# A Novel Machine Learning Approach for Diabetic Retinopathy Detection via Feature Importance Analysis

**Janhvi Chauhan[1], Dhaval Modi[2]**

[1, 2]Lecturer

[1, 2]Electronics & Communication Dept., Government Polytechnic, Ahmedabad

**Abstract**

**The human eye condition known as diabetic retinopathy damages the retina of the eye and can lead to total blindness. To prevent total blindness, diabetic retinopathy must be identified early. Diabetic retinopathy is identified by physical examinations such as optical consistency tomography, pupil dilation, and visual acuity testing. However, it may have an impact on the patients and is time-consuming. In light of these implications, this study uses a machine learning system to identify diabetic retinopathy in the human eye. The suggested approach uses classification algorithms on a number of characteristics of an existing Diabetic Retinopathy dataset, such as optical disk diameter, lesion-specific characteristics (microaneurysms, exudates), or the presence of hemorrhages. After that, the traits were taken out and applied to the ultimate decision-making process to determine whether diabetic retinopathy was present. The suggested approach made use of logistic regression and decision trees. For the prediction, use a support vector machine. The suggested approach outperformed the current efforts by achieving 89% accurate outcomes. This further demonstrates the vastness of the suggested approach.**

**Keywords: Hard Exudates, Logistic Regression, Feature Importance, Support Vector Machine**

## 1. Introduction

Diabetes mellitus, often known as diabetes, is the cause of diabetic retinopathy, also known as diabetic eye disease. This category of metabolic diseases is characterized by persistently elevated blood sugar levels. The retina is where the mellitus shows up. In many developed nations, diabetic eye disease is one of the leading causes of total blindness. While traditional procedures, such as dilatation of the pupil, are time-consuming and require the patient to suffer for a while, automated retinal image processing has made the diagnosis of retinal disorders considerably easier. It is currently one of the most serious and pervasive eye conditions. Among people of working age in developed nations, it is the most common cause of irreversible blindness [1]. When diabetes destroys the blood vessels inside the retina, blood and fluids flow into the surrounding tissue, resulting in diabetic retinopathy. Microaneurysms, bleeding, hard exudates, and cotton wool patches are all caused by this fluid leak [2], [3]. Patients may only become aware of this silent disease when retinal alterations have advanced to the point that therapy is challenging or impossible.

Therefore, the following contribution is made by this work. In order to assist individuals in identifying diabetic retinopathy at the primary level, this research suggests an automated diagnosis method for the condition based on a large and important characteristic taken from the DIARET-DB dataset. 15945 samples and 66 attributes that correspond to different symptoms of primary, mild, and severe diabetic retinopathy are included in the dataset. Area, bounding box, convex area, and regional intensity coefficients

that correspond to the maximum, minimum, and mean variance of the red and green planes, respectively, are a few of the properties. The most significant and vital features were extracted initially by using a feature selection technique. The accuracy of our model in identifying the presence of symptoms on a patient's eye was then determined by using those features to train some of the best machine learning models. The primary objective is to automatically categorize any retinal image's proliferative and non-proliferative diabetic retinopathy grade.

This suggested method's contribution is a completely automated, quick, and nearly precise DR detection. Therefore, our model is almost very successful in predicting the symptoms of Diabetic Retinopathy.

## 2. Literature review and Dataset

Due to the enormous number of cases of diabetic retinopathy that occur globally, aiding in the entire diagnosis process is necessary. The automatic identification of diabetic retinopathy saves a significant amount of time and effort. Thus, Maher et al. [4] looked at a decision-making support system. In this instance, a Support Vector Machine classifier was used. Diabetic Retinopathy was detected using a wide range of image processing techniques [5]–[8]. Nayek et al. [9] assessed and investigated a neural network classifier based on the region encircled by vessels and exudates. They attained a 93% detection accuracy. In order to distinguish between prolific and non-prolific fundus images, Acharya et al. constructed a support vector machine classifier and supplied it bi-spectral invariant characteristics [10]. To classify NPDR, an automated method was provided for the diagnosis system to categorize three categories of early symptoms: microaneurysms, red lesions, hemorrhages, and exudates and cotton wool spots of DR [11]. For microaneurysms, hard exudates, hemorrhages, and cotton wool patches, they obtained an accuracy of around 83%, 88.3% utilizing 430 photos, and another in the range of 85-87%, and 93% using 360 photographs.

### A. The DIARET-DB Dataset

Due to extensive research and studies on the automated diagnosis and early identification of diabetic retinopathy, a large number of datasets are available online on several platforms, including GitHub, Kaggle, and many more. We made use of the DIARETDB1 dataset, which has been extensively utilized for computer-aided diagnostic model evaluation and implementation. These 89 fundus photos were used to create this local dataset. It contains manually annotated pictures that demonstrate different degrees of diabetic retinopathy. This dataset contains region-based lesion information for six distinct classes, each with 15,945 samples and 66 characteristics. [12]

The class labels have the following meaning:

- Class 0: Bright non-lesion (Healthy Eye)
- Class 1: Hard exudates (Early DR)
- Class 2: Cotton wool spots (Early DR)
- Class 3: Red non-lesion (Early DR)
- Class 4: Microaneurysms (Early DR)
- Class 5: Hemorrhages (Severe DR)

An eccentricity, area, equivalent diameter, perimeter, orientation, and 16 region-based intensity coefficients—such as the mean, variance, and minimum and maximum of pixels in the red, green, intensity, and hue plans—are some of the most important characteristics among those 66 properties.

## 3. Methodology

We already know from the explanation above that our dataset consists of six classes, each of which expresses a label that is observed in the early stages of diabetic retinopathy. Because manual symptom

detection relies on years of experience, scientists and researchers are working hard to automate the entire process. That is our primary goal in our work. We start by preprocessing our dataset using feature scaling and feature selection techniques, then we use a number of machine learning algorithms to construct our model, and finally, we will assess our results.

### A. Algorithms

Four distinct kinds of advanced machine learning algorithms were used in our model, and Python 3.6 was the programmable language and Anaconda Platform was used for the implementation. We employed the following algorithms: Decision Tree, K-Nearest Neighbor, Support Vector Machine, and Logistic Regression [13] [14] [15]. These methods have their unique features and performance depending on the dataset, and they work well for various classifications. As previously stated, the most popular methods for classification problems are SVM and logistic regression.

### B. Preparing the Dataset

1. **Standardization**: The process of rescaling one or more attributes to have a mean value of 0 and a standard deviation of 1 is known as data standardization. The whole procedure reduces the disparity between the maximum and minimum value and provides a better display of the data points than previously possible because several of the attributes in the DIARETDB1 dataset have higher ranges.

$$X_{new} = \frac{x - \mu}{\sigma} \dots\dots\dots\dots\dots\dots\dots\dots (1)$$

2. **Normalization**: We don't know how our entire data is distributed because our dataset has a lot of features, some of which have extreme values. In this case, we applied normalization to every attribute in our whole dataset, with the exception of the label. Rescaling one or more qualities to the 0–1 range is what it is. Gaining a general understanding of our dataset's distribution is beneficial.

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \dots\dots\dots\dots\dots\dots\dots\dots (2)$$

3. **Feature Selection**: Training our model using every feature in our dataset would need a significant amount of computing time due to its large number of features. Selections are one of the key ideas that directly affect the model's performance in situations like these features. It can cut down on training time, enhance accuracy metrics, and prevent overfitting issues. Numerous feature selection methods exist, such as correlation matrix with heatmap, univariate selection, and others. In order to identify the most crucial features to take into account when developing our model, we have employed feature importance in this study.

4. **Feature Importance**: The model's feature importance attribute can be used to determine the relative relevance of each feature in the dataset. A tree-based classifier is included with this built-in class. Each aspect or attribute of the data is assigned a score; the greater the score, the more significant the information. The top 30 features for the dataset were extracted using the additional tree classifier. Every node in a decision tree is a condition on a single feature because this classifier is tree-based. This is intended to divide the dataset in half such that values with comparable answers end up on the same set. Impurity is the parameter that specifies this criterion. The ideal state is selected based on this metric. In classification problems, Gini Impurity is usually used. As a result, it is simple to determine how much each feature or attribute reduces the weighted impurity during decision tree training. Features can then be rated based on the average of the impurity decreases determined by each

characteristic. In the entire dataset, the most significant characteristics will reduce the weighted impurity less than the less significant features.

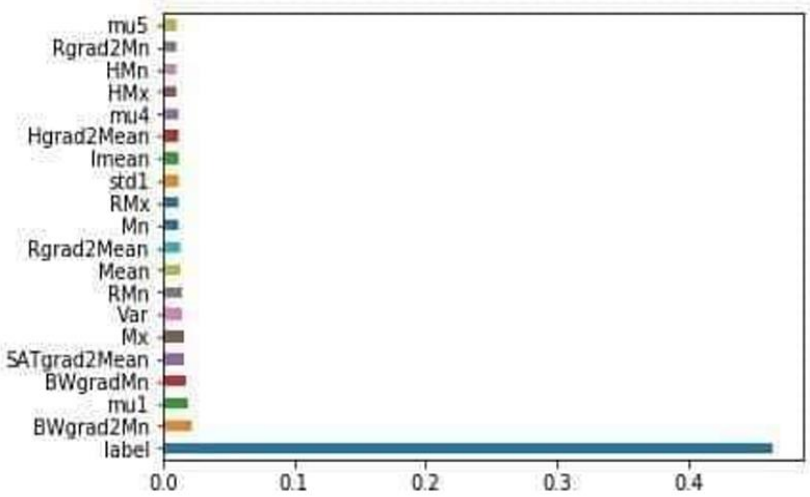


**Figure 1. Top features using feature importance technique.**

Scikit Learn uses Gini Importance to determine a node's importance for each decision tree, assuming that there are only two child nodes. A node's or feature's average over all trees indicates its final feature relevance. By adding up all of the feature importance values, the value is finally normalized to 0 to 1 [16–17].

To train our model, we chose the top 30 attributes. Thirty others were chosen over any number for a multitude of reasons. First of all, variables with more categories are preferred when features are chosen based on Gini impurity. Second, any characteristic in a dataset that has two or more correlated features can be used as a predictor. However, the significance of others is much diminished once any of them is utilized. The impurity that needed to be effectively eliminated was already eliminated by the first features. Since we initially considered cutting the features in half, 30 features were chosen as the final feature count since any other number would have produced less useful results.

It's time to construct our classifier now that we have all the necessary information. Four distinct kinds of advanced machine learning algorithms were used in our model, and Python 3.6 was the programmable language we used for the implementation. We employed the following algorithms: Decision Tree, K-Nearest Neighbor, Support Vector Machine, and Logistic Regression. These methods have their unique features and performance depending on the dataset, and they work well for various classifications. As previously stated, the most popular methods for classification problems are SVM and logistic regression.

## 4. Implementation

Our dataset was divided into 20% test data and 80% training data. Each of the previously extracted features using the feature significance technique was then subjected to a K-Nearest Neighbour Classifier test. Despite its poor accuracy of 65%, it provides an impressive score when all features are used. The train data was then subjected to a Decision Tree Classifier, which was subsequently assessed using the test data. The one vs. all approach was used in this instance because our dataset includes several class levels. This resulted in a 74% accuracy rate, which is marginally higher than what we obtained with the KNN model. Thirdly, the Logistic Regression Model was used to see if the current results might be improved further. Since multiclass categorization is the issue, a one-versus-all approach was used. The

accuracy of the logistic regression model was 84%. Therefore, it can be said that the Logistic Regression Classifier has achieved the best performance to yet, with an accuracy of 88% in predicting the DR level. Last but not least, the Support Vector Machine model was trained using the training data to see if it could perform on par with Logistic Regression. The accuracy of 89.3%, which is without a doubt the finest result of our investigation, was attained after using a few kernel gimmicks and a regularization value of C equals 5. At this stage, the support vectors were classified using a sigmoid kernel.

Therefore, compared to the other classifiers that were employed in our dataset, SVM and Logistic Regression exhibit superior performance.

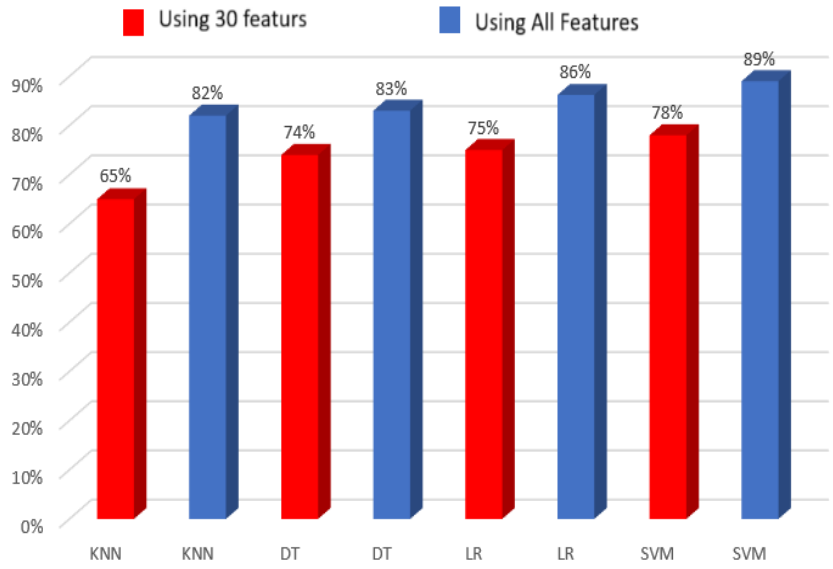


**Figure 3 Accuracy comparison**

## 5. Result and Discussion

Our study's objective was to develop a classification model that outperforms earlier research in this area in terms of metrics. The development of automated methods for categorizing the early symptoms of diabetic retinopathy has been the subject of an enormous amount of ongoing study. We employed three widely used machine learning metrics—accuracy, precision, and recall—to assess our work.

The contrast between previous research and our suggested approach is displayed in the table below.

| Reference | No. of classes | Method | Accuracy |
|---|---|---|---|
| Acharya et al. 2008 [10] | 5 | Higher order spectra | 82% |
| Acharya et al. 2008 [10] | 5 | Blood Vessel, Exudates, Microaneurysms, cotton wool spots | 86% |
| Proposed Method | 6 | Bright non-lesion, Hard exudates, Cotton wool spots, Red Non-lesion, Microaneurysms and Hemorrhages | 89% |

## 6. Conclusion

Building an automated system model that can effectively identify the early non-proliferative Diabetic Retinopathy (DR) symptoms in diabetic patients is the primary goal of this study. There is no known cure for DR. Although optical laser analysis is usually successful in preventing irreversible vision loss, it damages the retina. Early detection through screening is essential because the disease does not become irreversible until severe symptoms have developed. Thus, using supervised machine learning algorithms, this study suggested an automated technique for identifying and diagnosing early signs of diabetic retinopathy, such as microaneurysms, exudates, cotton wool patches, etc. By using a tree-based feature selection method instead of the pointless feature-building of previous approaches, the suggested method also outperforms them. In comparison to the current efforts, the suggested work also significantly improves precision and recall at the hierarchy level. The superiority of the suggested work to create an automated tool for diabetic retinopathy is demonstrated by all of these factors.

## REFERENCES


1. Wu, W. Zhu, F. Shi, S. Zhu, and X. Chen, "Automatic detection of microaneurysms in retinal fundus images," Computerized Medical Imaging and Graphics, vol. 55, pp. 106–112, 2017.
2. D. J. Browning, Diabetic retinopathy: evidence-based management. Springer Science & Business Media, 2010.
3. R. Maher, S. Kayte, and D. M. Dhopeshwarkar, "Review of automated detection for diabetes retinopathy using fundus images," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 3, pp. 1129–1136, 2015.
4. R. Maher, S. Kayte, D. Panchal, P. Sathe, and S. Meldhe, "A decision support system for automatic screening of non-proliferative diabetic retinopathy," International Journal of Emerging Research in Management and Technology, vol. 4, no. 10, pp. 18 24, 2015.
5. B. Singh and K. Jayasree, "Implementation of diabetic retinopathy detection system for enhance digital fundus images," International Journal of advanced technology and innovation research, vol. 7, no. 6, pp. 874–876,
6. E. M. Shahin, T. E. Taha, W. Al-Nuaimy, S. El Rabaie, O. F. Zahran, and F. E. A. El-Samie, "Automated detection of diabetic retinopathy in blurred digital fundus images," in Computer Engineering Conference (ICENCO), 2012 8th International. IEEE, 2012, pp. 20 25.
7. M. Gandhi and R. Dhanasekaran, "Diagnosis of diabetic retinopathy using morphological process and SVM classifier," in Communications and Signal Processing (ICCSP), 2013 International Conference on. IEEE, 2013, pp. 873–877.
8. N. Thomas and T. Mahesh, "Detecting clinical features of diabetic retinopathy using image processing," International Journal of Engineer-ing Research & Technology (IJERT), vol. 3, no. 8, pp. 558–561, 2014.
9. Nayak, J., Bhat, P. S., Acharya, U. R., Lim, C. M., and Kagathi, M., Automated identification of different stages
10. Acharya, U.R., Tan, P. H., Subramaniam, T., Tamura, T., Chua, K. C., Goh, S. C., Lim, C. M., Goh, S. Y., Chung, K. R., and Law, C. Automatic Identification of diabetic type 2 subjects with and without neuropathy using wavelet transform on pedobarograph. J. Med. Syst. 32(1):21–29, 2008.
11. Lee, S. C., Lee, E. T., Wang, Y., Klein, R., Kingsley, R. M., and Warn, A., Computer classification of nonproliferative diabetic retinopathy. Arch. Ophthalmol. 123(6):759–764, 2005.

**12.** Bihis, Matthew; Roychowdhury, Sohini, "A generalizedflow for multi-class and binary classification tasks: An Azure ML approach," in Big Data (Big Data), 2015 IEEE International Conference on , vol., no., pp.1728

**13.** S. M. Mazinani and K. Fathi, "Combining KNN and Decision Tree Algorithms to Improve Intrusion Detection System Performance," International Journal of Machine Learning and Computing, vol. 5, no. 6, pp. 476–479, 2015.

**14.** Logistic Regression for Machine Learning," Machine Learning Mastery, 06-Apr-2019. [Online]. Available: https://machinelearningmastery.com/logistic regression-for-machine-learning/.

**15.** L. Chen and L. Chen, "Support Vector Machine Simply Explained," Towards Data Science, 07-Jan 2019. [Online]. Available: https://towardsdatascience.com/support-vector machine-simply-explained-fee28eba5496

**16.** M. Shamsujjoha, and T. Bhuiyan, "A Content-based Image Retrieval Semantic Model for Shaped and Unshaped Objects", Journal of Computer Engineering, vol: 18, no: 1, pp:43-60, year: 2016

**17.** H. Abedy, F. Ahmed, M. N. Q. Bhuiyan, M. Islam, M. N. Y Ali, and M. Shamsujjoha "Leukemia Prediction from Microscopic Images of Human Blood Cell Using HOG Feature Descriptor and Logistic Regression", 16th Int'l Conf. ICT & Knowledge Engineering, pp: 1-6, year: 2018