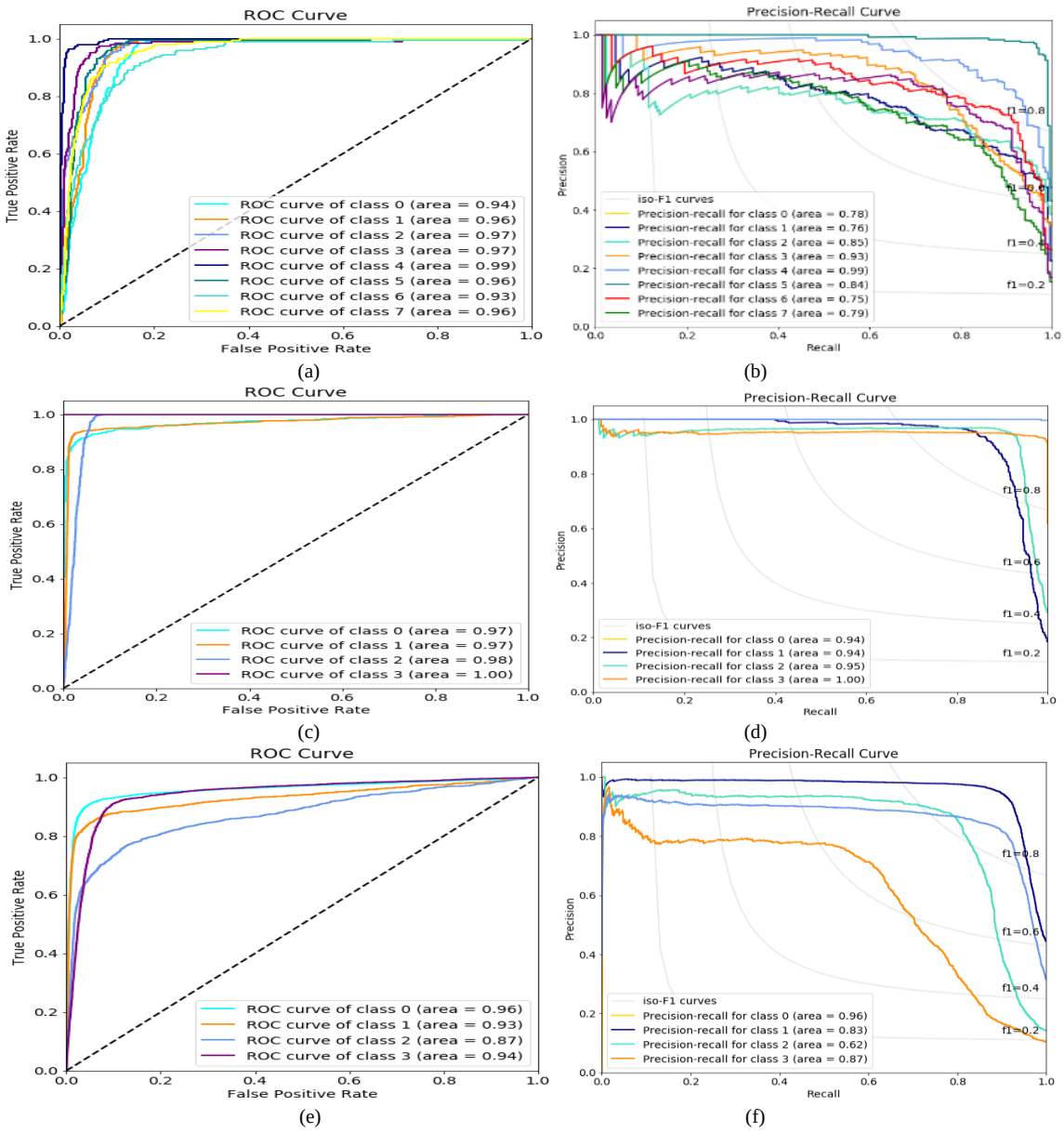


Fig. 3. ROC and PR Curves for the GLC Model. (a) ROC -KVASIR, (b) PR-KVASIR, (c) ROC-COVID-19, (d) PR- COVID-19, (e) ROC-ROCT, and (f) PR-ROCT.

Reconstruction allows visual verification of the model's output/performance and also works as a regularizer. The reconstructed images in Fig. A1 in Appendix A. are clearly showing that the network layers effectively used the instantiating parameters to reconstruct the input mages (GT). We also carried out predictions and reconstruction on the ROCT dataset as well as using the DR model to predict and reconstruct unseen images from the three datasets. The DR model misclassified 1% of the unseen images across the 3 datasets. These results, however, are omitted for brevity.

I. Model Complexity

Smaller deep learning models are required for efficient implementation on embedded devices such as FPGAs and mobile phones with limited memory [36]. Such models are also important for reducing overhead to make distributed online training and inference possible. The smaller the number of a model's trainable parameters, the less computationally complex the model is. This reduces the number of resources required by the model and also helps to prevent overfitting by ensuring that an l-layer capsule model has ln+k parameters required to exactly fit a d-dimensional dataset with n samples [37]. Our proposed model (see Fig. 1.) is deeper than the traditional CapsNet, but with a comparatively fewer number of parameters as shown in Table A3 in Appendix A. The values in Table A3 (Appendix) confirm that relatively smaller CapsNet models can represent complex real-life functions to outperform models with huge parameters [43, 18].

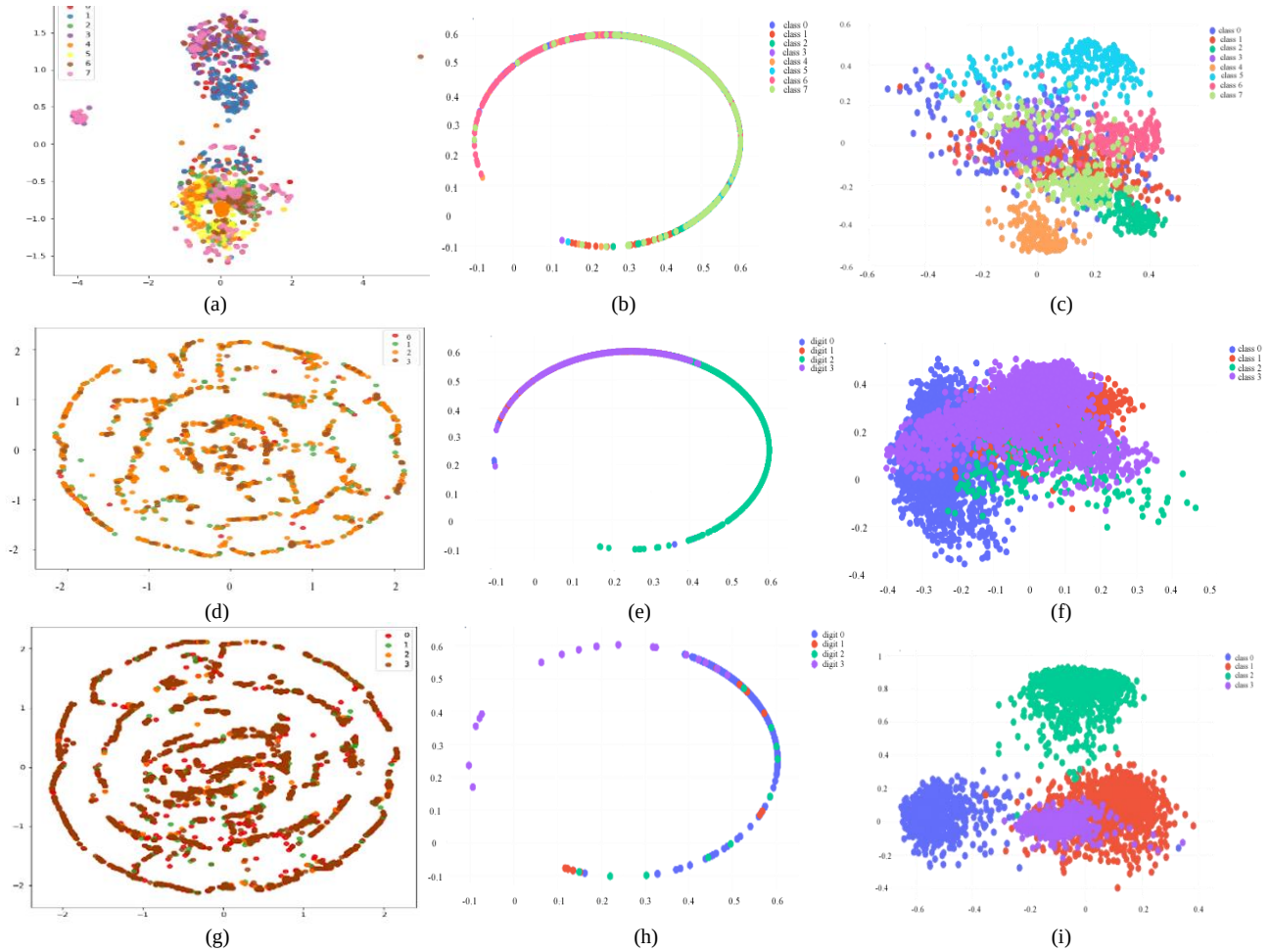


Fig. 4. *Tsne* Visualization of the Network's Raw and Learned Features at the Class Capsule Layer, (a) KVASIR Raw Test Set, (b) GLC clusters of KVASIR, (c) DR Clusters of KVASIR, (d) COVID-19 Raw Test Set, (e) GLC Clusters of COVID-19, (f) DR Clusters of COVID-19, (g) ROCT Raw Test Set, (h) GLC clusters of ROCT, and (i) DR Clusters of ROCT.

J. Performance of the Routing Process

We use the t-distributed stochastic neighbor embedding (TSNE) to visualize the network learned features at the class capsule layer. This method helps us to visually determine the level to which the network can differentiate between the different classes. Since primary capsules are coupled with secondary capsules with which there is a high agreement a_{ij} during routing, the features involved can be modeled as clusters. The compactness and separability of these clusters in the feature space indicate the performance of the routing algorithm at effectively making a distinction between the various classes. From Fig. 4., we observe that the clusters formed by the GLC model (second column); even though overlapping, are separable and some compact compared to those formed by the DR model (third column). These properties are linearly related to the performance of the routing algorithm and may be essential for further decision-making in case-by-case-based health applications.

We note that the reason for the GLC model forming circular clusters is that the routing algorithm is driven by K-means whose clusters are naturally circular from its use of the l_2 norm [39].

K. Feature Extraction

To uncover the network layers with good texture, edge, and shape feature extraction capabilities, we performed experiments to visualize the activation maps of the layers. This method is useful as it provides the opportunity to identify regions in the input image responsible for the activation of parts of the network. It also contributes to investigating whether a model is robust and can avoid failure through the inspection of the presence of layers with redundant features. Aside from the threat of overfitting resulting from model complexity, redundant layers are major contributors to a model's robustness and fault tolerance capabilities. On the other hand, through this method, redundant layers can be eliminated to improve the model's feature extraction to consequently reduce excessive oscillations and prolonged convergence during training [18]. This is a vital step for medical applications since it contributes significantly to the explainability and understandability of the "black box" [40] required to enhance confidence in model outputs for critical applications.

Fig. 5. Comparison of the Activation Maps of the Proposed GLC and the DR Models. (a) the First and Second Rows Show the Activation Maps for GLC and DR respectively on KVASIR, (b) Row One Shows the Activation Maps of GLC while Row Two Shows the Activation Maps of the DR Model for COVID-19, and (c) First and Second Rows are respectively the Activation Maps of GLC and DR Models for ROCT Dataset.

The feature maps in Fig. 5. show that the Gabor and LBP layers in the GLC have superior feature extraction capabilities than the convolutional layers. The Conv1 layer of the GLC network extracts some quality features since it is a higher-level layer with the ability to sample features from the lower-level layers (Gabor and LBP1) to represent advanced parts of the GT image. On the contrary, the Conv1 layer of the DR model is a lower-level layer, and with the difficulty of CNNs to extract quality features [18], it is not able to extract relevant features as required.

V. Conclusion and Future Work

In this paper, we propose a capsule network architecture with superior feature extraction capabilities for the recognition of medical conditions in medical images. The adoption of Local Binary Pattern (LBP), Gabor layers, and K-Means routing in an innovative architecture has dramatically improved the model's feature extraction capabilities leading to an appreciable performance while scaling up, preventing overfitting under class imbalance, and obtaining competitive validation and test accuracies. We further subjected the model through extensive visualization of layer activation maps, cluster of features, and ablation studies to enhance model interpretability and confidence for practical adoption. The results indicate that, it is possible to develop deep models to have smaller number of parameters (hence lower complexity) with huge potential for implementation on embedded devices with lower memories.

In the future, we will perform extensive experiments on these medical datasets for purposes of explainable artificial

intelligence (XAI). The aim will be to eliminate every ambiguity on model outputs to pave the way for its practical adoption in health.

REFERENCES

[1] P. Wieszczy, J. Regula, and M. F. Kaminski, "Adenoma detection rate and risk of colorectal cancer," Best Pract. Res. Clin. Gastroenterol., vol. 31, no. 4, pp. 441–446, 2017, doi: 10.1016/j.bpg.2017.07.002.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] M. Akbari et al., "Polyp Segmentation in Colonoscopy Images Using Fully Convolutional Network," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2018-July, pp. 69–72, 2018, doi: 10.1109/EMBC.2018.8512197.K. Elissa, "Title of paper if known," unpublished.

[3] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," Gut, vol. 66, no. 4, pp. 683–691, 2017, doi: 10.1136/gutjnl-2015-310912.

[4] R. Lalonde, P. Kandel, C. Spampinato, M. B. Wallace, and U. Bagci, "Diagnosing Colorectal Polyps in the Wild with Capsule Networks," Proc. - Int. Symp. Biomed. Imaging, vol. 2020-April, pp. 1086–1090, 2020, doi: 10.1109/ISBI45749.2020.9098411.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] A. F. M. Saif, C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "Abnormality Detection in Musculoskeletal Radiographs Using Capsule Network," *IEEE Access*, vol. 7, pp. 81494–81503, 2019, doi: 10.1109/ACCESS.2019.2923008.

[6] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv:1711.05225v3 [cs.CV]*, pp. 3–9, 2017.

[7] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, vol. 2017-Decem, no. NIPS 2017, pp. 3857–3867.

[8] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016, doi: 10.1109/TMI.2016.2535302.

[9] X. Zhang *et al.*, "Real-time gastric polyp detection using convolutional neural networks," *PLoS One*, vol. 14, no. 3, pp. 1–16, 2019, doi: 10.1371/journal.pone.0214133.t005.

[10] S. Sabour, A. Tagliasacchi, S. Yazdani, G. E. Hinton, and D. J. Fleet, "Unsupervised part representation by flow capsules," *arXiv:2011.13920v2 [cs.CV]*, 2021.

[11] A. Mobiny, H. Lu, H. V. Nguyen, B. Roysam, and N. Varadarajan, "Automated Classification of Apoptosis in Phase Contrast Microscopy Using Capsule Network," *IEEE Trans. Med. Imaging*, vol. 39, no. 1, pp. 1–10, 2020, doi: 10.1109/TMI.2019.2918181.

[12] P. Afshary, A. Mohammadiy, and K. N. Plataniotis, "BRAIN TUMOR TYPE CLASSIFICATION VIA CAPSULE NETWORKS," *arXiv:1802.10200v2 [cs.CV]*, 2018.

[13] P. Afshar, K. N. Plataniotis, and A. Mohammadi, "Capsule Networks for Brain Tumor Classification Based on MRI Images and Coarse Tumor Boundaries," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 1368–1372, 2019, doi: 10.1109/ICASSP.2019.8683759.

[14] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images," *Pattern Recognit. Lett.*, vol. 138, pp. 638–643, 2020, doi: 10.1016/j.patrec.2020.09.010.

[15] 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, 2018, doi: 10.1016/j.cell.2018.02.010.

[16] M. E. H. Chowdhury *et al.*, "Can AI Help in Screening Viral and COVID-19 Pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020, doi: 10.1109/ACCESS.2020.3010287.

[17] T. Rahman *et al.*, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-rays images," *Comput. Biol. Med.*, vol. 132, no. November 2020, p. 104319, 2020, doi: 10.1016/j.compbiomed.2021.104319.

[18] P. Mensah Kwabena, B. A. Weyori, and A. Abra Mighty, "Exploring the performance of LBP-capsule networks with K-Means routing on complex images," *J. King Saud Univ. - Comput. Inf. Sci.*, pp. 1–15, 2020, doi: https://doi.org/10.1016/j.jksuci.2020.10.006.

[19] H. Ren and H. Lu, "Compositional Coding Capsule Network with K-Means Routing for Text Classification," *arXiv:1810.09177v3 [cs.LG]*, 2018.

[20] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[21] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Electr. Eng. - Part III Radio Commun. Eng.*, vol. 93, no. 26, pp. 429–441, 1946, doi: 10.1049/ji-3-2.1946.0074.

[22] P. K. Mensah, B. A. Weyori, and A. A. Mighty, "Max-Pooled Fast Learning Gabor Capsule Network," in *IEEE International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, 2020.

[23] P. K. Mensah, B. A. Weyori, and A. A. Mighty, "Gabor Capsule Network for Plant Disease Detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, pp. 388–395, 2020.

[24] K. Pogorelov *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," *Proc. 8th ACM Multimed. Syst. Conf. MMSys 2017*, pp. 164–169, 2017, doi: 10.1145/3083187.3083212.

[25] P. Mooney, "Retinal OCT Images (optical coherence tomography)," *Kaggle*, 2018. [Online]. Available: https://www.kaggle.com/paultimothymooney/kermany2018. [Accessed: 19-Apr-2021].

[26] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms- A Classification Perspective*, vol. 5, no. 1. New York: Cambridge University Press, 2011.

[27] F. Provost, T. Fawcett, and R. Kohavi, "The Case Against Accuracy Estimation for Comparing Induction Algorithms," in *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 445–453.

[28] B. Mandal, S. Ghosh, R. Sarkhel, N. Das, and M. Nasipuri, "Using dynamic routing to extract intermediate features for developing scalable capsule networks," in *2nd International Conference on Advanced Computational and Communication Paradigms, ICACCP 2019*, 2019, pp. 1–6, doi: 10.1109/ICACCP.2019.8883020.

[29] R. Meyes, M. Lu, C. W. De Puiseau, and T. Meisen, "Ablation Studies in Artificial Neural Networks," *arXiv:1901.08644v2 [cs.NE]*, pp. 1–19, 2019.

[30] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on," *Z MedPhys*, vol. 29, no. 2, pp. 102–127, 2019, doi: 10.1016/j.zemedi.2018.11.002.

[31] F. J. Valverde-Albacete and C. Peláez-Moreno, "100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox," *PLoS One*, vol. 9, no. 1, 2014, doi: 10.1371/journal.pone.0084217.

[32] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in *Proceedings of the 23 rd International Conference on Machine Learning*, 2006, pp. 546–559, doi: 10.4135/9780857021113.n29.

[33] P. Domingos and P. Singla, "Discriminative training of Markov logic networks," *Proc. 20th Natl. Conf. Arti cial Intelligene*, vol. 20, p. 868{873, 2005.

[34] C. X. Ling, J. Huang, and H. Zhang, "AUC : A Better Measure than Accuracy in Comparing Learning Algorithms," *Adv. Artif. Intell. Can. AI 2003. Lect. Notes Comput. Sci. (Lecture Notes Artif. Intell.*, vol. 2671, pp. 329–330, 2003, doi: https://doi.org/10.1007/3-540-44886-1_25.

[35] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond Accuracy , F-Score and ROC : A Family of Discriminant Measures for Performance Evaluation," *LNAI 4304*, pp. 1015–1021, 2006.

[36] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SQUEEZENET: ALEXNET-LEVEL ACCURACY WITH 50 X FEWER PARAMETERS AND < 0 . 5MB MODEL SIZE," *arXiv1602.07360v4 [cs.CV]*, pp. 1–13, 2017.

[37] C. Zhang, B. Recht, S. Bengio, M. Hardt, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.

[38] C. W. Wu, "ProdSumNet: reducing model parameters in deep neural networks via product-of-sums matrix decompositions," *arXiv:1809.02209v2 [cs.LG]*, no. 1, pp. 1–10, 2018.

[39] B. Ojeda-Magaina, R. Ruelas, M. A. Corona-Nakamura, and D. Andina, "An Improvement to the Possibilistic Fuzzy C-Means Clustering Algorithm," in *World Automation Congress (WAC)*, 2006.

[40] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, no. October 2019, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

APPENDIX A

TABLE A1. NUMBER OF IMAGES AND THE DIVISIONS PER 80% TRAINING, 20% TEST FOR EACH DATASET

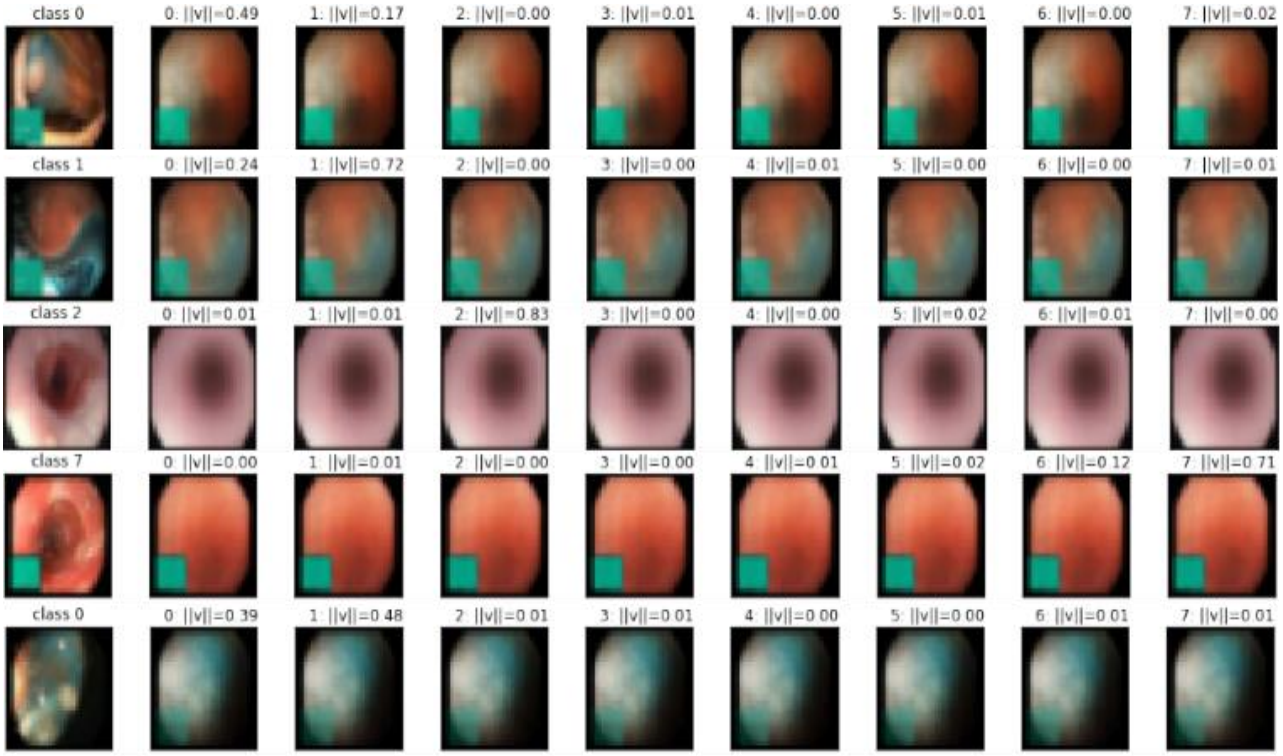
Dataset	Total Number of Images	Training Set	Validation Set	Test Set
KVASIR-2	8,095	6,476	1,619	100
COVID-19	82,570	66,056	16,514	100
ROCT	21,135	16,908	4,227	100

TABLE A2. COMPARISON OF MODEL ACCURACY TO THE TRADITIONAL CAPSNET. UNREPORTED VALUES ARE REPRESENTED BY (?)

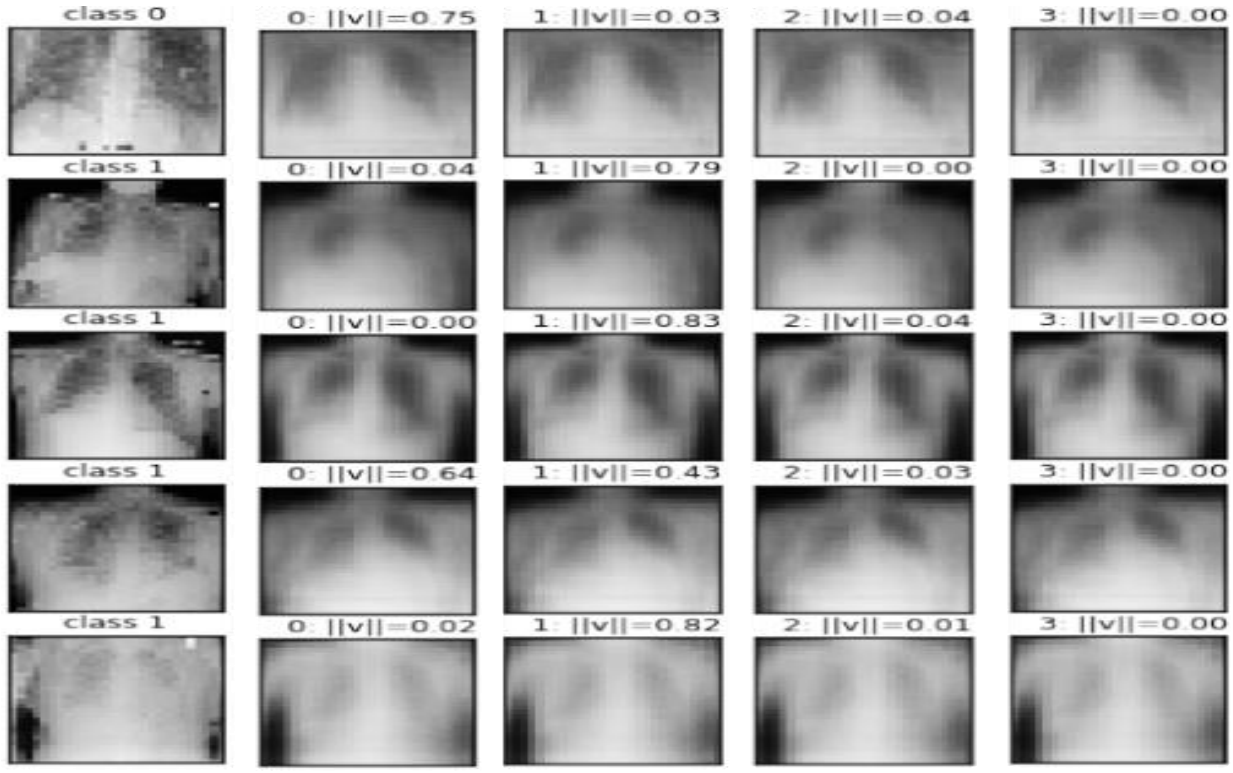
Model	KVASIR	COVID-19	ROCT
CNN [17]	?	91.30%	?
Transfer Learning [15]	?	?	93.40%
DR [7]	78.54%	65.15%	77.35%
GLC (ours)	**80.91%**	**91.96%**	**91.30%**

TABLE A3. COMPARISON OF MODEL PARAMETERS

Model	KVASIR		COVID-19		ROCT	
	Trainable	Non-Trainable	Trainable	Non-Trainable	Trainable	Non-Trainable
Traditional capsule (DR)	9,552,944	0	9,552,441	0	9,552,441	0
GLC (ours)	**8,323,640**	**0**	**8,302,136**	**0**	**8,302,136**	**0**
Difference	1,229,304	0	1,250,305	0	1,250,305	0



(a)



(b)

Fig. A1. The Reconstructed Images of the Proposed Model on (A) KVASIR and (B) COVID-19 Datasets.

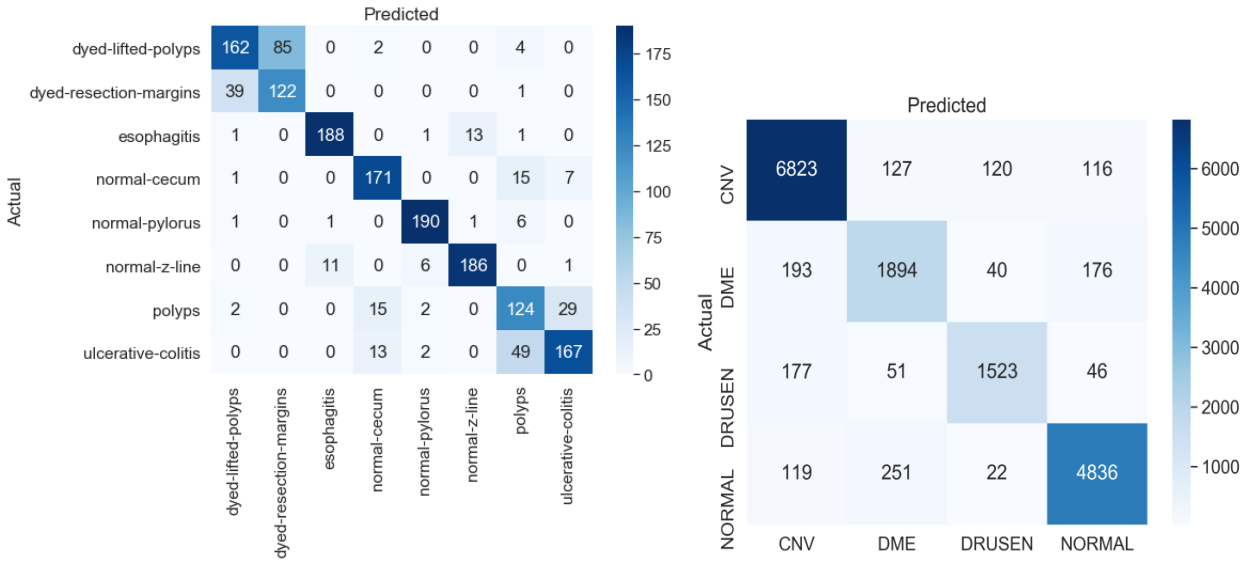


Fig. A2. Confusion Matrices of the Proposed Model on the KVASIR and COVID-19 Datasets.

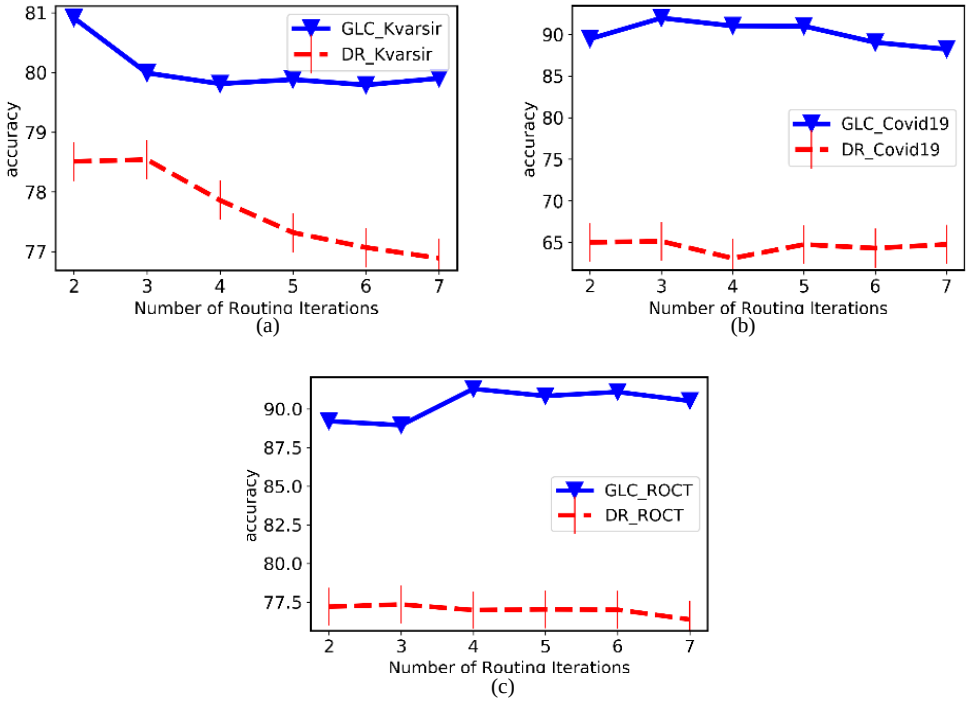


Fig. A3. Comparison of the Models' Ability to Scale on (a) KVASIR, (b) COVID-19, and (c) ROCT Datasets.