IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2021, accepted March 14, 2021, date of publication March 18, 2021, date of current version March 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3067195

Cervical Cancer Diagnosis Using Very Deep Networks Over Different Activation Functions

KHALED MABROUK AMER ADWEB[1], NADIRE CAVUS[1,3], AND BORAN SEKEROGLU[2], (Associate Member, IEEE)

[1]Department of Computer Information Systems, Near East University, 99138 Nicosia, Turkey
[2]Department of Information Systems Engineering, Near East University, 99138 Nicosia, Turkey
[3]Computer Information Systems Research and Technology Center, Near East University, 99138 Nicosia, Turkey

Corresponding author: Khaled Mabrouk Amer Adweb (adwebkhaled@gmail.com)

ABSTRACT Cancer prevention is mainly achieved by screening the transformation zones. Cervical pre-cancerous stages can be seen in three different types, and all can transform into cancer. Thus, it is crucial to intelligently screen cervical abnormality and have a robust system for detecting whether a cervix is in normal (healthy) or at a pre-cancerous stage. Deep learning showed great potentials when applied to biomedical problems, including medical image analysis, disease prediction, and image segmentation. Hence, in this paper, very deep residual learning based networks are designed in order to perform cervical cancer screening. Moreover, in this work, we highlight the importance of the activation functions on a residual network (ResNet)'s performance. Thus, three residual networks of the same structure are built with different activation functions. The employed models are trained and tested using a dataset of colposcopy cervical images, and the experimental results showed that designed residual networks with leaky and parametric rectified linear unit (Leaky-RELU and PRELU) activation functions performed almost equally in terms of accuracy where they reached accuracies of 90.2 and 100%, respectively. This achieved high accuracy was compared to other related works' results, and it showed an outperformance in screening the pre-cancerous and healthy colposcopy cervical images. Such an earlier and accurate diagnosis may help in preventing cervical cancer transformation.

INDEX TERMS Cervical cancer, residual learning, residual network, ResNet, activation functions.

I. INTRODUCTION

Cervical cancer develops in the cervix, the narrow entrance of the uterus. This cancer can mainly affect sexually active women aged between 30 and 45 [1]. It is estimated that 13800 of women will be diagnosed with cervical cancer in the United States [2] in 2020. The leading cause of this cancer is the human papillomavirus (HPV), i.e., a sexually transmitted virus [1]. Cervical cytology, or the so-called smear test, is the most common test used to detect cervical cancer. This test can help in detecting the cancer in its earlier stages, which helps in reducing the number of deaths [1]. This cancer can have five different stages and detecting it in its earlier stages improves the survival rate [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Charith Abhayaratne.

Over the past few decades, studies have been extensively conducted for the creation of computer-aided diagnosis (CAD) systems to help diagnosing diseases and analyzing medical images, and computer-assisted reading systems have been proposed for cervical cancer cell segmentation and classification [3]–[5]. These systems are based on automated image analysis algorithms and designed to select potential abnormal cells in a given cytology specimen to be classified. However, this task includes cytoplasm and nuclei segmentation, handcrafted features extraction, and cell classification. Cell segmentation is crucial in designing a CAD system; however, the large variations between normal and abnormal nuclei and the presence of cell clusters can form a significant obstacle to reaching an accurate segmentation of cells [3].

Moreover, engineered, and handcrafted features representing morphological characteristics of cells are costly and time inefficient. In addition to that, they may be error-prone as they

are limited to the understanding of the cervical cancer cytology and the finite imagination and visualization of human experts [6]. Overall, it has been noted that the current cervical cancer diagnosis systems that depend on handcrafted features are hindered by these aforementioned limitations, and others like ignoring some clue features and removing complementary information of the input image. Thus, there has been an urgent need for developing an automated technique for extracting different levels and discriminative features that can represent complex and distinguished information to describe cell abnormality. This technique of feature extraction has been developed in the deep learning networks, convolutional neural networks, which are biologically inspired structures, consist of powerful feature extraction tools [7]. These networks consist of several convolutions and pooling layers that act as feature extraction tools of effective capability in extracting features in a hierarchical way and at multiple levels, in a similar manner from the input space.

Recently, deep convolutional neural networks (DCNN) have shown remarkable performance in several medical tasks such as medical image analysis and understandings [4], [5], medical image classification [6], and cancer detection and identification [5], [6]. These networks have undergone significant upgrades and improvements over the years. This upgrade was mainly in terms of depth and performance, and it was found that the performance of CNNs varies proportionally to their corresponding depth, i.e., the number of hidden layers [4]–[6]. Hence, deep networks of various depths were proposed over the years, starting from the AlexNet [8] which consists of 8 layers, to VGGNet of 16 layers [9] and GoogleNet of 22 layers [10]. This increase in depth makes it difficult to train a network in terms of time, but the evolution of computers minimized this difficulty. On the other hand, going deeper and deeper, another problem has been encountered. It was noticed that training a very deep network (over than 18 layers) is associated with a vanishing gradient phenomenon due to the backpropagation algorithm used for training [11]. Thus, to suppress the issue, residual learning was proposed in 2016 [12]. This new concept of deep learning presented the shortcut connection or residual blocks in building the network. This concept is based on skipping connections between layers within the residual unit and an identity map of the input data instead of consecutive connections similar to AlexNet and GoogleNet.

Generally, all proposed residual networks (18, 50, 101, and 152 layers) use a non-saturating rectified linear unit (ReLU) as an activation function that allows complex relationships to be learned by the network [11], [12]. This function has been verified to a great outcome in many works [13]–[15]; however, researchers found that it can be associated with a problem called ''dying ReLU,'' i.e., it outputs zeros for the negative input values it receives [16]. This, therefore, badly affects the learning of the network. Hence, other activations networks were proposed in order to avoid the 'zeros' problem caused by ReLU and improve the learning process of the network. For such purposes, the Leaky rectified linear unit (Leaky-ReLU) [17] and Parametric rectified linear unit (PReLU) [17] were proposed in 2015 and 2016, respectively.

This paper addresses the cervical cancer screening using deep learning. The studies of diagnosing cervical cancer using deep learning are quite limited. Hence, there is a need for applying deep networks in detecting the pre-cancerous cervical colposcopy as this may prevent their transformation into cancer. In this context, a classification of colposcopy cervical images into pre-cancerous and healthy colposcopy images is carried out using a ResNet18 structure's inspired network. The network is built from scratch and trained on pre-cancerous and healthy images in order to detect patients that are more likely to develop cervical cancer. As this network is a newly designed and built, we attempted to investigate the best activation function that fits our model. Hence, we duplicated the same network's structure over three different activation functions (ReLU, Leaky-ReLU, and PReLU) and evaluated its performance. The developed residual network with ReLU activation function is denoted as ReLU-ResNet, and the network with Leaky-ReLU function is denoted as Leaky-ReLU-ResNet. The last is the residual network with Parametric ReLU function and denoted as PReLU-ResNet. Thus, in this paper, three networks with the same structure but with different activation functions were considered. As aforementioned, the developed networks are inspired by the ResNet18 architecture; however, more layers have been added in order to make them deeper, for optimal classification. The performance of the three designed networks is validated and tested on a cervical cancer diagnosis task, which is considered a challenging classification task due to the complexity and similarity of the pre-cancerous and healthy cervical colposcopy images.

The rest of the paper is organized as follows: Section 2 presents a deep insight into the residual learning, activations functions and dataset description. Section 3 discusses the problem formulation and the proposed models, in addition to the results discussion. Section 4 is a conclusion of the work.

II. MATERIALS AND METHODS

A. RESIDUAL LEARNING

Recently, deep networks started to go deeper, i.e., several hidden layers. This ''very'' depth seemed to be associated with some optimization difficulties during the learning process of networks. This problem is called the vanishing gradients, and it arises when a network is trained with a stochastic gradient descent algorithm [13], [17]. The idea is that as the networks go deeper, its calculated gradient of the loss function starts to decrease exponentially while it back propagates to initial layers. Hence, it may approach zero at some layers, and this makes the network hard to train. Small or zero gradient means that weights and biases may not be adequately tuned during every training pass, and this consequently leads to less convergence and high error value.

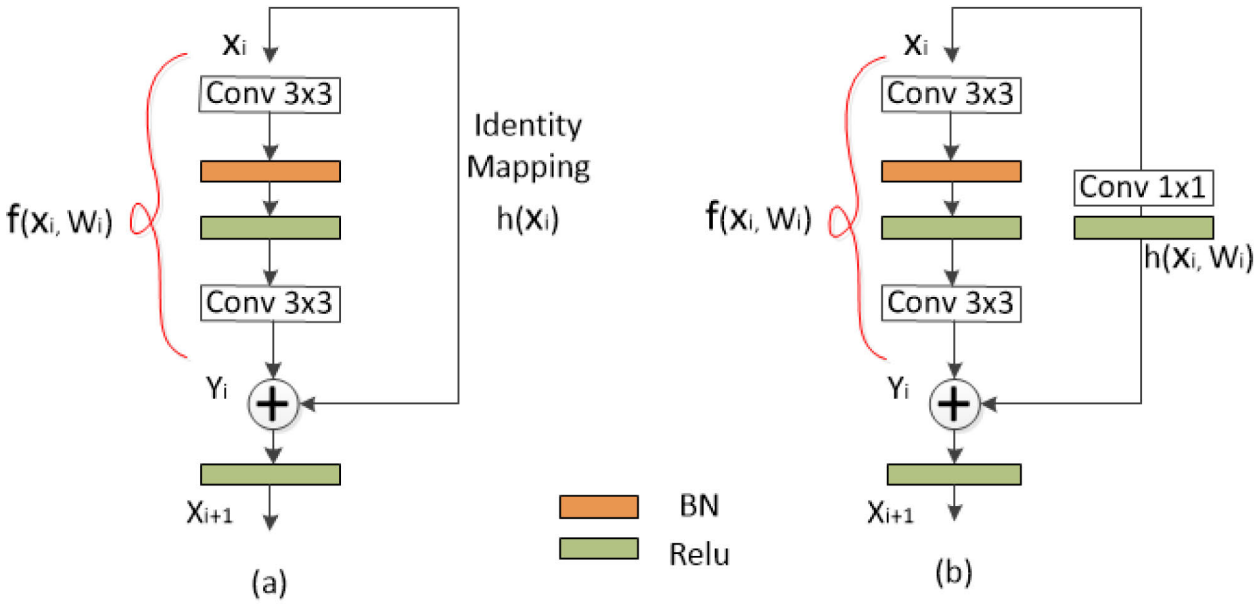


FIGURE 1. Residual units.

To overcome this issue, He *et al.* [12] proposed a residual learning concept designed for building and training very deep networks easily, without confronting the vanishing gradients problem. Residual networks (ResNets) introduced a new concept of connections called the skip or identity connections over some layers instead of connecting layers in a shallow way. This is based on residual units within the network, where the identity of input is mapped from the precedent hidden layers to higher ones. Hence, instead of learning the underlying mapping of input, the network is allowed to fit its residual mapping, as seen in Fig. 1(a).

There can be two types of residual mappings. The first one is the identity mapping-based shortcut connectivity (Fig. 1.a), in which the input encounters skip or short connection directly without passing through the convolution layer [12], [13]. The second residual block is the convolutional based identity mapping, i.e., short connection of input encounters a 1×1 convolution layer in the shortcut path (Fig. 1.b). Fig. 1 shows the residual units of a residual network structure.

The use of more identity mapping-based shortcut connectivity leads to better performance of the network in terms of computational complexity and training time; also, the convolution-based identity mapping results in a more straightforward propagation of information acquired during the forward and backward passes of the network learning [12].

Moreover, Fig. 1 shows that each residual unit contains two convolutional layers of 3×3 filter's size. These filters store the learnable parameters that are optimized during the training process of the whole network. In addition, each unit also has one batch normalization (BN) that normalizes the features map and one rectified linear unit (ReLU) layer that acts as an activation function for the network. The inputs and outputs are computed as follows [12]:

$$\boldsymbol{Y_i} = \boldsymbol{h}\,(\boldsymbol{X_i}) + \boldsymbol{f}\,(\boldsymbol{X_i} + \boldsymbol{W_i}) \tag{1}$$

$$\boldsymbol{X_{i+1}} = \boldsymbol{F}(\boldsymbol{Y_i}) \tag{2}$$

where $\mathrm{h}\,(\mathrm{X_i})$ represents the identity mapping of input $\mathrm{X_i}$, and F is the residual function. x, w, and y represent the input, weight, and output, respectively. Therefore, in the case of identity mapping-based shortcut connectivity, the identity mapping (Fig. 1.a) is defined as $\mathrm{h}\,(\mathrm{X_i}) = \mathrm{X_i}$. However, in the case of convolutional based identity mapping (Fig. 1.b), the identity mapping is defined as $\mathrm{h}\,(\mathrm{X_i}, \mathrm{W_i})$ and it is not the same as input as it is surpassed by a 1×1 convolution layer. In this case, input-output computations are defined as follows:

$$\boldsymbol{Y_i} = \boldsymbol{h}\,(\boldsymbol{X_i}, \boldsymbol{W_i}) + \boldsymbol{f}\,(\boldsymbol{X_i} + \boldsymbol{W_i}) \tag{3}$$

$$\boldsymbol{X_{i+1}} = \boldsymbol{F}(\boldsymbol{Y_i}) \tag{4}$$

where $h\,(X_i, W_i)$ represents the convolutional based identity mapping.

B. ACTIVATION FUNCTIONS

In this section, the three activation functions employed in this study, namely, ReLU, Leaky-ReLU, and PReLU, will be discussed in detail. Activation functions are the functions that set a specific output or "activation" for a given input, considering its weight. Preferred transform functions are the non-linear functions that can learn complex mapping functions [19]. Traditional and shallow neural networks with one or a very few hidden layers are used to perform well with the Sigmoid activation function. This non-linear activation function transforms the input values into a range of 0 and 1. Therefore, for any given positive or negative input, the output

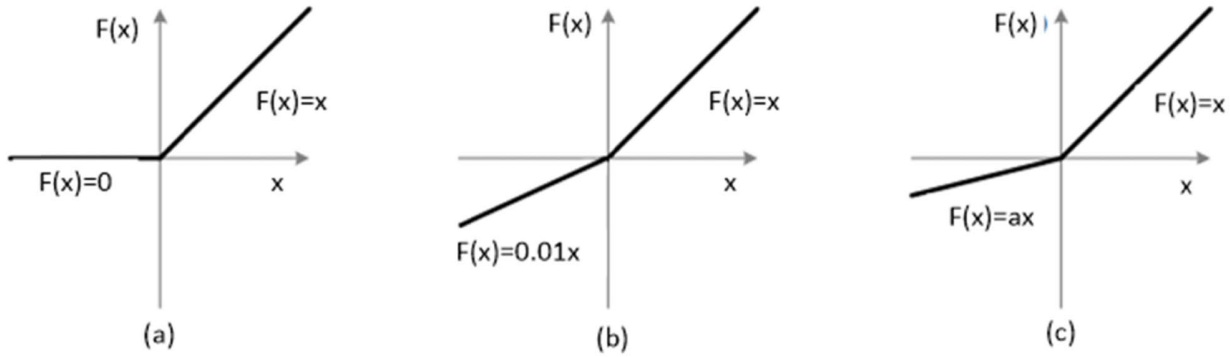


FIGURE 2. Activations functions: (a) ReLU, (b) LReLU, (c) PReLU.

is between 0 and 1. However, one major problem is encountered when using the Sigmoid function. The sigmoid function saturates with a given large positive or negative input; hence, the local gradient is minimal (almost zero). Therefore, during the backpropagation of error, this gradient is multiplied by that local gradient, and as a result, the gradient is vanished or killed. This is then called the vanishing gradient problem.

1) RECTIFIED LINEAR UNIT (ReLU)

A rectified linear unit (ReLU) was then proposed to overcome the problem caused by the Sigmoid activation function [20]. It has become very popular in the deep learning networks and leads to many breakthrough applications in different fields [8]–[10]. It is merely thresholded at zero input values, i.e., it zeros the negative input values, while keeping the positive input values, as seen in equations 5 and 6.

$$\boldsymbol{F(x) = max(0, X)} \tag{5}$$

Which has the gradient of

$$\boldsymbol{F'(x) = \left\{ \begin{matrix} 0 \; if \; x \leq 0 \\ x \; if \; x > 0 \end{matrix} \right\}} \tag{6}$$

where F is the ReLU activation output, and x is the input value.

ReLU boosted the performance of the deep convolutional neural networks as it helps networks to converge faster and not to be stuck in the vanishing gradient problem. On the other hand, ReLU can sometimes fall into "dying ReLU," where the input values of the ReLU neuron are stuck to negative values. In this case, ReLU output is always zero, and this deactivates the majority of the neurons. This, therefore, affects the learning of the network and results in a poor convergence and performance [21]. Setting the learning rate to minimal values can somehow solve this problem; however, other modified versions of ReLU were also proposed to avoid the zeros outputting of the negative input values.

2) LEAKY-ReLU

Xu *et al.* [22] proposed another activation function, namely, Leaky ReLU (Leaky-ReLU), to improve and modify the traditional ReLU and to solve more complex and non-linear functions. The main aim of Leaky-ReLU was to solve the problem associated with ReLU, i.e., "dying ReLU." Thus, this method suggests a slight leak of information in the part where output is always 0. This means that the gradient will be small but not zero; therefore, neurons won't be inactive as the weights will be adjusted. This function proposes that in the case of negative input, the corresponding output will be the input multiplied by a small number (0.01) in order not to shut off the neurons (Equation 7 and 8).

$$\boldsymbol{F(x) = \left\{ \begin{matrix} \alpha x \; if \; x \leq 0 \\ x \; if \; x > 0 \end{matrix} \right\}} \tag{7}$$

Which has the gradient of

$$\boldsymbol{F'(x) = \left\{ \begin{matrix} \alpha \; if \; x \leq 0 \\ 1 \; if \; x > 0 \end{matrix} \right\}} \tag{8}$$

3) PRELU

Applying Leaky-ReLU on CNNs [19], [22] showed that this function could have some benefits over ReLU, such as solving the vanishing gradient problem by avoiding the zero gradient's part and speeding up the network's learning and convergence. On the other hand, it was noticed by some studies [20] that Leaky-ReLU's impact on the network's accuracy is negligible; therefore, a new modified version of Leaky-ReLU was proposed by He *et al.* [17]. This new method is called a parametric rectified linear unit (PReLU), and it is similar to Leaky-ReLU. However, instead of having a predefined slope value of 0.01 for negative inputs, PReLU allows this coefficient to be learnable by the network, during training just like weights and biases. The output of this function is:

$$\boldsymbol{F(x) = \left\{ \begin{matrix} \alpha x \; if \; x \leq 0 \\ x \; if \; x > 0 \end{matrix} \right\}} \tag{9}$$

This function has the gradient of

$$\boldsymbol{F'(x) = \left\{ \begin{matrix} x \; if \; x \leq 0 \\ 0 \; if \; x > 0 \end{matrix} \right\}} \tag{10}$$

where F is the PReLU activation function, x is the input value, and α is the learnable coefficient. Fig. 2 shows the

TABLE 1. Original datasets.

	Healthy	Pre-cancerous
Dataset 1	-	4000
Dataset 2	800	-

three different activation functions. It is seen that PReLU can include them all, depending on the coefficient value, which is learnable in this case:

If $\alpha = 0$; F becomes ReLU

If $\alpha > 0$; F becomes Leaky-ReLU

If α is a learnable parameter, F becomes PReLU

In this paper, these three activation functions are applied in a residual network built for the diagnosis of cervical cancer. The network's performance is compared over the three aforementioned functions in terms of accuracy in order to select the optimal function for such residual network's classification task.

C. EXPERIMENTAL ANALYSIS

Dataset: Three residual networks of 18 layers are built, each with different activation functions. The networks are trained and validated with cervical pre-cancerous datasets that contain colposcopy images of healthy and pre-cancerous cervixes. The pre-cancerous stage cervical images are collected from the Intel and MobileODT Cervical Cancer Screening, which is a competition launched on Kaggle [18] to classify three types of cervical pre-cancer images (earlier stages of cervical abnormalities that may transform into cancer). This dataset provides three types of pre-cancer cervical images. In this study, all three types are considered as one class, which is pre-cancerous since our aim is to classify pre-cancerous or healthy cervixes. Normal or healthy images are collected from the Tripoli Hospital center, Libya [19]. Table 1 shows the number of healthy and pre-cancerous cervical images collected from both datasets. As it is seen in Table 1, 4,000 pre-cancerous images are collected from the first dataset, and 800 healthy images are collected from dataset 2. The learning scheme used for training and testing the network is 50:10:40, i.e., 50 % of the images are used for training the network, 10% for validation, and the remaining are used for testing purposes.

As noticed, the number of abnormal cervical images number is $\sim$ three times more than normal images in the created dataset. Hence, our classifier may tend to exhibit a bias towards the majority class (abnormal cervices), which is practically not desirable. Therefore, balancing the proportions of both classes is considered a common solution for this issue. Data augmentation is then applied only to healthy images in order to increase the number of images to $\sim$ 4000. Hence, shift translation and scale invariance are employed to balance the classes' proportions and to provide the network with the power of detecting the condition of disease at different angles and shifts. Therefore, the 800 original healthy cervical images are rotated at angle 90° and 180° and randomly translated up to two pixels horizontally and vertically. In total, a dataset of 3920 healthy cervical images is formed and shown in Table 2. As for preprocessing, our collected and augmented images were first converted to three-channel PNG format. Moreover, images were normalized to scale their pixel values to the range of 0 to 1, and resized to 224 $\times$ 224 pixels to fit the employed models' inputs.

TABLE 2. Datasets augmentation and learning scheme.

	Healthy	Pre-cancerous
Dataset 1	-	4000
Dataset 2	3920	-
Total	7920	
Train	2352	2400
Test	1568	1600

Fig. 3 shows a sample of images used for training the networks. The first row shows the pre-cancerous cervical images, which are three types, while the second row shows the healthy cervixes. As mentioned above, our aim in this work is to create a deep network from scratch and train it to classify cervical pre-cancerous and healthy images in order to diagnose cervical cancer in its earlier stages; so that it can be treated before transforming into cancer. Thus, we adopted the ResNet18's structure to build our network. Moreover, we investigated the performance of this model using three types of activations in order to discover which function can cause a slight impact on the network's accuracy using a specific dataset.

1) DEEP RESIDUAL NETWORK-BASED CERVICAL DIAGNOSIS

As outlined, three deep residual networks with three different activation functions are designed to diagnose cervical cancer. Our network architecture is inspired by the ResNet18 structure and is comprised of 50 layers with four residual and convolution blocks, as shown in Fig. 4. Each residual and convolution block contains 1 $\times$ 1 and 3 $\times$ 3 convolution operations (Conv) and rectified linear unit functions (ReLU). Moreover, batch normalization (BN) and dropout layers are also used for building more feasible designs and improving model generalization. Note that some layers are not shown in the figure due to simplicity.

The whole output of the residual blocks of all these layers undergoes a 3 $\times$ 3 max-pooling operation and then sent to the global averaging pooling layer [25]. This layer replaces the fully connected layer, and it converts every feature map into a value. The computed values are then fed into a Sigmoid function as our task is a binary classification task. Note that several studies have found that global average pooling reduces overfitting as, in this layer, there is no parameter to be optimized [12], [13], [25]. Moreover, this approach is more

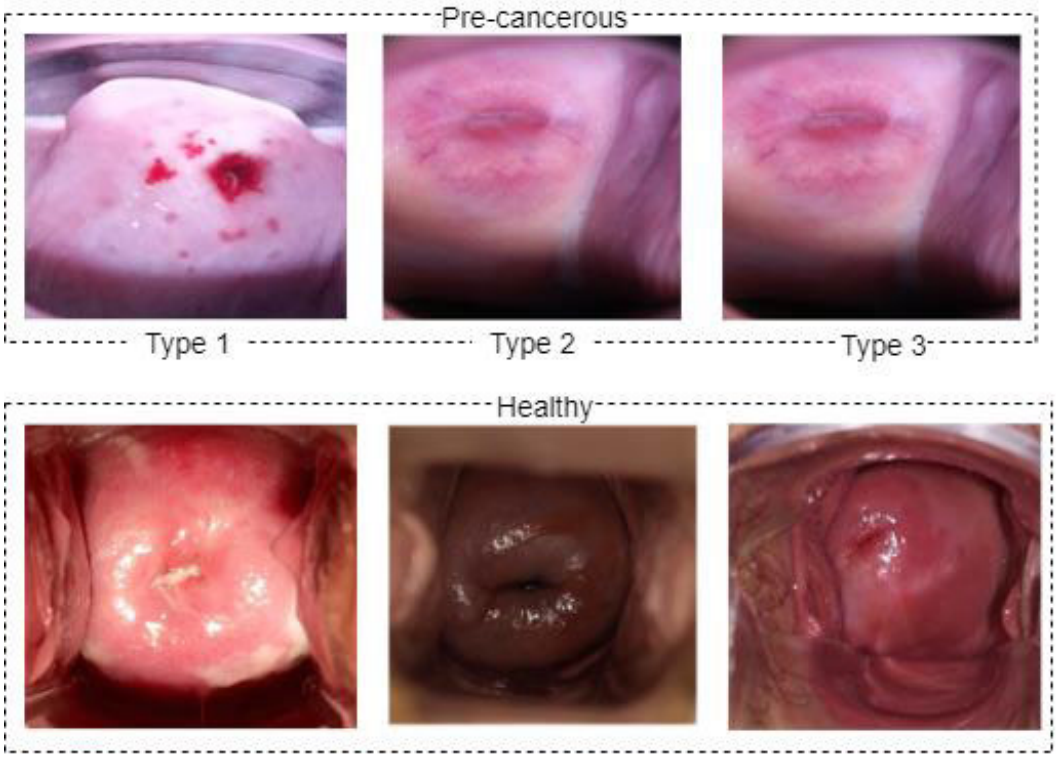


FIGURE 3. Cervical images. First row shows the pre-cancerous cervical images, and second row shows the healthy cervixes.

robust and efficient to spatial translation, and the shift of the input for it sums the spatial information of the input data.

This research addresses the activation functions of a convolutional network on a particular medical classification task; hence three ResNets are built, however, each with different activation functions. Fig. 4 shows only the ReLU based ResNet architecture; however, two other similar networks are also built with the same architecture but with different activation functions, one with Leaky-ReLU and the other with PReLU. Note that these functions are employed in the residual blocks, whereas a Sigmoid function is used for the fully connected layer since our cervical classification task is considered a binary task, i.e., pre-cancerous or healthy cervices.

2) TRAINING

As mentioned above, networks are trained using the same healthy and pre-cancerous images. A learning scheme of 60:40 is used to train and test the three developed models. This means that 60% (4752 images) of the data are used for training, and 40% are used as a held-out test (3168). The 10% of the train data was used to validate the networks to fine-tune their hyperparameters and reduce overfitting.

All employed ResNets were trained on the training dataset with a learning rate of 1e−4, 0–1 input image normalization, cross entropy loss function, growth rate of 12, block depth of 6, for 20 epochs on a MATLAB 2020a.

This selection of hyperparameters was based on the grid-hyperparameters search, inspired by the work of Huang et al [31].

The networks were simulated using Windows 64-bit desktop computer with an Intel Core i7 4770 Graphical processing unit (GPU) and 8 GB random access memory.

Networks are evaluated by calculating their training and testing accuracy using the following formula:

$$\textbf{Accuracy} = \frac{N}{T} \tag{11}$$

where N denotes the total number of correctly classified images, and T represents the total number of images. The loss function used in this work is a binary cross-entropy as cervical cancer screening is considered a binary classification problem in this study.

Note that more evaluations metrics such as sensitivity, specificity, and area under curve (AUC) are calculated as follows:

$$\textbf{Sensitivity} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} \tag{12}$$

$$\textbf{Specificity} = \frac{\textbf{TN}}{\textbf{TN} + \textbf{FP}} \tag{13}$$

where TP stands for true positive, and it indicates the number of correctly predicted positive classes. TN stands for true negative, and it indicates the number of correctly predicted negative classes. FP is the false positive, and it shows the number of incorrectly predicted negatives as positives by the

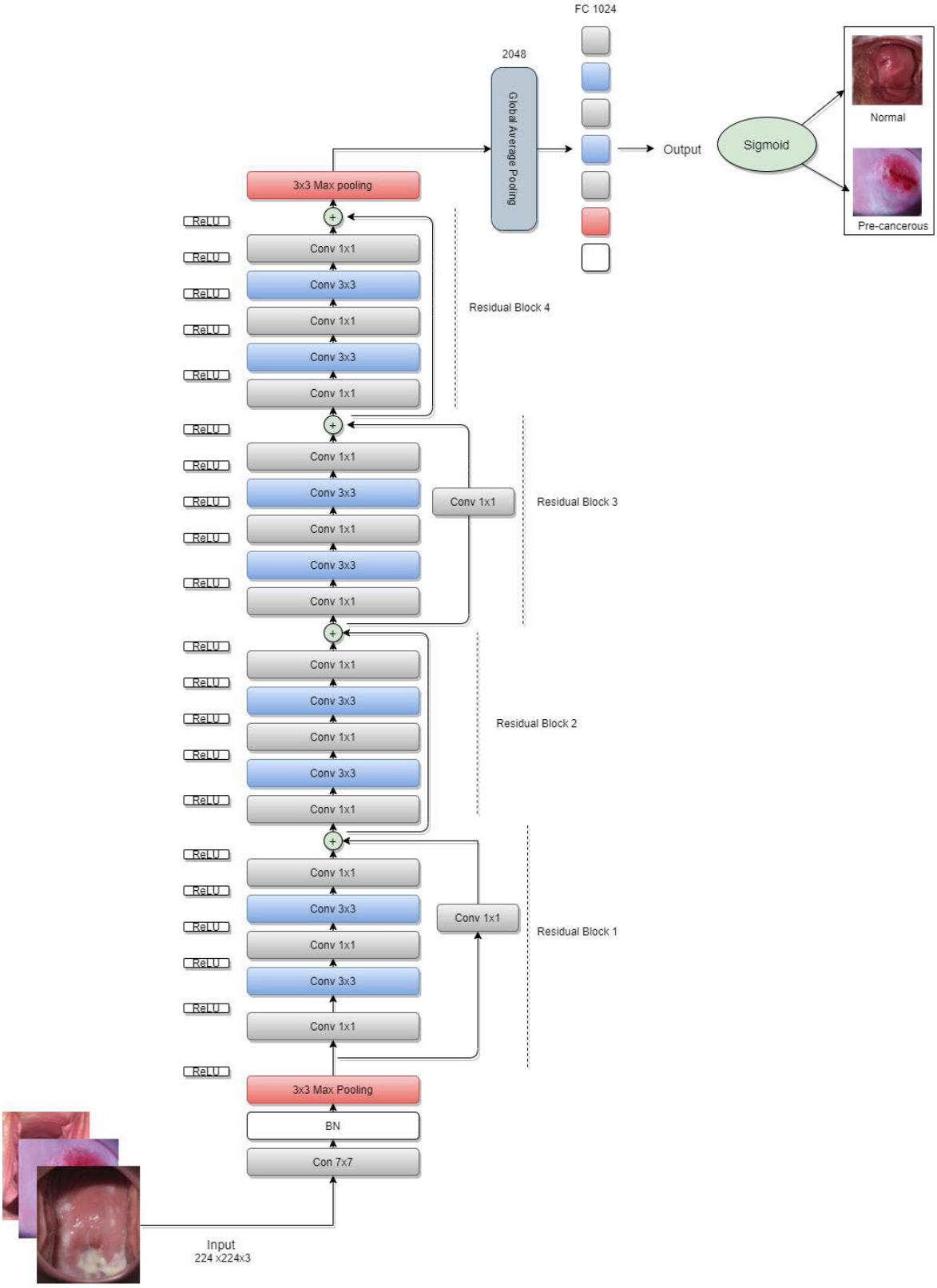


FIGURE 4. The developed ResNet18 architecture (with the ReLU activation function).

TABLE 3. Models learning parameters.

Learning parameters	ReLU-ResNet Values	Leaky-ReLU-ResNet Values	PReLU-ResNet Values
Training ratio	60%	60%	60%
Learning rate	0.0001	0.0001	0.0001
Max. number of epochs	30	30	30
Optimum epochs	20	20	20
Training accuracy	100%	100%	100%
Training time (Minutes)	125	119	122
Achieved mean square error (MSE)	0.0074	0.0014	0.0001

model, while FN is the false-negative, indicating the number of incorrectly predicted positive data as negative. AUC is the area under the Receiver Operating Characteristic curve (ROC), which is a graph that shows the performance of the network at thresholds. ROC plots the True positive rate versus the false positive rate.

Table 3 shows the setting of the learning parameters of the three different networks. A Stochastic gradient descent method is used to train the networks with a minibatch size of 64 images for every iteration to optimize the proposed models. Moreover, the learning rate and reducing factor of the fully connected layers of the three networks are set to 0.001 and 0.1, respectively. Epoch's number is selected based on the variations of the accuracy and validation error. In other words, the number of epochs is selected based on monitoring the validation error and performance of the networks as it can be associated with optimization problems, i.e., if it is high, overfitting may occur. Hence, a maximum number of 30 epochs is selected, and networks can stop learning whenever validation error starts to increase, and accuracy starts to saturate (Early stopping). Thus, we draw the learning curves of the three networks (Fig. 5). These curves show the variations of training classification accuracy against the increase of epochs. It is seen that the three networks reached their highest accuracies at epoch 20 and then started to saturate. Therefore, learning stops at epoch 20, the optimum epoch. It is also noted that the networks required approximately the same time to reach such maximum accuracies, which is 125 minutes, 119 minutes, and 122 minutes for ReLU-ResNet, Leaky-ReLU, and PRELU, respectively. This similarity, in time and optimum epochs, is because all models share the same structure with different activation functions. In terms of errors, PReLU_ResNet reached the lowest mean square error compared to other models.

Table 4 shows the loss values at the end of every residual bock of every employed network being trained, at epoch 2.

III. RESULTS AND DISCUSSION

In order to evaluate the feasibility of the three ResNet models in classifying cervical cancer, we conducted an experimental test of 40% of the remaining cervical data that were not seen before by the networks. Table 4 shows some testing results, such as classification accuracies and evaluation metrics.

As seen in Table 4, the ReLU-ResNet achieved the lowest classification accuracy, sensitivity, specificity, and area under curve (AUC) of 98.3%, 91.2%, 96.2%, and 96.9%, respectively. This may be due to its ReLU activation function, which, as discussed in Section 2, causes 'dying ReLU' during training that leads to a poor performance compared to other employed activations like Leaky-ReLU and PReLU. It is noticed that Leaky-ReLU and PReLU helped in improving the generalization capability of the networks by making the gradient small but not zero, which activates neurons and allows weights to be adjusted. Fig. 6 shows the confusion matrix and ROC graph of the PReLU-ResNet.

Fig. 7 shows the learned activations of the network that outperformed all other networks that is PReLU-ResNet. It shows the learned activations at the first convolution and max-pooling layers. It is seen that the network learned gradients and levels of abstractions in these two layers, but those features are kept to color, orientations, and edges since learning is still in the first layers.

TABLE 4. Loss values at epoch 2.

	ReLU-ResNet	Leaky-ReLU ResNet	PReLU-ResNet
Block 1	0.642	0.532	0.324
Block 2	0.413	0.246	0.259
Block 3	0.536	0.262	0.128
Block 4	0.125	0.103	0.052

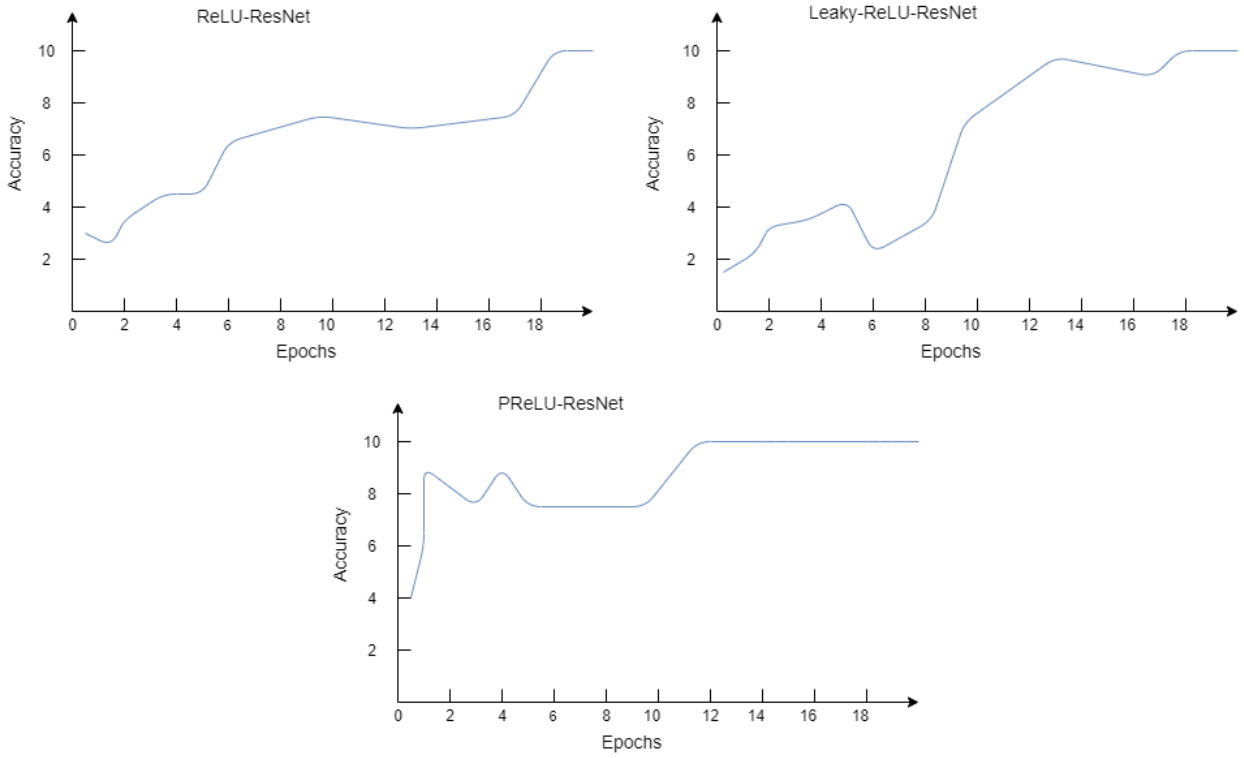


FIGURE 5. Learning curves and Loss for ResNet-50.

Nevertheless, once we go deeper (Fig. 7), networks seem to learn more complex and meaningful features such as objects and complete parts of cervices; as in deeper layers, learned features are built up by combining features from previous layers. In Fig. 8, we opt to show the learned features at a deeper layer and the layer of activations of each network. Thus, we choose the seventh activation function layer of each network and visualize its leaned features. As seen in the figure, the network with ReLU activation function (Fig. 8.a) appears to learn no meaningful features when compared to other networks of leaky-ReLU (Fig. 8.c) and PReLU functions (Fig. 8.b), which seem to learn more abstract and complicated features. This, therefore, affects the learning performance of the networks because, as seen in Fig. 7, ReLU-ResNet seems to be incapable of extracting the right features of cervical images that help in classifying whether the cervix is healthy or pre-cancerous. This is most likely the reason why this network achieved a lower accuracy and higher error than other networks.

Fig. 9 shows examples of some pre-cancerous (abnormal) cervical images which are misclassified by ReLU-ResNet as normal and correctly classified as pre-cancerous by Leaky-ReLU-ResNet and PReLU-ResNet. Fig. 10 shows samples of some misclassified healthy (normal) cervical images by ReLU-ResNet.

For more interpretability of what the network relies on to make a final decision concerning the class of an image, we visualize the Gradient Weight Class Activation Mapping (Grad-Cam) of some pre-cancerous images. Grad-Cam is a method for visualizing the highest activations' regions that were the reason for the network's final decision. The suspected regions associated with the predicted class are highlighted by heatmaps, in which a jet colormap shows the highest activation regions as deep red and the lowest activation regions as deep blue. Figure 10 shows the Grad-Cam of some testing pre-cancerous colposcopy images that were correctly classified by PReLU-ResNet.

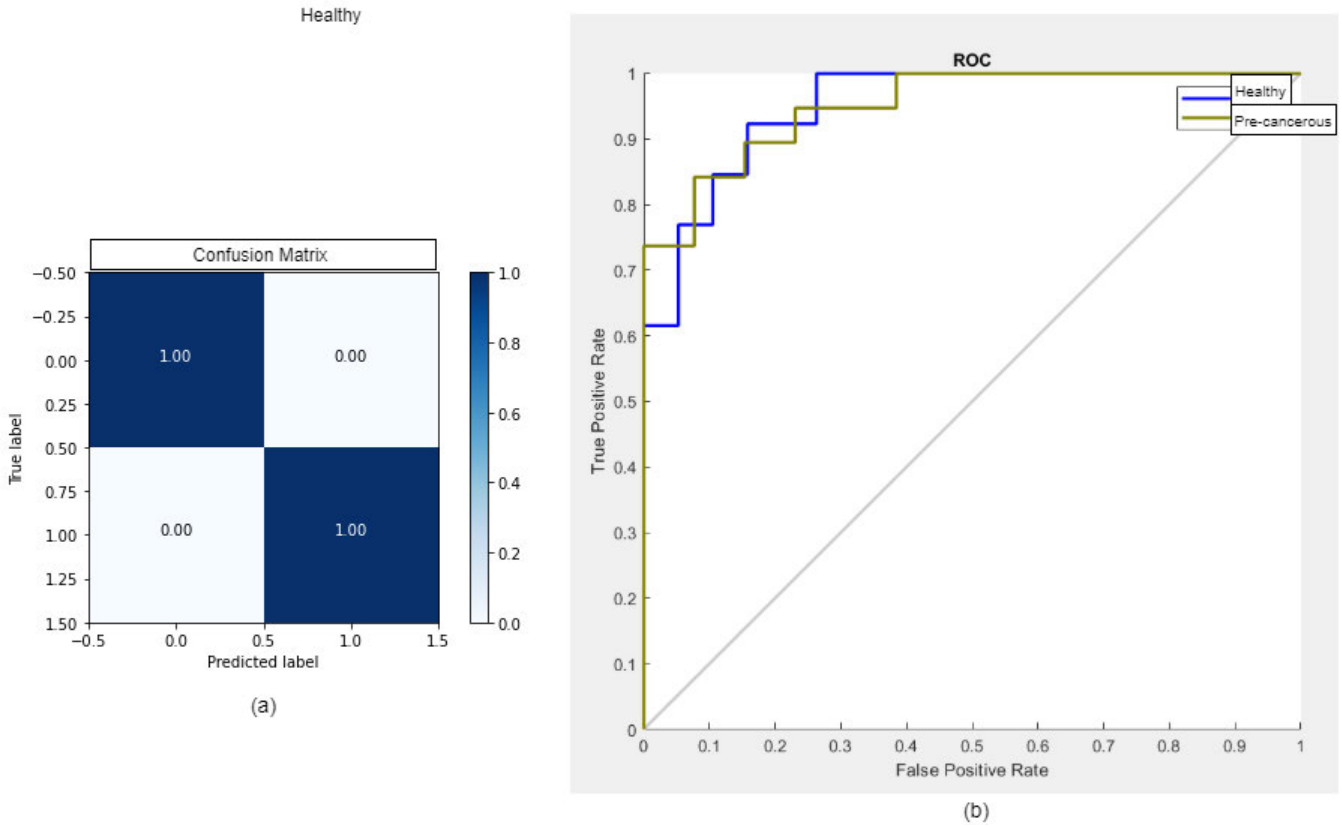


FIGURE 6. Confusion matrix and ROC of the PRELU-ResNet.

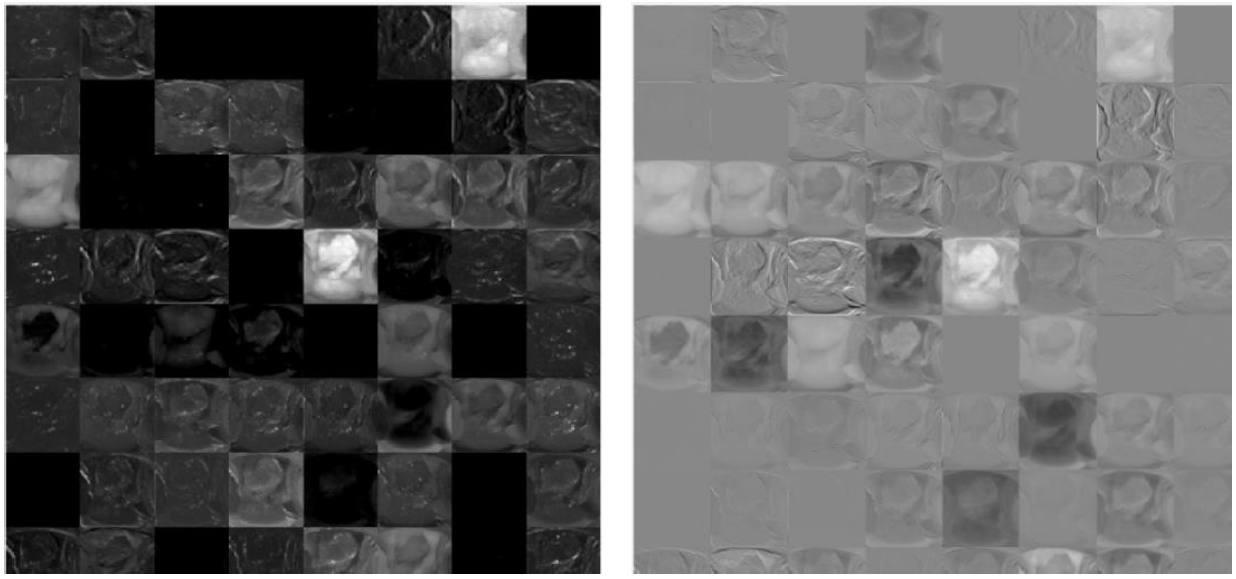


FIGURE 7. Learned features of PReLU-ResNet. (a) Pooling layer 1. (b) Convolution layer 1.

A. MODELS COMPARISON WITH EARLIER WORKS

Several studies have been conducted to diagnose cervical cancer [5], [14], [15], [26], [27]; however, colposcopy based cervical images classification using deep learning is quite limited. Generally, most of the published studies used microscopic data as inputs for their systems [5], [25]. These studies produced significant results, but they cannot be compared with ours as their database used for training and testing the

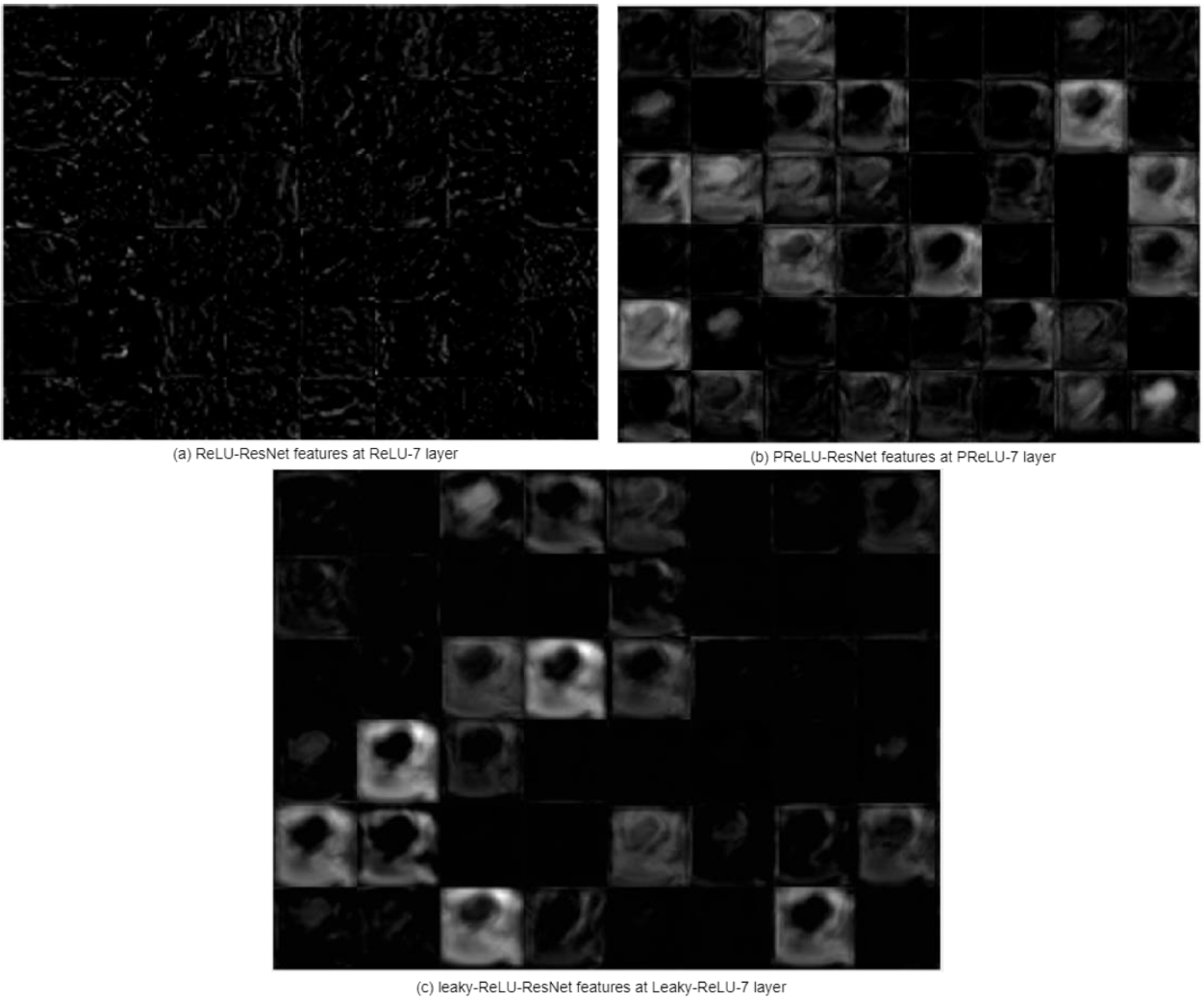


FIGURE 8. Learned features at activation function layers of all networks.

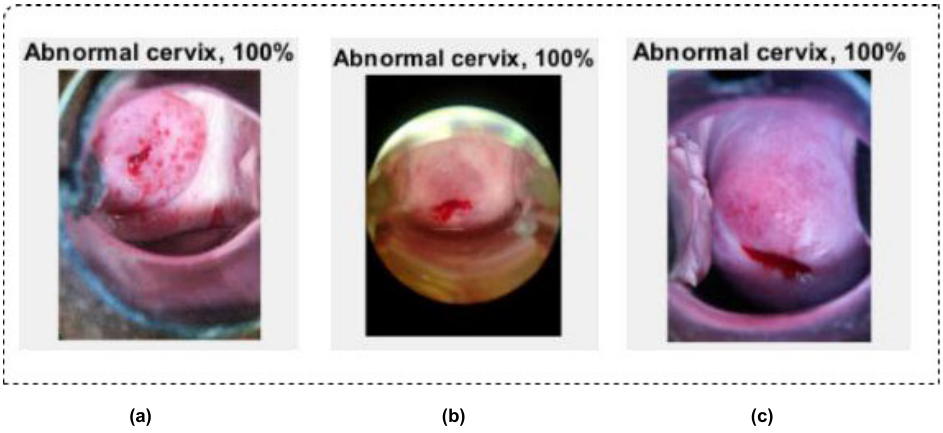


FIGURE 9. Examples of correctly classified abnormal cervical images by Leaky-ReLU-ResNet and PReLU-ResNet.

network's performances is of different type. On the other hand, some studies have used the same Kaggle database used in our study but for different purposes, i.e., to classify the three types of cervical cancer [14] or segment the uterine

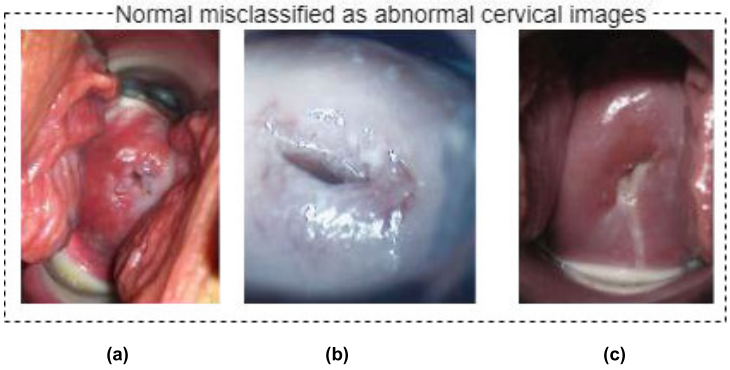


FIGURE 10. Examples of misclassified healthy cervical images by ReLU-ResNet.

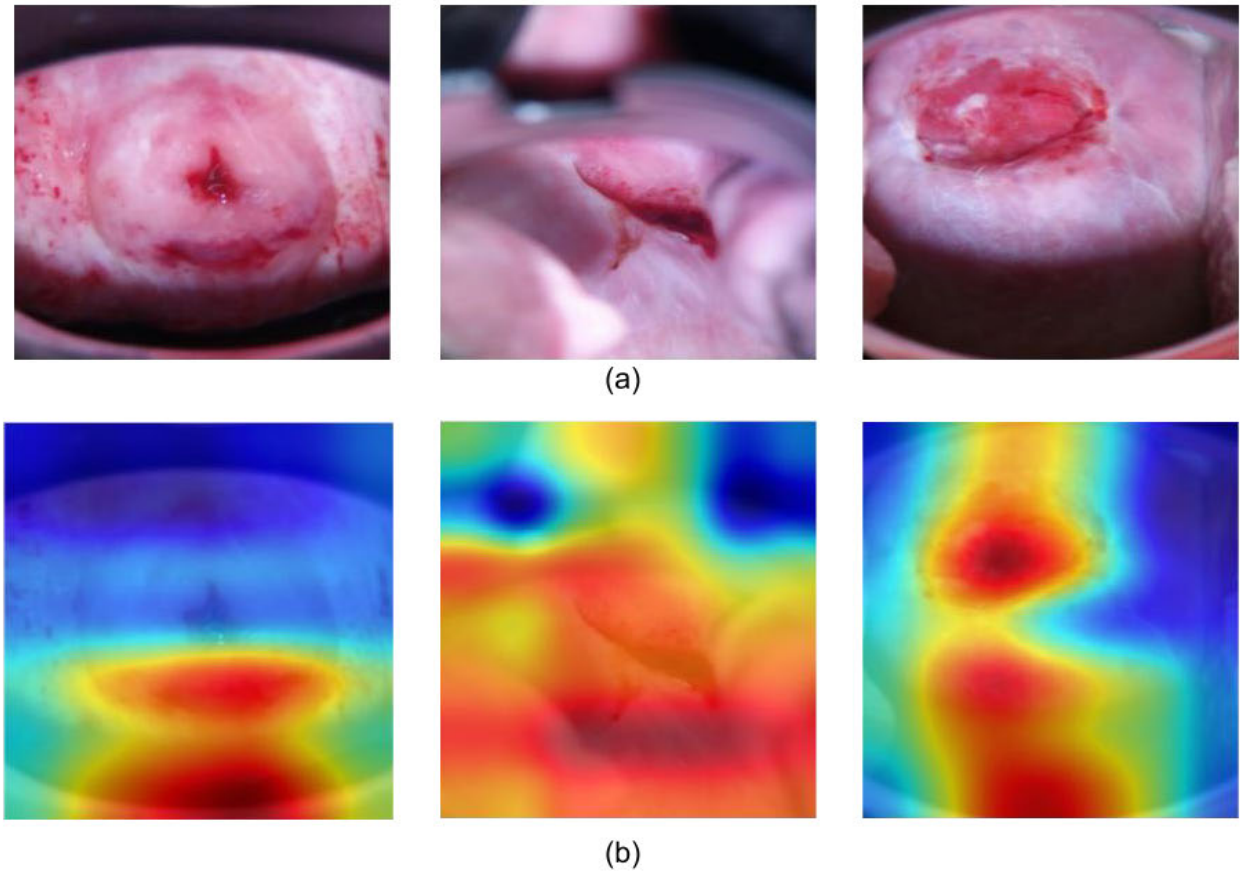


FIGURE 11. Grad-Cam of some pre-cancerous colposcopy images and their corresponding original images of the PReLU-ResNet.

cervix [28]. Thus, the two studies that used the same dataset and task were considered here to perform a robust comparison. Mustafa and Dauda [29] discussed the classification of colposcopy cervical images into cancerous or healthy. Their work utilized three different deep convolutional neural networks (DCNNs), each with different optimizers such as stochastic gradient descent (SGD), Root Mean Square Propagation (RMSprop), and Adaptive Moment Estimation (Adam). Those networks were all trained and tested using cancerous and healthy cervical images, and the result was compared to select the best optimizer. As the authors claimed, the Adam based CNN achieved the highest accuracy (90%) of classifying normal and cancerous cervices.

In another study, Yuan *et al.* [30] discussed the recognition of cervical squamous intraepithelial lesions recognition in colposcopy images. This study presented different applications using different neural networks, such as the U-Net model for segmenting the lesion in the cervix and the ResNet

TABLE 5. Learning parameters setting.

Learning parameters	ReLU-ResNet Values	Leaky-ReLU-ResNet Values	PReLU-ResNet Values
Testing ratio	40%	40%	40%
Accuracy	98.3%	99.2%	100%
Sensitivity	91.2%	95.6%	97.8%
Specificity	96.2%	96.3%	98.1%
AUC	93.4%	94.2%	96.9%

TABLE 6. Performance comparison with other related works.

	ReLU-ResNet	Leaky-ReLU-ResNet	PReLU-ResNet	Adam based CNN [29]	ResNet [30]
Maximum number of epochs	30	30	30	50	_
Accuracy	98.3%	99.2%	100%	90%	84.10%

model for classifying positive and negative colposcopy cervical cancer images. However, in this comparison, we are only interested in their ResNet classification model as it shares similar application with our models. Yuan *et al.* concluded that their ResNet model achieved an accuracy of 84.10% in differentiating positive colposcopy cervical cancer images from the negative ones. Table 5 shows a comparison of the proposed networks discussed in this paper with the aforementioned related works.

It can be seen from Table 5 that the developed residual learning based networks outperform the plain and traditional Adam convolution neural network [25] in terms of accuracy. It is concluded, by such comparison, that the residual learning approach can provide a great positive impact on a CNN as it boosts its generalization capability.

IV. CONCLUSION

In this paper, a very deep network is developed for the purpose of diagnosing cervical cancer using colposcopy images. The designed network is a residual learning-based network (ResNet) inspired by the ResNet18 architecture. Activations function drawbacks are also discussed in this work, and three different activation functions are employed to investigate the impact of activation function on the ResNet's performance. Hence, three networks were designed with three different networks. All networks were trained and tested using a dataset of raw cervical images.

The experimental results showed that all designed networks achieved high-accuracy generalization of cervical cancer diagnosis. However, the ReLU activation function network produced lower accuracy than other networks with the Leaky-ReLU and PReLU activation functions. Moreover, this network, ReLU-ResNet, also showed a weakness in extracting the rightful and complex features that distinguish the two cervical classes, leading to a poor generalization.

Finally, it was proved that Leaky-ReLU and PReLU activation functions could serve as an important booster for a designed residual learning-based network's performance. This performance improvement is mainly due to their ability to solve the 'dying ReLU' problem associated with the ReLU activation function.

Overall, this research shows that a deep learning system can achieve complex medical classification tasks like cervical screening. Such systems can help medical experts in the diagnosis of this disease and finding its pre-shape before its transformation into cervical cancer.

Furthermore, this system can be extended to diagnose the three different types of cervical cancer. Such a developed system could be very significant as it can stage cervical cancer, which helps develop the treatment plan tailored to patients with respect to their specific cancer type.

REFERENCES

[1] CDC. *Basic Information About Cervical Cancer*. Accessed: Apr. 21, 2020. [Online]. Available: https://www.cdc.gov/cancer/cervical/basic_info/index.htm

[2] American Society of Clinical Oncology (ASCO). (Jan. 2020). *Cancer Net Editorial Board*. [Online]. Available: https://www.cancer.net/cancer-types/cervical-cancer/statistics

[3] G. G. Birdsong, "Automated screening of cervical cytology specimens," *Hum. Pathol.*, vol. 27, no. 5, pp. 468–481, May 1996.

[4] M. Li, S. Dong, Z. Gao, C. Feng, H. Xiong, W. Zheng, D. Ghista, H. Zhang, and V. H. C. de Albuquerque, "Unified model for interpreting multi-view echocardiographic sequences without temporal information," *Appl. Soft Comput.*, vol. 88, Mar. 2020, Art. no. 106049.

[5] E. Bengtsson and P. Malm, "Screening for cervical cancer using automated analysis of PAP-smears," *Comput. Math. Methods Med.*, vol. 2014, Mar. 014, Art. no. 842037.

[6] A. Helwan, G. El-Fakhri, H. Sasani, and D. U. Ozsahin, "Deep networks in identifying CT brain hemorrhage," *J. Intell. Fuzzy Syst.*, vol. 35, no. 2, pp. 2215–2228, Aug. 2018.

[7] L. Zhang, H. Kong, C. Ting Chin, S. Liu, X. Fan, T. Wang, and S. Chen, "Automation-assisted cervical cancer screening in manual liquid-based cytology with hematoxylin and eosin staining," *Cytometry Part A*, vol. 85, no. 3, pp. 214–230, Mar. 2014.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[11] O. K. Oyedotun, A. E. R. Shabayek, D. Aouada, and B. Ottersten, "Improving the capacity of very deep networks with maxout units," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2971–2975.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual network," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 630–645.

[13] O. K. Oyedotun, A. E. R. Shabayek, D. Aouada, and B. Ottersten, "Training very deep networks via residual learning with stochastic input shortcut connections," in *Neural Information Processing*. Cham, Switzerland: Springer, 2017, pp. 23–33, doi: 10.1007/978-3-319-70096-0_3.

[14] R. Gorantla, R. K. Singh, R. Pandey, and M. Jain, "Cervical cancer diagnosis using cervixnet—A deep learning approach," in *Proc. IEEE 19th Int. Conf. Bioinf. Bioeng. (BIBE)*, Athens, Greece, Oct. 2019, pp. 397–404.

[15] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," *Future Gener. Comput. Syst.*, vol. 106, pp. 199–205, May 2020.

[16] L. Lu, Y. Shin, Y. Su, and G. Em Karniadakis, "Dying ReLU and initialization: Theory and numerical examples," 2019, *arXiv:1903.06733*. [Online]. Available: http://arxiv.org/abs/1903.06733

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1026–1034.

[18] *Intel & MobileODT Cervical Cancer Screening*. Accessed: Mar. 13, 2020. [Online]. Available: https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data

[19] *Tripoli Hospital Centre*. Accessed: May 5, 2020. [Online]. Available: https://apps.allianzworldwidecare.com/poi/hospital-doctor-and-health-practitioner-finder?PROVTYPE=HOSPITALS&TRANS=Hospitals%20in%20Tripoli,%20Libya&CON=Africa&COUNTRY=Libya&CITY=Tripoli&choice=en

[20] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*. [Online]. Available: http://arxiv.org/abs/1511.07289

[21] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018, *arXiv:1811.03378*. [Online]. Available: http://arxiv.org/abs/1811.03378

[22] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*. [Online]. Available: http://arxiv.org/abs/1505.00853

[23] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *Proc. Int. Conf. Mach. Learn.*, Atlanta, GA, USA, May 2013, vol. 28, no. 3, pp. 1319–1327.

[24] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Jun. 2013, vol. 30, pp. 1–3.

[25] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400

[26] P. Guo, Z. Xue, L. R. Long, and S. Antani, "Cross-dataset evaluation of deep learning networks for uterine cervix segmentation," *Diagnostics*, vol. 10, no. 1, p. 44, Jan. 2020.

[27] X. Zhang and S.-G. Zhao, "Cervical image classification based on image segmentation preprocessing and a CapsNet network model," *Int. J. Imag. Syst. Technol.*, vol. 29, no. 1, pp. 19–28, Mar. 2019.

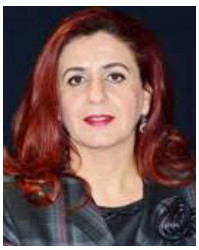
[28] A. Ghoneim, G. Muhammad, and M. S. Hossain, "Cervical cancer classification using convolutional neural networks and extreme learning machines," *Future Gener. Comput. Syst.*, vol. 102, pp. 643–649, Jan. 2020.

[29] S. Mustafa and M. Dauda, "Evaluating convolution neural network optimization algorithms for classification of cervical cancer macro images," in *Proc. 15th Int. Conf. Electron., Comput. Comput. (ICECCO)*, Abuja, Nigeria, Dec. 2019, pp. 1–5.

[30] C. Yuan, Y. Yao, B. Cheng, Y. Cheng, Y. Li, Y. Li, and X. Wang, "The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images," *Sci. Rep.*, vol. 10, no. 1, Jul. 2020, Art. no. 11639.

[31] D. Huang, T. T. Allen, W. I. Notz, and R. A. Miller, "Sequential Kriging optimization using multiple-fidelity evaluations," *Struct. Multidisciplinary Optim.*, 2006.

KHALED MABROUK AMER ADWEB received the master's degree from Near East University, Turkey, in 2015, where he is currently pursuing the Ph.D. degree with the Department of Management of Information Science. He has a strong background in machine learning, image processing, and computational intelligence.



NADIRE CAVUS is currently a Lecturer and the Chairperson with the Department of Computer Information Systems, Near East University, Turkey. She has got many scientific articles published by the worldwide famous journals. She is also in the editorial board and advisory board of several scientific journals. She also acts as a referee to these journals. Her research interests include information systems, health systems, networks, e-learning systems, and learning management systems (LMSs).

BORAN SEKEROGLU (Associate Member, IEEE) was born in Nicosia, Turkey, in 1980. He received the B.S., M.S., and Ph.D. degrees in computer engineering from Near East University, Nicosia, in 2001, 2004, and 2008, respectively. From 2002 to 2008, he was a Research Assistant with the AI Laboratory. He became an Assistant Professor in September 2009 and an Associate Professor in 2020. He is currently the Chairperson of the Information Systems Engineering Department, Near East University, and the Vice-Chairperson of the Artificial Intelligence Engineering Department, Near East University. He is the author of more than 50 articles, conference papers, and book chapters. His research interests include AI, machine learning applications, deep learning, image processing, and computer vision.