tvst

Article

Deep Learning for Automated Diabetic Retinopathy Screening Fused With Heterogeneous Data From EHRs Can Lead to Earlier Referral Decisions

Min-Yen Hsu[1–3], Jeng-Yuan Chiou[4], Jung-Tzu Liu[5], Chee-Ming Lee[1,2], Ya-Wen Lee[5], Chien-Chih Chou[6,7], Shih-Chang Lo[8,9], Edy Kornelius[8,9], Yi-Sun Yang[8,9], Sung-Yen Chang[5], Yu-Cheng Liu[5], Chien-Ning Huang[8,9], and Vincent S. Tseng[10,11]

[1] Department of Ophthalmology, Chung Shan Medical University Hospital, Taichung, Taiwan
[2] School of Medicine, Chung Shan Medical University, Taichung, Taiwan
[3] Biotechnology Center, National Chung Hsing University, Taichung, Taiwan
[4] Department of Health Policy and Management, Chung Shan Medical University, Taichung, Taiwan
[5] Computational Intelligence Technology Center, Industrial Technology Research Institute, Hsinchu, Taiwan
[6] Department of Ophthalmology, Taichung Veterans General Hospital, Taichung, Taiwan
[7] Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan
[8] Department of Internal Medicine, Division of Endocrinology and Metabolism, Chung Shan Medical University Hospital, Taichung, Taiwan
[9] Institute of Medicine, College of Medicine, Chung Shan Medical University, Taichung, Taiwan
[10] Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
[11] Institute of Data Science and Engineering, National Chiao Tung University, Hsinchu, Taiwan

Correspondence: Vincent S. Tseng, Institute of Data Science and Engineering, Department of Computer Science, National Chiao Tung University, 1001 University Road, Hsinchu, Taiwan.
e-mail: vincetsm@gmail.com

Received: September 9, 2020
Accepted: May 17, 2021
Published: August 17, 2021

Keywords: diabetic retinopathy; fundus image; multimodal data; neural network; data fusion

Citation: Hsu MY, Chiou JY, Liu JT, Lee CM, Lee YW, Chou CC, Lo SC, Kornelius E, Yang YS, Chang SY, Liu YC, Huang CN, Tseng VS. Deep learning for automated diabetic retinopathy screening fused with heterogeneous data from EHRs can lead to earlier referral decisions. Transl Vis Sci Technol. 2021;10(9):18, https://doi.org/10.1167/tvst.10.9.18

Purpose: Fundus images are typically used as the sole training input for automated diabetic retinopathy (DR) classification. In this study, we considered several well-known DR risk factors and attempted to improve the accuracy of DR screening.

Metphods: Fusing nonimage data (e.g., age, gender, smoking status, International Classification of Disease code, and laboratory tests) with data from fundus images can enable an end-to-end deep learning architecture for DR screening. We propose a neural network that simultaneously trains heterogeneous data and increases the performance of DR classification in terms of sensitivity and specificity. In the current retrospective study, 13,410 fundus images and their corresponding nonimage data were collected from the Chung Shan Medical University Hospital in Taiwan. The images were classified as either nonreferable or referable for DR by a panel of ophthalmologists. Cross-validation was used for the training models and to evaluate the classification performance.

Results: The proposed fusion model achieved 97.96% area under the curve with 96.84% sensitivity and 89.44% specificity for determining referable DR from multimodal data, and significantly outperformed the models that used image or nonimage information separately.

Conclusions: The fusion model with heterogeneous data has the potential to improve referable DR screening performance for earlier referral decisions.

Translational Relevance: Artificial intelligence fused with heterogeneous data from electronic health records could provide earlier referral decisions from DR screening.

Introduction

Certain factors, such as improper dietary habits, have led to an increased prevalence of diabetes worldwide. In 2019, the estimated global prevalence of diabetes in adults was 9.3%.[1] Between 2005 and 2014, the prevalence of diabetes in the adult population (20 to 79 years) of Taiwan was reported to increase by 0.3% each year (from 7.15% to 10.10%), representing approximately two million adults currently living with diabetes.[2] Among these adults, 25% also present with diabetic retinopathy (DR),[3] which can cause moderate to severe vision impairment and even blindness.[4] Moreover, 80% of those suffering from type 2 diabetes develop DR within 10 years.[5] However, DR-related vision impairment and blindness can be prevented if patients receive regular fundus examinations, which can lead to early diagnosis and treatment.[6,7] Poor adherence to these preventative examinations has been observed in Taiwan nevertheless, and this is believed to be due to two primary reasons. First, as DR is a chronic disease, most diabetic patients in the early stages of DR are not aware of the condition. Second, there is a shortage of ophthalmologists (approximately 7 per 100,000 people), and most rural areas have a severe lack of these specialists.[8] These issues thus result in low screening rates for DR.

To improve the adherence rate, several image-based diagnostic techniques for DR have been developed.[9–11] These techniques provide high classification performance and can be used by diabetes clinicians for early DR screening and referral, including those who don't have an ophthalmological background. Such solutions have been designed to enable patients with diabetes to obtain a prescription from their clinician and complete an eye examination at the same clinic, which in turn increases monitoring and regular follow-ups. These techniques used more than 80,000 images as part of their training models. In the product development stage, researchers designing such models may not need to collect such a large amount of image data if they also had access to historical electronic health records (EHRs), potentially reducing costs associated with annotation and training times.

EHRs often contain information on vital signs, laboratory results, clinical records, and previous medical care, which have been widely used in medical research and can be used to support a clinician's decision as part of early screening and diagnosis programs, or to predict disease progression and risk factors.[12,13] EHRs may provide complementary features for image data, and increase model interpretability. Several risk factors for DR, as determined by previous cross-sectional studies, can be retrieved from a patient's EHR, such as age, gender, body mass index (BMI), diabetes history, hypertension history, glycosylated hemoglobin (HbA1c), and systolic blood pressure.[5,14–17] Additionally, characteristics associated with different countries and ethnicities should be considered so clinicians may better understand and improve the classification results of deep learning models (e.g., the average HbA1c level may differ by race/ethnicity).[18] This broad scope of information can lead to improvements in the performance and robustness of models when compared with current techniques that do not consider EHR information when focusing on specific segments of a population. It could improve screening performance and lead to more precise medicine. Furthermore, a hybrid model combining heterogeneous features from image and nonimage data that extracts the most important features could potentially provide a broader picture for the outcome of interest.[12]

As a deep learning hybrid model can handle data obtained from different sources, it has been used in multimodal data fusion for comprehensive analysis and disease classification. A variety of different models have been proposed, such as a deep learning multimodal combining cervical images and nonimage information for cervical dysplasia diagnosis,[19] a text-image embedding network using both chest X-rays and free-text clinical reports of radiological scans for thorax disease classification,[20] and a hybrid decision support system combining a feedforward neural network, a classification and regression tree, and a hybrid wavelet neural network for the risk assessment of DR based on fundus imaging and related EHR data.[21] However, previous studies have mostly neglected the integration of fundus imaging and nonimaging information based on end-to-end convolutional neural networks (CNNs) for DR classification.

Previous studies mainly focused on the application of model training to either image or nonimage data, a knowledge gap remains concerning the development of a hybrid CNN model combining these heterogeneous datasets. To address this issue, we set out to design a fusion model with an end-to-end neural network which combines both CNN and multilayer perceptron (MLP) for training heterogeneous data simultaneously that could be used for early screening of DR severity classification. Our fusion model was designed to minimize the total loss from heterogeneous data. In this study, we developed an automated DR screening approach to distinguish between nonreferable DR (NRDR) and referable DR (RDR) based on the International Clinical Diabetic Retinopathy Disease Severity Scale.[22] No apparent DR and mild nonprolifera-

tive DR were defined as NRDR, whereas moderate and severe nonproliferative and proliferative DR were defined as RDR. Finally, we verified that the proposed model balanced both image and nonimage information to enhance interpretation and insight into DR classification, and then evaluated the results to determine whether the proposed model could be better than a single training source model.

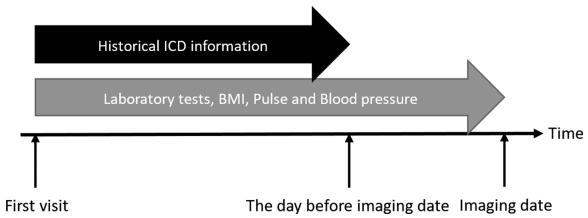


Figure 1. EHR information, including the patient's baseline characteristics, laboratory test results, and historical ICD information.

Methods

Dataset

In this retrospective study between 2013 and 2018, we collected 13,410 fundus images from 6,566 patients with a history of diabetes at the Chung Shan Medical University Hospital in Taiwan. All data were anonymized before processing and analyzed. We excluded data from grayscale, noncircular, and poor-quality images during training and analysis. Images were obtained from a digital nonmydriatic retinal camera (ZEISS VISUCAM 200) with 45-degree fields of view. The image resolutions ranged from 721 × 723 to 2846 × 4287 pixels, with 2055 × 2123 pixels as the predominant resolution, which corresponded to 96% of the images. Each fundus image was diagnosed clinically as either NRDR or RDR by a panel of ophthalmologists based on the International Clinical Diabetic Retinopathy Disease Severity Scale.[22] This study was approved by the Institutional Review Board at the Chung Shan Medical University Hospital and the Industrial Technology Research Institute. Informed consent was not required because of the retrospective nature of the study and the fact that the subject data used was not traceable to the patient.

We also used EHR data to support fundus image classification. This data included the patients' baseline characteristics (birth, gender, smoking status, hypertension, BMI, pulse, and blood pressure), laboratory test results (HbA1c, cholesterol, lipoprotein, total cholesterol, triglyceride, creatinine, estimated glomerular filtration rate, blood glucose, and urine protein), and disease history according to The International Classification of Diseases (ICD) 9th or 10th revision. Furthermore, we mapped the ICD codes into 136 categories for sparse dimension reduction based on the chronic condition indicator using clinical classification software (CCS).[23] Figure 1 shows that the records of birth, gender, and smoking status were collected once during the first patient visit, however, most laboratory test results and details of several vital signs were obtained from either the day before the imaging date or on the date of imaging. Measuring the time difference between them, the meaning of the time slot was considered. Historical ICD data was gathered before the imaging date. The patient's imaging age was then calculated as the difference between the imaging date and the date of birth.

The fusion data combining both imaging and EHR information was further split randomly into five distinct sets by anonymous subject identification numbers for fivefold cross-validation. All models were not trained using any of the patients included in the testing set. The whole dataset comprised 9775 images with NRDR (72.89%) and 3635 images with RDR (27.11%). Note that a physician may perform multiple fundus examinations for a patient on the same date, then have a referral or nonreferral recommendation for that patient in the real world. Therefore, to ensure a fair decision we checked the patient's every image taken on the same date return to a ground truth, giving the data comprised 7729 patients with NRDR (77.54%) and 2239 patients with RDR (22.46%).

Preprocessing

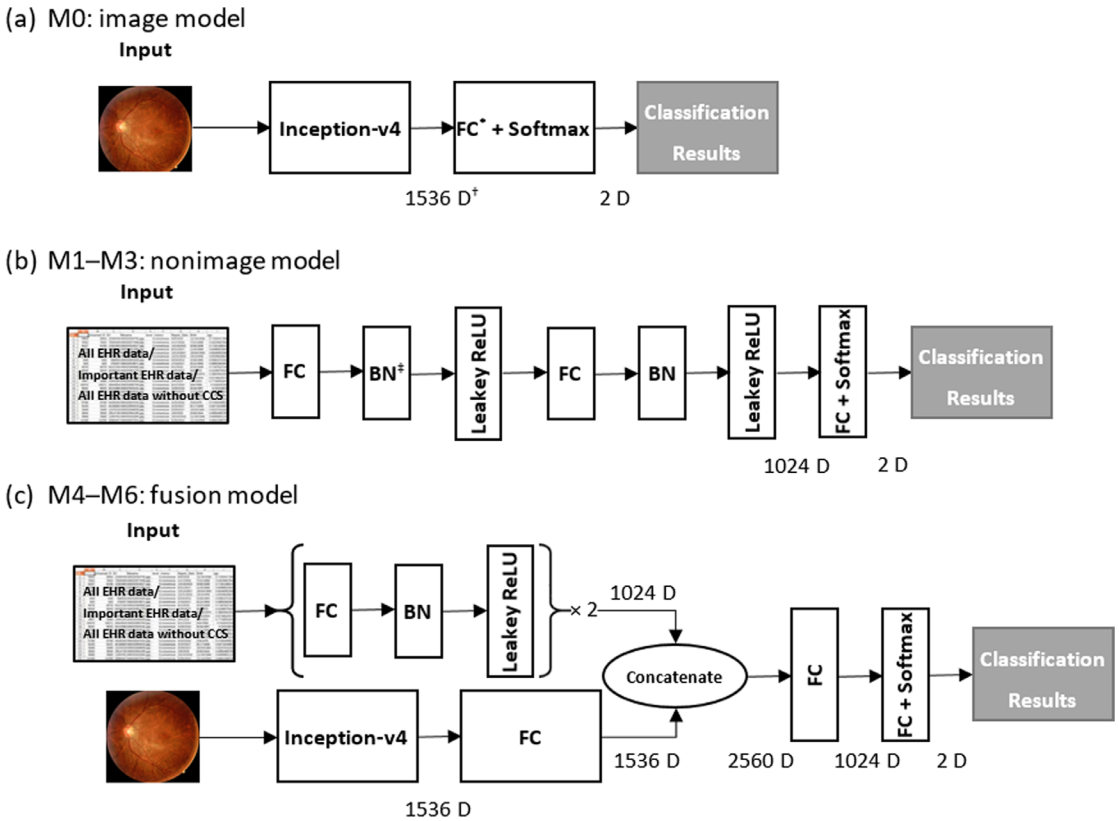
Several well-known preprocessing techniques were adopted to prepare quality data in the training process. For image processing, the circular boundary of the raw image was detected to circumscribe a square about the image, then the nonretinal background was cropped from the square boundary and the images were resized to squares of 299 pixels. For nonimage processing, improper data entry, symbols, texts, and outliers were removed, zero imputation was used to replace the missing value. Continuous data were normalized and rescaled to between 0 and 1. One-hot encoding was adapted to convert categorical data into a new form. Besides, we created the new variables for EHR data, which is raw data transformation. The difference between the imaging date and the date of examination data (laboratory test and vital signs) was mapped into a categorical format. Lab data were transformed into a normal or an abnormal group based on the criteria of the hospital.

Data Fusion

Before we moved on to the fusion architecture with heterogeneous data, we considered a logistic regression model (LR) as a baseline for training the nonimage data and compared it with the other algorithms, such as principal component analysis (PCA) with a logistic regression classifier, gradient boosting (GB), eXtreme Gradient Boosting (XGB), and MLP. Note that MLP was employed to detect RDR with CNN architecture for improving overall performance based on multimodal data in the proposed fusion model.

Widespread adoption of EHRs may result in difficulties with the practical application of the fusion model. As fundus imaging is typically used by specialists for DR severity screening, and several DR risk factors are associated with the onset and progression of DR, it is reasonable to assume that a correlation exists between fundus imaging and DR risk factors. A classifier we employed for risk factor selection was GB.[24] This is where the negative gradient of the loss function is used as a measurement of the previous round of base learners. In a new base learner, the loss in the previous round is corrected by fitting the negative gradient. In the final output, the variable influences from EHRs were further standardized and summed up to 100%. XGB was also considered for risk factor selection, however, it heavily relies on the features of the time difference in our case, and the time difference implies the physician's judgment and patient's revisit frequency. It varies depending on the hospital. We favored GB for practical use, in which the selected variables were the common features. The importance scores for EHR variables were presented in the next section. Consequently, three compositions of EHRs were used during the training and testing process to understand the influence of the compositions on model performance. First, all EHR data was applied; second, the important features were selected based on GB; third, the EHR data was used, except for the CCS information. It is possible that when one classifies for DR who ignore CCS because that information may not represent the complete physical condition of the patient.

As seen in Figure 2, we developed an image model (M0) that used a deep CNN inspired by Inception-v4[25] and three nonimage models (M1–M3)



*FC: fully connected layer; †D: Dimension; ‡BN: Batch normalization

Figure 2. The workflow of the (a) image model (M0), (b) the nonimage models (M1–M3), and (c) the fusion models (M4–M6) for DR screening.

that used MLP with all inputs from all EHR data, important EHR data, or all EHR data without CSS, respectively. Specifically, we repeated fully connected layers with batch normalization for constructing the MLP network. We also proposed several end-to-end fusion models (M4–M6) combining both image data and the different compositions of the EHRs for early DR screening (Fig. 2). Feature vectors from two heterogeneous branches were concatenated before two fully connected layers and a two-way softmax layer. Finally, this model was used to classify the patient's case as either NRDR or RDR.

Using data from Chung Shan Medical University Hospital, we trained the neural network models (M0–M6) and optimized the softmax loss function to minimize the overall loss using stochastic gradient descent. The network was designed with an input size of 278 (all EHR features) or its subset combination and a hidden size of 1024 for nonimage models. We used the images with a resolution of 299 × 299 pixels as the input for image and fusion models. The output was characterized by a 2-class DR severity grade for the tested data. The remaining hyperparameters in the fusion models were defined as follows: mini-batch of 16, L2 regularization with a factor of 0.00004, a momentum of 0.1, dropout probability of 20, and initial learning rate of 0.001, which automatically decreased by a factor of 30 in each epoch.

Data Analysis

To assess whether the EHR features in the NRDR and RDR groups were statistically different, we performed a two-sample test for equality based on functions in the R software (i.e., prop.test for categorical variables and *t*-test for continuous variables). The mean values of the numerical EHR features included the 2.5 and 97.5 percentiles, which formed the 95% confidence intervals. Bonferroni's correction[26] was used for multiple comparisons, where a P value below 0.0022 (= 0.05/23) was considered statistically significant given the 23 tested hypotheses.

In terms of model performance evaluation, we calculated the accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, and specificity for DR classification. Accuracy was defined as the proportion of images correctly classified out of the total number of evaluated images. Sensitivity was defined as the proportion of RDR images correctly detected. Specificity was defined as the proportion of NRDR images correctly detected. DeLong test[27] was used to compare various AUCs of two receiver operating characteristic curves. An assessment consensus from three ophthalmologists was considered the gold standard and was used as the reference for performance evaluation of the models. The higher predictive RDR probability of images on the same imaging date from one patient was chosen to support a referral decision for that patient.

Results

Baseline, Laboratory Tests, and Historical ICD Information in EHRs

Table 1 summarizes the characterization of diabetic patient cases as NRDR or RDR. There were 7729 and 2239 patients classified as NRDR and RDR, respectively. We analyzed the differences between the NRDR and RDR groups. Hyperglycemia determined based on ICD code was present in 88% and 72% of the patients from the NRDR and RDR groups, respectively. In contrast, the HbA1c level in the RDR group was significantly higher than that in the NRDR group. This shows that the ICD code may report the partial healthy status of the patients, and likewise, hypertension and hyperlipidemia between the two groups revealed a similar pattern versus SBP and cholesterol. We used the two-sample test for equality with Bonferroni's correction for multiple comparisons, confirming that most EHR features were statistically significant ($P < 0.0022$) between the two groups. Note that the data length differs according to the EHR parameters as they often suffer from missing values. No significant differences were observed between groups regarding BMI, DBP, HDL, LDL, TG, creatinine, and PRO 24 h.

Nonimage model M2 trained with the important EHR data (Table 2) selected based on GB. LR and M2 with important EHR produced similar AUC scores with a nonsignificant P value (= 0.1927) (Table 3). Besides, there was a statistically nonsignificant difference in AUCs for the nonimage models M1 and M2 (P value of 0.1949). This evidence shows that the sum of 80% of the importance score (Table 2) is a reasonable selection for defining the important EHR data. These important EHR data were considered as complementary information in the fusion model.

Model Performance

Table 3 shows the performance results from the baseline model (LR), machine learning models (PCA,

Table 1. Demographic and Clinical Characteristics of EHR Parameters in the Two Groups

Characteristic	NRDR	RDR	*P* Value
Number of samples	7729	2239	—
Age, mean ± SD (years)	57.40 ± 13.66	58.89 ± 12.76	0.0001
Sex, female (%)	3478 (45.00)	1109 (49.53)	0.0002
Smoking, at least once (%)	2033 (26.31)	448 (20.01)	< 0.0001
Hypertension, yes (%)	4510 (58.35)	1200 (53.60)	< 0.0001
Hyperglycemia, yes (%)	6822 (88.26)	1605 (71.69)	< 0.0001
Hyperlipidemia, yes (%)	4474 (57.89)	999 (44.62)	< 0.0001
BMI, mean ± SD	25.95 ± 4.69	25.53 ± 4.42	0.0023
Pulse, mean ± SD (bpm)	83.71 ± 12.71	85.73 ± 14.03	< 0.0001
SBP, mean ± SD (mmHg)	132.85 ± 17.52	137.56 ± 21.64	< 0.0001
DBP, mean ± SD (mmHg)	76.30 ± 11.66	77.23 ± 13.79	0.0054
HbA1c, mean ± SD (%)	7.16 ± 1.15	7.65 ± 1.26	< 0.0001
Cholesterol, mean ± SD (mg/dl)	170.18 ± 35.58	173.58 ± 37.81	0.0011
HDL, mean ± SD (mg/dl)	42.17 ± 10.11	41.67 ± 10.28	0.099
LDL, mean ± SD (mg/dl)	93.59 ± 28.34	96.04 ± 30.82	0.0039
TC/HDL, mean ± SD	4.09 ± 1.18	4.23 ± 1.24	0.0006
TG, mean ± SD (mg/dl)	132.90 ± 65.26	135.71 ± 66.70	0.1350
Creatinine, mean ± SD (mg/dl)	0.84 ± 0.24	0.86 ± 0.26	0.0071
eGFR, mean ± SD (mL/min/1.73 m^2)	91.73 ± 27.42	84.66 ± 31.84	< 0.0001
PBG, mean ± SD (mg/dl)	187.88 ± 76.62	215.37 ± 87.04	< 0.0001
FBG, mean ± SD (mg/dl)	140.86 ± 33.40	146.58 ± 41.56	< 0.0001
PRO random, mean ± SD (mg/dl)	83.88 ± 92.60	153.53 ± 134.06	< 0.0001
PRO 24 h, mean ± SD (mg/dl)	47.62 ± 75.28	96.76 ± 86.81	0.0021

SBP: systolic blood pressure; DBP: diastolic blood pressure; HDL: high density lipoprotein; LDL: low-density lipoprotein; TC/HDL: total cholesterol/high density lipoprotein; TG: triglyceride; eGFR: estimated glomerular filtration rate; PBG: postprandial blood glucose; FBG: fasting blood glucose; PRO random: random urine protein; PRO 24 h: 24-hour urine protein.

GB, and XGB), image model (M0), three non image models (M1–M3) with different EHR compositions (i.e., all EHR data, the important EHR data, or all EHR data without CCS), and three fusion models (M4–M6) with different EHR compositions for DR classification. In the evaluation metrics (accuracy, specificity, sensitivity, and AUC), AUC is used to compare overall performance with regard to these models, and the performance of the fusion models was better than that of the image model and the nonimage models. We also evaluated the influence of EHR composition on model performance and found that the AUCs for the models were 80% for LR and 78% to 81% for GB/M1/M2 (P value > 0.005), and that M3 which disregards CCS information showed the lower AUC (P value < 0.0001), which suggests that historical CCS information is critical when not using fundus image information. For the image model (M0), the values of the evaluation metrics are between those of the nonimage models and the fusion models. Since the image data included direct information on the fundus, DR determination without using image information tended to predict NRDR in nonimage models, thus resulting in lower sensitivity. Further increases in AUC values were observed when the image and non-image data were combined. Fusion models (M4–M6) showed significant results in improving the performance of M0. Similar AUCs were observed in the fusion models. Hence, the important EHR data may provide enough information and there may not be a need to use all EHR data for classification. Overall, using fusion models, which use heterogeneous data from both nonimage data and images to classify DR severity, improved overall performance (Fig. 3).

Models that were used to compare the performance included hybrid decision support systems,[21] and the best performing proposed fusion model was M5. Our model yielded results with an accuracy of 91.29%, sensitivity of 96.84%, specificity of 89.44%, and AUC of 97.96%. Overall, the proposed model performed well in terms of AUC compared with the results of the Skevofilakas et al. previous study (Table 4).

Table 2. The Sum of 80% of the Feature Importance Score

EHR Feature	Score
CCS 6.7	0.1334
CCS 3.3	0.0981
Age	0.0618
eGFR	0.0453
HbA1c	0.0451
FBG	0.0330
Creatinine	0.0299
SBP	0.0285
BMI	0.0266
PBG	0.0259
TC/HDL (category: normal)	0.0235
HDL	0.0212
DBP	0.0211
Height	0.0184
CCS 13.8	0.0179
TG	0.0178
LDL	0.0176
Pulse	0.0167
TC/HDL (category: N/A)	0.0157
FBG (time slot: <15 days)	0.0153
Cholesterol	0.0144
TC/HDL	0.0140
Weight	0.0126
CCS 1.3	0.0102
Hyperglycemia	0.0101
TC/HDL (time slot: N/A)	0.0083
Smoke (at least once)	0.0076
PRO random	0.0069
CCS 3.2	0.0065

Clinical records (CCS 1.3 [viral infection], CCS 3.2 [diabetes without complications, e.g., hyperglycemia or glycosuria], CCS 3.3 [diabetes with complications, e.g., chronic kidney disease or macular edema], CCS 6.7 [eye disorders, e.g., cataract or glaucoma], and CCS 13.8 [connective tissue disease]).

Discussion

In the current study, we proposed the combination of heterogeneous information to enhance the performance of binary DR screening. Very few previous studies have addressed data combinations, such as EHR information and fundus images, via a CNN.[12] As several medical records are associated with DR, as well as the fact that we must consider a patient's comprehensive information to increase data diversity, we devised an automated data fusion architecture that is jointly trained end-to-end and optimized through backpropagation. The resulting fusion model performed better than a model that used only fundus images. Therefore, the nonimage information could further enhance the diagnosis of RDR when there is a lack of imaging features. Comparing with nonimage model M1, M3 without CCS showed a lower AUC of 72% compared to the non-image model M1. It suggests that historical CCS information is critical when not using fundus image information. However, fusion model M6 slightly increases an AUC of 0.3% when comparing with M4; it may imply that the CCS information somehow interferes with the image information training. Overall, three fusion models with variations in the EHR composition used exhibited similar performance for all evaluation metrics. Hence, M5 with the important EHR data is enough for DR classification if the model is simultaneously trained with image data.

Several EHR datasets have been studied and the features can be expressed on the anatomical region of the fundus image, such as age, smoking, and HbA1c.[28] Rather than use the image data, the nonimage information provided a moderate contribution for DR classification. The nonimage model M1 with full EHR data yielded temperate performance with an AUC of 79%. This finding suggests that a nonimage model can reasonably estimate the risk of a diabetic patient developing RDR, leading to the prevention of referral delay if a patient cannot be photographed, which is the case for certain conditions, such as cataracts.

A previous study achieved a classification accuracy of 98% using a hybrid decision support system to detect the presence of DR; this study combined information from fundus images and an EHR.[21] However, the influence of EHR information was unclear. In the current study, we strengthened and verified the proposed fusion model for RDR detection and compared the fusion model results with those of independent image/nonimage models. Moreover, we examined the risk factors to better understand the contribution of EHR features and to achieve a more robust classification of the model's performance.

The present study had various limitations that must be addressed. Considering the correlation between two eyes from one patient as a paired input of architecture may improve the prediction performance. In particular, we found that the fusion model M5 expressed a moderate positive correlation between the predictive RDR probabilities from the left and right eye of the same patient (Spearman correlation coefficient = 0.4025). Besides, further progress is required concerning an analysis of time-dependent factors from laboratory tests and dimension reduction

Table 3. Performance Comparison for DR Classification

Model	Data	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	*P* Value
LR	All EHR	73.68 (72.82, 74.54)	71.70 (63.83, 73.57)	74.27 (73.30, 75.24)	80.01 (79.11, 80.91)	
PCA	All EHR	72.39 (71.51, 73.27)	69.79 (67.58, 71.40)	73.18 (72.19, 74.17)	80.24 (79.35, 81.13)	0.8701
GB	All EHR	83.08 (82.34, 83.82)	38.72 (36.70, 40.74)	96.52 (96.11, 96.93)	81.21 (80.34, 82.08)	0.4788
XGB	All EHR	83.18 (82.45, 83.91)	37.02 (35.02, 39.02)	97.16 (96.79, 97.53)	81.90 (81.05, 82.75)	0.0293
MLP (M1)	All EHR	79.32 (78.52, 80.12)	52.55 (50.48, 54.62)	87.43 (86.69, 88.17)	79.17 (78.25, 80.09	0.6257
MLP (M2)	Important EHR	78.67 (77.87, 79.47)	53.83 (51.77, 55.89)	86.20 (85.43, 86.97)	77.69 (76.73, 78.65)	0.1927
MLP (M3)	All EHR w/o CCS	71.80 (70.92, 72.68)	56.60 (54.55, 58.65)	76.40 (75.45, 77.35)	72.09 (71.01, 73.17)	< 0.0001
Inception-v4 (M0)	Image	88.42 (87.79, 89.05)	88.93 (87.63, 90.23)	88.25 (87.53, 88.97)	94.63 (94.22, 95.04)	
MLP + Inception-v4 (M4)	All EHR + image	91.24 (90.69, 91.79)	94.43 (93.48, 95.38)	90.18 (89.52, 90.84)	97.45 (97.18, 97.72)	< 0.0001
MLP + Inception-v4 (M5)	Important EHR + image	91.29 (90.74, 91.84)	96.84 (96.12, 97.56)	89.44 (88.75, 90.13)	97.96 (97.72, 98.20)	< 0.0001
MLP + Inception-v4 (M6)	All EHR w/o CCS + image	93.12 (92.62, 93.62)	86.79 (85.39, 88.19)	95.24 (94.77, 95.71)	97.74 (97.49, 97.99)	< 0.0001

The 95% confidence interval is listed in parentheses. Bonferroni's correction was used for multiple comparisons, where a *P* value of AUC difference for the baseline model (LR) and the other model below 0.005 (= 0.05/10) was considered statistically significant given the 10 tested hypotheses (including the DeLong test to compare AUCs for M1 and M2 with a corresponding *P* value of 0.1949).

Table 4. Assessment of Benchmark Comparison

Benchmark	Data Type	Dataset	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
Skevofilakas et al.[21]	Image and EHR data	55 diabetic patients	Hybrid decision support system	98	100	98	—
Proposed model M5	Image and EHR data	6566 diabetic patients	MLP+CNN	91.29	96.84	89.44	97.96

of the ICD codes. Patients may not have a complete set of tests from every visit to the physician, such that properly collecting and handling medical histories over time becomes very important. For example, a series of blood sugar control levels should be considered for DR research, including HbA1c, FBG, and PBG. However, instead of using longitudinal EHRs with a learned medical feature embedding matrix,[29] we only used the most recent record of laboratory results from the EHR information in the fusion model. From a clinical practice perspective, it is convenient if only a few EHR features are required for data fusion with the image model to enhance DR screening, such that this simplification can be an area of future research. On the other hand, we did not collect patient visual acuity, which represents their functional status. In this study, we only collected systemic parameters and gross anatomical changes (fundus image) for the fusion model. Further studies considering long-term follow-up from the early RDR and the effects of receiving prompt treatment may increase model usefulness.

Overall, the proposed neural network architecture for fusing images and related EHR information enables early diagnosis of RDR. The proposed fusion model can enhance classification performance and support a diagnosis by specialists. The model supports the combination of complementary image and nonimage data.

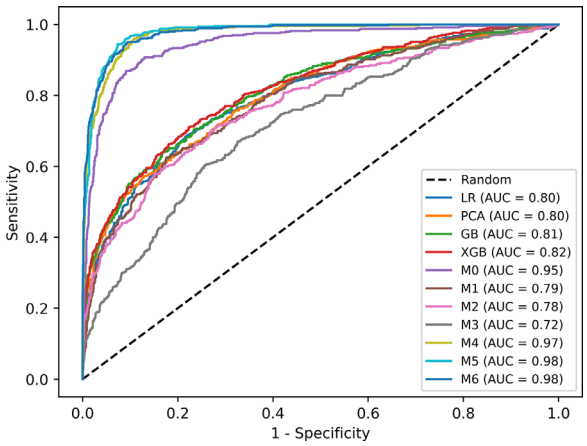


Figure 3. The relationship among the area under the receiver operating characteristic curves.

Acknowledgments

The authors thank the management staff at Chung Shan Medical University Hospital, Taichung, Taiwan for providing the data used in this study.

Supported by the Industrial Technology Research Institute (grant number: J367B82210). The funding agency participated in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Disclosure: **M.-Y. Hsu,** None; **J.-Y. Chiou,** None; **J.-T. Liu,** None; **C.-M. Lee,** None; **Y.-W. Lee,**

None; **C.-C. Chou,** None; **S.-C. Lo,** None; **E. Kornelius,** None; **Y.S. Yang,** None; **S.-Y. Chang,** None; **Y.-C. Liu,** None; **C.-N. Huang,** None; **V.S. Tseng,** None

References

1. Saeedi P, Petersohn I, Salpea P, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the international diabetes federation diabetes atlas, 9th edition. *Diabetes Res Clin Pract*. 2019;157:107843, doi:10.1016/j.diabres.2019.107843.
2. Sheen YJ, Hsu CC, YDer Jiang, Huang CN, Liu JS, Sheu WHH. Trends in prevalence and incidence of diabetes mellitus from 2005 to 2014 in Taiwan. *J Formos Med Assoc*. 2019;118(1650):S66–S73, doi:10.1016/j.jfma.2019.06.016.
3. Lin JC, Shau WY, Lai MS. Sex- and age-specific prevalence and incidence rates of sight-threatening diabetic retinopathy in Taiwan. *JAMA Ophthalmol*. 2014;132(8):922–928, doi:10.1001/jamaophthalmol.2014.859.
4. Flaxman SR, Bourne RRA, Resnikoff S, et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *Lancet Glob Health*. 2017;5(12):e1221–e1234, doi:10.1016/S2214-109X(17)30393-5.
5. Liao WL, Lin JM, Chen WL, et al. Multilocus genetic risk score for diabetic retinopathy in the Han Chinese population of Taiwan. *Sci Rep*. 2018;8(1):1–9, doi:10.1038/s41598-018-32916-y.
6. Quellec G, Charrière K, Boudi Y, Cochener B, Lamard M. Deep image mining for diabetic retinopathy screening. *Med Image Anal*. 2017;39:178–193, doi:10.1016/j.media.2017.04.012.
7. Abbas Q, Fondon I, Sarmiento A, Jiménez S, Alemany P. Automatic recognition of severity level for diagnosis of diabetic retinopathy using deep visual features. *Med Biol Eng Comput*. 2017;55(11):1959–1974, doi:10.1007/s11517-017-1638-6.
8. Hsu WM. Current status of ophthalmology in Taiwan. *J Clin Exp Ophthalmol*. 2015;6(5):5–7, doi:10.4172/2155-9570.1000485.
9. Abràmoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57(13):5200–5206, doi:10.1167/iovs.16-19964.
10. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410, doi:10.1001/jama.2016.17216.
11. Voets M, Møllersen K, Bongo LA. Reproduction study using public data of: development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLoS One*. 2019;14(6):1–11, doi:10.1371/journal.pone.0217541.
12. Lin WC, Chen JS, Chiang MF, Hribar MR. Applications of artificial intelligence to electronic health record data in ophthalmology. *Transl Vis Sci Technol*. 2020;9(2):13, doi:10.1167/tvst.9.2.13.
13. Cheng YT, Lin YF, Chiang KH, Tseng VS. Mining sequential risk patterns from large-scale clinical databases for early assessment of chronic diseases: a case study on chronic obstructive pulmonary disease. *IEEE J Biomed Health Inform*. 2017;21(2):303–311, doi:10.1109/JBHI.2017.2657802.
14. Tseng ST, Chou ST, Low BH, Su FL. Risk factors associated with diabetic retinopathy onset and progression in diabetes patients: a Taiwanese cohort study. *Int J Clin Exp Med*. 2015;8(11):21507–21515.
15. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye Vis*. 2015;2(1):17–42, doi:10.1186/s40662-015-0026-2.
16. Saleh E, Błaszczyński J, Moreno A, et al. Learning ensemble classifiers for diabetic retinopathy assessment. *Artif Intell Med*. 2018;85:50–63, doi:10.1016/j.artmed.2017.09.006.
17. Oh E, Yoo TK, Park EC. Diabetic retinopathy risk prediction for fundus examination using sparse learning: a cross-sectional study. *BMC Med Inform Decis Mak*. 2013;13(1):106, doi:10.1186/1472-6947-13-106.
18. American Diabetes Association. Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. *Diabetes Care*. 2018;41(Supplement 1):S13–S27, doi:10.2337/dc18-S002.
19. Xu T, Zhang H, Huang X, Zhang S, Metaxas DN. Multimodal deep learning for cervical dysplasia diagnosis. In: Joskowicz L, Unal G, Wells W, eds. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*. Vol. 9901. 2nd ed. Cham: Springer; 2016:115–123, doi:10.1007/978-3-319-46723-8_14.
20. Wang X, Peng Y, Lu L, Lu Z, Summers RM. TieNet: text-image embedding network for common thorax disease classification and reporting in

chest X-Rays. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2018:9049–9058, doi:10.1109/CVPR.2018.00943.
21. Skevofilakas M, Zarkogianni K, Karamanos BG, Nikita KS. A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus. *2010 Annu Int Conf IEEE Eng Med Biol Soc EMBC'10*. 2010:6713–6716, doi:10.1109/IEMBS.2010.5626245.
22. AAO. International clinical diabetic retinopathy disease severity scale, https://docplayer.net/20784048-International-clinical-diabetic-retinopathy-disease-severity-scale-detailed-table.html. Published 2002.
23. Healthcare Cost and Utilization Project. Beta chronic condition indicator (CCI) for ICD-10-CM, https://www.hcup-us.ahrq.gov/toolssoftware/chronic_icd10/chronic_icd10.jsp. Published 2019. Accessed May 3, 2019.
24. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367–378, doi:10.1016/S0167-9473(01)00065-2.
25. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: *arXiv preprint arXiv:1602.07261*, 2016, https://arxiv.org/abs/1602.07261.
26. Nahler G, Nahler G. Bonferroni correction. In: *Dictionary of Pharmaceutical Medicine*. Vienna: Springer; 2009, doi:10.1007/978-3-211-89836-9_140.
27. Delong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845.
28. Poplin R, Varadarajan AV., Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158–164, doi:10.1038/s41551-018-0195-0.
29. Che Z, Cheng Y, Sun Z, Liu Y. Exploiting convolutional neural network for risk prediction with medical feature embedding. In: *arXiv preprint arXiv:1701.07474*, 2017, https://arxiv.org/abs/1701.07474.