Original Article

HEALTH SCIENCES
MEDICINE

DOI: 10.32322/jhsm.1502050
J Health Sci Med. 2024;7(4):467-471

Prediction of retinopathy through machine learning in diabetes mellitus

Tarık Keçeli, Nevruz İlhanlı, Kemal Hakan Gülkesen

Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University, Antalya, Turkiye

Cite this article as: Keçeli T, İlhanlı N, Gülkesen KH. Prediction of retinopathy through machine learning in diabetes mellitus. *J Health Sci Med.* 2024;7(4):467-471.

Received: 27.06.2024 ◆ **Accepted**: 22.07.2024 ◆ **Published**: 30.07.2024

ABSTRACT

Aims: Development of a machine learning model on an electronic health record (EHR) dataset for predicting retinopathy in people with diabetes mellitus (DM), analysis of its explainability.

Methods: A public dataset based on EHR records of patients diagnosed with DM located in İstanbul, Turkiye (n=77724) was used. The categorical variable indicating a retinopathy-positive diagnosis was chosen as the target variable. Variables were preprocessed and split into training and test sets with the same ratio of class distribution for model training and evaluation respectively. Four machine learning models were developed for comparison: logistic regression, decision tree, random forest and eXtreme Gradient Boosting (XGBoost). Each algorithm's optimal hyperparameters were obtained using randomized search cross validation with 10-folds followed by the training of the models based on the results. The receiver operating characteristic (ROC) area under curve (AUC) score was used as the primary evaluation metric. SHapley Additive exPlanations (SHAP) analysis was done to provide explainability of the trained models.

Results: The XGBoost model showed the best results on retinopathy classification on the test set with a low amount of overfitting (AUC: 0.813, 95% CI: 0.808-0.819). 15 variables that had the highest impact on the prediction were obtained for explainability, which include eye-ear drugs, other eye diseases, Disorders of refraction, Insulin aspart and hemoglobin A1c (HbA1c).

Conclusion: Early detection of retinopathy based on EHR data can be successfully detected in people with diabetes using machine learning. Our study reports that the XGBoost algorithm performed best in this research, with the presence of other eye diseases, insulin dependence and high HbA1c being observed as important predictors of retinopathy.

Keywords: Diabetic retinopathy, diabetes mellitus, machine learning, electronic health records

INTRODUCTION

Diabetes mellitus (DM) is a noncommunicable disease that is caused by the insufficient production of the insulin hormone within the pancreas or the inability of the human body to effectively use the produced insulin. Categorized under different stages to indicate its severity, diabetes is known to be a devastating disease that may take many years to be noticed and contributes to significant health problems to a person; such as vision impairment, kidney failure and stroke.[1] Along with the severity of the disease, the prevalence of diabetes is also expected to be increased, with the number of people with diabetes aged 20-79 years predicted to be increased to 642 million by 2040.[2]

Under the complications caused by diabetes, diabetic retinopathy is a major example. This medical condition is defined by the presence of retinal hemorrhages, microaneurysms, cotton wool spots and/or prior photocoagulation scars.[3] In terms of blindness, it has been reported that retinopathy was observed globally in approximately 3 million cases; and compared to under-corrected refractive error, cataract, age-related macular degeneration and glaucoma; retinopathy was the smallest contributor to blindness in 2020, but also the only cause of blindness that showed a global increase in age-standardized prevalence.[4] Retinopathy is one of the most common complications in type-1 diabetes and it has been reported that after 15-20 years since an individual's diagnosis, almost all patients would have some degree of retinopathy.[5]

To minimize the risk of vision impairment and blindness; preventative measures such as early detection by screening, effective management and compliance to guidelines is suggested.[6] It has been reported that patients who receive consistent care, have lower rates of low vision and blindness.[7] It is also important to note that the management of retinopathy, especially vision threatening variants require the expertise and skills of trained ophthalmologists or retinal specialists[8] and with reported low density values of ophthalmologists in many countries, access to the treatment of vision threatening diabetic retinopathy may be difficult based on the patient's location.[9]

Corresponding Author: Tarık Keçeli, tarikkeceli141@gmail.com

Several studies have been conducted to predict retinopathy based on electronic health records (EHR). Liu et al.[10] utilized an extreme learning machine approach on a dataset containing EHR data of 1093 patients and reported a classification area-under-curve (AUC) performance metric of 88.34%. Ogunyemi et al.[11] used EHR dataset comprising 40631 people with diabetes to predict retinopathy by training five machine learning algorithms and presented the univariate analyses of dataset variables. Their results showed that their deep learning model performed best and was able to achieve an AUC score of 0.8 on their external validation set. They also presented the most significant predictors they've observed which included insulin dependence, blood urea nitrogen and systolic blood pressure. Saleh et al.[12] used ensemble classification techniques based on uncertainty models using an EHR dataset of 2323 people with diabetes Their fuzzy random forests approach obtained an accuracy of 84% while their dominance-based rough set balanced rule ensemble approach showed an accuracy of 77%.

The aim of this study is to develop a machine learning model predicting retinopathy on a diabetes dataset, and examine the model for understanding the variables predicting retinopathy development, and discuss the possibility of developing an early-diagnosis tool for retinopathy in people with diabetes.

METHODS

For the development of the machine learning model, a public dataset containing 107 variables, composed of electronic health records of 77724 patients diagnosed with diabetes mellitus in 2017, located in İstanbul Province, Turkiye was used.[13] The dataset was originally created for the purpose of glycemic control prediction in diabetes mellitus patients, and it also includes information of retinopathy diagnosis. The retrospective analysis on the dataset, model development, SHapley Additive exPlanations (SHAP) analysis and visualizations were done using the Pandas[14,15] (v. 2.0.2), scikit-learn[16] (v. 1.3.0), shap[17] (v. 0.44.1), and matplotlib[18] (v. 3.7.1) modules in the Python (v. 3.11.3) programming language respectively. This study did not require ethics committee approval, as the data was sourced from a public dataset. All procedures were carried out in accordance with the ethical rules and the principles of the Declaration of Helsinki.

The dataset contained no missing values (106 variables in total). A random seed value of 4564 was used in the development runs for the reproducibility of results. The target variable used in this study was the "retinopathy" variable of the dataset. The train (n: 54406) and test (n: 23318) sets were created with a 70/30 split with stratification on the target variable. Both sets had a target class distribution of 13,9%, with the training set containing 7610 entries of retinopathy-positive patients and the test set containing 3262 retinopathy-positive entries. Afterwards, the numerical variables in the train and test sets were standardized in the preprocessing step. Categorical variables remained intact.

As candidate algorithms for baseline evaluation of prediction performance; logistic regression (LR), decision tree (DT), random forest (RF) and eXtreme gradient boosting (XGBoost) algorithms were selected. Before training the algorithms (except LR, which was chosen as a baseline algorithm for performance comparison), each algorithm's hyperparameters were optimized by using randomized search cross validation with 10-folds and 50 iterations in the train set (n: 54406). During this phase, a random combination of selected hyperparameters for the algorithm was produced and the hyperparameters were changed in each iteration by the help of the algorithm. For each hyperparameter combination, the mean of AUCs of 10-fold cross-validation was calculated. The best performing hyperparameters were selected based on their receiver operating characteristic (ROC) AUC score, and used for the training of the final model on the complete training set (n: 54406) for each candidate algorithm. The ROC AUC score was chosen as the primary evaluation metric of the models; along with f-score, Matthew's correlation coefficient (MCC), and precision-recall AUC (PR-AUC) recorded for the reporting of the results on the train and test sets. For the explainability of a model's prediction, SHAP values were calculated.

RESULTS

Obtained results are presented in Table. The results showed that the XGBoost model achieved the highest predictive performance with an AUC score of 0.813 (95% CI: 0.808-0.819) on the training set and 0.799 on the test set, followed by the RF model which achieved an AUC score of 0.784 on

Table. Model performance metrics on the training and test sets

	Training set (n= 54406)							Test set (n=23318)						
	AUC	PR AUC	MCC	Accuracy	F-score	Sensitivity	Specificity	AUC	PR AUC	MCC	Accuracy	F-score	Sensitivity	Specificity
XGBoost	0.813 (0.808-0.819)	0.461	0.34	0.74	0.44	0.72	0.74	0.799 (0.791-0.807)	0.420	0.32	0.68	0.41	0.78	0.66
RF	0.784 (0.779-0.790)	0.381	0.31	0.72	0.41	0.70	0.72	0.783 (0.776-0.792)	0.377	0.31	0.72	0.41	0.70	0.72
DT	0.753 (0.747-0.759)	0.361	0.26	0.63	0.36	0.76	0.61	0.749 (0.740-0.758)	0.353	0.26	0.63	0.36	0.75	0.62
LR	0.779 (0.774-0.785)	0.372	0.31	0.71	0.41	0.70	0.71	0.781 (0.772-0.789)	0.376	0.30	0.70	0.40	0.72	0.70

AUC: Area under curve, PR-AUC: Precision recall-area under curve, MCC: Matthews correlation coefficient, XGBoost: eXtreme gradient boosting, RF: Random forest, DT: Decision tree, LR: Logistic regression

the training set and 0.783 on the test set. According to the obtained statistics, the developed XGBoost model was reported as the superior model. The ROC AUC and PR AUC plots of the models are presented in Figure 1, 2 respectively.

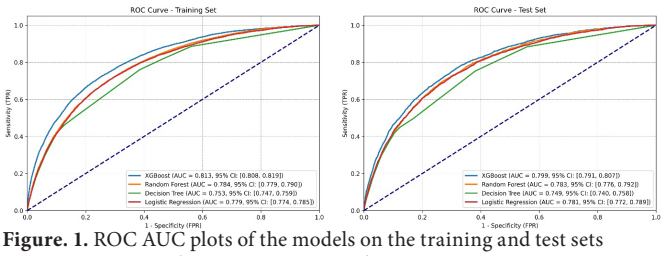

Figure. 1. ROC AUC plots of the models on the training and test sets
ROC: Receiver operating characteristic AUC: Area under curve

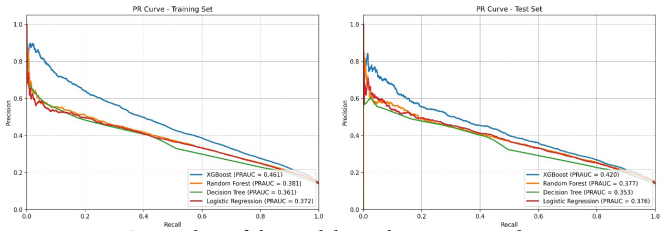

Figure. 2. PR Curve plots of the models on the training and test sets
PR: Precision recall

Model Interpretation

In terms of model interpretation, 15 variables that had the most impact in calculating the prediction were obtained using a SHAP analysis. The analysis results showed that the variable of "eye-ear drugs", which contained information on whether the patient takes eye and/or ear drugs had the highest impact on the prediction of retinopathy, followed by the variables "other eye diseases"; which described whether the patient had other eye-related diagnosis, and "disorders of refraction" (Figure 3).

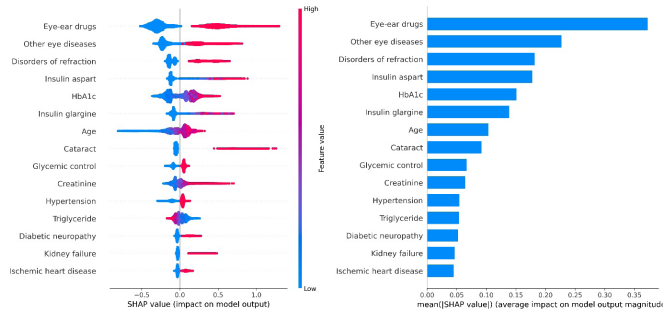

Figure 3. SHAP summary and feature importance plots
SHAP: SHapley Additive exPlanations

DISCUSSION

With this study, an analysis of retinopathy prediction based on EHR data of patients using statistics and machine learning approaches was done. Our findings show that for the task of predicting retinopathy in patients with diabetes, an XGBoost model can achieve a notable predictive performance of 0.799, as observed in our test set. With a considerable predictive performance, the developed model shows that early retinopathy detection with data from electronic health records is a feasible approach for early diagnosis. Additionally, SHAP analysis shows that most important predictors of diabetic retinopathy are presence of other eye diseases, insulin dependence and a high level of Hemoglobin A1c (HbA1c).

The performance of our XGBoost model is similar to previous studies. On the other hand, it has been reported that deep learning models trained on retinal fundus images for retinopathy detection has been successfully developed with good prediction metrics.[19] However, a predictive model based on EHR data would be useful, as obtaining fundus images of people is not always feasible. A reliable prediction of a patient's potential retinopathy diagnosis based on their electronic health records without the need of medical image analysis may provide this feasibility.

The variables that had the biggest impact on the prediction was observed to be, in descending order; "eye ear drugs", "other eye diseases", "disorders of refraction", "insulin aspart", "HbA1c", "insulin glargine", "age", "cataract", "glycemic control", "creatinine", "hypertension", "triglyceride", "diabetic neuropathy", "kidney failure", and "ischemic heart disease". The summary plot showed that for every variable, except "triglyceride", an increase or occurrence was directly proportional to a retinopathy-positive prediction. These variables and results can be categorized under the following sections for further discussion:

Eye-related Complaints Indicate a Higher Risk of Retinopathy

The variables under this observation are "eye ear drugs", "other eye diseases", "disorders of refraction", and "cataract". The usage of eye and ear drugs may be interpreted as an indicator of patients being treated for eye diseases, as can be seen the frequent diagnosis of "other eye diseases". All these predictors show that DM may cause various problems in the eye, showing a simultaneous increase of incidence in eye diseases. Conditions such as glaucoma, age-related macular degeneration, and diabetic macular edema often coincide with retinopathy, leading to damaging effects that accelerate disease progression and impair visual function. The development of cataracts has been shown to have a proportional impact on the risk of retinopathy. Cataracts have an impact visual perception but may also induce inflammatory responses and oxidative stress within the eye, further increasing retinal damage.[20]

Indicators Showing that Diabetes has Progressed

The variables under this observation are "insulin aspart", "HbA1c", "insulin glargine", "glycemic control", "creatinine", "diabetic neuropathy", "kidney failure", and "ischemic heart disease". Medications such as insulin aspart and insulin glargine are commonly used in the management of diabetes while also serve as indicators for disease severity and insulin resistance. Elevated levels of these observations often correlate with advanced stages of diabetes. A more severe condition of diabetes is likely to be predictive of retinopathy in this instance. HbA1c, a widely utilized measure of long-term glucose control, offers insights into the overall management of diabetes and its impact on retinopathy progression. Elevated HbA1c levels may indicate suboptimal glycemic control over an extended period, thus inclining individuals to microvascular complications, including retinopathy.

Glycemic control variable in our study shows several high serum HbA1c measurements in a person.

Additionally, renal function, as reflected by markers like creatinine, serves as an indicator of systemic diabetic complications, including nephropathy and retinopathy. The relationship between kidney function and retinal health underlines the importance of comprehensive diabetic care and regular screening protocols. Diabetic neuropathy, characterized by peripheral nerve damage secondary to chronic hyper-glycemia, poses a significant risk factor for retinopathy progression. The neurovascular axis plays a critical role in maintaining retinal homeostasis, and disruptions in peripheral nerve function can worsen retinal ischemia and neurodegeneration. Furthermore, the onset of kidney failure and ischemic heart disease indicates a systemic decline in vascular health, worsening the microvascular complications associated with retinopathy. These comorbidities strengthen the multifactorial aspect of retinal disease progression and emphasize the need of a complete approach to diabetes management.

Risk Factors Based on the Patient

Advancing age is an important risk factor for retinopathy development and progression. Age-related structural and functional changes within the retina contribute to increased vulnerability to retinal pathologies. Moreover, aging-related alterations in vascular integrity and neurotrophic support mechanisms predispose older individuals to microvascular dysfunction and retinal ischemia, strengthening the progression of retinopathy.

Hypertension, characterized by persistently elevated blood pressure levels, exerts profound effects on retinal microvasculature and vascular autoregulation. Chronic hypertension induces arteriolar remodeling, endothelial dysfunction, and increased vascular permeability, culminating in retinal vascular abnormalities and exacerbation of retinopathy. The coexistence of hypertension and diabetes further potentiates retinal microvascular damage, underscoring the importance of stringent blood pressure control in retinopathy management.

Limitations

It is also important to consider the limitations of this study. The used data was originally collected from electronic health records of patients in İstanbul, Turkiye. The dataset belongs to a population with certain genetic and environmental conditions and may not be universal. External validation would show the universal value of obtained models. In addition, the dataset does not include some important data such as body mass index, and physical activity. A model with more variables would have possibly a higher performance. Another problem is related to diabetic retinopathy diagnosis in this dataset. The dataset included EHR data of already diagnosed patients, and may not include silent cases or cases in their early stages.

In future studies, for additional validation, a potentially better performance and a more generalized model, a larger dataset containing more diverse patient data and additional variables can be considered. Machine learning techniques improved and continue to improve in a fast pace. New machine learning techniques may be more successful in the future.

CONCLUSION

In conclusion, presence of retinopathy can be successfully detected in people with diabetes. The best model for this purpose seems XGBoost. Other eye diseases, insulin dependence and high HbA1c are important predictors of retinopathy.

ETHICAL DECLARATIONS

Ethics Committee Approval

Due to the use of a public dataset, written patient consent was not taken. This study did not require ethics committee approval, as the data was sourced from a public dataset.

Informed Consent

Because the study was designed retrospectively, no written informed consent form was obtained from patients.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

Acknowledgements

This study has been presented orally on 30.05.2024 in the 15th Turkish Congress of Medical Informatics held in Karadeniz Technical University, Trabzon, Turkiye.

REFERENCES

1. World Health Organization (WHO). Diabetes. World Health Organization. Published May 4, 2023. Accessed February 29, 2024. https://www.who.int/news-room/fact-sheets/detail/diabetes
2. Ogurtsova K, Da Rocha Fernandes JD, Huang Y, et al. IDF diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Res Clin Pract*. 2017;128:40-50.
3. Early treatment diabetic retinopathy study research group. Grading diabetic retinopathy from stereoscopic color fundus photographs-an extension of the modified airlie house classification. ETDRS report number 10. Early treatment diabetic retinopathy study research group. *Ophthalmology*. 1991;98(5 Suppl):786-806.
4. Steinmetz JD, Bourne RRA, Briant PS, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the right to Sight: an analysis for the global burden of disease study. *Lancet Glob Health*. 2021;9(2):144-160.

5. Aiello LP, Gardner TW, King GL, et al. Diabetic retinopathy. *Diabetes Care*. 1998;21(1):143-156.

6. Wong TY, Sabanayagam C. Strategies to tackle the global burden of diabetic retinopathy: from epidemiology to artificial intelligence. *Ophthalmologica*. 2020;243(1):9-20.

7. Sloan FA, Grossman DS, Lee PP. Effects of receipt of guideline-recommended care on onset of diabetic retinopathy and its progression. *Ophthalmology*. 2009;116(8):1515-1521.

8. The diabetic retinopathy study research group. Indications for photocoagulation treatment of diabetic retinopathy: diabetic retinopathy study report no. 14. *Int Ophthalmol Clin*. 1987;27(4): 239-253.

9. Teo ZL, Tham YC, Yu M, Cheng CY, Wong TY, Sabanayagam C. Do we have enough ophthalmologists to manage vision-threatening diabetic retinopathy? A global perspective. *Eye (Lond)*. 2020;34(7):1255-1261.

10. Liu L, Wang M, Li G, Wang Q. Construction of predictive model for type 2 diabetic retinopathy based on extreme learning machine. *Diabetes Metab Syndr Obes*. 2022;15:2607-2617.

11. Ogunyemi OI, Gandhi M, Lee M, et al. Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system. *JAMIA Open*. 2021;4(3):1-10.

12. Saleh E, Błaszczyński J, Moreno A, et al. Learning ensemble classifiers for diabetic retinopathy assessment. *Artif Intell Med*. 2018;85:50-63.

13. Gülkesen KH, Ülgü MM, Mutlu B, et al. Machine learning for prediction of glycemic control in diabetes mellitus. *Mendeley Data*; 2022. doi: 10.17632/rr4rzzrjfc.2

14. The pandas development team. pandas-dev/pandas: pandas (v2.0.2). Zenodo; 2023. doi: 10.5281/zenodo.7979740

15. McKinney W. Data structures for statistical computing in python. In: Van Der Walt S, Millman J, eds. *Proceedings of the 9th* Python in Science Conference; 2010:56-61.

16. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12(85):2825-2830.

17. Lundberg SM, Lee SI. A Unified approach to interpreting model predictions. In: Guyon I, Luxburg U Von, Bengio S, et al., eds. Advances in neural information processing systems 30. Curran Associates, Inc.;2017.

18. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90-95.

19. 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.

20. Alabdulwahhab KM. Senile cataract in patients with diabetes with and without diabetic retinopathy: a community-based comparative study. *J Epidemiol Glob Health*. 2022;12(1):56-63.