International Journal of Medical Informatics 132 (2019) 103926

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Bimodal learning via trilogy of skip-connection deep networks for diabetic retinopathy risk progression identification

Cam-Hao Hua[a], Thien Huynh-The[b], Kiyoung Kim[c], Seung-Young Yu[c], Thuong Le-Tien[d], Gwang Hoon Park[a], Jaehun Bang[a], Wajahat Ali Khan[a], Sung-Ho Bae[a,*], Sungyoung Lee[a,*]

[a] Department of Computer Science and Engineering, Kyung Hee University, Gyeonggi-do 17104, South Korea
[b] ICT Convergence Research Center, Kumoh National Institute of Technology, Gumi, South Korea
[c] Department of Ophthalmology, Kyung Hee University Medical Center, Kyung Hee University, Seoul 02447, South Korea
[d] Department of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology, Ho Chi Minh City, 700000, Vietnam

ARTICLE INFO

Keywords:
Bimodal learning
Diabetic Retinopathy risk progression
EMR-based attributes
Fundus photography
Retinal fundus image
Trilogy of skip-connection deep networks

ABSTRACT

Background: Diabetic Retinopathy (DR) is considered a pathology of retinal vascular complications, which stays in the top causes of vision impairment and blindness. Therefore, precisely inspecting its progression enables the ophthalmologists to set up appropriate next-visit schedule and cost-effective treatment plans. In the literature, existing work only makes use of numerical attributes in Electronic Medical Records (EMR) for acquiring such kind of DR-oriented knowledge through conventional machine learning techniques, which require an exhaustive job of engineering most impactful risk factors.

Objective: In this paper, an approach of deep bimodal learning is introduced to leverage the performance of DR risk progression identification.

Methods: In particular, we further involve valuable clinical information of fundus photography in addition to the aforementioned systemic attributes. Accordingly, a Trilogy of Skip-connection Deep Networks, namely Tri-SDN, is proposed to exhaustively exploit underlying relationships between the baseline and follow-up information of the fundus images and EMR-based attributes. Besides that, we adopt Skip-Connection Blocks as basis components of the Tri-SDN for making the end-to-end flow of signals more efficient during feedforward and backpropagation processes.

Results: Through a 10-fold cross validation strategy on a private dataset of 96 diabetic mellitus patients, the proposed method attains superior performance over the conventional EMR-modality learning approach in terms of Accuracy (90.6%), Sensitivity (96.5%), Precision (88.7%), Specificity (82.1%), and Area Under Receiver Operating Characteristics (88.8%).

Conclusions: The experimental results show that the proposed Tri-SDN can combine features of different modalities (i.e., fundus images and EMR-based numerical risk factors) smoothly and effectively during training and testing processes, respectively. As a consequence, with impressive performance of DR risk progression recognition, the proposed approach is able to help the ophthalmologists properly decide follow-up schedule and subsequent treatment plans.

1. Introduction

According to World Health Organization (WHO), Diabetic Retinopathy (DR) is the top-five and -four causes of vision impairment and blindness on earth, respectively. It is originated by the long-term impact of diabetes mellitus (DM) which results in adverse changes of nerves and blood vessels of the patients' retina. In other words, DR is considered a pathology of retinal vascular complications. Therefore, early and accurate detection of the DR symptoms can help the ophthalmologists easily form optimal treatment plans for DR progression prevention and management.

Historically, the severity of DR has been graded by structural changes within color fundus photography such as retinal hemorrhages, microaneurysms, intraretinal microvascular abnormalities [1].

* Corresponding authors.
E-mail addresses: hao.hua@oslab.khu.ac.kr (C.-H. Hua), thienht@kumoh.ac.kr (T. Huynh-The), pourma@naver.com (K. Kim), syyu@khu.ac.kr (S.-Y. Yu), thuongle@hcmut.edu.vn (T. Le-Tien), ghpark@khu.ac.kr (G.H. Park), jhb@oslab.khu.ac.kr (J. Bang), wajahat.alikhan@oslab.khu.ac.kr (W.A. Khan), shbae@khu.ac.kr (S.-H. Bae), sylee@oslab.khu.ac.kr (S. Lee).

https://doi.org/10.1016/j.ijmedinf.2019.07.005
Received 4 April 2019; Received in revised form 4 June 2019; Accepted 6 July 2019
1386-5056/

Nowadays, recent advancements in imaging modalities have enabled us to visualize retinal neurodegeneration and microvascular changes as early events of clinically visible DR-related signs [2–4]. Those neuronal and microvascular parameters extracted from retinal images have been identified as indicators of DR risk progression. Additionally, count of microaneurysms, caliber of retinal vessels, and inner retinal layer thickness [5–7] were reported to be associated with early stage of DR progression as well.

Moreover, a number of biochemical pathways has been proposed that the pathogenesis of DR includes glucose-mediated microvascular damage [8]. There are also emerging evidences suggesting that DR may share common linkages with various systemic vascular complications [9]. Thus, summarized from previous clinical studies [10–13], diabetes duration, hemoglobin A1c (HbA1c), hypertension, age, cholesterol, cardiovascular symptoms, and kidney functions are considered as independent risk factors for the progression of DR.

In short, DR-based complications and its progression can be detected by different modalities, comprising (i) Fundus photography taken by nonmydriatic fundus camera, where the representation of retinal blood vessel, appearances of microaneurysms, exudates, hemorrhages, etc. are concerned clinical signs, and (ii) Electronic Medical Records (EMR) based attributes, where selective numerical measures like diabetes duration, hypertension index, Body Mass Index (BMI), HbA1c, Ganglion Cell-Inner Plexiform Layer (GC-IPL) thickness, to name a few, are utilized for the pathogenesis identification or follow-up planning. Subsequently, inspecting the changes of these multi-modal factors by time enables the ophthalmologists to identify whether there is any risk progression of DR for setting up proper next-visit schedule and cost-effective treatment plans.

Lately, with the powerful advancement of GPUs and fast-paced growth of numerical & visual data, deep learning has been empowered in diverse fields, ranging from the image classification [14–17] and segmentation [18–20] problems in computer vision area to systemic attributes processing in medical data analysis domain [21–24]. In terms of DR-oriented studies, two major machine learning based research directions are currently of interest, i.e., identifying DR progression from risk factors [25–29] and detecting severity scales of DR based on retinal fundus images [30–36]. As for the former, a great number of risk factors leads to difficulties in selecting properly impactful inputs as well as making training progress of the built model converge. Regarding the latter, for high accuracy of ranking the DR severity level on fundus images of different screening time, large-scale dataset requirement and expensive resource consumption are significant concerns.

Therefore, to our best knowledge, utilizing deep learning to effectively incorporate meaningful information from various modalities (e.g. EMR-based attributes and fundus photography) for certain objectives of diagnosis, treatment or follow-up (e.g. DR risk progression identification) remains an open research area. It is worth noting that such kind of multimodal learning approach has been exploited in other fields such as image-text matching [37], audio and visual signals joint representation [38], multi-type medical image processing [39], to name a few, but not thoroughly investigated on both fundus photography and EMR-based attributes in studies about DR risk progression.

Remarkably, there are three considerable advantages of such approach as follows. Firstly, it is hypothesized that retinal fundus images possess clinical features in terms of both explicit and implicit representation of blood vessel, microaneurysms, exudates, hemorrhages, etc; which bring in extra valuable information of DR progression that may not be explicitly reported in the EMR. Secondly, such multimodal strategy can reduce the dependence of exhaustively determining which EMR-based risk factors having strong effects on the progression of DR. Thirdly, the proposed method lowers the necessity of expensively constructing and training the designated deep learning model with large-scale dataset of fundus photography.

Motivated by those observations, this paper proposes a Trilogy of Skip-connection Deep Networks (Tri-SDN) which is capable of comprehensively extracting useful knowledge from bimodalities, i.e., retinal fundus images and EMR-based attributes, for identifying DR risk progression effectively. In concrete, the proposed architecture consists of: (i) a Convolutional Neural Network (CNN) followed by a global average pooling (GAP) layer and a subsequent Deep Neural Network (DNN) having Skip Connection Blocks (SCB), called SDN, for encoding salient features of lesions existing in both baseline and follow-up fundus images; (ii) a SDN extracts latent features which represent inter-relationships between systemic attributes of interest (e.g. Hb1Ac, GC-IPL thickness, BMI, HTN,...) as well as intra-relationships between baseline and follow-up values of each attribute in the EMR; and (iii) another SDN classifies DR risk progression by processing concatenated version of fundus-based and EMR-based features.

Then, the proposed architecture is evaluated on an internal dataset of 96 DM patients collected at Kyung Hee University Medical Center, Seoul, South Korea. Through impressive performance in terms of accuracy, sensitivity/recall, precision, specificity, and area under the receiver operating characteristics, the proposed Tri-SDN shows effectiveness in assisting ophthalmologists the task of identifying DR progression.

In brief, the main contributions delivered by this work are summarized as follows

- We propose an approach of bimodal learning that exhaustively utilizes both baseline and follow-up information of retinal fundus images and EMR-based attributes for identifying the DR risk progression, from which the ophthalmologists can efficiently deploy appropriate next-visit schedule and cost-effective treatment plan.
- We introduce the Trilogy of Skip-connection Deep Networks (Tri-SDN) that combines latent features of different modalities (i.e., color images and numerical risk factors) smoothly and effectively during training and testing processes, respectively.
- The proposed bimodal learning method is able to suppress the necessity of large-scale fundus image dataset involvement and the costly task of engineering most impactful EMR-based risk factors.
- We execute a 10-fold cross validation strategy to evaluate the proposed model on an internal dataset of 96 DM patients, who may suffer from DR development or vice versa, collected at Kyung Hee University Medical Center, Seoul, South Korea. Accordingly, the experimental results and corresponding ablation studies show the effectiveness of the proposed Tri-SDN.

The remainder of this paper is distributed as follows. Related work on EMR-based attributes and retinal fundus images utilization in the domain of DR are given in Section 2. Then, Section 3 provides comprehensive description of the proposed Tri-SDN. Subsequently, a benchmark dataset with corresponding implementation details and experimental results are presented in Section 4. Finally, Section 5 summarizes the materials delivered by this work.

2. Related work

Recently, in order to deal with DR-related issues such as risk progression detection, follow-up scheduling, severity classification, most existing work takes into account either EMR-based attributes or retinal fundus photography as follows.

2.1. EMR-based attributes exploitation

Several works already reported typical predictive models for the risk assessment of DR development by processing various systemic attributes. Since there is a large amount of risk factors impacting on the progression of DR for subsequently determining suitable treatment plan and follow-up schedule, constructing an efficient clinical decision support system (CDSS) with selective EMR-based attributes attracts numerous researches [25–29].

Table 1
Highlighted differences between the proposed approach and several related work. The abbreviations are interpreted as follows: 'cML' means conventional machine learning techniques; 'DL' means deep learning technique; 'Fundus Img.' means fundus images; 'EMR Att.' means EMR-based attributes; (.) in 'Fundus Img.' and 'EMR Att.' columns indicates the number of fundus images and utilized risk factors, respectively.

Publication	Approach		Dataset size	Input		Output
	cML	DL		Fundus Img.	EMR Att.	
Skevofilakas et al. [25]	✓		55		✓ (7)	DR progression
Bajestani et al. [26]	✓		200		✓ (4)	Time-span DR
Piri et al. [27]	✓		1.4M		✓ (5)	DR detection
Eleuteri et al. [28]	✓		11,806		✓ (5)	DR progression
Romero-Aroca et al. [29]	✓		2,323		✓ (8)	DR screening interval
Choi et al. [30]	✓	✓	400	✓		Retinal disease
Yang et al. [31]	✓	✓	320	✓		Retinal disease
Gulshan et al. [32]		✓	128,175	✓		DR grade
Rahim et al. [33]	✓		600	✓		DR grade
Pratt et al. [34]		✓	85,000	✓		DR grade
Trivino et al. [35]		✓	88,000	✓		DR grade
Zhou et al. [36]		✓	35,126	✓		DR grade
Ours		✓	96	✓ (192)	✓ (22)	DR progression

For instance, in [25], a CDSS was proposed in which wavelet neural network, decision trees, and Bayesian neural network are integrated to predict the progression of DR on type 1 DM patients. The considered EMR-based attributes include age, DM duration, HbA1c, total cholesterol, triglycerides, hypertension, and treatment duration.

Moreover, the study in [26] aimed at assisting the ophthalmologists to judiciously determine the time span between DM diagnosis and the development of DR for further healthcare actions. In addition, the authors showed the effectiveness of utilizing fuzzy regression model on a small-scale dataset with the involvement of four attributes, i.e., blood pressure, fasting blood sugar, HbA1c, and diagnosis age for their prespecified purpose.

Meanwhile, with a large amount of dataset, a hybrid strategy [27] of logistic regression, decision tree, random forest, and neural network was proposed to diagnose early DR signs from critical risk factors (diabetic neuropathy, hematocrit, blood urea nitrogen, creatinine serum, glucose serum plasma) and subsequently aid appropriate screening intervals.

Besides that, the authors in [28] introduced a risk calculation engine (RCE) for estimation of risk progression, which is then used to individualize proper screening intervals for each particular patient. The RCE took into consideration of five risk factors (i.e., DM duration, HbA1c, age, systolic BP, total cholesterol) from 11,806 diabetic patients for the predefined continuous-time Markov process.

In recent times, another CDSS [29] utilized fuzzy random forest consisting of 200 trees for personalizing follow-up schedule based on such input risk factors as age, gender, DM duration, hypertension, BMI, HbA1c, estimated glomerular filtration, and microalbuminuria. The constructed system was trained by dataset of 2,323 type 2 DM patients with expected outputs among four predefined classes, i.e., next visit of six months, one year, two years, and three years.

2.2. Retinal fundus photography exploitation

Obviously, features extracted from fundus photography contribute valuably clinical information for detecting various types of retinal diseases. Since these can be grouped as a traditional image classification problem, deep learning based models [30–36] can handle effectively given a reasonable amount of training dataset.

Concretely, in the case of common retinal diseases recognition given large quantity of supervised classes, the CNNs take charge of feature extractor followed by a conventional classifier like random forest [30] or principal component analysis plus support vector machine [31].

On the other hand, with respect to the DR domain, retinal fundus images are mainly utilized for grading the severity of DR such as non-DR, mild DR, moderate DR, severe DR, and proliferative DR according to the International Clinical Diabetic Retinopathy Disease Severity Scale. Regarding the utilization of conventional machine learning algorithms, fuzzy-based techniques are of great interest in the domain. For instance, Rahim et al. [33] introduced an ensemble of fuzzy filtering, histogram equalization, and edge detection for effectively extracting useful features and classifying the DR grades in 600 fundus images accordingly. In terms of more advanced deep learning technique, [34] and [35] employed 13-layer and 11-layer CNNs (which are constructed from common stacks of convolution, rectified linear unit (ReLU) activation, max pooling, fully connected (FC) layers), respectively. Meanwhile, the authors in [32] fine-tuned the light-weight Inception-v3 architecture [16] for such kind of DR detection issue.

Different from the aforementioned approaches wherein full-sized retinal images are taken into account, the multiple instance learning method [36] executed a two-stage patch-based training process. At the first stage, numerous patches were initially generated from the raw fundus images and then fed into the pretrained AlexNet [14] for predicting patch-wise DR grades, of which the results were aggregated to form the final probability map. As for the second stage, the authors utilized multi-scale versions of the original input, each of which passes through the procedure of the first stage, for forming the globally averaged probability map of final DR grade.

It should be noted that training with large-scale dataset ($>$ 80,000 images) plays a critical role for the impressive performance in the existing fundus image-based work. To summarize, we highlight major differences between the proposed approach and the aforementioned related work in Table 1. Additionally, readers may refer to [40] for a comprehensive survey of the DR related literature.

3. Methodology

In this section, we firstly deliver overall description of the proposed architecture shown in Fig. 1, wherein the basis component called Skip-Connection Block (SCB) is introduced. Then, the following sub-sections sequentially provide details of deep neural networks applied for handling retinal fundus images, EMR-based attributes, and the cross-modal features concatenated from previously extracted ones, respectively.

3.1. Basis component of Tri-SDN: skip-connection block

As illustrated in Fig. 1, the proposed Tri-SDN firstly exploits vital information from retinal fundus images and EMR-based attributes by using a pretrained CNN followed by a SDN and another SDN, respectively, in parallel. Then, the extracted latent features across different modalities are aggregated so that a third SDN effectively acquires high-

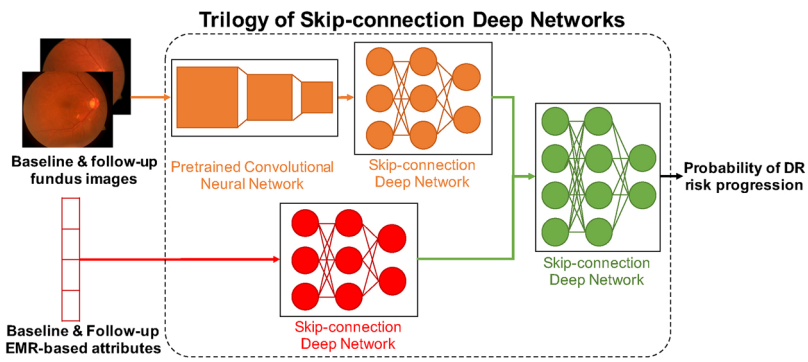


Fig. 1. Overall architecture of the proposed method - Bimodal learning via Trilogy of Skip-connection Deep Networks (Tri-SDN) for DR risk progression identification. In the Tri-SDN, one pre-trained convolutional neural network followed by a skip-connection deep network is used for encoding meaningful features of baseline and follow-up fundus images; in parallel, another skip-connection deep network is utilized for exploring interactions between baseline and follow-up EMR-based attributes; finally, one deep neural network also with skip-connection mechanism is employed to learn the hybrid features concatenated from outputs of the two previous networks for identifying the progression of DR.

level representation of the cross-modal features for DR risk progression identification.

It is obvious that in order to recognize the progression of DR, both baseline and follow-up measures of the concerned risk factors need to be taken into account, which makes the number of considered inputs increase significantly. Hence, the learning model should be designed to be deep enough for efficiently capturing essential knowledge from the given inputs. However, as remarked in [17,41], training DNNs using large number of layers along with Sigmoid activation leads to the issue of vanishing or exploding gradients, which makes the learning progress not converge properly and lowers final performance accordingly. Therefore, we adopt the mechanism of residual connections [17] along with ReLU activation function in the proposed SDNs for overcoming the above-mentioned problems.

Concretely, every SDN involves a predefined number of Skip-Connection Blocks (SCB), each of which comprises two FC layers (having trainable weights and biases of $\{W_1, W_2 \in R^{d\times d}; b_1, b_2 \in R^{d\times 1}\}$, respectively) with ReLU activation in the middle, and a subsequent element-wise summation operator followed by another ReLU activation layer. As illustrated in Fig. 2, let a d-length feature vector $X = [x_1, x_2, \ldots, x_d]^T$, i.e., $X \in R^{d\times 1}$, be input of a SCB, the final output Y is achieved through following feedforward operations

$$H_1 = W_1 X + b_1 \tag{1}$$

where $H_1 = [h_{11}, h_{12}, \ldots, h_{1d}]^T$ denotes the output of the first FC layer, then the feature vector of ReLU-style activated neurons $H_1' = [h_{11}', h_{12}', \ldots, h_{1d}']^T$ is obtained by

$$h_{1i}' = \max(h_{1i}, 0), \quad \forall\, i = 1, \ldots, d \tag{2}$$

Afterwards, the feature extraction process continues with

$$H_2 = W_2 H_1' + b_2 \tag{3}$$

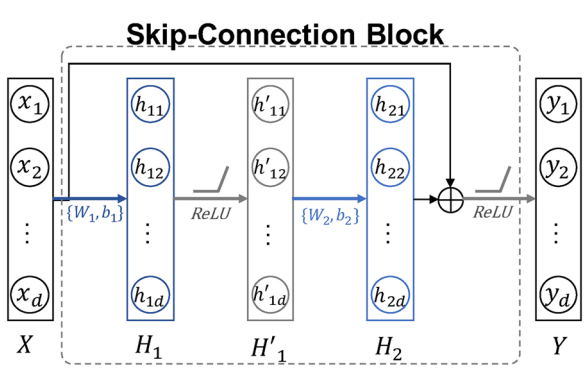


Fig. 2. Structure of a Skip-Connection Block (SCB), wherein an arbitrary feature vector $X \in R^{d\times 1}$ is fed into two FC layers (having trainable weights and biases $\{W_1, W_2 \in R^{d\times d}; b_1, b_2 \in R^{d\times 1}\}$, respectively) with ReLU activation in the middle. Then, the final output $Y \in R^{d\times 1}$ of such SCB is inferred by the element-wise summation between input feature vector X and output H_2 of the second hidden layer followed by ReLU activation.

where $H_2 = [h_{21}, h_{22}, \ldots, h_{2d}]^T$ indicates the output feature vector of the second FC layer. Notably, in conventional DNN, H_2 is continuously fed into the next layer. Meanwhile, in the proposed network, the desired output Y is acquired from the element-wise summation between this vector and the original input X followed by a ReLU activation function as follows

$$\begin{aligned} Y &= \mathrm{ReLU}(X + H_2) \\ &= [\max(x_1 + h_{21}, 0), \ldots, \max(x_d + h_{2d}, 0)] \\ &= [y_1, \ldots, y_d]^T \end{aligned} \tag{4}$$

Clearly, such scheme of skip-connection as formulated in (4) offers two advantages over the conventional structure in terms of both feedforward and backpropagation processes. With respect to the former, this mechanism drives the output Y to acquire newly essential features different from those existed in the input X. As for the latter, it avoids the corresponding gradient signals flown through Y to X from being vanished thanks to the existence of direct stream to X.

In brief, due to taking into account the bimodal inputs, at least three deep networks are needed for effectively acquiring valuable features from each single modality and the aggregated ones. Thus, it is necessary to define a strategy be able to ensure the smooth flow of signals between the three networks in both feedforward and backpropagation stages for ease of feature encoding and optimization of the whole model's parameters, respectively. As a consequence, we opt for taking the described SCBs as the basis components of the proposed Tri-SDN in this work for efficiently learning essential feature representation from a large amount of risk factors existing in both fundus photography and EMR.

3.2. Fundus image modality learning

In order to extract DR-based features from the baseline and follow-up retinal fundus images, the detailed architecture demonstrated in Fig. 3 is executed.

At first, we utilize the 50-layer ResNet [17] pretrained with lmageNet dataset [42] as a fixed feature extractor thanks to its powerful representation. Generally, the ResNet architecture consists of one initial convolutional block (named Conv-1 in Fig. 3) followed by four residual blocks (named Res-block 1, Res-block 2, Res-block 3, Res-block 4, respectively), each of which includes multiple convolution layers organized in terms of residual/skip connections manner. Note that Res-block 3 and Res-block 1 hold the greatest and least number of layers, respectively. Along the feedforward process of such pretrained CNN as shown in Fig. 3, there are two observations, i.e., (i) spatial size of final outputs of the residual blocks constantly decreases by two times while the corresponding channel dimension numerously rises, and (ii) more distinctive and abstract features representing key information in the original images are encoded in depth-wise manner. However, it is clear that such kind of retinal images is beyond the scope of pretrained categories in ImageNet dataset, thus, taking into account feature maps learned from high-level layers (such as those in Res-block 4) certainly

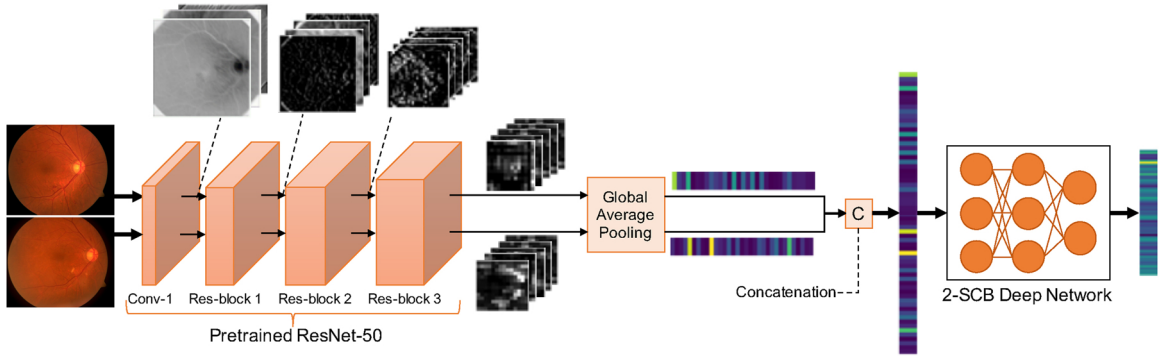


Fig. 3. Detailed workflow of the deep network learning baseline & follow-up retinal fundus images. Given these two images, a pretrained ResNet-50 [17] plays as a fixed feature extractor, in which the corresponding final outputs of the third residual block (Res-block 3) are considered for DR-oriented characteristics exploitation. Concretely, those feature maps are converted into vector formats in depth-wise manner using Global Average Pooling. Then, the concatenated version of the resulting feature vectors passes through a deep network having two Skip-Connection Blocks to infer the DR-oriented fundus based feature vector for further processes.

leads to unexpected biases and/or losses in terms of encoded informative responses. Consequently, we opt for picking up feature maps containing reasonable contextual information, of which spatial resolution has stride of 16 compared to that of the input images, finalized from the Res-block 3 (as illustrated in Fig. 3) for next stage.

Afterwards, the extracted baseline and follow-up feature maps are converted into corresponding vectors so that the subsequent SDN can selectively learn latent fundus-based risk factors that exhibit the development of DR. It should be noted that the reason for this conversion is to ensure the smooth aggregation with feature vectors learned from EMR-based attributes for the third SDN. As aforementioned, since every channel of the feature maps of interest performs specific depiction of essential characteristics with respect to the original inputs, GAP is applied to convert the two considered feature maps (i.e., the baseline f_{ib} and the follow-up f_{if} feature maps) into vector formats (i.e., g_{ib} and g_{if}, respectively) by spatially averaging in depth-wise manner as following operation

$$\mathcal{G}^d = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} f^d \tag{5}$$

where f is either f_{ib} or f_{if}; f^d specifies the dth channel of the processed feature maps f having spatial resolution of $H \times W$; $\mathcal{G}^d$ denotes the representative response for corresponding channel of the feature map f. Then, the resulting vector is formed as follows

$$g = [\mathcal{G}^1, ..., \mathcal{G}^d, ..., \mathcal{G}^D]^T \tag{6}$$

where $g \in R^{D \times 1}$ is either g_{ib} or g_{if}, D stands for the depth size of the processed feature map. Notably, $D = 1024$ in this work since the chosen features are outputs of the Res-block 3 of the ResNet-50. Next, the newly decoded vectors g_{ib} and g_{if} are concatenated in order to exploit the interactions between baseline and follow-up input images as below

$$G = C[g_{\mathrm{ib}}, g_{\mathrm{if}}] = [\mathcal{G}^1_{\mathrm{ib}}, ..., \mathcal{G}^D_{\mathrm{ib}}, \mathcal{G}^1_{\mathrm{if}}, ..., \mathcal{G}^D_{\mathrm{if}}]^T \tag{7}$$

where $C[.]$ signifies the concatenation operator.

Finally, a deep network having two SCBs is employed to thoroughly learn fundus-based essential characteristics in terms of DR risk progression. In details, input of this network is the fundus-based feature vector $G \in R^{2D \times 1}$ firstly fed into a single FC layer with trainable parameters of $\{W_{is} \in R^{2D \times d_i}; b_{is} \in R^{2D \times 1}\}$ followed by a ReLU layer. Then, the freshly inferred feature vector of d_i-length passes through two successive SCBs before being finalized by another FC layer (also attached with a ReLU layer behind) having setting of $\{W_{ie} \in R^{d_i \times d_{h1}}; b_{ie} \in R^{d_i \times 1}\}$. As a ultimate result of such 2-SCB deep network, the DR-oriented fundus based feature vector with dimension of d_{h1}, namely f_{idr}, is then involved in the third SDN mentioned in Section 3.4.

3.3. EMR-based attributes modality learning

Generally, numerical attributes in the EMR carry a critical role in determining the severity and/or progression of DR, which accordingly leads to the importance of identifying risk factors majorly associated with the DR-based diagnosis results. By referring to existing work and our own assessment, the proposed approach of learning EMR-based attributes modality involves 22 risk factors including Gender, Insulin, HTN, baseline DR grade, Age, Duration, GC-IPL thickness, GC-IPL thinning rate, CAN score, Mean velocity, HbA1c, BUN, Cr, Chole, TG, HDL, LDL, SBP, DBP, BMI, Microalbu, eGFR, of which several interpretations are enumerated in Table 2, for structuring the input vector. It is worth noting that regarding HbA1c, BUN, Cr, Chole, TG, HDL, LDL, SBP, DBP, BMI, Microalbu, eGFR, both baseline and follow-up values are taken into account for obtaining higher-fidelity information of DR risk progression. Eventually, a vector having dimension of 34, namely $f_e \in R^{34 \times 1}$, is fed into a predefined SDN for gaining latent inter-relationships between all elements of the input as well as intra-relationships between baseline and follow-up values of each clinical attribute of interest (i.e., HbA1c,..., eGFR).

Identical to the above-mentioned 2-SCB deep network, the one used in this phase as presented in Fig. 4 also consists of one FC layer with trainable parameters of $\{W_{es} \in R^{34 \times d_e}; b_{es} \in R^{34 \times 1}\}$ followed by two SCBs and finally one FC layer having trainable parameters of $\{W_{ee} \in R^{d_e \times d_{h2}}; b_{ee} \in R^{d_e \times 1}\}$. Accordingly, the retrieved DR-oriented EMR based feature vector with length of d_{h2}, namely f_{edr}, is utilized for the

Table 2
Interpretations of several concerned EMR-based attributes.

EMR-based Attributes	Interpretations
Insulin	Added hormone to balance blood sugar level
HTN	Hypertension
baseline DR grade	Diabetic Retinopathy severity
Duration	Period of suffering from diabetes mellitus
GC-IPL	Ganglion Cell-Inner Plexiform Layer
CAN score	Cardiac Autonomic Neuropathy score
Mean Velocity	Mean conduction Velocity of peripheral nerves
HbA1c	Glycated Hemoglobin
BUN	Blood Urea Nitrogen
Cr	Creatinine
Chole	Cholesterol
TG	Triglyceride levels
HDL	High-Density Lipoprotein
LDL	Low-Density Lipoprotein
SBP	Systolic Blood Pressure
DBP	Diastolic Blood Pressure
BMI	Body Mass Index
Microalbu	Urine Microalbuminuria
eGFR	Estimated Glomerular Filtration Rate

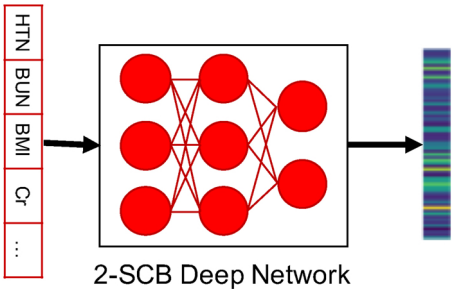


Fig. 4. Workflow of the deep network learning baseline & follow-up EMR-based attributes. The values of interest are organized in form of vector and then fed into a 2-SCB deep network for extraction of DR-oriented EMR based feature vector.

third SDN described in next section.

3.4. Concatenated features learning

To this end, two fundus images and EMR attributes based feature vectors, which represent meaningful DR-oriented information, are aggregated for cross-modal learning in order to attain optimal performance of DR risk progression recognition. As depicted in Fig. 5, these two features are firstly concatenated as follows

$$f_{\mathrm{hdr}} = C[f_{\mathrm{idr}}, f_{\mathrm{edr}}] \tag{8}$$

where the output vector $f_{hdr} \in R^{(d_{h1}+d_{h2})\times 1}$ subsequently goes through a 4-SCB deep network (which is constructed from two FC layers centered by four successive SCBs and a final classification layer at the end) to yield probability of the DR progression. According to the retrieved certainty degree, the ophthalmologists can easily establish appropriate next-visit schedule and treatment plans. It should be noted that the two FC layers at the beginning and end of this network are set up with $\{W_{hs} \in R^{(d_{h1}+d_{h2})\times d_h}; b_{hs} \in R^{(d_{h1}+d_{h2})\times 1}\}$ and $\{W_{he} \in R^{d_h \times d_h}; b_{he} \in R^{d_h \times 1}\}$, respectively, and together followed by the ReLU activation. About the final binary classification layer, a dropout layer [43] with dropping rate of 50% along with a FC layer with parameters of $\{W_{cl} \in R^{d_h \times 2}; b_{cl} \in R^{d_h \times 1}\}$ are utilized for combating overfitting issue and identifying the risk of DR progression, respectively. Remarkably, all the notations defined in this section are summarized in Table 10 for more convenience of following-up.

4. Experiments

This section initially provides characterization of the dataset used to benchmark the proposed approach in this study. Subsequently, evaluation metrics to validate the final classification performance are mentioned and then the enumeration of implementation details is given. Finally, we present experimental results in order to show the effectiveness and potentiality of the proposed methodology in terms of

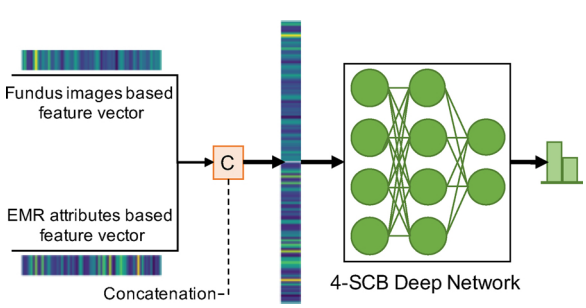
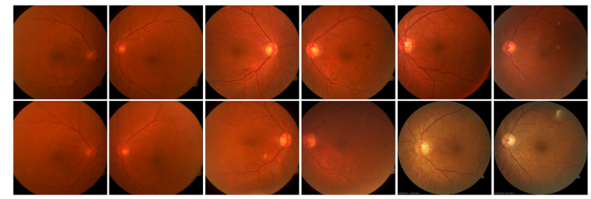


Fig. 5. Workflow of the deep network learning features concatenated from fundus and EMR based feature vectors. The hybrid features are then fed into a 4-SCB deep network for DR risk progression identification.



(a)

TG	TG(fu)	HDL	HDL(fu)	LDL	LDL(fu)	SBP	SBP(fu)	DBP	DBP(fu)	BMI	BMI(fu)
92	82	39	43	76	68	128	123	81	78	27.5	27.5
277	99	44	53	216	74	130	116	80	60	23.1	21.5
105	115	41	35	78	83	115	121	73	66	21.6	22.7
140	139	39	28	106	64	151	125	82	70	28	28.68
188	139	28	33	110	82	120	134	70	60	30.5	30.95
61	65	66	69	102	103	131	121	75	66	22.8	21.63
130	128	41	39	45	53	113	100	73	58	21.2	20.42
108	218	32	36	88	107	142	137	77	79	22.6	23.95
122	142	51	43	119	109	137	114	96	63	24.2	23.81
55	48	48	42	82	68	132	127	72	70	26	24.68

(b)

Fig. 6. Several fundus- and EMR-based samples of the benchmark dataset. (a) Top: baseline samples; Bottom: corresponding follow-up samples. (b) Example values of baseline and follow-up TG, HDL, LDL, SBP, DBP, BMI of 10 DM patients.

DR risk progression identification, which assists the ophthalmologists to decide suitable treatment and follow-up schedule for each individual DM patient.

4.1. Dataset

The historical record of DM patients used in this study is collected from Kyung Hee University Medical Center, Seoul, South Korea. There are totally 96 patients with full records of baseline and follow-up EMR-based attributes of interest (totally 22 risk factors) and retinal fundus images. Correspondingly, the ground-truth classification labels of DR progression for model training are determined and agreed by heuristics of five experienced ophthalmologists in the Department of Ophthalmology. Fig. 6 respectively presents several baseline and corresponding follow-up samples of fundus images (Fig. 6a) as well as EMR-based attributes' values (Fig. 6b) in the benchmark dataset. Note that the period for collecting follow-up data with respect to the baseline ones arbitrarily ranges from 6 to 40 months. Also, the number of raw fundus images with respect to DR grades as well as resolution is further enumerated in Table 3, which shows the challenging DR severity progression issue and resolution difference.

Moreover, it is usual that the DM patients may be screened by different nonmydriatic fundus cameras and conditions at each time of checking-up, which leads to the dissimilarity of inherent properties between taken fundus images of the same person. As a consequence, the dataset evaluated in this paper is challenging due to the diverse representation of numerical factors as well as varied luminance and chrominance in the considered fundus photography.

Regarding the considered 22 EMR-based attributes, their values can be split into three different types, i.e., nominal (Gender, Insulin, HTN), categorical (baseline DR grade), and numerical (the others) data. Accordingly, corresponding normalization approaches as presented in Table 4 are executed to make the model training converge smoothly. Specifically, values of Gender, Insulin, HTN are normalized into either −1 or 1 due to their nominal property. About baseline DR grade, we categorize its severity as 0, 1, 2, 3, 4 corresponding to None, Mild, Moderate, Severe, and Proliferative DR. Meanwhile, Gaussian normalization with mean of zero and standard deviation of one is applied onto the remaining numerical risk factors.

With respect to the input fundus images, we firstly implement

Table 3
The number of raw fundus images with respect to DR grades as well as resolution in the dataset.

Property	DR grade					Resolution		
	None	Mild	Moderate	Severe	Proliferative	1326 × 1594	2656 × 3192	3608 × 3608
Baseline	28	26	42	0	0	56	54	82
Follow-up	13	27	25	27	4			

Table 4
Different data types of the considered EMR-based attributes and corresponding normalization values.

Data Type	EMR-based Attributes	Normalization Value
Nominal data	Gender, Insulin, HTN	−1, 1
Categorical data	baseline DR grade	0, 1, 2, 3, 4
Numerical data	Others	Gaussian $\sim \mathcal{N}(0, 1)$

channel-wise mean subtraction for color normalization, then down-sample the resolution and finally implement center-cropping to the size of 224 × 224, of which the results are delivered to the pretrained ResNet-50 for the extraction of fundus-based information beside the aforementioned EMR-based ones.

4.2. Evaluation metrics

In order to assess the effectiveness of the proposed Tri-SDN on the above-mentioned dataset, we utilize common evaluation metrics such as accuracy (Acc), sensitivity/recall (Sen), precision (Pre), specificity (Spe), and Area Under the Receiver Operating Characteristics (AUROC). Given rates of true negative (TN), false positive (FP), false negative (FN), and true positive (TP) calculated from the retrieved confusion matrix, the corresponding formulas of those measures are as follows

$$\begin{aligned} \text{Acc} &= \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}} \\ \text{Sen} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Pre} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Spe} &= \frac{\text{TN}}{\text{TN} + \text{FP}} \end{aligned}$$

4.3. Implementation details

In this work, Pytorch [44] and Scikit-learn [45] are taken into account for training and evaluating the proposed deep architecture on one NVIDIA 1080TI GPU, respectively. Each batch of eight patients' data is fed into the proposed network sequentially. It should be noted that parameters of the constructed architecture are initialized by He's approach [46]. Then, for each training sample, the compatibility between the resulting prediction scores (s) of DR progression and corresponding ground-truth labels (g) is measured by a predefined cross-entropy loss function as follows

$$\mathcal{L} = -[o \log(s) + (1 - o)\log(1 - s)] \tag{9}$$

where $o = 1$ if the predicted class of the considered sample is identical to its actual label g, otherwise $o = 0$. Based on the loss averaged from all training samples, Adam optimizer [47] (with learning rate, decay rate of moving average of gradient's first and second moment assigned at 0.001, 0.9, and 0.999, respectively) is adopted for the purpose of optimizing the initialized parameters of the proposed model in an end-to-end manner. In addition to the aforementioned dropout tactic, the weight decay of 0.0001 is involved to boost the generalization capability of the Tri-SDN. Finally, the 10-fold cross validation strategy is utilized to evaluate the proposed method. In specific, for each fold, we train the constructed model in 20 epochs and then execute the testing phase with predefined evaluation metrics. The reported results are mean and standard deviation values of the corresponding fold-wise metrics.

4.4. Experimental results

To this end, we deliver ablation studies of bimodal learning, skip-connection mechanism, and the chosen set of EMR-based attributes respectively to express power of the proposed architecture for DR risk progression recognition through experimental results. Specifically, such ablation studies enable us to move toward the objective of making deep learning better interpretable to the domain expert for increasing trustfulness of artificial intelligence in biomedical area [48]. Finally, performance comparison with the state-of-the-art methods and subsequent discussion are presented respectively.

4.4.1. Ablation study for bimodal learning

In specific, we perform three different network architectures to show the effectiveness of fundus images involvement in addition to conventional EMR-based attributes for the DR risk progression prediction by the retrieved experimental results. Firstly, a network of five hidden layers with 256 units at each layer, namely EMR-DN, is taken into account to only exploit the EMR-based attributes for the binary classification of DR progression. Then, a trilogy of conventional deep networks involving both baseline and follow-up fundus images as well as EMR-based attributes, which is called Tri-DN, is considered. Notably, in this architecture, both sub-networks learning fundus photography and EMR-based risk factors have same setting of five 256-node layers while the third one, which extracts concatenated features, is constituted by 10 layers of 128 units. Finally, the similar version with additional embedding of skip-connection mechanism, a.k.a. the proposed Tri-SDN described in Section 3, is validated.

As quantitatively presented in Table 5, the Tri-DN reaches higher mean Acc, Pre, Spe, and AUROC of 3.2%, 3.6%, 7.7%, and 2.8%, respectively, while keeping same Sen compared to those of the EMR-DN. This implies that the aforementioned hypothesis, in which the latent correlations between baseline and follow-up fundus images can powerfully contribute to the prediction of DR progression for further appropriate follow-up scheduling, works effectively. However, as noted before, such kind of bimodal learning leads to the increase in (i) the complication of network structure (which consists of three sub-networks for learning image-based, numerical, and hybrid feature vectors,

Table 5
Experiment 1 - Quantitative results of different strategies for DR risk progression identification on the benchmark dataset. EMR-DN means conventional deep network only learning EMR-based attributes for the prediction; Tri-DN means conventional deep networks with bimodal learning; Tri-SDN means skip-connection deep networks with bimodal learning. Note that boldface numbers indicate the best performance of corresponding metrics.

Strategy	Acc (%)	Sen (%)	Pre (%)	Spe (%)	AUROC (%)
EMR-DN	83.3 ± 1.1	**96.5 ± 0.7**	79.7 ± 1.3	64.1 ± 3.5	83.6 ± 2.0
Tri-DN	86.5 ± 0.9	**96.5 ± 1.5**	83.3 ± 1.2	71.8 ± 3.0	86.4 ± 1.5
Tri-SDN	**90.6 ± 0.7**	**96.5 ± 1.2**	**88.7 ± 1.0**	**82.1 ± 3.0**	**88.8 ± 1.4**

respectively) and (ii) the total number of hidden layers for better learning capacity, which requires an exhaustive job of hyperparameters tuning for ensuring the training stage to converge smoothly and optimally. Therefore, we opt for applying the proposed SCB in the subnetworks, which turns Tri-DN into Tri-SDN, in order to resolve the issue of training convergence and accordingly leverage the final classification performance. Apparently, the experimental results in Table 5 show the effectiveness of skip-connection mechanism in deep and complicated networks. In specific, except for Sen, the mean performance of other metrics increases significantly, i.e., 2.4–10.3%. Moreover, it can be observed from the results that average standard deviation values of Acc, Sen, Pre, and AUROC are approximately 1.0–1.5% while that of Spe fluctuates around 3.0%. Correspondingly, the reasons are twofold: (i) the misclassification of no progression into having progression class is dominant in a couple of validation folds, and (ii) the dataset size is kind of small (96 patients) while the involved attributes are diverse, which can be considered as an indirect affect. In addition, it should be noted that the performance of mean values is mainly discussed for further ablation studies.

Especially, it is known that the Receiver Operating Characteristic (ROC) and the associated AUROC metric are the highest-fidelity measures showing how well a model differentiates a patient experiencing the DR development from those who do not undergo such progression. Thus, corresponding plots of the three encountered strategies, i.e., EMR-DN, Tri-DN, and Tri-SDN, are further presented in Fig. 7a.

Obviously, comparison of mean ROC between the three strategies illustrated in Fig. 7a proves the superior performance of the Tri-SDN over the remaining two networks. From such observation, it can be stated that the proposed Tri-SDN offers two advantages to leverage the task of DR risk progression prediction, i.e., (i) taking into consideration of meaningful depth-wise features extracted from fundus images, and (ii) associating with SCBs for ease of training plus performance optimization.

4.4.2. Ablation study for in-depth role of skip-connection mechanism

Previous experiment has shown the robustness of SDNs in coordination with the cross-modal learning manner compared to the conventional approaches. This sub-section further provides an in-depth analysis of the contribution of particular SDNs at fundus image modality, EMR-based attributes modality, and concatenated features learning stages to the final performance. Concretely, we deploy different strategies in three learning phases as follows

- SDD: one 2-SCB SDN is used to acquire features of fundus images, the EMR-based and concatenated features learning phases are taken over by two conventional DNNs of 5 and 10 layers, respectively.
- DSD: one 2-SCB SDN at the phase of learning EMR-based features, the fundus-based and concatenated features learning phases are taken over by two conventional DNNs of 5 and 10 layers, respectively.
- DDS: one 4-SCB SDN at the phase of learning concatenated features, the remainders are learned by conventional 5-layer DNNs.
- DSS: one conventional 5-layer DNN at the phase of learning fundus-based features, the EMR-based and concatenated features learning phases are taken over by 2-SCB and 4-SCB SDNs, respectively.
- SDS: one conventional 5-layer DNN at the phase of learning EMR-based features, the fundus-based and concatenated features learning phases are taken over by 2-SCB and 4-SCB SDNs, respectively.
- SSD: one conventional 10-layer DNN at the phase of learning concatenated features, the remainders are learned by 2-SCB SDNs.

Note that there are 256 and 128 hidden nodes in the 5-layer and 10-layer DNNs, respectively, as default settings.

Consequently, the corresponding performance in terms of predefined evaluation metrics and ROC is reported in Table 6 and Fig. 7b, respectively. Regarding the in-depth role of skip-connection mechanism

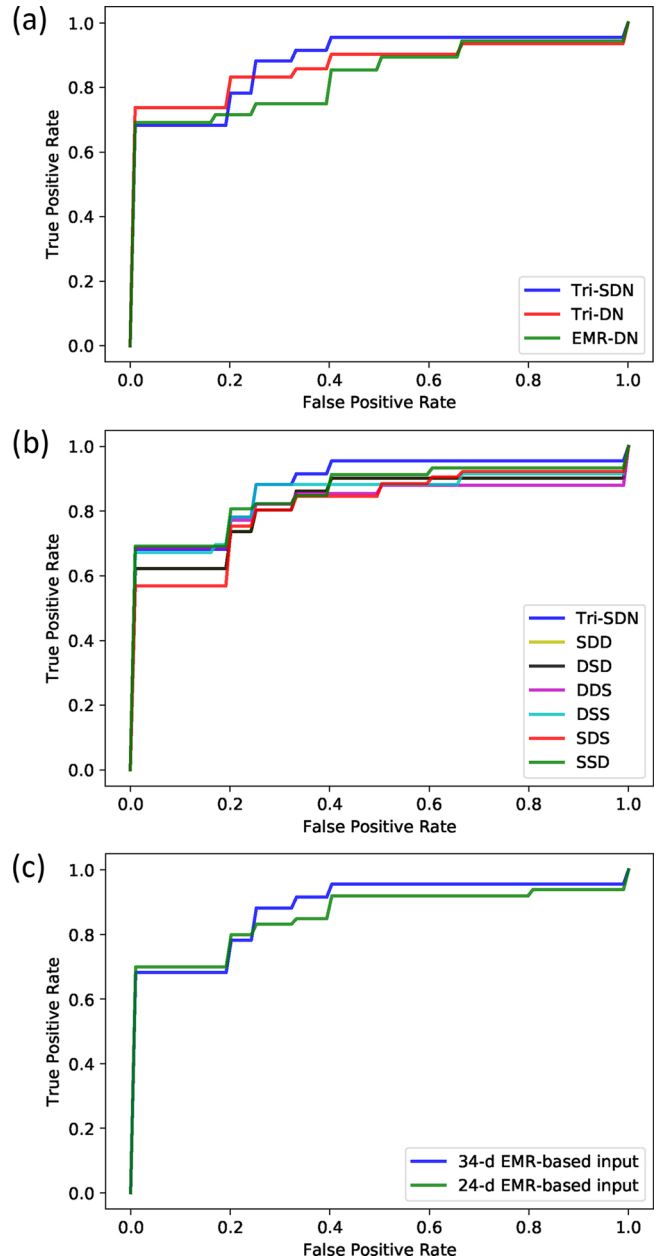


Fig. 7. Comparisons of Mean ROC between different strategies. (a) EMR-DN, Tri-DN, and Tri-SDN. (b) SDD, DSD, DDS, DSS, SDS, SSD, and Tri-SDN. (c) using EMR-based inputs having dimension of 34 and 24.

Table 6
Experiment 2 - Quantitative results of different strategies (i.e., SDD, DSD, DDS, DSS, SDS, SSD, and Tri-SDN) for DR risk progression identification on the benchmark dataset. Note that boldface numbers indicate the best performance of corresponding metrics.

Strategy	Acc (%)	Sen (%)	Pre (%)	Spe (%)	AUROC (%)
SDD	86.5 ± 0.8	91.2 ± 1.3	86.7 ± 1.2	79.5 ± 3.0	82.7 ± 2.2
DSD	87.5 ± 0.9	91.2 ± 1.3	88.1 ± 1.3	82.1 ± 1.8	88.0 ± 1.1
DDS	89.6 ± 0.7	**96.5 ± 0.9**	87.3 ± 1.0	79.5 ± 2.9	82.7 ± 2.9
DSS	88.5 ± 0.8	91.2 ± 1.3	**89.7 ± 1.0**	**84.6 ± 3.0**	84.7 ± 1.5
SDS	87.5 ± 0.7	91.2 ± 1.3	88.1 ± 1.0	82.1 ± 3.0	81.6 ± 2.1
SSD	85.4 ± 0.9	94.7 ± 1.0	83.1 ± 1.3	71.8 ± 3.1	86.0 ± 1.9
Tri-SDN	**90.6 ± 0.7**	**96.5 ± 1.2**	88.7 ± 1.0	82.1 ± 3.0	**88.8 ± 1.4**

Table 7
Experiment 3 - Quantitative results of different number of involved EMR-based attributes for DR risk progression identification on the benchmark dataset. Note that the proposed Tri-SDN is utilized in this experiment and boldface numbers indicate the best performance of corresponding metrics.

Input Dim.	Acc (%)	Sen (%)	Pre (%)	Spe (%)	AUROC (%)
24	86.5 ± 1.1	94.7 ± 1.1	84.4 ± 1.5	74.4 ± 3.0	86.1 ± 1.4
34	**90.6 ± 0.7**	**96.5 ± 1.2**	**88.7 ± 1.0**	**82.1 ± 3.0**	**88.8 ± 1.4**

at the learning phases, it is noticeable that using SDN to acquire meaningful features from EMR-based attributes is crucial for the high recognition performance in terms of AUROC despite different number of SDNs used in the proposed architecture. Particularly, the corresponding strategies (i.e., DSD, DSS, and SSD) yield average AUROC of 86.2%, which is much higher than the value of 82.3% averaged from AUROCs of the remaining ones (i.e., SDD, DDS, and SDS). We argue that the massive involvement of 22 risk factors makes the gradient signals difficult to reach globally optimal target, which is then effectively managed by the usage of SDN at this learning phase.

Furthermore, since a large number of layers should be used for maximizing learning capability of concatenated features, additionally utilizing SDN (i.e., DSS) results in impressive Pre (89.7%) and Spe (84.6%). Finally, for an optimal all-round performance, we opt for using SDNs in all learning phases of interest, a.k.a. Tri-SDN, and attain best performance with higher Acc of 1.0–5.2%, Sen of up to 5.3%, and especially AUROC of 0.8–7.2% in comparison with those of the six strategies discussed in this experiment. Note that AUROC is the best indicator of overall performance as mentioned before. Moreover, it can be realized that the ROC of Tri-SDN represented in Fig. 7b mostly lies on top of the rest.

4.4.3. Ablation study for different number of involved EMR-based attributes

The importance of SDN in particular and Tri-SDN in common has been comprehensively studied in aforementioned experiments. To this end, we further investigate how performance of the proposed deep model is affected by fewer number of input EMR-based attributes. As previously mentioned, the considered systemic risk factors can be divided into two groups, i.e., (i) those with standalone values like Gender, Insulin, HTN, baseline DR grade, Age, Duration, GC-IPL thickness, GC-IPL thinning rate, CAN score, Mean velocity, and (ii) those having baseline and follow-up values such as HbA1c, BUN, Cr, Chole, TG, HDL, LDL, SBP, DBP, BMI, Microalbu, eGFR. Subsequently, we target to compare the performance when using full set of chosen systemic attributes (as done in earlier experiments) and only the latter group as EMR-modality inputs of the proposed Tri-SDN. In other words, these EMR-related inputs have dimensions of 34 and 24, respectively, while the corresponding fundus-based ones remain unchanged.

From the obtained results exhibited in Table 7 and Fig. 7c, it is undeniable that the utilization of all chosen systemic attributes, i.e., 34-d EMR-based input vector, totally dominates the approach only using the specified 12 risk factors in all measurements. This leads to the fact that although the concerned attributes with standalone values seem to be meaningless with respect to the risk progression of DR, its combination with the ones having both baseline and follow-up measures plays a critical role in the proposed architecture for achieving optimal prediction performance. In short, it can be implied that the standalone attributes and basis profile information of the DM patients hold a significant importance besides the meaningful features underlying the fundus photography for efficiently identifying the risk progression of DR.

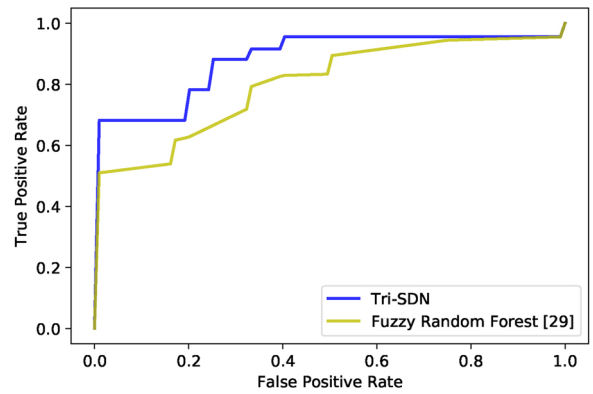


Fig. 8. Comparisons of Mean ROC between the proposed Tri-SDN and the Fuzzy Random Forest-based approach [29] that uses EMR-based attributes only. Note that color print is recommended for fully visual representation of strategies' ROC.

4.4.4. Performance comparison with state-of-the-art methods

Previously tested strategies have shown the effectiveness of the proposed Tri-SDN. Accordingly, we further compare our bimodal learning mechanism with the state-of-the-art methods exploiting single modality for DR-related problems (i.e., using EMR-based attributes only [29] and fundus photography only [35]) on our benchmark dataset. Particularly, we apply fuzzy random forest (FRF) introduced in [29] into the 34-d EMR-based input vector for predicting DR risk progression. Meanwhile, the 11-layer CNN proposed in [35] is adopted to classify DR grades of the baseline and associated follow-up fundus images, which can then be manually utilized for inferring the corresponding DR progression status. Hence, the mean ROC and AUROC metrics are not reported for this case. Finally, it is worth noting that the compared methods in this sub-section are realized by our own implementation with the same 10-fold cross validation procedure.

It can be perceived from Table 8 that the proposed method outperforms the compared ones by a large margin for all the evaluation metrics. Furthermore, a significant difference between our Tri-SDN's mean ROC and that of the FRF-based approach is demonstrated in Fig. 8. Obviously, since the benchmark dataset size is quite small, the re-implemented ones, especially the 11-layer CNN model, cannot optimally generalize useful DR-related information. Moreover, it is also quite difficult for FRF to handle such large number of EMR-based risk factors in our dataset. On the other hand, by the bimodal-learning oriented approach, the proposed Tri-SDN can effectively suppress the necessity of large-scale dataset consideration while being able to seamlessly acquire comprehensive information from numerous systemic features.

Table 8
Experiment 4 - Performance comparison with the state-of-the-art techniques on the benchmark dataset. Note that the boldface numbers indicate the best performance of corresponding metrics in terms of DR risk progression identification.

Approach	Acc (%)	Sen (%)	Pre (%)	Spe (%)	AUROC (%)
Fuzzy Random Forest [29]	78.1 ± 1.6	84.2 ± 2.1	80.0 ± 1.2	69.2 ± 3.0	80.3 ± 1.8
11-layer CNN [35]	72.9 ± 2.1	81.1 ± 3.8	72.9 ± 2.2	62.8 ± 4.4	-
Tri-SDN	**90.6 ± 0.7**	**96.5 ± 1.2**	**88.7 ± 1.0**	**82.1 ± 3.0**	**88.8 ± 1.4**

Table 9
Statistics of recognition performance in terms of different progression rates on specific baseline and follow-up DR grades. Note that baseline and follow-up categories are linked by '-' (dash) in 'DR grades' columns; 'Correct Pred. Ratio' stands for correct prediction ratio of *a/b*, where *a* and *b* are the number of correct and total predictions, respectively.

No Progression		With Progression			
		One-level Progression		Two-level Progression	
DR grades	Correct Pred. Ratio	DR grades	Correct Pred. Ratio	DR grades	Correct Pred. Ratio
None-None	10/13	None-Mild	11/12	None-Moderate	3/3
Mild-Mild	13/ 15	Mild-Moderate	7/7	Mild-Severe	4/4
Moderate-Moderate	13/15	Moderate-Severe	22/23	Moderate-Proliferative	4/4

4.4.5. Discussion on the recognition performance in terms of different progression rates on specific baseline and follow-up DR grades

In this section, we examine how well the proposed Tri-SDN performs in terms of different progression rates on specific baseline and follow-up DR grades by summarizing statistics of interest in Table 9. Since all the baseline DR grades are either None, Mild, or Moderate as shown in Table 3, the benchmark dataset introduces three scenarios of baseline - follow-up DR progression levels: no progression (None-None, Mild-Mild, and Moderate-Moderate); one-level progression (None-Mild, Mild-Moderate, and Moderate-Severe); and two-level progression (None-Moderate, Mild-Severe, and Moderate-Proliferative).

Obviously, the proposed approach can perfectly handle the cases of two-level progression thanks to the distinctive difference of characteristics between both baseline and follow-up fundus images as well as EMR-based attributes. Meanwhile, the failures mainly occur in the remaining cases, especially the no-progression ones. It is argued that the proposed Tri-SDN seems to be sensitive to some unexpectedly latent properties which are unnecessary for the consideration of DR risk progression. Thus, in order to alleviate such biases learning issue, it is potential to additionally involve other retinal-based materials such as Fluorescein Angiography (FA) and Optical Coherence Tomography Angiography (OCTA) with the proposed combination mechanism, which can generate a more stable bound of underlying conditions discriminating between none and one-level progression.

Table 10
List of notations defined in Section 3.

Notation	Meaning
Skip-Connection Block (SCB)	
$X \in R^{d\times 1}$	Input of SCB
$\{W_1 \in R^{d\times d},\ b_1 \in R^{d\times 1}\}$	Parameters of 1st FC layer
$H_1' \in R^{d\times 1}$	Activated features after 1st FC layer
$\{W_2 \in R^{d\times d},\ b_2 \in R^{d\times 1}\}$	Parameters of 2nd FC layer
$H_2 \in R^{d\times 1}$	Output of 2nd FC layer
$Y \in R^{d\times 1}$	Output of SCB
Fundus Image Modality Learning	
$f_{ib} \in R^{w\times h\times D}$	Baseline feature map
$f_{if} \in R^{w\times h\times D}$	Follow-up feature map
$\mathcal{G}^d$	Average response of dth channel of f_{ib} or f_{if}
$g_{ib} \in R^{D\times 1}$	Baseline feature vector
$g_{if} \in R^{D\times 1}$	Follow-up feature vector
$G \in R^{2D\times 1}$	Concatenation feature vector
$\{W_{is} \in R^{2D\times d_i},\ b_{is} \in R^{2D\times 1}\}$	Parameters of 1st FC layer in 2-SCB network
$\{W_{ie} \in R^{d_i\times d_{h1}};\ b_{ie} \in R^{d_i\times 1}\}$	Parameters of last FC layer in 2-SCB network
$f_{idr} \in R^{d_{h1}\times 1}$	DR-oriented fundus based feature vector
EMR-based Attributes Modality Learning	
$f_e \in R^{34\times 1}$	EMR-based feature vector
$\{W_{es} \in R^{34\times d_e};\ b_{es} \in R^{34\times 1}\}$	Parameters of 1st FC layer in 2-SCB network
$\{W_{ee} \in R^{d_e\times d_{h2}};\ b_{ee} \in R^{d_e\times 1}\}$	Parameters of last FC layer in 2-SCB network
$f_{edr} \in R^{d_{h2}\times 1}$	DR-oriented EMR based feature vector
Concatenated Features Learning	
$f_{hdr} \in R^{(d_{h1}+d_{h2})\times 1}$	Concatenated feature vector
$\{Whs \in R^{(d_{h1}+d_{h2})\times d_h}; b_{hs} \in R^{(d_{h1}+d_{h2})\times 1}\}$	Parameters of 1st FC layer in 4-SCB network
$\{Whe \in R^{d_h\times d_h\times 1}; bhe \in R^{d_h\times 1}\}$	Parameters of last FC layer in 4-SCB network
$\{W_{cl} \in R_h^{d\times 2}; b_{cl} \in R^{d_h\times 1}\}$	Parameters of the classification layer

5. Conclusion

This study introduced a bimodal learning approach using Trilogy of Skip-connection Deep Networks for DR risk progression prediction. Specifically, the proposed technique performed an efficient combination of two different modalities, i.e., fundus photography and EMR-based numerical attributes, for thoroughly exploiting vital DR-oriented information since it is hypothesized that fundus images carry latent clinical representations of retinal vessel, microaneurysms, exudates, hemorrhages, etc, which are not explicitly addressed in EMR. According to the experimental results of various strategies, the involvement of features extracted from color fundus images, the selective consideration of typical EMR-based attributes along with the utilization of skip-connection mechanism in the Tri-SDN improved the performance of DR risk progression recognition significantly (in terms of well-known evaluation metrics like Acc, Sen, Pre, Spe, and AUROC). However, in order to operate effectively, the proposed model requires full set of pre-specified fundus images and EMR-based attributes as desired inputs, which heavily depends on affordability of a particular patient with respect to all of the necessary clinical experiments.

In the future, we aim to additionally take into account underlying context in other modalities of retinal image such as FA and OCTA by using deep learning for enabling the DR-oriented computer-aided systems to assist the ophthalmologists more efficiently in terms of follow-up scheduling and corresponding treatment plans decision.

Conflict of interest

None declared.

Authors' contributions

Cam-Hao Hua proposed and elaborated the methodology, implemented and analyzed the experiments, and wrote the paper; Sung-Ho Bae conceptualized the core idea and provided technical comments during implementation stage as well as for paper presentation; Thien Huynh-The and Thuong Le-Tien suggested ablation studies for the experiments besides providing additional comments on figures visulization and technical details; Gwang Hoon Park and Jaehun Bang gave advice for existing work re-implementation and in-depth analysis and discussion of the experiments; Wajahat Ali Khan contributed in English proofreading and finalized content flow in the manuscript; Kiyoung Kim and Seung-Young Yu participated in data acquisition and delivered medical-related details for inputs finalization and the write-up; Sungyoung Lee supervised the whole process and provided advisory comments.

Summary table

What was already known on the topic:

- Diabetic Retinopathy (DR) based complications and its progression can be detected by either fundus photography or Electronic Medical Records (EMR) based attributes.
- Inspecting the changes of these multi-modal factors by time enables the ophthalmologists to identify whether there is any risk progression of DR for setting up proper next-visit schedule and cost-effective treatment plans.

What this study added to our knowledge:

- An approach of bimodal learning that exhaustively utilizes both baseline and follow-up information of retinal fundus images and EMR-based attributes for identifying the DR risk progression.
- The Trilogy of Skip-connection Deep Networks that combines features of different modalities (i.e., color images and numerical risk factors) smoothly and effectively during training and testing processes, respectively.
- The proposed bimodal learning method is able to suppress the necessity of large-scale fundus image dataset involvement and the costly task of engineering most impactful EMR-based risk factors.

Acknowledgment

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-0-01629) supervised by the IITP (Institute for Information & communications Technology Promotion). This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00655).

References

[1] Grading diabetic retinopathy from stereoscopic color fundus photographs-an extension of the modified airlie house classification: Etdrs report number 10, Ophthalmology 98 (5, Supplement) (1991) 786–806.
[2] J. Chhablani, A. Sharma, A. Goud, H.K. Peguda, H.L. Rao, V.U. Begum, G. Barteselli, Neurodegeneration in type 2 diabetes: Evidence from spectral-domain optical coherence tomography neurodegeneration in type 2 diabetes, Invest. Ophthal. Visual Sci. 56 (11) (2015) 6333–6338.
[3] M. Pedrosa, J.M. Silva, J.F. Silva, S. Matos, C. Costa, Screen-dr: Collaborative platform for diabetic retinopathy, Int. J. Med. Inform. 120 (2018) 137–146.
[4] S. Vujosevic, A. Muraca, M. Alkabes, E. Villani, F. Cavarzeran, L. Rossetti, S. De Cilla', Early microvascular and neural changes in patients with type 1 and type 2 diabetes mellitus without clinical signs of diabetic retinopathy, Retina 39 (3) (2019) 435–445.
[5] M.L. Rasmussen, R. Broe, U. Frydkjaer-Olsen, B.S. Olsen, H.B. Mortensen, T. Peto, J. Grauslund, Microaneurysm count as a predictor of long-term progression in diabetic retinopathy in young patients with type 1 diabetes: the Danish cohort of pediatric diabetes 1987 (dcpd1987), Graefe's Arch. Clin. Exp. Ophthalmol. 253 (2014) 199–205.
[6] R. Klein, K.E. Lee, L. Danforth, M.Y. Tsai, R.E. Gangnon, S.E. Meuer, T.Y. Wong, C.Y. Cheung, B.E. Klein, The relationship of retinal vessel geometric characteristics to the incidence and progression of diabetic retinopathy, Ophthalmology 125 (11) (2018) 1784–1792.
[7] K. Kim, E.S. Kim, S.-Y. Yu, Longitudinal relationship between retinal diabetic neurodegeneration and progression of diabetic retinopathy in patients with type 2 diabetes, Am. J. Ophthalmol. 196 (2018) 165–172.
[8] L.Z. Heng, O. Comyn, T. Peto, C. Tadros, E. Ng, S. Sivaprasad, P.G. Hykin, Diabetic retinopathy: pathogenesis, clinical grading, management and future developments, Diabetic Med. 30 (6) (2013) 640–650.
[9] S.K. Lynch, M.D. Abràmoff, Diabetic retinopathy is a neurodegenerative disorder, Vision Res. 139 (2017) 101–107 diabetic Retinopathy - an Overview.
[10] Y. Jeon Kim, J.-G. Kim, J.-Y. Lee, K. Sub Lee, S. Geun Joe, J.-Y. Park, M.-S. Kim, Y. Hee Yoon, Development and progression of diabetic retinopathy and associated risk factors in korean patients with type 2 diabetes: The experience of a tertiary center, J. Korean Med. Sci. 29 (2014) 1699–1705.
[11] A.J. Jenkins, M.V. Joglekar, A.A. Hardikar, A.C. Keech, D.N. O'Neal, A.S. Januszewski, Biomarkers in diabetic retinopathy, Rev. Diabet. Stud. 12 (1-2) (2015) 159–195.
[12] C. Cardoso, N.C. Leite, E. Dib, G. Salles, Predictors of development and progression of retinopathy in patients with type 2 diabetes: Importance of blood pressure parameters, Scientific Reports 7.
[13] V. Schreur, F. van Asten, H. Ng, J. Weeda, J.M.M. Groenewoud, C.J. Tack, C.B. Hoyng, E. Jong, C. Klaver, B. Jeroen Klevering, Risk factors for development and progression of diabetic retinopathy in Dutch patients with type 1 diabetes mellitus, Acta Ophthalmol. 96 (2018) 459–464.
[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, Curran Associates Inc, USA, 2012, pp. 1097–1105.
[15] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556. arXiv:1409.1556.
[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2818–2826.
[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 770–778.
[18] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, Springer International Publishing, 2015, pp. 234–241.
[19] C.-H. Hua, T. Huynh-The, S. Lee, Convolutional networks with bracket-style decoder for semantic scene segmentation, 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2018) 2980–2985.
[20] M.A. Al-antari, M.A. Al-masni, M.-T. Choi, S.-M. Han, T.-S. Kim, A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification, Int. J. Med. Inform. 117 (2018) 44–54.
[21] Q. Suo, H. Xue, J. Gao, A. Zhang, Risk factor analysis based on deep learning models, Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '16, ACM, New York, NY, USA, 2016, pp. 394–403.
[22] A. Rajkomar, E. Oren, K. Chen, A.M. Dai, N. Hajaj, P.J. Liu, X. Liu, M. Sun, P. Sundberg, H. Yee, K. Zhang, G.E. Duggan, G. Flores, M. Hardt, J. Irvine, Q.V. Le, K. Litsch, J. Marcus, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S.L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N.H. Shah, A.J. Butte, M. Howell, C. Cui, G.S. Corrado, J. Dean, Scalable and accurate deep learning with electronic health records, npj Digital Medicine 1 (2018) 1–10.
[23] C. Cao, F. Liu, H. Tan, D. Song, W. Shu, W. Li, Y. Zhou, X. Bo, Z. Xie, Deep learning and its applications in biomedicine, Genomics, Proteomics Bioinform. 16 (1) (2018) 17–32.
[24] M. Wainberg, D. Merico, A. Delong, B.J. Frey, Deep learning in biomedicine, Nature Biotechnol. 36 (2018) 829 EP -.
[25] M. Skevofilakas, K. Zarkogianni, B.G. Karamanos, K.S. Nikita, A hybrid decision support system for the risk assessment of retinopathy development as a long term complication of type 1 diabetes mellitus, 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology (2010) 6713–6716.
[26] N.S. Bajestani, A.V. Kamyad, E.N. Esfahani, A. Zare, Prediction of retinopathy in diabetic patients using type-2 fuzzy regression model, Eur. J. Oper. Res. 264 (3) (2018) 859–869.
[27] S. Piri, D. Delen, T. Liu, H.M. Zolbanin, A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble, Decision Support Syst. 101 (2017) 12–27.
[28] A. Eleuteri, A.C. Fisher, D.M. Broadbent, M. García-Fiñana, C.P. Cheyne, A. Wang, I.M. Stratton, M. Gabbay, D. Seddon, S.P. Harding, for the Individualised Screening for Diabetic Retinopathy (ISDR) Study Group, Individualised variable-interval risk-based screening for sight-threatening diabetic retinopathy: the liverpool risk calculation engine, Diabetologia 60 (11) (2017) 2174–2182.
[29] P. Romero-Aroca, A. Valls, A. Moreno, R. Sagarra-Alamo, J. Basora-Gallisa, E. Saleh, M. Baget-Bernaldiz, D. Puig, A clinical decision support system for diabetic retinopathy screening: Creating a clinical support application, Telemedicine and e-Health 25 (1) (2019) 31–40.
[30] J.Y. Choi, T.K. Yoo, J.G. Seo, J. Kwak, T.T. Um, T.H. Rim, Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database, PLOS ONE 12 (11) (2017) 1–16.
[31] C. H. Yang, J. Huang, F. Liu, F. Chiu, M. Gao, W. Lyu, I. Lin, J. Tegnér, A novel hybrid machine learning model for auto-classification of retinal diseases, CoRR abs/1806.06423. arXiv:1806.06423.
[32] 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, D.R. Webster, Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, JAMA 316 (22) (2016) 2402–2410.
[33] S.S. Rahim, V. Palade, C. Jayne, A. Holzinger, J. Shuttleworth, Detection of diabetic retinopathy and maculopathy in eye fundus images using fuzzy image processing, Brain Informatics and Health, Springer International Publishing, 2015, pp. 379–388.
[34] H. Pratt, F. Coenen, D.M. Broadbent, S.P. Harding, Y. Zheng, Convolutional neural networks for diabetic retinopathy, Procedia Comput. Sci. 90 (2016) 200–205 20th Conference on Medical Image Understanding and Analysis (MIUA 2016).
[35] M. C. A. Trivino, J. Despraz, J.A. L. Sotelo, C.A. Peña, Deep learning on retina images as screening tool for diagnostic decision support, CoRR abs/1807.09232. arXiv:1807.09232.
[36] L. Zhou, Y. Zhao, J. Yang, Q. Yu, X. Xu, Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images, IET Image Processing 12 (4) (2018) 563–571.

[37] L. Ma, Z. Lu, L. Shang, H. Li, Multimodal convolutional neural networks for matching image and sentence, 2015 IEEE International Conference on Computer Vision (ICCV) (2015) 2623–2631.
[38] R. Abreu, J.d. Santos, E. Bezerra, A bimodal learning approach to assist multi-sensory effects synchronization, 2018 International Joint Conference on Neural Networks (IJCNN) (2018) 1–8.
[39] D. Lu, K. Popuri, G.W. Ding, R. Balachandar, M.F. Beg, Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images, Sci. Rep. 8 (1) (2018) 5697.
[40] N. Asiri, H.A. Abualsamh, Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey, CoRR abs/1811.01238.
[41] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Vol. 9 of Proceedings of Machine Learning Research, Italy (2010) 249–256.
[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, A.C. Berg, F. Li, Imagenet large scale visual recognition challenge, CoRR abs/1409.0575.
[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.
[44] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, NIPS-W (2017).
[45] F. Pedregosa, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[46] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (2015) 1026–1034, https://doi.org/10.1109/ICCV. 2015.123.
[47] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980. arXiv:1412.6980.
[48] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainabilty of artificial intelligence in medicine, Wiley Interdisciplinary Rev.: Data Mining Knowledge Discovery 0 (0) (2019) e1312, https://doi.org/10.1002/widm.1312.