UNIVERSITY OF WATERLOO

Faculty of Mathematics

Machine Learning for Characterizing the Biomedical Features of Diabetic Retinopathy in Patients with Type 2 Diabetes Mellitus

University of Waterloo-Faculty of Mathematics

Waterloo, ON.

Prepared by
Chuan Shi

Computational Physiology Team
ID 20817410

Table of Contents

List of Figures

Abstract

Background

Diabetic retinopathy is the major cause of blindness among patients with diabetes mellitus. A considerable amount of studies on the risk factors of diabetic retinopathy show that the long duration of diabetes is the most significant risk factors of diabetic retinopathy. However, other risk factors have been identified with varying importance in past studies. The purpose of this study is to identify the risk factors of diabetic retinopathy in type 2 diabetes mellitus by machine learning algorithms. Machine learning is a new and rapidly evolving method to analyze the interaction among significant features, which has been widely applied in the medical field. Correspondingly, machine learning can be used to build prediction models to characterize the risk of diabetes mellitus. (Zhang et al., 2020). The purpose of this report is to apply the machine learning algorithms to predict the presence of diabetic retinopathy in patients with type 2 diabetes mellitus (T2DM) and determine the discriminative features.

Risk assessment models for diabetic retinopathy in patients with T2DM were developed using three machine learning algorithms, including random forest (RF), light gradient boosting machine (LightGBM), and logistic regression (LR). The model performance was measured in an area under the receiver

operating characteristic curve (AUC), sensitivity, specificity, F1 score, and area under the precision-recall curve.

Participants and Data set

503 southern Chinese patients with type 2 diabetes mellitus.

(Zhuang et al., 2019)

1.0 Introduction

Diabetic retinopathy is a retinal complication of diabetes and it is prevalent among patients with type 2 diabetes. DR is the major cause of blindness in type 2 diabetes patients. According to the WHO, 4.8% of all blindness cases globally are attributed to DR (Resnikoff et al., 2004). Diabetes is growing faster than population growth and people are getting it at younger ages nowadays, so there is enough and emergent reason for society and medical experts to find effective prevention and treatment of diabetic retinopathy, since

Diabetic retinopathy is diagnosed by clinical ophthalmic examination and image evaluation. Non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR) are two forms of diabetic retinopathy (PDR). History diabetic retinopathy is another name of NPDR (BDR). As DR progresses, it moves from the less severe Non-proliferative form to the sever proliferative stage. The test targets of diagnosis in the ophthalmic examination of diabetic retinopathy are listed below:

- Abnormal blood vessels
- Swelling, blood, or fatty deposits in the retina
- Growth of new blood vessels and scar tissue
- Bleeding in the clear, jelly-like substance that fills the center of the eye (vitreous)

- Retinal detachment
- Abnormalities in your optic nerve

To the suggestions from Taiwan diabetic association, T2D patients need to perform screening of fundus examination annually and perform more frequently if they already have diabetic retinopathy. However, the screening rate is low since many patients do not realize they have diabetic retinopathy in the early stage. Once they develop proliferative diabetic retinopathy, they lost vision suddenly. Also, the accuracy of the diagnosis of DR especially for the stage of DR by the fundus photograph is not high enough and many people are building deep learning models to help assist the diagnosis of DR. Thus, identifying the interpretable biomedical features is beneficial for the medical practitioner.

In this study, patient characteristic and laboratory data were included to build the prediction model of diabetic retinopathy, and the methods applied in this study to identify the risk factors was by the similar way of a study to characterizing the risk factors of type 2 diabetic mellitus.

This project is carried to identify the important biomedical features which show correlation with diabetic retinopathy in patients with type 2 diabetes mellitus. In this way, the clinician is able to early detect early diabetic retinopathy for patients with T2D by their blood tests. Through the results in the risk

assessment models, machine learning algorithms can be used to assist the diagnosis and treatment of diabetic retinopathy.

2.0 Method

Study population

The raw data set consists of 503 patients with diabetes in southern China. The patients had undergone ophthalmic consultation between December 2017 and November 2018 at the Guangdong Provincial People's Hospital's Department of Endocrinology. This study included patients with T2DM (by the WHO criteria (Alberti et al., 1998)) and reports from the Early Treatment Diabetic Retinopathy Study (ETDRS) 35-degree 7-standard fields color retinal photographs (Topcon TRC; Topcon, Tokyo, Japan) on them. Any other ocular condition that could impair ocular circulation (e.g., glaucoma, endophthalmitis, retinal vascular occlusion, age-related macular degeneration, refractive error >3 diopters, eye trauma), any serious systemic disorders (e.g., myocardial infarction, cerebral infarction, connective tissue disorder), or a history of prior intravitreal injection or dialysis were omitted. (Zhuang et al., 2019). Patients' medical records were used to obtain all of the medical information. Sex, age, diabetes mellitus (DM) length, height, weight, and blood pressure were among the demographic and physical data collected. BMI was determined by weight divided by height squared. A systolic blood pressure of 140 mm Hg or diastolic blood pressure of 90 mm Hg is considered hypertension. (Zhuang et al., 2019).

Assessment of DR

According to the description of the diagnosis of diabetic retinopathy on the source of data set used in this study (Zhuang et al., 2019), DR and DME (diabetic macular edema) were diagnosed both on the clinical ophthalmic examination and image evaluation by two trained graders. And the examination will be further confirmed by a fundus expert if the graders have a different diagnosis on the same patient. However, the data set also includes some patients with undiagnosed patients.

Machine Learning Algorithm

Random Forests

Random forest (RF) is an ensemble learning theory. At training time, RF generates a large number of decision trees for randomly splitting data. A subset containing K attributes is randomly selected from the attribute set of each node in the base decision tree, and then an optimal attribute is selected from the subset for partitioning. (Svetnik, V. et al. 2003)

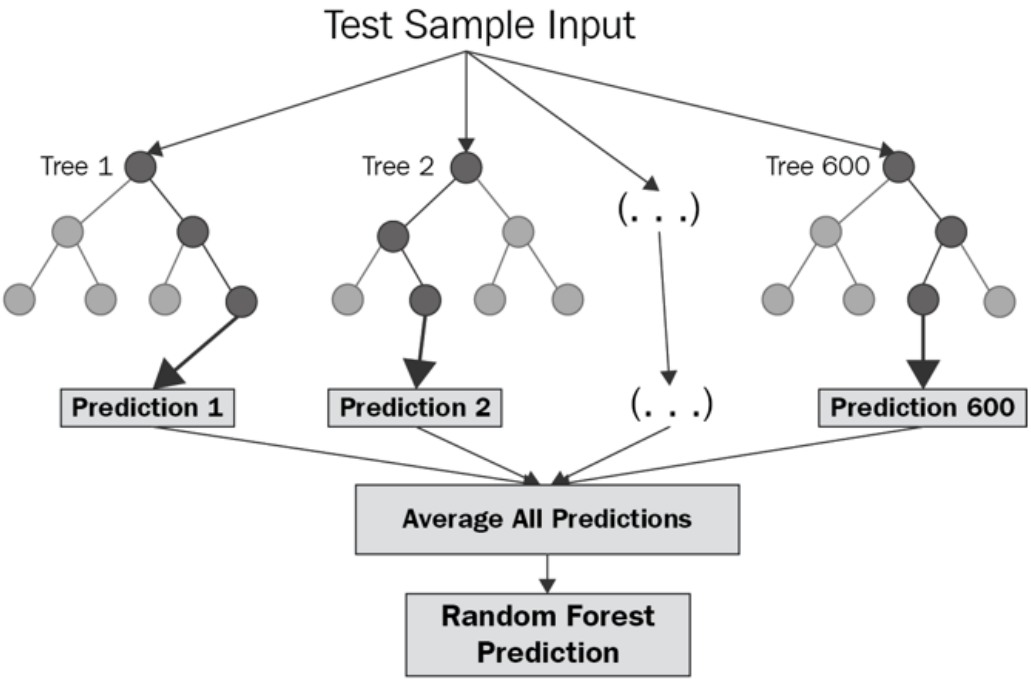

Figure 1. Random Forest Structure https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

Light Gradient Boosting Machine

Light GBM is a gradient boosting framework that uses a tree-based learning algorithm. Gradient boosting machine (GBM) is an iterative algorithm in which different classifiers are trained for the same training set, and then these weak classifiers are combined to form a stronger final classifier. Each iteration is implemented into a weak classifier to solve the established shortcomings of weak classifier combinations through a sequence of iterations to improve the classification performance. When training each weak classifier, GBM uses the residual of training data fitted by the previous weak classifier to improve the model. (Zhang et al., 2020)

Logistic Regression

Logistic regression (LR) is a type of generalized linear regression analysis that aims to find the best model for describing the relationship between dependent and independent predictors. (Bagley et al., 2001) The probability of an individual developing diabetic retinopathy is p(Y=1|X) = p(X)p(Y=1|X) = p(X). Then, the formula of the LR model is defined as follows. (Zhang et al., 2020)

$$logit(p) = ln\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (1)$$

and equivalently, after exponentiating both sides:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k} \quad (2)$$

The probability of an individual developing T2DM is

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}} \quad (3)$$

Where $X = (X_1, X_2 \cdots X_k)$ represents the risk factors, $\beta = (\beta_1, \beta_2 \cdots \beta_k)$ are the coefficients estimated by using the method of maximum likelihood.

Figure 2- Mathematical Derivation of Logistic Regression https://www.nature.com/articles/s41598-020-61123-x#ref-CR39

Feature Importance Measurements

Mean decrease in impurity (Gini) importance

Random forests are treated as “black box” prediction model, but the importance Metrics associated with each feature can be measured as output. Mean decrease in impurity matric is used to show the improvement in the "Gini gain" for the classification problem. It incorporates a weighted mean of individual tress' improvement in the splitting criterion by each feature.

The Gini impurity index:

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i) = 1 - \sum_{i=1}^{n_c} p_i^2$$

Where n_c is the number of classes in the target variable and p_i is the ratio of this class.

Permutation Feature Importance

This method will randomly shuffle each feature and compute the change in the model's performance. The features which impact the performance the most are the most important ones.

Odds Ratio

the odds ratio represents the constant effect of a predictor X, on the likelihood that one outcome will occur. Which is used in the coefficient effect size for logistic regression.

Shapley Additive Explanations

It is a game theoretic method which uses the Shapley values for local data point from game theory to estimate how each feature contributes to the prediction. Mathematical explanation:

$$\phi_i = \sum_{S \subseteq M \setminus i} \frac{|S|!(|M| - |S| - 1)!}{|M|!}[f(S \cup i) - f(S)]$$

A key part of this is the difference between the model's prediction with the feature i, and the model's prediction without feature i.
S refers to a subset of features that doesn't include the feature for which we're calculating ϕ_i.
$S \cup i$ is the subset that includes features in S plus feature i.
$S \subseteq M \setminus i$ in the Σ symbol is saying, all sets S that are subsets of the full set of features M, excluding feature i.

Shapley, Lloyd S. "A value for n-person games."

Figure 3

Project Procedures and Model Performance Measurements

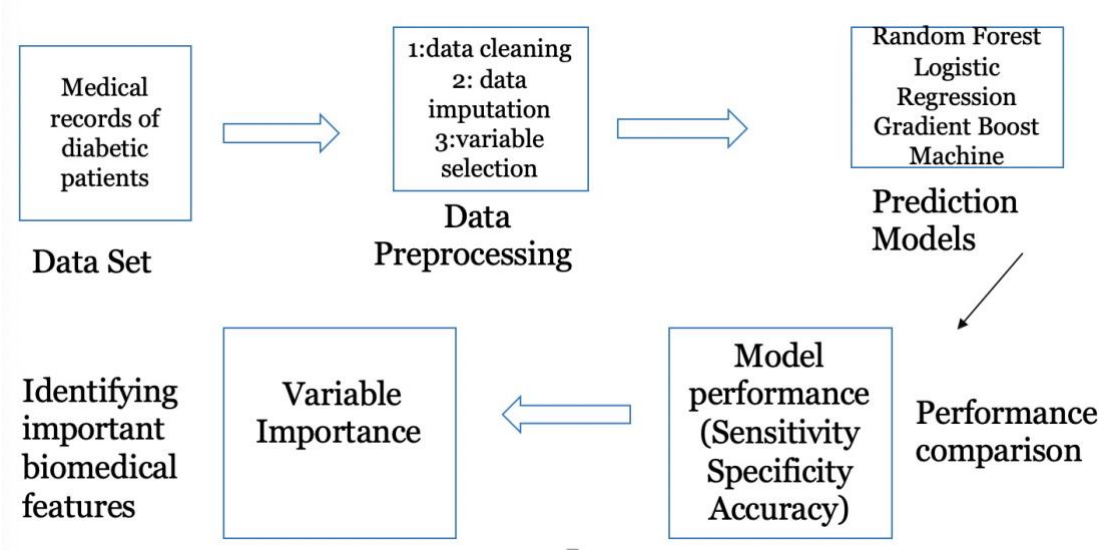

Figure 4

The models were trained by stratified 10-fold cross-validation. This is an extensive version of regular k-fold cross validation which splits the training set with the same ratio of the target classes in the full dataset. This method avoids the overfitting and the unbalanced distribution for training the predictive model. For the comparison between the models, accuracy, F1-score, and area under the receiver operating characteristics (ROC) curve were applied as the measurements. The accuracy refers to the accuracy of the model prediction performance on the test set. F1-score is calculated from the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive. Precision is also known as positive predictive value, and recall is also known as sensitivity in diagnostic binary classification. The Area under curve (AUC) is

used to evaluate discrimination which refers to the model's ability to identify who is at risk of developing diabetic retinopathy and who is not.

3.0 Results

Descriptive statistical analysis

The data set contains 503 patients. Removing 2 patients under 18 to avoid bias. For the rest data set 424 patients have a final diagnosis of the presence of diabetic retinopathy and 77 patients have an uncertain presence of diabetic retinopathy. The 424 patients (257 no DR and 167 DR) were included first in the statistical analysis and model training. For this study, the explanations and the abbreviation of features:

General data:

sex: 0 for female; 1 for male

age(years) : range from 14 to 92

duration(years) : duration of diabetes mellitus (years)

hbp: whether or not the patients have hypertension 1 for the existence of hypertension, 0 is not. Hypertension is defined as

systolic blood pressure ≥140 mm Hg or diastolic blood pressure ≥90 mm Hg

sbp：systolic blood pressure dbp: diastolic blood pressure (mm Hg)

HbA1c(%): Hemoglobin A1C normal: Below 6.0%; prediabetes:6.0% to 6.4% Diabetes:6.5% or over

Renal data:

bun(mmol/L): Blood urea nitrogen.The normal range is 2.1–7.1 mmol/L

urea (Serum urea)

utp (mg/L):Urinary total protein.

ualb (mg/L): Urinary albumin

Ucr (μmol/L): urinary creatinine

UACR (mg/g）: urine albumin-to-creatinine ratio

UPCR (mg/g) : urinary protein/Ucr

eGFR (mL/min/1.73 m^2): estimated glomerular filtration rate

Blood liqid:

NEFA (mmol/L): non-esterified fatty acid

HDL(mmol/L): high-density lipoprotein

LDL (mmol/L): low-density lipoprotein

TRIG (mmol/L)： triglycerides

CHOL(mmol/L): total cholesterol

Lpa (mg/L) : lipoprotein a

APOA (g/L) : apolipoprotein A

APOB (g/L): apolipoprotein B

Others:

Uric (μmol/L） : uric acid

ALT (U/L) : alanine aminotransferase

AST (U/L) : aspartate transaminase

Che (U/L): acetylcholinesterase

ALB (g/L) : serum albumin

TP (g/L) : total protein

ddimer (μg/L): a fibrin degradation product

VitB12 (μmol/L): vitamin B12

The distribution of categorical features:

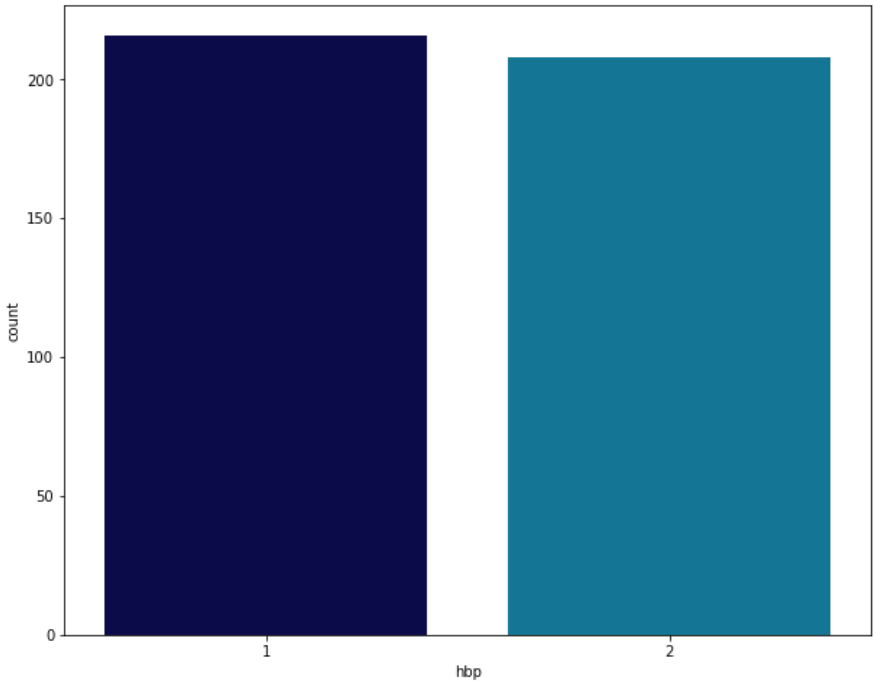

Figure 5

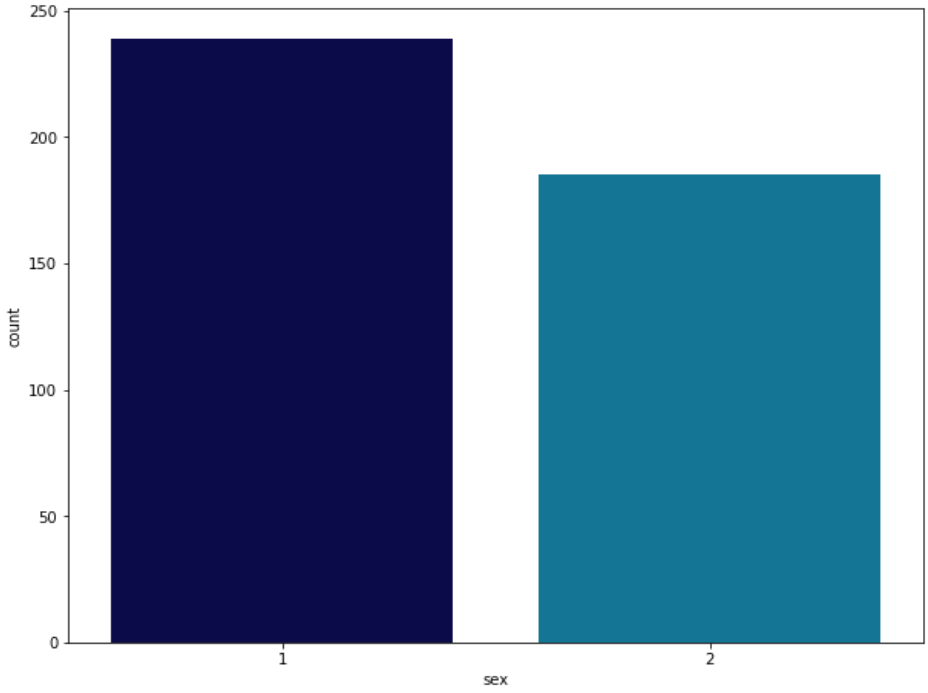

Figure 6

239 males and 185 females

208 patients have hypertension and 216 patients do not have hypertension.

Data imputation

```
0   sex        424 non-null
1   age        424 non-null
2   duration   424 non-null
3   hbp        424 non-null
4   sbp        424 non-null
5   dbp        424 non-null
6   hb1ac      419 non-null
7   uric       416 non-null
8   bun        424 non-null
9   urea       424 non-null
10  NEFA       406 non-null
11  HDL        423 non-null
12  LDL        423 non-null
13  TRIG       423 non-null
14  CHOL       422 non-null
15  Lpa        389 non-null
16  APOA       389 non-null
17  APOB       389 non-null
18  ALT        424 non-null
19  AST        424 non-null
20  ChE        423 non-null
21  ALB        423 non-null
22  TP         423 non-null
23  ddimer     421 non-null
24  VitB12     356 non-null
25  utp        413 non-null
26  ualb       417 non-null
27  ucr        417 non-null
28  UACR       417 non-null
29  UPCR       413 non-null
30  eGFR       424 non-null
```

Figure 7

In the sample data, there are a decent number of missing values and abnormal zero values which are likely to cause the bias of machine learning prediction. Thus, there are some methods to

impute the data and the random-forest-based imputation was chosen for this research (Sam Wilson 2020).

Statistical significance table

pvalue	mean(dr=1)	std(dr=1)	mean(dr=0)	std(dr=0)	
ChE	4.144804e-01	8276.796407	2135.138100	8443.859922	2005.921719
age	5.390473e-02	60.413174	12.897766	58.081712	13.701802
duration	1.455421e-10	12.383234	7.274925	8.066148	7.635207
sbp	3.816561e-04	143.281437	23.305433	135.287938	19.058780
dbp	4.355434e-01	80.149701	12.471958	80.198444	11.636567
hb1ac	4.940133e-01	9.709880	2.276049	9.739689	2.398010
uric	4.789568e-02	381.798263	119.832225	361.466926	105.492668
bun	6.594827e-06	9.857246	27.159325	5.900506	2.409572
urea	5.566713e-04	111.158683	92.421082	77.735953	29.533253
NEFA	7.993300e-06	0.340120	0.190304	0.423191	0.206009
HDL	1.473066e-02	1.066347	0.316731	1.003891	0.333413
LDL	1.010015e-02	3.369880	1.110072	3.105447	0.900924
TRIG	4.933668e-01	2.215928	2.429350	2.529494	5.940556
CHOL	1.089077e-02	5.269701	1.852956	4.853696	1.389414
Lpa	4.868849e-03	237.730539	252.094036	188.225681	201.008054
APOA	1.594781e-01	1.172575	0.248133	1.142607	0.238235
APOB	9.814339e-02	0.967485	0.305166	0.925603	0.257518
ALT	2.055497e-05	19.580838	13.527451	27.501946	35.834703
AST	3.908673e-04	19.874251	9.254878	25.031128	27.784589

pvalue	mean(dr=1)	std(dr=1)	mean(dr=0)	std(dr=0)	
ALB	4.936936e-06	35.699401	5.688940	38.108171	3.930228
TP	3.048509e-02	64.461677	6.526745	65.643969	5.714323
ddimer	1.163047e-04	840.598802	1812.616993	472.412451	364.705062
VitB12	2.167953e-02	462.281437	279.635712	396.101167	227.126766
utp	2.679788e-07	835.771737	1449.621427	228.873696	707.072739
ualb	1.155685e-14	377.559940	708.663960	57.723268	228.339899
ucr	7.702465e-09	7.391737	9.876842	9.890973	5.967271
UACR	1.055061e-20	623.075903	1162.528063	97.978890	455.044625
UPCR	5.352715e-17	1409.395587	2521.524228	358.791371	1757.746437
eGFR	8.393216e-07	75.566909	40.727489	90.028899	27.863894

Table 1

Shapiro-Wilk test was used to test the normality for the continuous features and independent t-test and Mann-Whitney U test were used to compare the distribution of the features in group of patients with DR and without DR. Features show statistical significance : 'duration', 'sbp', 'uric', 'bun', 'urea', 'NEFA', 'HDL', 'LDL', 'CHOL', 'Lpa', 'ALT', 'AST', 'ALB', 'TP', 'ddimer', 'VitB12', 'utp', 'ualb', 'ucr', 'UACR', 'UPCR', 'eGFR'. The threshold value for statistic tests are 0.05 of p-value.

Comparison of model performance

Three Model performance on the stratified 10-fold cross-validation. The hyperparameters of models were tuned by 400 trails.

Models	Random Forest	Light Gradient Boosting Machine	Logistic Regression
Accuracy out-of-fold	0.761	0.788	0.7286
F1 score out-of-fold	0.637	0.710	0.607
Accuracy on the test set	0.788	0.777	0.741
F1 score	0.678	0.655	0.607
Hyperparameter after tuning	{'n_estimators': 39, 'max_depth': 28, 'min_samples_split': 6, 'min_samples_leaf': 4}	{'colsample_bytree': 0.7993039578263481, 'learning_rate': 0.32277549340076406, 'max_depth': 24, 'min_child_samples': 42, 'min_child_weight': 0.19969898786372456, 'n_estimators': 138, 'num_leaves': 375,	{'C': 5.533253688880515, 'intercept_scaling': 1.458859796107597, 'max_iter': 935}

		'reg_alpha': 1.6794776473680096e-05, 'reg_lambda': 0.2674863869394413, 'subsample': 0.9745537209676508, 'subsample_freq': 1}	

Table 2

From the table above, the light Gradient Boosting Machine achieved the highest accuracy and F1 score (Accuracy:0.788 F1 score: 0.710) on the cross-validation. Random Forest achieved the highest accuracy and F1 score on the test set (Accuracy: 0.788 F1 score 0.678).

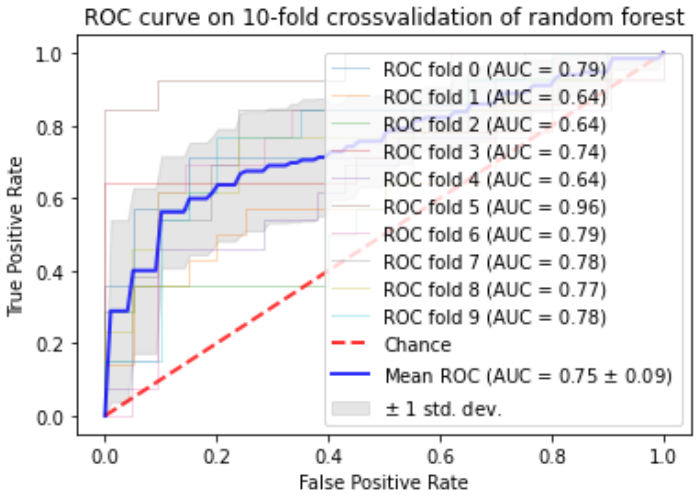

Figure 8

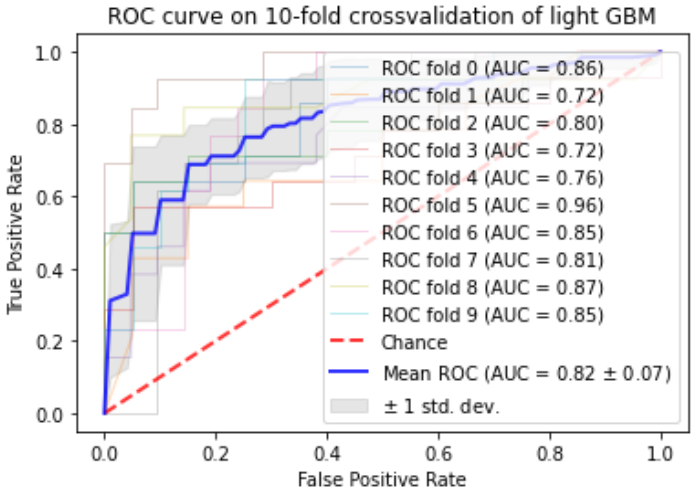

Figure 9

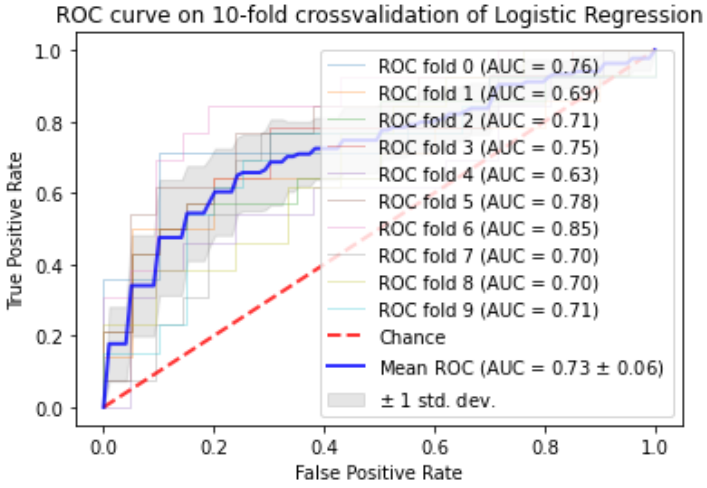

Figure 10

From the ROC curves, light GBM achieved the best AUC by the stratified 10-fold cross-validation (AUC: 0.82).

Feature Importance Analysis

The top-10 ranked variables by the variable importance for each algorithm

Rank	RF (Gini importance)	RF (Permutation Importance)	LightGBM (Gini importance)	LightGBM (Permutation Importance)	LR (Coefficient Effect Size)	LR (Permutation Importance)
1	Urinary albumin	urinary creatinine	urine albumin-to-creatinine ratio	urinary creatinine	Urinary albumin	Urinary albumin
2	urine albumin-to-creatinine ratio	non-esterified fatty acid	duration of diabetes mellitus	urine albumin-to-creatinine ratio	total cholesterol	D-dimer
3	urinary protein/Ucr	duration of diabetes mellitus	urinary creatinine	duration of diabetes mellitus	serum albumin	total cholesterol
4	Urinary total protein	alanine aminotransferase	low-density lipoprotein	non-esterified fatty acid	non-esterified fatty acid	Sex

5	urinary creatinine	aspartate transaminase	non-esterified fatty acid	low-density lipoprotein	low-density lipoprotein	Serum urea
6	duration of diabetes mellitus	D-dimer	Age	serum albumin	Sex	duration of diabetes mellitus
7	non-esterified fatty acid	systolic blood pressure	urinary protein/Ucr	Blood urea nitrogen	Age	vitamin B12
8	alanine aminotransferase	Age	aspartate transaminase	alanine aminotransferase	vitamin B12	total protein
9	estimated glomerular filtration rate	high-density lipoprotein	serum albumin	urinary protein/Ucr	duration of diabetes mellitus	uric acid
10	systolic blood pressure	urine albumin-to-creatinine ratio	lipoprotein A	lipoprotein A	aspartate transaminase	apolipoprotein B

Table 3

The features rank of the frequency in top-11 are duration of diabetes mellitus(6), urine albumin-to-creatinine ratio(4), urinary creatinine(4), non-esterified fatty acid(4), Urinary albumin(3), urinary protein/ucr(3), alanine aminotransferase(3), low-density lipoprotein(3), age(3), aspartate transaminase(3), serum albumin(3). And the top 10 features for each

algorithm are also listed above. Notice that some features are linear dependent, but they were important in improving the performance of prediction models and can be used to demonstrate different body functions. The long duration of diabetes mellitus has been identified to be the risk factor of diabetic retinopathy in both statistical analysis and clinical researches. The longer the patients suffer from diabetes mellitus, the more likely they are to have diabetic retinopathy. UACR (urine albumin-to-creatinine ratio) was calculated by urinary albumin and urinary creatinine and it can be clinically used to evaluate renal function. This group of features shows that there is a correlation between diabetic retinopathy and renal function. This result is corresponding with the result from “Association of diabetic retinopathy and diabetic macular edema with renal function in southern Chinese patients with type 2 diabetes mellitus” (Zhuang et al., 2019). Moreover, low-density lipoprotein is also tested to be the risk factor of DR in Zhuang's study by statistical analysis. In Tan's research conducted in 2019 (Tan et al., 2019) about the risk factors of an early stage of diabetic retinopathy, age was tested to be the risk factor for patients with type 2 diabetes mellitus. For the serum albumin, Moctezuma MY et al showed that less than 3 g/dL serum albumin in Mexican patients with type 2 diabetes mellitus is associated with retinopathy. However, whether or not aspartate transaminase, alanine aminotransferase, and non-esterified fatty acid correlate with diabetic retinopathy and whether or

not they will directly or indirectly cause the progression of diabetic retinopathy requires further research.

Variable ranking based on the mean rank of all models based on the shapley additive explanations approach.

Model		RF	LightGBM	LR	Mean Rank
Feature Importance Rank	Urinary albumin	1	11	1	4.33
	duration of diabetes mellitus	6	2	7	5
	urinary creatinine	4	3	10	5.67
	non-esterified fatty acid	10	5	3	6
	urinary protein/Ucr	5	4	13	7.33
	sex	13	10	2	8.33
	low-density lipoprotein	15	6	4	8.33
	urine albumin-to-creatinine ratio	3	1	22	8.67
	urinary total protein.	2	19	8	9.67

	total cholesterol	9	22	5	12

Table 4

Among the top-10 features across all methods were Urinary albumin , duration of diabetes mellitus, urinary creatinine, non-esterified fatty acid, urinary protein/Ucr, sex, low-density lipoprotein, urine albumin-to-creatinine ratio, urinary total protein. total cholesterol. For new noticeable features, whether or not gender is a risk factor of diabetic retinopathy in T2DM patients has not reached an agreement. The relationship between total cholesterol and diabetic retinopathy also requires more research.

4.0 Conclusions

According to the model performance and feature rankings, on the one hand, the performance of the prediction model shows that the prediction models are valid and credible in a research study, on the other hand, multiple papers are showing that the top-ranked features play an important role in the development of diabetic retinopathy. By the results, there is a correlation between diabetic retinopathy and renal functions. Chen et al and Zhuang et al showed that UACR is associated with the development of diabetic retinopathy. the duration of diabetes has already been identified as the risk factors of diabetes retinopathy (Sekioka R et al., 2015). When diabetic retinopathy occurs in type 2 diabetes mellitus patients, systemic conditions including renal function and blood lipids should be enhanced as much as possible while ocular conditions are treated. (Zhuang et al., 2020) The study also provides the source of the data set used in this research. Low-density lipoprotein (LDL) is the risk factor in the study by the statistical analysis conducted in the study for finding the association between renal function and diabetic retinopathy. (Zhuang et al., 2020). Non-esterified fatty acids were the important feature in machine learning models but there is not enough evidence to show there is any biological or clinical

association between it and diabetic retinopathy. In this project, the features importance ranking is a valid method to identify the highly relevant biomedical features with diabetic retinopathy. In the prediction features, random forest and light gradient boosting machine were all promising classifiers for diabetic retinopathy. The result can also relieve the burden of DR screening caused by a large number of diabetic patients and a shortage of ophthalmologists in China.

5.0 Discussion

Limitation

First, the data set was from a cross-sectional study, so I was not able to analyze the follow-up data of patients. Second, the models included were limited and it might be able to have a more accurate result with more prediction models. Third, the machine learning algorithms are "black box", the results generated by machine learning algorithms require clinical and biological validation and explanation from prospective studies and randomized controlled trials. Fifth, the number of data set is relatively small, I need to have access to a larger data set. Sixth, all the results require clinical and biological proofs.

Next steps

1. There were 77 patients with an uncertain diagnosis of DR classified by three algorithms. And each algorithm had 10 models trained by 10-fold cross-validation. Thus, there were different results for the data set. The data set was used to do the statistical significance tests with the 424 patients. However, the size of the data set is still small. And bootstrap sampling is a good way to compare them.

2. The diabetic macula edema requires further analysis

3. It is more important and complex to analyze the progression of diabetic retinopathy based on the stages of diabetic retinopathy.

4. Conducting a study on the fundus images to predict the stages of diabetic retinopathy is also doable.

References

Zhang, L., Wang, Y., Niu, M. et al. (2020)Machine learning for characterizing the risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. Sci Rep 10, 4406. https://doi.org/10.1038/s41598-020-61123-x

Xuenan Zhuang et al., (2019) Association of diabetic retinopathy and diabetic macular edema with renal function in southern Chinese patients with type 2 diabetes mellitus: a single-center observational study.Doi: 10.1136/bmjopen-2019-031194

Serge Resnikoff et al., (2002) Global data on visual impairment in the year 2002
PMCID: PMC2623053

Alberti KG, Zimmet PZ, (1998) Definition ZPZ. Definition, diagnosis, and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a who consultation. Diabet Med 1998;15:539–53. 10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S

Sam Wilson (2020) Multiple Imputation with Random Forests in Python https://towardsdatascience.com/multiple-imputation-with-random-forests-in-python-dec83c0ac55b

Svetnik, V. et al. (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inform. Comput. Sci. 43, 1947–1958.

Bagley, S. C., White, H. & Golomb, B. A. (2001) Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J. Clin. Epidemiol. 54, 979–985.

Breiman, Leo. (2001) “Random Forests.” Machine Learning 45 (1). Springer: 5-32.

Lundberg, S., Lee, S. I. (2017) A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, 4765–4774.

Sekioka R, Tanaka M, Nishimura T, et al. (2015) Serum total bilirubin concentration is negatively associated with increasing severity of retinopathy in patients with type 2 diabetes mellitus. J Diabetes Complications 2015;29:218–21.

Shapley, Lloyd S. “A value for n-person games.” Contributions to the Theory of Games 2.28 (1953): 307–317.

Tan, F., Chen, Q., Zhuang, X. et al. Associated risk factors in the early stage of diabetic retinopathy. Eye and Vis 6, 23 (2019). https://doi.org/10.1186/s40662-019-0148-z

Moctezuma MY, Rodríguez LL, Parra RJA Association of serum albumin with severity of diabetic retinopathy (2012)