JOURNAL OF
EMERGING INVESTIGATORS

Article

Machine learning for retinopathy prediction: Unveiling the importance of age and HbA1c with XGBoost

Ramana Ramachandran[1], **Rajesh H. Honnutagi**[2]
[1]Mallya Aditi International School, Bangalore, Karnataka, India
[2]Professor of Medicine, Shri B. M. Patil Medical College, Hospital and Research Center, Vijayapura, Karnataka, India

SUMMARY

Retinopathy is a major microvascular complication of both diabetes and hypertension. It can lead to blindness if left untreated. As one in three adults in India is estimated to be either diabetic or hypertensive, their future risk of developing retinopathy is high. The purpose of our study was to examine the correlation of glycosylated hemoglobin (HbA1c), blood pressure (BP) readings, and lipid levels with retinopathy. Our main hypothesis was that poor glycemic control, as evident by high HbA1c levels, high blood pressure, and abnormal lipid levels, causes an increased risk of retinopathy. We screened 119 Indian patients using a Fundus camera for retinopathy and obtained their recent BP, HbA1c, and lipid profile values. We then applied the XGBoost machine learning algorithm to predict the presence or absence of retinopathy from their lab values. We were able to predict retinopathy with high accuracy from these key biomarkers. Further, using Shapely Additive Explanations (SHAP), we identified the top two features that were most important to the model as age and HbA1c. This indicates that older patients with poor glycemic control are more likely to show presence of retinopathy. Hence, these high-risk individuals can be targeted for early screening and intervention programs to prevent the progression of retinopathy to blindness.

INTRODUCTION

As of 2021, approximately 537 million people worldwide are living with diabetes (1). In India, there are approximately 75 million people coping with diabetes (2). Diabetes is a multi-faceted illness, causing many complications when left untreated. Diabetic patients can present with a wide range of symptoms such as fatigue, weight loss, and increased appetite. Hypertension is another chronic illness, which is a major health crisis in India. Hypertensive patients commonly present with headaches, neck pain, and breathlessness with activity. Almost 30% of adults in India are estimated to be hypertensive (3). Hyperlipidemia is another common condition that coexists in both hypertensive and diabetic patients. Hyperlipidemic patients are usually asymptomatic.

Retinopathy is a major microvascular complication of diabetes and hypertension, which can lead to blindness if left untreated. It is a complex disease with multiple risk factors, including high blood pressure (BP), poor glycemic control, and abnormal lipid levels. Studies suggest a link between retinopathy and poorly controlled hypertension, diabetes, and hyperlipidemia (4-6). Several studies suggest a modest association between abnormal lipid levels and the severity of retinopathy (7, 8). However, the relationship between abnormal lipid level and the development of retinopathy is poorly understood. Thus, understanding the correlation of these risk factors with retinopathy is critical for managing and treating this disease. Furthermore, the screening process for estimating the risk of retinopathy is not sufficient or effective today as the disease has often progressed to blindness in many patients before retinopathy is detected (9). Therefore, there is an urgent need for quick and early identification of high-risk individuals who are likely to have retinopathy so that controlling their hypertension and diabetes can help arrest the progression of retinopathy to blindness.

Typically, studies that seek to find the retinopathy risk factors use conventional screening techniques such as a dilated fundus examination, which requires on-site ophthalmologists and cumbersome equipment. The Simple, Mobile-based Artificial Intelligence Algorithm in the detection of Diabetic Retinopathy (SMART) India study was a large cross-sectional screening study for diabetic retinopathy that did retinal screening without dilation and estimated the prevalence of diabetic retinopathy in India (10). Thus, it appears that simpler screening techniques such as the use of a Fundus camera without dilation are sufficient for detecting retinopathy.

The main objective of this study was to examine the correlation of BP control, glycosylated hemoglobin (HbA1C), and lipid levels with the severity of retinopathy in the urban Indian population. By doing so, we hoped to identify high risk patients who can be quickly screened using a Fundus camera without dilation so that they can be promptly treated. We collected key biomarkers from 119 Indian patients and used the biomarkers to predict retinopathy using the XGBoost machine learning algorithm (11). We achieved an overall accuracy of 93% as measured by the Area Under the Curve (AUC) metric. We further identified that age and HbA1c were primary determinants of retinopathy by using Shapely Additive Explanations (SHAP).

In summary, our findings suggest that older patients with poorly controlled blood sugar levels are more likely to exhibit retinopathy. Consequently, by targeting these high-risk individuals, it is possible to implement early screening and intervention programs to prevent the progression of retinopathy to blindness.

RESULTS

In late June 2023, we screened 119 Indian patients living in urban environments for retinopathy using a Fundus camera. The Fundus camera captured images of the interior surface of the eye, including the retina, retinal vasculature, and optic

https://doi.org/10.59720/23-184

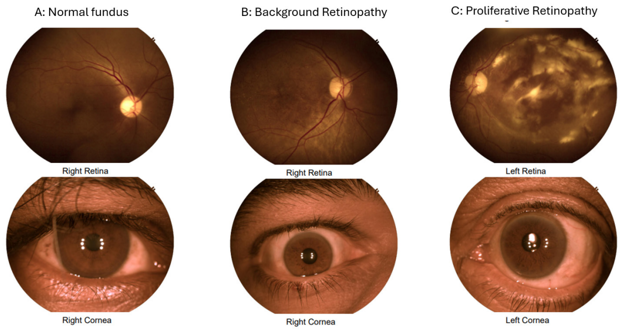

Figure 1: Fundus camera captures retinopathy in patients. Sample photographs of the retina of three patients captured using a Fundus camera showing (A) a normal fundus, **(B) b**ackground retinopathy, and **(C)** proliferative retinopathy.

disk (**Figure 1**). A licensed ophthalmologist analyzed the figures and made the diagnosis of presence or absence of retinopathy. We also obtained the patients' recent BP, HbA1c, and lipid profile values. Blood pressure consists of two readings, systolic and diastolic, with ideal levels being 120 mm Hg and 80 mm Hg, respectively. An HbA1C test measures the average blood glucose level over the past three months and is an indicator of how well diabetes is controlled, with levels below 6.5% showing ideal control. Lipid measurement consists of Low-Density Lipoprotein (LDL) and High-Density Lipoprotein (HDL) levels, with optimal levels being below 100 mg/dl and above 40 mg/dl, respectively. We wanted to see if patients with poor control of these biomarkers exhibited a higher likelihood of retinopathy. Along with the patients' age and gender, we created an anonymized dataset for analysis (**Table 1**). Note that the dataset is imbalanced with approximately four negative classes for each positive retinopathy class. Such imbalances are typical of medical datasets since positive cases are often rarer than negative cases, but this can be handled via careful choice of hyper-parameters of the machine learning algorithms.

We applied the XGBoost machine learning algorithm to make predictions about the presence or absence of retinopathy using the above dataset (11). We chose XGBoost as our machine learning algorithm because of its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression on tabular datasets (11). We compared the model's prediction accuracy with the diagnosis made by a licensed ophthalmologist as ground truth. We used the AUC because it is widely used to quantify the overall performance of a binary classification model, especially in imbalanced datasets.

Gender	Age	HbA1c	TC	TG	HDL	LDL	SBP	DBP	Retinopathy Positive
Male	51	5.9	145	76	46	84	140	80	0
Male	66	7.9	119	202	25	54	130	80	1
Male	63	12.4	212	35	26	115	140	80	0
Female	42	9.4	146	210	29	75	150	80	0
Male	45	5.5	151	113	28	101	140	80	0

Table 1: Five patient sample dataset. The features are gender, age, HbA1c, Total Cholesterol (TC), Triglycerides (TG), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Systolic Blood Pressure (SBP), and Diastolic Blood Pressure (DBP). These nine features are used to predict the last column, which indicates the presence (1) or absence (0) of Retinopathy.

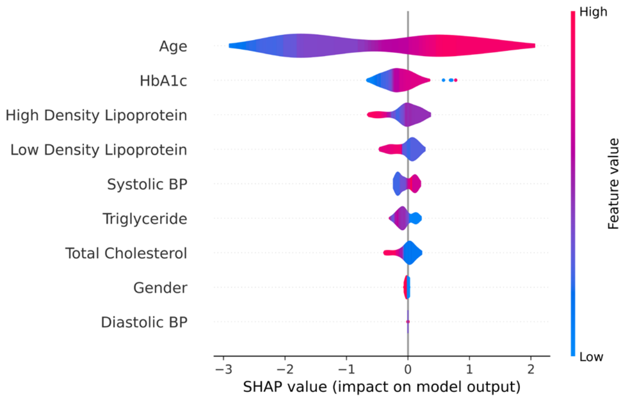

Figure 2: SHAP summary violin plot for the XGBoost model shows Age and HbA1c are the most important features for predicting retinopathy. The color represents the value of the feature from low (blue) to high (red). The features are ordered according to their importance. Thus, age is the most important feature for predicting Retinopathy and higher age contributes positively to Retinopathy. Similarly, HbA1c is the second most contributing feature and higher HbA1c values correlate with Retinopathy. Finally, Gender and Diastolic BP do not seem to have any predictive value for Retinopathy.

We used five-fold cross-validation to select hyper-parameters and evaluate model performance. In five-fold cross-validation, the dataset is randomly divided into five folds or equal portions, and the model is trained on four folds and validated on the remaining fold. Finally, this procedure is repeated five times and the average result for the training and validation sets over the five runs are reported. With the best hyper-parameters and after two iterations, XGBoost's mean AUC on the training set was 93.4% and on the validation set was 85.5%. We then used the same hyper-parameters to train XGBoost on the full dataset and achieved an overall model AUC of 93.1%. AUC values over 80% indicate that the model has excellent discrimination power (12). The model's hyper-parameters were tuned so that the positive class (retinopathy present) was given higher weightage to reduce the likelihood of missing retinopathy cases, even at the cost of a few false positives. Thus, over the entire dataset of 119 patients, the model's prediction resulted in 25 true positives (100% recall of retinopathy positive patients), 81 true negatives, 13 false positives, and 0 false negatives.

Finally, we used SHAP to explain the predictions of XGBoost. SHAP provides a unified framework based on cooperative game theory to assign importance values to features or variables in a predictive model (13). It enables users to gain a deeper understanding of the factors driving a model's predictions and helps build trust in the model's decision-making process. For our XGBoost model, the most important features predictive of retinopathy according to SHAP were age, HbA1c, HDL, LDL, and systolic BP with the first two being most prominent (**Figure 2**). Specifically, we found that age > 50 and HbA1C > 8 were the most significant predictors of retinopathy (**Figure 3**). HDL, LDL, and systolic BP were modestly predictive, with higher HDL and LDL negatively correlated and systolic BP positively correlated with retinopathy (**Figure 4**).

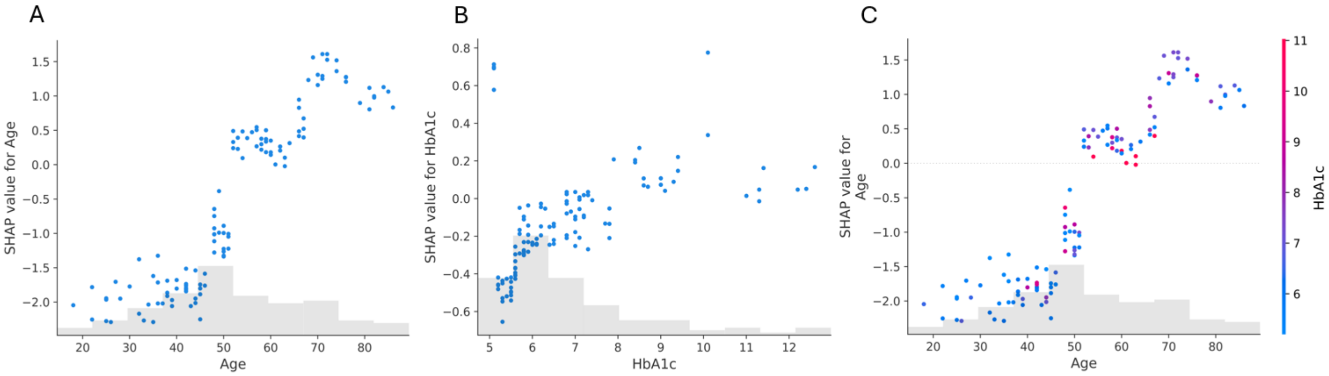

Figure 3: Scatter plots of SHAP value show age and HbA1c are positively correlated with retinopathy. (A) Age is highly positively correlated with retinopathy (Pearson Correlation Coefficient R2 of 0.90). **(B)** HbA1c has positive correlation with retinopathy (R2 = 0.51) with HbA1c values > 8% having a positive predictive value. Note the different scales of the y-axis for Figure 3A and 3B, showing that age has much higher SHAP values than HbA1c. **(C)** Interaction effects between age and HbA1c using color with red indicating higher HbA1c values.

Our analysis of the XGBoost machine learning model suggests that retinopathy screening could be prioritized towards patients older than 50 years, especially the ones with HbA1c values above 8, as these patients are identified as having a higher risk of exhibiting retinopathy. By targeting screening of these high-risk patients, one could identify and treat retinopathy in a scalable, cost-effective manner in developing countries with large, susceptible populations.

DISCUSSION

The effect of age on the prevalence and severity of retinopathy is still unclear in the literature (14). Previous studies have shown conflicting evidence on the impact of age on diabetic retinopathy. While some have reported that older age is a risk factor for the progression of retinopathy (15), others identified younger age at diagnosis as a risk factor (16). These differences may be explained by confounders such as genetic variations, lifestyle differences and the type of study population. Our study shows that age is the strongest predictor of retinopathy. This may be due to the absence of the above confounding factors in our study.

Since diabetes is a known risk factor for retinopathy, blood sugar levels measured by HbA1c are a good proxy to measure diabetic control in patients. Not surprisingly, previous studies have indicated that the risk of retinopathy increased with HbA1c levels (4, 10). Our study identifies HbA1c as the second most important predictor of retinopathy. Thus, our study augments these prior studies and suggests that age along with HbA1c is a strong predictor of retinopathy.

The role of lipids in the pathogenesis of retinopathy remains controversial today. Higher total cholesterol, triglyceride, HDL, and LDL levels all associated with increased risk of retinopathy (7). In the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR), a significant association was found between high total cholesterol levels and retinopathy (17). Similarly, in the Early Treatment Diabetic Retinopathy Study (ETDRS), patients with high total cholesterol or high LDL levels were found more likely to be diagnosed with retinopathy (18). Interestingly, in the United Kingdom Prospective Diabetes Study (UKPDS), triglyceride and LDL levels were found to have no association with retinopathy. Instead, they found that higher HDL levels correlated with retinopathy (19). Cardiovascular Health Study (CHS) showed a relationship of retinopathy with total cholesterol and LDL levels (20). However, there was no correlation with HDL or triglyceride levels. Finally, in the SMART study, the authors found no significant relationship between total cholesterol levels and retinopathy (10).

In contrast to the above findings, our study found a modest negative association between high HDL and high LDL levels

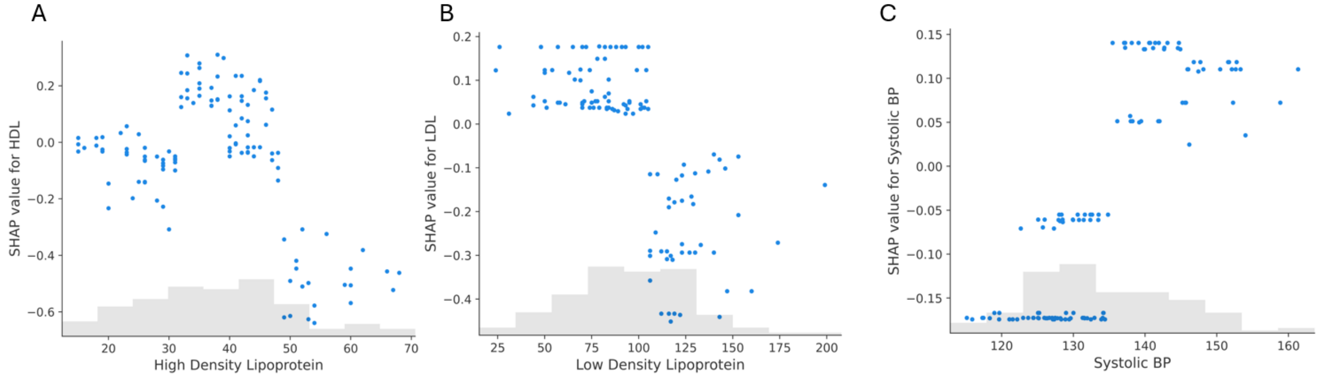

Figure 4: Scatter plots of SHAP values show HDL and LDL are negatively correlated and Systolic BP is positively correlated with retinopathy. (A) HDL is negatively correlated with retinopathy (R2 = -0.46) with HDL values > 50 mg/dL having the most protective impact. **(B)** LDL has negative correlation with retinopathy (R2 = -0.70). **(C)** Systolic BP has a positive correlation with retinopathy (R2 = 0.82).

with retinopathy. While high HDL can have a protective, anti-inflammatory effect, thereby decreasing the risk of developing retinopathy, the negative association of high LDL in our study is hard to explain. It could be that the effect is small and is not visible in a study of such modest size as ours (**Figure 4**). Thus, the effect of lipid levels on retinopathy remains to be understood.

Our study also indicated that retinopathy was positively associated with systolic blood pressure but had no association with diastolic blood pressure. This association has also been seen in previous studies (5, 6). Yu et al. found an association between high blood pressure in general and increased risk of retinopathy while Van Leiden et al. specifically suggested systolic blood pressure in particular increases the risk of retinopathy (5, 6). However, the importance of systolic blood pressure was relatively low in our study (**Figure 4**). More studies are necessary to better understand the importance of systolic blood pressure for risk of retinopathy.

Furthermore, our study found no association between the risk of retinopathy and gender. This finding corroborates the results from other studies such as the SMART India study (10).

As part of tuning the hyper-parameters of the XGBoost algorithm, we were careful to give higher prediction weightage to the retinopathy positive class compared to the retinopathy negative class. This ensured that false negatives were zero for our dataset while false positives were slightly high. Such tuning is critical since missing true positives such as predicting patients with retinopathy as normal can be highly detrimental. More study is needed to ensure that hyper-parameters tuned in this way does not result in a significant increase in false positives.

Finally, given that our study size is modest and is focused on an urban Indian population, larger studies with diverse populations are needed to shed more light on this important area. As part of future work, we would like to collect patient biomarkers from larger as well as more diverse populations to build stronger evidence for our hypothesis.

In summary, we developed an XGBoost machine learning model that had an AUC of 93% in identifying retinopathy. Further, we used SHAP to identify age and HbA1c as the major predictors of retinopathy in the population we studied. Since age is a non-modifiable risk factor, this suggests that controlling HbA1c is critical for controlling the progress of retinopathy. Further, retinopathy screening camps that prioritize older patients with higher HbA1c levels can be a scalable, cost-effective option, especially in developing countries.

MATERIALS AND METHODS

This study was first proposed to and approved by the Institutional Ethical Committee, BLDE, SHRI B.M. Medical College, Vijayapura, India. After the approval, we obtained informed consent and collected key biomarker data such as HbA1c, BP, and lipid values from 119 patients for analysis.

XGBoost model

To validate the hypothesis, the XGBoost machine learning algorithm was used (11). XGBoost stands for "Extreme Gradient Boosting" and is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. Patients were classified into retinopathy positive or negative classes based on their biomarkers.

AUC was used as the metric for evaluating the performance of binary classification models. In binary classification, the AUC metric measures the area under the receiver operating characteristic (ROC) curve. The ROC curve is created by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds. The true positive rate represents the proportion of actual positive samples correctly classified as positive while the false positive rate represents the proportion of actual negative samples incorrectly classified as positive.

The AUC metric provides a measure of how well a model can distinguish between positive and negative samples across all possible classification thresholds. It ranges from 0 to 1, where an AUC of 0.5 indicates that the model's predictions are no better than random; an AUC of 0.8 indicates that a model has outstanding predictive power and an AUC of 1 represents a perfect classifier that can perfectly separate the two classes (12). It is worth noting that the AUC metric is insensitive to the specific classification threshold chosen and is suitable for imbalanced datasets, where the number of samples in the positive and negative classes differs significantly, such as the dataset used in the study.

Further, since our dataset is small and contains only 119 patients, K-fold cross-validation was used (21). In the K-fold cross-validation method, the dataset is divided into k subsets or folds. The model is trained and assessed k times, with each iteration utilizing a distinct fold as the validation set. By averaging the performance metrics obtained from each fold, the model's generalization ability can be accurately estimated. Specifically, K=5 was used to select hyper-parameters and evaluate XGBoost's model performance.

The objective used to optimize the task was "binary:logistic". XGBoost has a number of hyper-parameters that can be tuned. Two hyper-parameters were tuned, namely, scale_pos_weight, which is the scaling value to be used for the positive class, and alpha, a regularizer that is used to reduce the complexity of the model so that it does not overfit on the training dataset. Since four negative class examples for each positive class example existed in the dataset, scale_pos_weight was set to be 4. Finally, alpha was set to be 5 since that gave the best AUC on the 5-fold cross validation. These tuned hyper-parameters were then used to train an XGBoost model on the entire dataset. This resulted in an AUC of 93.1% and the model was able to correctly predict all 25 positive examples.

SHAP analysis

The model was then analyzed using SHAP, which is an approach for explaining the predictions of machine learning models (13). The main idea behind SHAP is to measure the contribution of each feature to the difference between the expected model output and the actual prediction for a specific instance. It calculates feature importances by considering all possible combinations of features and their interactions. By utilizing SHAP, feature importance values that reflect the individual and interactive effects of features in a model prediction can be obtained. These importance values provide insights into the model's decision-making process and help understand the relative influence of different features on the predictions.

SHAP takes in the trained XGBoost model and the dataset and identifies the top biomarkers that resulted in the model's predictions. A SHAP summary plot to list the top biomarkers and their importance was used (**Figure 2**). The individual features were then analyzed using the SHAP scatter plots and the specific age and HbA1c thresholds that were useful for predicting retinopathy were identified (**Figure 3**). Finally, the scatter plots were developed for other biomarkers such as HDL, LDL, and Systolic BP (**Figure 4**).

The python code used for training the XGBoost model, evaluating its predictive power and performing the SHAP analysis is listed in the **Appendix**.

ACKNOWLEDGEMENTS

We would like to thank Forus Health for their support in providing Fundus Camera for data collection.

Received: July 25, 2023
Accepted: October 11, 2023
Published: September 5, 2024

REFERENCES

1. "Diabetes: Facts and Figures." *International Diabetes Federation*. idf.org/about-diabetes/facts-figures/. Accessed 31 Aug. 2023.
2. "India Diabetes Report 2000 – 2045." *Diabetes Data Portal*. diabetesatlas.org/data/en/country/93/in.html. Accessed 31 Aug. 2023.
3. Ramakrishnan, S. and K. Gupta. "Prevalence of hypertension among Indian adults: Results from the great India blood pressure survey." *Indian Heart Journal*, 2020, https://doi.org/10.1016/j.ihj.2019.09.012.
4. Memon, Wasim R., et al. "Diabetic retinopathy: frequency at level of HbA1C greater than 6.5%." *Professional Med J* 2017;24(2):234-238, https://doi.org/10.29309/TPMJ/2017.24.02.510.
5. Yu, T., et al. "Retinopathy in older persons without diabetes and its relationship to hypertension." *Arch Ophthalmology*. 1998;116(1):83–89, https://doi.org/10.1001/archopht.116.1.83.
6. Van Leiden, Henrik A., et al. "Blood pressure, lipids, and obesity are associated with retinopathy: the Hoorn study." *Diabetes Care*. 2002 Aug;25(8):1320-5, https://doi.org/10.2337/diacare.25.8.1320.
7. Liu, Zhenzhen, et al. "Association between increased lipid profiles and risk of diabetic retinopathy in a population-based case-control study." *J Inflamm Res.* 2022 Jun 10;15:3433-3446, https://doi.org/10.2147/JIR.S361613.
8. Bryl, Anna, et al. "The effect of hyperlipidemia on the course of diabetic retinopathy-literature review." *J Clin Med*. 2022 May 13;11(10): 2761, https://doi.org/10.3390/jcm11102761.
9. Kropp, Martina, et al. "Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation." *EPMA Journal* 14, 21–42, 2023. https://doi.org/10.1007/s13167-023-00314-8.
10. Raman, Rajiv, et al. "Prevalence of diabetic retinopathy in India stratified by known and undiagnosed diabetes, urban–rural locations, and socioeconomic indices: results from the SMART India population-based cross-sectional screening study." *The Lancet Global Health* 10.12 (2022): e1764-e1773, https://doi.org/10.1016/S2214-109X(22)00411-9.
11. Chen, T., and C. Guestrin. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM, https://doi.org/10.1145/2939672.2939785.
12. Hosmer, David W., et al. Applied Logistic Regression. *Wiley*, 2013. https://doi.org/10.1002/9781118548387
13. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30, 2017, https://doi.org/10.48550/arXiv.1705.07874.
14. Tan, Colin, et al. "Is age a risk factor for diabetic retinopathy?" *British Journal of Ophthalmology* 94.9 (2010): 1268-1268, https://doi.org/10.1136/bjo.2009.169326.
15. Stratton, I. M., et al. "UKPDS 50: risk factors for incidence and progression of retinopathy in Type II diabetes over 6 years from diagnosis." *Diabetologia* 44 (2001): 156-163, https://doi.org/10.1007/s001250051594.
16. Klein, Ronald, et al. "The Wisconsin Epidemiologic Study of Diabetic Retinopathy: III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years." *Archives of ophthalmology* 102.4 (1984): 527-532, https://doi.org/10.1001/archopht.1984.01040030405011.
17. Klein, Barbara EK, et al. "The Wisconsin Epidemiologic Study of Diabetic Retinopathy: XIII. Relationship of serum cholesterol to retinopathy and hard exudate." *Ophthalmology* 98.8 (1991): 1261-1265, https://doi.org/10.1016/S0161-6420(91)32145-6.
18. Chew, Emily Y., et al. "Association of elevated serum lipid levels with retinal hard exudate in diabetic retinopathy: Early Treatment Diabetic Retinopathy Study (ETDRS) Report 22." *Archives of ophthalmology* 114.9 (1996): 1079-1084, https://doi.org/10.1001/archopht.1996.01100140281004.
19. Kohner, Eva M., et al. "United Kingdom Prospective Diabetes Study, 30: diabetic retinopathy at diagnosis of non–insulin-dependent diabetes mellitus and associated risk factors." *Archives of ophthalmology* 116.3 (1998): 297-303, https://doi.org/10.1001/archopht.116.3.297.
20. Klein, Ronald, et al. "The relation of atherosclerotic cardiovascular disease to retinopathy in people with diabetes in the Cardiovascular Health Study." *British journal of ophthalmology* 86.1 (2002): 84-90, https://doi.org/10.1136/bjo.86.1.84.
21. Trevor, Hastie, et al. "The elements of statistical learning: data mining, inference, and prediction. Chapter 7." *Springer*, 2009.

https://doi.org/10.59720/23-184

Appendix

```
#Python code for model training and explanation below
import pandas as pd
# import XGBoost
import xgboost as xgb
from xgboost as cv
from sklearn.metrics import roc_auc_score, confusion_matrix
import shap
import matplotlib.pyplot as pl
%matplotlib inline
# Load dataset
dataset = pd.read_csv('retinopathy_data.csv', sep=',|/', engine='python')

# train XGBoost model
X1 = dataset.drop(['RetinopathyPositive'], axis=1)
y1 = dataset['RetinopathyPositive']
data_dmatrix = xgb.DMatrix(data=X1,label=y1)

# Tune hyper-parameters scale_pos_weight and alpha, optimizing AUC
params = {"objective":"binary:logistic",'scale_pos_weight': 4, 'alpha': 5}
xgb_cv = cv(dtrain=data_dmatrix, params=params, nfold=5, num_boost_round=100,
    early_stopping_rounds=10, metrics="auc", as_pandas=True, seed=123)
print(xgb_cv)

# Use hyper-parameters to train model on full dataset
bst = xgb.train(dtrain=data_dmatrix, params=params, num_boost_round=50)

# Use bst, the XGBoost model, to evaluate predictions
y_pred = bst.predict(data_dmatrix)
predictions = [round(value) for value in y_pred]
print('XGBoost model auc: {0:0.4f}'. format(roc_auc_score(y1, predictions)))

# Calculate true negative, false positive, false negative, true positive
tn, fp, fn, tp = confusion_matrix(y1,predictions).ravel()
```

```
print(tn,fp,fn,tp)

# Explain the XGBoost model using SHAP
explainer = shap.TreeExplainer(bst)
shap_values = explainer(X1)
shap.summary_plot(shap_values,X1, plot_type='violin')
shap.plots.scatter(shap_values[:,"Age"]) # etc.

# Pearson Correlation Coefficient Calculation
Import numpy as np
np.corrcoef(shap_values[:,”Age”].data,shap_values[:,”Age”].values) #age. etc.
```