Automatic Retinal Disease Classification Using Machine Learning and AI

Tapesh Santra
Email: tapesh.santra@gmail.com

Abstract*—Some retinal diseases can lead to vision loss and blindness. Early diagnosis can help patients get preventive care. Yet, due to lack of infrastructure and resources millions of patients do not avail such diagnosis. In this paper, I explore the possibility of developing an automatic retinal disease classifier using computer vision algorithms. Two different classes of algorithms are tested; (a) traditional computer vision approach of hand crafting features followed by developing machine learning (ML) models (b) automatic feature engineering and classification using more modern convolutional neural networks (CNN). The above algorithms were used to build both multi-class classifiers, i.e. the models which are trained to identify the correct disease, and binary classifiers, i.e. models that are trained to determine if a patient has a specific disease or not. A set of 600 pre-labelled retinal scan images were used to train the models. Both the ML and CNN models had relatively modest success in the multiclass scenario. However, the ML models were found to be reliably accurate in binary classification scenario, achieving >90% accuracy in identifying cataract.

Keywords*— Machine learning, computer vision, retina disease, image analysis, deep learning, artificial intelligence

I. INTRODUCTION

Research in the field of medical image analysis has shown promise to bring automated diagnosis in the healthcare industry. Especially, areas such as ophthalmology, i.e. eye disease, can benefit immensely from advances in the field of medical image analysis research. This is because diagnosis of many eye diseases rely on visually inspecting retinal scans by domain experts. It is a tedious and time consuming process and can largely be automated using modern machine learning (ML) and artificial intelligence (AI) algorithms. While ML/AI may not replace opthalmologists overnight, this technology can (a) make their jobs easier by assisting them to diagnose patients (b) can help deliver eyecare in places (e.g. developing countries/remote locations) where there is a shortage of opthalmologists.

II. RELATED WORK

Retinal image analysis is a well-studied field of research among image processing and computer vision practitioners. Here I mention a few relevant works.

M. Yang et al. [1] developed a back propagation neural network (BPNN) algorithm for detecting cataracts from a set of 504 retinal fundus images. They hand engineered features to deal with small sample numbers. They used several image transformation and manipulation techniques such as top-bottom hat transformation [2], trilateral filters [3] etc. to extract texture feature of the retinal image. These features were then used to classify retinal fundus images into medium, mild, normal and severe cataract.

J. Nayak et al. [4] used image processing and Support vector machine (SVM) to build automatic cataract detection model from 174 retinal fundus images. Following image processing steps such as mean intensity normalization, reference frame selection etc. they extracted features like Big Ring Area, Small Ring Area, Edge Pixel Count and Object Perimeter. The features were then used in Support Vector Machines for detecting cataract cases.

L. Guo et al. [5] used wavelet transform, cosine transform, multiclass discriminant analysis [6] to develop an automatic cataract classification and grading system from 445 images. Two feature extraction approaches was used wavelet transformation and the sketch with discrete cosine transformation (DCT).
W. Fan et al. [7] used top bottom hat transformation, followed by wavelet and sketch-based feature extractions and boosting and bagging classifiers to build a model for cataract grading and classification from 445 images. They also used PCA (principal component analysis) [8] reducing the dimensions of the extracted features.
M. Caixinha et al. [9] also used PCA to build a cataract classification model from 220 retinal images. They relied on Naïve Bayes, fisher linear discriminant (FLD), K Nearest-Neighbors (KNN) and support vector machine (SVM) algorithms for the actual classification task.
S. Kolhe at el. [10] used histogram equalization (CLAHE) followed by feature extraction using skeletonization with Discrete Cosine Transformation and Discrete wavelet transformation. They also used Principle Component Analysis (PCA) to reduce feature dimension and Support Vector Machine (SVM) [11] to build the classifier from 261 images.
J.J. Yang et al. [12] used set sketch, wavelet and texture-based features to build ensemble learning model of cataract classification from 1239 pictures.

While there are many other noteworthy works which were not mentioned above, a few common themes appear among the above works.

- Extensive image processing and feature engineering has been used to develop the models, especially in cases where the training datasets had small number of samples
- SVM and ensemble methods are classifiers of choice for many researchers
- Most of the above works are limited to detecting and grading cataract. Other diseases such as glaucoma and retinal detachment received limited attention.

III. CONTRIBUTION OF THIS PAPER

In this paper I shall explore the following:

1. The feasibility of developing deep learning models for retinal disease classification from a small set of images (600 images).
2. Exploring features and image processing techniques that are relatively less explored in the filed of retinal image analysis
3. Feasibility of detecting retinal diseases other than cataract for instance, glaucoma, retinal detachment etc.

IV. DATASET

We used a dataset from Kaggle **(**https://www.kaggle.com/jr2ngb/cataractdataset/version/2**).** The dataset contains 300 images of normal retinas, 100 images of cataract, 100 images of glaucoma and 100 images of retina disease. The images are of high resolution (2464 X 1632 pixels). A representative image from each class is shown in Figure 1

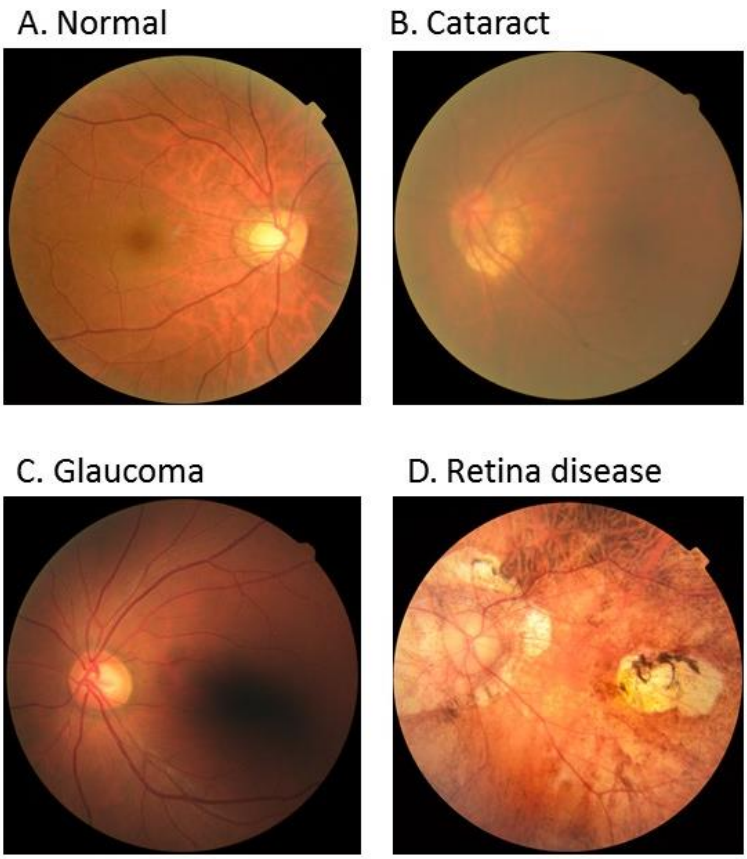

Figure 1: Images of normal retina (A), cataract (B), glaucoma (C) and retinal disease(D)

V. METHODOLOGY

Below, I provide brief details of the image processing, feature engineering, data augmentation steps, as well as the CNN and ML algorithms used in the study.

A. Deep convolutional Neural Networks (CNN)

A.1 Architecture:

We used two different CNN architectures (a) a smaller version of ResNet50 [13] with a smaller set of filters, and (b) a custom built 7 layers (4 convolutional + pooling layers, 3 fully connected layer) CNN [14]. Each identity and convolution block of the reduced ResNet50 model had 16,16 and 32 filters respectively. The custom CNN had 16,16,32,32 filters of size 5 X 5 respectively in its convolution layers, and 4 neurons each in its fully connected layers. Each convolution layer was followed by BatchNormalization, 'relu' activation and max-pooling layers of size 3X3. The code for both implementations are in link.

A.2 Pre-processing steps for deep CNN implementations

The following steps were used for image pre-processing:

- Resized the images to 128 X 128
- Normalized their histograms by stretching the red, green blue channels of the images to cover the full range 0-255. This is done by
    - Extracting red, green and blue channels of the images
    - Pixel value in each channel are then transformed by
        - R_transformed=(255/(R_range+0.0001)*(R_original-min_R))
        - G_transformed=(255/(G_range+0.0001)*(G_original-min_G))
        - B_transformed=(255/(B_range+0.0001)*(B_original-min_B))
        - Where R/G/B_range = max_R/G/B – min_R/G/B; R, G, B represent the pixel values in the red, green and blue channels respectively.

A.3 Data augmentation steps for deep CNN implementations

The following steps were used for data augmentation:

- Flipped the original images from left to right and used the flipped image as a separate data point
- Flipped the original images from up to down and used the flipped image as a separate data point
- Added 5% to the pixel values of the original and flipped images, i.e R/G/B = R/G/B + 0.05* R/G/B
- Subtracted 5% from the pixel values of the original and flipped images, i.e. R/G/B = R/G/B - 0.05* R/G/B

B. Feature extraction and machine learning

B.1 Pre-processing and feature engineering for ML methods

The following image pre-processing and feature engineering pipeline was created to extract features from images.

- Histograms of the images were normalized as described above
- The following features were created after histogram normalization:
    - Mean, Standard deviation, Entropy, Energy and Kurtosis (see [15] for definition) of Red, Green, Blue channels and Grayscaled versions of each image (20 features)
    - 16 bin histograms for the Red, Green and blue channels (48 features)
    - Images were then binarized using adaptive Gausian thresholding (See Figure 2 for examples of binarized retinal scans) with windowsize=11 and buffer=5 and the following features were calculated:
        - Ratio between the number of white pixels to the total number of pixels in the binarized images (1 feature)
        - Number of contours (except those with area >40000, there were image outlines) (1 feature)
        - 9 bin histograms of the contour areas (9 features)
    - The above resulted in a total of 79 features per image

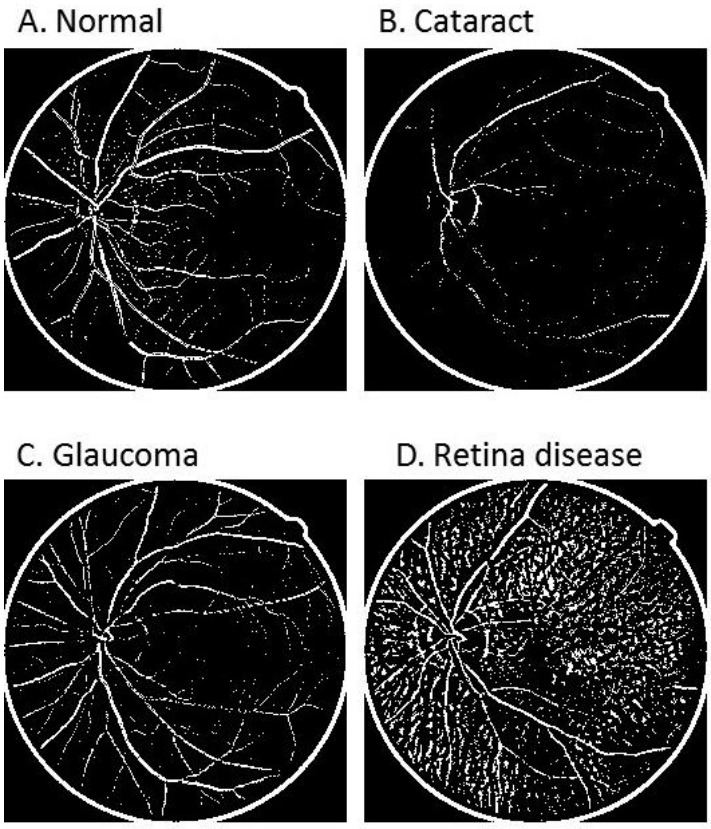

Figure 2: Thresholded images of normal retina (A), cataract (B), glaucoma (C) and retinal disease(D)

B.2 ML models

The following types of classifiers were used to develop ML models (see [16] for details of these algorithms)

- Multinomial Logistic regression (with L2 penalty)
- Random Forest (100 trees with maximum depth of 4 for each tree)
- Gradient Bossting classifier (100 trees, maximum depth=2, learning rate=0.015)
- Support Vector Machine

We tried two different settings

1. **Multiclass classification:** In this case the classifier is supposed to identify one of the four cases (normal, cataract, glaucoma and retina disease) given an image of a retina.

2. **One vs all binary classification:** In this case, three classifiers were developed, each triend to detect only one disease, cataract, glaucoma or retina disease.

VI. RESULTS

A. Deep convolutional Neural Networks (CNN)

The deep CNNs had limited success in classifying retinal diseases. While training the models, it seemed that they trained well initially, but within a few epochs (training iterations), their accuracy on training datasets started to surpass that on the test datasets (see Figure 3), and training the models for longer didn't improved test accuracy. The cross entropy loss for both training and test data, on the other hand, kept reducing and eventually stabilized. This suggests that the dataset has too few images to train parameter heavy CNNs. We tried the following measures to prevent such overfitting:

- Augmented the dataset using common augmentation methods as described in section A.2
- Increased l1 and l2 regularizations
- Reduced the number of parameters by using smaller number of filters in the convolutional layers

However, the above steps didn't have significant impact on the accuracy of the models.

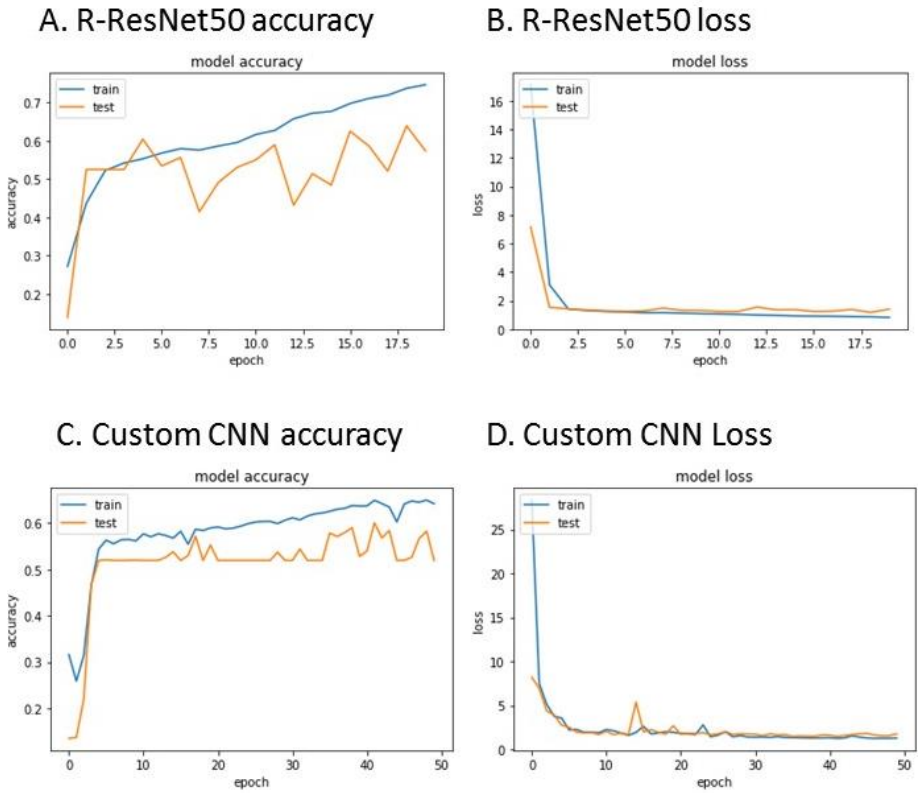

Figure 3: Accuracy and loss during training of R-ResNet50 (reduced ResNet50) (shown in panels A, B respectively) and a custom CNN (shown in panels C, D respectively).

The test accuracies of both CNNs were < 60%. The accuracy was calculated as the ratio between the number of correctly classified images divided by total number of images in the dataset. While, the above accuracy is still better than random, it begs the question whether more traditional approach of hand engineering features and then training ML models may give us better accuracy, especially given the small sample size of the datset.

B. ML models

B.1 Multiclass classification

As described above, four ML algorithms, logistic regression, random forest, gradient boosting, SVM were used for this exercise. The accuracies of these algorithms are shown in Table 1. The results are not significantly different from those of the CNNs, with the exception that, in case of ML models, there are little difference between training and test accuracies, suggesting less overfitting. To further investigate the performances, I plotted multi-class ROC curves for each model. These are shown in Figure 4.

ML Algorithm	Training accuracy	Test accuracy
Logistic Regression	64%	57%
Random Forest	62%	59%
Gradient Boosting	69%	61%
Support Vector Machines	55%	56%

Table 1: Multiclass classification accuracies of four ML models.

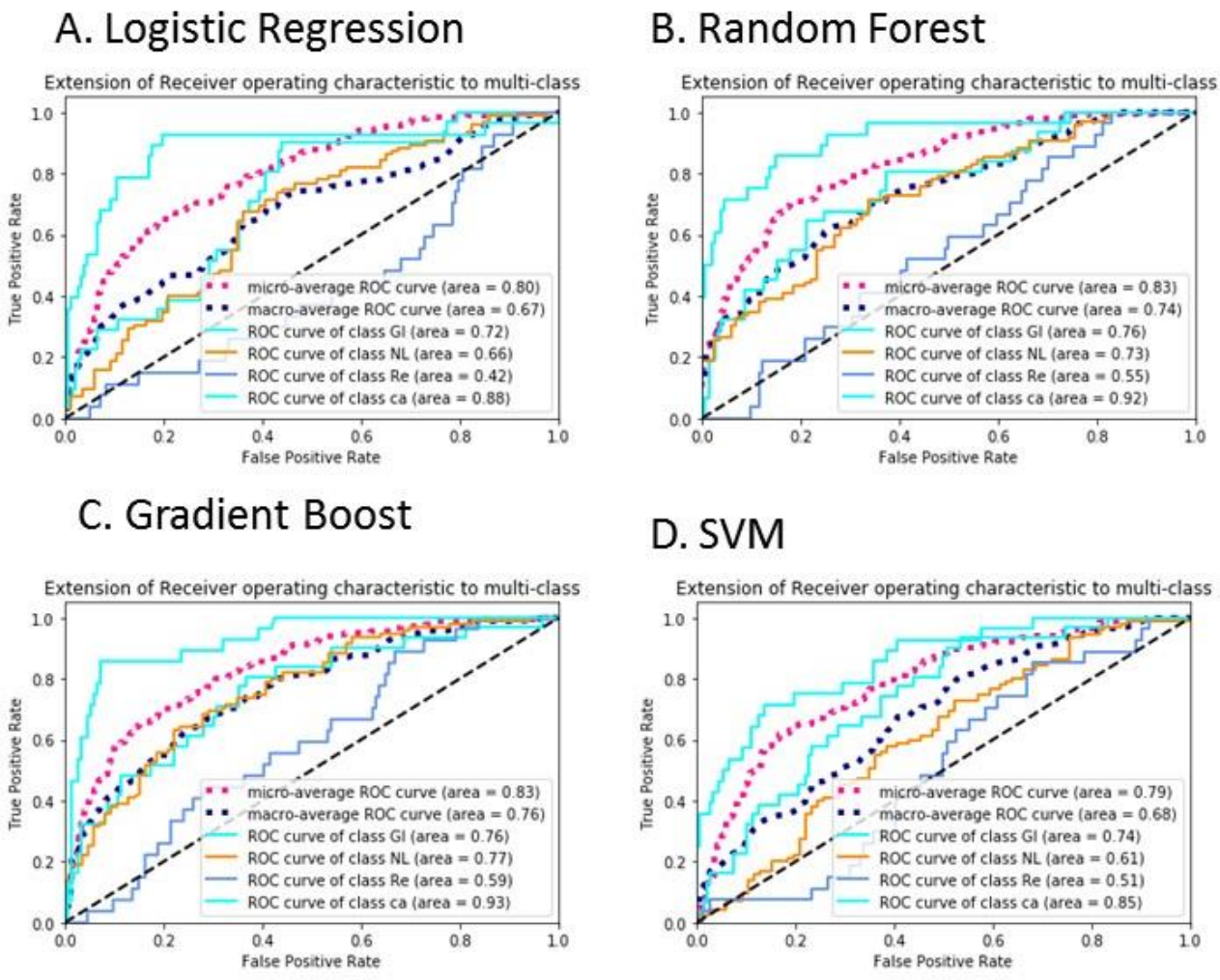

Figure 4: Multiclass ROCs for Logistic Regression (A), Random Forest (B), Gradient Boosting (C), SVM (D). Here NL, Gl, Re, ca represent normal, glaucoma, retina disease and cateract respectively.

The ROC curves and the area under these curves revealed few important insights.

- Firstly, the overall multiclass ROCs (micro and macro averaged ROCs) all had area >0.5, suggesting that the classifiers had better than random performance.
- Secondly, the ROCs of individual classes varied significantly, i.e. the classifiers were better at detecting certain diseases than others. For instance, the ROC curves for cataract and retinal disease classes consistently had the highest and lowest areas respectively.

The above results indicate that the ML models have significantly different abilities in identifying different retinal diseases. It begs the question, whether building separate classifier for each disease may have more success than building one multiclass classifier for all diseases.

B.2 One vs all binary classification

Next, we built binary classifiers using the same four ML algorithms mentioned above. The classifiers were trained to identify only one disease at a time. The performances of the classifiers are given in Table 2. This time, the models had >80 percent accuracy in all scenarios. They were most efficient in detecting cataracts with two models showing >=90% accuracy. These results indicate that the dataset that is used for this study may not be ideal for developing robust multiclass classifiers, but can still be used to develop reliable binary classifiers. This can arguably be due to small sample size and/or having biased samples in the dataset.

	Test accuracy		
ML Algorithm	**Glaucoma**	**Retina disease**	**Cataract**
Logistic Regression	85%	83%	88%
Random Forest	83%	85%	91%
Gradient Boosting	85%	85%	90%
Support Vector Machines	83%	85%	85%

Table 2: Binary classification accuracies of four ML models.

VII. FUTURE WORK AND CONCLUSION

In future, I want to try the following:

- Use transfer learning to re-train a trained ResNet50 network for classifying retinal diseases
- Use generative models to augment data for training deep CNN models
- Use more advanced image processing techniques for pre-processing retinal images
- Use wavelet, cosine and texture based features for developing ML models
- Use meta classifiers to improve classification accuracy

.

In conclusion, thinking purely from the perspective of accuracy, it seems that the binary classifiers have a better chance of being useful in clinic than the multiclass classifiers. However, the logistics of deploying such models in real life scenario is another challenge. For instance, it's one thing that a patient is diagnosed to have a certain eye disease out of many possibilities, it is an altogether different challenge if a patient is flagged for multiple eye diseases by many binary classification models. It's not clear whether and how the logistics of deploying these models in real life scenario are considered by the research community who are working in the filed of medical image analysis. Regardless, the direction of research in this area, should be informed by the downstream deployment challenges.

VIII. REFERENCES

[1] M. Yang, J.J. Yang, Q. Zhang , Y. Niu, J. Li, "Classification of Retinal Image for Automatic Cataract Detection", 15th International Conference On e-Health Networking Application and Services, pp. 674-679, 2013.

[2] Y. Liang, L. He, C. Fan, F. Wang, and W. Li, "Preprocessing Study of Retinal Image Based on Component Extraction", Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education, pp. 670-672, 2008.

[3] C. Tomasi and R. Manduchi. "Bilateral filtering for gray and color images", Computer Vision, Sixth International Conference on IEEE, pp. 839-846, 1998.

[4] J. Nayak, "Automated Classification of Normal, Cataract and Post Cataract Optical Eye Images using SVM Classifier", in Proceedings of the World Congress on Engineering and Computer Science, no.1, pp. 23-25, 2013.

[5] L. Guo, J.J. Yang, L. Peng, J. Li, Q. Liang, "Computer-aided healthcare system of cataract classification and grading based on fundus image analysis", Computers in Industry, no. 69, pp.72-80, 2015.

[6] T. Li, S. Zhu, M. Ogihara, "Using discriminant analysis for multi-class classification: an experimental investigation", Knowledge and Information Systems, vol.10, no. 4, pp. 453-472, 2006.

[7] W. Fan, R. Shen, Q. Zhang, J.J. Yang, J. Li, "Principal component Analysis Based Cataract Grading and Classification", IEEE 17th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 459-462, 2015.

[8] I.T. Jolliffe, Principal Component Analysis, 2nd, ed., 2002.

[9] M. Caixinha, E. Velte, M. Santos, F. Perdigao, Joao Amaro, Marco Gomes and Jaime Santos, "Automatic Cataract Classification based on Ultrasound Technique using Machine Learning: A comparative Study", 2015 International Congress on Ultrasonics, 2015 ICU Metz 70, pp. 1221-1224, 2015.

[10] L. T. Chylack, S. Kolhe, "Remote Automated Cataract Detection System Based On Fundus Images", International Journal Of Innovative Research in Science,Engineering and Technology, vol. 5, no. 6, 2016.

[11] M.Caixinha, E.Velte, M. Santos, J.B. Santos, "New approach for objective cataract classification based on ultrasound techniques using multiclass SVM classifiers." Ultrasonics Symposium (IUS) 2014 IEEE International, pp. 2402-2405, 2014.

[12] J.J. Yang, J. Li, R. Shen, Y. Zeng, J. He, J. Bi, Y. Li, Q. Zhang, L. Peng, Q.Wang, "Exploiting ensemble learning for automatic cataract detection and grading." Computer methods and programs in biomedicine, no. 24, pp. 45-57, 2016.

[13] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[14] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[15] Kumari, Pallavi, and Nibedita Jagadev. "Automatic Cataract Classification on Retinal Image using Support Vector Machine."

[16] Bonaccorso, Giuseppe. Machine Learning Algorithms: Popular algorithms for data science and machine learning. Packt Publishing Ltd, 2018.