Improving Diabetic Diagnosis and Prevention with Machine Learning on Retinal Imaging

Yushan Min^{1,*}

¹American Heritage School Boca/Delray, Florida, United States

Abstract. If the retinal images show evidences of abnormalities such as change in volume, diameter, and unusual spots in the retina, then there is a positive correlation to the diabetic progress. Mathematical and statistical theories behind the machine learning algorithms are powerful enough to detect signs of diabetes through retinal images. Several machine learning algorithms: Logistic Regression, Support Vector Machine, Random Forest, and Neural Networks were applied to predict whether images contain signs of diabetic retinopathy or not. After building the models, the computed results of these algorithms were compared by confusion matrixes, receiver operating characteristic curves, and Precision-Recall curves. The performance of the Support Vector Machine algorithm was the best since it had the highest true-positive rate, area under the curve for ROC curve, and area under the curve for Precision-Recall curve. This conclusion shows that the most complex algorithms doesn't always give the best performance, the final accuracy also depends on the dataset. For this dataset of retinal imaging, the Support Vector Machine algorithm achieved the best results. Detecting signs of diabetic retinopathy is helpful for detecting for diabetes since more than 60% of patients with diabetes have signs of diabetic retinopathy. Machine learning algorithms can speed up the process and improve the accuracy of diagnosis. When the method is reliable enough, it can be utilized in diabetes diagnosis directly in clinics. Current methods require going on diets and taking blood samples, which could be very time consuming and inconvenient. Using machine learning algorithms is fast and non-invasive compared to the existing methods. The purpose of this research was to build an optimized model by machine learning algorithms that can improve the diagnosis accuracy and classification of patients at high risk of diabetes using retinal imaging.

1 Introduction

The purpose of this research is to build an optimized model by machine learning algorithms that can improve the diagnosis accuracy and classification of patients at high risk of diabetes using retinal imaging & OCT scans. This method has multiple advantages over existing methods. Diabetes has brought unbearable pain and suffering to patients around the world. The cause of this disease is multi-factored yet not definite. However, the effects that lead to this disease is gradual, hence predictable, and preventable to a certain degree by adjustments at the early stage of the disease.

There are currently four officially sanctioned methods for diagnosing diabetes: 1) 2-month average blood sugar level test (HbA1C); 2) blood sugar level test after 8 hours without food (FPG); 3) Oral Glucose Tolerance Test where blood sugar level is measured 2 hours before and after taking a special sweet drink (OGTT); 4) Casual Plasma Glucose Test where blood sugar level is measured at any time but requires multiple measurements to confirm. [1] The HbA1C test monitors a patient's blood sugar control, it also measures the percentage of hemoglobin A1C which gives the patient's average glucose level over a two-to-three-month period.

Since it is difficult to know which A1C tests are accurate, a new laboratory-based test, the Tina-quant HbA1cDx assay, is developed and can be used to both accurately diagnose diabetes and monitor blood glucose control. But both the A1C test and the Tina-quant HbA1cDx assay have a few limitations. These tests are not allowed to be used to diagnose diabetes during pregnancy and should not be used to monitor diabetes in patients with hemoglobinopathy, hereditary spherocytosis, malignancies, or severe chronic, hepatic, and renal disease. [2] The Fasting Plasma Glucose Test (FPG) requires the patient to fast (nothing to eat or drink except water) for eight hours. Then blood samples are drawn from the patient and the plasma is combined with other substances to determine the amount of glucose in the plasma. To ensure the results of this test, the Oral Glucose Tolerance Test might be done subsequently. The Oral Glucose Tolerance Test (OGTT) measures how well the body handles a standard amount of glucose. Similarly, blood samples are needed from the patient. The patient's blood is drawn before and two hours after the patient drinks a large, premeasured beverage containing glucose, then the glucose levels contained in the patient's blood samples are compared, observing how well the body processed sugar. But because of the

* Corresponding author: jessiemin2005@gmail.com

poor reproducibility and limited use in clinical practice, the OGTT test is not recommended to be used routinely to diagnose diabetes. [3] All these methods involve taking blood samples which is invasive where the tests and the preparation are also lengthy, requiring hours to months of time as well as tedious labor from medical examiners. Furthermore, many require multiple tests to confirm as blood sugar level fluctuates widely based on diet, time of day, and other factors.

On the other hand, the impacts on organs by diabetes are evident and gradual. Occurring in more than 60% of patients with diabetes, diabetic retinopathy is now the leading cause of blindness in the US [4]. Blood sugar irregularities cause certain damages to the retina, the light-sensitive membrane at the back of the eyes that is crucial for clear vision. Even if a person is not at the full stage of diabetes yet, the retina may have already sustained various degrees of deterioration. Retinal damage has two forms: non-proliferative and proliferative. Non-proliferative retinopathy is the early stage which only has mild symptoms. In this stage the blood vessels in the retina are weakened. The proliferative retinopathy is the more advanced form with circulation problems depriving the retina of oxygen. Moreover, weak blood vessels can start to grow in the retina and leak into the back of the eye, clouding vision. Diabetic Retinopathy causes the eye to have abnormal blood vessels and “cotton wool” spots [4]. By examining retinal imaging and OCT scans for signs of damage can assist in diabetes diagnosis and prediction.

Currently retinal imaging and OCT scans are mainly used for eye disease diagnosis and not diabetes diagnosis. The reading of the images is done by human doctors who may not notice minute anomalies. This provides opportunities for the computational power of Artificial Intelligence, specifically machine learning algorithms to speed up the reading and improve the accuracy of diagnosis. When the method is reliable enough, it can be utilized in diabetes diagnosis directly in clinics. It is fast and non-invasive compared to the existing methods. Most retinal imaging machines now use digital imaging, allowing for taken images to be viewed instantaneously. This makes the process more convenient because of immediate determination instead of lengthy preparations and results waiting for patients. [5]

Google has done pioneering work on diabetic retinopathy with their Automated Retinal Disease Assessment (ARDA) tool. In the lab, this system analyzes images to indicate the condition and obtained more than 90% accuracy. But it would only go so far in the lab, therefore, it will be used in Thailand this year to screen diabetic patients for blindness prevention. [6] The developed algorithm is complex and requires much knowledge in computer vision so this research will strive to simplify certain aspects of the algorithm without affecting accuracy.

This research will also push this method to general diabetic prediction at early stages instead of being restricted in the eye disease diagnosis area. Furthermore, this research will try to aggregate certain geo, social, and habitual datasets to enhance confidence of the prediction. If the retinal imaging shows a display of abnormalities

such as change in volume, diameter, and unusual spots in the retina, then there is a positive correlation to the diabetic progress. Mathematical and statistical theories behind the machine learning algorithms are powerful enough to detect signs of diabetes through retinal images.

2 Dataset

This data set contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. The total number of patients is 1151 in this dataset. The types of features from retina images are listed as follows:

- Image quality assessment
- Pre-screen of retinal abnormality
- Microaneurysm (MA) detection at different confidence interval
- Microaneurysm (MA) detection at different confidence interval for exudates
- Euclidean distance of the centre of macula to optic disc
- Diameter of the optic disc

This is the link to this public dataset:
<https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debreceen+Data+Set>

3 Methods

3.1 Logistic Regression

The Logistic regression model is used to predict the probability of dichotomous outcomes that can only have a value of 0 or 1 (non-success or success, respectively). Logistic Regression predictions are based on one or more predictors. The Logistic Regression model is shown in equation (1):

$$g(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (1)$$

$$g(z) = \frac{1}{1 + e^{-z}} \quad (2)$$

which is also known as Sigmoid Function where $g(x)$ or $g(z)$ is the probability of the positive outcome (usually as 1). There are multiple ways to train a logistic regression model which would attempt to make the “s-shaped” line to fit the data and finding the decision boundaries. Methods such as iterative optimization algorithms called Gradient Descent, Conjugant Gradient, BFGS, and L-BFGS are useful for training logistic regression models. The Cost/Loss Function is used to measure the performance of a classification model whose output is a probability value between 0 and 1. The loss function is in equation (3):

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_{\theta}(x^{(i)}), y^{(i)}) \quad (3)$$

And the cost function is shown as follows:

$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases} \quad (4)$$

In total, equation (5) shows the simplified form after combining these two formulations

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log h_{\theta}(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_{\theta}(x^{(i)})) \right] \quad (5)$$

The optimization algorithm, Gradient Descent, is used to minimize the cost. Gradient Descent is applied by repeating equation (6):

$$\theta_j: \theta_j - \alpha \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} \quad (6)$$

where α is the learning rate. Overfitting is a common issue in Logistic Regression, where the model is too specific with the application of too many features, it would fail to generalize to new examples. To prevent overfitting, it may be helpful to reduce the number of features or apply regularization functions.

3.2 Support Vector Machine

The goal of the Support Vector Machine is to find the decision boundary in the middle of the two support vectors created based on the outermost data point of distinct features. The lost function of Support Vector Machine is shown in equation (7)

$$\min_{\theta} C \sum_{i=1}^m [y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T x^{(i)})] + \frac{1}{2} \sum_{i=1}^n \theta_i^2 \quad (7)$$

with the hypothesis:

$$h_{\theta}(x) \begin{cases} 1 & \text{if } \theta^T x \geq 1 \\ 0 & \text{if } \theta^T x \leq -1 \end{cases} \quad (8)$$

There is a slight difference between SVM model and the Logistic Regression model: if $y=1$, we want $\theta^T x \geq 1$ (not just ≥ 0), and if $y=0$, we want $\theta^T x \leq -1$ (not just < 0).

3.3 Decision Tree

Decision trees can be applied to both regression and classification problems. A decision tree is very easy to explain since it can visually represent decision-making models. Some people believe that decision trees can closely mirror human decision-making processes. A decision tree is drawn upside-down with the roots at the top. Each split is called an internal node while each end is called a terminal node/leaf. A decision tree can be partitioned into a corresponding feature space, where each rectangular region is a leaf of the decision tree. We choose simple rectangular models since it easily interprets the resulting predictive model. The steps for building a regression tree are:

- Divide the predictor space from the set of possible values for $X_1, X_2, X_3, \dots, X_p$ into J distinct and non-overlapping regions $R_1, R_2, R_3, \dots, R_J$.
- For every observation that falls into the region R_j , we make the same prediction.

The goal is to find the regions that minimizes the RSS which is given by

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - y_{R_j})^2 \quad (9)$$

Where y_{R_j} is the mean response for the training observations within the j^{th} region. This is a *greedy* approach since for each step in the top-down process, the best split is made that step. This approach is also known as *recursive binary splitting*. Tree Pruning is a commonly used technique when the result decision tree model is overfitting. When the decision tree is too complex, it needs pruning to become a smaller tree. A classification tree is very similar to a regression tree, except classification trees are used to predict a qualitative response. We predict that each observation belongs to the most commonly occurring class. RSS cannot be obtained in classification tree, instead, we calculate the classification error rate by using the fraction of the training observations in the region that do not belong to the most common class:

$$E = 1 - \max(p_{mk}) \quad (10)$$

Where p_{mk} represents the proportion of training observations in the m_{th} region that are from the k_{th} class. There are two other methods used to calculate error rates are Gini Index and Cross Entropy. The Gini Index and Cross Entropy both measure the total variance across K classes, also known as the node purity. They are similar since they will take on a small value if the m_{th} node is pure.

3.4 Random Forest

Bootstrap aggregating, also known as bagging, is used for reducing the variance without increasing the bias by a lot. This is done by taking the average of the resulting predictions from separate prediction models built by each training set. We average all the predictions to obtain

$$f_{bag}(x) = \frac{1}{B} \sum_{b=1}^B f^b(x) \quad (11)$$

which can be applied to regression trees. For classification trees, we record the class predicted by each of the B trees, and take the majority vote; the overall prediction is the most commonly occurring class. The feature importance factor is another critical factor in the Random Forest algorithm. The Gini impurity is to calculate the fraction of positive and negative samples in each node split. The Gini importance factor is therefore calculated by the decrease of the Gini impurity for each feature.

3.5 Neural Network

Neural networks are one set of algorithms that mimic how neurons function in human brains. Like a neuron, artificial neural networks have many connected layers of inputs and outputs. Neural networks can learn much

more complex functions and nonlinear decision boundaries since they can have many layers, making them more flexible to use. A neuron can be a binary logistic regression unit with f , a nonlinear activation function (Sigmoid); w , weights; and b , bias with the function:

$$h_{w,b}(x) = f(w^T x + b) \quad (12)$$

Neural network has three major components, the input layer, hidden layer, and output layer. The hidden layers can have multiple layers with different number of nodes. These neural networks are simply formed by connecting multiple neurons together. The inputs feedforward through the neurons to get to the outputs at the end. Training a network is trying to minimize the loss since lower loss means better predictions. The loss function for classification is:

$$J(\theta) = -\frac{1}{m} [\sum_{i=1}^m \sum_{k=1}^K y_k^{(i)} \log(h_{\theta}(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - (h_{\theta}(x^{(i)}))_k)] + \frac{\lambda}{2m} \sum_{l=1}^{L-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (\theta_{ji}^{(l)})^2 \quad (13)$$

Then we use backpropagation, calculating the partial derivative's error gradient by working backwards, to adjust the weight and bias parameters. There are different types of complicated neural networks used now, such as Convolution Neural Networks (CNNs) used for image detection and computer vision and Recurrent Neural Network used for Natural Language Processing are both very powerful.

4 Results

Figure 1 shows the distribution of the labels which are 0 and 1. 0/negative meaning there is no signs of diabetes present while 1/positive means signs of diabetes are detected. In this histogram, it shows the number of 0's to be about 550 and the number of 1's to be around 600. These numbers show this is a balanced dataset.

Fig. 1. Distribution of Labels.

Table 1. Comparison of ROCAUC and PRAUC's mean values with Standard Deviation values in parentheses.

	ROC AUC (STD)	PR AUC (STD)
Logistic Regression	0.8225 (0.0304)	0.8675 (0.0238)
Support Vector Machine	0.8319 (0.0410)	0.8702 (0.0290)
Random Forest	0.7689 (0.0413)	0.8109 (0.0271)
Neural Network	0.7924 (0.0639)	0.8428 (0.0421)

This table shows the mean values of the area under the curve for ROC curves and Precision-Recall curves; and the standard deviation values in parentheses. By comparing these values, the Support Vector Machine had the greatest mean ROCAUC value of 0.8319, followed by 0.8225 (Logistic Regression) > 0.7924 (Neural Network) > 0.7689 (Random Forest). The Support Vector Machine also had the largest mean PRAUC value of 0.8702, followed by 0.8675 (Logistic Regression) > 0.8428 (Neural Network) > 0.8109 (Random Forest).

Fig. 2. Comparison of Confusion Matrixes.

The confusion matrixes are tables that are used to describe the performances of the four different types of classification models with true positives, true negatives, false positives, and false negatives. Figure 1 shows the confusion matrixes calculated from Logistic Regression, Support Vector Machine, Random Forest, and Neural Networks. According to the results, Support Vector Machine achieved the best performance with the true-positive rate of 0.71, the false-positive rate of 0.19, precision of 0.81, and recall of 0.71. Logistic Regression has similar true-negative values but 24 less than Support Vector Machine's true-positive value. The true-negative value of Random Forest is less than Support Vector Machine's by 53, which is a relatively large difference. Neural Network's true-positive value is 19 less than Support Vectors Machine's, which shows that the Support Vector Machine evaluated the best results.

For the comparison of ROC curves, the larger the area-under-the-curve is, the better the model performed.

Figure 3 compares the ROC curves of four machine learning algorithms mentioned before, it is shown that the curve line of the Support Vector Machine is above all the other model's curve lines. Support Vector Machine also has the largest AUC value of 0.83 which is greater than 0.82 (Logistic Regression) > 0.79 (Neural Network) > 0.77 (Random Forest).

Fig. 3. Comparison of ROC Curves.

For the comparison of Precision-Recall curves, the larger the area-under-the-curve is, the better the model performed. Figure 4 compares the Precision-Recall curves of Logistic Regression, Support Vector Machine, Random Forest, and Neural Network, it is shown that the curve line of the Support Vector Machine is above all the

other model's curve lines. Support Vector Machine also has the largest AUC value of 0.87 which is like Logistic Regression's AUC of 0.87 by rounding to the nearest hundredth, but greater than 0.84 (Neural Network) > 0.81 (Random Forest).

Fig. 4. Comparison of Precision-Recall Curves.

Figure 5 compares the feature importance factor used specifically in Random Forest. The feature importance factor is computed by the Gini Importance Factor which is normalized to 1. The feature "microaneurysm (MA) detection at different confidence interval for exudates" has the highest importance value at about 0.5, while the feature "image quality assessment" has the lowest importance value close to 0.

Fig. 5. Feature importance factors for Random Forest.

5 Discussion

5.1 Conclusion

In this paper, Logistic Regression, Support Vector Machine, Random Forest, and Neural Network algorithms were applied to classify whether images contain signs of diabetic retinopathy or not. The powerful machine learning algorithms were able to detect signs of diabetic retinopathy through retinal imaging. The performance of each model is concluded by confusion matrixes based on their true positive rate and false positive rate (Figure 2), receiver operating characteristic curve (Figure 3), and Precision-Recall

curve (Figure 4). According to the results, Support Vector Machine achieved the best performance with the true-positive rate of 0.71, the false-positive rate of 0.19, precision of 0.81, and recall of 0.71. Logistic Regression has similar true-negative values but 24 less than Support Vector Machine's true-positive value. The true-negative value of Random Forest is less than Support Vector Machine's by 53, which is a relatively large difference. Neural Network's true-positive value is 19 less than Support Vectors Machine's, which shows that the Support Vector Machine evaluated the best results. Comparing the ROC curves, the curve line of the Support Vector Machine is above all the other model's curve lines. Support Vector Machine also has the largest AUC (area under the curve) value of 0.83 which is greater than 0.82 (Logistic Regression) > 0.79 (Neural

Network) > 0.77 (Random Forest). By comparing the Precision-Recall curves, Support Vector Machine also has the largest AUC value of 0.87 which is like Logistic Regression's AUC of 0.87 by rounding to the nearest hundredth, but greater than 0.84 (Neural Network) > 0.81 (Random Forest). This conclusion shows that most complex algorithms doesn't always work the best for all cases, the algorithm must fit the dataset.

The hypothesis, if the retinal imaging shows a display of abnormalities such as change in volume, diameter, and unusual spots in the retina, then there is a positive correlation to the diabetic progress, was best supported by the results from four machine learning algorithms used in this study: Logistic Regression algorithm, Support Vector Machine, Random Forest algorithm, and Neural Network algorithm. Support Vector Machine algorithm generated the best performance than others.

5.2 Applications

This research of detecting signs of diabetic retinopathy using machine learning algorithms would be important to the modern society since there are many who struggle with diabetes. Detecting signs of diabetic retinopathy is helpful for detecting for diabetes since more than 60% of patients with diabetes have signs of diabetic retinopathy. Machine learning algorithms can speed up the process and improve the accuracy of diagnosis. When the method is reliable enough, it can be utilized in diabetes diagnosis directly in clinics. Current methods require going on diets and taking blood samples, which could be very time consuming and inconvenient. Using machine learning algorithms is fast and non-invasive compared to the existing methods.

5.3 Limitations

The main limitation of this research is the limited amount of data that were able to be used for the algorithms. This data set was only gathered from one location, meaning that people from different locations may differ. This could be improved by combining datasets from different locations. The retinal images taken by the current technology may not have been very clear or some may have low resolution. These images could be a limitation since there are large amounts of images being used, and it is a difficulty to pick out specific images that may have low resolution, so they remained in the dataset that was used.

5.4 Error Analysis

There may be errors in the taken retinal images since there are large amounts of photos used. It is a challenge to obtain perfect resolution and shapes in each photo used in the dataset. There may be a few photos with unclear spots which may have caused error in the dataset. There may also be slight errors in the algorithms since there are random factors. Parameters could've also been chosen differently based on each model. There is no

general error in the algorithm since it has been checked over many times. The calculations of the models also did not have issues since it was calculated by the computer.

5.5 Error Analysis

This research could be improved in the future by many ways. More advanced machine learning algorithms could be applied on more datasets. Not only more advanced algorithms could be used, but also more complex dataset could be gathered and used. The data could be gathered from different locations and different hospitals. The retinal images used for the datasets could also be improved by using more advanced technology and taking pictures with high resolution for each image. Future research could also be based on some different features gathered from the dataset. For future research, the algorithms could also be coded on different computers or programs that may improve the speed of calculations.

References

1. Belvoir Media Group. (2011, June). Identify diabetes early to prevent complications. *Healthy Years*, 8(6). Gale Virtual Reference Library. <https://link.gale.com/apps/doc/A258358658/GPS?u=delray3411&sid=GPS&xid=aed8c300>
2. World Health Organization. (2013). First A1c test labelled for diagnosing diabetes. *WHO Drug Information*, 27(2). Gale Virtual Reference Library. <https://link.gale.com/apps/doc/A342769554/GPS?u=delray3411&sid=GPS&xid=27025b35>.
3. Davidson, M. B. (2001). How do we diagnose diabetes and measure blood glucose control? View 1 (diagnosing): A clinical basis for the diagnosis of diabetes. *Diabetes Spectrum*, 14(2), 67-82.
4. Belvoir Media Group. (2019). Protect Your Sight if You Have Diabetes: As many as 45 percent of Americans with diabetes have a condition called diabetic retinopathy, but many are not aware of it. *Focus on Healthy Aging*, 22(11).
5. Mark Allen Group. (2004). Optometric imaging systems Part 3: Retinal imaging systems. *Optician*.
6. Heaven, W. D. (2020). Google's medical AI was super accurate in a lab. Real life was a different story. *MIT Technology Review*. <https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/>