IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 July 2022, accepted 17 September 2022, date of publication 26 September 2022, date of current version 6 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210179

RESEARCH ARTICLE

NSL-MHA-CNN: A Novel CNN Architecture for Robust Diabetic Retinopathy Prediction Against Adversarial Attacks

OTHMANE DAANOUNI, BOUCHAIB CHERRADI, AND AMAL TMIRI
LaROSERI Laboratory, Chouaib Doukkali University, El Jadida 24000, Morocco
Corresponding author: Bouchaib Cherradi (bouchaib.cherradi@enset-media.ac.ma)

ABSTRACT Convolution Neural Network (CNN) models have gained ground in research activities particularly in medical images used for Diabetes Retinopathy (DR) detection. X-ray, MRI, and CT scans have all been used to validate CNN models, with classification accuracy generally reaching that of trained doctors. It is mandatory to evaluate the strength of CNN models used in medical tasks against adversarial attacks especially in healthcare, that is to say, the security of such models is becoming extremely relevant to the diagnosis as this latter will guide high-stakes decision-making. However, little study has been conducted to better comprehend this issue. This paper focuses on MobileNet CNN architecture in order to investigate its vulnerability against fast gradient sign methods (FGSM) adversarial attacks. For this end, a Neural Structure Learning (NSL) and a Multi-Head Attention (MHA) have been used to effectively reduce the vulnerability against attack by end-to-end CNN training with adversarial neighbors that produce adversarial perturbations on optical coherence tomography (OCT) images. With suggested model NSL-MHA-CNN, there has been an ability to maintain model performance on adversarial attack without increasing cost of training. Through theoretical assistance and empirical validation, it was possible to examine the stability of MobileNet architecture and demonstrate its susceptibility, particularly to adversarial attack. The experiments in this paper show that indiscernible degrees of perturbation $\varepsilon < 0.01$ were sufficient to cause a task failure resulting to misclassification in majority of the time. Moreover, empirical simulation shows that the proposed approach advanced in this paper can be an effective method to defense against adversarial attack at level of CNN model testing.

INDEX TERMS Fast gradient sign method, malicious attack, MobileNet, multi-head attention, neural structure learning, vulnerability in CNN.

I. INTRODUCTION

Deep Learning (DL) has recently delivered cutting-edge performance in a variety of applications without the guideline for manual feature extraction [1], in particular in the area of pattern recognition. Recently, multiple deep learning models, especially those CNNs achieved several human competitive results [2]. CNNs models are used in a variety of application in medical field specially in Computer Aided Diagnosis (CAD) systems for diseases prediction and classification such as: Skin cancer classification using photographic images [3], Pneumonia and COVID-19 prediction [4], [5], [6], [7], Diabetic retinopathy detection on OCT images of the retina [8], [9], [10], [11], brain tumor detection using high performance computing [12], [13], [14] and CNN [15], heart disease prediction [16], [17], [18], [19], [20] Parkinson's disease detection [21], [22], etc.

Despite that CAD high diagnostic performances [23], recent advances in adversarial examples have revealed numbers of security concerns [24], [25], mainly the CAD systems are usually fragile to adversarial attacks [26]. For example, simple perturbation of the input image makes little difference to the human eye, but it can mislead the CNNs models to have opposite conclusions. Furthermore, a common problem

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

during the data acquisition phase is image noise, which can implicitly form an adversarial attack. For example, particle contamination in dermoscopy and endoscopy, as well as metal/respiratory artifacts lenses of CT scans, significantly decreases the quality of acquired imagery.

Although neural networks are locally unstable with the respect to small perturbation producing considerably distinct output [27], [28], DL models are vulnerable to adversarial attack. Nevertheless, largely focused investigations on adversarial attacks were on non-medical images, while such attacks on medical images are relatively unknown [29], [30].

In this work, the focus has been put on MobileNet CNN model and its susceptibility on adversarial attack. First, the vulnerability of the model against FGSM attacks is investigated. Second, based on the investigation results, a novel approach to train model is proposed, using adversarial neighbors by leveraging structured signals in addition to feature inputs. The signal structure is implicitly induced by adversarial perturbation by taking a small amount of carefully designed perturbation. Based on the reverse gradient direction and applying that perturbation to the original sample, a structure connecting the sample with its adversarial neighbors is obtained. On the other hand, the attention model has become an integral part of DL, leading to impressive performance in image classification of DR [31], [32] and captioning. Furthermore, attention performance has been improved by MHA [33] which appeals for the ability to receive knowledge from various representation subspaces at multiple locations simultaneously. In this paper, the proposed model incorporates MHA, which execute attention functions on distinct representation subspaces of the input sequence. As a result, multiple attention heads can collect different aspects of the input. [34], incorporate MHA mechanisms with structure signals boost the performance of the proposed models and show promising result.

The primary contributions of this work are summarized as follows:

- Literature review and comparative study on existing methods to reduce vulnerability of CNN models against adversarial attacks.
- The evaluation and analysis vulnerability of MobileNet model in regard to adversarial attacks on DR images.
- The proposing of a novel defensive model (NSL-MHA-CNN) against adversarial attacks with NSL and MHA, which preserves the DR accurate prediction results.
- The examination of the novel methodology and its evaluation with different state-of-the-art techniques and verifying its effectiveness.

The remainder of the article is structured as follows:

Section II reviews related work in adversarial examples, Section III present the material methods used for base model, in Section IV, we provide an explication of our experiment attack with different attack scenario and description of proposed methods to defend against adversarial vulnerability. Section V contains some results and discussion. Section VI summaries this paper and provide further perspective.

TABLE 1. Summarized accuracies in the paper (Joel *et al.*).

Model	Attack	CT (%)	Mammogram (%)	MRI (%)
VGG16	No Attack	76.33	75.40	86.33
	FGSM (0.004)	47.33	26.19	41.54
VGG16+ Adv training	No Attack	73.67	70.37	86.33
	FGSM (0.004)	73.13	51.85	41.54

II. RELATED WORK

This section includes a review of existing work on adversarial attack methods on medical and non-medical images, meanwhile, with an inspection of important medical image analysis tasks where adversarial attacks present a major vulnerability and security challenge.

The authors in [35] studied the vulnerability of the VGG16 model for medical images including CT scans, Mammography and MRI images and for non-medical images using MNIST and CIFAR-10 Dataset. First, the paper examined the sensitivity of the VGG16 model across three methods FGSM, PGD and BIM attack with different perturbations in order to maximize classifier error, while minimizing the perturbation. The study shows that the medical image was more susceptible to adversarial perturbation thus accuracy of model on CT, Mammogram et MRI images drop using FGSM Attack with maximum perturbation size of 0.004. The study introduced some approach to defend the vulnerability by using adversarial training, with the same FGSM configuration accuracy of model on CT, Mammogram et MRI images; nevertheless, the study concluded that the approach has limited effectiveness against adversarial attacks on medical images. The detailed accuracies reached in the paper are presented in Table 1.

In the black box assault scenario, more recent works [36] suggest a genetic method for creating adversarial samples without access to the model's weights. The study aims to investigate the vulnerability of several ML models such as CNN, MPL, SVM and others to adversarial examples, using Keras Library with MNIST Data Set and DEAP library to implement genetic algorithms. The experiment revealed that most of models either deep or shallow are affected by vulnerability to adversarial attacks, that are likely to be shared by several other models.

In the paper [37], the authors investigate the vulnerability of different DNN models based on three medical images classification skin cancer, referable diabetic retinopathy and pneumonia classification, using various models' architecture VGG16, VGG19, ResNet50, Inception ResNetV2, DenseNet 121, and DenseNet 169 with universal adversarial perturbation with and without target attack. In order to evaluate the vulnerability of DNN, the inception v3 model is used for Skin

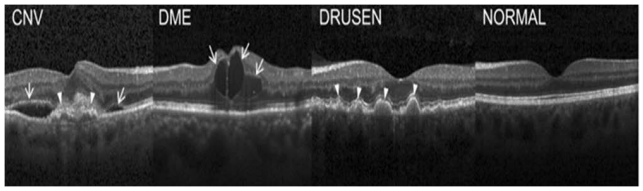

FIGURE 1. Four DR classes.

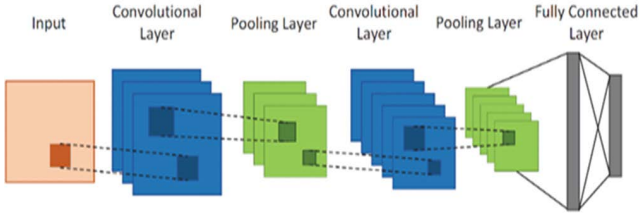

FIGURE 2. Different component of convolution neural network.

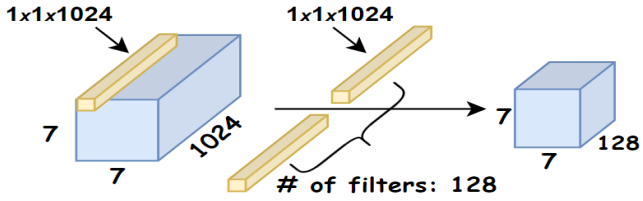

FIGURE 3. Overview of network in network with 1 × 1 × 1024 convolution filter size.

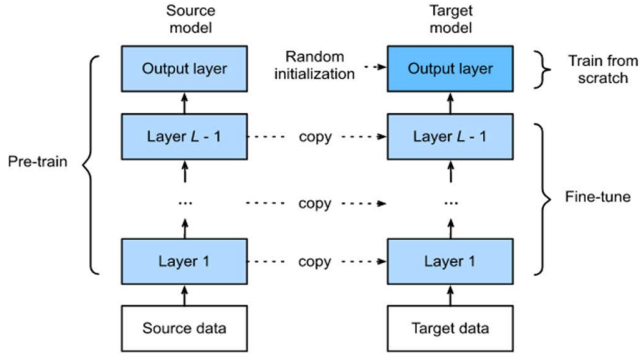

FIGURE 4. Fine tuning model steps.

lesion, OCT, Chest X-ray. The Fooling rate (Rf) and Success rate (Rs) metric are used with perturbation (p) equal to 2.

To improve the resilience of the DNN against any attack, the paper proposed adversarial retraining with fine tuning. However, the impact was limited to non-target UAPs, and while the target UAP's vulnerability was mitigated, it was not completely avoided. Unfortunately, retraining the adversarial requires expensive calculations.

The authors in [38] present a medical adversarial attack approach based on three oncology medical images including diabetic retinopathy grading, artifact detection, and lung segmentation. To classify fundus images, they used ResNet-50 integrated with graph convolutional network and Unet for segmentation. With the aim to improve the attack

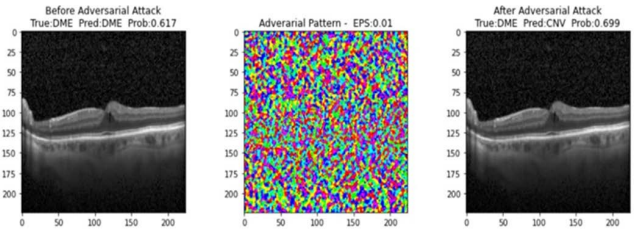

FIGURE 5. Example of adversarial pattern with epsilon = 0.01.

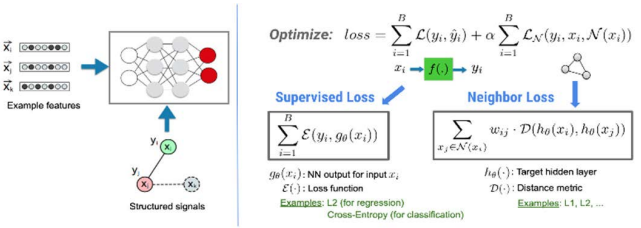

FIGURE 6. Neural structure learning formula.

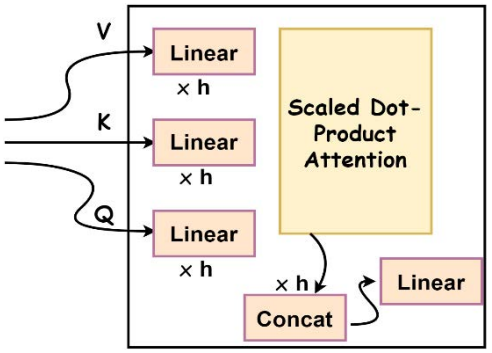

FIGURE 7. Multi-head attention mechanism architecture.

performance of the deviation loss, the suggested Stabilized Medical Image Attack (SMIA) incorporates a stabilization loss term.

Many research addressed the security of DL models however, few studies have poorly evaluated the strength of CNN models used in medical tasks. For that reason, this present paper proposes NSL-MHA-CNN approach and evaluates it in terms of complexity, light weight and accuracy against adversarial attack.

III. MATERIALS AND METHODS

This section will showcase several strategies and datasets used to accomplish diabetic retinopathy classification using a CNN model in the setting of an adversarial attack.

A. DATASET

There are 207.130 OCT images divided into two sets with 42823 training images were collected from 4.686 patients and 1.000 testing images were collected from 633 patients (250 images from each category) [4]. Fig.1 demonstrates various DR recorded in OCT images and classified.

B. CONVOLUTIONAL NEURAL NETWORK (CNN)

Convolutional Neural Networks (CNNs) are similar to regular Neural Networks in that they are most typically used to

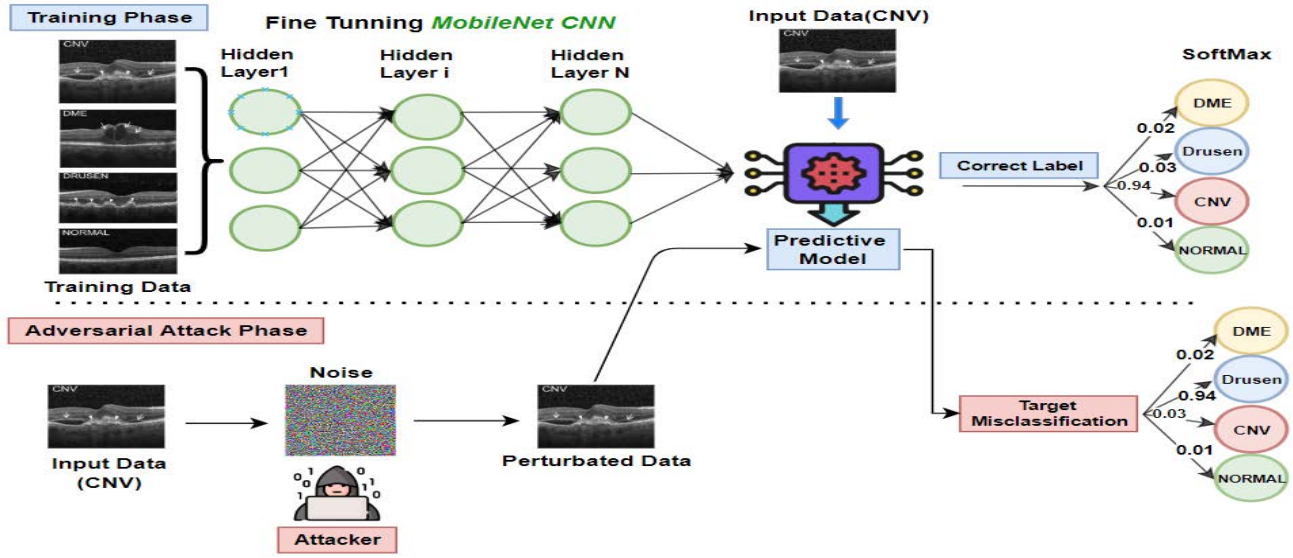

FIGURE 8. Overview of attack architecture.

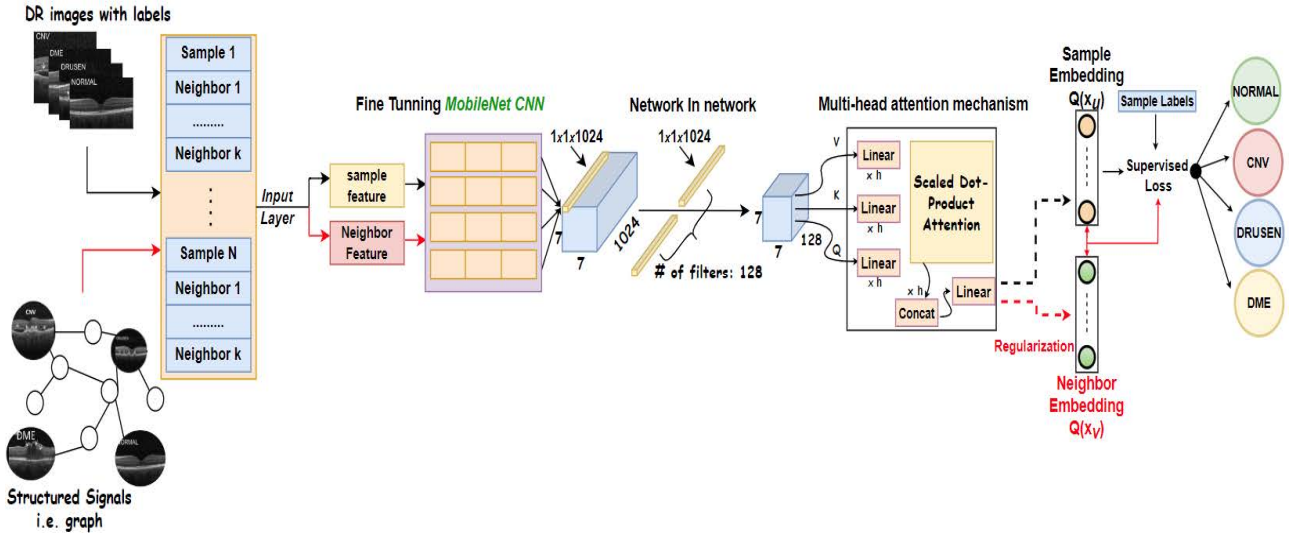

FIGURE 9. Different component of proposed architecture.

analyze visual vision such as handwriting [28] and classification [29] with specific convolution and pooling operations for automated feature recognition and extraction.

C. NETWORK IN NETWORK

Increasing number of parameters and feature maps in depth CNN models resulting in a performance reduction, to address this issue, a Network-in-Network approach presented in [30] to minimize dimensionality and the number of feature mappings.

Inception is one of the network architectures which employs this technique [31].

D. FINE TUNING

In general, fine-tuning means making minor changes to a process in the interest to get the high performance with desire output [32]. Training a neural network from scratch is time and resource consuming; with insufficient data the fine-tuning method can be an effective solution, thus, most of the data can be integrated from previous models. With fine-tuning, it is possible to provide ease of transferring knowledge by not restricted to retraining the classifier stage (the fully connected layers), but retrain also the feature extraction stage (the convolutional and pooling layers).

E. MobileNet ARCHITECTURE

To minimize the number of parameters in CNN models built for mobile and embedded vision applications, depth wise convolution and pointwise convolution are used to construct the MobileNet CNN.

F. FAST GRADIENT SIGN METHOD (FGSM)

The Fast Gradient Sign Method (FGSM) is a simple algorithm to generate adversarial images proposed by [16] in

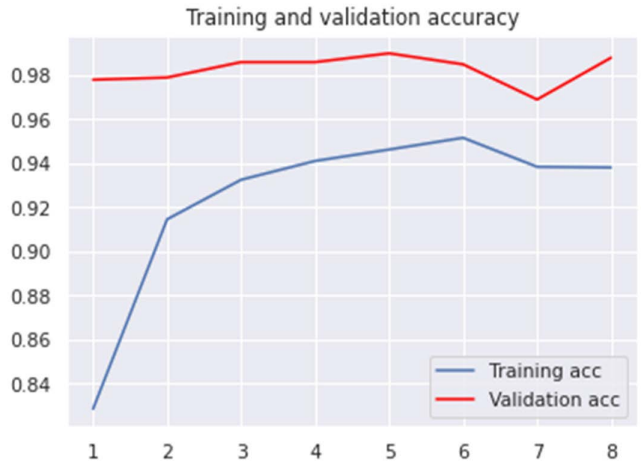

FIGURE 10. Base model training and validation accuracy.

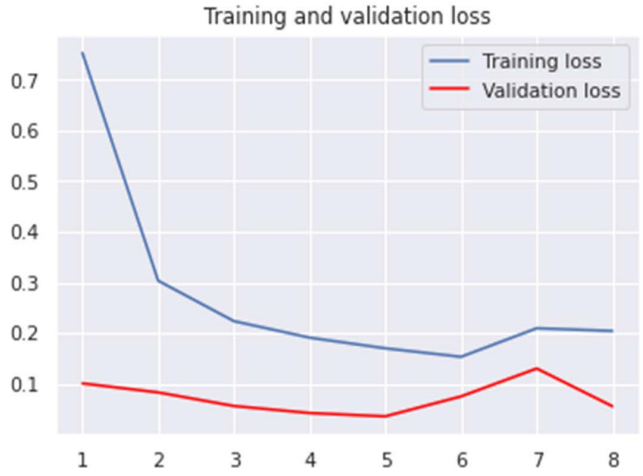

FIGURE 11. Base model training and validation loss.

TABLE 2. Performance metrics reached in the paper (Hirano *et al.*).

Attack	Metrics	Skin Lesion (%)	OCT (%)	Chest X-ray (%)
No Attack	Accuracy	87.7	95.5	97.6
Non target uap	Rf	92.2	70.2	81.7
Target uap	Rs	NV: 93.3 MEL: 94.4	NM: 84.1 CNV: 95.9	N: 96.1 Pneumonia: 93.3

TABLE 3. Summarized performance reached in the work Of (Qi, Gong *et al.*) under different perturbations.

Epsilon	Metrics	Clean (%)	FGSM (%)	SMIA (%)
0.005	iou	22.86	12.03	11.34
	mAp	43.36	26.65	15.92
0.01	iou	22.86	6.82	4.8
	mAp	43.36	17.28	3.47
0.015	iou	22.86	5.27	1.18
	mAp	43.36	13.52	3.25

TABLE 4. Percentage of parameters distribution in different layer.

Layer	Mult-Adds	Parameters
Conv 1 x 1	94.86%	74.59%
Conv dw 3 x 3	3.06%	1.06%
Conv 3 x 3	1.19%	0.02%
Fully Connected	0.18%	24.33%

TABLE 5. Confusion matrices for classification.

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

TABLE 6. Different base model performance metrics.

Class	Accuracy	Precision	Recall	Support
CNV	0.99	0.98	1	250
DME	0.99	1	1	250
DRUSEN	0.98	0.98	0.98	250
NORMAL	0.98	1	0.98	250

their paper to enhance a neural network robustness against input perturbation. The goal is to determine the amount that each pixel in the image contributes to the loss value and add appropriate perturbations in order to create a new image called adversarial image which maximizes the loss. The fast gradient sign method will be given by:

$$x^{adv} = x + \varepsilon \times sign(\nabla_x J(\theta, x, y)) \quad (1)$$

where:

- x^{adv} : Adversarial image.
- x : Original input image.
- y : Original input label.
- ε : Multiplier to ensure the perturbations are small.
- θ : Model parameters.
- J : Loss.

G. NEURAL STRUCTURE LEARNING NSL

The consistent development achieved in the field of computer vision has resulted in some spectacular accomplishments across several disciplines. Despite these incredible accomplishments, multiple studies have shown how sensitive these models are, to even imperceivable small changes during collections of input data as result of camera misalignment, vibration or out of sample example that can mislead the models.

In order to overcome those problems and generally improve models against corrupted and perturbed data, a form of neural structure learning called adversarial regularization has been adopted [38]. Forming structure learning dynamically is done by creating adversarial neighbors that

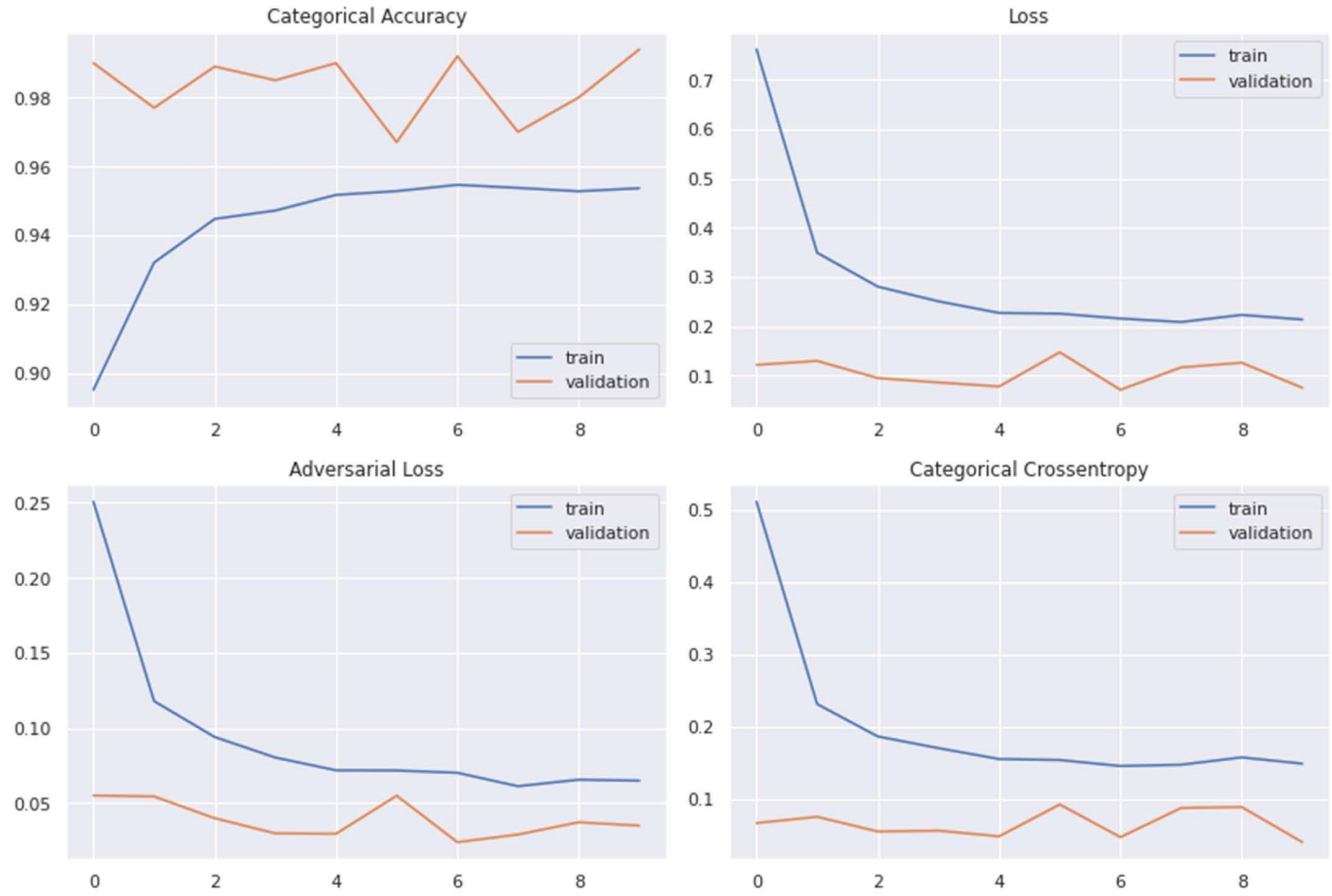

FIGURE 12. Training and validation of NSL-MHA-CNN after 9 epochs.

represent the similarity among all the images. The adversarial neighbors are generated [39] by taking a small amount of designed perturbation based on the reverse gradient direction and applying that perturbation to the original sample.

After the adversarial neighbor sample is generated, an edge is added to connect the sample with its adversarial neighbor to dynamically construct the structure in order to be used in the neural network. The neural network learns to maintain a structure by keeping the similarity between a sample and its neighbors and won't be confused by the small perturbation.

H. MULTI-HEAD ATTENTION MECHANISM

Attention mechanism has recently contributed to impressive result in the area of deep learning, which were initially developed in end to end machine translation applications [40], [41] using recurrent neural networks (RNN), image captioning and speech recognition tasks.

It is a powerful method that can assist models in achieving better classification result by selecting essential features. Recently, attention performance has been further improved by multi-head mechanism [33] according to current studies show that MHA is more effective than the single attention function [42], for its capacity to mutually receive information from several representation subspaces in an effort to obtain strong feature representation.

IV. PROPOSED ARCHITECTURE

The suggested attack architecture for both base model and NSL-MHA-CNN model are described in this section, as well as the suggested defensive architecture for enhancing security and defense against adversarial attacks.

A. ATTACK ARCHITECTURE USED IN OUR EXPERIMENTS

Fig 8 depicts an overview of the experiment attack adopted for this paper, in which the vulnerability of machine learning systems is highlighted from the following perspectives.

1. **Training phase** where the base model is used to create a model with parameters by learning to generate the appropriate output from the training data.
2. **Attack phase** that reviews the threat and attack faced by our base model system as shown in Fig 8 adversarial example attack that occur during the test phase using fast gradient sign method applied to our test data set.

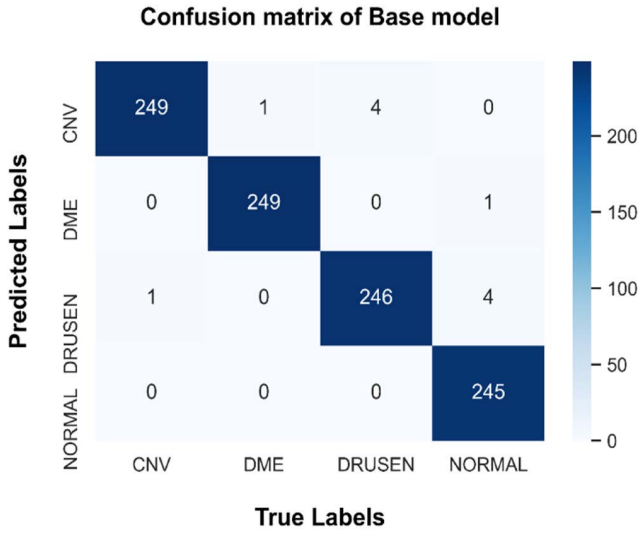

FIGURE 13. Confusion matrix of base model on testing DR dataset.

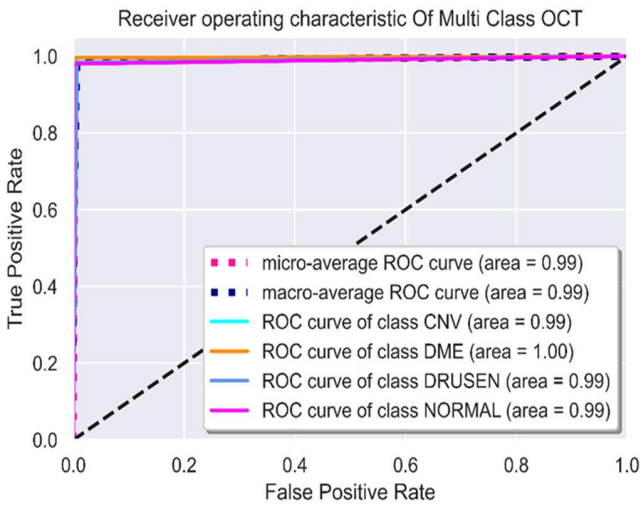

FIGURE 14. Receiver operating characteristic performance of base model.

TABLE 7. NSL-MHA-CNN performance under different metrics.

Class	Accuracy	Precision	Recall	Support
CNV	0.98	1	0.98	250
DME	1	0.99	1	250
DRUSEN	0.99	0.99	1	250
NORMAL	1	1	1	250

B. AN OVERALL VIEW OF THE PROPOSED ARCHITECTURE

In this section, the approach to defense against adversarial attack is presented, by detailing different component for robust DR prediction.

Fine tuning (FT) is a process that aims to update the weights of a previously model and correctly retrained to a specific task, which is prediction of DR. Moreover, this CNN architecture denoted NSL-MHA-CNN model employed for DR classification is a novel approach based on

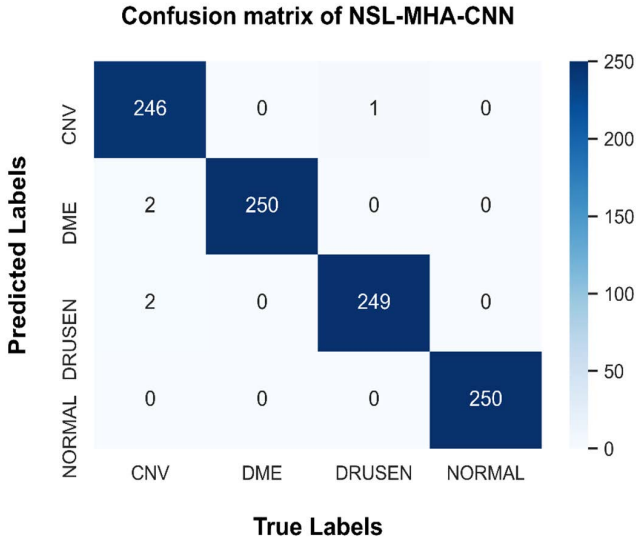

FIGURE 15. Confusion matrix of NSL-MHA-CNN on testing DR Dataset.

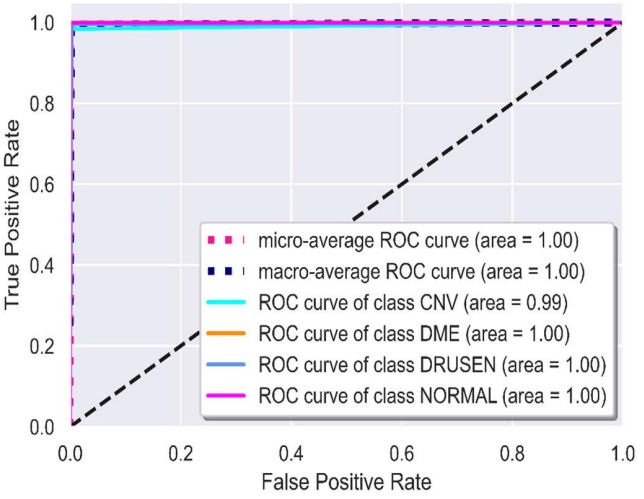

FIGURE 16. NSL-MHA-CNN Receiver Operating Characteristic performance.

the modification of MobileNet architecture with fine-tuning combined with a neural structure signal and Multi Head attention mechanism.

The last fully connected layer was restituted with a Multi-Head Attention, combined with end-to-end training with neural structure learning called Adversarial regularization by forming structure learning dynamically by creating adversarial neighbors. In this manner, 6.555.478 trainable parameters in the NSL-MHA-CNN model have been achieved. Fig 9 describes the architecture with various component and their workflow. fully connected layer was restituted with a Multi-Head Attention, combined with end-to-end training with neural structure learning called Adversarial regularization by forming structure learning dynamically by creating adversarial neighbors. In this manner, 6.555.478 trainable parameters in the NSL-MHA-CNN model have been

TABLE 8. Evaluation metrics of base model under attack.

Class	Accuracy	Precision	Recall	Support
CNV	0.98	0.76	0.98	250
DME	0.94	0.88	0.94	250
DRUSEN	0.72	0.80	0.72	250
NORMAL	0.74	0.99	0.74	250

TABLE 9. Base model average performance under different perturbation.

Perturbation(ε)	Accuracy	Precision	Recall	F1-score
0	0.99	0.99	0.99	0.99
0.01	0.84	0.85	0.84	0.84
0.03	0.78	0.80	0.77	0.77
0.05	0.76	0.78	0.76	0.75
0.1	0.71	0.74	0.71	0.70
0.3	0.73	0.76	0.73	0.73
0.5	0.71	0.73	0.70	0.70

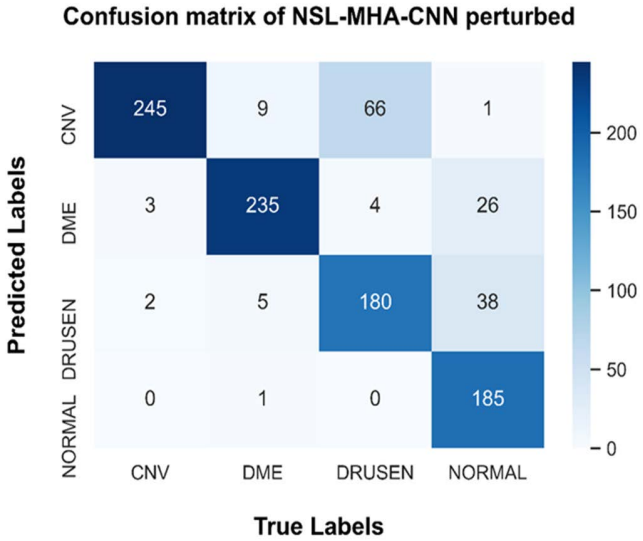

FIGURE 17. Confusion matrix of base model with epsilon perturbation = 0.01.

achieved. Fig 9 describes the architecture with various component and their workflow.

V. RESULTS AND DISCUSSION

A. METRICS FOR PERFORMANCE EVALUATION

There has been an identification of the measures most frequently used to assess CNN performance: the mean clues included in the evaluation are accuracy, precision, recall additionally to confusion matrix and ROC (Receiver Operating Curve) to provide a helpful evaluation of the model's classification performance.

1) ROC (RECEIVER OPERATING CURVE)

The ability to distinguish between classes is one of the statistical measures used to evaluate model performance. The Area Under Curve (AUC) of an optimal classifier is close to 1, if it is close to 0.5, the result is equivalent to random guessing [43].

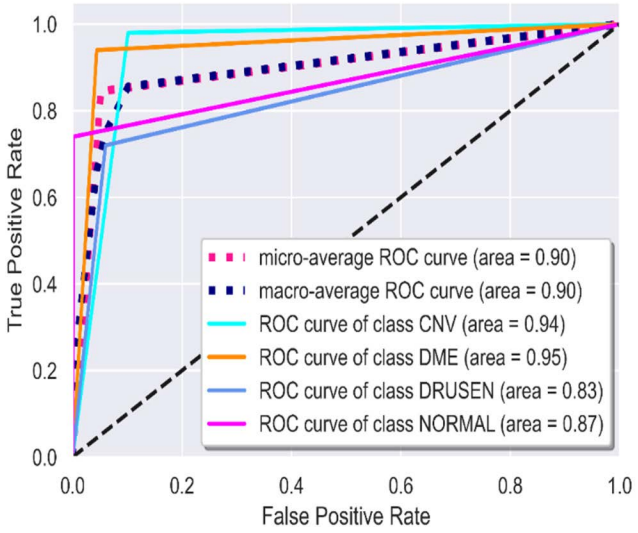

FIGURE 18. Receiver operating characteristic performance of base model with epsilon perturbation = 0.01.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{FP + TP} \tag{3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

2) CONFUSION MATRIX

Confusion matrix is a well-known metric of visualizing the performance of prediction model used while solving binary as well as for multiclass classification problems. It presents very simple, yet efficient performance measures [44]

B. EXPERIMENTAL RESULTS

1) BASE MODEL TRAINING

After training the base model as shown in Fig 10 and Fig 11, the model reaches optimal performance within 8 epochs.

2) NSL-MHA-CNN TRAINING

After training our NSL-MHA-CNN architecture with end-to-end training with Adversarial loss as shown in Fig12.

3) PERFORMANCE EVALUATION WITHOUT ATTACK

a: BASE MODEL

The performance evaluation of the base model is described in Section 3.1, and in order to present more about the model's efficiency and performance, the testing scores are presented in Table 6, the confusion matrix and roc curve are used.

The Accuracy achieved by base model reach 99%, and as shown in Fig 14, the DME class obtains the highest ROC (area = 1).

Fig.14 presents evaluation results with regard to Receiver Operating Characteristic performance for the testing DR dataset.

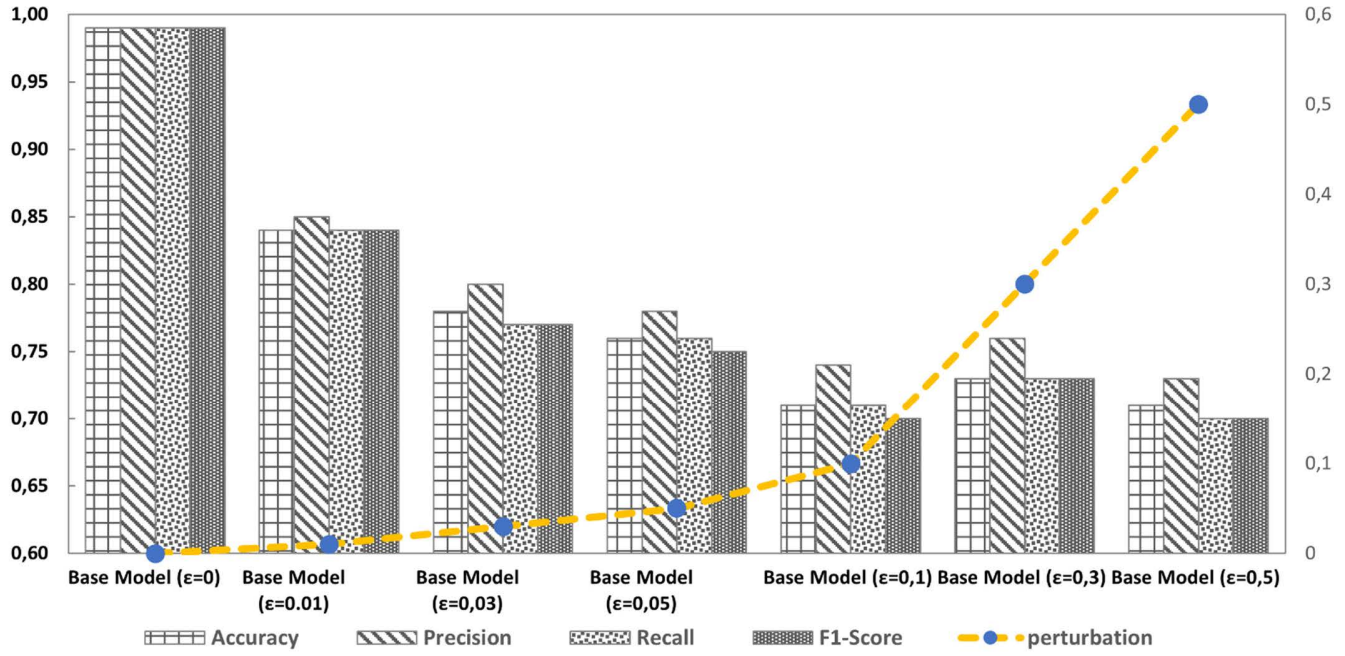

FIGURE 19. Different metric of base model performance under attack.

TABLE 10. Evaluation metrics of NSL-MHA-CNN under attack.

Class	Accuracy	Precision	Recall	Support
CNV	0.96	0.98	0.97	250
DME	1	0.95	0.98	250
DRUSEN	0.96	0.99	0.98	250
NORMAL	1	1	1	250

b: NSL-MHA-CNN MODEL

The NSL-MHA-CNN model was tested by using the testing set with the same configuration as previous, Table 7 summarize the performance metrics.

Fig. 15 shows the results of four class predictions on the confusion matrix of the test DR dataset.

NSL-MHA-CNN attained 99% accuracy, and as shown in Fig.16 the classes DME, DRUSEN and NORMAL achieved 100% area under curve.

4) PERFORMANCE EVALUATION WITH ATTACK

a: BASE MODEL

In this part, the base model vulnerability is highlighted and the way this model may easily be misled by a FGSM attack.

1) **Example of attack with perturbation epsilon** ($\varepsilon = 0.01$)
2) **Overview of attack with different perturbation**

The performance of base model was investigated according to different perturbation; in order to emphasize the vulnerability of base model to adversarial attack, the performance metric was calculated by taking the average performance of four classes CNV, DME, DRUSEN and NORMAL.

TABLE 11. NSL-MHA-CNN model average performance under different perturbation.

Perturbation(ε)	Accuracy	Precision	Recall	F1-score
0	0.99	0.99	0.99	0.99
0.01	0.98	0.98	0.98	0.98
0.03	0.98	0.98	0.98	0.98
0.05	0.98	0.98	0.98	0.98
0.1	0.98	0.98	0.98	0.99
0.3	0.96	0.96	0.96	0.96
0.5	0.92	0.92	0.92	0.92

TABLE 12. Models' performance in terms of accuracy and runtime are compared.

Models	Accuracy (%)	Runtime	Time/Epoch	Total Parameters
Base Model	98	7 h 12 min 56 s	54 min 7 s	6.574.404
NSL-MHA-CNN	99	18 h 10 min 8 s	1h 47 min 30 min	6.577.622

From Table 9 and Fig. 19, there can be deduced that as the perturbation increases, the performance consistently decreases. The drop in accuracy and F1-score from 0.99, to 0.84 respectively with only small perturbation epsilon = 0.01; this performance drop indicates the effectiveness of the advanced approach in this paper to highlight the vulnerability of base model.

b: NSL-MHA-CNN MODEL

1) **Example of attack with perturbation epsilon** ($\varepsilon = 0.01$)

TABLE 13. Compared result with some related work.

Works	Methods	Database	Attacks Applied	Accuracy (%)
[35]	VGG16 + Adv Training	CT	FGSM ($\epsilon = 0.004$)	73.13
[45]	ResNet + Adversarial Training with Feature Scattering	CIFAR10	White-box Attack ($\epsilon = 8$)	78.4
[46]	Darknet53 + Mixed Adversarial Training (MAT)	DR	FGSM	DR1: 99.83 DR2: 74.94 DR3: 97.09
[47]	Model F+Defense-GAN-Rec	MNIST	FGSM ($\epsilon = 0.10$)	98.64 ± 0.0011
[37]	Inception V3 +Adv retraining	OCT	Non target uap (p=2)	CNV: 97 DRUSEN: 93 DME: 100 NORMAL: 58
Proposed Model	**NSL-MHA-CNN**	OCT	FGSM ($\epsilon = 0.01$)	CNV: 96 DRUSEN: 96 DME: 100 NORMAL: 100

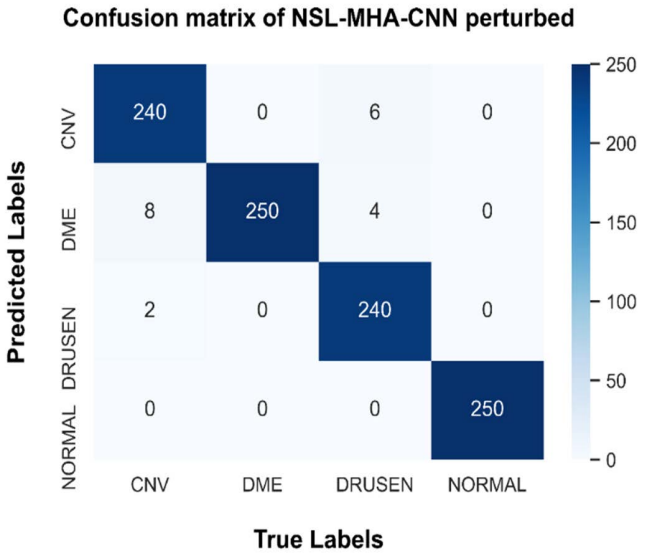

FIGURE 20. Confusion matrix of NSL-MHA-CNN model with epsilon perturbation = 0.0.

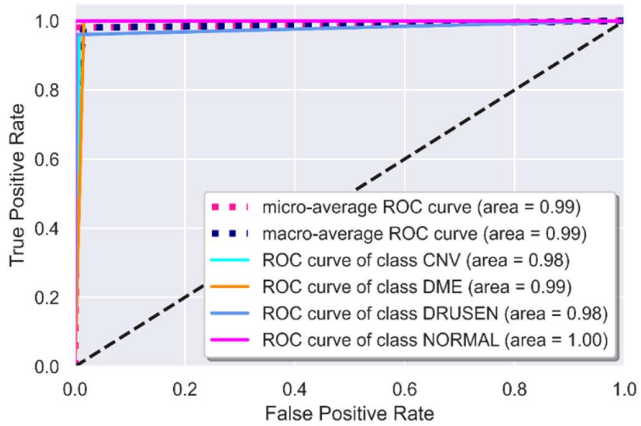

FIGURE 21. Receiver operating characteristic performance with epsilon perturbation = 0.01.

2) Overview of the attack with different perturbations values

Based on this view, it can be observed that incorporating NSL with MHA can help improve the robustness of CNN model against adversarial attack.

In particular, it can also be noticed that even if perturbation values increase, the performance metrics are still higher, and that may result in improved generalization, as evidenced by increased test set accuracy.

5) RESULTS COMPARISON WITH BASE MODEL

Last but not least, Fig 23 demonstrates the advancement in terms of cutting-edge DR detection, a comparative performance between base model and proposed model. According to the experiments, the suggested approach in this paper shows promising result against adversarial attack and get stable performance with different epsilon perturbation.

6) THE EXECUTION TIME RESULTS

Training a CNN on a large DR image is a challenging and expensive task that can take many hours to days to complete. Both the quality of the training data and the choice of the algorithm are central to the model training phase. Table 12 presents the runtimes and time by epoch for both base model and our NSL-MHA-CNN model.

As shown in Table 12, the training time of NSL-MHA-CNN increased and that due to Adam optimizer, that optimize each independent variable in the objective function in the case of NSL four variables: Accuracy, Loss, Adversarial loss

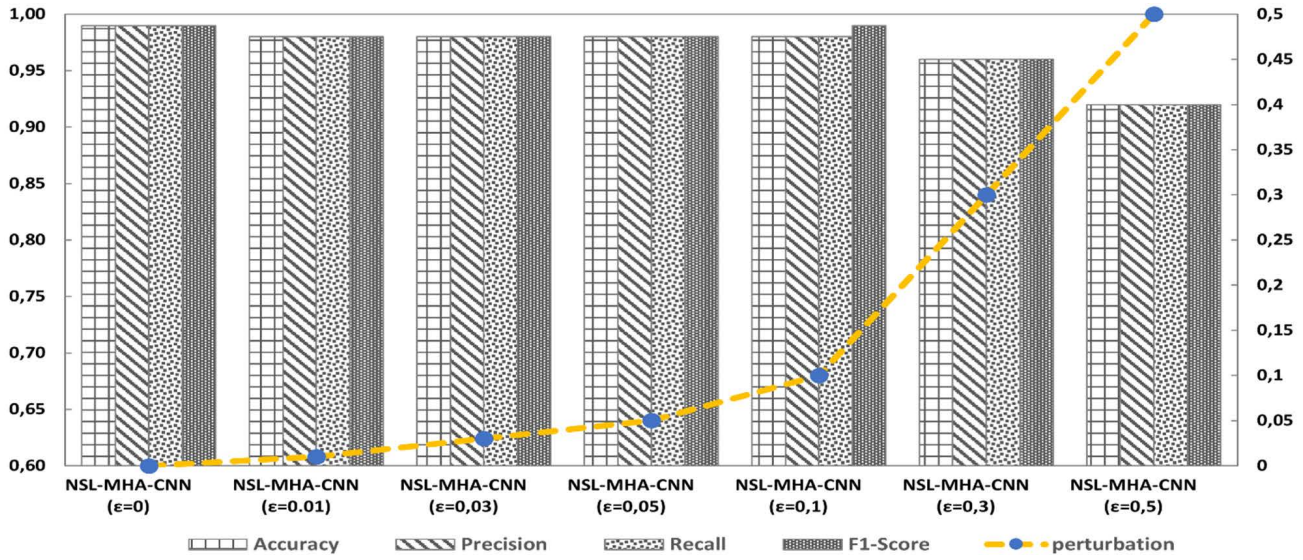

FIGURE 22. Different metric of NSL-MHA-CNN performance under attack.

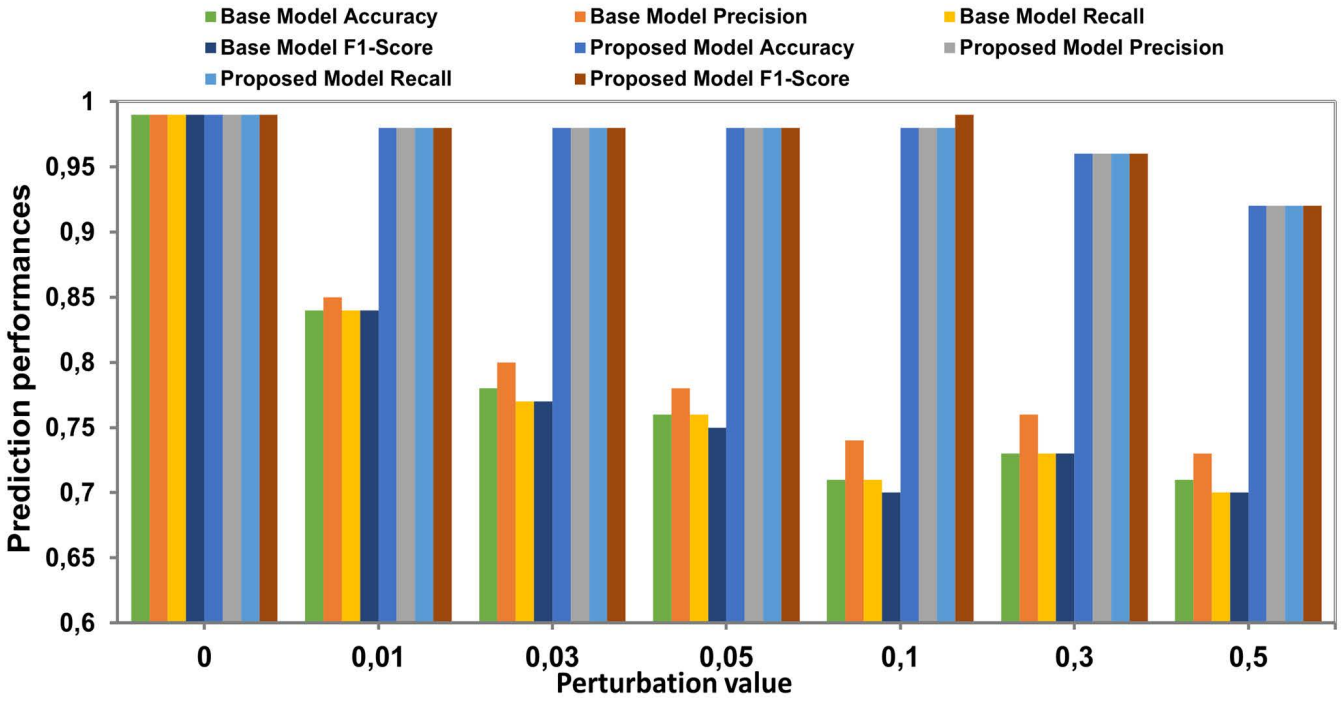

FIGURE 23. NSL-MHA-CNN model performance under different perturbation.

and Cross Entropy- instead of two variables in base model accuracy and loss which have its own learning rate.

7) DISCUSSION

A Novel CAD system for robust diabetes retinopathy detection based on NSL and MHA approach proposed in this research. The goal is to investigate the vulnerability of base model especially in medical images task by using different attack perturbation. Based on those investigations, an NSL-MHA-CNN model as well as a comparative performance conducted with the base model are proposed.

In order to emphasize this study, a further comparison with state of the art work are presented in Table 13.

Despite the fact that the results of this paper seem promising, there have been many challenges such as lack of studies about vulnerability in medical CNN specially in DR prediction as well as computation complexity of training different models with Cloud GPUs has expensive cost. Further

investigation of the NSL-MHA-CNN model in the IOT environment will be considered for future works.

VI. CONCLUSION AND PERSPECTIVES

Machine Learning security is one of the most-debated academic areas, as it continues to generate a number of security concerns. Hence, it worked as a motive to evaluate and analyze vulnerability of MobileNet model in regard to adversarial attacks on DR images. Considering that the base model cannot defense against adversarial attack, the experiments presented in the paper show that indiscernible degrees of perturbation $\varepsilon < 0.01$ were sufficient to cause a task failure resulting to misclassification in majority of the time.

This paper proposes a novel NSL-MHA-CNN DR classification model by introducing neural structure learning with multi head attention, the strategy is to fine-tune Mobile-Net with multi head attention with end-to-end neural structure learning. The FGSM Attack is used to demonstrate the efficacy of the suggested solution against adversarial attack. The proposed models show promising results by achieving 98% accuracy with 0.05 epsilon perturbation. With this proposed novel approach, it is possible to maintain the model performance on adversarial attack without increasing cost of training. Hopefully, this work will give a complete technique for constructing safe, resilient and private CNN systems.

REFERENCES

[1] F. Shaheen, B. Verma, and M. Asafuddoula, "Impact of automatic feature extraction in deep learning architecture," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2016, pp. 1–8, doi: 10.1109/DICTA.2016.7797053.

[2] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[3] 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, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[4] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, and J. Dong, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.

[5] S. Hamida, O. El Gannour, B. Cherradi, A. Raihani, H. Moujahid, and H. Ouajji, "A novel COVID-19 diagnosis support system using the stacking approach and transfer learning technique on chest X-ray images," *J. Healthcare Eng.*, vol. 2021, pp. 1–17, Nov. 2021, doi: 10.1155/2021/9437538.

[6] O. El Gannour, S. Hamida, B. Cherradi, M. Al-Sarem, A. Raihani, F. Saeed, and M. Hadwan, "Concatenation of pre-trained convolutional neural networks for enhanced COVID-19 screening using transfer learning technique," *Electronics*, vol. 11, no. 1, p. 103, Dec. 2021, doi: 10.3390/electronics11010103.

[7] H. Moujahid, B. Cherradi, M. Al-Sarem, L. Bahatti, A. B. A. M. Y. Eljialy, A. Alsaeedi, and F. Saeed, "Combining CNN and grad-cam for COVID-19 disease prediction and visual explanation," *Intell. Autom. Soft Comput.*, vol. 32, no. 2, pp. 723–745, 2022, doi: 10.32604/iasc.2022.022179.

[8] V. Srinadh, B. Maram, and V. Gampala, "Prediction of retinopathy in diabetic affected persons using deep learning algorithms," in *Proc. 6th Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2022, pp. 1285–1291, doi: 10.1109/ICOEI53556.2022.9777193.

[9] O. Daanouni, B. Cherradi, and A. Tmiri, "Diabetes diseases prediction using supervised machine learning and neighbourhood components analysis," in *Proc. 3rd Int. Conf. Netw., Inf. Syst. Secur.*, New York, NY, USA, Mar. 2020, pp. 1–5, doi: 10.1145/3386723.3387887.

[10] O. Daanouni, B. Cherradi, and A. Tmiri, "Predicting diabetes diseases using mixed data and supervised machine learning algorithms," in *Proc. 4th Int. Conf. Smart City Appl.*, 2019, p. 85.

[11] C. Harshitha, A. Asha, J. L. S. Pushkala, and R. N. S. Anogini, "Predicting the stages of diabetic retinopathy using deep learning," in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Jan. 2021, pp. 1–6, doi: 10.1109/ICICT50816.2021.9358801.

[12] N. Ali, B. Cherradi, A. Abbassi, O. Bouattane, and Y. Mohamed, "GPU fuzzy c-means algorithm implementations: Performance analysis on medical image segmentation," *Multimedia Tools Appl.*, vol. 77, pp. 1–23, Aug. 2018, doi: 10.1007/s11042-017-5589-6.

[13] O. Bouattane, B. Cherradi, M. Youssfi, and M. O. Bensalah, "Parallel C-means algorithm for image segmentation on a reconfigurable mesh computer," *Parallel Comput.*, vol. 37, nos. 4–5, pp. 230–243, Apr. 2011, doi: 10.1016/j.parco.2011.03.001.

[14] N. A. Ali, A. E. Abbassi, and B. Cherradi, "The performances of iterative type-2 fuzzy C-mean on GPU for image segmentation," *J. Supercomput.*, vol. 78, no. 2, pp. 1583–1601, Feb. 2022, doi: 10.1007/s11227-021-03928-9.

[15] A. Naseer, T. Yasir, A. Azhar, T. Shakeel, and K. Zafar, "Computer-aided brain tumor diagnosis: Performance evaluation of deep learner CNN using augmented brain MRI," *Int. J. Biomed. Imag.*, vol. 2021, pp. 1–11, Jun. 2021, doi: 10.1155/2021/5513500.

[16] O. Terrada, B. Cherradi, A. Raihani, and O. Bouattane, "A novel medical diagnosis support system for predicting patients with atherosclerosis diseases," *Informat. Med. Unlocked*, vol. 21, 2020, Art. no. 100483, doi: 10.1016/j.imu.2020.100483.

[17] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[18] O. Terrada, B. Cherradi, S. Hamida, A. Raihani, H. Moujahid, and O. Bouattane, "Prediction of patients with heart disease using artificial neural network and adaptive boosting techniques," in *Proc. 3rd Int. Conf. Adv. Commun. Technol. Netw. (CommNet)*, 2020, p. 6, doi: 10.1109/CommNet49926.2020.9199620.

[19] O. Terrada, B. Cherradi, A. Raihani, and O. Bouattane, "Atherosclerosis disease prediction using supervised machine learning techniques," in *Proc. 1st Int. Conf. Innov. Res. Appl. Sci., Eng. Technol. (IRASET)*, Apr. 2020, pp. 1–5, doi: 10.1109/IRASET48871.2020.9092082.

[20] B. Cherradi, O. Terrada, A. Ouhmida, S. Hamida, A. Raihani, and O. Bouattane, "Computer-aided diagnosis system for early prediction of atherosclerosis using machine learning and K-fold cross-validation," in *Proc. Int. Congr. Adv. Technol. Eng. (ICOTEN)*, Jul. 2021, pp. 1–9, doi: 10.1109/ICOTEN52080.2021.9493524.

[21] W. Wang, J. Lee, F. Harrou, and Y. Sun, "Early detection of Parkinson's disease using deep learning and machine learning," *IEEE Access*, vol. 8, pp. 147635–147646, 2020, doi: 10.1109/ACCESS.2020.3016062.

[22] O. Asmae, R. Abdelhadi, C. Bouchaib, S. Sara, and K. Tajeddine, "Parkinson's disease identification using KNN and ANN algorithms based on voice disorder," in *Proc. 1st Int. Conf. Innov. Res. Appl. Sci., Eng. Technol. (IRASET)*, Meknes, Morocco, Apr. 2020, pp. 1–6, doi: 10.1109/IRASET48871.2020.9092228.

[23] M. U. Emon, R. Zannat, T. Khatun, M. Rahman, M. S. Keya, and E. Ohidujjaman, "Performance analysis of diabetic retinopathy prediction using machine learning models," in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Jan. 2021, pp. 1048–1052, doi: 10.1109/ICICT50816.2021.9358612.

[24] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.*, vol. 2, no. 6, pp. 305–311, Jun. 2020, doi: 10.1038/s42256-020-0186-1.

[25] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[26] M. Xue, C. Yuan, H. Wu, Y. Zhang, and W. Liu, "Machine learning security: Threats, countermeasures, and evaluations," *IEEE Access*, vol. 8, pp. 74720–74742, 2020, doi: 10.1109/ACCESS.2020.2987435.

[27] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, vol. 307, pp. 195–204, Sep. 2018, doi: 10.1016/j.neucom.2018.04.027.
[28] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.
[29] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020, doi: 10.1016/j.eng.2019.12.012.
[30] T. Dai, Y. Feng, B. Chen, J. Lu, and S.-T. Xia, "Deep image prior based defense against adversarial examples," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108249, doi: 10.1016/j.patcog.2021.108249.
[31] O. Daanouni, B. Cherradi, and A. Tmiri, "Self-attention mechanism for diabetic retinopathy detection," in *Emerging Trends in ICT for Sustainable Development*. Cham, Switzerland: Springer, 2021, pp. 79–88, doi: 10.1007/978-3-030-53440-0_10.
[32] O. Daanouni, B. Cherradi, and A. Tmiri, "Automatic detection of diabetic retinopathy using custom CNN and Grad-CAM," in *Advances on Smart and Soft Computing*. Singapore: Springer, 2021, pp. 15–26, doi: 10.1007/978-981-15-6048-4_2.
[33] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10. Accessed: Nov. 8, 2021. [Online]. Available: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
[34] A. Raganato and J. Tiedemann, "An analysis of encoder representations in transformer-based machine translation," in *Proc. EMNLP Workshop BlackboxNLP, Analyzing Interpreting Neural Netw. NLP*, 2018, pp. 287–297, doi: 10.18653/v1/W18-5431.
[35] M. Z. Joel, S. Umrao, E. Chang, R. Choi, D. Yang, J. Duncan, A. Omuro, R. Herbst, H. Krumholz, and S. Aneja, "Adversarial attack vulnerability of deep learning models for oncologic images," *MedRxiv*, pp. 1–23, Feb. 2021. [Online]. Available: https://www.medrxiv.org/content/10.1101/2021.01.17.21249704v3, doi: 10.1101/2021.01.17.21249704.
[36] P. Vidnerová and R. Neruda, "Vulnerability of classifiers to evolutionary generated adversarial examples," *Neural Netw.*, vol. 127, pp. 168–181, Jul. 2020, doi: 10.1016/j.neunet.2020.04.015.
[37] H. Hirano, A. Minagi, and K. Takemoto, "Universal adversarial attacks on deep neural networks for medical image classification," *BMC Med. Imag.*, vol. 21, no. 1, Jan. 2021, doi: 10.1186/s12880-020-00530-y.
[38] G. Qi, L. Gong, Y. Song, K. Ma, and Y. Zheng, "Stabilized medical image attacks," 2021, *arXiv:2103.05232*.
[39] T. D. Bui, S. Ravi, and V. Ramavajjala, "Neural graph learning: Training neural networks using graphs," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Feb. 2018, pp. 64–71, doi: 10.1145/3159652.3159731.
[40] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
[41] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[42] J. Wang, X. Peng, and Y. Qiao, "Cascade multi-head attention networks for action recognition," *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102898, doi: 10.1016/j.cviu.2019.102898.
[43] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Caspian J. Internal Med.*, vol. 4, no. 2, pp. 627–635, 2013.
[44] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "Multiclass confusion matrix reduction method and its application on net promoter score classification problem," *Technologies*, vol. 9, no. 4, p. 81, Nov. 2021, doi: 10.3390/technologies9040081.
[45] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," 2019, *arXiv:1907.10764*.
[46] S. Lal, S. U. Rehman, J. H. Shah, T. Meraj, H. T. Rauf, R. Damaševičius, M. A. Mohammed, and K. H. Abdulkareem, "Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition," *Sensors*, vol. 21, no. 11, p. 3922, Jun. 2021, doi: 10.3390/s21113922.
[47] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," 2018, *arXiv:1805.06605*.

OTHMANE DAANOUNI was born in El Jadida, Morocco, in 1991. He received the B.Sc. degree in software development and the M.Sc. degree in information systems engineering from the Faculty of Science, Chouaib Doukkali University, in 2014 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the LAROSRI Laboratory, Faculty of Science. His current research interests include machine learning, computer vision and medical image analysis, and machine learning vulnerability. He works actually as a Senior Executive for transformation digital of the public administration at interior ministry of Morocco.

BOUCHAIB CHERRADI was born in El Jadida, Morocco, in 1970. He received the B.S. degree in electronics and the M.S. degree in applied electronics from the ENSET Institute of Mohammedia, Morocco, in 1990 and 1994, respectively, the D.E.S.A. Diploma degree in instrumentation of measure and control (IMC) from Chouaib Doukkali University, El Jadida, in 2004, and the Ph.D. degree in electronics and image processing from the Faculty of Science and Technology, Mohammedia. His research interests include applications of massively parallel architectures, cluster analysis, pattern recognition, image processing, fuzzy logic systems, artificial intelligence, machine learning, and deep learning in medical and educational data analysis. He currently works as an Associate Professor at the CRMEF-El Jadida. In addition, he is also an Associate Research Member of the Signals, Distributed Systems and Artificial Intelligence (SSDIA) Laboratory, ENSET Mohammedia, Hassan II University of Casablanca (UH2C), and LaROSERI Laboratory, on leave from the Faculty of Science El Jadida, Chouaib Doukali University. He is a supervisor of several Ph.D. students.

AMAL TMIRI was born in Azemmour, Morocco, in 1972. She received the B.S. degree in physiques, in 1995, the D.E.S.A. Diploma degree in static physiques from the Faculty of Science UM5, Rabat, Morocco, in 1999, and the Ph.D. degree in physiques from the Faculty of Science, El Jadida, Morocco. Her research interests include applications of computer vision, pattern recognition, image processing, and machine learning. She works actually as an Associate Professor in computer science at ENSAM, UM5, Rabat. In addition, she is also an Associate Research Member of the LaROSERI Laboratory, on leave from the Faculty of Science El Jadida Morocco, Chouaib Doukali University, Morocco. She is a supervisor of several Ph.D. students. She is also the President of IEEE GLOBAL 5G-IOT Summit, organized in Marrakech, Morocco, during the WINCOM'18 Conference, and a member of the Organizing Committee of the 19th IEEE Mediterranean Electronic Conference (IEEE MELECON 18), Marrakech, in 2018.