International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

A Critical Insight into Pragmatic Manifestation of Diabetic Retinopathy Grading and Detection

Muhammad Samer Sallam, Rashidah Funke Olanrewaju, Ani Liza Asnawi

Abstract: ***Nowadays, artificial intelligence applications invade all of the fields including medical applications field. Deep learning, a subfield of artificial intelligence, in particular, Convolutional Neural Networks (CNN), have quickly become the first choice for processing and analyzing medical images due to its performance and effectiveness. Diabetic retinopathy is a vision loss disease that infects people with diabetes. This disease damages the blood vessels in the retina, hence, leads to blindness. Due to the sensitivity and complications involved in managing diabetics, designing and developing automated systems to detect and grade diabetic retinopathy is considered one of the recent research areas in the world of medical image applications. In this paper, the aspects of deep learning field related to diabetic retinopathy have been discussed. Various concepts in deep learning including traditional Artificial Neural Network (ANN) algorithm, ANN drawbacks in context of computer vision and image processing applications, and the best algorithm to overcome ANN drawbacks, CNN, have been elucidated along with the architecture. The paper also reviews an extensive summary of some works in the current research trend and future applications of the DL algorithms in medical image analysis for DR detection and grading. Furthermore, various research gabs related to building such automated systems for medical image analysis have been conferred – such as imbalance dataset which is considered one of the main performance issues that should be handled, the need of high performance computational resources to train deep and efficient models and others. This is quite beneficial for researchers working in the domain of medical image analysis to handle DR.***

Keywords: Convolutional Neural Networks, Retinal Fundus Images Classification, Diabetic Retinopathy

I. INTRODUCTION

Diabetes is a chronic disease causing the sugar (glucose) level in the blood to arrive to dangerous high levels. Glucose is vital to the body because it's an important source of energy for the cells that make up muscles and tissues. Meanwhile, insulin, a hormone that comes from the pancreas, is secreted into the bloodstream to burn the sugar in order to lower it in the blood.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Muhammad Samer Sallam*, , Department of Computer and Information Engineering, International Islamic University, Kuala Lumpur, Malaysia. Email: samersallam92@gmail.com
Rashidah Funke Olanrewaju, Department of Computer and Information Engineering, International Islamic University, Kuala Lumpur, Malaysia. Email: frashidah@iium.edu.my
Ani Liza Asnawi, Department of Computer and Information Engineering, International Islamic University, Kuala Lumpur, Malaysia. Email: aniliza@iium.edu.my

Insufficient production of insulin or an inability of the body to correctly use insulin causes diabetes. Mainly there are two types of this disease (insulin-dependent and non-insulin-dependent diabetes, or juvenile onset and adult onset diabetes).

The number of people diagnosed to have diabetes increases over last years. According to International Diabetes Federation (IDF) Atlas 7th edition in 2015 more than 415 million people worldwide are affected by diabetes. Also, According to IDF Western Pacific (WP) Atlas 8th Edition in 2017, there are approximately 158.8 million adults in the age range 20-79 years were living with diabetes in the IDF WP region, representing 9.5% of the population. More than 54% of the cases are without diagnoses. Approximately two thirds of the cases of adults with diabetes in the WP region live in urban areas. In Malaysia, there were over 3.492.600 cases of diabetes in 2017, and the number increases according to the official website of IDF [1]

Through time, diabetes could cause eye diseases which eventually lead to the blindness. One of these diseases is diabetic retinopathy (DR), also known as diabetic eye disease. DR is considered the main cause of blindness in the mid-age. Early diagnosis and detection of DR help provide the required treatment which could prevent or at least delay blindness. Most guidelines advice to do annual screening for no DR or mild DR and repeat screening in 6 months for moderate DR. Screening for DR is a very important step to detect the effects of DR since DR most often has no early warning signs. DR disease evolves through time and mainly has two stages. In the first stage, which is called non- proliferative diabetic retinopathy (NPDR), there are no symptoms and the patient can see clearly without any problems. In this stage, DR is detected by fundus photography [2]. This photography helps show narrowing or blocked retinal blood vessels. In the second stage, which is called proliferative diabetic retinopathy (PDR), new abnormal blood vessels start to form and appear at the back of the eye. These blood vessels burst and bleed. In most of cases when bleeding happens for the first time, it will not be severe where it will leave some blood spots on the patient visual field. These spots disappear after few hours. However, if the disease is not treated, the bleeding could happen again after a few days or weeks with bigger blood spots which could blur the vision. In severe bleeding case, the patient will be able to discriminate between dark and light areas only, and the spots need longer periods to disappear reaching to months or years in some cases.

Fig. 1 shows DR severity categories where specialists classify the severity into four categories (1 - Mild DR 2 - Moderate DR 3 - Severe DR 4 - Proliferative DR.).

Understanding the retinal fundus images is not an easy task, and it requires a specialist to check these images manually to check if there are any strange spots or signs of DR which makes the processing time consuming even for well-trained doctors [2]. With this hard to be detected, quick to evolve, and slow to be diagnosed disease, the need for an automated system to detect this disease cannot be underestimated.

In order to automate the process of DR detection, researchers have found that artificial intelligence (AI) is a very promising direction especially after the successful applications of Artificial Neural Networks (ANN) in different domains [3] and [4]. Nowadays, computer vision field is one of the current research fields benefiting from the great abilities of a special type of ANN which is known as Convolutional Neural Network (CNN). CNN is a class of deep feed forward ANN mainly used in analysing images contents. The main strength point in CNN is its two main components where one component is for features extraction, and the other one is for patterns recognition. These two components are trained together in order to achieve the required task, which means, there is no need for human intervention in order to choose and design features for recognition. Usually features extraction process is considered one of the hardest tasks in traditional machine learning projects and researches. CNN is one of the neural network architecture that overcomes this hard and essential process for any machine learning system. Therefore, this paper mainly focuses on CNN usage for diabetic retinopathy detection and grading problem.

The rest of this review article is organized as follow. In section 2, literature review is discussed where in section 2.1 there is a brief introduction about artificial intelligence, machine learning and deep learning. In section 2.2, related works in the field of machine learning and deep learning are discussed. In section 2.3, the most important research gaps in the field are highlighted. Finally, section.3 is the conclusion of this review article.

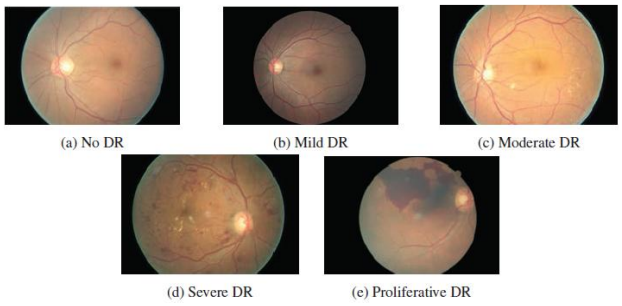

Fig. 1 DR severity categories [5]

II. LITERATURE REVIEW

A. Background

Artificial intelligence, AI, is the science that automates the tasks and functions which require human intelligence. An AI system could be implemented through classical programming by building a rules-based system for the problem under study, or it can be implemented through teaching the machine using algorithms under machine learning science. Deep learning is a subfield of machine learning and the "deep" refers to a model of multiple successive layers used to represent the knowledge. This model is known as artificial neural network ANN. Fig. 2 explains the relationship between these related sciences.

In traditional AI systems, the program data and rules are provided to get outputs, whereas, in machine learning systems, we require some amount of data, the output expected from the data, and a way to train the machine on this data in order to get the rules, and then these rules will be used to process the new data. The different concepts of traditional AI and machine learning systems can be seen in Fig. 3.

It has been mentioned that ANN is the core algorithm in deep learning field. ANN could be defined as a biologically-inspired model of building computer programs that have the ability to learn and create connections according to the available data. ANN consists of a set of overlapping layers above each other. The three main layers in any neural network are the input layer, the hidden layer, and the output layer.

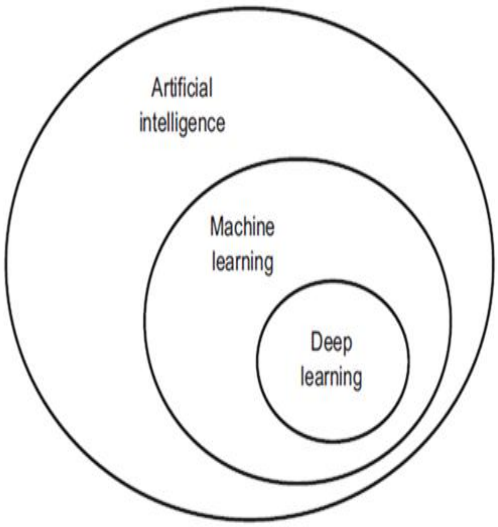

Fig. 2 Artificial intelligence, machine learning and deep learning fields [6]

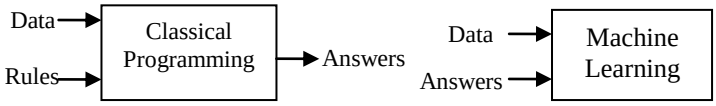

Fig. 3 Classical Programming vs. Machine Learning [6]

As can be seen from Fig. 4, these layers consist of a group of neurons. The neurons in the input layers receive the input **X**. This input signal will flow over the hidden layers one after one until it reaches the output layer which produces the expected output $\hat{\mathbf{y}}$.

In order to make the neural network works, training process should be applied. The training process depends on a training set, an activation function, a loss function, and an optimization algorithm.

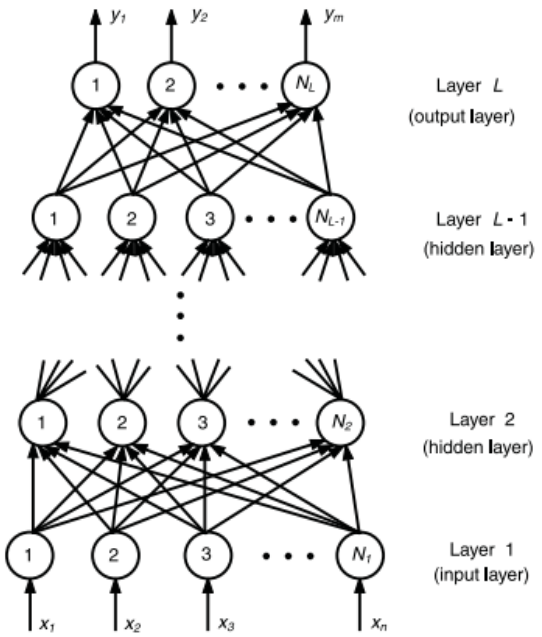

Fig. 4 The standard diagram of the neural network [7]

The first one is the training set where each record of this set has its own properties, which represent the value of **X** - the input of the neurons in the input layer, and the value of the correct output for this record y. Each **X** record will be fed into the neurons of the first hidden layer where each neuron has a set of weights **w** and a bias **b** which can change during the learning process. The weighted average of the values of vector **X** is calculated during every iteration of the training process. After that, the result is processed by a non-linear activation function to have the output of each neuron as it is shown in Fig. 5.

Activation function is one of the main elements of the neural network. The nonlinear feature of activation gives the possibility of performing many complex functions during the training process. Activation function is selected based on its effect on the speed of the training process. There are two types of activation function - saturated and non-saturated. Saturated activation functions are sigmoid and tanh, whereas non-saturated are ReLU and its variants.

Now, loss function represents the difference between the result of neural network and the optimal solution to the problem. The goal of learning process is to reduce the value of the loss function by changing the value **b** and **w** during every iteration of learning process.

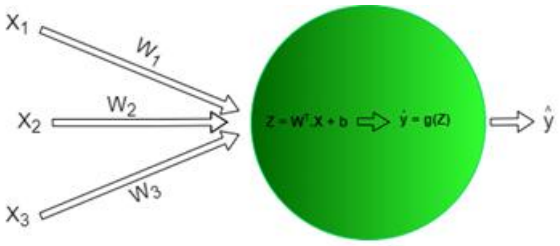

Fig. 5 The structure of the neuron

Finally, an optimization algorithm is used to find the best value of **b** and **w**. This algorithm aims to reduce the value of the loss function during the training process.

One of the main drawbacks of the traditional ANN is that it is fully connected network which increases the number of trained parameters rapidly when the size of the input or the number of layers and neurons increase. Also, ANN receives the input as a vector which makes it not suitable for computer vision applications since the image is represented as a 2-d or 3-d matrix. One way to overcome this problem is by vectorizing the image – convert the 2-d matrix into 1-d vector. However, vectorizing the image leads to lose the spatial relationship between pixels in the image which could degrade the performance of the neural network.

In order to overcome these drawbacks, convolutional neural network (CNN) has been invented. The term convolution is a mathematical operation giving the integral of the point-wise multiplication of the two vectors as a function of the amount that one of the original vectors is translated [6]. CNN is one of the types of neural networks that works particularly in image processing, and has shown great results in solving problems of classification and recognition of faces, classification of handwritten numbers, and other computer vision application. The most important features of CNN are its ability to handle very large training sets, its ability to receive the image as a matrix at the input which protects the spatial relationships between the pixels, and finally using neurons of limited number of weights through the network. These networks are characterized by their structure and characteristics but share the same general features of the previously mentioned neural networks: the neurons, the input layer, the hidden layers, the activation function, the loss function, and the optimization algorithm.

CNN contains two main parts: (i) convolutional part which is responsible for feature extraction and (ii) prediction part which is responsible for producing the final output of CNN. Referring to Fig. 6, the convolutional part of CNN mainly contains the input layer, convolutional layer, activation layer, and pooling layer whereas the prediction layer contains the fully connected layers.

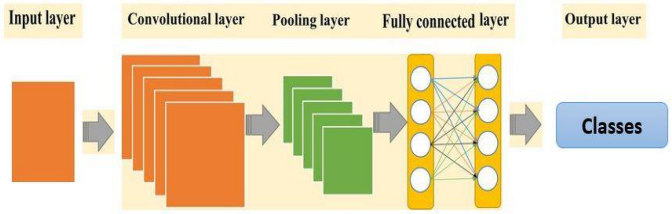

Fig. 6 The structure of CNN [8]

B. Related works:

Once it was possible to capture medical images and load these images to computers, researchers started to think how image analysis could be automated [9]. At the beginning, from 1970s to 1990s, rule-based systems to achieve specific targets were implemented using basic pixel processing (edge detector for example). These systems contained multiple if-else statements, which made the performance very brittle. These systems were known mainly as Good Old-Fashion Artificial Intelligence (GOFAI) [9].

By the end of 1990s supervised learning techniques started to appear where datasets of medical images were used to train and develop analysis systems. In this era, the concept of features extraction appeared where features by human must be extracted from medical images, and then they be fed to the trained system. This era is considered the shift from system built totally by the human to systems trained by features extracted from the datasets.

Features extraction requires domain knowledge and technical skills, which makes the process hard and sometimes not achievable. Therefore, the logical direction to think in is to make computers learns the features that represent datasets.

This idea is the core of deep learning science and algorithms; models should learn the features first, and then map these features into a decision (healthy/unhealthy for example).

The most promising type of models in medical image analysis is CNN. The first work to embark on CNN in medical application was in 1995 by [10]. However, the CNN achieved the first success in a real world application in 1998 in LeNet [11] where the CNN was used for hand-written digit recognition. Although CNNs achieved success in some applications, they did not attract a lot of researchers in that period of time. Training of these CNNs requires huge datasets and high computational resources and both of these two requirements were not present. Therefore, the winter of neural networks started and remained until December 2012 when AlexaNet (a CNN architecture developed and trained by [12]) won the ImageNet competition with a large margin compared with the other machine learning models.

The community of medical image analysis researchers has noted this great development, and the era of transition from handcrafted features to systems that use datasets to learn features has started. Firstly, most of applications including diabetic retinopathy detection started to appear in workshops and conferences articles, and later on in journals articles. Most of CNN applications in medical image analysis appeared in 2015, 2016, and 2017 according to a survey article published in July 2017 [4]. It is clear from the Fig. 7 that number of publications about using CNN in medical image analysis is increasing from year to year since 2015.

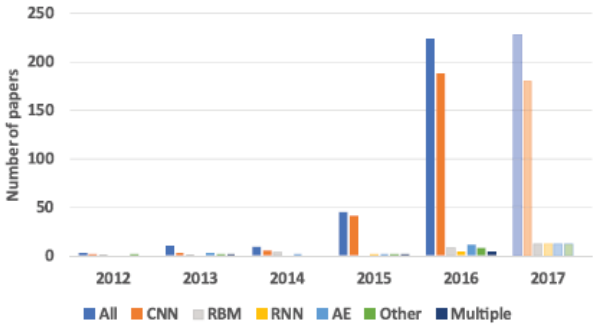

Fig. 7 Number of the papers published in the field of medical images analysis applications from 2012 [4]

In the next section the research works related to diabetic retinopathy detection and grading will be discussed and summarized. Mainly, these works have been divided into two directions: (i) machine learning direction and (ii) deep learning direction.

- **Traditional machine learning for detection of diabetic retinopathy**:

These methods focus on extracting some hand tuned features from small scale datasets in order to do the classification process. Traditional methods are considered from the old style of learning. Therefore, this section sheds the light on some aspects of these methods without delving in the details.

In [13] the authors have used the concept of ensemble learning between multiple classifiers like logistic regression, SVM, KNN, etc in order to build a classifier able to classify the image of retinal fundus into health or not healthy. To train this classifier, a traditional set of features have been extracted like image quality assessment, the diameter of optic disk, and lesion exudates. Messidor dataset of 1200 images which is publicly available has been used for the training. The sensitivity of this classifier has been in the range of 80%

Another research in this area has been done by [14]. In this research, three traditional algorithms have been implemented and tested to classify fundus image into healthy or not healthy. To train these classifiers, a feature set from exudates detection have been extracted from a dataset of 100 fundus images where 80 images used for training and the remaining for testing. To extract these features, RGB image has been converted into LUV color space after segmenting the necessary part of image. After that, statistical feature like mean, standard deviation, and edge strength have been computed. The authors have reported the accuracy for the three classifiers where ELM has outperformed all of the other algorithms and achieved 90% of accuracy.

[15], the authors have developed three classifiers models PNN, SVM, and Bayesian model. To train these models, a feature set from a dataset of 139 fundus images has been extracted. This set has been extracted by traditional method like green channel manipulation, thresholding, and morphological operation like dilation to compute the radius and the diameter of the optic disk. The accuracy has been for PNN 89.65%, for SVM 94.4%, and for Bayesian model 97.6%.

Studies in [16] and [17] have focused on segmenting the exudates in the fundus image firstly using Fuzzy C-means clustering algorithm. After segmentation, traditional features like color, size, mean, standard deviation and contrast. between areas have been extracted. After getting the features set a simple ANN has been used and trained to classify the fundus images. For the first article, the sensitivity and specificity for exudates detection are 86% and 99% respectively whereas for the second article they have been 92% and 82% respectively. Although the evaluation metrics have been relatively high, it should be mentioned that the tests have been done on very small dataset, less than one hundred images. From previous review, it is clear that the direction in these traditional methods mainly focuses on exudates segmentation. Then, some features from the segmented area are collected. These features are fed into traditional classifier like SVM, KNN, or decision tree. Segmentation step at the beginning could harm the classification accuracy since it could include unnecessary area or it could remove important parts from the images. Also, computing simple features like mean and standard deviation are not robust since it could be badly affected if the light conditions are different between images.

Finally, designing a feature set considered suitable for different datasets of different people from different countries is semi impossible. Therefore, there is a need to an automated way to determine and compute these features.

- **Deep learning for detection of diabetic retinopathy*:***

These methods have focused on using deep neural network architectures like convolutional neural networks with large scale dataset in order to do the classification process.

[5] have developed a CNN from scratch to detect DR from fundus images. The CNN structure has contained ten convolution layers with increasing number of filters when moving toward deeper layers. These layers, which represent the feature extraction part, have been followed by two fully connected layers and finally a softmax layer for classification. All of the layers have been provided with ReLU activation function. All of the convolutional layers have been followed by max pooling layer. A dropout layer with a probability of 0.5 has been used after the last convolutional layer and the first fully connected layer to reduce the over-fitting. Kaggle dataset, which is published online, has been used for the training. Color normalization has been done on the images to mitigate the effect of lighting between different images .They have worked on multi class classification problem, where the input is a fundus image and the output could be one of these five classes (normal, mild DR, moderate DR, severe DR, and proliferative DR classes). The network has been trained using stochastic gradient descent with Nesterov momentum, and categorical cross-entropy function has been used as a loss function. The authors have published the confusion matrix and showed the accuracy 75 %, the specificity 95 % and the sensitivity 30 %. The dataset set is highly imbalanced where most of the images are from the first class. Getting such high value of specificity is because the model could classify most of the images to be from the first class and this is clear from the low value of sensitivity. Working on grading disease is not an easy problem. This paper is considered one of few papers handling the problem from this perspective. However, it is clear that the sensitivity is very low since the authors have not considered the problem of imbalanced dataset when the model has been trained.

[18] The authors have worked on grading of diabetic retinopathy and diabetic macular edema. In this research, they have advocated the concept of transfer learning from a very common neural network know as Inception-v3 to do multiple binary classification where the image could be considered (1) moderate or worse DR, (2) sever or worse DR, etc. For training, stochastic gradient descent algorithm has been used where the neural network has been trained on a dataset of 28175 retinal fundus images. These images have been graded for diabetic retinopathy by a board of 54 US licensed ophthalmologists between May and December 2015. In order to speed up the training process, batch normalization layer [19]has been used. For weights initialization, weights of the Inception trained on ImageNet [20] dataset have been used. In order to evaluate the neural network, two datasets have been used where the first one is the EyePACS-1 which contains 9963 images from 4997 patients, and the second one is Messidor-2 data which contains 1748 images from 874 patients. For these two datasets respectively the sensitivity has been 90.3% and 87% and the specificity has been 98.1% and 98.5%. It is clear from the results, that using transfer learning from a pertained neural network has achieved high results from sensitivity and specificity perspective. Nothing about confusion matrix has been mentioned. Although the results could be considered satisfying, further research is necessary to evaluate the performance of the neural network in order to check whether the output leads to a positive outcome compared with current ophthalmologists.

In [21], the researchers have worked on early detection of diabetic retinopathy by detection the exudates in the retinal fundus images. Exudates existence is considered an early sign of diabetic retinopathy. Usually these exudates will not be in retinal blood vessels, bright borders, or in optic disk. Therefore, the authors has integrated theirs algorithm with segmentation step before detection step. In the segmentation step the optic disk and blood vessels has been detected and removed from the image which in turn has enhanced the performance since it has decreased false positive rate. In the detection step the problem has been formulated as a binary pixel classification problem. To solve this problem CNN has been used and trained to classify each pixel into exudate or non-exudate pixel. The CNN architecture has had 4 convolutional layers and 1 fully connected layer to give the classification result. In order to train and test the neural network a dataset of 50 retinal fundus images has been used. In this datasets, blood vessels and optic disk are marked along with pathological changes like dot and blot hemorrhages, and hard and soft exudates. The dataset is split into two disjoint datasets for training and testing. For evaluation, F1 measure has been computed and the value achieved has been 78%. Using segmentation stage to improve the performance of the classification is a very good idea since it decreases false positive and lets the algorithm process the important areas only. However, this step increases the computational complexity of the algorithm. For evaluation side the author has reported only F1 measure on a dataset of 50 images. The number of images is very small and it is hard to trust in the performance and generalize the evaluation results on a dataset of 50 images only.

In the previous research [21], the authors have detected hemorrhages which is an early sign of diabetic retinopathy, and they have formulated the problem as pixel classification problem, in [22] the researchers have advocated the same direction but they have formulated the problem as patch classification problem. In order to classify these patches, they have built and trained a CNN of five convolutional layers with ReLU as an activation function. At the end of the neural network there has been a fully connected layer of 1024 neurons followed by a logistic regression unit which produces 0 if there are no hemorrhages in the patch and 1 vice versa. In order to train the neural network, a subset of kaggle dataset has been used. This subset has contained 6679 images. This dataset has been split into 60-20-20 training, mentoring, and test set. Also, for test purpose the publicly available Messidor dataset of 120 images has been used. The images inside these two datasets have been split firstly into positive and negative samples. From these samples, patches of 41*41 resolutions have been cropped from the images. In order to improve the training efficiency and speed up the training, the authors have proposed a selection algorithm to select the samples during the training process.

In this algorithm, misclassified samples have been selected with a higher probability in the next training iteration. In order to guarantee that the neural network will not over fit, the mentoring dataset has been used to monitor on line the performance of the neural network. Using this algorithm for sample selection during the training decreases the training time from 170 epochs into 60 epochs where the training process has been stopped when the performance reach to a maximum level. For performance evaluation, ROC values have been reported for the two test datasets (89.4% and 97.2%). It is clear that the selection algorithm has speed-up the training process in a very efficient way. Also, this have used decision tree classification algorithm in order to do the classification task with binary cross entropy as a loss function. In order to evaluate thesystem, they have used 5-fold cross validation on the dataset and reported the average values of area under curve AUC sensitivity and specificity. They have reported 94% sensitivity and 98% specificity and these results are considered competitive comparing with the state of art algorithm. Also, the model structure is not very deep, which mean it is easy to run such model on normal computer or mobile phone. However, using cross validation method in order to train the model is time consuming and it is not recommended in the context of deep learning since each training cycle requires a lot of time. Also, the algorithm has tended to fail in detecting DR in early stages where the images could be very similar to the normal image.

Furthermore, [23] have developed the model from scratch. They have tested different architecture of the neural network ranges from 9-18 layers, and the convolution kernel size ranges from 1 to 5. The final architecture has contained 8 convolutional layers and 2 fully connected layers and one soft max layer for classification. The authors have used Kaggle dataset for training the neural network. However, they have used only 800 images for training and 200 images for the testing. Also, they have dealt with the problem as a binary classification problem where they have ignored the different levels of disease severity. In order to increase the size of the dataset they have used data augmentation with five types of transformation (rotation, flipping, shearing, rescaling, and translation). They have compared their architecture with the traditional algorithm gradient boosting trees-based (GBM) classification and showed that they have achieved better accuracy comparing with different methods of features extraction with GBM. They have reported 94.5 % of accuracy, but they have not computed sensitivity, specificity, or confusion matrix. Although some important performance metrics have not been mentioned and the architecture has been very traditional, CNN accuracy has outperformed the traditional algorithm where it achieved better accuracy and saved the times and efforts required for designing and extracting new features.

In [24], the problem of DR has been considered as a binary class classification problem where the two classes are(referable/non referable) or simply (healthy/unhealthy). Mainly, two datasets have been used for training and evaluation Kaggle and Messidor-2 datasets. For the model, VGG16 model structure has been used where it has 16 trainable layers including convolution layers and fully connected layers. The model has been trained from scratch. The images have been resized to the resolution 540*540 pixel and pre-processed for lightness and noise problem. Table - I shows for Messidor-2 and Kaggle datasets, Area Under Curve (AUC), sensitivity and specificity metrics values for high sensitivity, and high specificity operating point (OP).

Table - I: Evaluation results of the proposed model in [24]

Data set	AUC	High sensitivity OP		High specificity OP	
		Sensitivity	Specificity	Sensitivity	Specificity
Messidor-2	0.97	99%	71%	87%	92%
Kaggle	0.92	92%	72%	80%	92%

From the results above, it is clear that this model has achieved quite good performance as a binary-class model with AUC 97% for Messidor-2 and 92% for kaggle dataset. However, these high values do not reflect the real performance for different grades of DR (mild- moderate-severe-PDR).

In [25], the authors have compared between two pre-trained models to solve the problem of DR as binary-class classification problem and as multi-class classification problem. These two models are Alexanet and GoogleNet. Models for 2-class, 3-class, and 4-class have been trained and evaluated. Kaggle dataset and Messidor-1 has been used for training and evaluation where 550 images in total have been used for testing. In order to pre-process the images, contrast limited adaptive histogram equalization has been applied. First of all, 2-class (No DR – Severe DR) models have been trained first. The best model has been GoogleNet achieving 95% sensitivity and 96% specificity. After that, a 3-class model using GoogleNet (No DR – Mild – Severe) have been trained. The sensitivity for No DR and severer DR has been in the range 90%. However, the sensitivity of the mild DR is around 7% which mean that the model has been highly confused between mild DR and the other classes. Finally, 4-class models using Alexanet and GoogleNet have been trained. The models performance has degraded a lot since the dataset is highly imbalanced. The best accuracies for 2-class (DR – No DR), 3-class, and 4-class models have been respectively equal to 74.5%, 68.8%, and 57.2%. This research has shown very good results for binary-class models. However, from the reported results, it is clear that DR grading is not an easy task and as long as number of classes to be detected increases the performance starts to degrade.

In [26] the authors have proposed a deep learning-based pipeline for DR grading. Kaggle dataset has been used for training and testing where the dataset has been split into 80% and 20% respectively. This research has supported the concept of fine tuning where different deep pre-trained models have been fine-tuned and evaluated. Mainly VGG19, ResNet101, and Densnet121 have been used. The accuracy has been 35 %, 32%, and 38% and the Kappa score has been 49%, 44%, and 54% respectively. After that Densenet121, which has achieved the best performance, has been used as a feature extractor for the images. Based on these features, lightGBM [27] classifier has been trained and evaluated.

The best accuracy has been 65% and the best kappa has been 82%. Also, the research has reported the classification report which shows the details of recall, precision, and F1 Score for each one of the available classes. This research is considered one of very few researches that supported the concept of transfer learning and using the deep neural network as features extractor which is considered a very promising direction nowadays. The authors mainly have focused on three deep pre-trained models but there are a lot of them need to be tested and evaluated.

In [28], the authors have proposed a system based on combination between image processing techniques and deep learning algorithms to diagnose DR. MESSIDOR-1 dataset of 400 images has been used where 300 images has been used for training and the remaining for testing. The problem has been handled as a binary-class classification problem (healthy/unhealthy). For image pre-processing the research has acknowledged, image resizing to the size (150 * 225) and trying histogram equalization (HE) and contrast limited adaptive histogram equalization (CLAHE). For deep learning model, a new CNN model of eight layers has been implemented and trained from scratch. Thorough comparison between the model trained without HE or CLAHE, model trained with HE, and model trained with CLAHE has been conducted. Reported results have shown that the model with CLAHE has outperformed the other models with 97% of accuracy, 94% for sensitivity, precision and F1 score. This research has highlighted using HE and CLAHE as two techniques for image pre-processing and from the results it is clear that they are promising. However, dataset of 400 images is very small for a problem like DR classification. Therefore, experiments with a bigger dataset should be conducted and evaluated.

Next, research in the field of deep learning and fundus images processing will be discussed in this section. In [29], the authors have worked on retinal vessel segmentation problem. The problem has been formulated as boundaries detection problem and solved by using an architecture called DeepVessel. This architecture mainly has had two layers. The first one has been a CNN to learn a hierarchical image representation where it contains 4 fully convolutional layers. The second layer has been a Conditional Random Field (a recurrent neural network) to model the interaction between different pixels and take into account non-local pixel correlations.

In [30] the authors have worked on Optic Cup (OC) segmentation from colour fundus image. OC segmentation is kind of fundus morphological analysis which is very important to diagnose the glaucoma – one of the leading cause of blindness in the world. In order to do the OC segmentation, the problem has been formulated as a binary-class classification problem where the image has been divided into multiple square patches of size 51*51 and each patch has been labelled as (0 or 1) according to if it belongs to the OC area or not. In order to do the classification, CNN has been used where this CNN architecture has had 3 convolutional layers, 1 max pooling layer, 1 fully connected layer, and finally one sigmoid unit to give the classification result. This architecture has been tested on a public dataset known as Drishti-GS. The authors have reported 93.7% of F-score which is slightly high value and gives a promising direction to use CNN on such a task.

In [31] the researchers have focused on the problem of Image Quality Assessment (IQA) of retinal fundus images. IQA is considered an important task for screening system to detect eye disease like diabetic retinopathy since it helps classify the images into trustable and not trustable images according to the quality. In order to solve the problem of IQA, two types of features have been collected and used for the classification. The first one has been unsupervised type and has been built by computing saliency map where this map represents the correlation between neighbouring areas in the fundus images. The second type has been supervised and has been built by using CNN to learn the necessary features. The CNN architecture contains 5 convolutional layers, 3 fully connected layers, and one softmax layer to classify the image into gradable or upgradable. These CNN has been trained on a dataset of 9653 upgradable retinal images and 11347 gradable images. All of the pictures have been normalized to the range 0 – 1 and resized to the size 512*512. Stochastic gradient descent has been used as an algorithm to train the neural network. After finishing the training process, the features from the last fully connected layer have been combined with the unsupervised features to train a random forest classifier to give the final result. Table – II shows the summary of all of the presented researches about DR detection and grading using deep learning algorithms - its findings and shortcomings.

Table - II: Summary of deep learning researches on diabetic retinopathy grading and detection

Reference	Findings	Shortcomings
[5]	A multi-class problem solved by their own structure. It is one of very few papers handling multi class problem	Have not considered the problem of imbalanced dataset
[18]	A multi class problem solved by transfer learning from Inception-v3. It has achieved very good sensitivity and specificity	Nothing about confusion matrix has been mentioned
[21]	Segmentation and binary class problem solved by their own structure. Excluding unnecessary parts from the image which decreased false positive	Segmentation increases the computation complexity – very small dataset for training and testing
[22]	Segmentation and binary class problem solved by their own structure. New algorithm to speed up the training by selecting the best points for the next epochs which is very good for unbalanced dataset	Selection algorithm has been tested on one dataset only. Therefore, it needs more tests to verify its performance
[32]	A binary class problem solved by fine tuning from GoogleNet. It has proposed a new algorithm to determine number of layers to be retrained	The algorithm is repetitive consuming a lot of time to finish training - Small dataset for testing
[33]	A multi class problem solved by transfer learning from OF model with high accuracy about 90%	Needs to be tested on larger datasets

[34]	A binary class problem solved by their own structure. The model was light and did not need a lot of computational resources	Unable to detect DR in early stages
[23]	A binary class problem solved by their own structure. It gave clear indication that deep learning algorithms outperforms traditional machine learning algorithm	It reported only the accuracy metric
[24]	A binary class problem solved by training VGG16 from scratch with high sensitivity	VGG16 has a lot of trainable parameters
[25]	A binary class problem and multi-class problem solved by fine tanning pre-trained Alexanet and GoogleNet with high accuracy for binary-class model (in the range of 90 %) and degraded performance for multi-class model	The performance of multi-class model is very poor.
[26]	A multi class problem solved by fine-tuning VGG16, ResNet101, and Densenet121. Then Densenet121 has been used as features extractors for lightGBM classifier. The best accuracy is 65%	Only three pre-trained models have been considered. Other pre-trained models should be studied
[28],	A binary class problem solved by using HE and CLAHE for pre-processing and training a model from scratch with high accuracy 97%	The dataset used for training and evaluation is quite small

C. Research gaps

From all the reviewed research works, it has been concluded that several research gaps in the field exist. These research gaps are divided into: (i) problems related to the dataset like imbalanced dataset and how to protect the patient privacy (ii) and problems related to the AI model itself like difficulties in model predictions interpretation, lack of computational resources to train big models, and lack of people trustiness in AI models results in the field of medical image analysis. The following subsections will discuss these gaps in details.

- **Imbalanced data set**

For satisfying results, deep learning projects usually requires huge amount of data to be trained on. Getting such amount of data is not an easy task, and the task becomes more complicated when the problem is related to the medical fields. Also, another problem appears in the medical fields is that the number of healthy cases is bigger usually than non-healthy cases which leads eventually to unbalance in the dataset. This unbalance could harm the performance of the training process if not processed. Therefore, it is very important to process this problem to protect the model from bias toward healthy cases.

- **Privacy**

Privacy is a very important concern in medical applications which require data exchange among people. A retinal fundus image contains the human retina which considered one of the bio-metrics which can be used to identify the human personality. Therefore, even though the meta-data of images have been removed, it is still possible to extract the human identity from the image itself.

- **Model predictions interpretation**

Even with an accurate model, people need to understand why this result occurs. Understanding the result becomes more important when the problem is related to the human health. Unfortunately, up to this moment models provide the final result without providing any other answers .Therefore, depending on such models still requires more time to find the best way to explain the final result.

- **Trustiness**

This problem related to all deep learning applications that could harm the human or cause human life losing like autonomous car systems and computer diagnosis systems. In the medical fields, there are a lot of doubts about the accuracy of the trained models where even if the model gives very good results, there is nothing guarantees that the model will not make mistakes on other datasets gathered from different people from different countries and geographical locations. Therefore, we still need advanced ways to evaluate the performance of the model and trustiness degree in the result.

- **Lack of computational resources**

The complexity of medical images requires building complex models for detection and recognition. However, building such models requires huge computational resources which could not be available always. Therefore, most of the researchers have implemented relatively small models for small datasets. Also, most of the researches have simplified the problem and handled it as a binary classification problem whereas DR has different stages and different level of severity. Very few researchers have advocated the concept of transfer learning and fine tuning of pre-trained models to overcome this issue but without any trial to harnessing the abilities of multiple pre trained models together.

It is clear that there are a lot of challenges facing the development of fully automated and reliable system for DR detection and grading. Although there are a lot of research works trying to solve the problem of DR detection as binary classification problem, there are very few research works trying to solve the problem of DR grading as multi-class classification problem which makes the doors open for new research ideas in this direction. For imbalanced dataset problem, it is necessary to develop new loss functions and models architectures which consider this problem since it is considered one of the main performance-degrading reasons in the current models. Also, extensive efforts should be made from the researchers to improve the models predictions interpretation which shades the light inside the neural network black box. In addition, the neural networks are hungry for computational resources and data samples. To overcome these problems, researchers should consider transfer learning and fine-tuning of pre-trained models which transfer the knowledge from other models to DR detection and grading problem. Transfer learning decreases the model training time and the required amount of the computational resources since training will not start from totally random model. Also, it requires fewer amounts of data to achieve satisfying results comparing with starting from scratch.

Finally, cooperation between ophthalmologists community and computer scientists is highly demanded to provide the required data for building such automated systems and to evaluate the performance of the systems after implementation which increases the quality of these systems which in turn increases people trust in such systems.

III. CONCLUSION

Diabetic retinopathy is a disease infects the retina of people who have diabetes. It is a chronic disease which could harm the human vision or cause total blindness. AI-based automated systems are a promising solution to help huge number of people with diabetes to get the required screening easily and quickly. DL, a subfield of artificial intelligence, achieved the most promising results in the field up to this moment. In this article, the different concepts related to DL have been explained. The main algorithm under DL, ANN, its shortcomings in computer vision applications, and its new type, CNN, to cope ANN drawbacks have been explained in details. To determine the new research directions in the field, extensive presentation of some research works have been done where model structure, dataset, training process, and results have been summarized and discussed. At the end of this article, the most important research gaps have been highlighted to show where researcher's efforts could be made in future in this field.

Finally, with this exponentially increasing in the number of people with diabetes, integration of AI with eye healthcare systems becomes very important to implement effective and low-cost solutions to help ophthalmologists in diagnosis process, and to increase levels of patients' satisfaction.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Higher Education Malaysia (Kementerian Pendidikan Tinggi) under KNOWLEDGE TRANSFER PROGRAMME - RESEARCH INITIATIVE GRANTSCHEME (KTP-RIGS SDG) 2019 number IRG19-005- 0005. Also, we would like to thank Dr. Muhanad Hreh for providing us with information regarding diabetic retinopathy and answering any medically related questions. (Muhanad Hreh, M.D. Pikeville Medical Center, KY, USA).

REFERENCES

1. "IDF Website." [Online]. Available: https://idf.org/our-network/regions-members/western-pacific/members/108-malaysia.html. [Accessed: 14-Aug-2019].
2. "Diabetic Retinopathy Detection | Kaggle." [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection. [Accessed: 17-Aug-2019].
3. O. Faust, R. Acharya U., E. Y. K. Ng, K. H. Ng, and J. S. Suri, "Algorithms for the automated detection of diabetic retinopathy using digital fundus images: A review," J. Med. Syst., 2012.
4. G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis. 2017.
5. H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," in Procedia Computer Science, 2016.
6. F. Chollet, Deep Learning with Python & Keras. 2018.
7. Q. J. Zhang, K. C. Gupta, and V. K. Devabhaktuni, "Artificial neural networks for RF and microwave design - From theory to practice," IEEE Trans. Microw. Theory Tech., 2003.
8. N. Q. K. Le and V.-N. Nguyen, "SNARE-CNN: a 2D convolutional neural network architecture to identify SNARE proteins from high-throughput sequencing data," PeerJ Comput. Sci., vol. 5, p. e177, Feb. 2019.
9. M. Williams and J. Haugeland, "Artificial Intelligence: The Very Idea," Technol. Cult., 1987.
10. S.-C. B. Lo, S.-L. a Lou, J.-S. Lin, M. T. Freedman, M. V Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection," IEEE Trans. Med. Imaging, 1995.
11. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, 1998.
12. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst., 2012.
13. K. Bhatia, S. Arora, and R. Tomar, "Diagnosis of diabetic retinopathy using machine learning classification algorithm," in Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies, NGCT 2016, 2017.
14. P. R. Asha and S. Karpagavalli, "Diabetic retinal exudates detection using Extreme Learning Machine," in Advances in Intelligent Systems and Computing, 2015.
15. P. R. and A. P., "DIAGNOSIS OF DIABETIC RETINOPATHY USING MACHINE LEARNING TECHNIQUESR., P., & P., A. (2013). DIAGNOSIS OF DIABETIC RETINOPATHY USING MACHINE LEARNING TECHNIQUES. ICTACT Journal on Soft Computing. https://doi.org/10.21917/ijsc.2013.0083," ICTACT J. Soft Comput., 2013.
16. A. Sopharak and B. Uyyanonvara, "Automatic exudates detection from diabetic retinopathy retinal image using fuzzy C-means and morphological methods," in Proceedings of the 3rd IASTED International Conference on Advances in Computer Science and Technology, ACST 2007, 2007.
17. A. Osareh, M. Mirmehdi, B. Thomas, and R. Markham, "Automatic Recognition of Exudative Maculopathy using Fuzzy {C}-Means Clustering and Neural Networks," in Medical Image Understanding and Analysis, 2001.
18. V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA - J. Am. Med. Assoc., 2016.
19. S. Ioffe and C. Szegedy, "Batch {Normalization}: {Accelerating} {Deep} {Network} {Training} by {Reducing} {Internal} {Covariate} {Shift}," arXiv1502.03167 [cs], 2015.
20. O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," Int. J. Comput. Vis., 2015.
21. P. Prentašić and S. Lončarić, "Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion," Comput. Methods Programs Biomed., 2016.
22. M. J. J. P. Van Grinsven, B. Van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images," IEEE Trans. Med. Imaging, 2016.
23. K. Xu, D. Feng, and H. Mi, "Deep Convolutional Neural Network-Based Early Automated Detection of Diabetic Retinopathy Using Fundus Image.," Molecules, 2017.
24. A. Rakhlin, "Diabetic Retinopathy detection through integration of Deep Learning classification framework," pp. 1–11, 2017.
25. C. Lam, D. Yi, M. Guo, and T. Lindsey, "Automated Detection of Diabetic Retinopathy using Deep Learning.," AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci., vol. 2017, pp. 147–155, 2018.
26. Y. Wang, W. Fan, C. K. Reddy, and Y. Wang, "A Deep Learning Based Pipeline for Image Grading of Diabetic Retinopathy A Deep Learning Based Pipeline for Image Grading of Diabetic Retinopathy," 2018.
27. G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," no. Nips, pp. 1–9, 2017.
28. D. J. Hemanth, O. Deperlioglu, and U. Kose, "An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network," Neural Comput. Appl., vol. 0, 2019.
29. H. Fu, Y. Xu, S. Lin, D. W. K. Wong, and J. Liu, "Deepvessel: Retinal vessel segmentation via deep learning and conditional random field," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.
30. Y. Guo, B. Zou, Z. Chen, Q. He, and Q. Liu, "Optic Cup Segmentation Using Large Pixel Patch Based CNNs," pp. 129–136, 2016.

31. D. Mahapatra, “Retinal Image Quality Classification Using Neurobiological Models of the Human Visual System,” 2016.
32. D. Worrall, C. Wilson, and G. Brostow, “Automated Retinopathy of Prematurity Case Detection with Convolutional Neural Networks,” Proc. Deep Learn. Data Labeling Med. Appl., 2016.
33. P. Burlina, D. E. F. N. Joshi, and Y. W. N. M. Bressler, “DETECTION OF AGE-RELATED MACULAR DEGENERATION VIA DEEP LEARNING,” in 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016.
34. R. Gargeya and T. Leng, “Automated Identification of Diabetic Retinopathy Using Deep Learning,” Ophthalmology, 2017.

AUTHORS PROFILE

Muhammad Samer Sallam, is currently a master student and researcher in Electrical and Computer Engineering Department, Faculty of Engineering, International Islamic University Malaysia (IIUM). He received his Bachelor degree from Computer and Automation Engineering department, Faculty of Mechanical and Electrical Engineering, Damascus University in Syria. His current research interests include optimization algorithms, machine learning and deep learning for medical images applications, computer vision, image processing, big data analytics, data science, geo-spatial data analysis, and social media data analysis. He started his career as an algorithm engineer in Rachis Systems, Kuala Lumpur, Malaysia. Then, he has been promoted to be the data science team lead in the company. After two years, he has been hired by Quaking Aspen in Dublin, Ireland as Senior Data Scientist. Also, he conducts training courses related to data science and artificial intelligence for corporates and universities in Kuala Lumpur, Malaysia.

Rashidah Funke Olanrewaju, a Nigerian citizen born in Kaduna, Nigeria. She received the BSc. Hons degree in Software Engineering from the University of Putra Malaysia, in 2002, and the MSc and Ph.D. degrees in Computer & Information Engineering from the International Islamic University Malaysia (IIUM) Kuala Lumpur, in 2007 and 2011, respectively. She is currently an Associate Professor at Department of Electrical and Computer Engineering, International Islamic University Malaysia where she is leading the Software Engineering Research Group (SERG). She is an executive committee member of technical associations like: IET UK, IEEE Women in Engineering, Arab Research Institute of Science and Engineers, Nigeria Computer Society, Malaysia Society of Cryptology Research, editorial board member in SCIREA journal of Computer etc. She represents her university, IIUM, at Malaysian Society for Cryptology Research. Her current in hand projects revolve around: Fintech; financial Institution security measures, creation and deployment of solutions protecting networks, systems and information assets for diverse companies & organizations, MapReduce Optimization Techniques, Compromising Secure Authentication and Authorization Mechanisms, Secure Routing for adhoc networks, Formulating Bio-Inspired Optimization Techniques and Artificial Intelligent systems.

A.L. Asnawi is currently an Assistant Professor in Electrical and Computer Dept, Faculty of Engineering, International Islamic University Malaysia (IIUM). She received her Phd from School of Electronics and Computer Science, University of Southampton, United Kingdom, in 2012. She obtained her M.Eng in Communication and Computer Engineering from University Kebangsaan Malaysia (UKM), and her Bachelor degree from International Islamic University Malaysia (IIUM). Her current research interests include wireless communication, software defined radio, software engineering, empirical software engineering, Agile methods software processes, machine learning, big data analytics and smart farming. She is a senior member for IEEE, Executive Committee for IEEE Computer Society Malaysia, a registered member for Board of Engineers Malaysia (BEM) and The Institution of Engineers Malaysia (IEM).