An Intelligent System for Early Detection of Eye Diseases that Lead to Irreversible Vision Loss

Navya Ramakrishnan

Jasper High School, 6800 Archgate Dr, Plano, Texas, 75024, U.S.A.; nr24.ramk@gmail.com

ABSTRACT: Based on a 2018 World Health Organization survey an estimated 217 million people live with moderate to severe vision impairment. Glaucoma and diabetic retinopathy (DR) are two eye diseases that cause a large percentage of vision impairment. Glaucoma affects 60.5 million people globally and is a leading cause of blindness. Similarly, DR, an eye condition that affects diabetics, is the fastest growing cause of blindness with nearly 415 million at risk worldwide. Treatments of these diseases are incredibly tricky and require sophisticated instruments, so the best countermeasure is early detection. One paradigm of early detection systems is statistical inference, specifically machine learning. In this work, the effectiveness of machine learning classifiers on early detection of glaucoma and diabetic retinopathy is demonstrated. On a dataset of over 300 glaucoma patients, the best model, the random forest classifier, showed an accuracy of 97% with equally impressive precision, recall, and F1 score. It is also shown that ensemble methods work particularly well for prediction of diabetic retinopathy. The models are packaged into a website where doctors can input patients' features to make quick and efficient predictions to aid in their diagnosis. This system can improve detection capabilities of these debilitative eye diseases and prevent vision impairment.

KEYWORDS: Vision impairment; Glaucoma; Diabetic Retinopathy; Machine learning; Ensemble learning.

Introduction

An estimated 217 million people have moderate to severe vision impairment, and 36 million people are blind according to 2018 World Health Organization data.[1] Glaucoma, a leading cause of blindness, clogs the eye's drainage system which causes pressure to build up inside the eye. High pressure damages the optic nerve which is responsible for carrying visual information to the brain. Optic nerve damage can lead to blindness.[2] It is important to explore this issue because glaucoma affects 60.5 million people globally.[3] As vision loss from glaucoma is irreversible, early detection and timely treatment are critical to managing the disease. The diagnosis of glaucoma in its early stages is challenging. Misdiagnosis can lead to failure in identifying those with the condition until significant vision loss has occurred.[4,5] The regular eye checkup for early detection takes a great deal of time for ophthalmologists, which prompts the development of an automatic disease diagnosis system.

Similarly, Diabetic Retinopathy (DR) is an eye condition that affects people with diabetes and is the fastest growing cause of blindness.[6] DR occurs when high blood sugar levels damage blood vessels in the retina. These blood vessels can swell, cause leakage, or, in some cases, completely close, preventing blood from passing through. Sometimes new, abnormal blood vessels grow on the retina. All of these changes can cause vision loss.[7] Doctors use retinal fundus images to diagnose DR. The manual evaluation of these images is slow and demands substantial resources. Regular screening is required for early detection of the disease which makes automatic detection techniques a very attractive alternative or supplement to current medical practices.

Machine learning, a branch of artificial intelligence, is defined as a set of methods that can learn from data, detect patterns in data, and make decisions with minimal human intervention.[8] Classification problems are a particular subclass of machine learning problems in which models use data to distinguish between distinct classes or categories. Hence, it is suitable for diagnosis of glaucoma or diabetic retinopathy. In the last few decades, many studies have been done on the automation of detection and prediction of glaucoma using different machine learning techniques as well as deep learning. In one of the studies, a comparison of the performance of the three machine learning classification models, neural network (NN), naïve Bayes (NB), and support vector machine (SVM), was done.[9] This study concluded that the NN had the best performance with an accuracy of 87.8% using only nine ocular parameters. The selected parameters enabled the trained NN to classify glaucomatous optical discs with relatively high performance without requiring color fundus images. Other studies have used deep learning methods to detect glaucoma in colored retinal fundus images. In one such study, the features were extracted from the raw images by Convolutional Neural Network (CNN) and fed to the SVM to classify the images into normal or abnormal. CNN distinguishes between normal and glaucomatous patterns for diagnostic decisions with an accuracy, specificity, and sensitivity of 88.2%, 90.8%, and 85.0% respectively.[10]

Goal

The goal of this study is to develop a software model that would help detect severe eye diseases like glaucoma and diabetic retinopathy at their early stages using different machine learning algorithms. The developed software would be used in

DOI: 10.36838/v2i4.6

a web application with strong predictive power which would be very helpful to clinicians.

Results and Discussion

Various classification models including Logistic Regression (LR), Linear Discriminant Analysis (LDA), k-nearest neighbor (KNN), Classification and Regression Trees (CART), Gaussian Naive Bayes (NB), Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting (GB) were tested to develop the software model. The patient dataset contained the information about patients as well as healthy individuals. The dataset was shuffled and 70% of the dataset was used as training dataset and the remaining 30% was test dataset. To determine the best models, each model was evaluated on the unseen test dataset using various metrics including accuracy, precision, recall, and F1 score. The best performing classification model was selected to create the web application.

1. Testing the Glaucoma Dataset

Using the training data, a 10-fold cross validation was performed to train and evaluate the classification models. The average of the results obtained from ten iterations of cross validation were given in Table 1 for each classifier. The classification models were then used to validate the new, unseen test dataset and the predictions of each model was compared with the ground truth diagnosis. A summary of these results is displayed in Figure 1.

Table 1. Performance of the Models with glaucoma training dataset (10-fold Cross Validation).

Classifier	Average value obtained from 10 iterations of cross validation			
	Accuracy	Precision	Recall	F1 score
LR	0.92	0.95	0.92	0.93
LDA	0.90	0.94	0.89	0.92
KNN	0.90	0.93	0.90	0.92
CART	0.92	0.94	0.93	0.93
NB	0.89	0.96	0.85	0.90
SVM	0.69	0.66	1.00	0.79
RF	0.93	0.93	0.93	0.94

The RF model showed the highest performance considering all four metrics. The high recall means the model has a high ability to identify patients with Glaucoma. It also means that the RF model shows a small false negative ratio. In the medical field, the false negative ratio is often more important than the false positive ratio. The ROC/AUC of each classification model is given in Figure 2. RF again showed a high value of AUC. RF builds multiple decision trees and merges them to get a more accurate and stable prediction.

2. Testing the Diabetic Retinopathy Dataset

Using the training data, a 10-fold cross validation was performed to train and evaluate the classification models. The results are given in Table 2. None of the classification models showed an excellent prediction accuracy of the DR data set which led to the author's attempt at using ensemble learning.

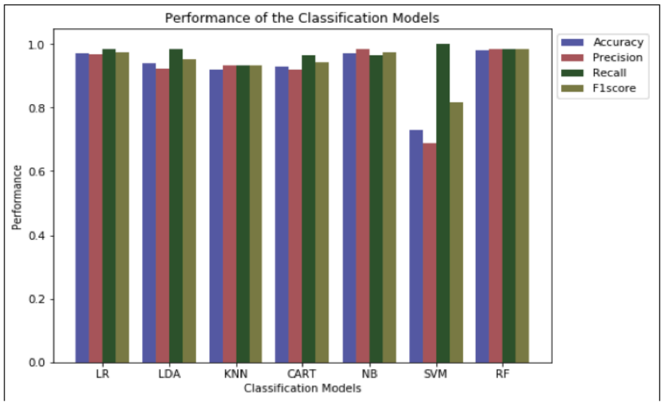


Figure 1. Performance of the models with Glaucoma Test Dataset.

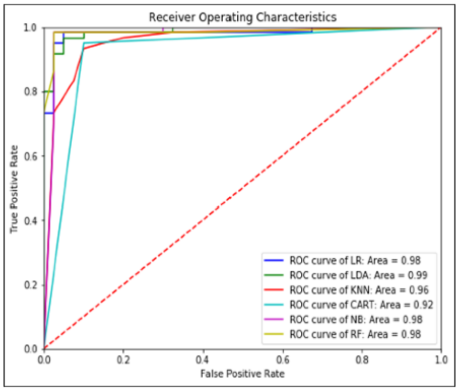


Figure 2. ROC/AUC with Glaucoma Test Dataset.

Table 2. Performance of the Models with the DR training dataset (10-fold Cross Validation).

Classifier	Average value obtained from 10 iterations of cross validation			
	Accuracy	Precision	Recall	F1 score
LR	0.74	0. 79	0.71	0.74
LDA	0.71	0.80	0.63	0.69
KNN	0.65	0.69	0.63	0.65
CART	0.58	0.60	0.63	0.60
NB	0.57	0.81	0.27	0.40
SVM	0.58	0.57	0.89	0.69
RF	0.65	0.72	0.58	0.64
GB	0.67	0.70	0.68	0.69

A voting classifier (VC) ensemble model was developed with the four best-performing base classifier models (LR, LDA, RF, and GB). The ensemble model was used to validate the test dataset and the predictions were compared with the ground truth diagnosis. The performance of the ensemble model and the base classifier models in validating the new, unseen test dataset are given in Figure 3. The ROC/AUC curve is plotted in Figure 4.

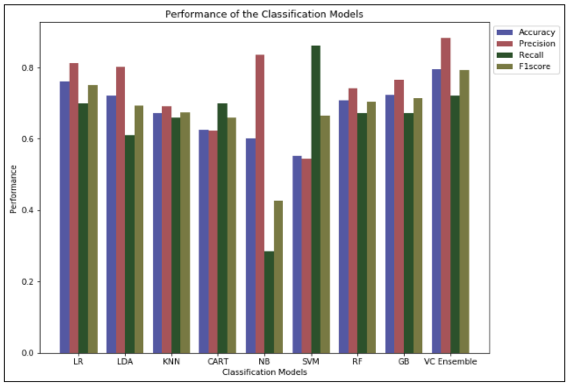


Figure 3. Performance of the models with DR Test Dataset.

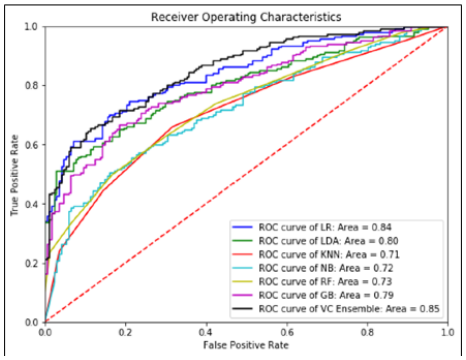


Figure 4. ROC/AUC of the models with DR Test Dataset.

3. Web Application Development

Using the highest-performing algorithms for the glaucoma dataset, RF, and diabetic retinopathy dataset, VC ensemble, a web application with an intuitive user interface was developed. The frontend of the web application was created with JavaScript and backend was with Python. The knowledge acquired by the best-performing model in the case of each disease was saved for the backend. Users can input a csv file with the patients' features and generate the output classification for each patient.

Methods

Performance Evaluation Criteria of Classification Models

A binary classification model classifies each data sample into one of two classes: a true and a false class. This gives rise to four possible classifications for each sample; a true positive, a true negative, a false positive, and a false negative. [11, 12]

1. True positive (TP): the patient has a disease and the prediction is positive.

2. False positive (FP): the patient does not have a disease, but the prediction is positive.

3. True negative (TN): the patient does not have a disease and the prediction is negative

4. False negative (FN): the patient has a disease, but the prediction is negative.

Metrics such as accuracy, precision, recall, F1 score, and Receiver Operating Characteristic (ROC) are used to evaluate the performance of the classification models in this research.

The accuracy of a diagnosis model refers to the model's ability to correctly identify patients with the disease and without the disease.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

The precision of a diagnosis model refers to the ratio of correctly predicted positive observations (disease) to the total predicted positive observations.

$$\text{Precision} = \frac{TP}{TP+FP}$$

The recall (sensitivity) of a diagnosis model refers to the ability of the test to correctly identify patients with the disease.

$$\text{Recall} = \frac{TP}{TP+FN}$$

The F1 score of a diagnosis model is the harmonic mean of precision and recall.

$$\text{F1 score} = 2 \times \frac{precision \bullet recall}{precision+recall}$$

Receiver Operating Characteristic (ROC)

ROC graphs are constructed by plotting the TP rate against the FP rate. The diagonal line from the bottom left corner to the top right corner represents the random classifier performance (Figure 5). The number of FP responses produced by a model mapped onto this line is equal to the number TP responses produced. Classifiers that fall in the region to the right of the random performance line have a performance worse than the random classifier, meaning it consistently produces more FP responses than TP responses.

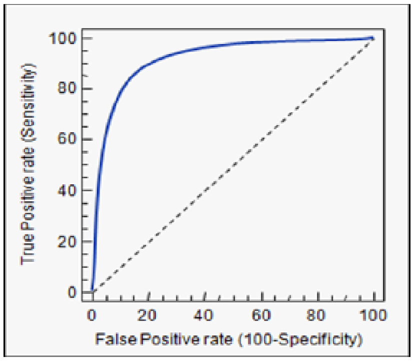


Figure 5. ROC Curve.[13]

The point in the top left corner denotes perfect classification: 100% TP rate and 0% FP rate.[11,12] The area under the curve (AUC) can have any value between 0 and 1. The area under the dashed line (ROC curve of a random classifier) is 0.5. When

AUC is higher than 0.5, the model exceeds the random classifier whereas the prediction model is perfect when AUC = 1.

The methodology used for this research is given in Figure 6. The following classification models were compared in their performance across glaucoma prediction and diabetic retinopathy prediction: LR, LDA, KNN, CART, GNB, SVM, RF, and GB. After training the model on the training dataset, the model is tested on a new unseen test dataset. The web application development centered on the highest performing algorithms from each task and allows users to input csv files with patient data for analysis. The website then generates the classification for each patient.

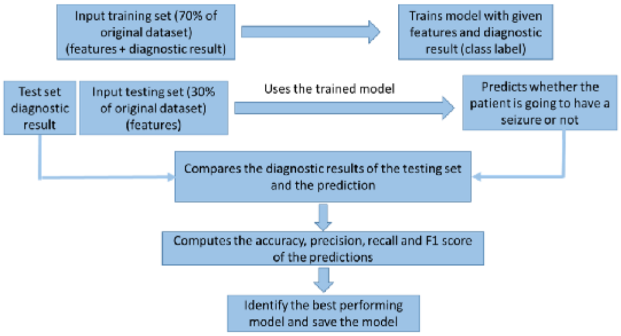

Input training set (70% of original dataset) (features + diagnostic result)
Trains model with given features and diagnostic result (class label)
Test set diagnostic result
Input testing set (30% of original dataset) (features)
Uses the trained model
Predicts whether the patient is going to have a seizure or not
Compares the diagnostic results of the testing set and the prediction
Computes the accuracy, precision, recall and F1 score of the predictions
Identify the best performing model and save the model

Figure 6. Methodology.

A 10-fold cross-validation is used to evaluate the performance of each model (Figure 7). The advantage of this method over repeated, random sub-sampling is that all observations are used for both training and validation and each observation is used for validation exactly once.[14]

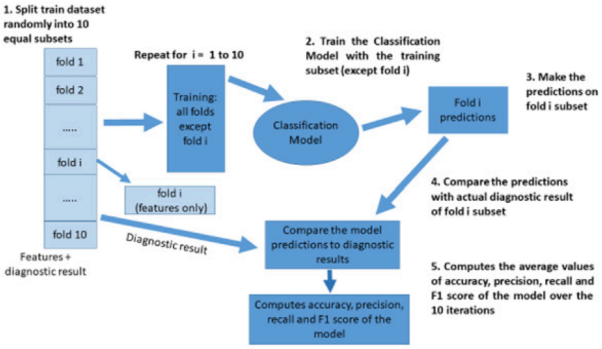

1. Split train dataset randomly into 10 equal subsets
fold 1
fold 2
.....
fold i
.....
fold 10
Features + diagnostic result
Repeat for i = 1 to 10
Training: all folds except fold i
fold i (features only)
Diagnostic result
2. Train the Classification Model with the training subset (except fold i)
Classification Model
Fold i predictions
3. Make the predictions on fold i subset
4. Compare the predictions with actual diagnostic result of fold i subset
Compare the model predictions to diagnostic results
Computes accuracy, precision, recall and F1 score of the model
5. Computes the average values of accuracy, precision, recall and F1 score of the model over the 10 iterations

Figure 7. 10-Fold Cross Validation

Ensemble Learning

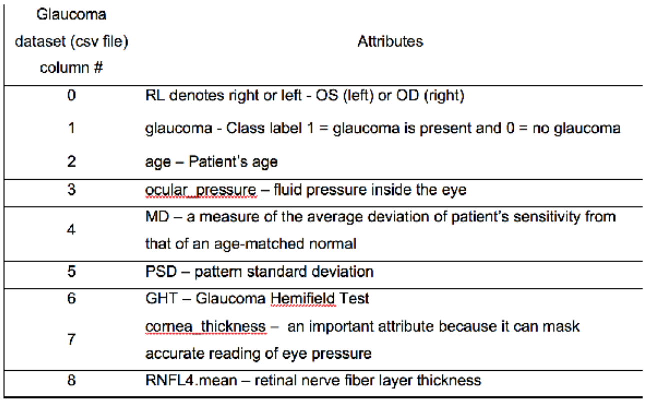
While a single model can capture the relationships within the data, several studies have found that single models have a ceiling in terms of performance.[20] Many practitioners of machine learning use ensemble learning to improve performance on particular tasks. An ensemble contains a number of learners which are usually known as base learners. Base learners are generated from training data by a base learning algorithm such as a decision tree, CART or other machine learning algorithm. The generalization ability of an ensemble is often much stronger than that of base learners. Ensemble learning is able to boost weak learners into strong learners which can make very accurate predictions. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm to improve performance over a single estimator. In contrast to ordinary machine learning approaches, which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses and combine them for prediction.

Typically, an ensemble is constructed in two steps. First, a number of base learners are produced, which can be generated in a parallel style or in a sequential style where the generation of a base learner has influence on the generation of subsequent learners. Then, the base learners are combined for use where among the most popular combination schemes are majority voting for classification and weighted averaging for regression. Generally, to get a good ensemble, the base learners should be as accurate and diverse as possible.[15,20]

Patient Datasets

1. Glaucoma Dataset: The database includes eye examination records of glaucoma and normal cases. The records contain Retinal Nerve Fiber Layer (RNFL) thickness, ocular pressure, patient age, and Visual Field (VF) test parameters (Table 3). There are two datasets available, a training set containing 399 cases and a test dataset containing 100 cases.[16,17] The training dataset is used to train and validate the various classification models. Then, each classification model is used to validate the test dataset and the model with the highest performance is identified.

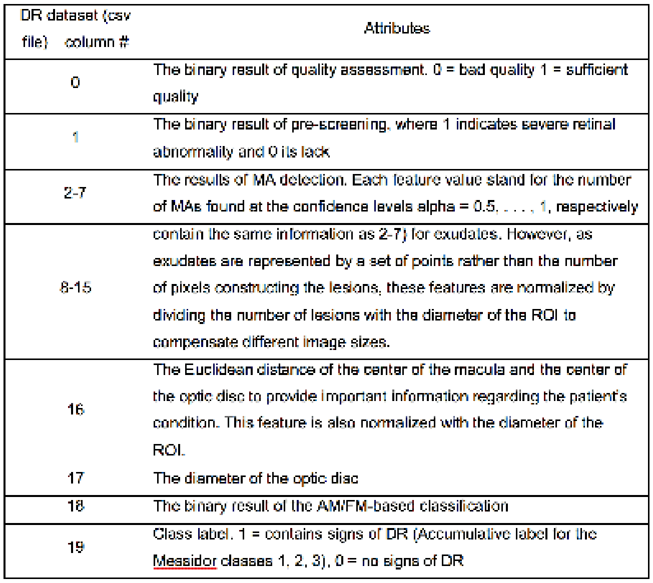
Table 3. Attribute Information of the Glaucoma dataset.



Glaucoma dataset (csv file) column #	Attributes
0	RL denotes right or left - OS (left) or OD (right)
1	glaucoma - Class label 1 = glaucoma is present and 0 = no glaucoma
2	age – Patient's age
3	ocular_pressure – fluid pressure inside the eye
4	MD – a measure of the average deviation of patient's sensitivity from that of an age-matched normal
5	PSD – pattern standard deviation
6	GHT – Glaucoma Hemifield Test
7	cornea_thickness – an important attribute because it can mask accurate reading of eye pressure
8	RNFL4.mean – retinal nerve fiber layer thickness

2. Diabetic Retinopathy Dataset: This dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. The dataset contains 1151 cases including examination records of both diabetic retinopathy and normal cases.[18] All features represent either a detected lesion, a descriptive feature of an anatomical part, or an image-level descriptor (Table 4). The features are extracted from the original image set, Messidor.[19]

Table 4. Attribute Information of the DR Dataset.



DR dataset (csv file) column #	Attributes
0	The binary result of quality assessment. 0 = bad quality 1 = sufficient quality
1	The binary result of pre-screening, where 1 indicates severe retinal abnormality and 0 its lack
2-7	The results of MA detection. Each feature value stand for the number of MAs found at the confidence levels alpha = 0.5, . . . , 1, respectively
8-15	contain the same information as 2-7) for exudates. However, as exudates are represented by a set of points rather than the number of pixels constructing the lesions, these features are normalized by dividing the number of lesions with the diameter of the ROI to compensate different image sizes.
16	The Euclidean distance of the center of the macula and the center of the optic disc to provide important information regarding the patient's condition. This feature is also normalized with the diameter of the ROI.
17	The diameter of the optic disc
18	The binary result of the AM/FM-based classification
19	Class label. 1 = contains signs of DR (Accumulative label for the Messidor classes 1, 2, 3), 0 = no signs of DR

Conclusion

Diagnosing glaucoma is challenging, especially in the early stages of the disease. It usually takes a very long time to determine whether a patient has glaucoma. The proposed model will assist in early detection of glaucoma during routine eye checkups. The classifications of patients after providing a csv file with patients' features, or test results, can be generated within a matter of seconds using the developed web application. Earlier detection allows glaucoma patients to undergo earlier treatment and reduces the likelihood of complications including blindness. The ensemble model could be used for early screening of diabetic retinopathy among diabetic patients. The web application would be a great support to doctors to evaluate more patients quickly. In the future, this type of web application could be integrated into hospitals and eye care centers for quick, reliable glaucoma and diabetic retinopathy testing that can potentially catch these diseases in their earliest stages. An automatic feature extraction scheme from the eye images could be developed in the future to use with the proposed model to improve the performance.

Acknowledgements

I would like to thank my parents for encouraging me throughout this project. I would also like to thank Dr. Ben Colston for explaining details about the eye diseases, examination, and diagnosis procedures. I am also grateful to my science teachers for their valuable suggestions and support.

References

1. Vision impairment and blindness. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

2. Funding Cutting-Edge Research and Educational Outreach. https://glaucomafoundation.org/aboutglaucoma/glaucoma/

3. Quigley, H. A.; Broman, A. T. The number of people with glaucoma worldwide in 2010 and 2020. http://www.ncbi.nlm.nih.gov/pubmed/16488940

4. What is Glaucoma? - All Articles. https://www.glaucoma.org/glaucoma/archive.php

5. Facts about Glaucoma. https://nei.nih.gov/health/glaucoma/glaucoma_facts

6. IDF diabetes atlas - 8th edition. https://www.diabetesatlas.org/

7. What Is Diabetic Retinopathy? https://www.aao.org/eye-health/diseases/what-is-diabetic-retinopathy

8. Machine Learning: What it is and why it matters. https://www.sas.com/en_us/insights/analytics/machine-learning.html

9. An, G.; Omodaka, K.; Tsuda, S.; Shiga, Y.; Takada, N.; Kikawa, T.; Nakazawa, T.; Yokota, H.; Akiba, M. Comparison of Machine-Learning Classification Models for Glaucoma Management. Journal of Healthcare Engineering 2018, 2018, 1–8.

10.B. AI-Bander, W. AI-Nuaimy, M. A. AI-Taee, and Y. Zheng, "Automated glaucoma diagnosis using deep learning approach," in Proceedings of 14th International Multi-Conference on Systems, Signals & Devices, Marrakesh, Morocco, March 2017.

11.A Review on Evaluation Metrics for Data Classification https://www.researchgate.net/publication/275224157_A_Review_on_Evaluation_Metrics_for_Data_Classification_Evaluations

12. Aditya Mishra. Metrics to Evaluate your Machine Learning Algorithm. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

13. Schoonjans, F. (2017, May 09). ROC curve analysis with MedCalc. https://www.medcalc.org/manual/roc-curves.php

14. Dr. Gary. "What does cross-validation do with models on each subset?" https://community.rapidminer.com/discussion/3943/what-does-cross-validation-do-with-models-on-each-subset

15. Zhi-Hua Zhous Publications - Nanjing University. https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/publication_toc.htm

16. Kim, S. J.; Cho, K. J. Development of machine learning models for diagnosis of glaucoma. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177726

17. Kim; Jin, K.; Cho. Data from: Development of machine learning models for diagnosis of glaucoma. https://datadryad.org/resource/doi:10.5061/dryad.q6ft5

18.https://archive.ics.uci.edu/ml/datasets/ Diabetic Retinopathy Debrecen Data Set

19.PATRY, G.; GAUTHIER, G.; LAY, B.; ROGER, J.; ELIE, D.; DON JON, A.; MAFFRE, H. Messidor. http://www.adcis.net/en/third-party/messidor/

20. Mangiameli, P.; West, D.; Rampal, R. Model selection for medical diagnosis decision support systems. https://www.sciencedirect.com/science/article/pii/S0167923602001434

Authors

Navya Ramakrishnan is a sophomore at Jasper High School, Plano, Texas. She is interested in computer science and its applications in biomedical field. She plans to pursue higher studies in computer science or biomedical engineering.