ISSN 2063-5346

# BINARY CLASSIFICATION OF DIABETIC RETINOPATHY USING RANDOM FOREST CLASSIFIER

**Srilaxmi Dasari[1],Boo. Poonguzhali[2],ManjulasriRayudu[3] K. Muthukumar [4] R. Dhanalakshmi [5]**

**Article History: Received: 01.02.2023 Revised: 07.03.2023 Accepted: 10.04.2023**

## Abstract

Diabetic retinopathy affects millions of individuals globally. If untreated, this condition, mostly affects the retina of the eye, results in permanent blindness. Therefore, it's crucial to identify Diabetic retinopathy in early stage to protect patients from going blind. This article proposes a novel ML approach that uses the mutual information technique for selecting the optimized features. . The efforts are made at the feature extraction and feature selection stage to select optimized feature set for the classification. The dataset for diabetic retinopathy is initially subjected to the ML algorithms Nearest neighbor classifier (NNC ),Naive Bayes classifier(NBC),Decision Tree classifier(DTC).The Random forest classifier outperforms them all in accuracy with an average performance of 75% after best feature selection and 66% before mutual information technique.

**Keywords—**Diabetic Retinopathy, Machine Learning , Random Forest Classifier , Retinal Images , Information Gain

**[1]** Research Scholar, Department of Electronics and Instrumentation, Annamalai University, Annamalai Nagar, Chidambaram, India.

**[2]** Assistant Professor, Department of Electronics and Instrumentation, Annamalai University, Annamalai Nagar, Chidambaram, India.

**[3]** Professor **,** Department of Electronics and Instrumentation, Vnr Vignana Jyothi Institute of Engineering and Technology, Telangana, India

**[4]**Assistant Professor, Department of Electronics and Instrumentation, Annamalai University, Annamalai Nagar, Chidambaram, India.

**[5]**Associate Professor, Department of Electronics and Communication, Thanthai Periyar Government Institute of Technology, Vellore,India.

[1]hemasrisrm@gmail.com,[2]bookuzhali@gmail.com,[3]manjulasree_r@vnrvjiet.in,[4]muthukumar.Kalyanam@gmail.com,[5]dhanavishnu02@gmail.com

**DOI:10.31838/ecb/2023.12.s1-B.226**

## I. INTRODUCTION

Diabetes has been a prevalent disease for a long time mostly regarded with the body's blood sugar levels and how it can cause heart conditions. Still, another lesser-considered state is one called "Diabetic Retinopathy". As the name suggests, this is a disease that affects the eyes, particularly the retina.DR is the main contributor of visual impairment in the majority of industrialized and developing nations, particularly among working people in [11].Diabetic retinopathy causes significant damage to the blood vessels in the retina, which may cause it to either rupture or exude or completely get shut, preventing blood flow. This may result in partial, or in some cases, complete blindness. Regular checkups are for this very reason, very important. The symptoms do not seem as serious in the first stages, and one can be indifferent towards them, but spotting this in the early stages may prevent one from suffering with complete blindness. The causes of diabetic retinopathy is usually excessive sugar level in the blood. This high blood sugar level causes damage to all blood vessels in the body but the damage can be more significant in parts of the body like the eye where the vessels are thin and fragile, therefore more prone to ruptures and leakages. To be more specific, the sugar content in the blood changes the viscosity of the blood, and in diabetic retinopathy, this sugar blocks the minute blood vessels from supplying blood to the retina. This triggers a repairing mechanism in the eye where new blood vessels might be formed which do not quite function as well as their counterparts and are more fragile. They leak or bleed easily, causing retinal damage. Diabetic retinopathy can broadly be classified into two types: "Non-proliferative diabetic retinopathy (NPDR)" and., "Proliferative diabetic retinopathy (PDR)". Early diabetic retinopathy refers to the initial stages of the condition where the blood vessels are getting progressively blocked off completely, causing stagnation and bulging of the vessel which sometimes cause extrudes leaking into the retina. This may also progress into leaking fluid into central part of the retina called the macula, which leads to the symptom of decrease in vision. Proliferative diabetic retinopathy is when the original blood vessels are blocked completely and new blood vessel formation takes place. As mentioned, these vessels are fragile and may protrude blood into the jelly like vitreous of the eye. Paired with the formation of new blood vessels, is also the development of scar tissue which causes retina to gradually pull away from the cavity of the back of the eye. Since these blood vessels are a result of a defense trigger, they might not perform all the functions that the original blood vessels performed, causing a probable blockage of flow of fluids from the eye. Vitreous hemorrhage, Retinal detachment, Glaucoma, Macular edema, macular ischemia, and blindness are conditions that are associated with the progression of diabetic retinopathy in a patient. The identification of markers of blood vessel ruptures, fluid leakages, swelling up of vessels which are associated with diabetic retinopathy on a case to case basis is not humanly possible. This is where the idea of machine learning comes in. Models can be trained to identify markers and abnormalities from a normal no-disease fundus images. The cases that stand out as an abnormality can be further analyzed by a doctor and researched on a case-to-case basis. The models can be trained with the specific markers associated with various stages so that they can be classified based on the severity and the progression of the disease. The initial stage is a binary classification that classifies into the two broad categories of whether or not a patient has diabetic retinopathy. If the result is true (or 1), the fundus image is sent to another model which has a pre-defined classification of 5 stages of the disease based on the progression of the seriousness of retinal damage. This second model is an

multi stage model with a higher share of images which is extensively trained by dividing into training, validation and testing sets to improve the model accuracy. Manual diagnosis of eye conditions by ophthalmologists used to take a lot of time [16] hence automated techniques were used.

## II.LITERATURE SURVEY

This section discusses the survey of current methodologies with the goal of identifying Diabetic Retinopathy (DR). In this survey, the benefits and drawbacks of current approaches are also discussed. The use of Artificial Neural Network (ANN), K-means clustering, and random forest (RF) algorithms for early diabetes prediction was covered in discussion of machine learning techniques [1, 2]. The ANN outperformed these algorithms with an accuracy of 75.7% [1], whereas DT, SVM, LDA, and NB approaches were also used. The LDA performed well, including hypertension and prehypertension, with an accuracy rate of 79% [2]. A GA- and SVM-based strategy was proposed to identify heart disease [4] and GA based systems were explored in [3] where SVM Classifier is used for Binary classification and results were merged and fed into GA to detect the presence of diabetic retinopathy. S Bandyopadhyay et al .[5] identified Diabetic Retinopathy (DR) by extracting the Blood vessel using morphological segmentation, whereby this technique required less computing time to mine the finest vessels. K-Nearest Neighbor classifiers, which were used to identify the final DR platforms, took the retrieved Blood vessel features as input. Wang Z et al. [6] have introduced a innovative architecture that classifies images as abnormal or healthy, nonreferable or referable diabetic retinopathy by achieving greater AUC. While Chetoui M et al. [7]discusses the use of various texture features as LESH (Local-Energy based Shape Histogram) for Diabetic retinopathy and LTP (Local Ternary Pattern) by Veerashetty S et al. in [17].Vadloori et al. [8]forecasted the occurrences of the DR using DR dataset and a variety of machine learning classification techniques were used to predict. Pragathi.P et al[9]proposed integrated approach using support vector machine classifier(SVM) ,moth flame optimization technique(MFO), Brownlee Et al.[12] proposed Base models for Diabetic retinopathy, also referred to as level-0 models, are included in the stacking, along with a meta-model that integrates the level-0 model prediction. Kamal, M et al.[10] clearly explained complete review on DR Detection techniques.SK et al.[13] predicted DR early using Ensemble Machine learning classifier. Gharaibeh N et al.[14]recognized Bright lesion in Retinal Images using a hybrid SVM Naïve Bayes Classifier .Barkana B.Det al.[15] depicted performance analysis of statistical features in Blood vessel segmentation using Fuzzy logic.

## III.PROPOSED METHODOLOGY

The Block Diagram shown in Fig .1 depicts about the machine learning approach for detection of Diabetic Retinopathy using Random Forest Classifier .The optimized features are selected from the Debrecen Dataset using mutual information technique after normalization process .

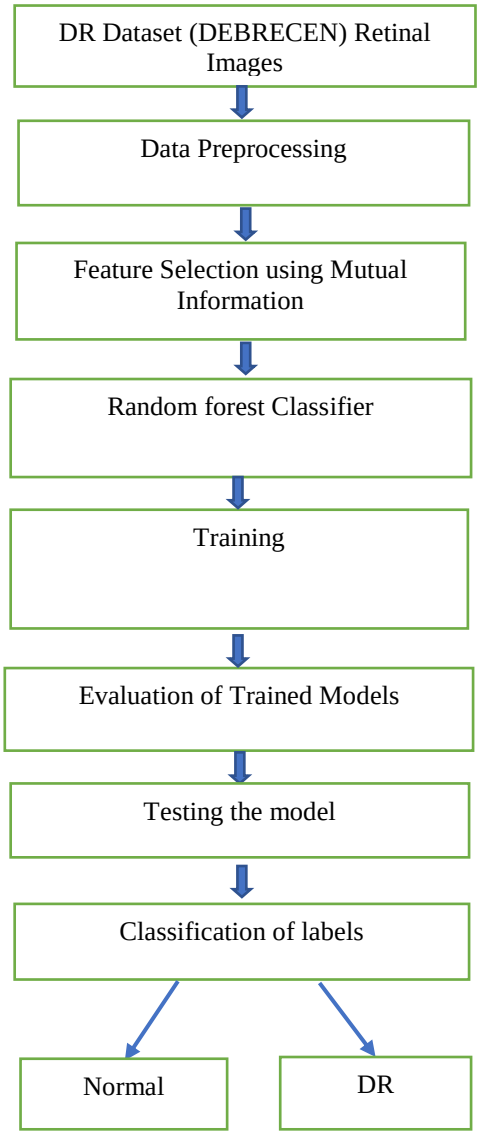


**Fig.1.** Block Diagram for proposed Blended ML Approach

3.1 Dataset Description (Debrecen Dataset)

The Debrecen Dataset is represented by 20 features in the diabetic retinopathy dataset. Extracted Features from an image will show whether Diabetic retinopathy is present or not. In Table 1, all 20 attributes are shown with numbers ranging from 0 to 19.The values associated with the image quality make up feature 0 in the dataset. An image is considered to have good quality if lesions can be found effectively in it, otherwise, it is said to have poor quality. Binary numbers 1 and 0 are used to indicate this characteristic. The specifics of the pre-screening are included in Feature 1. A number of 1 indicates a serious problem, whereas a value of 0 indicates the retina is normal. The amount of values relating to micro aneurysms that were found in features 2 to 7 makes up this feature micro aneurysms (MA). The confidence intervals used to identify these MA values range from 0.5 to 1. The exudate-related normalized values make up Features 8 to 15. These are depicted as having the same features as those from 2 to 7. Between the centers of the macula and the optic disc, Feature 16 provides the Euclidean distance information. The specifics of the optic disc diameter are contained in Feature number 17. The binary values associated with the classification based on AM and FM modulations. The binary values associated with class labels relates Feature 19. Diabetic retinopathy symptoms are represented by the value 1 no symptoms are represented by the value 0.The Features for Diabetic retinopathy Grading are given in the below Table1.

Table 1

Features for DR Grading

| Feature | Feature Description |
|---|---|
| 0 | Quality OF Image is described as binary values 1 and 0. Where 1 denotes good quality, and 0 denotes bad quality |
| 1 | Pre-screening information is described as binary values 1 and 0. Where 1 denotes severe abnormality in the retina, and 0 denotes No abnormality in retina |
| 2-7 | These features describe the number of Microaneursym values detected. Each value of Feature stands for the no. of the Microaneursym obtained at levels $\alpha = 0.5, \ldots .1$ respectively |
| 8-15 | Contains Same description as (2–7) for exudates. These are normalized by dividing no.of lesions with the ROI Diameter to compensate |

various sizes of image.

16 Information on the euclidean distance between the macula and optic disc centres is provided by this feature.

17 This feature provides information on the diameter of the optic disc

18 The amplitude modulation (AM) and frequency modulation (FM)-based classification's binary values (FM)

19 Binary values 1 and 0 are used to represent class labels,1=Presence of DR, 0=Lack of presence of DR

### 3.1 Data Preprocessing

The retinal images require preprocessing because the vessels' low background contrast, high illumination, ambiguous vessel residuals, etc. Green channel is separated from the image due to its differentiation in high background contrast. Later, it is improved using the CLAHE contrast enhancement approach.

### 3.2 Normalization

Normalization is the process of scaling Technique applied in Machine Learning during preparation of data to vary the attributes of columns of numeric in the applied dataset to use on a standard scale .Normalization is used when the dataset which is consisting of various attributes possess various ranges .Performance metrics are very much enhanced using Normalization. Mathematically Normalization is calculated using the formulae given below .

Xn =$(\boldsymbol{X} - \boldsymbol{Xmin})|\boldsymbol{Xmax} - \boldsymbol{Xmin})$

Where Xn=Normalization Value

Xmax=Maximum Feature Value

Xmin=Minimum Feature Value

### 3.3 Feature Extraction and Selection

In the present work, totally 19 features are extracted out which 2to 7 features represent Microaneursym values and 8 to 15 features represent Exudates which are diseased features of Diabetic RetinopathyAND THE Information Gain of each figure is represented in Fig 2 . Using different feature selection techniques, such as Threshold variance, Correlation coefficient, Information gain redundant, correlated features are filtered and the best features are chosen. These techniques improve a model's ability to Generalise, and may also improve a classifier's overall accuracy, and duplicated features also make the model more complex. Thus, the optimum set of 14 features are chosen for categorization using the aforementioned techniques. The information gain chart shown in Fig 3 illustrates the significance and applicability of these 14 Features . The optimized 14 features shown in Fig 3 in descending Information gain order are Image quality feature, and Diseased Features of Diabetic retinopathy .

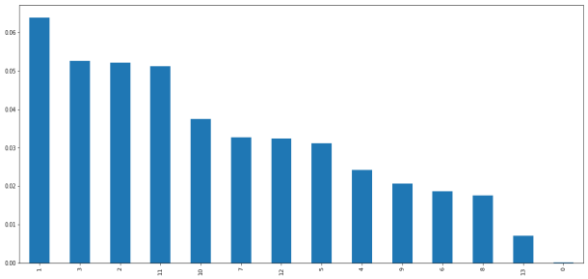

**Fig .2**.Information Gain of each Feature

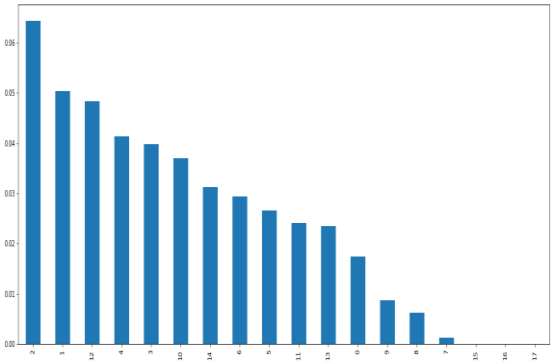

**Fig 3**.Optimized Features

### 3.4 Classification

The effectiveness of machine learning algorithms heavily depends on the quality of the features. The target variable should have a high correlation with the optimized set of Features while there should be little correlation between them. As part of our experiment, we used a variety of machine learning (features-based) classifiers, including Random Forest (RF) [18]–[20], Decision Tree, Naive Bayes (NB), Adaboost (AB) ,Guassian Process(GP) , and Quadrant Discriminative Analysis (QDA). In this study, we demonstrated a Random Forest (RF) classifier that was most accurate when categorizing Diabetic Retinopathy as normal and abnormal. The RF is an ensemble-based supervised technique that can be used for a variety of applications. It bases target class prediction on the majority vote of the ensembled collection of decision trees. The dataset is divided 80:20 for training and testing, and cross-validated 10 times.

## IV. EXPERIMENTS AND RESULTS

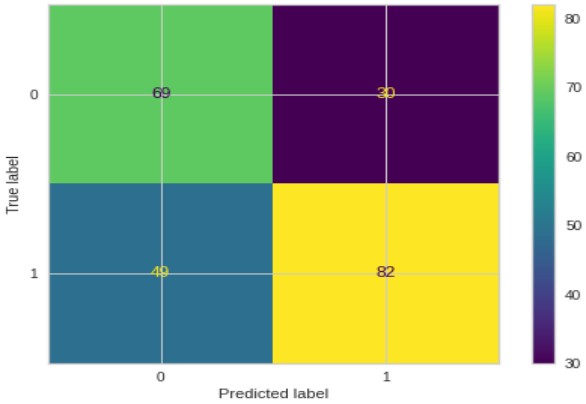


**Fig.4**.RF Confusion Matrix Using Unoptimized Features

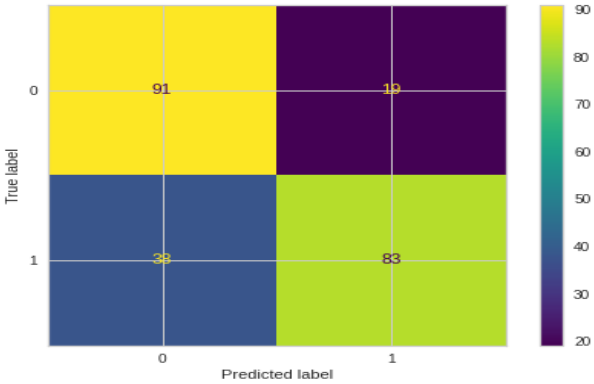


**Fig5**.RF Confusion Matrix Using Optimized Features

```
FP= [38. 19.]
FN= [19. 38.]
TP= [91. 83.]
TN= [83. 91.]
TPR= [0.82727273 0.68595041]
TNR= [0.68595041 0.82727273]
PPV= [0.70542636 0.81372549]
NPV= [0.81372549 0.70542636]
FPR= [0.31404959 0.17272727]
FNR= [0.17272727 0.31404959]
FDR= [0.29457364 0.18627451]
ACC= [0.75324675 0.75324675]
```

**Fig.6.**RF Performance Metrics Using Optimized Features

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.58 | 0.70 | 0.64 | 99 |
| 1 | 0.73 | 0.63 | 0.67 | 131 |
| accuracy | | | 0.66 | 230 |
| macro avg | 0.66 | 0.66 | 0.66 | 230 |
| weighted avg | 0.67 | 0.66 | 0.66 | 230 |

**Fig.7**.RF Metrics Using UnOptimized Features

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.71 | 0.83 | 0.76 | 110 |
| 1 | 0.81 | 0.69 | 0.74 | 121 |
| accuracy | | | 0.75 | 231 |
| macro avg | 0.76 | 0.76 | 0.75 | 231 |
| weighted avg | 0.76 | 0.75 | 0.75 | 231 |

**Fig.8.**RF Metrics Using Optimized Features

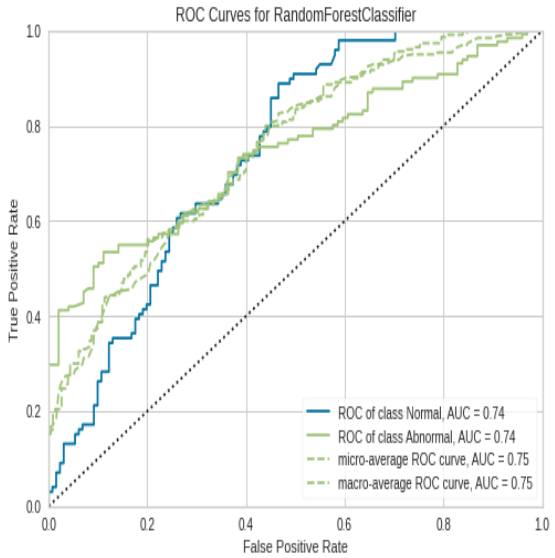


**Fig.9**.RF Metrics Using Optimized Features

The Random Forest Confusion Matrix of Optimized and Unoptimised Features is shown in Fig.4. and Fig.5.The Random Forest Metrics using UnOptimized and Optimized Features is shown in Fig.7. and Fig.8. The performance of the Random Forest classifier in categorizing the diabetic retinopathy abnormality is seen on RoC plots, which show the relationship between the true-positive rate and the false-positive rate. RoC plot, Fig 9 depict the RF classifier performance for an optimized Feature set.

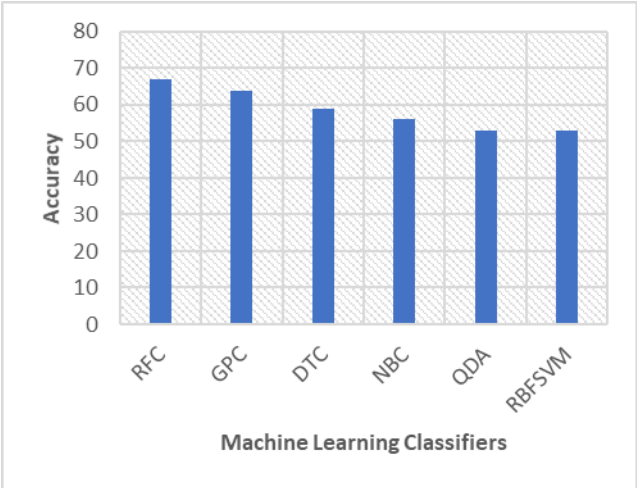


**Fig.10**. Summary of Machine Classifiers

The performance of various machine learning classifiers such as Random Forest ,Gaussian Process ,Decision Tree, Naïve Byaes ,Quadriatic Discriminant Analysis, Radial Basis Function Support Vector Machine with unoptimized features is shown in Fig.10.Random Forest Classifier outperformed highest accuracy when compared to other classifiers. Comparison to the paper cited in [21] the accuracy score obtained by KNN Classifier is 65 percent which is less than the proposed Random forest Classifier as described in Table 2.

Table 2

Comparision of Machine learning Classifiers

| S.No | Classifier | Accuracy |
|---|---|---|
| 1 | KNN | 65 |
| 2 | Random Forest (Proposed) | 75 |

## V. CONCLUSION

In this work, a unique method for classification system of Diabetic retinopathy images into normal and abnormal category is proposed. An

optimised fourteen-dimensional feature vector is created for Classification Design. The feature vector association between each other has shown low Correlation and a strong correlation with target class variable. RF classifiers obtained 75% accuracy in classifying binary grades of diabetic retinopathy into healthy and abnormal rate .The model also has a low false positive rate and a high true positive rate.

Datasets created during and/or analysed during the current investigation are available in the Diabetic Retinopathy Debrecen Data Set, which is located at https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy +Debrecen+Data+Set.

## VI. REFERENCES


[1] T. Mahboob Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. Imtiaz Baig, et al., "A model for early prediction of diabetes," Informatics Med. Unlocked, vol. 16, pp. 1–6, 2019.

[2] H. Chirath and C. Charith, "A Machine learning approach to predict diabetes using short recorded photoplethysmography and physiological characteristics," Artif. Intell. Med.,vol. 11526, pp. 322–327, 2019.

[3] R. A. Welikala, M. M. Fraz, J. Dehmeshki, A. Hoppe, V. Tah, S. Mann, et al., "Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy," Computerize. Med Imag. Graphic., vol. 43, pp. 64–77, 2015.

[4] C. G. Babu and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," Cluster Comput., vol. 22 pp. 14777–14787, 2019

[5] Bandyopadhyay S, et al. Gradation of diabetic retinopathy using KNN classifier by morphological segmentation of retinal vessels. In: Sreenivasa Reddy M, Viswanath K and Shiva Prasad KM (eds) International Proceedings on Advances in Soft Computing, Intelligent Systems, and Applications. Singapore: Springer, 2018, pp.189–198.

[6] Wang Z, Yin Y, Shi J, Fang W, Li H, Wang X (2017) Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. Int Conf Med Image Comput Comput-Assist Intervent Berlin, Germany: Springer, pp. 267–275

[7] Chetoui M, Akhloufi MA, Kardouchi M (2018) Diabetic retinopathy detection using machine learning and texture features. 2018 IEEE Canadian conference on electrical & computer engineering (CCECE), Quebec City, QC, pp. 1–4.

[8] Vadloori, S., Huang, Y. P., & Wu, W. C. (2019). Comparison of various data mining classification techniques in the diagnosis of diabetic retinopathy. Acta Polytechnica Hungarica. Advance online publication. doi:10.12700/APH.16.9.2019.9.3

[9] Pragathi, P.; Rao, A.N. An effective integrated machine learning approach for detecting diabetic retinopathy. Open Comput. Sci. 2022, 12, 83–91. [CrossRef].

[10] Kamal, M.M.; Shanto, M.H.I.; Mirza Mahmud Hossan, M.; Hasnat, A.; Sultana, S.; Biswas, M. A Comprehensive Review on the Diabetic Retinopathy, Glaucoma and Strabismus Detection Techniques Based on Machine Learning and Deep Learning. Eur. J. Med. Health Sci. 2022, 24–40. [CrossRef].

[11] Mishra, A.; Singh, L.; Pandey, M. Short Survey on machine learning techniques used for diabetic retinopathy detection. In Proceedings of the IEEE 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater

Noida, India, 19–20 February 2021; pp. 601–606. [CrossRef]

[12] Brownlee, J. Stacking Ensemble Machine Learning with Python. In Machine Learning Mastery; Machine Learning Mastery: San Francisco, CA, USA, 2020; Availableonline: https://machinelearningmastery.com/stacking-ensemble-machine-learning-withpython/ (accessed on 21 May 2022).

[13] SK, S. A Machine Learning Ensemble Classifier for Early Prediction of Diabetic Retinopathy. J. Med. Syst. 2017, 41, 201. [CrossRef]

[14] Gharaibeh, N.; Al-Hazaimeh, O.M.; Abu-Ein, A.; Nahar, K.M. A Hybrid SVM NAÏVE-BAYES Classifier for Bright Lesions Recognition in Eye Fundus Images. Int. J. Electr. Eng. Inform. 2021, 13, 530–545. [CrossRef]

[15] Barkana, B.D.; Sariçiçek, I.; Yildirim, B. Performance analysis of descriptive statistical features in retinal vessel segmentation via fuzzy logic, ANN, SVM, and classifier fusion. Knowl. Based Syst. 2017, 118, 165–176. [CrossRef]

[16] Khatri, M. Diabetes Complications. Available online:https://www.webmd.com/diabetes/diabetes-complications (accessed on 18 May 2022).

[17] Veerashetty S, Patil NB (2020) Novel LBP based texture descriptor for rotation, illumination and scale invariance for image texture analysis and classification using multi-kernel SVM. Multimedia Tools and Applications 79 (15-16):9935–9955

[18] Kuppusamy, M. M. Basha and C. -L. Hung, ”Retinal Blood Vessel Segmentation using Random Forest with Gabor and Canny Edge Features,” 2022 International Conference onSmart Technologies and Systems for Next Generationomputing (ICSTSN), 2022, pp. 1-4, doi: 10.1109/ICSTSN53084.2022.9761339.

[19] Zaaboub, Nihel, and Ali Douik. "Early Diagnosis of Diabetic Retinopathy using Random Forest Algorithm." 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE, 2020.

[20] Nirmala, K., Narayanan Venkateswaran, and C. Vinoth Kumar. "HoG based Naive Bayes classifier for glaucoma detection." TENCON 2017-2017 IEEE Region 10 Conference. IEEE, 2017.

[21] Apoorva Hegde, K R Sumana.”Comparitive Study of Diabetic Retinopathy Detection Using Machine Learning Techniques.”IJRASET Volume 10, Issue VIII, August 2022,doi: 10.22214/ijraset.2022.46101