International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 05 | May - 2021 ISSN: 2582-3930

Detection of Diabetic Retinopathy Using Machine Learning

Bharat Naga Sunil Karri[1], Appikatla Chaitanya Sai Krishna[2],VaradaKowsik[3], Tetali Lakhsman Reddy[4]

[1,2,3,4]UG Scholar, Department of Electronics & Communication Engineering, GITAM University, Vizag, India

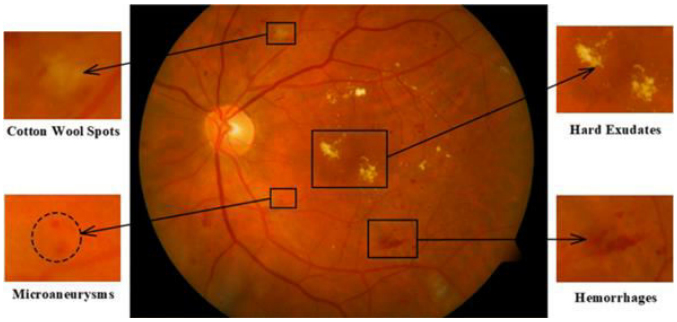
---

Abstract-Diabetic Retinopathy(DR) is one of the leading causes of sight inefficiency for diabetic patients. The clinical diagnostic results and several outcomes of eye testing methods revealed a set of observations that eases the decision-making in diabetic retinopathy for the doctor, therapist. Machine learning, a branch of artificial intelligence, is applied in clinical data analytic. It can detect patterns in data and then use these uncovered patterns to predict future data or perform some decision-making under uncertainty. In the case of DR, finding the correlation between the depth of affection and the clinical result is critical, as several parameters need to be considered for optimal decision-making by the therapist. We formulate the correlation matrix and find the depth of the relationship between the parameters. We will discover the accuracy of each algorithm and apply the algorithm with the highest precision and deploy the model.

Key Words:Machine Learning, Diabetic Retinopathy, Algorithms, Accuracy, correlation, Model

1.INTRODUCTION

Diabetic Retinopathy(DR) is a diabetic eye disease that affects light-sensitive tissues like the Retina that lines the back of the eye. It is a common cause of vision loss among people with diabetes and one of the leading causes of blindness among working-age adults. It refers to retinal changes in patients suffering from diabetes mellitus. With the increase in the life expectancy of diabetes, the incidence of Diabetic Retinopathy has also increased. Diabetic Retinopathy is a leading cause of blindness. This disease is also known as diabetic eye disease. Diabetic Retinopathy affects the retinal blood vessels and causes them to bleed or leak fluids, thus distorting the vision. Diabetic Molecular Edema is a consequence of Diabetic Retinopathy, causing the swelling in the Retina called the Macula. Therefore, early detection, timely treatment, and follow-up care are essential to protect against vision loss. Poor metabolic control is less critical than duration but is nevertheless relevant to the development and progression of Diabetic Retinopathy. The sex ratio is more in females than in males (4:3). Pregnancy may accelerate the changes in Diabetic Retinopathy. Hypertension, when associated, may accelerate the changes of diabetic Retinopathy. Other risk factors include smoking, anemia, obesity, and hyperlipidemia. Various Machine Learning algorithms are applied to clean data, and accuracy is calculated for each algorithm.



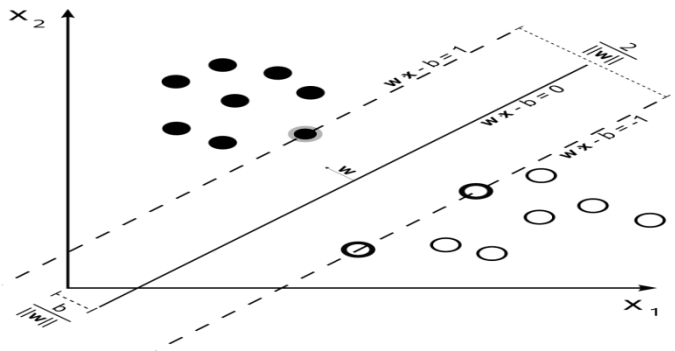
2. Machine Learning

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.[2] Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

APPROACHES

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning



PYTHON

Python is an interpretedhigh-levelgeneral-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

LIBRARIES

Python's large standard library, commonly cited as one of its greatest strengths, provides tools suited to many tasks. For Internet-facing applications, many standard formats and protocols such as MIME and HTTP are supported. It includes modules for creating graphical user interfaces, connecting to relational databases, generating pseudorandom numbers, arithmetic with arbitrary-precision decimals, manipulating regular expressions, and unit testing.

Some parts of the standard library are covered by specifications (for example, the Web Server Gateway Interface (WSGI) implementation wsgiref follows PEP 333), but most modules are not. They are specified by their code, internal documentation, and test suites. However, because most of the standard library is cross-platform Python code, only a few modules need altering or rewriting for variant implementations.

```
from scipy.io import arff

import pandas as pd

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.linear_model import LogisticRegression

from sklearn import svm

from sklearn.neighbors import KNeighborsClassifier
```

DATASET

This dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. All features represent either a detected lesion, a descriptive feature of an anatomical part, or an image-level descriptor.

We collected 1151 instances of data from the UCI repository which consists of 1151 rows and 20 columns.

Explanation of each column:

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 indicates lack.

2-7) The results of MA detection. Each feature value stands 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 for different image Sizes.

16) The Euclidean distance of the center of the macula and the center of the optic disc 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 cl

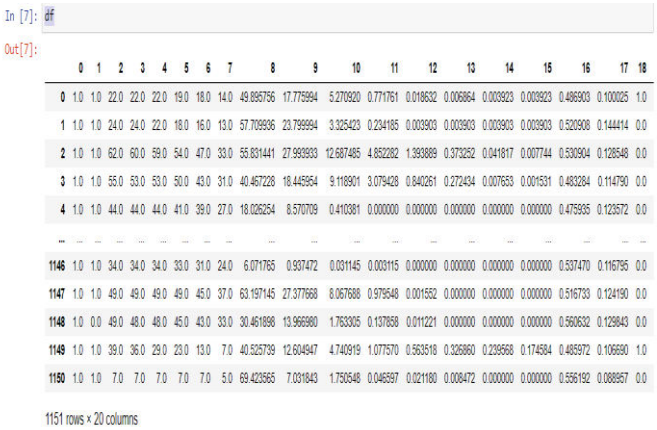


Fig -1: Data Set

Charts

The count of individuals is separated by who are suffering from Diabetic Retinopathy and who are not is represented visually using seaborn.

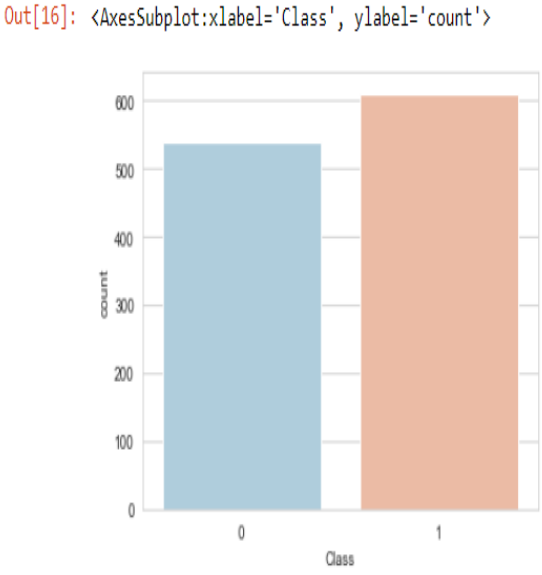


Fig -2: Data Visualization using Seaborn

Correlation Matrix: The relation between the columns is calculated by correlation factor.The depth of relation between the class and column can be seen visually.

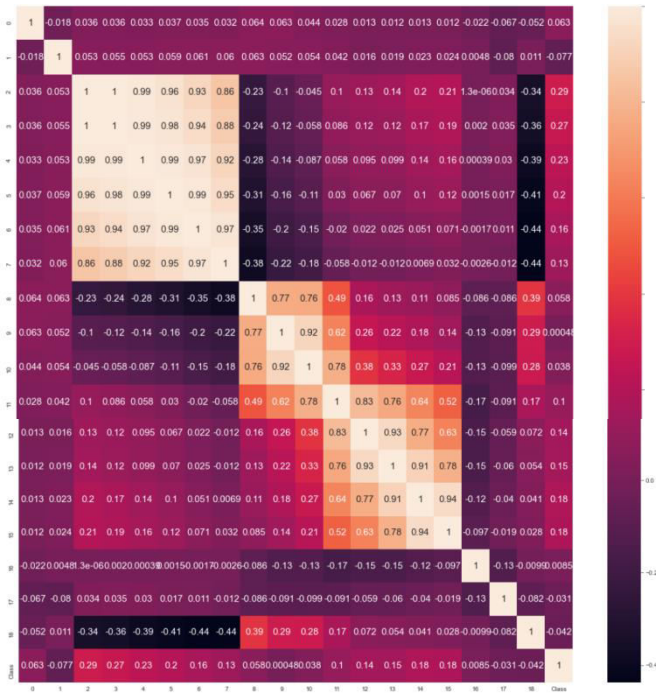


Fig -3: Correlation Matrix

K NEAREST NEIGHBOURS:

In statistics, the *k*-nearest neighbors algorithm (*k*-NN) is a non-parametricclassification method first developed by Evelyn Fix and Joseph Hodges in 1951,and later expanded by Thomas Cover. It is used for classification and regression. In both cases, the input consists of the *k* closest training examples in the data set. The output depends on whether *k*-NN is used for classification or regression:

- In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

Logistic Regression:

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

Gradient Boosting:

- Gradient Boosting and Random Forest algorithms are ensemble techniques.
- In gradient boosting the learning happens by optimizing the loss function.
- A Gradient Boosting Machine combines the predictions from multiple decision trees to generate the final predictions.

Random Forest:

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

The working of the algorithm can be better understood by the below example:

```
In [55]: rfc=RandomForestClassifier()

In [56]: rfc.fit(x1_train,y1_train)
Out[56]: RandomForestClassifier()

In [57]: y_pre=rfc.predict(x1_test)

In [58]: rfc.score(x1_train,y1_train)
Out[58]: 0.8760869565217392
```

Fig -4: Result

CONCLUSION

As Machine Learning created a huge impact on the health sector we used various Machine Learning algorithms on the dataset using python.By using this model we can detect Diabetic Retinopathy in a person with an accuracy of 87%.The algorithm used in this model was Random Forest which is an ensemble technique.We checked the model by giving sample data as input and it has predicted successfully.

REFERENCES

[1] Balint Antal, Andras Hajdu, " An ensemble-based system for automatic screening of diabetic retinopathy, Knowledge-Based Systems 60",2014, 20-27

[2] M.Akhil Jabbar, B.L Deekshatulua, Priti Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm",CoRR, volume abs/1508.02061, arXiv:1508.02061,2015.

[3] Mihaela GHEORGHE, "A Support Vector Machine Approach For Developing Telemedicine Solutions: Medical Diagnosis", Network Intelligence Studies, Issue 5, pp.43-48, 2015.

[4] V. Anuja Kumari, R.Chitra, "Classification of Diabetes Disease Using Support Vector Machine",International Journal of Engineering Research and Applications. Vol. 3, pp. 1797-1801, ISSN: 2248-9622,2013.