Improving Diabetic Diagnosis and Prevention with Machine Learning on Retinal Imaging

Yushan Min
American Heritage School Boca/Delray

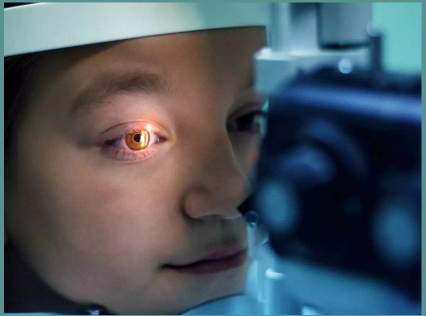
Abstract

The purpose of this research was to build an optimized model by machine learning algorithms that can improve the diagnosis accuracy of classifying patients at high risk of diabetes using retinal imaging. If the retinal imaging shows evidence 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 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 matrices and graphs. The performance of the Support Vector Machine algorithm was the best with a 75% accuracy. This conclusion shows that the most complex algorithms don't always give the best performance and the final accuracy also depends on the dataset. Detecting signs of diabetic retinopathy is helpful for detecting diabetes since more than 45% of American 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.

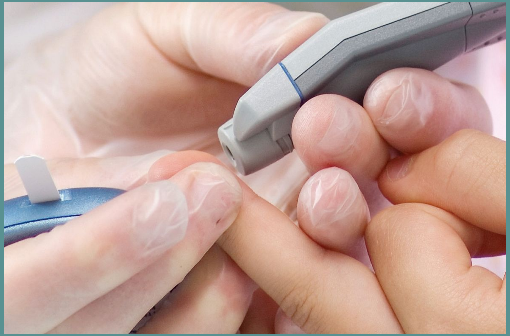
Research Question

Can machine learning algorithms speed up the process and improve the accuracy of diabetes diagnosis by using retinal imaging?

Background Information: Detecting Diabetes

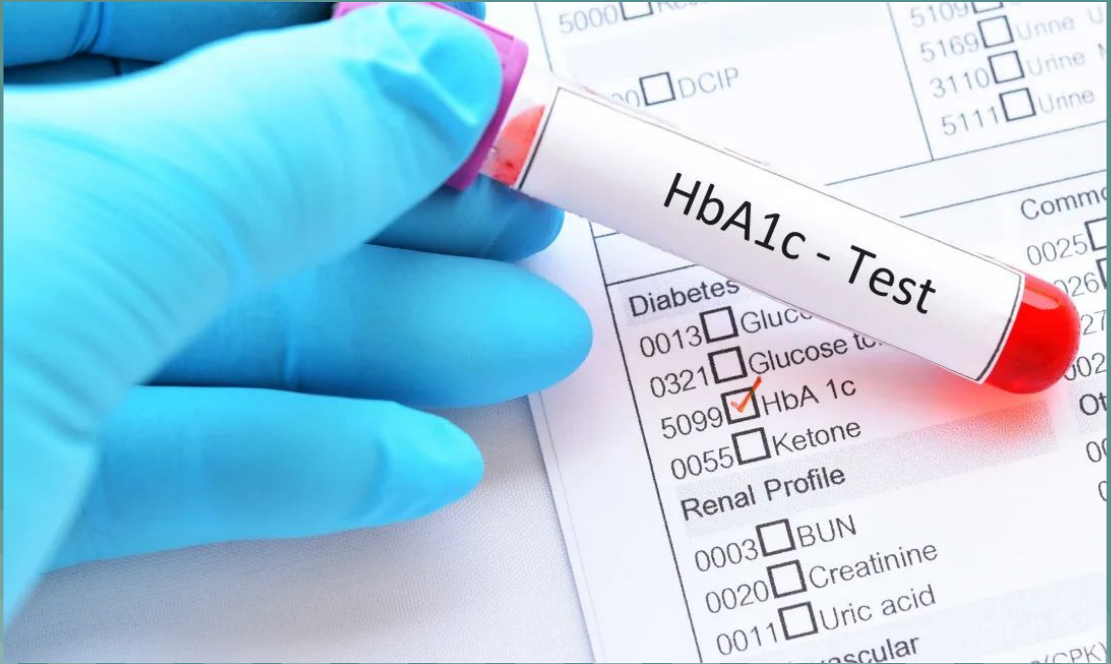


https://health.economictimes.indiatimes.com/news/diagnostics/artificial-intelligence-can-accurately-detect-diabetic-eye/71607332



https://decisionsindentistry.com/article/detecting-type-2-diabetes-pediatric-populations/

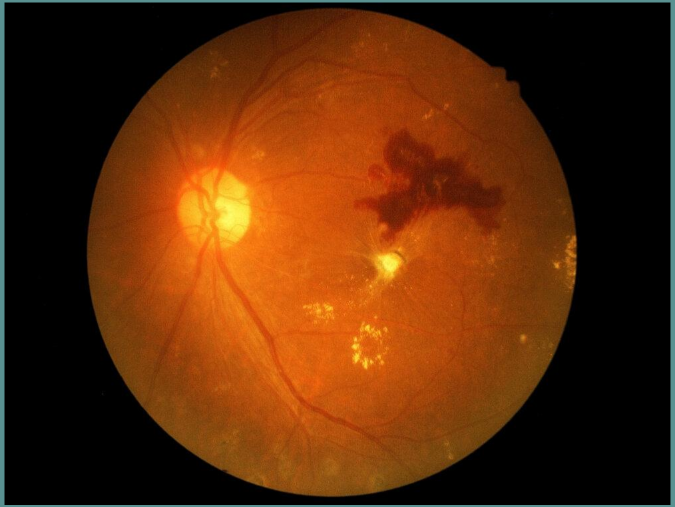
- HbA1C Test
- Oral Glucose Tolerance Test
- Fasting Plasma Glucose Test
- Casual Plasma Glucose Test



https://www.diabetes.co.uk/hba1c-meters/

Background Information: Diabetic Retinopathy

(Unusual Spots)



Messidor Images Dataset

http://www.adcis.net/en/third-party/messidor/

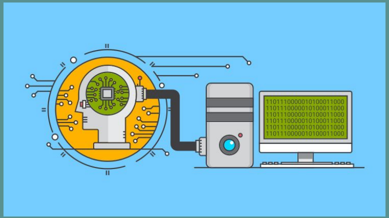
#1

Leading cause of legal blindness in the US

45%

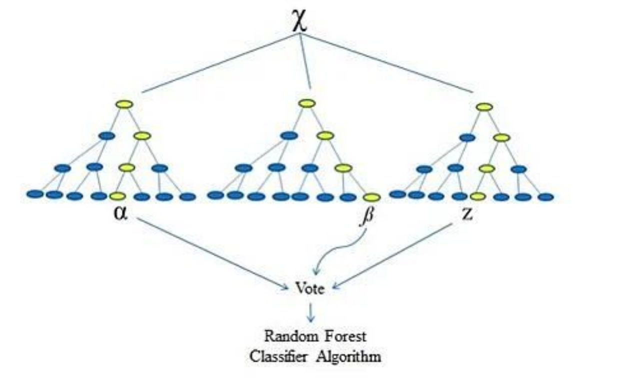
Percentage of Americans with diabetes with signs of DR

Background Information: Machine Learning Algorithms

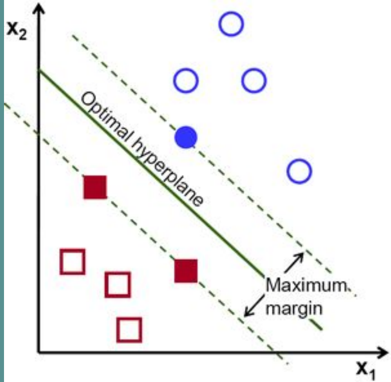


https://www.iberdrola.com/innovation/machine-learning-automatic-learning

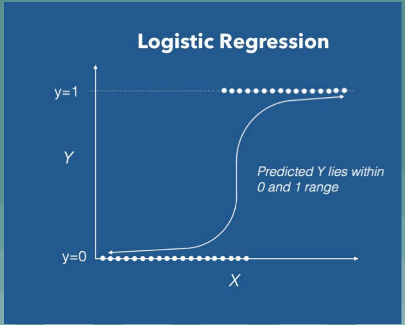
- Logistic Regression
- Support Vector Machine
- Random Forest
- Neural Networks



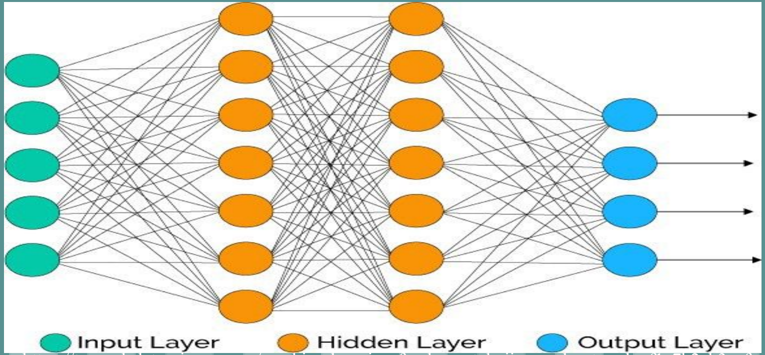
https://www.rxdatascience.com/blog/machine-learning-for-pharma-using-random-forest



https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47



https://jeppbautista.wordpress.com/2019/01/27/theory-to-application-logistic-regression-from-scratch-using-python/



https://towardsdatascience.com/machine-learning-fundamentals-ii-neural-networks-f1e7b2cb3eef

Data Collection

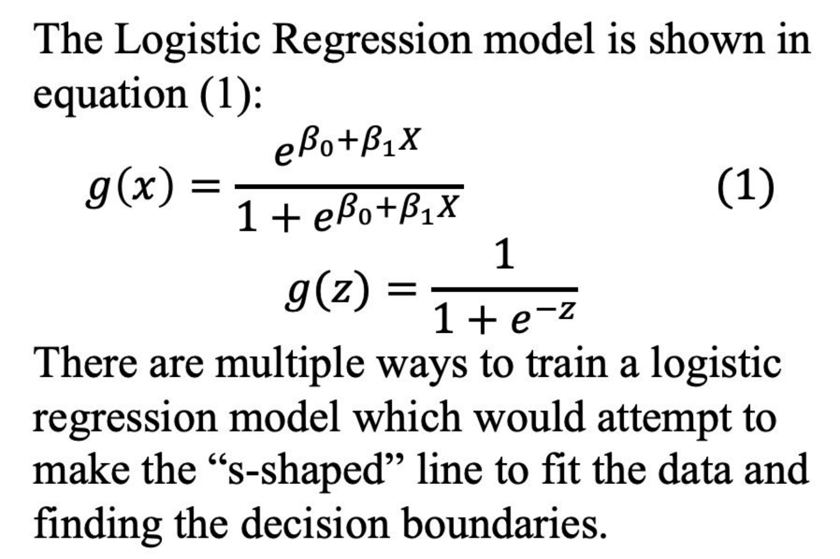
This data set contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. The total number of patients is 1151 in this dataset. The types of features from retina images are listed as follows:

- 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 center 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+Debrecen+Data+Set

Methods:
Logistic Regression & Support Vector Machine

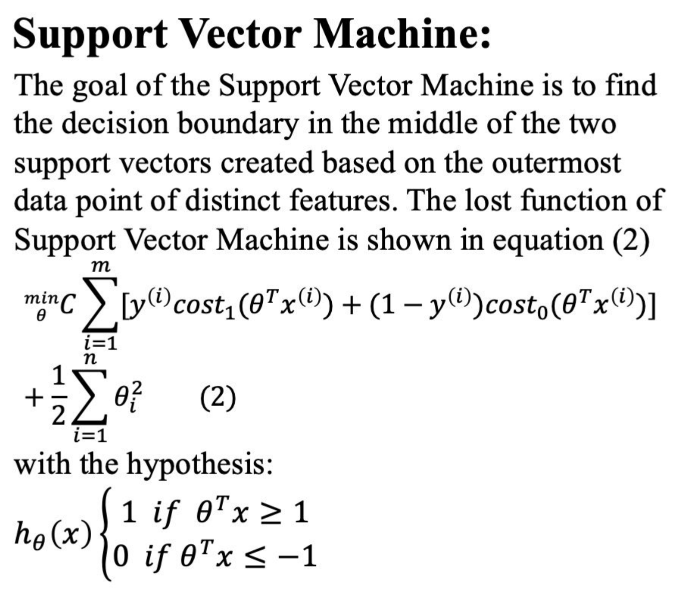


The Logistic Regression model is shown in equation (1):

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

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

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.



Support Vector Machine:

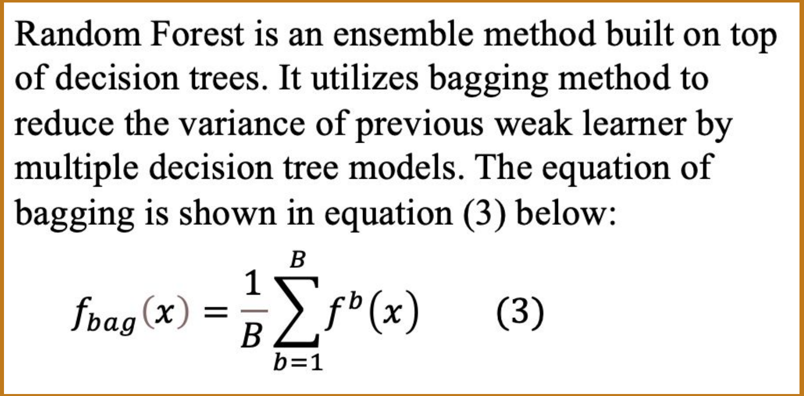
The goal of the Support Vector Machine is to find the decision boundary in the middle of the two support vectors created based on the outermost data point of distinct features. The lost function of Support Vector Machine is shown in equation (2)

$$\min_{\theta} C \sum_{i=1}^{m} [y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)})]$$
$$+ \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \qquad (2)$$

with the hypothesis:

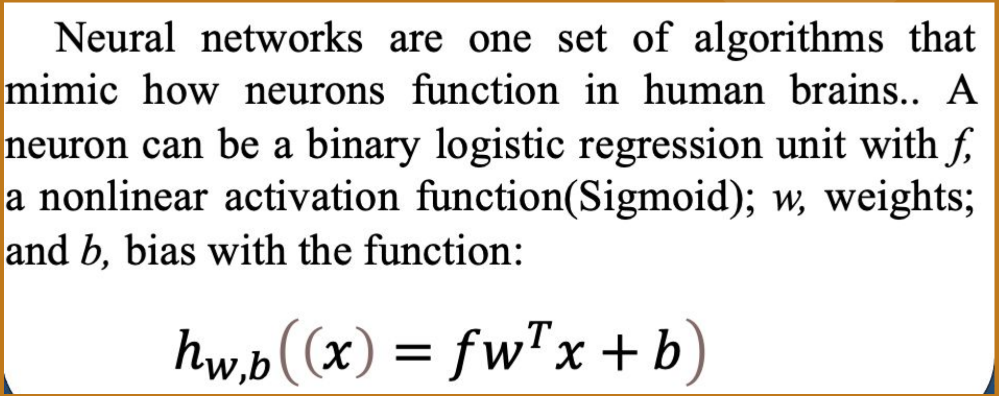
$$h_\theta(x) \begin{cases} 1 \;\; if \;\; \theta^T x \geq 1 \\ 0 \;\; if \;\; \theta^T x \leq -1 \end{cases}$$

Methods:
Random Forest & Neural Networks



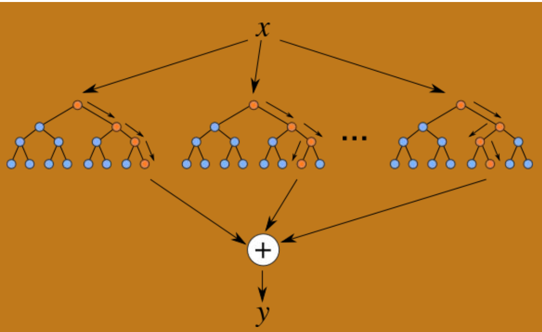
Random Forest is an ensemble method built on top of decision trees. It utilizes bagging method to reduce the variance of previous weak learner by multiple decision tree models. The equation of bagging is shown in equation (3) below:

$$f_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{b}(x) \qquad (3)$$

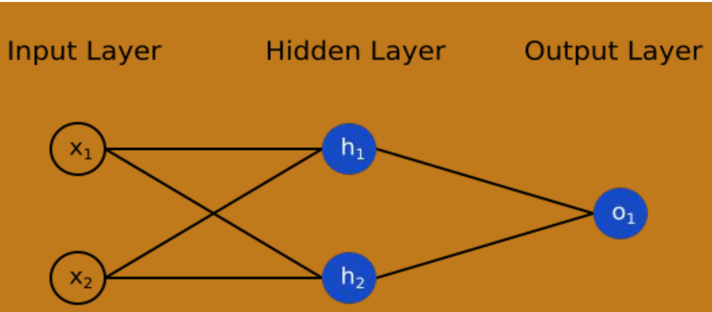


Neural networks are one set of algorithms that mimic how neurons function in human brains.. A neuron can be a binary logistic regression unit with *f*, a nonlinear activation function(Sigmoid); *w*, weights; and *b*, bias with the function:

$$h_{w,b}((x) = fw^{T}x + b)$$



https://levelup.gitconnected.com/random-forest-regression-209c0f354c84



https://towardsdatascience.com/machine-learning-for-beginners-an-introduction-to-neural-networks-d49f22d238f9

Data Analysis:

	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)

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

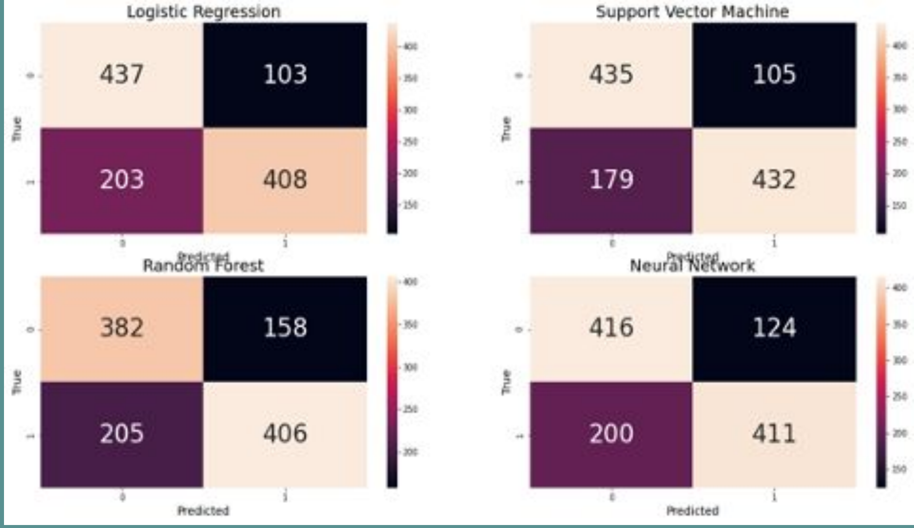


Figure 1. Comparison of Confusion Matrices

Data Analysis:

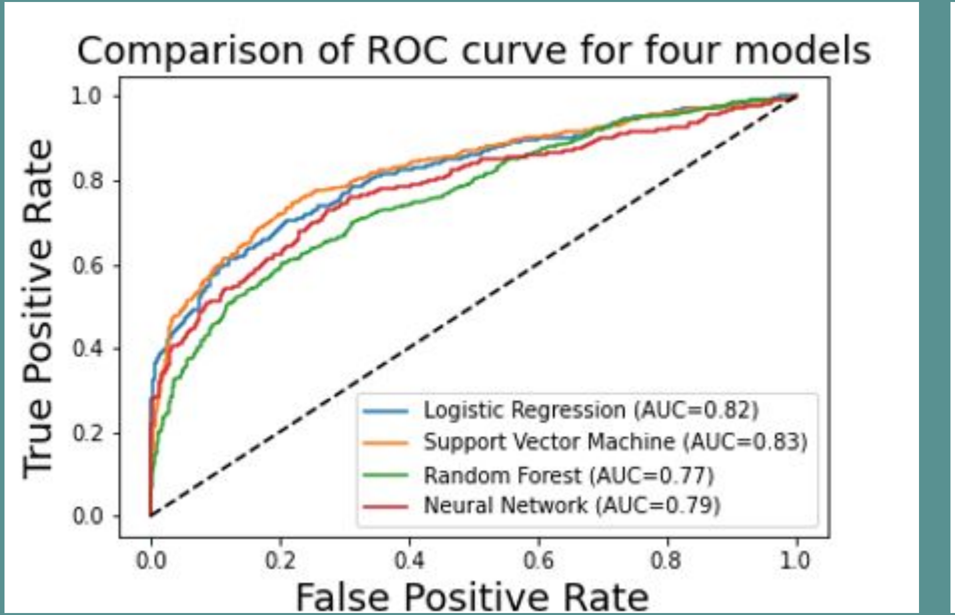


Figure 2. Comparison of ROC Curves

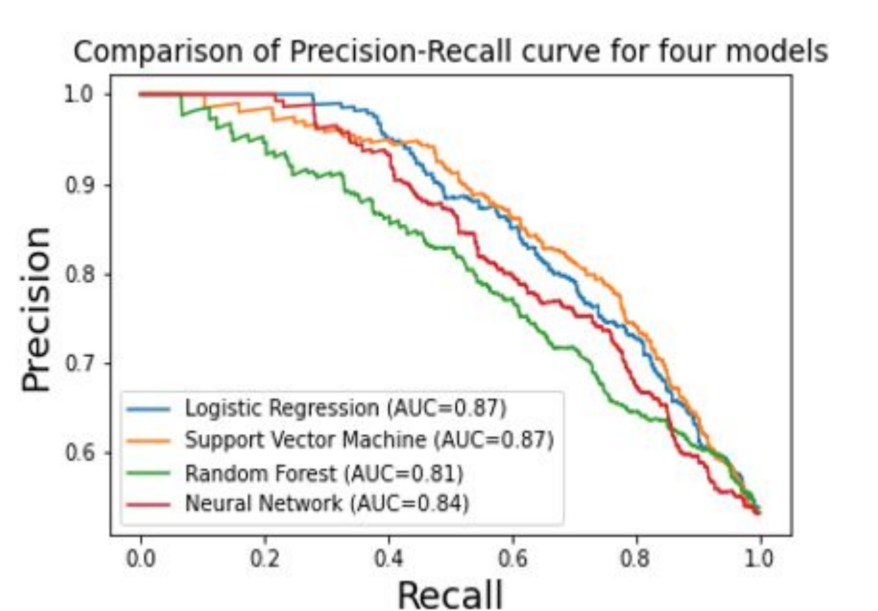


Figure 3. Comparison of Precision-Recall Curves

Discussions

In this project, 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 hypothesis, if the retinal imaging show 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 Support Vector Machine used in this study. The results achieved were equal to or greater than the accuracy of the current methods. This conclusion shows that most complex algorithms doesn’t always work the best for all cases, the algorithm has to fit the dataset. The main limitation of this research is the limited amount of data that were able to be used for the algorithms. 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. 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.