IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 19, 2020, accepted June 14, 2020, date of publication June 19, 2020, date of current version July 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3003890

Skin Lesions Classification Into Eight Classes for ISIC 2019 Using Deep Convolutional Neural Network and Transfer Learning

MOHAMED A. KASSEM[1], KHALID M. HOSNY[2], AND MOHAMED M. FOUAD[3]

[1]Department of Robotics, Faculty of Artificial Intelligence, Kafrelsheikh University, Kafr El-Sheikh 33516, Egypt
[2]Department of Information Technology, Faculty of Computers and Informatics, Zagazig University, Zagazig 44511, Egypt
[3]Department of Electronics and Communication, Faculty of Engineering, Zagazig University, Zagazig 44511, Egypt

Corresponding author: Khalid M. Hosny (k_hosny@yahoo.com)

ABSTRACT Melanoma is a type of skin cancer with a high mortality rate. The different types of skin lesions result in an inaccurate diagnosis due to their high similarity. Accurate classification of the skin lesions in their early stages enables dermatologists to treat the patients and save their lives. This paper proposes a model for a highly accurate classification of skin lesions. The proposed model utilized the transfer learning and pre-trained model with GoogleNet. The model parameters are used as initial values, and then these parameters will be modified through training. The latest well-known public challenge dataset, ISIC 2019, is used to test the ability of the proposed model to classify different kinds of skin lesions. The proposed model successfully classified the eight different classes of skin lesions, namely, melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and Squamous cell carcinoma. The achieved classification accuracy, sensitivity, specificity, and precision percentages are 94.92%, 79.8%, 97%, and 80.36%, respectively. The proposed model can detect images that do not belong to any one of the eight classes where these images are classified as unknown images.

INDEX TERMS Melanoma classification, skin lesions, convolution neural network, GoogleNet; ISIC 2019, bootstrap multiclass SVM, transfer learning.

I. INTRODUCTION

In 2018 [1], the WHO reported that there are more than 14 million new cancer patients and more than 9.6 million deaths over the world because of cancer. These statistics show that cancer is the leading cause of human death [2], [3]. Skin Cancer initially occurs on the upper layer of the skin, the epidermis, where it is noticeable and can be seen by human eyes [4]. Skin cancer is one of the significant contributors to the cause of death over the world [5].

Different types of skin cancers have been discovered. Melanoma is a well-known kind of skin cancer, which usually is the most malignant lesion compared to other skin lesions types [6], [7]. Melanoma is one of the fastest spreading skin cancers where recent studies show that the number of skin cancer patients increased year by year [8], [9]. Automatic computer-aided systems for accurate classification of skin lesions are beneficial to saving human life. In the present era, Computer-Aided Diagnosis (CAD) systems become a necessity for preliminary diagnosis of different diseases. Image-based CAD systems utilize skin lesions images without any other medical information to provide dermatologists with an accurate diagnosis [10]. Furthermore, an image-based CAD system could classify different skin lesions based on features extracted from the colors in images of the skin [11]. Based on its accuracy, a CAD system could lead to early diagnosis of skin cancer and then open the door for treatments to save human life [12].

Dermoscopy is the most well-known method of skin imaging, which showed an improvement in the diagnosis of melanoma compared to that of the naked eye [13]. Nevertheless, the benchmark datasets of skin cancer are limited and contain a few numbers of classes. Besides, the available number of images in these classes is limited. Three different issues make the automatic identification of melanoma from dermoscopy images a challenging task. First, although the skin lesions belong to different classes, the characteristics of these lesions, such as size, texture, color, and shape,

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

are very similar, which makes classification a challenging task. Second, there is a high correlation between melanoma and non-melanoma lesions. Third, environmental conditions, such as hair, veins, and illumination [14].

Several attempts made to overcome these challenging issues. Oliveira *et al.* [15] used low-level hand-crafted features to differentiate between melanoma and non-melanoma lesions. Unfortunately, this trial led to wrong results due to the high visual similarity, significant intra-class variations, and dermoscopic artifacts [16]. Pathanbet *et al.* [17] and Shimizu *et al.* [18] segmented the input images to remove the background and unnecessary contents to improve the classification of the skin lesions. Nonetheless, poor results were obtained when segmentation and classification procedures focused on features at low levels, which results in low discrimination rates [19].

Many techniques, such as ABCDE rule, genetic algorithms, support vector machines (SVMs), and artificial neural networks (ANNs), proposed to assess the skin lesion and classify it as either melanoma or benign [20]–[24]. All of these procedures were efficient, cost-effective, and less painful than conventional medical techniques.

It is undisputed that deep learning and CNNs are the preferred techniques in many computer vision applications [25]. Esteva *et al.* [26] utilized a pre-trained deep CNN to classify benign nevi from malignant melanomas, which outperformed dermatologist discrimination rates. Hosny *et al.* [27] used the transfer learning with AlexNet to classify the images of the Ph2 dataset as defined in [28] into three classes called Atypical Nevus, Common Nevus, and Melanoma. Hosny *et al.* [29], [30] used image augmentation and transfer learning with different pre-trained deep neural networks (DNN) to get a significant improvement in the classification rates with the well-known datasets of skin lesions, Derma quest [31], Derma IS [32], and MED-NODE [33]. Gessert *et al.* [34] tried to ensemble a multi-resolution efficientNets with loss balancing. They achieve an accuracy rate of ∼ 63.4%, as listed in the ISIC 2019 leaderboard.

Most available standard datasets for skin lesions [31]–[33] are small datasets and contain small numbers of ∼ 200 images and consist of only two types of skin lesions (two classes). In contrast, the Ph2 dataset [28] consists of three lesions type. The International Skin Imaging Collaboration (ISIC) released four challenging datasets namely ISBI 2016 [35], ISIC 2017 [36], ISIC 2018 [37], and ISIC 2019 [38]. The first challenge, ISBI 2016, was the first challenge, which consisted of two classes with images ∼1200. In the second challenge, ISIC 2017, the number of images and classes increased to ∼2000 images while the number of classes increased to three. ISIC 2018 contains ∼10000 images and divided into seven classes of skin lesions. These classes are nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis Intra Epithelial Carcinoma (AKIEC), Benign Keratosis (BKL), Dermatofibroma (DF), and Vascular lesion (VASC). The most recent challenging dataset, ISIC 2019, contains 25,331 images divided into eight classes–the seven classes as defined in ISIC 2018 with an additional new class called Squamous Cell Carcinoma (SCC).

The contributions in this work can be summarized through the following points:

1. We modified the architecture of the GoogleNet by adding more filters to each layer, to enhance features and reduce noise.
2. We replaced the last three layers in GoogleNet in two different ways:
   - **a.** The last three layers have been dropped out and replaced with a new fully-connected, SoftMax, and classification output layers. By using the SoftMax, the proposed model can work through binary and multi-class. The probabilities output of SoftMax ranges from 0 to 1. In binary classification, these probabilities summed to be one. In multi-class classification, the possibility for every class will be an indicator for each class, but the target class will be the high probability.
   - **b.** The second way was done by dropping out the last two layers only and keeping the original fully connected layer of GoogleNet as features extractors. Then, a bootstrap multi-class support vector machine is used to classify images. This change aims to detect unknown images (outliers).
3. In the proposed model, no pre-processing such as noise reduction, segmentation, or enhancement is used. The weights of All layers are fine-tuned; also, the proposed model does not overfit even with some classes containing a small number of images.
4. To the best knowledge of the authors, there is no published work for the classification of the ISIC 2019 challenge. In this paper, we classified this challenging dataset with high-performance measures in two different ways.
5. The modified GoogleNet achieved higher classification rates than other deeper architectures.

The rest of this work is organized as follows: The utilized methods are described in section 2; the datasets, the performed experiments, and the discussion are described and discussed in section 3, and finally, the conclusion is presented in section 4.

II. METHODS

Traditional neural networks, in general, consist of three layers: input, output, and one single hidden layer. Training methods for these networks encountered a problem, such as the values of the gradient may equal zero or close to zero when weights update. This problem is called the vanishing gradient [39], [40]. Deep convolutional architecture is utilized to overcome the difficulties resulting from traditional methods of learning [41], [42].

A. GOOGLE-NET

Hosny *et al.* [29], [30] performed several experiments using different architectures such as Alex-net, Resnet, VGG,

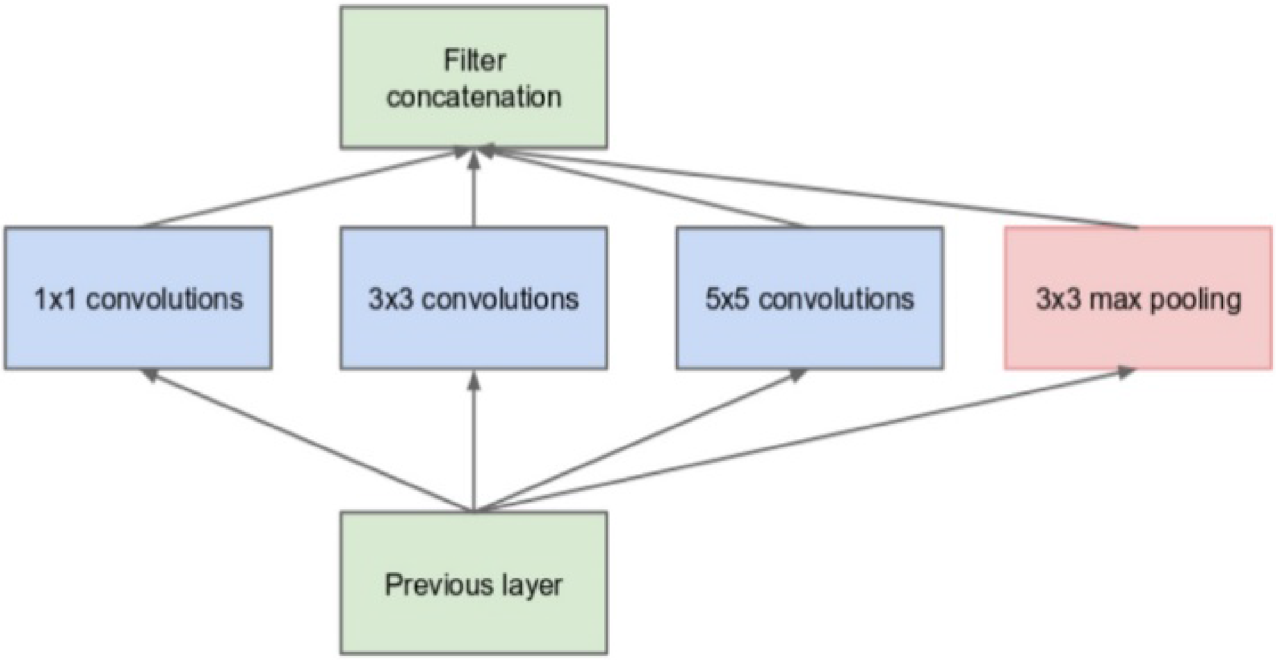


FIGURE 1. Inception model in GoogleNet [43].

and GoogleNet. They showed that the VGG could not work using the same hardware resources and required high-configuration hardware and large memory. GoogleNet outperformed the other architectures.

Increasing the network size or its depth is the simplest way to improve deep neural network performance where the depth refers to the number (levels) of network layers. Successful training of deeper models required a large number of labeled data. There are two drawbacks in this way. First, this process involves the evaluation of a large number of parameters. Such parameters may lead to architecture overfitting, especially when a limited labeled dataset is used for training. The second drawback is concerned with the computational cost, where this cost increased when using a network with many hidden layers.

GoogleNet [43] is a deep convolution network. It is also known as inception architecture. The main idea behind inception architecture is how to estimate and distribute the dense components in the optimum sparse structure of a convolutional network. Based on that, the deepest network activations are either unnecessary (zero value) or redundant due to the correlations between them. The most powerful deep-network architecture would have a sparse connection between the activations, which means that it is not necessarily a connection between all 512 output channels and all 512 input channels. These connections theoretically eliminated using different techniques, which led to a decrease in weight/connection sparse. The 2014 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC14) [44] proposed to be classified by GoogleNet.

GoogleNet built an innovation module called inception, which approximates a sparse CNN by a conventional dense construction, as displayed in Fig. 1. As mentioned above, a small number of neurons is sufficient. The number of convolution filters and the width of a specific kernel size are also kept low. In GoogleNet, various sizes of convolutions are used to extract data and features with different scales ($5 \times 5, 3 \times 3, 1 \times 1$). Another essential element of this module is a (1×1) bottleneck convolutional layer, as displayed in the figure. This process helps to reduce the computation requirement significantly, as explained below in Fig. 1.

Let us consider the GoogleNet's first inception module as an example with 192 channels as input. It has 128 and 32 filters for each group with size 3×3 and 5×5, respectively. Thus, for the 5×5 filter, the computation order is $25 \times 32 \times 192$. It is expandable when going deeper by increasing the width of both the network and the filter. A 1×1 convolution is used with only 16 filters to minimize the dimensions of the input channel. The computations will reduce from $25 \times 32 \times 192$ to $16 \times 192 + 25 \times 32 \times 16$. All of these changes enable a large width and depth to the network.

The fully-connected layers are replaced with a simple pooling global average layer. This change in GoogleNet reduces the total number of parameters significantly.

B. TRANSFER LEARNING

Transfer learning is a machine learning technique, which means a model is used with another task related to what it has trained for the first time. Domain adaptation and transfer learning refer to the condition in which the training of one set is used to improve generalization in another setting [45], [46]. Due to the enormous or massive resources required for deep learning and the extensive number of images, transfer learning is a well-known approach in deep networks. In addition to the limited numbers of available skin lesions datasets, these datasets are not suitable to train a deep network from the beginning, due to their small number of images. Transfer learning was the best solution to overcome this problem.

Transfer learning is achieved in three steps. First, selecting the pre-trained model; the GoogleNet is trained using ImageNet. In the second step, several layers forming the end of the GoogleNet are replaced. In the third step, we reuse and adapt the model layers for new tasks by using fine-tuned layers.

The GoogleNet architecture was adapted to classify skin images by replacing several layers in two ways. First, the last three layers are replaced with a new fully-connected, SoftMax, and classification output layer. The size of the output fully connected layer is $N \times 2048$, where N refers to the class number, which equals 8. Fig. 2 explains the difference between the original model of GoogleNet and the proposed model. Second, only the last two layers have been removed instead of removing all three layers. We use the original fully connected layers with GoogleNet architecture to extract the features, and then these features are classified using a multi-class SVM. The multi-class SVM evaluates the similarity score of an image to different classes by computing the similarity score for all classes when multi-class SVM is used to predict a new image. The utilization of the multi-class SVM is shown in Fig. 2.

Fine-tuning the weights of the model is the last step. Only the last layers (transferred layers) significantly lead to lower results compared with fine-tuning all model layers (0.37 versus 0.17, respectively) [47]. Consequently, fine-tuning the entire network, helps in improving the classification performance significantly. In this paper, weights for all layers in the proposed model are randomly initialized, where these weights are automatically adjusted. Researchers attempted to remove more than three layers. The quality of the classification measures was less than the last three layers.

III. EXPERIMENTAL RESULTS AND DISCUSSION

An IBM computer equipped with a Core i7 processor and 16 GB DDRAM with a GeForce MX150 NVIDIA graphics card was used to perform the experiments. An x64-bit MATLAB 2018 was used to execute the program. The maximum training epochs number was set to 40, while the mini-batch size was 6; the initial learning rate was 0.00001, and the momentum was 0.9.

A. DATASET

The well-known dataset from the ISIC 2019 challenge was used to test and evaluate the proposed model. The ISIC 2019 dataset contains images of HAM10000 and the BCN_20000. HAM10000 contains ∼10000 images with a size of 600×450. This dataset was the older challenge of ISIC 2018. While the BCN_20000 contains ∼19424 images of size 1024 × 1024. The test set includes an additional unknown class that was not presented in the training dataset. The limitations of ISIC 2019 discussed in this paper with different experiments to gain the best performance. GoogleNet was used for automatic identification of skin lesions with transfer learning (knowledge). Experimental results have been made in two ways, as discussed in the following subsection.

The performance of the proposed model was evaluated using five quantitative measures, accuracy, sensitivity, specificity, and precision [48]. These measures are computed from Fig. 3 as follows:

$$\text{Accuracy} = \frac{t_p + t_n}{t_p + f_p + f_n + t_n} \tag{1}$$

$$\text{Sensitivity(TPR or Recal)} = \frac{t_p}{t_p + f_n} \tag{2}$$

$$\text{Specificity(TNR)} = \frac{t_n}{f_p + t_n} \tag{3}$$

$$\text{Precision(PPV)} = \frac{t_p}{t_p + f_p} \tag{4}$$

$$F1Score = 2 \times \frac{\text{Precision} \times \text{sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{5}$$

where $t_p, f_p, f_n,$ *and* t_n refers to true positive, false positive, false negative, and true negative, respectively. The acronyms, TPR, TNR, and PPV, refer to true positive rate, true negative rate, and a positive predictive value. The true negative should be large, while false-positive rates should be small, resulting in most points falling in the left part of the receiver operating characteristic (ROC) curve [48].

B. CLASSIFYING KNOWN CLASSES USING MODIFIED GoogleNet BY REPLACING THE LAST THREE LAYERS

The dataset is divided into three parts. The first part was 80% of the dataset for training, the second part was 10% for validation, and the third was 10% for testing. ISIC 2019 consists of 8 classes namely: Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AKIEC), Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), and Squamous Cell Carcinoma (SCC). ISIC 2019 consists of 25,331 images where these images are distributed with a different number of images: MEL images is 4,522; NV is 12,875; BCC is 3,323; AKIEC is 867; BKL is 2,624; DF is 239; VASC is 253; and SCC is 628. Fig. 4 shows sample images for the different types of skin cancer. This dataset is one of the most challenging tasks to classify various images into eight classes with an imbalanced number of images in each class.

CUDA has been used to implement the model over GPU. The authors removed some layers and replaced them with new layers to fulfill the main purpose of this paper to classify eight skin lesions in the challenge of ISIC 2019. In the proposed model, the last three layers are replaced with three new layers to be suitable for the required task here; to classify eight classes instead of classifying 1000 classes in the GoogleNet. Weights of all layers were initialized randomly; during the training process, the weights of these layers automatically updated, as discussed in the following sections.

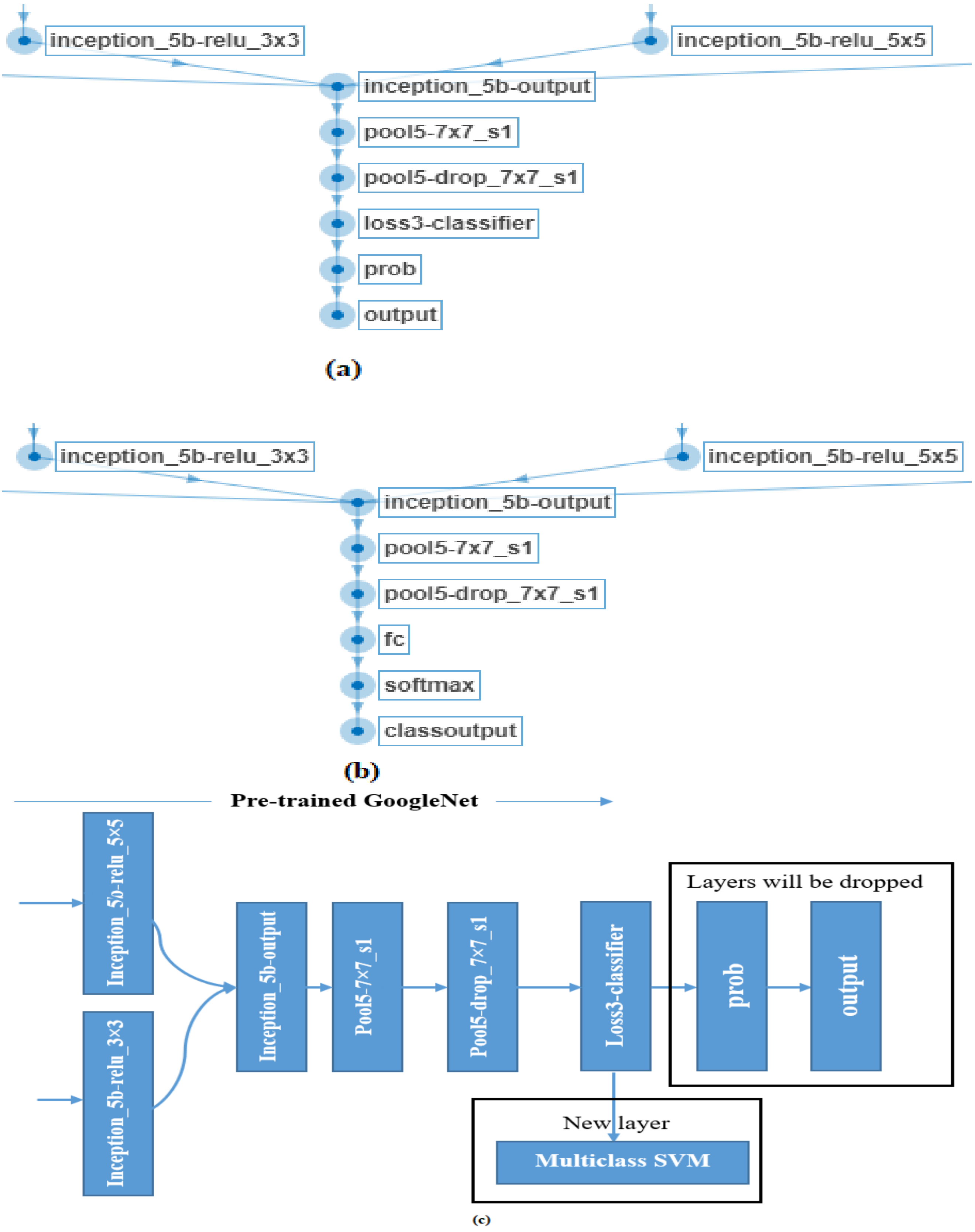
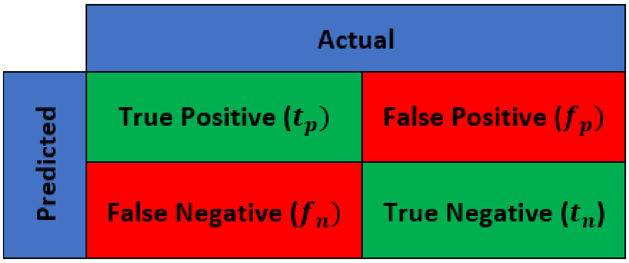


FIGURE 2. (a) original GoogleNet, (b) modified GoogleNet by dropping out last three layers, and (c) modified GoogleNet by replacing last two layers with multiclass SVM.



	Actual	
Predicted	True Positive (t_p)	False Positive (f_p)
	False Negative (f_n)	True Negative (t_n)

FIGURE 3. Confusion matrix used to compute performance measures.

The total number of images is 25,331. The authors trained and validated the pre-trained GoogleNet model after applying the transfer learning using 80% equals 20,265 images, and 10 % equals 2,531 images of the ISIC 2019 dataset, respectively, while the rest of the dataset 10% equals 2,531 images used for testing. The confusion matrix of the testing in this experiment is shown in Fig. 5. The performance measures of this experiment were:

$$\begin{aligned} \text{Accuracy} &= 93.31\% \\ \text{Sensitivity} &= 53.3\% \\ \text{Specificity} &= 95.37\% \\ \text{Precision} &= 62.5\% \\ \text{F1 \ Score} &= 57.53\% \end{aligned}$$

Sensitivity and precision are very low; in a medical CAD system, it is unacceptable because it is related to human life. The reason behind the lower measure for sensitivity and precision is the imbalance between classes in the dataset. There is a huge gap in the number of images in each class. The NV class contains 12,875 images while AKIEC contains 867 images, DF contains 239 images, VASC contains 253 images, and the SCC class contains 628 images. The big gap in the number of images results in a biased classification to the class that contains the largest number of images.

Therefore, the authors suggested image augmentation for solving this problem. The second experiment was carried out using ISIC 2019 dataset, where authors augmented the smallest classes, AKIEC, DF, VASC, and SCC. Different augmentation techniques were utilized where vertical and horizontal-shift, vertical and horizontal-flip, and rotation were applied. The number of images in AKIEC, DF, VASC, and SCC become 1660, 1673, 1518, and 1884, respectively. The second experiment was carried out using the augmented classes, AKIEC, DF, VASC, and SCC. The total number of images in this experiment is 30,079. The confusion matrix for this experiment is shown in Fig. 6 when dividing the dataset for 80% equals 24,067 images, 10% equals 3,006 images, and 10% equals 3,006 images for training, validating, and testing, respectively. The performance measures were:

$$\begin{aligned} \text{Accuracy} &= 94.2\% \\ \text{Sensitivity} &= 74.5\% \\ \text{Specificity} &= 96.5\% \\ \text{Precision} &= 73.62\% \\ \text{F1 \ Score} &= 74.02\% \end{aligned}$$

We observed a significant improvement in the sensitivity and precision compared with the first experiment. The imbalance gap in the number of images in each class was reduced. Despite the image augmentation in the small classes, there is still a considerable difference that makes classifiers biased to the class containing the most significant number of images. So, additional image augmentation steps were performed where the number of images in all classes increased to be equal to the number of images in the largest class, NV, which contains 12,875 images. This process leads to the best result, where the classification performance approached 100%.

The second way is to make the number of images in each class equal to each other. The third experiment was carried out using the ISIC 2019 dataset, where the authors reduced

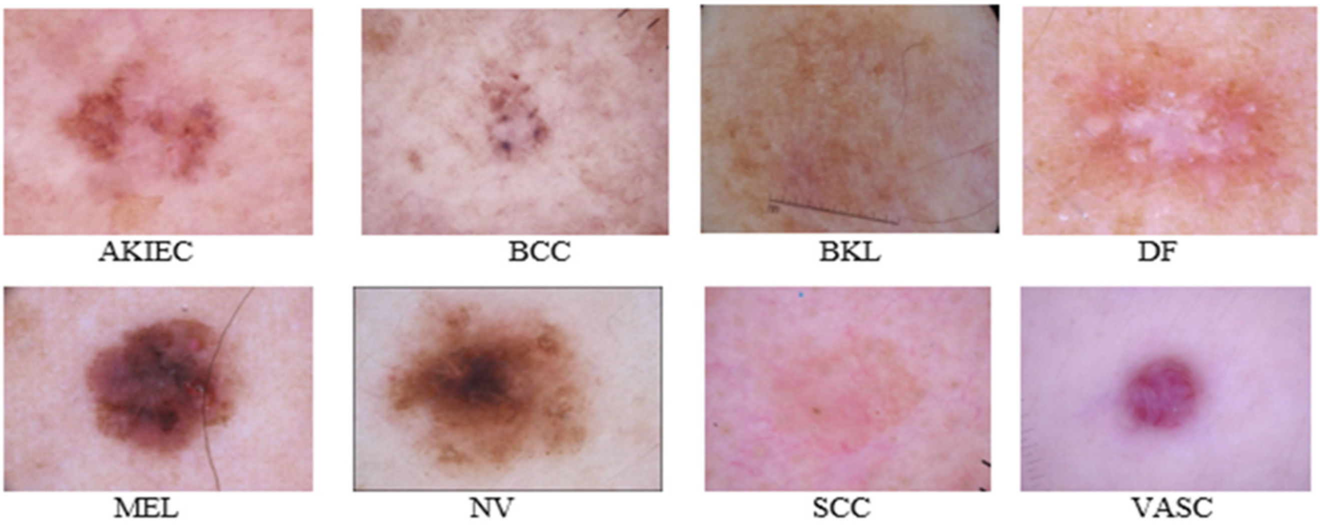
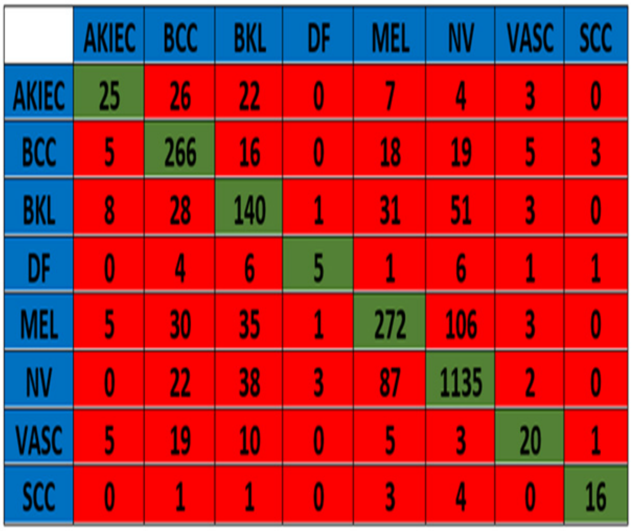
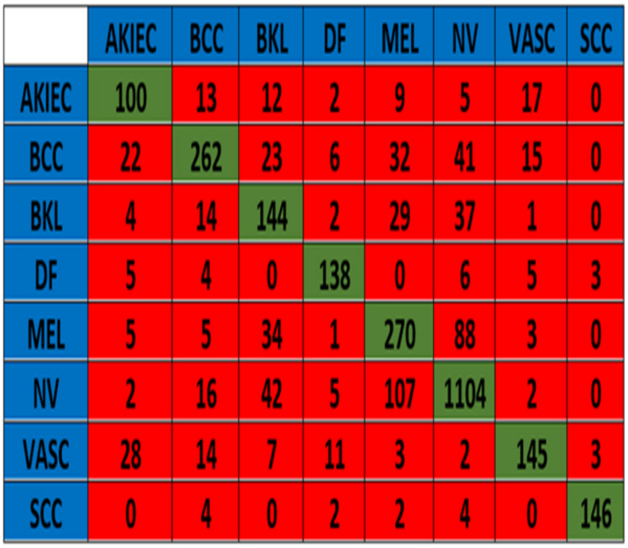


FIGURE 4. ISIC 2019 different skin lesions examples.



	AKIEC	BCC	BKL	DF	MEL	NV	VASC	SCC
AKIEC	25	26	22	0	7	4	3	0
BCC	5	266	16	0	18	19	5	3
BKL	8	28	140	1	31	51	3	0
DF	0	4	6	5	1	6	1	1
MEL	5	30	35	1	272	106	3	0
NV	0	22	38	3	87	1135	2	0
VASC	5	19	10	0	5	3	20	1
SCC	0	1	1	0	3	4	0	16

FIGURE 5. Confusion matrix (8 × 8) for the first experiment.



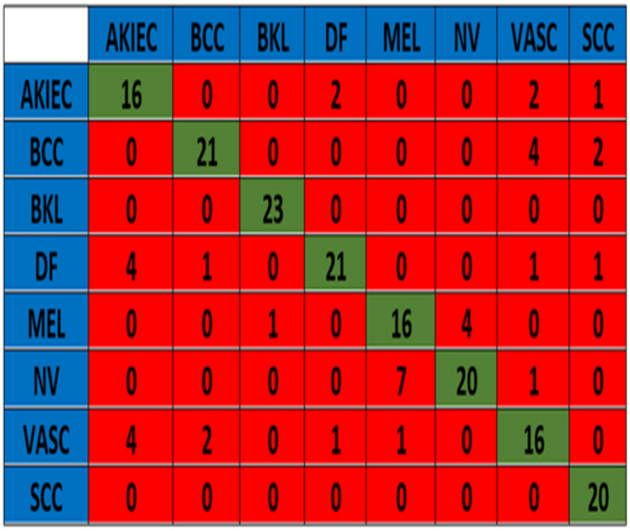
	AKIEC	BCC	BKL	DF	MEL	NV	VASC	SCC
AKIEC	100	13	12	2	9	5	17	0
BCC	22	262	23	6	32	41	15	0
BKL	4	14	144	2	29	37	1	0
DF	5	4	0	138	0	6	5	3
MEL	5	5	34	1	270	88	3	0
NV	2	16	42	5	107	1104	2	0
VASC	28	14	7	11	3	2	145	3
SCC	0	4	0	2	2	4	0	146

FIGURE 6. Confusion matrix (8 × 8) for the second experiment.

the number of images in all classes to be equal to the number of images in the smallest class, DF. The authors eliminated images randomly to reduce the number of images to 239 images.

The total number of images in this experiment is 1,912. The proposed model of modified GoogleNet trained, validated, and tested using 80% equals 1,530 images, 10% equals 191 images, and 10% equals 191 images, respectively. The confusion matrix of this experiment shown in Fig. 7, where the obtained results are:

$$\begin{aligned}\text{Accuracy} &= 94.92\%\\ \text{Sensitivity} &= 79.8\%\\ \text{Specificity} &= 97\%\\ \text{Precision} &= 80.36\%\\ \text{F1 Score} &= 80.07\%\end{aligned}$$



	AKIEC	BCC	BKL	DF	MEL	NV	VASC	SCC
AKIEC	16	0	0	2	0	0	2	1
BCC	0	21	0	0	0	0	4	2
BKL	0	0	23	0	0	0	0	0
DF	4	1	0	21	0	0	1	1
MEL	0	0	1	0	16	4	0	0
NV	0	0	0	0	7	20	1	0
VASC	4	2	0	1	1	0	16	0
SCC	0	0	0	0	0	0	0	20

FIGURE 7. Confusion matrix (8 × 8) for the third experiment.

We observed that the values of all measures, especially the sensitivity and precision, increased compared with the first experiments. Table 1 gives an overview of the obtained results for the performed experiments. Fig. 8 shows the obtained ROC curve, while Fig. 9 displays the plotted results. Based on Table 1, it is clear that the lowest performance occurred when using the original dataset in the first experiment. In the second experiment, the obtained results improved. The third experiment shows the best values of performance measures. In this experiment, the number of images was reduced by random elimination to match the number of images in the smallest class. The sensitivity increased by 26% from 53.3% to 79.8%.

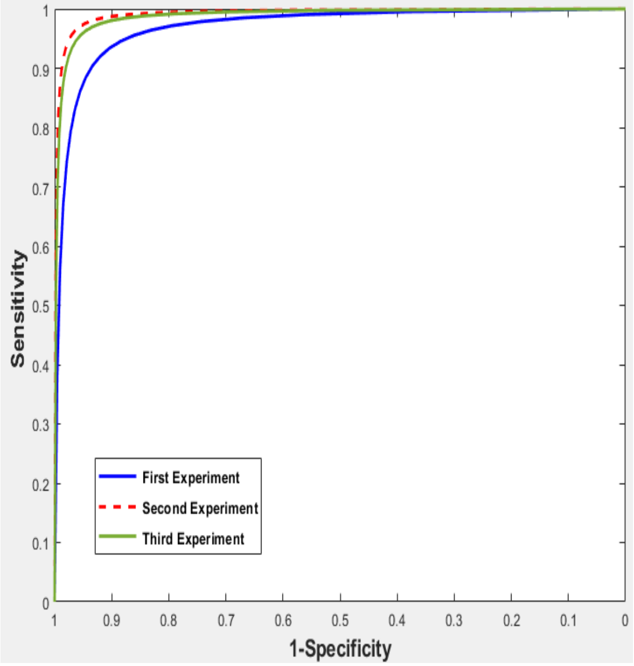


FIGURE 8. Roc for proposed models.

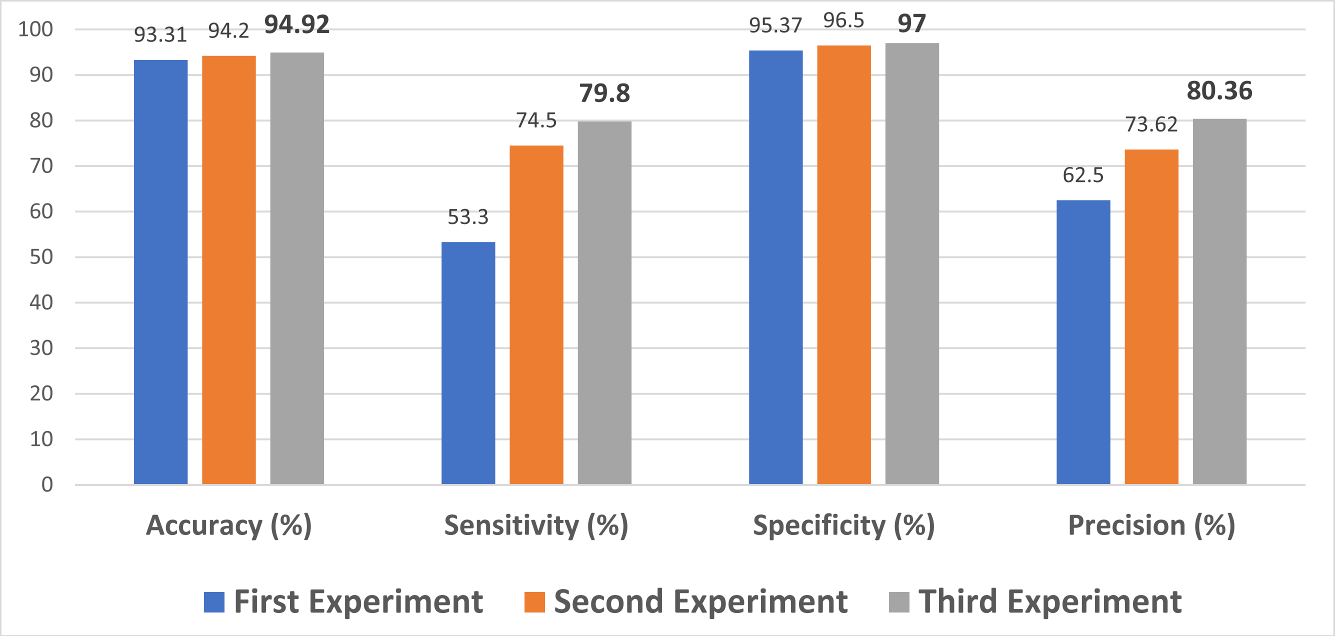


FIGURE 9. Visualization of the performance metrics for the proposed model.

TABLE 1. Proposed model accuracy.

Experiments	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Precision (%)**
First	93.31	53.3	95.37	62.5
Second	94.2	74.5	96.5	73.62
Third	**94.92**	**79.8**	**97**	**80.36**

Similarly, the precision increased by 17% from 62.5% to 80.36; the specificity increased from 95.37% to 97%. Finally, the accuracy moved from 93.31% to 94.92%.

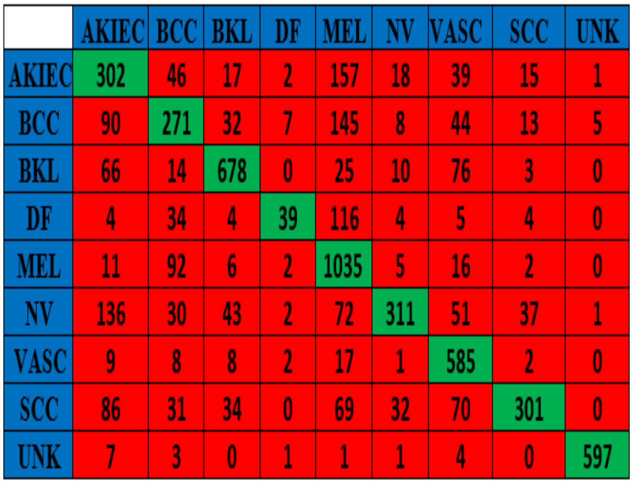
C. OUT-OF-DISTRIBUTION IMAGES DETECTION USING GoogleNet AND BOOTSTRAP MULTI-CLASS SVM

The hardest challenges in ISIC 2019 are the imbalanced number of images in each class and the detection of unknown images (outliers). A bootstrap multi-class SVM aggregation has been used to handle the problem of imbalanced classes, which helps to reduce overfitting. A single bootstrap sampling on the images datastore has been performed which balanced out classes. It worked by replacement sampling with the weight according to the imbalanced class. The second problem during the training phase is that there are no samples of images identified as strange images or outliers. So, a robust model that can detect any outliers without negatively affecting the overall accuracy is required. The second proposed model is based on GoogleNet and transfers learning. In this model, we apply transfer learning by removing only the last two layers, SoftMax and classification output, named (prob) and (output), respectively. We kept the fully-connected layer to extract the image features.

The bootstrap multi-class SVM was used to classify these features. The trained bootstrap multi-class SVM saved from the training process to classify new images (test set). The saved classifier was used to classify the ISIC 2019 test set and compute the similarity score for each image with different classes. In the test set, we added different outlier images (unknown) by adding external images. The outlier images are collected from [49] and some healthy skin images. There are two different ways used to enable the proposed model to detect strange images. First, the proposed model is trained with the nine classes, the eight known classes plus to the new class named unknown images (UNK). Different augmentation approaches, such as rotation, translation, and flipping, have been applied. The total number of images in this experiment is 60,162. The dataset has been divided into 80% equals 48,130 images, 10% equals 6,015 images, and 10% equals 6,015 images for training, validation, and testing, respectively. The confusion matrix of this experiment shown in Fig. 10, where the obtained results are:

$$\begin{aligned}\text{Accuracy} &= 92.99\%\\ \text{Sensitivity} &= 70.44\%\\ \text{Specificity} &= 96\%\\ \text{Precision} &= 62.78\%\\ \text{F1 Score} &= 66.39\%\end{aligned}$$

The second way to detect the outlier images: Instead of train the proposed model on external images (unknown), a conditional rule has been applied. If the similarity score is less than value, then this image is defined as an unfamiliar image. Any image with a similarity score of less than 0.5 for the eight defined classes will designate as unknown. We choose this value because the proposed model can't identify this image with certainty to one of the eight classes that were identified during training. The next algorithm



	AKIEC	BCC	BKL	DF	MEL	NV	VASC	SCC	UNK
AKIEC	302	46	17	2	157	18	39	15	1
BCC	90	271	32	7	145	8	44	13	5
BKL	66	14	678	0	25	10	76	3	0
DF	4	34	4	39	116	4	5	4	0
MEL	11	92	6	2	1035	5	16	2	0
NV	136	30	43	2	72	311	51	37	1
VASC	9	8	8	2	17	1	585	2	0
SCC	86	31	34	0	69	32	70	301	0
UNK	7	3	0	1	1	1	4	0	597

FIGURE 10. Confusion matrix (9 × 9) for the unknown class.

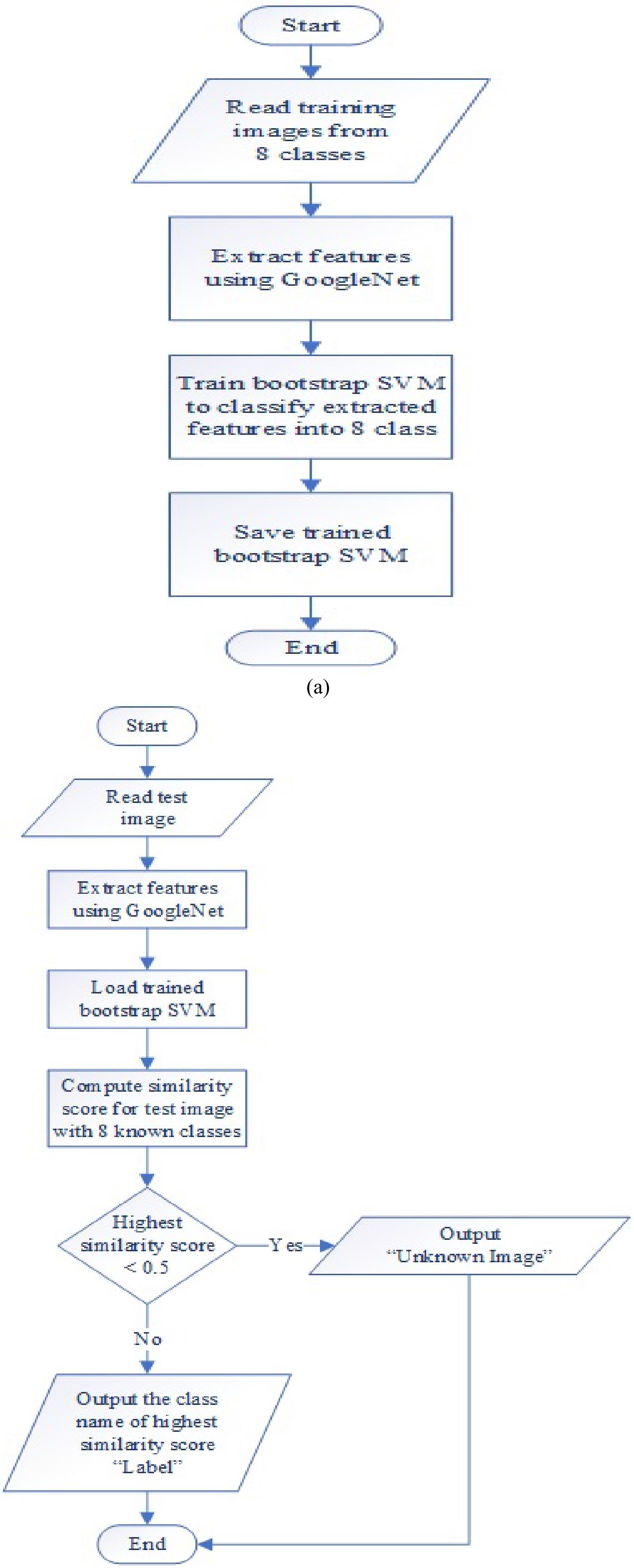
explains the overall process of detecting unknown images. Fig. 11 illustrates the process of training and testing for the proposed model to detect outliers.

Algorithm 1 Detection of Desired Class or Unknown Image

```
Input: detection of the unknown image from an external source
Output: Detect the desired label or unknown image
1    Load trained classifier
2    For $I = 1$; $I <=$ total test set Do
3       Extract features using fully connected layer in
          pre-trained model GoogleNet
4       For $T = 1$; $T <=$ total number of classes Do
5         Compute similarity score $SS(I, T)$
6         Save $SS(I, T)$ go to step 4; compute SS to
          next class
7       END
8       Get highest $SS(I, T)$
8          IF highest $SS(I, T) >= 0.5$
9             Output the class name "Label."
10         ELSE
11            Output "unknown image."
12         END-IF go to step 2; load next image
13   END
```

The total number of images in this experiment is 25,331. The proposed model trained and validated using 80% equals 20,265 images, and 10% equals 2,531 images of the ISIC 2019 dataset, respectively, while the rest of the dataset 10% equals 2,531 images used for testing. The obtained results of this experiment using the modified GoogleNet with bootstrap SVM to detect outliers are:

$$\text{Accuracy} \approx 81\%$$
$$\text{Sensitivity} \approx 74\%$$
$$\text{Specificity} \approx 84\%$$
$$\text{Precision} \approx 77\%$$
$$\text{F1 Score} = 75.47\%$$


Start
Read training images from 8 classes
Extract features using GoogleNet
Train bootstrap SVM to classify extracted features into 8 class
Save trained bootstrap SVM
End
Start
Read test image
Extract features using GoogleNet
Load trained bootstrap SVM
Compute similarity score for test image with 8 known classes
Highest similarity score < 0.5
Yes
Output "Unknown Image"
No
Output the class name of highest similarity score "Label"
End

(a)

(b)

FIGURE 11. Training and testing to detect outliers, (a) training process, and (b) testing process.

The values of the performance measures are less than those of the previous experiments for many reasons. The first one because the test performed using the ISIC 2019 test set.

The second one is the unknown images and the threshold that applied to the similarity score, which may ignore the image with uncertainty related to one of the previously trained classes. The third one is the use of traditional machine learning methods such as multi-class SVM, which give a lower performance measure [30].

A comparative study is performed with ISIC 2019 leaderboard. The highest performance measure was achieved by Gessert *et al.* [34], which tried to ensemble a multi-resolution efficientNets with loss balancing. Gessert and his coauthors achieved an accuracy rate of ~63%, as listed in the ISIC 2019 leaderboard. Table 2 summarizes the comparative study, while Fig. 12 shows the obtained ROC.

TABLE 2. Comparative study.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
Gessert et al. [34]	63	73	-	-
Proposed Method	81	74	84	77

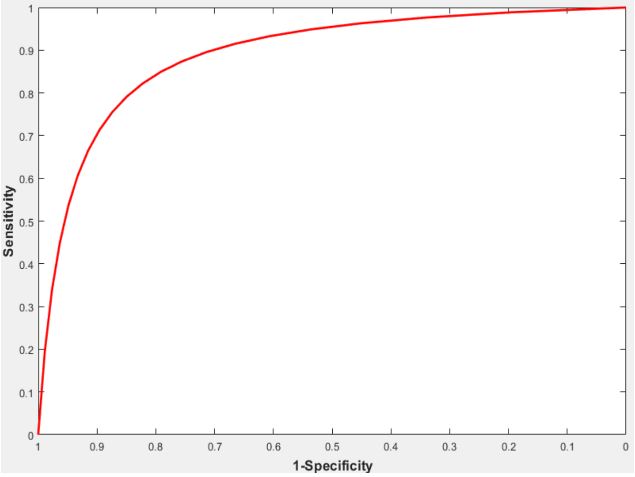


FIGURE 12. Roc for the proposed model.

IV. CONCLUSION

A method for ISIC 2019 challenge dataset has been developed here by using transfer learning and pre-trained deep neural network GoogleNet. The proposed method can classify eight different types of lesions accurately, even with the imbalance of images between classes. The performance measures of the proposed methods were accuracy, sensitivity, specificity, and precision, while the values of these measures were 94.92%, 79.8%, 97%, and 80.36%, respectively. The performance of the proposed method increased when the number of images in all classes decreased to overcome the problem of imbalance in images between classes. We noticed that when the weights of all architecture layers were fine-tuned, the performance measures become higher than fine-tuning only the replaced layers. Another model proposed to detect unknown images using GoogleNet and multi-class SVM. In the same way of using GoogleNet here, the authors tried to use VGG19, but it can't be trained or tested using the same device. It requires a specific hardware specification that all researchers in various countries cannot meet.

REFERENCES

[1] (2018). *Cancer*. Accessed: Mar. 22, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Y.-E. Choi, J.-W. Kwak, and J. W. Park, "Nanotechnology for early cancer detection," *Sensors*, vol. 10, no. 1, pp. 428–455, Jan. 2010.

[3] (2017). *National Cancer Institute. Cancer Statistics*. Accessed: Oct. 27, 2019. [Online]. Available: https://www.cancer.gov/aboutcancer/understanding/statistic

[4] (2018). *National Cancer Institute. Skin Cancer (Including Melanoma)-Patient Version*. Accessed: Oct. 27, 2019. [Online]. Available: https://www.cancer.gov/types/skin

[5] H. E. Kanavy and M. R. Gerstenblith, "Ultraviolet radiation and melanoma," *Seminars Cutaneous Med. Surg.*, vol. 30, no. 4, pp. 222–228, Dec. 2011.

[6] C. K. Bichakjian, A. C. Halpern, T. M. Johnson, A. F. Hood, J. M. Grichnik, S. M. Swetter, H. Tsao, V. H. Barbosa, T. Y. Chuang, M. Duvic, and V. C. Ho, "Guidelines of care for the management of primary cutaneous melanoma," *J. Amer. Acad. Dermatol.*, vol. 80, no. 1, pp. 208–250, 2019.

[7] *American Cancer Society. Facts & Figures*, Amer. Cancer Soc., Atlanta, Ga, USA, 2020.

[8] R. J. Friedman, D. S. Rigel, and A. W. Kopf, "Early detection of malignant melanoma: The role of physician examination and self-examination of the skin," *CA A, Cancer J. Clinicians*, vol. 35, no. 3, pp. 130–151, 1985.

[9] Z. Apalla, D. Nashan, R. B. Weller, and X. Castellsagué, "Skin cancer: Epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches," *Dermatol. Therapy*, vol. 7, no. S1, pp. 5–19, Jan. 2017.

[10] M. Q. Khan, A. Hussain, S. U. Rehman, U. Khan, M. Maqsood, K. Mehmood, and M. A. Khan, "Classification of melanoma and nevus in digital images for diagnosis of skin cancer," *IEEE Access*, vol. 7, pp. 90132–90144, 2019.

[11] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Med. Imag. Graph.*, vol. 31, nos. 4–5, pp. 198–211, Jun. 2007.

[12] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Syst. Appl.*, vol. 46, pp. 139–144, Mar. 2016.

[13] M. Binder, "Epiluminescence microscopy. A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists," *Arch. Dermatol.*, vol. 131, no. 3, pp. 286–291, Mar. 1995.

[14] C. Barata, M. E. Celebi, and J. S. Marques, "A survey of feature extraction in dermoscopy image analysis of skin cancer," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1096–1109, May 2019.

[15] R. B. Oliveira, A. S. Pereira, and J. M. R. S. Tavares, "Computational diagnosis of skin lesions from dermoscopic images using combined features," *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6091–6111, Oct. 2019.

[16] M. E. Celebi, Q. Wen, H. Iyatomi, and K. Shimizu, "A state-of-the-art survey on lesion border detection in dermoscopy images," *Dermoscopy Image Analysis* vol. 10, pp. 97–129, Sep. 2015.

[17] S. Pathan, K. G. Prabhu, and P. C. Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review," *Biomed. Signal Process. Control*, vol. 39, pp. 237–262, Jan. 2018.

[18] K. Shimizu, H. Iyatomi, M. E. Celebi, K.-A. Norton, and M. Tanaka, "Four-class classification of skin lesions with task decomposition strategy," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 274–283, Jan. 2015.

[19] F. Peñaranda, V. Naranjo, G. Lloyd, L. Kastl, B. Kemper, J. Schnekenburger, J. Nallala, and N. Stone, "Discrimination of skin cancer cells using Fourier transform infrared spectroscopy," *Comput. Biol. Med.*, vol. 100, pp. 50–61, Sep. 2018.

[20] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[21] D. Ruiz, V. Berenguer, A. Soriano, and B. Sánchez, "A decision support system for the diagnosis of melanoma: A comparative approach," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 15217–15223, Nov. 2011.

[22] A. Murugan, S. A. H. Nair, and K. P. S. Kumar, "Detection of skin cancer using SVM, random forest and kNN classifiers," *J. Med. Syst.*, vol. 43, no. 8, p. 269, Aug. 2019.

[23] Z. Ma and J. M. R. S. Tavares, "A novel approach to segment skin lesions in dermoscopic images based on a deformable model," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 2, pp. 615–623, Mar. 2016.

[24] S. B. Sulistyo, W. L. Woo, and S. S. Dlay, "Regularized neural networks fusion and genetic algorithm based on-field nitrogen status estimation of wheat plants," *IEEE Trans. Ind. Informat.*, vol. 13, no. 1, pp. 103–114, Feb. 2017.

[25] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.

[26] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[27] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Skin cancer classification using deep learning and transfer learning," in *Proc. 9th Cairo Int. Biomed. Eng. Conf. (CIBEC)*, Dec. 2018, pp. 90–93.

[28] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH^2—A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 5437–5440.

[29] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Classification of skin lesions using transfer learning and augmentation with alex-net," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0217293.

[30] K. M. Hosny, M. A. Kassem, and M. M. Fouad, "Skin melanoma classification using deep convolutional neural networks," in *Deep Learning for Computer Vision: Theories and Applications*, M. Hassaballah and A. Awad, Eds. Boca Raton, FL, USA: CRC Press, 2020.

[31] (2012). *DermQuest*. Accessed: Nov. 23, 2019. [Online]. Available: http://www.dermquest.com

[32] (2012). *Dermatology Information System*. Accessed: Nov. 23, 2019. [Online]. Available: http://www.dermis.net

[33] I. Giotis, N. Molders, S. Land, M. Biehl, M. F. Jonkman, and N. Petkov, "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images," *Expert Syst. Appl.*, vol. 42, no. 19, pp. 6578–6585, Nov. 2015.

[34] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," 2019, *arXiv:1910.03910*. [Online]. Available: http://arxiv.org/abs/1910.03910

[35] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," 2016, *arXiv:1605.01397*. [Online]. Available: http://arxiv.org/abs/1605.01397

[36] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," 2017, *arXiv:1710.05006*. [Online]. Available: http://arxiv.org/abs/1710.05006

[37] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Dec. 2018, Art. no. 180161.

[38] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "BCN20000: Dermoscopic lesions in the wild," 2019, *arXiv:1908.02288*. [Online]. Available: http://arxiv.org/abs/1908.02288

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[40] W. Jia, M. Yang, and S.-H. Wang, "Three-category classification of magnetic resonance hearing loss images based on deep autoencoder," *J. Med. Syst.*, vol. 41, no. 10, p. 165, Oct. 2017.

[41] S. Wang, J. Yang, G. Liu, S. Du, and J. Yan, "Multi-objective path finding in stochastic networks using a biogeography-based optimization method," *Simulation*, vol. 92, no. 7, pp. 637–647, Jul. 2016.

[42] S. Wang, Y. Jiang, X. Hou, H. Cheng, and S. Du, "Cerebral micro-bleed detection based on the convolution neural network with rank based average pooling," *IEEE Access*, vol. 5, pp. 16576–16583, 2017.

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

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[45] E. S. Olivas, J. D. M. Guerrero, M. M. Sober, J. R. M. Benedito, and A. J. S. Lopez, "Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques," in *Information Science Reference*, vol. 2. Philadelphia, PA, USA: IGI, 2009.

[46] S. Lu, Z. Lu, and Y.-D. Zhang, "Pathological brain detection based on AlexNet and transfer learning," *J. Comput. Sci.*, vol. 30, pp. 41–47, Jan. 2019.

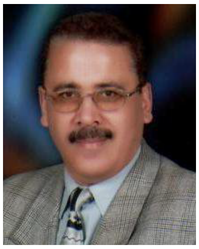
[47] S. Wang, Z. Lu, L. Wei, G. Ji, and J. Yang, "Fitness-scaling adaptive genetic algorithm with local search for solving the multiple depot vehicle routing problem," *Simulation*, vol. 92, no. 7, pp. 601–616, Jul. 2016.

[48] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[49] *Dermnet Skin Disease Atla*. Accessed: May 13, 2020. [Online]. Available: http://www.dermnet.com/

MOHAMED A. KASSEM was born in Mansoura, Egypt, in 1987. He received the B.Sc. and M.Sc. degrees in information technology from the Faculty of Computers and Information, Mansoura University, Mansoura, Egypt, in 2008 and 2015, respectively. He is currently pursuing the Ph.D. degree in information technology with the Faculty of Computers and Information, Zagazig University, Egypt, under the supervision of Prof. Khalid M. Hosny and Prof. Mohamed M. Fouad. He is also a Senior Assistant Lecturer with the Department of Robotics and Intelligent Machines, Faculty of artificial intelligence, Kafrelshiekh University, Egypt.

KHALID M. HOSNY was born in Zagazig, Egypt, in 1966. He received the B.Sc., M.Sc., and Ph.D. degrees from Zagazig University, Egypt, in 1988, 1994, and 2000, respectively. From 1997 to 1999, he was a Visiting Scholar with the University of Michigan, Ann Arbor, and the University of Cincinnati, Cincinnati, OH, USA. He is currently a Professor of information technology with the Faculty of Computers and Informatics, Zagazig University. He published three edited books and more than 70 articles in international journals. His research interests include image processing, pattern recognition, multimedia, deep learning, and computer vision. He is a Senior Member of ACM. He is an editor and scientific reviewer for more than 35 international journals.



MOHAMED M. FOUAD is currently a Professor of electronics and communication with the Department of Electronics and Communication, Faculty of Engineering, Zagazig University, Egypt.