IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 1, 2021, accepted January 16, 2021, date of publication January 19, 2021, date of current version January 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052870

Automatic Diabetic Retinopathy Grading System Based on Detecting Multiple Retinal Lesions

EMAN ABDELMAKSOUD[1], SHAKER EL-SAPPAGH[2,3], SHERIF BARAKAT[1], TAMER ABUHMED[4], AND MOHAMMED ELMOGY[5], (Senior Member, IEEE)

[1]Information Systems Department, Faculty of Computers and Information, Mansoura University, Mansoura 35516, Egypt
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
[3]Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Banha 13518, Egypt
[4]College of Computing, Sungkyunkwan University, Suwon 16419, South Korea
[5]Information Technology Department, Faculty of Computers and Information, Mansoura University, Mansoura 35516, Egypt

Corresponding author: Tamer Abuhmed (tamer@skku.edu)

This work was supported in part by the National Research Foundation of Korea grant funded by the Korean Government, Ministry of Science and ICT, under Grant NRF-2016R1D1A1A03934816, and in part by the Chowis Company Ltd.

ABSTRACT Multi-label classification (MLC) is considered an essential research subject in the computer vision field, principally in medical image analysis. For this merit, we derive benefits from MLC to diagnose multiple grades of diabetic retinopathy (DR) from various colored fundus images, especially from multi-label (ML) datasets. Therefore, ophthalmologists can detect early signs of DR as well as various grades to initiate appropriate treatment and avoid DR complications. In this paper, we propose a comprehensive ML computer-aided diagnosis (CAD) system based on deep learning technique. The proposed system's main contribution is to detect and analyze various pathological changes accompanying DR development in the retina without injecting the patient with dye or making expensive scans. The proposed ML-CAD system visualizes the different pathological changes and diagnoses the DR grades for the ophthalmologists. First, we eliminate noise, enhance quality, and standardize the sizes of the retinal images. Second, we differentiated between the healthy and DR cases by calculating the gray level run length matrix average in four different directions. The system automatically extracts the four changes: exudates, microaneurysms, hemorrhages, and blood vessels by utilizing a deep learning technique (U-Net). Next, we extract six features, which are the gray level co-occurrence matrix, areas of the four segmenting pathology variations, and the bifurcation points count of the blood vessels. Finally, the resulting features were afforded to an ML support vector machine (SVM) based on a classifier chain to differentiate the various DR grades. We utilized eight benchmark datasets (four of them are considered ML) and six different performance evaluation metrics to evaluate the proposed system's performance. It achieved 95.1%, 91.9%, 86.1%, 86.8%, 84.7%, 86.2% for accuracy, area under the curve, sensitivity, specificity, positive predictive value, and dice similarity coefficient, respectively. The experiments show encouraging results as compared with other systems.

INDEX TERMS Multi-label computer-aided diagnosis (ML-CAD), multi-label classification (MLC), deep learning (DL), U-Net, diabetic retinopathy (DR).

I. INTRODUCTION

Diabetes is a chronic disease characterized by blood glucose level elevation. This elevation leads over time to severe damage of the human blood vessels (BV), eyes, and nerves [1]. Diabetic retinopathy (DR) is mostly one of the common complications of diabetes. It is a progressive disease that can cause permanent blindness without warning [2]. By 2040, the studies estimate that diabetes will affect about 642 million adults overall the world, while DR affects one from every three people with diabetes [3]. Another study ensures that by 2030, the number of people with DR will grow to 191 million [4].

The main characteristics and signs of the DR are microaneurysms (MA), hemorrhages (HM), exudates (EX), venous loops (VL), venous reduplication (VR), and neovascularization (NV). The occurrence of one/two or all of these features in the retina determines the DR stages [5]. In the initial

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

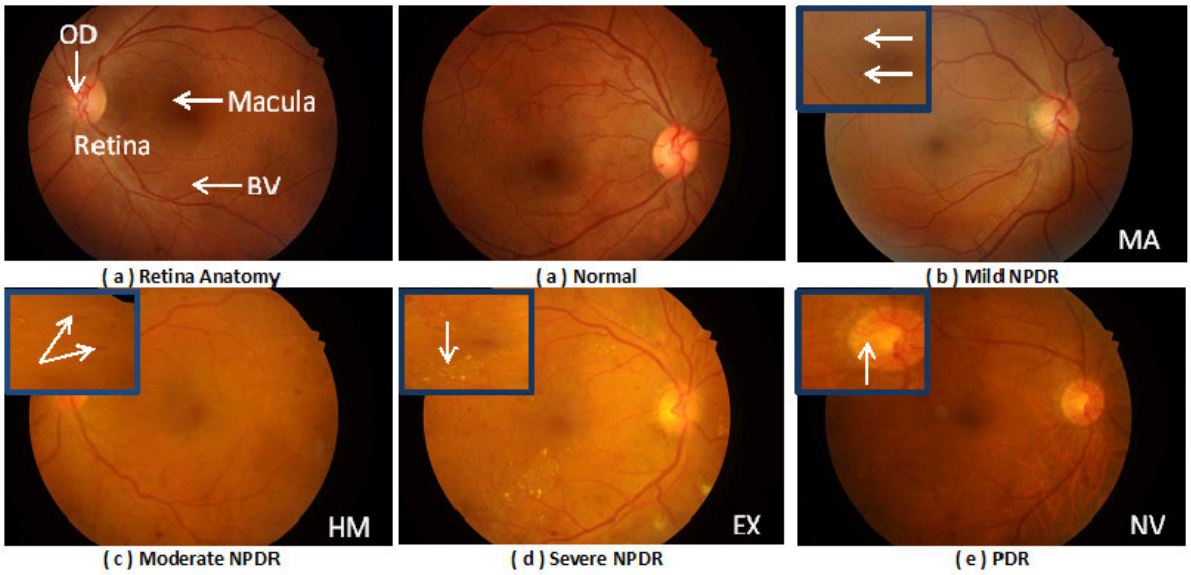

FIGURE 1. The various grades of DR from IDRiD dataset: (a) Normal retina anatomy, (b) Normal case, (c) Mild, (d) Moderate, (e) Severe (NPDR), and PDR.

DR grades, patients are generally without notable symptoms, but, in advance, patients may suffer from symptoms that include distortion, blurred vision, and progressive visual severity loss. Therefore, DR grades can be categorized into non-proliferative DR (NPDR) and proliferative DR (PDR). On the other hand, NPDR can be branched to subgrades, which are mild, moderate, and severe. Mild is indicated by appearing small MA, whereas moderate reflects appearing HM and/or EX. The severe NPDR reflects increasing in retinal ischemia by appearing small, abnormal, and weak BV, which are called NV. This severe grade is setting the stage for the PDR. Fig. 1 shows the various DR grades by annotating the retina anatomy and pathological changes, such as HM and EX. Besides, NVs increase the area of BV and cause ischemia in the retina. Fig. 1 shows some examples of different DR grades in fundus images.

HM appear similar to MA if they are small. On the contrary, MA appears similar to HM if it is large on wide BV. The physician can distinguish between the two signs in the clinic by injecting the patients with fluorescein dye. The MA, in this case, takes the same white color as the BV but HM not. Another solution is that the patient pays for an OCTA scan, which may be centered only on the retina. The third pathological change sign is EX. It is resulted from the breakdown of the blood-retina barrier, allowing leakage of serum proteins and lipids from the BV. On the other hand, NV is the PDR main mark. It often occurs near the optic disc (OD). It is called NV of the disc (NVD). When NV occurs within three disc diameters of the major BV, it is called NV elsewhere (NVE).

There are multiple ocular imaging modalities used to depict the retina to help ophthalmologists detect ocular diseases. Previously, fluorescein angiography (FA) based on dye used in detecting retina vascular diseases. Then, fundus autofluorescence (FAF) became commonly used in macular degeneration and pattern dystrophies. By 1990, optical coherence tomography (OCT) was invented to visualize the retina's layers. It was developed mainly to detect macular diseases and choroidal NV. OCT's subjective and insensitive to the small retinal thickness and macular breaks [6]. Recently, OCT angiography (OCTA) is developed, which depends on motion contrast from the blood flow. In OCTA, there is no need for dye injection. It is safe and non-invasive, but it works in a small field of view (FOV) and unable to show leakage. After this brief review, we can conclude that color fundus photography remains the most applicable imaging modality, especially in DR. The main merits depict the retinal BV, OD, macula, and vascular abnormalities. It allows an objective comparison of the retina and optic nerve appearance [7].

It is noteworthy that deep learning (DL) has a vitally important role in detecting DR and its different stages. In the last few years, many systems classified the DR to different stages from mild to PDR by utilizing DL based on the color fundus imaging modality, such as [8]. DL's idea was inspired by brain neuronal connectivity. This connectivity enables the brain to process large amounts of data. Besides, it extracts meaningful patterns depending on bygone experiences with the same inputs. Moreover, DL is able to model data at various abstractions. Deep convolutional neural networks (CNN) has been at the forepart of DL. Recently, CNN has achieved great success in many real-life applications [9], [10], especially in medical image analysis and multi-label (ML) classification [11], [12]. In detecting DR and its grades, the ophthalmologists and developers face many challenges and problems. They can be summarized in the following points:

- DR detection is accomplished by involving a well-trained physician.

- The manual retina's structural changes and BV abnormalities detection may be inconsistent and time-consuming, it depends on physician's experience [13].
- Previous automated systems [14] were developed to solve such problems based on the hand-crafted features tools. These tools are sensitive to the contrast of fundus images. Besides,there are noise, artifacts, and illumination in fundus images.
- There is feature similarity between the eye anatomies and DR lesions. For example, HM takes the same color as BV. On the other side, it may be like an MA if it is small. EX takes the same color and shape as OD.
- Feature Extraction and segmentation steps are the workload and burden the developers.
- Recent improvements in biomedical image analysis are based on DL, which could be exploited to enhance the Computer-aided Diagnosis systems (CAD)s' performance. Moreover, many DL models fall into overfitting.
- A deep fine-tuned CNNs are very useful in medical image analysis, and even outperform the fully trained CNNs, especially in limited training set [15].

All of these challenges motivated us to present a novel ML-CAD system based on conventional and DL techniques to automatically detect DR grades accurately by utilizing different ML color fundus images. The proposed system starts with the preprocessing phase, in which the system removes noise, enhances contrast, and resizes the color fundus images to a standard size. In the binary classification phase, we differentiate the healthy from the DR grades by extracting 11 descriptors of the gray level run length matrix (GLRLM) in four directions and feeding the feature vector to the support vector machine (SVM) classifier. To visualize the DR signs for the ophthalmologists, we made the segmentation phase for the DR case images. We resulted in four segmenting images (BV, EX, MA, and HM) for each DR image by the customized U-Net DL model. Then, the bifurcation points (BP) are extracted and counted from the BV network image. After that, we extracted six features from the four segmented DR images. Finally, we utilized the multi-label SVM (MLSVM). MLSVM is SVM based on the classifier chain (CC)) to diagnose the four DR grades (mild NPDR, moderate NPDR, severe NPDR, and PDR). We validated each of these phases to ensure robustness and accuracy. The proposed ML-CAD system improved the accuracy of classifying the DR grades from eight various benchmark datasets.

The proposed methodology comprises a series of contributions, which are listed as follows:

- The problems of low quality, contrast enhancement, and various resolutions and sizes of the utilized datasets were solved.
- We present a comprehensive system that used DL and conventional methods to classify healthy and DR grades.
- We utilized a customized, robust, and automated method for segmenting the four pathological variations (BV, EX, MA, and HM) rather than using many supervised segmentation methods for each sign's detection.
- The ophthalmologists are provided with four accurate segmenting images of main pathological DR signs and the overall diagnosing results. The other systems concentrated on one or two signs' segmentation or making direct diagnosing without visualizing the different pathological variations.
- The proposed system extracts seven various essential features from the segmenting pathological variations.
- We utilized ML classification (MLC) by MLSVM based on problem transformation to diagnose the different DR grades. The utilized ML classifier provides the flexibility to future grading based on other lesion detection.
- The proposed ML-CAD system was applied on eight benchmark datasets with different cameras' settings, various patients (children, adults, men, women, and elderly people), and different noise, quality, and illumination levels.
- We validated the proposed ML-CAD system by comparing it with other systems. Moreover, we utilized six different performance measures.

We organized the rest of the manuscript into six sections. Section II presents the background of the current literature' reviews. It also focuses on recent studies' limitations and how the proposed ML-CAD systems tame these limitations. Section III presents the proposed ML-CAD system framework phases. Section IV describes the conducted experiments. Section V presents the discussion and the analytical comparison between the proposed ML-CAD system and others. Finally, Section VI concludes our work and highlights our future research directions.

II. RELATED WORK

Many researchers have worked on DR detection and diagnosis by utilizing retinal fundus images. For instance, Brian *et al.* [16] differentiated HM from EX to detect DR. They first detected OD from the green channel. To improve the image contrasts, they utilized the contrast limited adaptive histogram equalization (CLAHE). Thereafter, the authors segmented the EX by combining CLAHE with Gabor filters, followed by thresholding. They adopted the circular Hough transform (CHT) approach, followed by thresholding to extract HM. The authors segmented only EX and HM signs without diagnosing the DR grade.

Atlas and Parasuraman [17] extracted gray level co-occurrence matrix (GLCM), GLRLM, and speeded up robust features (SURF). They classified normal and DR images. They utilized the adaptive neuro-fuzzy inference system (ANFIS) to extract HM. The authors only segmented HM as a sign of DR, but the DR grade cannot be diagnosed using HM alone. Orlando *et al.* [18] applied the SVM technique to extract BV. According to the related distance of pixels, they weighed the pairwise interactions. They utilized a 2D Gabor filter and unary potentials of line detector to standardize all the images. However, the authors ignored

the merits of automated segmentation, which can affect the results negatively.

Fadafen *et al.* [19] extracted EX by morphological after excluding OD. The authors utilized edge and feature-based detection to detect BV through brightness, width, and direction.Their results were dependent on the human visual system, which are sensitive to intensity and directions. Although EX is considered a strong sign for DR detection, the authors could not classify the DR grades. Moreover, they did not utilize a contrast enhancement technique. Safitri and Juniati [20] diagnosed normal and DR. The authors enhanced the contrast by CLAHE. They segmented the BV by thresholding and the matched filter. Finally, they utilized the box-counting technique for fractal dimension and k-nearest neighbor (KNN) for classification. However, the results were dependent on the fractal dimension values. Their performance measures did not sufficient in ML imbalanced dataset. Abdelmaksoud *et al.* [21] classified the healthy and the DR grades by extracting EX, MA, HM, and BV. They utilized matched filter with a first-order Gaussian derivative filter and some morphological operations. They extracted the GLCM, areas of lesions, and BP counts. Finally, they utilized MLSVM classifier. It isn't easy to extract many signs from the fundus images using conventional methods. It burdens the developer, especially in large datasets. Therefore, it is crucial to utilize DL methods.

Recently, several researchers have focused their attention on detecting DR grades based on DL to save the effort of extracting and selecting features by handcrafted feature-based methods, such as Abr 'amoff al. [22] evaluated the analysis software of the IDx-DR device. The device takes OD, macula centered images for each eye. It outputs the grade of the DR. It depends on AlexNet and Oxford Visual Geometry. The authors modified the device's system to be applied on public datasets. But, they did not capable of detecting PDR separately from macular edema (ME) as well as the IDx-DR device. Moreover, the diagnostic drift in differentiating between HM and MA affected detecting mild and moderate grades. Bellemo *et al.* [23] made a combination of two CNNs: an adapted visual geometry group network (VGGNet) and a residual neural network (ResNet) to classify the images based on gradient-descent (GD). The two models' probability output scores were summed, and then, they made the final classification by thresholding the output scores due to the sensitivity (SEN) and specificity (SPE). Their model gave higher accuracy but required more computational time.

Unlike [22] and [23], Mansour [24] utilized AlexNet to extract BV features. He utilized connected component analysis (CCA) for feature extraction and selection. The AlexNet was 5- convolution (CONV) layers and two- fully connected (FC) layers. Finally, he utilized an SVM classifier to classify the DR classes. The author segmented only BV from the images and ignored other lesions, such as HM, MA, and EX that are essential in detecting mild and moderate cases.Gadekallu *et al.* [25] utilized principal components analysis (PCA) and dimensional reduction by firefly. The authors made normalization by using the Standardscalar. Their model was being overfitted when it was implemented on a small dataset. Hagos *et al.* [26] utilized the pre-trained Inception-V3 model to classify the DR into two classes: normal and abnormal. The authors made the preprocessing by cropping and resizing the images. They utilized a softmax classifier, stochastic GD (SGD) optimizer. They assigned the learning rate (lr) to be 5×10^{-4} and utilized the cosine loss function. The authors just classified the presence/absence of DR, while different grades of DR need to be differentiated. The same idea of hagos *et al.* [26] was proved in Tymchenko *et al.* [27] work. They made some augmentation processes such as horizontal and vertical flipping, transposing, and rotation. The authors utilized EfficientNet models based on pretrained ImgeNet. Although they utilized more than one dataset to validate the model, it is not enough to validate their real-life model. Therefore, they intended to utilize Shapley Additive exPlanations (SHAP) method in the future to visualize features that give the physician the assessment ability of the stages.

Xu *et al.* [28] introduced a system for detecting only DR's presence/absence. They presented A CNN model with 10- CNN layers. After each two CNN layers, they inserted one max pooling (MP) layer then the FC layers. They utilized the SGD optimizer and the softmax classifier. They cared about the preprocessing by doing data augmentations, but they utilized a small dataset.Pratt *et al.* [29] presented a CNN model that included 10-layers of CNN, three FC layers, and classification to five classes by the softmax classifier. They used the rectified linear unit (ReLU) as an activation function. Besides, they used batch normalization (BN) and used MP to occur after each CONV layer. Unlike [28], the authors ignored the preprocessing stage in their proposed system. In fact, noise affected their classification results. Moreover, it is necessary to utilize more than one dataset to achieve reality and robustness.

Butt *et al.* [30] built CNN models like [28] and [29], but the difference in their work was that they built three CNN modules based on RGB channels. They separated the RGB fundus images to R, G, and B and supplied each one in a distinct model. The authors concluded that the second model with the B channel gave better accuracy than the models with R and G channels.Li *et al.* [31] utilized fractional MP in CNN to detect five classes of DR. The authors processed the images by rescaling and clipping. They utilized an SVM classifier with a modified recognition rate. Although the authors built two CNN models with different layers to get different feature spaces and combine the best predictions by SVM classifier, they need sufficient and balanced groups of images, such as [29]. They failed to predict classes 3 and 4 accurately and hardly differentiated class 0 from class 1 in testing new data.

From the previous review of the current literature utilized conventional methods and DL architectures, we can conclude their main limitations in diagnosing DR grades from color fundus images as follows:

- Most studies focused on detecting the DR presence/absence and ignored the DR grades. On the other hand, the studies which focused only on segmenting the DR signs, satisfied with segmenting only one or two of DR pathological variations (EX, HM, BV, and MA).
- Some studies proposed the DR grades diagnosis. These models were conservative, and they were not applicable in the real world because of the insufficient and imbalanced datasets. Besides, they fall into overfitting.
- A lot of state-of-the-art systems diagnosed the DR grades without segmenting and visualizing the different variations of DR for the ophthalmologists.
- Most studies ignored preprocessing steps, while the noise and low contrast affect the segmentation and classification accuracy.

To conquer these restrictions and obstacles, we present a comprehensive ML-CAD system. It mainly depends on the problem transformation MLC. It means that the system transforms the problem into sub-problems. The number of derived problems is the same as the class labels number. MLC idea depends on label correlations, which can result in unprecedented labels from the existing labels. The proposed system evicts noise by using the median filter. On the other hand, it boosts the contrast and handles the illumination problem by histogram equalization for brightness preservation based on a dynamic stretching technique (HEBPDS) [32]. The system makes the preprocessing steps without losing the images' features. It segments 4 DR signs (HM, EX, BV, and MA) from various colored fundus images by utilizing a customized U-Net model.

The segmentation is a significant phase for the DR grades diagnosis for the developer and the ophthalmologists. It is essential to visualize the main DR four pathological variations for the ocular specialists. It helps the ophthalmologists and lessens the burden of the patient. The ophthalmologist can observe the BV network with NV and MA without injecting the patient with dye or paying for an expensive scan. It provides them with HM and EX and diagnoses the disease grade to do the right treatment in time. The system combines conventional and DL methods to get benefits from them in diagnosing DR grades accurately.

Seven different important features are extracted: GLRLM, GLCM, regions of interest (ROIs) areas, and BP of BV from the four segmenting images. MLSVM classifier is used to classify the five various DR grades: normal, mild NPDR, moderate NPDR, severe NPDR, and PDR. Our system can be applied to various color fundus images with different cameras' settings, qualities, noise, illumination on different patients. Eventually, we validated our ML-CAD system by making many different rapprochements with other systems and methods. We build the proposed framework to segment and diagnose the DR grades based on our system in Abdelmaksoud *et al.* [21]. In the proposed framework, we extend and promote our previous system [21]. We can summarize the difference in the following five points. First, we made a binary classification that depends on hand-crafted feature extraction. This phase is used to distinguish healthy from DR cases. Second, we added some post-processing steps to prepare images for the segmentation phase. Third, we segmented the fundus images by utilizing a customized, universal DL U-Net model. In fact, shallow classification models' performance depends on the quality of the features fed into them. On the other hand, classification is mainly based on the accuracy of the segmentation phase. Therefore, we customized the U-Net model by establishing it deeper and customized its hyperparameters to provide precise results. Fourth, we increased the features that we extract to classify the DR grades accurately. We extracted 11 descriptors of the GLRLM on four directions, which are 0^o, 45^o, 90^o, and 135^o for each image. Finally, we evaluated the performance using six different performance metrics and compared it with many current conducted systems and methods.

III. THE PROPOSED ML-CAD SYSTEM

The primary objective of the proposed ML-CAD system is to detect DR, present the four DR signs, and classify the five various healthy and disease grades from different eight RGB fundus datasets (four of them are considered ML) that contain multiple DR lesions. Segmentation of DR signs and the DR grades diagnosis help the ocular specialists observe different disease variations and make the right treatment decisions.

The proposed ML-CAD system utilizes GLRLM to retrieve texture features from four different angles of the preprocessed fundus images. The output features are used to differentiate healthy and DR cases. The system also visualizes the segmenting lesions in four separated images for each entered DR case image. The ML-CAD system also extracts six features from the segmented DR signs. It then selects the most correlated and significant features values to locate each sign's peculiar characteristics deeply. The proposed system depends on MLC problem transformation. Finally, in validation,we utilized six performance measures to validate the proposed ML-CAD system.

Fig 2 shows the architecture of the proposed ML-CAD system, which consists of eight phases. First, the preprocessing phase eliminates noise and enhances images. Second, the feature extraction phase is implemented to retrieve the entered fundus images' main characteristics. Third, the binary classification phase uses the previous feature vector to classify the images into normal and DR cases. This phase is significant as it provides only the DR cases to the next phases to reduce time, memory space, and effort. Then, we make the post-processing phase, which contains three steps. First, resizing all the images and their ground truth (GTs) to be in a standard size of 512×512 and enable validation between the predicting and the segmenting images. Second, creating the mask for each image to be excluded in the segmentation. Third, we utilized the IDRiD dataset to train the U-Net model on the three lesions GT's (MA, EX, and HM). But the IDRiD dataset's GTs are in RGB. Therefore, we have to binarize them. The fifth phase in the proposed framework is

TABLE 1. A summery of some current studies, AUC: Area under curve, ACC: Accuracy, SPE: Specificity, SEN: Sensitivity, and DSC: Dice coefficient.

Study	year	Analysis Type	The Methodology	The dataset	Performance
Biran et al. [16]	2016	Segmenting Ex and HM	Gabor filter and CHT followed by thresholding	DRIVE, STARE	depended on observing the resulting segmenting images without calculating even the similarity with the GT.
Safitri et al. [20]	2017	Segmenting BV and DR detection	box counting and KNN	MESSIDOR	ACC 89.17%
Atlas and Parasuraman [17]	2018	Segmenting HM and DR detection	Region growing. GLCM, GLRLM and SURF for feature extraction. Binary classification by ANFIS	50 (MESSIDOR)	average ACC in HM segmentation 92.56%, ACC of DR detection 63%
Fadafen et al. [19]	2018	EX segmentation	Morphological operations	DIARETDB1	AUC 90.12%
Abràmoff et al. [22]	2016	DR detection	IDx-DR version X2.1.	Messidor-2	SEN 96.8%, SPE 87%, NPV 99%, AUC 98%
Bellemo et al [23]	2019	DR classification	VGG and ResNet	Private	AUC 97.3%, SEN 92.25%, SPE 89.04%
Mansour [24]	2018	DR grading	AlexNet DNN	KAGGLE	ACC 95.26%
Gadekallu et al. [25]	2020	DR classification	DNN	Private	ACC 97%, PRE 96%, SEN 92%, SPE 95%
Hagos et al. [26]	2019	DR detection	Inception-V3	KAGGLE	ACC 90.9%
Tymchenko et al. [27]	2020	DR classification	EfficientNet	APTOS 2019	kappa score 0.925
Xu et al. [28]	2017	DR classification	CNN and XGB	KAGGLE	ACC 94.5%
pratt et al. [29]	2016	DR grading	CNN	DIARETDB0,1	ACC 75%, SEN 95%
Butt et al. [30]	2019	DR grading	CNN	EyePACS	ACC 97.08%
Li et al. [31]	2019	DR grading	DCNN	KAGGLE	ACC 86.17%, SEN 89.30%, SPE 90.8%
Abdelmaksoud et al. [21]	2020	EX, MA, HM, BV segmentation and DR grading	matched filter with first order gaussian derivative, morphological operation and MLSVM	DRIVE, STARE, MESSIDOR and IDRiD	ACC 89.2%, AUC 85.20%, SEN 85.1%, SPE 85.2%, PPV 92.8%, DSC 88.7%

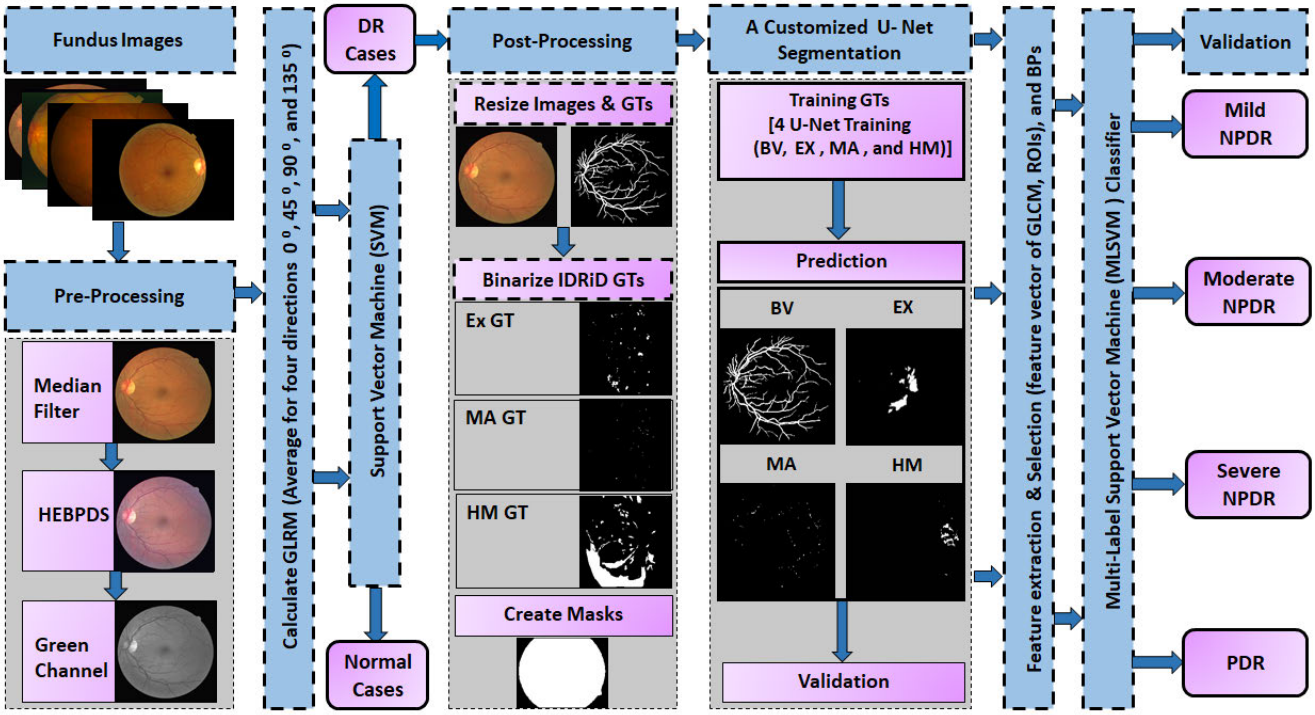

FIGURE 2. The proposed ML-CAD system for detecting and diagnosing healthy and DR grades from color fundus images.

the segmentation by utilizing the customized U-Net model. In this phase, we train the U-Net model on the DRIVE dataset BV's GT to predict the vasculature network of the other seven datasets. By doing the four training, the model produces four segmenting images for each DR case. We validate the resulting segmenting images with their GTs and the experts. After

that, we performed the feature extraction on the ROIs. The resulting feature vector is fed to the classification, the seventh phase of the proposed framework. We utilized the SVM based on CC classifiers, which are called MLSVM, to classify the other four DR grades. Finally, we evaluated binary classification, segmentation, and final MLC performance by utilizing six various performance metrics. We evaluated the overall proposed system by comparing it with other systems and methods. The proposed ML-CAD system phases are demonstrated in detail in the following subsections.

A. PREPROCESSING

This phase is crucial in any medical system as the medical images are characterized by artifacts, noise, and insufficient quality that vary from one modality to the other. In this respect, fundus images suffer from low contrast, illumination, and noise. Therefore, the proposed system includes some steps in the preprocessing phase to enhance the quality and remove the noise. First, we apply the median filter to strip noise [33]. Then, we utilized HEBPDS [32] to enhance the contrast of the fundus images. At the end of the preprocessing phase, we extracted the green channel from the RGB enhanced image to use it in the feature extraction and the binary classification by GLRLM and SVM.

B. GLRLM EXTRACTION

In the first extraction phase of the proposed framework, we utilized GLRLM to extract the features of the green channel of all processed images [34]. The resulting feature vector is used to differentiate normal and DR cases. To illustrate how the GLRLM works, we represented it by (gl, rl, θ), where gl is the gray level, rl is the run length, and θ is the direction angle. It is a way of testing an image across a given direction to find the pixels with the same gray level values. Thus, it gives the homogeneous runs' size for each gray level. Many different GLRLM matrices can be computed for a single image as we utilized 11 matrices of them. Each matrix is calculated for each selected direction of the preprocessed image. Therefore, we calculated 11 GLRLM matrices in four different directions 0º, 45º, 90º, and 135º. We computed these 11 matrices' average in 4 directions to get a single averaged GLRLM matrix for each image. The main GLRLM construction for processed fundus images and the feature vector steps' measurement is shown in two algorithms, which are found in [35]. GLRLM can be represented by Eq. 1.

$$GLRLM(\theta) = g(i,j)|\theta, 0 \le i \le N_{gl}, 0 \le j \le rl_{max} \quad (1)$$

where j is the number of elements, i is the intensity in the direction θ, N_{gl} is the maximum gl, and rl_{max} is the maximum length. We calculated 11 texture feature descriptors, which are short run emphasis (SRE), long run emphasis (LRE), short run low gray-level emphasis (SRLGLE), short run high gray-level emphasis (SRHGLE), long run low gray-level emphasis(LRLGLE), long run high gray-level emphasis (LRHGLE), run percentage (RP), low gray-level run emphasis (LGLRE), high gray-level run emphasis (HGLRE), run length non-uniformity normalized (RLNN), and run length non-uniformity(RLN) [36]. Eqs. 2 − 12 show the computations of the aforementioned 11 feature descriptors.

$$\mathbf{SRE} = \frac{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r}\frac{(g(i,j)|\theta)}{i^2}}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r} g(i,j)|\theta} \quad (2)$$

$$\mathbf{LRE} = \frac{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r}(g(i,j)|\theta)j^2}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r} g(i,j|\theta)} \quad (3)$$

$$\mathbf{SRLGLE} = \frac{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r}\frac{(g(i,j)|\theta)}{i^2\,j^2}}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r} g(i,j)|\theta} \quad (4)$$

$$\mathbf{SRHGLE} = \frac{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r}\frac{(g(i,j)|\theta)i^2}{j^2}}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r} g(i,j)|\theta} \quad (5)$$

$$\mathbf{LRLGLE} = \frac{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r}\frac{(g(i,j)|\theta)j^2}{i^2}}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r} g(i,j)|\theta} \quad (6)$$

$$\mathbf{LRHGLE} = \frac{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r}(g(i,j)|\theta)i^2j^2}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r} g(i,j|\theta)} \quad (7)$$

$$\mathbf{HGLRE} = \frac{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r}(g(i,j)|\theta)i^2}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r} g(i,j|\theta)} \quad (8)$$

$$\mathbf{LGLRE} = \frac{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r}\frac{(g(i,j)|\theta)}{j^2}}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r} g(i,j)|\theta} \quad (9)$$

$$\mathbf{RLN} = \frac{\sum_{j=1}^{N_r}((\sum_{i=1}^{N_l} g(i,j)|\theta))^2}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r} g(i,j)|\theta} \quad (10)$$

$$\mathbf{RLNN} = \frac{\sum_{j=1}^{N_r}((\sum_{i=1}^{N_l} g(i,j)|\theta))^2}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_r}(g(i,j)|\theta)^2} \quad (11)$$

$$\mathbf{RP} = \sum_{i=1}^{N_l}\sum_{j=1}^{N_r}\frac{(g(i,j)|\theta)}{N_p} \quad (12)$$

where N_l is the number of discrete intensities in the image, N_r is the number of discrete rl, and N_P is the number of pixels. SRE calculates the short runs distribution. The higher value of SRE marks accurate textures. LRE measures the long runs distribution. The higher value of LRE marks poor textures. SRLGLE assures runs in the upper left quadrant of GLRLM, where SRL and low gl are located. LRLGLE measures the joint distribution of long rl with lower gl values. The LRHGLE method measures the joint distribution of the long rl with higher gl values. The HGLRE method measures the distribution of higher gl values. The higher value indicates more concentration of high gl values in the image. The LGLRE method measures the distribution of the low gl values. The higher value indicates more concentration of low gl values. The RLN method measures the similarity of the rl throughout the image. The lower value indicates more homogeneity among rl in the image. The RLNN method is the normalized version of the RLN. RP calculates the percentage of the number of realized runs and the maximum number of

potential runs. The highly uniform ROI volumes produce a low run percentage.

C. BINARY CLASSIFICATION

The calculated feature vector is supplied to an SVM classifier to distinguish normal from DR cases. SVM divides the data points into two classes. The hyperplane gives a margin to separate the two data groups into (0) or (1). The distance between the points and the separation line should be far enough, so the points are called support vectors. After generating the model, the classifier diagnoses the test set of images. The DR cases are labeled (1) and supplied to the next phases for detecting the exact DR grade, whereas the normal cases are labeled (0). This phase is very important for the segmentation phase. There is no need to segment the healthy cases, which reduces the overall processing time, memory space, and effort. The primary and essential reason to construct this phase is that the significant contribution is to visualize all the pathological variations to the ophthalmologists besides giving them the grade of each case. Therefore, it is no need to provide the ophthalmologists with unnecessary segmenting images for each healthy case. It is sufficient to tell them about the normal or healthy cases without giving them the BV, black (EX, HM, and MA) images. The four DR signs are not found in healthy cases. Of course, this phase prevents overlapping, obfuscation, and confusion for the ophthalmologists.

D. POST-PROCESSING

In this phase, we prepare the images for the customized U-Net segmentation phase. Because we utilize different images with diversity in resolution, noise, contrast, and illumination, we had to build the post-processing. This phase is an extension of the preprocessing phase. In this subsection, we have to resize all the preprocessed RGB images and their GTs to a standard size of 512×512. The GTs images are resized to be validated with the predicted ones. On the other hand, IDRiD dataset is very large. All images are in resolution 4288×2848. This step also saves memory space and reduces the processing time. The second step in this phase is the IDRiD's GTs binarization because IDRiD GTs are in RGB, as shown in Fig. 3. Finally, we extract masks of all images to be excluded in the segmentation process.

E. SEGMENTATION

In this subsection, we first give a brief illustration of the CNN identifications and operations. Then, we demonstrate in detail the U-Net architecture and its hyperparameters that are changed to improve training and validation accuracy. The segmentation phase by the U-Net model positively affects the classification of the four DR grades (mild NPDR, moderate NPDR, severe NPDR, and PDR). In addition, it is considered a universal method to extract the four pathological changes by training the model on each sign's GTs.

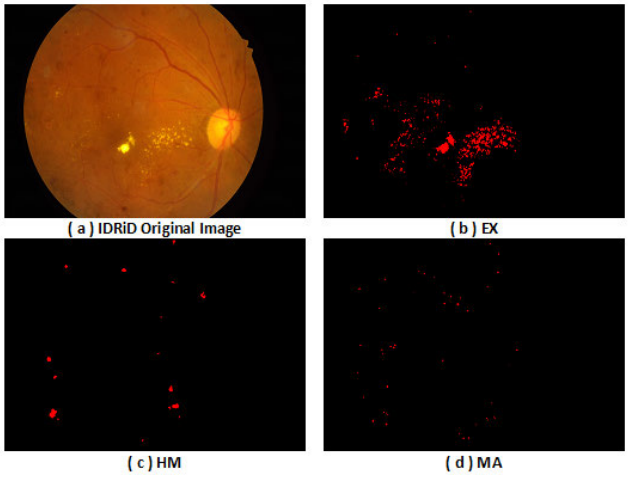

FIGURE 3. An example of IDRiD dataset with HM, EX, and MA GTs without binarization process: (a) The original image, (b) EX GT, (c) HM GT, and (d) MA GT.

1) CNN

Generally, the human brain detects the features to differentiate and categorize the objects around the human. In the same context, CNN work as the brain. It cannot categorize the objects without detecting the detailed features. CNN includes a set of operations, such as convolution, activation, pooling, flattening, and full connection. It is a feedforward multilayered hierarchical network. The connectivity between its neurons is inspired by visual cortex organization. The individual neurons are organized in such a way that they respond to the overlapping regions. The main positive characteristics of CNNs are they contain some of the best learning algorithms for grasping the image contents. They can learn good internal representation from unstructured, raw data. Also, they give a promising performance in ML classification problems, which lie in the correlations between labels or label dependency. Therefore, CNN can exploit spatial correlation in data to produce new hidden features from the obvious features. The CNN topology is split into multiple learning phases, such as convolutional and sub-sampling layers. Each layer uses kernels and performs multiple transformation processes [37].

2) CNN COMPONENTS

A general CNN architecture consists of CONV, pooling, and FC layers. Each component may at least consist of one layer. FC layer may be substituted by the global average pooling (GAP) layer. This layer reduces the overfitting because there are no parameters to be optimized and represent one feature map for each class or category. Besides, different mapping functions, different regulatory units such as BN and dropout (DO) are also embedded in the architecture to optimize CNN performance and avoid overfitting. It is very necessary to concentrate on the CNN components arrangement. This organization plays a vital role in designing new architectures and achieving satisfactory performance. The CNN components are stated as follows:

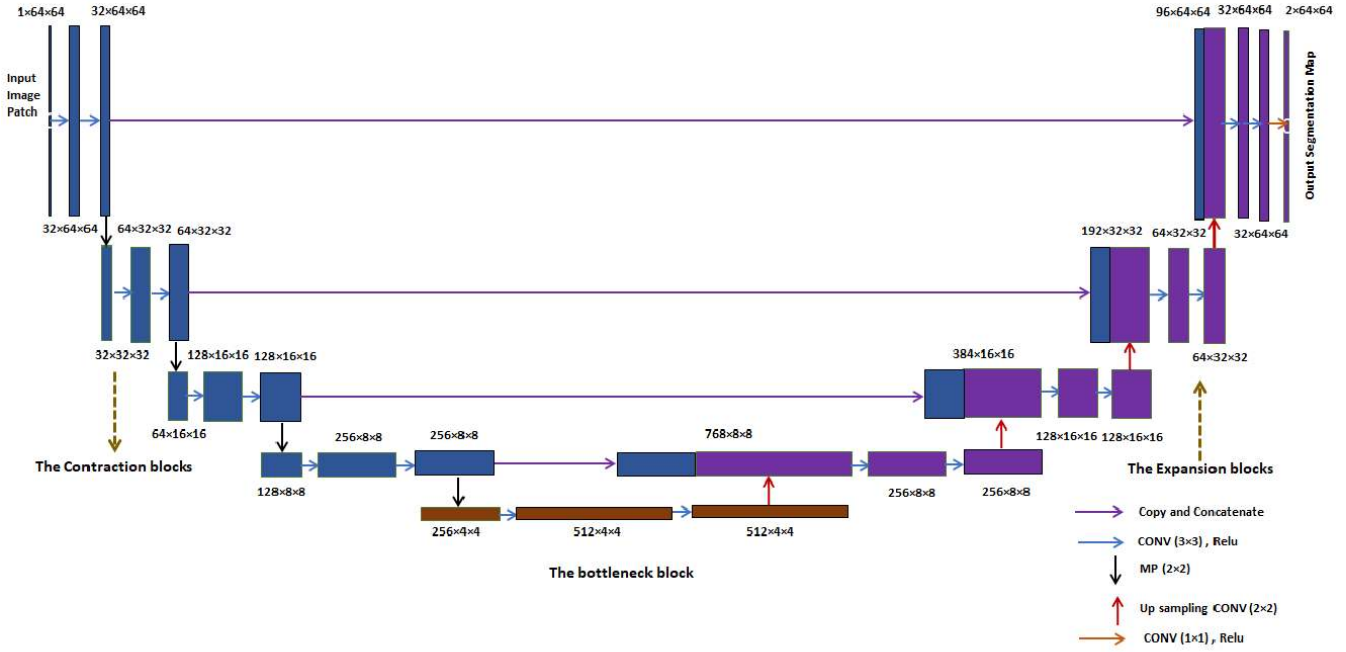

FIGURE 4. The U-Net architecture to extract EX, HM, BV, and MA.

- CONV [38], [39]: This layer is parameterized by a set of learnable filters (kernels). The kernel divides the image into small blocks. These blocks are recognized as receptive fields. The kernel is applied across the input tensor, which is represented as an array of numbers. An element-wise is calculated by the product between each kernel element and the input tensor at each tensor location. After that, the product is summed to obtain the output value in the corresponding position of the output tensor. The output is called a feature map. Zero paddings are utilized greatly in recent CNN architectures. It is used for retaining in-plane dimensions to apply more layers because, without zero paddings, each successive feature map would get smaller after the CONV process.
- Hyperparameters and down-sampling: A stride is a distance between two successive kernel positions. To achieve down-sampling, a stride is selected to be larger than one. A pooling is another process to perform down-sampling. The summarized hyperparameters are kernels size, number, padding, which can also be used for down-sampling, and stride. Finally, the CONV layer output is passed to the non-linear activation function such as Hyperbolic tangent (tanh), sigmoid, and ReLU.
- Pooling (PO) [40]: This process decreases the subsequent learnable parameters number, reduces the feature-map size, regulates the CNN network complexity, reduces overfitting, and increases the generalization. Po in CNN can be MP, average pooling (AP), global pooling (GP), global average pooling (GAP), L2, overlapping, or spatial pyramid pooling (SPP).
- FC [41]: It is known as a dense layer in which the output of CONV and PO layers is flattened and transformed into a 1D array. The learnable weight connects each input with each output. In the classification task, the final FC layer's output represents the output of the network. It is the probability of each class. Generally, the output nodes have the same number of classes.

3) U-NET ARCHITECTURE

It is a CNN model that is used to localize the abnormalities areas. If CNN is used to learn the image feature map to exploit new feature maps and convert the image to feature vector, the U-Net construct the image from this feature vector [42]. As shown in Fig. 4, the U-Net architecture consists of three phases that make the architecture take the (U) shape, which are contraction (down), bottleneck (the middle bottom), and expansion (up). In each phase, we can increase or collapse the number of the blocks. In the utilized architecture, we added three blocks in the contraction stage after the input. Each block includes two CONV (3×3) layers with RELU activation and followed by one MP (2×2) layer. The number of kernels is duplicated after each block as we started with 32 kernels and increased to 512 kernels or feature maps in the bottleneck phase. After that, the architecture starts the expansion phase by doing the up-sampling CONV (2×2) and RELU activation. This phase consists of three blocks as well as the contraction blocks. Each block includes two CONV (3×3) with RELU activation and followed by up-sampling CONV (2×2). The kernels or feature maps number are reduced to the half after each block. Finally, one CONV (1×1) is added to result in the segmentation maps. We trained

the architecture four times on the BV, EX, MA, and HM GTs and predicted four segmenting images from each input.

We trained and evaluated the model by 10-fold cross-validation with 30 epochs and 100 steps for each epoch. In prediction, the batch size was 16, stride (2, 2) for concatenation, ''same'' padding, and dropout equals (0.1). We optimized the model by Adam optimizer with the learning rate (l_r) equals to $1e-3$. Finally, we utilized a sigmoid function and binary-cross entropy. The utilized U-Net architecture is shown in Algorithm 2. C is convolution, U is up-sampling, 2@ is two consecutive convolutions, 1@ ConvT is one convolution transpose, plus (+) is a concatenation of the output of 1 convT layer of the expansion, and the feature maps of the contraction in the same level.

After segmenting EX, BV, HM, and MA, we validated the results due to six performance measures, as illustrated in detail in the next section.In addition, we compare the resulting segmentation of the proposed ML-CAD system with the universal customized DL segmentation model with other current segmentation methods.

F. FEATURE EXTRACTION AND SELECTION

We applied this phase by utilizing conventional hand-crafted methods. The reason is that we need to complete the four lesion segmentation process we made. We segmented each disease sign to diagnose it in its early and advanced grades carefully. Therefore, we cared about appearing even small MA that formulates the early mild grade for the ocular specialist or physician. It is crucial to extract features from these small signs. The thing that needs more auditing and supervising, while other systems, such as Lam *et al.* [43], could not diagnose the mild grade by using CNN.

For BV images, we utilized GLCM to extract 12 different feature descriptors, as proposed by Gadkari [44]. The GLCM describes the texture features. GLCM computes the frequency of appearance of pixel pairs with specified values in a spatial relation in the processed image. We skeletonize the BV network. Then, we determine BP with red marks and dismiss dummy, terminal branches and points.

Meanwhile, we calculated the BV, MA, EX, and HM areas. There are four fields recorded in the feature vector. The feature vector consists of 12 GLCM descriptors, 4 ROIs areas, and BP count for each DR image. We applied PCA technique. It is utilized to describe the extracted features with low dimensional space without information loss [45] by defining the most correlated values.

G. ML CLASSIFICATION

We utilized the MLSVM technique. MLSVM is based on SVM with a kernel of a radial basis function (RBF). We added the four class labels (mild NPDR, moderate NPDR, severe NPDR, and PDR) to the feature vector. The normal or healthy grade was defined before the segmentation phase by the binary SVM. In the first-class label (mild NPDR), all the images with a mild grade are defined by 1, and the others are 0, and so on for the other class labels. So, the MLSVM classifier builds one binary classifier for every class label based on the predictions of preceding classifiers in the chain.

Algorithm 1: MLC Phase of the ML-CAD System

```
Data: Label matrix, data matrix, K-Fold, SVM kernel,
      and previous labels.
Result: Model and predictions.
Calculate Ch = random permutation(classes NO.);
Set Previous labels=zeros;
for I= Ch do
    if Index = [] then
        OPERATE SVM training (data, label, SVM
          STRUCT);
        RETURN model;
        OPERATE SVM predicting (label, test, model,
          SVM (type));
        RETURN Predicted labels;
        PUT Previous labels (I) = Predicted labels;
        PUT post Index = Model (label) = 1;
    else
        OPERATE SVM training (data, label (Index),
          SVM STRUCT);
        GET Predicted labels;
        PUT Previous labels = Predicted labels;
        PUT post Index = Model (label) =1;
    end
    PUT Index = [Index, I];
end
```

According to the MLC idea, the correlations among the labels are significant in producing new labels. The SVM classifier based on CC achieves the correlation by aggregating the binary classifiers' predictions that were built. CC makes the aggregation in a chaining order strategy. CC prompts additional features for the instances. In addition, it randomly prompts the connections among class labels. These correlations are specified by the alteration. In testing, the binary classifiers are applied. Then, the classifiers' outputs form the label features of the chain structure. Finally, the technique aggregates both responses and computes the prediction. We validated the results by the k-fold cross-validation technique to avoid overfitting.

IV. EXPERIMENTAL RESULTS

This section is divided into four subsections. The first subsection is the dataset description, which illustrates the settings of the nine utilized datasets (eight for training and testing while the last one is for training only MA GTs). Second, the performance metrics are discussed, which are utilized to evaluate the system's segmentation and classification phases. Third, DR sign segmentation is divided into two parts. One for BV segmentation and Bifurcation points (BP) extraction and the second for MA, EX, and HM segmentation. Finally, the ML classification results subsection presents the results of grading the DR cases.

Algorithm 2: Segmentation by Using U-Net

Data: batch size, inputs, activation, DO, classifier, optimizer, and learning rate.
Result: Model and predictions.
SET kernel k=3, Stride S=2;
SET PO=2, padding PA=same;
Stage1: Contraction path
Block 1:
C1= 2@Conv(K, inputs, PA), filters F=32;
MP1= MP(C1, PO);
Block2:
C2= 2@Conv(k, MP1, PA), F=64;
MP2= MP(C2, PO);
Block3:
C3= 2@Conv(K, MP2, PA), F=128;
MP3= MP(C3, PO);
Block4:
C4= 2@Conv(K, MP3, PA), F=256;
MP4= MP(C4, PO);
Stage2: Bottleneck point
C5= 2@Conv(K, MP4, PA), F=512;
[Stage3: Expansion path] Block1:
U6= 1@ConvT(K=2, S, PA, C5), F=256;
U6= U6+C4;
C6= 2@Conv(K=3, U6, PA), F=256;
Block2:
U7= 1@ConvT(k=2, S, PA, C6), F=128;
U7= U7+ C3;
C7=2@Conv(K=3, U7, PA), F=128;
Block3:
U8= 1@ConvT(k=2, S, PA, C7), F=64;
U8= U8 + C2;
C8= 2@Conv(K=3, U8, PA), F=64;
Block4:
U9= 1@ConvT(K=2, S, PA, C8), F=32;
U9= U9 + C1;
C9= 2@Conv(K=3, U9, PA), F=32;
C10= 1@Conv (K=1, classifier, C9), F=1;
SET outputs=C10;
GET model (inputs, outputs);
Compile model;

A. DATASETS' DESCRIPTION

We applied our proposed ML-CAD system on eight standard datasets: HRF [46], ChaseDB1 [47], DIARETDB0 [48], DIARETDB1 [49], STARE [50], MESSIDOR [51], DRIVE [52], and IDRiD [53]. Table 2 lists the main features of the used eight benchmark datasets and the last one for training the U-Net model on the MA sign.

- DRIVE dataset [52]: It consists of 40 retinal images. Twenty for training and the rest for testing. The experts manually diagnosed them as seven cases have DR, and the other 33 are healthy cases. Each set contains a field of view (FOV) masks.
- High-Resolution Fundus (HRF) dataset [46]: It consists of 30 images in total. Fifteen cases are healthy, and the others are DR cases. It has BV ground truth (GT) for each healthy and DR images, which are manually segmented by a clinical expert.From our point of view, the dataset is very important in training and validating models because each GT image includes full thick and thin vessels that perform the complete vasculature.
- STARE dataset [50]: It includes 400 images. We utilized only twenty of them because they have BV GTs. These images were divided in balance, 10 images are healthy, and the others are DR cases.
- CHASEDB1 dataset [47]: It consists of 28 images, which are manually segmented to BV by two experts. The advantage of this dataset is that its images were collected from 14 children. It will give our ML-CAD system variety in training and testing. Besides, all images are paired with the same person. The dataset is split into 20 images for training and 8 for testing.
- DIARETDB0 dataset [48]: It consists of 130 images, 20 of them are normal, and the rest contain DR signs. The GTs are about *.dot files, including DR sign's name.
- IDRiD dataset [53]: We utilized two parts of this dataset, which are segmentation and disease grading. It contains 81 in JPEG format. It has a GT of 4 lesions, which are HM, MA, hard EX, and soft EX in TIF format. These images are pixel-level annotated. They were split into 54 and 27 for training and testing, respectively. EX and HM are found in 80 different images, while MA is found in 81 images. In DR grading, the dataset contains 516 images. The images are split into 413 and 103 as training and testing sets, respectively.
- DIARETDB1 dataset [49]: It includes 89 images. The experts ensured that only five of them are normal, and the rest contain at least mild NPDR. The dataset has GT for all images in hard EX, soft EX, HM, and MA signs.
- MESSIDOR dataset [51]: It includes 1200 images. Two medical experts specified the DR grades and ME. We utilized the 100 images of the dataset base1. We separated it evenly for training and testing.
- E-ophtha dataset [54]: It contains 82 and 381 color fundus images for EX and MA, respectively. In EX, the images with the sign are 47, and the others are healthy. In MA, the images with the sign are 148 images, and the others are healthy. We utilized it in training the model on the MA signs to be tested on the other eight datasets because it performs better than IDRiD in detecting MA.

This work was implemented by using MATLAB R2018a and python 3.7. We ran our experiments on a core i5/2.4 GHz computer with 8 GB RAM and an NVIDIA/ (1 GB VRAM) VGA card. As described in Table 2, DRIVE, HRF, STARE, and chasedb1 datasets have no GT for MA, EX, and HM, but has GT for BV. The DIARETDB1 dataset has no GT for BV but has lesion level GT for MA, EX, and HM.

TABLE 2. The main specifications of the nine used benchmark datasets (The ninth is used for training U-Net on MA GTs.

dataset	Type	Fundus Images	Used Camera	Resolution	Format	GT	GT level	Experts	Notes
DRIVE	Binary classes	400	Canon CR-5 45^0 FOV	768×584	JPEG	Yes: BV	pixel	2	40 were chosen randomly, 7 of them have mild NPDR, and the others are normal. The images are for 25 to 90 years of age
HRF	Binary classes	30	Canon CR-1 45^0 FOV	3504×2336	JPG	Yes: BV	pixel	1	Were separated evenly to healthy and DR.
STARE	Multi classes	400	TopCon TRV 50 with 35^0 FOV	700×605	PPM	Yes: BV	pixel	2	The 20 of them had GTs were separated evenly to normal and DR
CHASEDB1	Binary class	28	Nidek NM-200-D with 30^0 FOV	999×960	JPG	Yes: BV	pixel	2	from 14 children, paired for the same person.
DIARETDB0	Multi-Label	130	digital fundus camera woth 50^o FOV	1500×1152	PNG	Yes: only sign's name in Dot file	lesion	unknown	20 of them are normal and the rest are DR cases.
IDRiD	Multi-Label	597	AKowa VX-10 alpha with 50^o FOV	4288×2848	JPG	Yes: Ex, HM, MA, and grading CSV	pixel	unknown	pixel level annotation
DIARETDB1	Multi-Label	89	50^0 FOV with various digital fundus camera	1500×1152	PNG	Yes: EX, HM, and MA	lesion	4	5 are normal and the rest contain at least mild NPDR
MESSIDOR1	Multi-Label	1200	Topcon TRC NW6 with 45^0 FOV	1440×960	TIFF	Yes: grading CSV	Image	3	Normal: MA, HM = 0, mild: 0 < MA <= 5, HM = 0 , moderate:5 < MA < 15/ 0 < HM < 5, NV = 0, severe:MA >= 15/ HM >=5/ NV = 1, EX for DME.
E-ophtha	Binary	463	unknown fundus camera settings	1440×960	JPG	Yes: EX, and MA	pixel	unknown	we trained the U-Net model on MA GTs and give best performance than IDRiD in MA detection

DIARETDB0 dataset has no GT for BV segmentation but has .dot files that include the lesions occurrence in each image. The MESSIDOR dataset has no EX, BV, HM, and MA GTs. The experts provide specific values for each sign in rules. These rules help in the diagnosis process. Finally, IDRiD has no BV GTs. It has a pixel level GTs of HM, EX, and MA.

The DIARETDB1 dataset shows a GT type of lesion level. Three retinal experts indicated EX, MA, and HM by a manual annotation that was done by using a single pixel, the lesion center, or using a coarse boundary. The manual annotation was done by drawing a disk over the lesion, which covers the entire lesion region, Therefore, the annotation does not mark specific lesion regions' contours. Therefore, the dataset is not a pixel-level. Because of this fact, we couldn't use DIARETDB1 in training U-Net. Instead, we utilized the IDRiD dataset in training. It is a pixel-level annotation of typical DR lesions, as shown in Fig. 3.

Therefore, we segmented the BV and compared the result with GTs of the CHASEDB1, DRIVE, HRF, and STARE dataset. The complete BV segmentation performance is measured to ensure the custom U-Net model's accuracy in BV segmentation. After that, we predict the other four datasets that have no BV GTs. We trained the U-Net model on the IDRiD dataset's training set to segment two lesions (EX, and HM). We segmented the two lesions from the testing set of the IDRiD dataset and compared the results. While we check the performance, we apply the model on the other seven datasets with no EX and HM GTs. But for the last lesion (MA), we used the E-ophtha dataset for training the model as it gives better performance. In the same way, we predicted MA lesion on IDRiD and the others. Finally, we measured the ROIs areas of the eight datasets. We compared the DR grades diagnosis performance for each dataset using the formerly substantive rules in MESSIDOR and IDRiD datasets.

B. PERFORMANCE METRICS

We utilized six different measures to evaluate the performance of the proposed ML-CAD system, i.e., SEN, specificity (SPE), DSC, accuracy (ACC), positive predictive value (PPV), and area under the curve (AUC), which are listed in Eqs. 13, 14, 15 [55], 16, 17, and 18 [56] and [57].

SEN is the rate of true positive (TP). SPE is the proportion of the true negatives (TN). The technique may be accurate without being sensitive, or it may be sensible without being specific. ACC is the ratio of true results, either TP or TN overall images. False positive (FP) is the ratio of false predictive or incorrect positive predictions. False negative (FN) is the ratio of incorrect negative predictions. DSC measures the resemblance between the predictions and GT. PPV is the proportion of the correct positive predictions over the correct and incorrect positive predictions. Finally, AUC is nearly half

of the summing of the SEN and SPE.

$$SEN/RE = \frac{TP}{TP + FN} \tag{13}$$

$$SPE = \frac{TN}{TN + FP} \tag{14}$$

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{15}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$PPV = \frac{TP}{TP + FP} \tag{17}$$

$$AUC \approx 0.5\,(SEN + SPE) \tag{18}$$

C. BINARY CLASSIFICATION RESULTS

In this section, we feed the feature vector resulting from the GLRLM to the binary SVM to differentiate the normal and DR. We have two graded datasets, such as MESSIDOR and IDRiD. Their images are graded to 5 grades from normal to PDR. On the other hand, we have the HRF dataset, which its images are obviously differentiated to only normal and DR cases. In STARE datasets, we have the diagnose code of the ocular disease in each image where DR is one of them. DIARETDB0 has *.dot files that include the sign that is found in the image. Of course, the dot file of the image containing NaN is normal; otherwise, DR. The other datasets, such as DRIVE, CHASEDB1, and DIARETDB1, have no graded and not detected except by observation and the number of normal and DR images of them. Therefore, it is reasonable to train the model on the well-defined labels (normal and DR) dataset, which is HRF, as we need the binary classification. As shown in Table 3, we also trained the others to select the best-trained model to predict the unknown labels. We trained the SVM model on each feature vector of HRF, IDRiD, MESSIDOR, STARE, and DIARETDB0 with 5-fold cross-validation.

From Table 3, we can notice that for the HRF dataset, the binary DS comes in the first order, the ML IDRiD dataset comes in the second order. The second ML dataset, the DIARETDB0 is ranked in the third order. The model on the STARE dataset is somewhat more balanced than in the previous one. The ML MESSIDOR dataset comes in the last order. From the results presented in Table 3, we generated the SVM model that is trained and tested on the HRF dataset to predict the other unknown labels. The images that are labeled as DR cases or label 1 are supplied to the next phase.

D. SEGMENTATION RESULTS

In this subsection we present the BV and BP segmentation results in part IV-D1 and the other lesions segmentation in part IV-D2 as following:

1) BV AND BP SEGMENTATION

This part includes segmenting the BV network from the color fundus images in the eight datasets and recording performance measures metrics. After removing the noise and enhancing the resized images' contrast, we trained the U-Net model on the DRIVE dataset (training set). After that, we tested the other datasets. The average training ACC is 95.48%. Fig. 5 presents the examples of applying the customized U-Net model on four datasets that have BV GTs in addition to the BP extraction. The BP extraction was done by skeletonizing the BV images and omitting the dummy branches and BP. After that, marking the BP and resulting in BP's count. The increasing BP counts indicate the appearance of NVs. It is noteworthy that the BV network size also leads to the NVs occurrence. On the other hand, Fig. 6 presents the examples of applying the customized U-Net model on the four ML datasets that have not BV GTs in addition to BP.

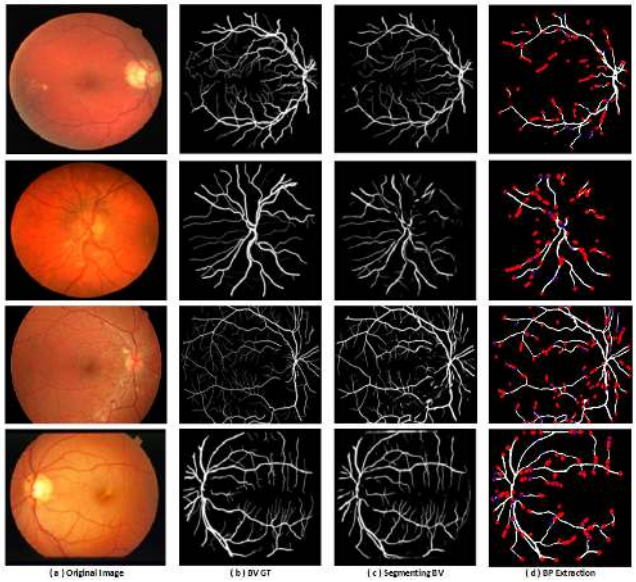

FIGURE 5. Examples of BV segmentation by using U-Net model on DRIVE, CHASEDB1, HRF, and STARE datasets, respectively: (a) The original images, (b) BV GT, (c) BV segmentation, and (d) BP extraction.

Table 4 shows the six performance measures (AUC, ACC, SPE, SEN, DSC, and PPV) of the BV segmentation by U-net model on the four datasets (DS) that have BV GTs (DRIVE, STARE, CHASEDB1, and HRF datasets). We compared the resulting BV with the BV GTs of the four aforementioned datasets. Besides, we compared the BV segmentation results of the proposed ML-CAD system by U-Net model with current five methods. In Table 4, we present the comparison between the proposed system, soares *et al.* [58], B-COSFIRE filter [59], Abdelmaksoud *et al.* [21], Gao *et al.* [60], and Adapa *et al.* [61]. Soares *et al.* [58] utilized 2D morlet wavelet transform in multiple scales with GMM. B-COSFIRE filter [59] calculates the weighted geometric mean of input collinearly aligned DoG filters. Abdelmaksoud *et al.* [21] combined a matched filter with a first-order Gaussian derivative and Coye Filter. Gao *et al.* [60] utilized U-Net with Gaussian matched filter. Adapa *et al.* [61] utilized gray level, shape, and Zernike moment features to differentiate between BV and background pixels.

TABLE 3. The average performance of the proposed ML-CAD system in the binary classification phase.

DS	ACC(%)	AUC(%)	SEN(%)	SPE(%)	PPV(%)	DSC(%)	PCA(%)
HRF	96.66	99	100	93.3	93.75	96.7	PC1: 95.1, PC2:4.8, PC3:0.1
DIARETDB0	92.3	95	94	33	88.03	91.15	PC1:85.3, PC2:11.2, PC3:3.1, PC4:0.4
STARE	90	92	90	90	90.2	90	PC1:95, PC2: 5
MESSIDOR	75	72	86	50	80	82.7	PC1:95, PC2:5
IDRiD	96.1	96	100	80	95.3	97.6	PC1:92.9, PC2:7.1

TABLE 4. Average performance of the BV segmentation by using U-net model.

Method	DS	ACC(%)	AUC(%)	SEN(%)	SPE(%)	PPV(%)	DSC(%)
AbdelMaksoud et al. [21]	DRIVE	95.61	-	62.45	98.79	-	71
Soares et al. [58]		94.6	95.9	-	-	-	-
B-COSFIRE filter [59]		94.4	96.1	76.5	97.04	-	-
Gao et al. [60]		96.3	97.7	78	98.7	89	-
Adapa et al. [61]		94.5	93.9	69.9	98.1	-	-
The Proposed system		96.58	97.84	72.58	98.89	86.29	78.84
Abdelmaksoud et al. [21]	STARE	96.14	-	69.84	97.63	-	68.23
Soares et al. [58]		97.7	96.5	-	-	-	-
B-COSFIRE filter [59]		94.9	95.6	77.1	97.01	-	-
Adapa et al. [61]		94.8	95	62.9	98.3	-	-
The Proposed system		95.55	94.93	66.1	97.93	72.25	69.04
B-COSFIRE filter [59]	CHASEDB1	93.8	94.8	75.8	95.8	-	-
The Proposed System		96.17	95.08	56.75	98.94	79.05	66.07
B-COSFIRE filter [59]	HRF	96.4	95	75	97.4	-	-
The proposed system		95.6	95.30	70.14	98.25	85.1	76.2
The proposed system	Averages	95.97	95.78	66.4	98.51	80.6	72.5

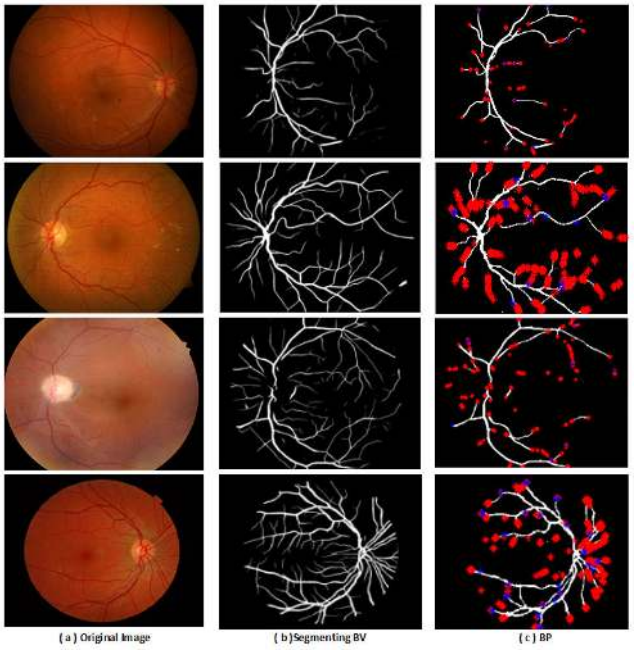

FIGURE 6. Examples of BV segmentation by U-Net model on DIARETDB0, DIARETDB1, IDRiD, and MESSIDOR datasets, respectively: (a) The original images, (b) The segmenting BV, and (c) The BP extraction.

We divided Table 4 into four parts according to the four datasets. In DRIVE dataset, the system accomplished 96.56%, 97.84%, 72.58%, 98.89%, 86.29%, and 78.84% for ACC, AUC, SEN, SPE, PPV and DSC, respectively. In CHASEDB1 dataset, it achieved 96.17%, 95.08%, 56.75%, 98.94%, 79.05%, and 66.07% for ACC, AUC, SEN, SPE, PPV and DSC, respectively. In STARE dataset, the proposed system achieved 95.55%, 94.93%, 66.1%, 97.93%, 72.25%, and 69.04% for ACC, AUC, SEN, SPE, PPV and DSC, respectively. In HRF dataset, the system achieved 95.6%, 95.30%, 70.14%, 98.25%, 85.1%, and 76.2% for ACC, AUC, SEN, SPE, PPV and DSC, respectively.

By comparing the results in DRIVE, we can notice that the proposed system outperforms the others in ACC, AUC, SPE, and DSC metrics, but B-COSFIRE filter [59] and Gao *et al.* [60] are higher in SEN and SPE. B-COSFIRE filter [59] is higher than ours by approximately 4 for SEN, while Gao *et al.* [60] is highest by approximately 5.5,and 2.7 for SEN, and PPV respectively. In STARE, the soares *et al.* [58] is higher in ACC and AUC by difference of 2.2 and 1.6 respectively. But, the average ACC and AUC of our proposed system is greater than the averages of them in DRIVE and STARE datasets.

In CHASEDB1 and HRF, the proposed system outperforms the B-COSFIRE filter [59] in all matrices except in SEN. It achieved 75%.

Finally, we can conclude that the proposed system achieved 95.97% 95.78%, 66.4%, 98.51%, 80.6%, and 72.5% for averages of ACC, AUC, SEN, SPE, PPV, and DSC, respectively in BV segmentation.

2) EX, MA, AND HM SEGMENTATION

In this part, we present the results of segmenting the other three lesions (EX, MA, and HM) using A customized U-Net model. To segment the three lesions, we trained the model on the IDRiD GTs and produced the weights that can be loaded

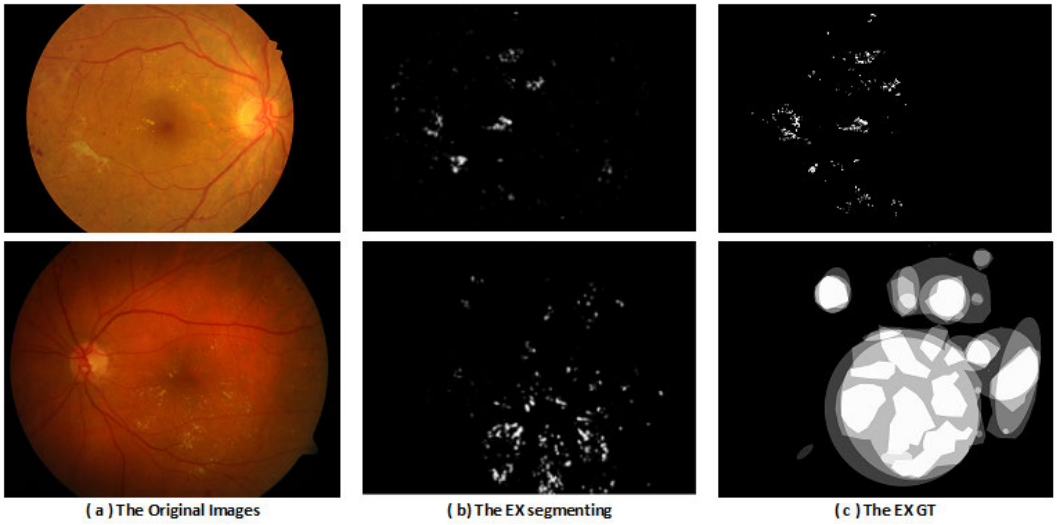

FIGURE 7. Examples of EX segmentation by U-Net model on IDRiD and DIARETDB1 datasets compared with their GTs. First row represents IDRiD dataset and the second one represents the DIARETDB1 dataset: (a) The original images, (b) The segmenting EX, and (c) The EX GT.

to predict the other datasets. Fig. 7 shows EX segmentation on two datasets (IDRiD and DIARETDB1), which have the EX GTs for validation. Training the U-Net model with IDRiD is more accurate than training with the DIARETDB1, as illustrated before in subsection IV-A. Figs. 8 and 9 show the EX segmentation results on the other six datasets that have not GTs of EX sign. Finally, we present a complete example of segmenting EX, MA, and HM using the U-Net model (23 layers) in Fig. 10.

Table 5 shows the comparisons between the proposed system, Abdelmaksoud *et al.* [21], Yan *et al.* [62], Kou *et al.* [63], and Khojasteh *et al.* [64] in EX, MA, and HM segmentation on the IDRiD and DIARETDB1 datasets that have EX, MA, and HM GTs. Abdelmaksoud *et al.* [21] utilized wavelet and morphological operation in EX, MA, and HM segmentation. Kou *et al.* [63] used residual U-Net in EX, and MA segmentation. Khojasteh *et al.* [64] built CNN model to segment the three signs. Yan *et al.* [62] utilized UNICOM feature. They combined intensity uniqueness and spatial compactness characteristics together.

In Table 5, we can observe that the proposed ML-CAD system with the customized U-Net gives a full performance in detecting the EX signs for all DR entered images. In MA, and HM the system achieved better performance. The system does a better performance in the IDRiD dataset than in DIARETDB1 dataset. The results of detecting EX and MA in DIARETDB1 are very near. Notably, the proposed system and Abdelmaksoud *et al.* [21] give the same results in EX detection from IDRiD, but other than that, the proposed system outperforms the other systems in the six metrics.

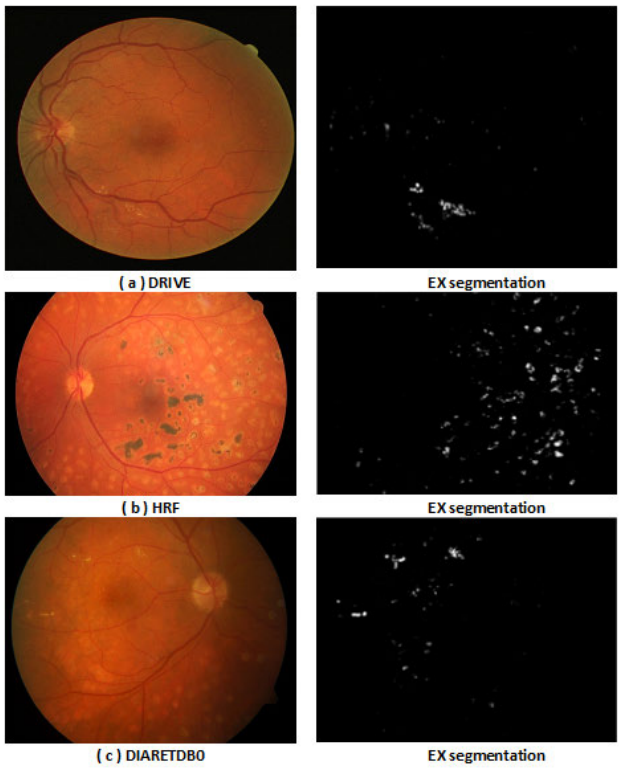

FIGURE 8. Examples of EX segmentation by U-Net model on DRIVE, HRF, and DIARETDB0 datasets. The left column is the original images and the right one is the EX segmenting images.

3) THE ML CLASSIFICATION RESULTS

After segmenting the DR images and producing five images (BV, BP, EX, MA, and HM) for each one image of the tested datasets, now, the role is to grade the DR images into four grades (mild NPDR, moderate NPDR, severe NPDR, and PDR). For BV images, we compute the 12 descriptors of the GLCM, which are stated in [65]. Then, the BP count, which is stated while extracting the BP from the BV, is recorded in the

TABLE 5. The performance of MA, EX, and HM detection for the IDRiD and DIARETDB1 dataset.

Method	dataset	DR lesion	ACC(%)	AUC(%)	SEN(%)	SPE(%)	PPV (%)	DSC(%)
Kou et al. [63]	IDRiD	EX	99	98.6	95	93.9	-	-
		MA	98	98.01	92.9	93.5	-	-
		HM	-	-	-	-	-	-
Abdelmaksoud et al. [21]	IDRiD	EX	100	-	100	100	-	100
		MA	97.8	-	100	80	-	98.7
		HM	96.7	-	100	72.7	-	98.2
Khojasteh et al. [64]	DIARETDB1	EX	98	-	96	98	94	-
		MA	94	-	85	96	83	-
		HM	90	-	84	92	85	-
Yan et al [62]	DIARETDB1	EX	-	96.4	89.1	98	-	-
		MA	-	96.4	97.8	95.5	-	-
		HM	-	98.1	-	-	-	-
The proposed system	IDRiD	EX	100	100	100	100	100	100
		MA	99.9	98	99.8	99.5	98.5	99.9
		HM	99.1	97.5	100	98	98.2	99
	DIARETDB1	EX	98.72	97.8	97	96.2	95	96
		MA	98	97	96	98	94	95
		HM	96	97.5	95	97	96.2	96

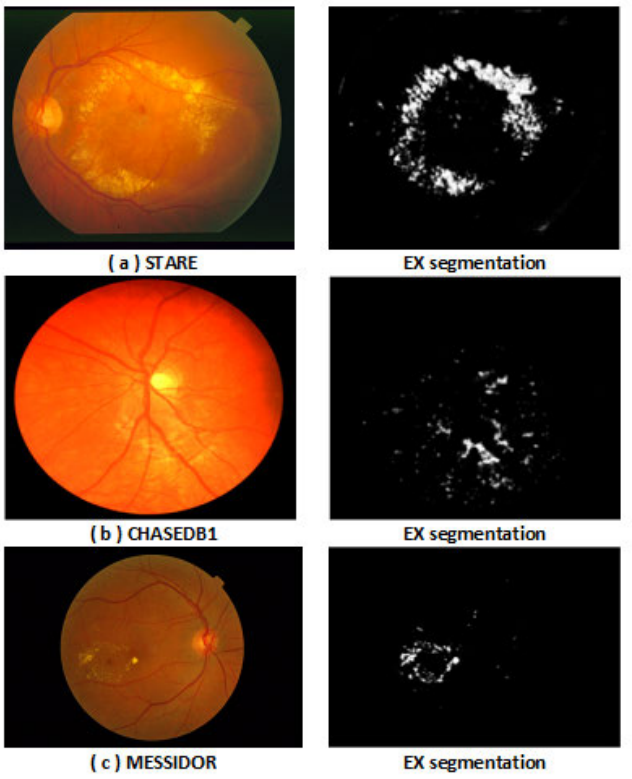

FIGURE 9. Examples of EX segmentation by U-Net model on STARE, CHASEDB1, and MESSIDOR respectively, The left column is the original images, and the right one is the EX segmenting images.

feature vector file. The four segmenting images of EX, BV, HM, and MA are characterized as the ROIs are in white color on a black background. We calculated the areas of the white pixels in each image for each sign. In this respect, we added the ROIs areas' results in the feature vectors. We used the 10-folds cross-validation technique to avoid overfitting.

Table 6 presents the ML-CAD system grading results via the utilized datasets. We compared the averages of the six performance measures for Abdelmaksoud *et al.* [21], decision tree (DT) classifier [66], Gaussian naive bayes (NB) [67], logistic regression (LR) [67], random forest (RF) [67], ML-k nearest neighbor (ML-KNN) [66], label power set (LP) [68], and classifier chain (CC) [68]. DT, GaussianNB, LR, and RF are based on ML binary relevance (BR) classifier [68]. Table 7 presents the comparisons of the averages of the six performance measures for the proposed ML-CAD system and the other aforementioned ML classifiers.

In ML classification, we can notice from table 6 that the proposed ML-CAD system achieved total averages of 95.05%, 91.85%, 86.11%, 86.8%, 84.7%, and 86.2% for ACC, AUC, SEN, SPE, PPV, and DEC respectively.

From Table 7, we can observe that the proposed ML-CAD system outperforms Abdelmaksoud *et al.* [21] and the seven ML classifiers in DR grading. Except in PPV and DSC, Abdelmaksoud *et al.* [21] is greater than the proposed one by a difference of 7.3% for PPV and a small difference for DSC, which equals 1.2%. In ACC, CC comes in the third order, while ML-KNN in the fourth and RF in the fifth, then LP in sixth, GaussianNB in the seventh, DT in the eighth and LR in the final order. DT gives better results in SEN, SPE, PPV, and DSC. It comes in the third order. Fig. 11 shows, the comparison between the eight classifiers, and the proposed ML-CAD system in DR grading due to the six measures.

V. DISCUSSION

In this section, we discuss and compare between kou *et al.* [63], luo *et al.* [69], Abdelmaksoud *et al.* [21] system and the proposed ML-CAD system. kou *et al.* [63] improved U-Net in order to detect the two early signs of DR; EX, and MA. Their system can be considered as a special step in diagnosing DR grades. But, the authors extracted only EX, and MA. HM and NV are very important signs in completing the grading of DR. The same observation is for luo *et al.* [69]. They utilized

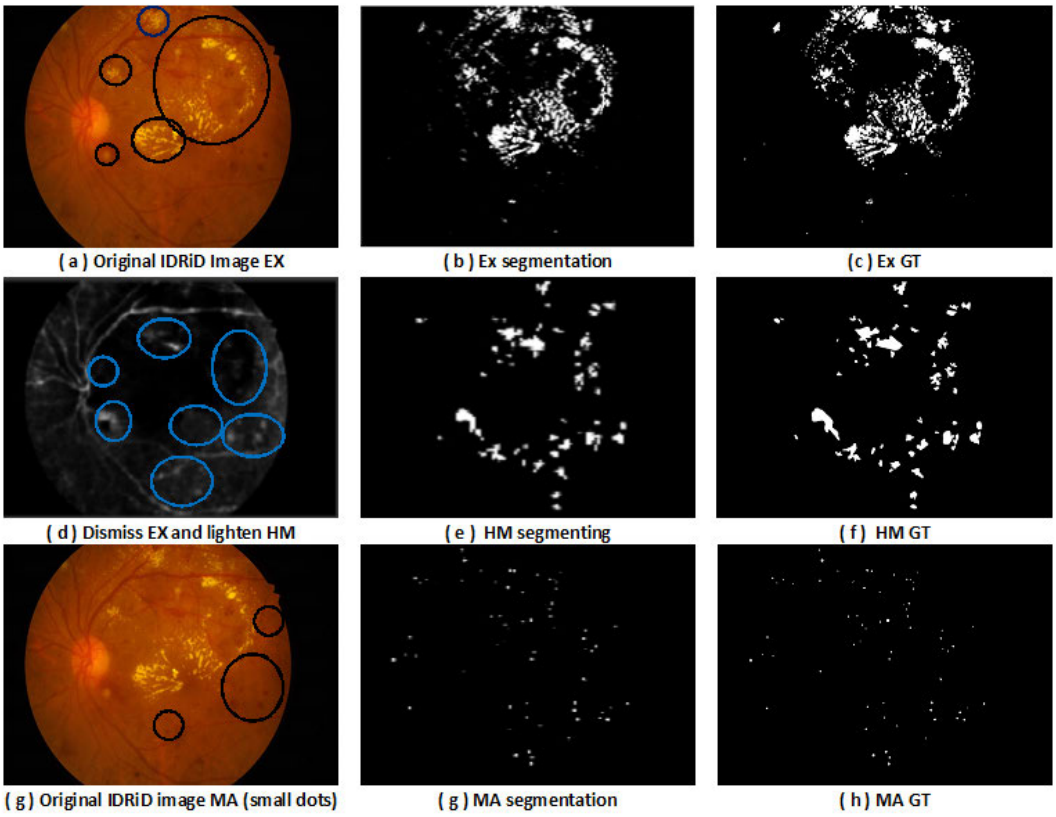

FIGURE 10. Examples of MA, EX, and HM segmentation by U-Net model on IDRiD dataset with its GTs.

TABLE 6. The average of the six metrics for the proposed ML-CAD system DR grading on all eight datasets.

DS	ACC(%)	AUC(%)	SEN(%)	SPE(%)	PPV(%)	DSC(%)
DRIVE	96.3	97.8	79.4	99.1	86.5	79.4
STARE	96.2	98	80.9	92.93	80.1	90.8
MESSIDOR	92.1	93.2	92.9	91	90.2	91.4
IDRiD	95.1	88.5	97.3	79.8	91.2	93.2
CHASEDB1	96.4	97.5	75.7	97.4	83.05	74.1
HRF	96.05	97	77.1	95.8	82.3	77.5
DIARETDB1	95.5	69.4	98.8	40.2	96.5	97.64
DIARETDB0	92.8	93.4	86.8	98.3	82.5	85.3
Total Averages	95.05	91.85	86.11	86.8	84.7	86.2

TABLE 7. The comparison between the proposed ML-CAD system and others.

Method	ACC(%)	AUC(%)	SEN(%)	SPE(%)	PPV(%)	DSC(%)
Abdelmaksoud et al. [21]	89.2	85.20	85.1	85.2	92.8	88.7
DT-BR [66]	71.7	70.1	78.4	89.8	80.8	80.2
LP [68]	76.5	76.4	55.5	84.9	70.7	62.3
GaussianNB-BR [67]	75.6	72.4	81	64.1	83.4	82
LR-BR [67]	64.03	57	77.6	39.1	72	74.2
RF-BR [67]	77.6	70.5	50	90.8	49.2	79.7
CC [68]	79.8	74.9	67.7	82.3	68.4	65.7
ML-KNN [66]	78.1	66	50	83.9	50.4	81.5
The proposed ML-CAD system	95.05	91.85	86.11	86.8	84.7	86.2

U-Net with denseNet to extract the BV. BV is extracted in order to notice the retina abnormality. It determines the NV that leads to determining the severe NPDR and PDR. Our proposed system extracts the four lesions and diagnoses the DR grade. These two systems not need user interaction, and hand crafted feature extraction. Moreover, they not utilized

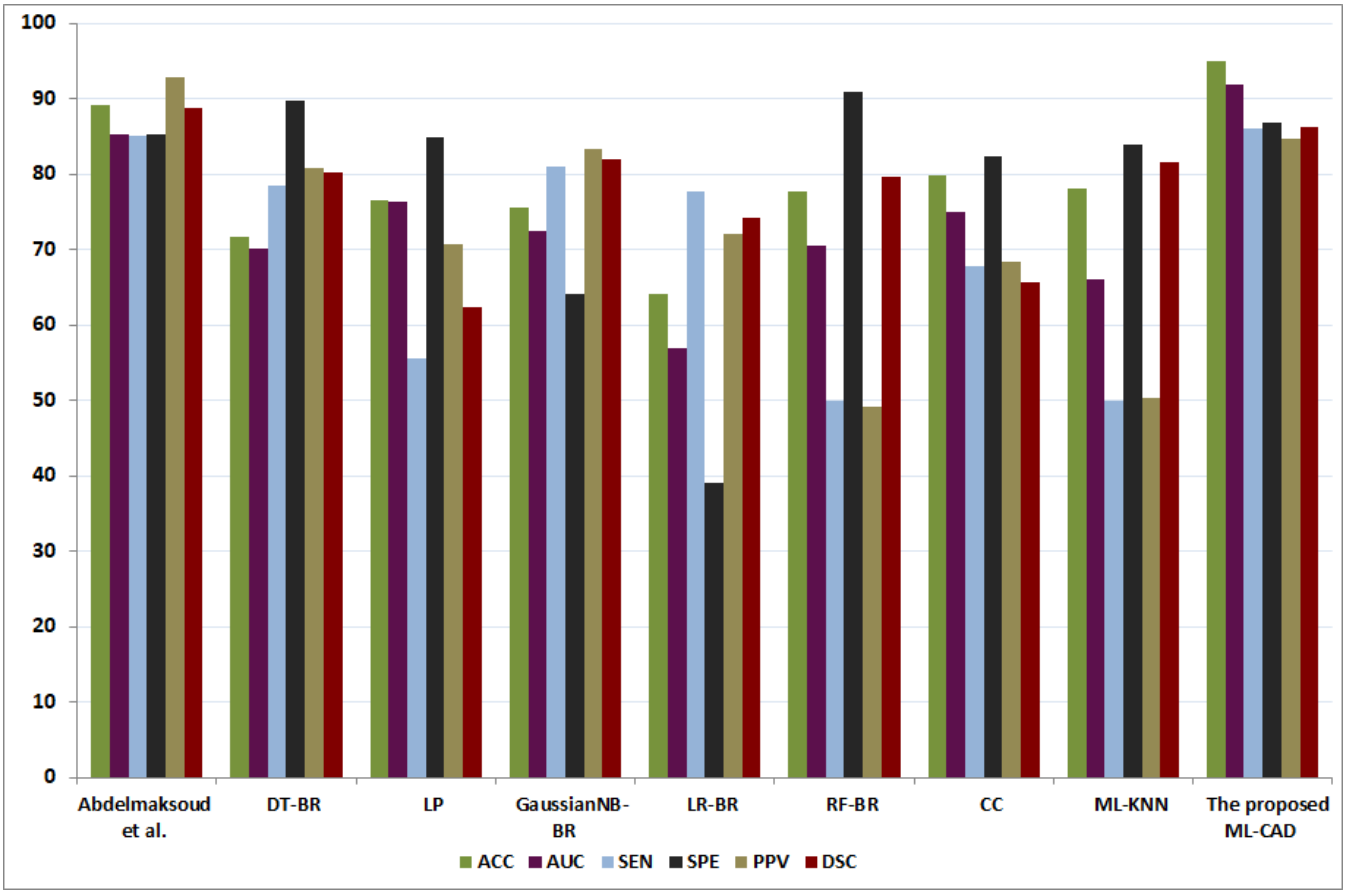

FIGURE 11. The comparison between the seven ML classifiers, abdelmaksoud *et al.*, and the proposed ML-CAD system due to the six measures (ACC, AUC, SEN, SPE, PPV, and DSC).

any ML classifiers in order to classify the DR grades. They segmented all the images (healthy and DR cases), which may happen confusion if the ophthalmologists checked for diagnosing the absence/presence of DR.

Therefore, we take care this problem in the proposed ML-CAD system. It extracts the GLRLM feature of the color fundus images in four degrees and utilized binary SVM classifier to differentiate between the healthy and DR cases before segmentation. Then it makes the segmentation using deep learning CNN model (U-Net) model to segment the four retinal pathological changes. In the proposed system, there is no need for user interaction besides, it utilized a universal segmentation model unlike the previous system. So, the proposed ML-CAD system prevents the ophthalmology confusion and lessen the burden on the developer. The proposed ML-CAD system was applied on eight datasets (DRIVE, STARE, CHASEDB1, HRF, DIARETDB0, DIARETDB1, MESSIDOR, and IDRiD), four of them have been ML. The thing that makes the proposed ML-CAD system is reliable, robust, and can be applied on the real world. It is applied on various color fundus images with different cameras' settings and from different patients (children and adults), (paired and unpaired), different qualities, noise and illumination. Table 9 shows the analytical comparison between the proposed ML-CAD system and the three aforementioned systems.

In segmentation phase, We trained U-Net on CHASEDB1 dataset and used the resulting weights in testing the other seven datasets. It resulted better BV segmentation in ML datasets (Messidor, IDRiD, DIARETDB0, and DIARETDB1). We trained the U-Net model on DRIVE datasets and it resulted best BV segmentation in (STARE, CHASEDB1, and HRF). Training U-Net on IDRiD in case of EX, and HM, gives best results in testing the others.

In training the U-Net on MA, It gave training ACC of nearly 74.6% and loss 0.569 by utilizing different optimizers such as Adam, Adamax, Adagard, RmsProp and SGD with different learning rate (lr). Therefore, we had to customize the U-Net hyper parameters as shown in table 8 and train it again on e-ophtha dataset to detect the MA in other eight datasets. The main advantage of the e-ophtha dataset is that it includes the images that only contain the MA signs and detected as DR. On the other hand, the IDRiD is a ML dataset in which each image contains at least 3 signs.

TABLE 8. Training ACC and Loss of U-Net model based on the hyperparameter values.

GT(DS)	Training ACC(%)	Training Loss	Item	Parameters
IDRiD	74	0.569	Optimizer DO steps per epoch Epochs batch size loss Classifier	Adam ($lr = le - 3$) 0.1 100 30 16 binary cross entropy sigmoid
IDRiD	74.56	0.6655	Optimizer DO steps per epoch batch size Classifier	Adagrad ($lr = le - 6$) 0.1 100 16 sigmoid
IDRiD	74.68	0.556	Optimizer	SGD($lr = 0.01, decay = le - 6, momentum = 0.9$), RmsProp ($lr = le - 4$)
IDRiD	69.92	0.624	Optimizer	Adamax ($lr = le - 6$)
DIARETDB1	95	0.05	Optimizer DO steps per epoch batch size Classifier	Adam ($lr = le - 3$) 0.1 100 16 sigmoid
e-ophtha	99.9	0.0016	Optimizer DO steps per epoch batch size Classifier	Adam (lr=le-3) 0.1 50 16 sigmoid

TABLE 9. Analytical comparison between the current systems [21] and the proposed ML-CAD system for analysis and diagnosing health and DR grades.

Item	Abdelmaksoud et al. [21]	The proposed ML-CAD	kou et al. [63]	luo et al. [69]
purpose	visualizing (EX, BV, HM, and MA) and diagnosing health and DR grades	same purpose	Only EX, and MA	only BV segmentation
BV segmentation	matched filter with first-order Gaussian derivative and Coye Filter	CNN (U-Net model)	No	U-Net with DenseNet
EX,MA, and HM segmentation	Morphological operations and thresholding	CNN (U-Net model)	residual U-Net	No
user interaction	Yes	No	No	No
datasets	4 - (2 ML)	8-(4 ML)	3-(2 ML)	2
Preprocessing	Yes	Yes	Yes	Yes
Postprocessing	No	Yes	No	No
Pre- binary classification	No	Yes (Binary SVM classifier to diagnose healthy and DR cases)	No	No
Hand-Crafted Feature Extraction	BP counts, (BV, EX, MA, and HM) ROIs and GLCM	GLRLM on four degree for binary classification, then the same previous system features	No	No
ML Classifier	Yes: (SVM based on CC)	Yes: the same classifier	No	No
Segmentation performance measures	4 (ACC, SEN, SPE, and DSC)	6 (ACC, SEN, SPE, DSC, AUC, and PPV)	4 (ACC, SEN, SPE, and AUC)	5(ACC, SEN, SPE, DSC, and AUC)
Classification performance measures	6 (ACC, SEN, SPE, DSC, AUC, and PPV)	same Performance measures	No	No
ophthalmologist confusion	Yes (segments all the healthy and DR cases)	No (segments only images if DR cases)	Yes	Yes

VI. CONCLUSION

We developed a novel ML-CAD system that can be applied on varied datasets to diagnose diabetic retinopathy grades. We used nine public benchmark datasets; DRIVE, CHASEDB1, STARE, HRF, IDRiD, DIARETDB1, MESSIDOR, and E-ophtha. At first, the proposed system filters and enhances the contrast. Then, it utilizes 11 texture feature descriptors by using GLRLM to determine the normal and DR images. Then, prepares the DR images by postprocessing steps for U-Net model. The U-Net model is trained four times on the four variations (hemorrhages, exudates, Blood Vessels, and microaneurysms). The system extracts 6 features; 2 for BV using GLCM with 11 descriptors and bifurcation point's count, 4 ROIs areas computations. Then, the system

utilized the MLSVM for ML classification depending on the problem transformation. Finally, we computed 6 performance matrices averages of the proposed ML-CAD system. Our system proved that it is reliable and robust. It can be applied on the real world as it can be applied on different color fundus images with different cameras' settings, and different patients.

In the future, we aim to apply the proposed ML-CAD system on another retinal diseases such as glaucoma. In addition, we intent to apply it on the other imaging modalities such as OCTA that can collect different diseases features simultaneously such as Diabetic retinopathy and glaucoma. We want to develop disease-based system not lesion-based system for only one disease.

ACKNOWLEDGMENT

This work was validated and approved by Dr. Hatem Abdelkawy. He is an Assistant Professor with the ophthalmology Department, Faculty of Medicine, Al-Azhar University, Egypt. The authors thank him for his great effort and time.

COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by:

- https://idrid.grand-challenge.org/
- https://drive.grand-challenge.org/
- http://cecas.clemson.edu/ ahoover/stare/
- https://blogs.kingston.ac.uk/retinal/chasedb1/
- https://www5.cs.fau.de/research/data/fundus-images/
- https://www.it.lut.fi/project/imageret/diaretdb1/
- https://www.it.lut.fi/project/imageret/diaretdb0/
- http://www.adcis.net/en/third-party/messidor/
- http://www.adcis.net/en/third-party/e-ophtha/

Ethical approval: was not required as confirmed by the license attached with the open access data.

Conflict of interest: the authors declare no conflict of interest.

REFERENCES

[1] *World Health Organization, Diabetes*. Accessed: Dec. 4, 2020. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1

[2] L. Pizzarello, A. Abiose, T. Ffytche, R. Duerksen, R. Thulasiraj, H. Taylor, H. Faal, G. Rao, I. Kocur, and S. Resnikoff, "Vision 2020: The right to sight: A global initiative to eliminate avoidable blindness," *Arch. Ophthalmol.*, vol. 122, no. 4, pp. 615–620, 2004.

[3] T. Y. Wong and C. Sabanayagam, "Strategies to tackle the global burden of diabetic retinopathy: From epidemiology to artificial intelligence," *Ophthalmologica*, vol. 243, no. 1, pp. 9–20, 2020.

[4] N. Congdon, Y. Zheng, and M. He, "The worldwide epidemic of diabetic retinopathy," *Indian J. Ophthalmol.*, vol. 60, no. 5, p. 428, 2012.

[5] Z. Li, S. Keel, C. Liu, Y. He, W. Meng, J. Scheetz, P. Y. Lee, J. Shaw, D. Ting, T. Y. Wong, H. Taylor, R. Chang, and M. He, "An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs," *Diabetes Care*, vol. 41, no. 12, pp. 2509–2516, Dec. 2018.

[6] T. Hassan, M. U. Akram, B. Hassan, A. Nasim, and S. A. Bazaz, "Review of OCT and fundus images for detection of macular edema," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Sep. 2015, pp. 1–4.

[7] M. U. Saeed and J. D. Oleszczuk, "Advances in retinal imaging modalities: Challenges and opportunities," *World J. Ophthalmol.*, vol. 6, no. 2, pp. 10–19, 2016.

[8] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[9] S. El-Sappagh, T. Abuhmed, S. M. R. Islam, and K. S. Kwak, "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data," *Neurocomputing*, vol. 412, pp. 197–215, Oct. 2020.

[10] S. El-Sappagh, T. Abuhmed, and K. S. Kwak, "Alzheimer disease prediction model based on decision fusion of CNN-BiLSTM deep neural networks," in *Intelligent Systems and Applications*, K. Arai, S. Kapoor, and R. Bhatia, Eds. Cham, Switzerland: Springer, 2020, pp. 482–492.

[11] G. Lim, V. Bellemo, Y. Xie, X. Q. Lee, M. Y. T. Yip, and D. S. W. Ting, "Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: A review," *Eye Vis.*, vol. 7, no. 1, pp. 1–13, Dec. 2020.

[12] T. Abuhmed, S. El-Sappagh, and J. M. Alonso, "Robust hybrid deep learning models for Alzheimer's progression detection," *Knowl.-Based Syst.*, vol. 213, Feb. 2021, Art. no. 106688.

[13] K. Y. Tey, K. Teo, A. C. Tan, K. Devarajan, B. Tan, J. Tan, L. Schmetterer, and M. Ang, "Optical coherence tomography angiography in diabetic retinopathy: A review of current applications," *Eye Vis.*, vol. 6, no. 1, pp. 1–10, 2019.

[14] S. M. S. Islam, M. M. Hasan, and S. Abdullah, "Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images," 2018, *arXiv:1812.10595*. [Online]. Available: http://arxiv.org/abs/1812.10595

[15] A. C. Tan, G. S. Tan, A. K. Denniston, P. A. Keane, M. Ang, D. Milea, U. Chakravarthy, and C. M. G. Cheung, "An overview of the clinical applications of optical coherence tomography angiography," *Eye*, vol. 32, no. 2, pp. 262–286, 2018.

[16] A. Biran, P. S. Bidari, and K. Raahemifar, "Automatic method for exudates and hemorrhages detection from fundus retinal images," *Int. J. Comput. Inf. Eng.*, vol. 10, no. 9, pp. 1599–1602, 2016.

[17] G. L. Atlas and K. Parasuraman, "Detection of retinal hemorrhage from fundus images using ANFIS classifier and MRG segmentation," *Biomed. Res.*, vol. 29, no. 7, pp. 1489–1497, 2018.

[18] J. I. Orlando, E. Prokofyeva, and M. B. Blaschko, "A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 16–27, Jan. 2017.

[19] M. K. Fadafen, N. Mehrshad, and S. M. Razavi, "Detection of diabetic retinopathy using computational model of human visual system," *Biomed. Res.*, vol. 29, no. 9, pp. 1956–1960, 2018.

[20] D. W. Safitri and D. Juniati, "Classification of diabetic retinopathy using fractal dimension analysis of eye fundus image," *AIP Conf.*, vol. 1867, Aug. 2017, Art. no. 020011.

[21] E. AbdelMaksoud, S. Barakat, and M. Elmogy, "A comprehensive diagnosis system for early signs and different diabetic retinopathy grades using fundus retinal images based on pathological changes detection," *Comput. Biol. Med.*, vol. 126, Nov. 2020, Art. no. 104039.

[22] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Invest. Ophthalmol. Vis. Sci.*, vol. 57, no. 13, pp. 5200–5206, 2016.

[23] V. Bellemo, Z. W. Lim, G. Lim, Q. D. Nguyen, Y. Xie, M. Y. Yip, H. Hamzah, J. Ho, X. Q. Lee, W. Hsu, M. L. Lee, L. Musonda, M. Chandran, G. Chipalo-Mutati, M. Muma, G. S. W. Tan, S. Sivaprasad, G. Menon, T. Y. Wong, and D. S. W. Ting "Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in africa: A clinical validation study," *Lancet Digit. Health*, vol. 1, no. 1, pp. e35–e44, 2019.

[24] R. F. Mansour, "Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy," *Biomed. Eng. Lett.*, vol. 8, no. 1, pp. 41–57, Feb. 2018.

[25] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, I.-H. Ra, and M. Alazab, "Early detection of diabetic retinopathy using PCA-firefly based deep learning model," *Electronics*, vol. 9, no. 2, p. 274, Feb. 2020.

[26] M. T. Hagos and S. Kant, "Transfer learning based detection of diabetic retinopathy from small dataset," 2019, *arXiv:1905.07203*. [Online]. Available: http://arxiv.org/abs/1905.07203

[27] B. Tymchenko, P. Marchenko, and D. Spodarets, "Deep learning approach to diabetic retinopathy detection," 2020, *arXiv:2003.02261*. [Online]. Available: http://arxiv.org/abs/2003.02261

[28] K. Xu, D. Feng, and H. Mi, "Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image," *Molecules*, vol. 22, no. 12, p. 2054, Nov. 2017.

[29] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Comput. Sci.*, vol. 90, pp. 200–205, Jan. 2016.

[30] M. M. Butt, G. Latif, D. N. F. A. Iskandar, J. Alghazo, and A. H. Khan, "Multi-channel convolutions neural network based diabetic retinopathy detection from fundus images," *Procedia Comput. Sci.*, vol. 163, pp. 283–291, Jan. 2019.

[31] Y.-H. Li, N.-N. Yeh, S.-J. Chen, and Y.-C. Chung, "Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network," *Mobile Inf. Syst.*, vol. 2019, pp. 1–14, Jan. 2019.

[32] M. A. Rahman, S. Liu, S. Lin, C. Wong, G. Jiang, and N. Kwok, "Image contrast enhancement for brightness preservation based on dynamic stretching," *Int. J. Image Process.*, vol. 9, no. 4, p. 241, 2015.

[33] E. Abdel-Maksoud, M. Elmogy, and R. Al-Awadi, "Brain tumor segmentation based on a hybrid clustering technique," *Egyptian Informat. J.*, vol. 16, no. 1, pp. 71–81, Mar. 2015.

[34] X. Tang, "Texture information in run-length matrices," *IEEE Trans. Image Process.*, vol. 7, no. 11, pp. 1602–1609, Nov. 1998.

[35] E. AbdelMaksoud, S. Barakat, and M. Elmogy, "A multi-label computer-aided diagnoses system for detecting and diagnosing diabetic retinopathy," in *Proc. 14th Int. Conf. Comput. Eng. Syst. (ICCES)*, Dec. 2019, pp. 373–386.

[36] S. V. da Rocha, G. B. Junior, A. C. Silva, A. C. de Paiva, and M. Gattass, "Texture analysis of masses malignant in mammograms images using a combined approach of diversity index and local binary patterns distribution," *Expert Syst. Appl.*, vol. 66, pp. 7–19, Dec. 2016.

[37] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020.

[38] V. Chandore and S. Asati, "Automatic detection of diabetic retinopathy using deep convolutional neural network," *Int. J. Adv. Res. Ideas Innov. Technol.*, vol. 3, no. 4, pp. 633–641, 2017.

[39] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. Kruthiventi, and R. V. Babu, "An introduction to deep convolutional neural nets for computer vision," in *Deep Learning for Medical Image Analysis*, S. K. Zhou, H. Greenspan, and D. Shen, Eds. New York, NY, USA: Academic, 2017, ch. 2, pp. 25–52.

[40] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Into Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018.

[41] M. Ahmadi, S. Vakili, J. M. P. Langlois, and W. Gross, "Power reduction in CNN pooling layers with a preliminary partial computation strategy," in *Proc. 16th IEEE Int. New Circuits Syst. Conf. (NEWCAS)*, Jun. 2018, pp. 125–129.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[43] C. Lam, D. Yi, M. Guo, and T. Lindsey, "Automated detection of diabetic retinopathy using deep learning," *AMIA Summits Transl. Sci.*, vol. 2018, no. 1, p. 147, 2018.

[44] D. Gadkari, *Image Quality Analysis Using GLCM*. Orlando, FL, USA: Univ. Central Florida, 2004.

[45] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 374, no. 2065, Apr. 2016, Art. no. 20150202.

[46] T. Kohler, A. Budai, M. F. Kraus, J. Odstrcilik, G. Michelson, and J. Hornegger, "Automatic no-reference quality assessment for retinal fundus images using vessel segmentation," in *Proc. 26th IEEE Int. Symp. Comput.-Based Med. Syst.*, Jun. 2013, pp. 95–100.

[47] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2538–2548, Sep. 2012.

[48] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "DIARETDB0: Evaluation database and methodology for diabetic retinopathy algorithms," in *Machine Vision and Pattern Recognition Research Group*, vol. 73. Lappeenranta, Finland: Lappeenranta Univ. Technol., 2006, pp. 1–17.

[49] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "The diaretdb1 diabetic retinopathy database and evaluation protocol," in *Proc. BMVC*, vol. 1, 2007, pp. 1–10.

[50] M. H. Goldbaum. (1975). *STructured Analysis of the Retina Project*. Accessed: Dec. 4, 2020. [Online]. Available: http://cecas.clemson.edu/~ahoover/stare

[51] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed database: The messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.

[52] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[53] P. Porwal *et al.*, "Diabetic retinopathy: Segmentation and grading challenge workshop," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, Oct. 2018, p. 1.

[54] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, and A. Chabouis, "Teleophta: Machine learning and image processing methods for teleophthalmology," *Irbm*, vol. 34, no. 2, pp. 196–203, 2013.

[55] V. Thada and J. Vivek, "Comparison of Jaccard, dice, cosine similarity coefficient to find best fitness value for Web retrieved documents using genetic algorithm," *Frontiers Comput. Sci.*, vol. 2, no. 4, pp. 202–205, 2013.

[56] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[57] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informat.*, vol. 16, no. 1, pp. 1–25, Aug. 2020.

[58] J. V. B. Soares, J. J. G. Leandro, R. M. Cesar, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1214–1222, Sep. 2006.

[59] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov, "Trainable COSFIRE filters for vessel delineation with application to retinal images," *Med. Image Anal.*, vol. 19, no. 1, pp. 46–57, Jan. 2015.

[60] X. Gao, Y. Cai, C. Qiu, and Y. Cui, "Retinal blood vessel segmentation based on the Gaussian matched filter and U-Net," in *Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, 2017, pp. 1–5.

[61] D. Adapa, A. N. J. Raj, S. N. Alisetti, Z. Zhuang, and G. Naik, "A supervised blood vessel segmentation technique for digital fundus images using Zernike moment based features," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0229831.

[62] Q. Yan, Y. Zhao, Y. Zheng, Y. Liu, K. Zhou, A. Frangi, and J. Liu, "Automated retinal lesion detection via image saliency analysis," *Med. Phys.*, vol. 46, no. 10, pp. 4531–4544, Oct. 2019.

[63] C. Kou, W. Li, Z. Yu, and L. Yuan, "An enhanced residual U-Net for microaneurysms and exudates segmentation in fundus images," *IEEE Access*, vol. 8, pp. 185514–185525, 2020.

[64] P. Khojasteh, B. Aliahmad, and D. K. Kumar, "Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms," *BMC Ophthalmol.*, vol. 18, no. 1, pp. 1–13, Dec. 2018.

[65] E. A. A. Maksoud, M. Ramadan, S. Barakat, and M. Elmogy, "A computer-aided diagnoses system for detecting multiple ocular diseases using color retinal fundus images," in *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 19–52.

[66] A. Pakrashi, D. Greene, and B. MacNamee, "Benchmarking multi-label classification algorithms," in *Proc. 24th Irish Conf. Artif. Intell. Cogn. Sci. (AICS)*, Dublin, Ireland, Sep. 2016, 2016, pp. 1–13.

[67] *Scikit Multiclass and Multilabel Algorithms*. Accessed: Dec. 4, 2020. [Online]. Available: https://scikit-learn.org/stable/modules/multiclass.html

[68] M. Pushpa and S. Karpagavalli, "Multi-label classification: Problem transformation methods in tamil phoneme classification," *Procedia Comput. Sci.*, vol. 115, pp. 572–579, Jan. 2017.
[69] Z. Luo, Y. Zhang, L. Zhou, B. Zhang, J. Luo, and H. Wu, "Micro-vessel image segmentation based on the AD-UNet model," *IEEE Access*, vol. 7, pp. 143402–143411, 2019.

EMAN ABDELMAKSOUD received the B.Sc. degree from the Faculty of Computers and Information, Zagazig University, Zagazig, Egypt, and the M.Sc. degree from the Faculty of Computers and Information, Mansoura University, in 2016, where she is currently pursuing the Ph.D. degree with the Information Systems Department. She has authored approximately ten research publications in journals, book chapters, and conference proceedings. Her current research interests include computer vision, medical image analysis, artificial intelligence, machine learning, and biomedical engineering.

SHAKER EL-SAPPAGH received the bachelor's degree in computer science from the Information Systems Department, Faculty of Computers and Information, Cairo University, Egypt, in 1997, the master's degree from Cairo University, in 2007, and the Ph.D. degree in computer science from the Information Systems Department, Faculty of Computers and Information, Mansura University, Mansura, Egypt, in 2015. In 2003, he joined the Department of Information Systems, Faculty of Computers and Information, Minia University, Egypt, as a Teaching Assistant. Since June 2016, he has been with the Department of Information Systems, Faculty of Computers and Information, Benha University, as an Assistant Professor. He is currently a Postdoctoral Fellow with the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain. He is very interested in diseases' diagnosis and treatment researches. He has publications in clinical decision support systems and semantic intelligence. His current research interests include machine learning, medical informatics, (fuzzy) ontology engineering, distributed and hybrid clinical decision support systems, semantic data modeling, fuzzy expert systems, and cloud computing. He serves as a Reviewer for many journals.

SHERIF BARAKAT received the B.Sc. and M.Sc. degrees from the Faculty of Science, Mansoura University, Mansoura, Egypt, and the Ph.D. degree from the Faculty of Science, Helwan University, Helwan, Egypt, in 2003. He is currently a Professor with the Information Systems Department, Faculty of Computers and Information, Mansoura University. He has authored/coauthored over 50 research publications in peer-reviewed reputed journals, book chapters, and conference proceedings. His current research interests include machine learning, pattern recognition, and artificial intelligence.

TAMER ABUHMED received the Ph.D. degree in information and telecommunication engineering from Inha University, in 2012. He is currently an Assistant Professor with the College of Computing, Sungkyunkwan University, South Korea. His research interests include biomedical applications, information security, network security, the Internet security, and machine learning and its application to medical, security, and privacy problems.

MOHAMMED ELMOGY (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Faculty of Engineering, Mansoura University, Mansoura, Egypt, and the Ph.D. degree from the Informatics Department, MIN Faculty, Hamburg University, Hamburg, Germany, in 2010. From July 2016 to August 2019, he worked as a Visiting Researcher with the Department of Bioengineering, University of Louisville, Louisville, KY, USA. He is currently an Associate Professor with the Information Technology Department, Faculty of Computers and Information, Mansoura University. He advised more than 30 master and doctoral graduates. He has authored/coauthored over 200 research publications in peer-reviewed reputed journals, book chapters, and conference proceedings. His current research interests include computer vision, medical image analysis, machine learning, pattern recognition, and biomedical engineering. He is a Professional Member of the ACM Society. He served as a Technical Program Committee Member for many workshops and conferences. He has also served as a Reviewer for various international journals.