Classification of Retinopathy using Machine Learning

Swetha S[1], Abarna A[2], Sornadevi G[3], Kamalesh S [4]

[1]-[3]Student, Department of IT, Velammal College of Engineering and Technology, Tamilnadu, India.
[4]Professor, Department of IT, Velammal College of Engineering and Technology, Tamilnadu, India.

Abstract -*Retinopathy is the most common cause of blindness of the eye depends on hypertension, diabetes, prematurity of a person. For this reason, early detection of retinopathy is of critical importance. In this study, the digital image processing and a machine learning-based approach plays an important role in the medical field for detection of various symptoms of diseases for the early detection of retinopathy from retinal images. Here we are using feature extraction and machine learning algorithm "K-Nearest Neighbor" used for predicting and classifying disease on retinal images. The accuracy is based on the algorithm which we are going to use and the datasets and splitting them into train set and test set.*

***Key Words*: Diabetic retinopathy, Hypertensive retinopathy, Retinopathy of prematurity, Healthy retinopathy Image processing, Convolution neural network.**

1. INTRODUCTION

Retinopathy is any damage to the retina of the eyes, which may cause vision impairment. Retinopathy often refers to retinal vascular disease, or damage to the retina caused by abnormal blood flow. Retinopathy occurs when blood vessels in the back of the eye, the retina, become damaged. When the blood vessels become damaged they can leak and these leaks can cause dark spots on our vision. The main causes of retinopathy tend to be sustained high blood glucose levels and high blood pressure.

Hypertension is probably the best known systemic condition associated with non-diabetic retinopathy. In people with hypertension, retinopathy is often referred to as hypertensive retinopathy, although this definition has sometimes been expanded to include retinal arteriolar signs such as arteriovenous nicking and focal and generalized arteriolar narrowing. Retinopathy has been found to be present in about 11% of hypertensive non-diabetic people over 43 years of age. An appreciable proportion (average 6%) of normotensive non- diabetic people may also have retinopathy.1-3 In two population based studies, more than 50% of the participants with non- diabetic retinopathy did not have a history of hypertension.[2],[3] Retinopathy may thus represent the cumulative effects of elevated blood pressure throughout life in people not classified as having hypertension.

Diabetic retinopathy also known as diabetic eye disease is a medical condition in which damage occurs to the retina due to diabetes mellitus. It is a leading cause of blindness in developed countries.

[1] Diabetic retinopathy affects up to 80 percent of those who have had diabetes for 20 years or more.[2] At least 90% of new cases could be reduced with proper treatment and monitoring of the eyes.[3] The longer a person has diabetes, the higher chances of developing diabetic retinopathy.[4] Each year in the United States, diabetic retinopathy accounts for 12% of all new cases of blindness. It is also the leading cause of blindness in people aged 20 to64.

The first stage, called non-proliferative diabetic retinopathy (NPDR), has no symptoms. Patients may not notice the signs and have 20/20 vision. The only way to detect NPDR is by fundus photography, in which microaneurysms(microscopic blood-filled bulges in the artery walls) can be seen. If there is reduced vision, fluorescein angiography can show narrowing or blocked retinal blood vessels clearly (lack of blood flow or retinal ischemia).

Macular edema, in which blood vessels leak their contents into the macular region, can occur at any stage of NPDR. Its symptoms are blurred vision and darkened or distorted images that are not the same in both eyes. Ten percent (10%) of diabetic patients will have vision loss related to macular edema. Optical Coherence Tomography can show areas of retinal thickening due to fluid accumulation from macular edema.

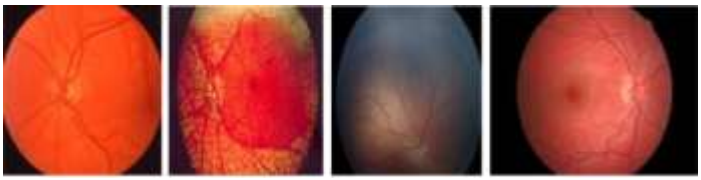
[5]In the second stage, abnormal new blood vessels (neovascularization) form at the back of the eye as part of proliferative diabetic retinopathy (PDR); these can burst and bleed (vitreous hemorrhage) and blur the vision, because these new blood vessels are fragile. The first time this bleeding occurs, it may not be very severe. In most cases, it will leave just a few specks of blood, or spots floating in a person's visual field, though the spots often go away after a few hours.

These spots are often followed within a few days or weeks by a much greater leakage of blood, which blurs the vision. In extreme cases, a person may only be able to tell light from dark in that eye. It may take the blood anywhere

from a few days to months or even years to clear from the inside of the eye, and in some cases the blood will not clear.

Retinopathy of prematurity (ROP), also called retrolental fibroplasia (RLF) and Terry syndrome, is a disease of the eye affecting prematurely born babies generally having received intensive neonatal care, in which oxygen therapy is used on them due to the premature development of their lungs. It is thought to be caused by disorganized growth of retinal blood vessels which may result in scarring and retinal detachment. ROP can be mild and may resolve spontaneously, but it may lead to blindness in serious cases. Thus, all preterm babies are at risk for ROP, and very low birth-weight is an additional risk factor. Both oxygen toxicity and relative hypoxia can contribute to the development of ROP.

Healthy corresponds to people who do not suffer from any of the above retinopathy types.



a)Hypertensive b)Diabetic C)Premature D)Healthy

Fig-1: Retinopathy fundus images

2. LITERATURE SURVEY

2.1Retinal vascular development in premature infants

Retinopathy of prematurity (ROP) is a proliferative retinal vascular disease affecting the premature infant with an incompletely vascularized retina. The spectrum of ophthalmological findings in ROP exists from minimal sequel, which do not affect vision, to bilateral retinal detachment and total blindness. With the increased survival of very small infants, retinopathy of prematurity has become one of the leading causes of childhood blindness. Over the past two decades, major advances have been made in understanding the pathogenesis of ROP, to a large extent as a result of changes in clinical risk factors (oxygen and non-oxygen related) and characteristics observed in ROP cases. This article provides a literature review on the evolution in clinical characteristics, classification and treatment modalities and indications of ROP. Special attention is hereby paid to the neonatal factors influencing the development of ROP and to the necessity for everyone caring for premature babies to have a well-defined screening and treatment protocol for ROP. Such screening protocol needs to be based on a unit- specific ROP risk profile and, consequently, may vary between different European regions.

2.2 An international classification of retinopathy of prematurity

Because of modern life-support systems capable of keeping tiny premature infants alive, retinopathy of prematurity has recurred. No classification system currently available adequately describes the observations of the disease being made today. A new classification system, the work of 23 ophthalmologists from 11 countries, is presented in an attempt to meet this need. It emphasizes the location and the extent of the disease in the retina as well as its stages

2.3 Automated detection of diabetic retinopathy on digital fundus images

Diabetic retinopathy (DR) is a condition where the retina is damaged due to fluid leaking from the blood vessels into the retina. In extreme cases, the patient will become blind. Therefore, early detection of diabetic retinopathy is crucial to prevent blindness. The main stages of diabetic retinopathy are non-proliferate diabetes retinopathy (NPDR) and proliferate diabetes retinopathy (PDR). In this study, we propose a system for automated classification of normal, and abnormal retinal images by automatically detecting the blood vessels, hard exudates micro aneurysms, entropy and homogeneity. The objective measurements such as blood vessels area, exudates area, micro aneurysms area, entropy and homogeneity are computed from the processed retinal images. These objective measurements are finally fed to the artificial neural network (ANN) classifier for the automatic classification. Different approaches for image restoration are tested and compared on Fundus images. The effect of restoration on the automatic detection process is investigated in this paper

2.4 U-net: Convolutional networks for biomedical image segmentation

There is large consent that successful training of deep networks requires many thousand annotated training samples. In this paper, we present a network and training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. Using the same network trained on transmitted light microscopy images (phase contrast and DIC) we won the ISBI cell tracking challenge 2015 in these categories by a large margin. Moreover, the network is fast. Segmentation of a 512x512 image takes less than a second on a recent GPU.

3. PROPOSED WORK FOR RETINOPATHY

In this section, the proposed system for retinopathy is described. Classifying retinal disease based on machine learning technique “K-nearest neighbor algorithm”.

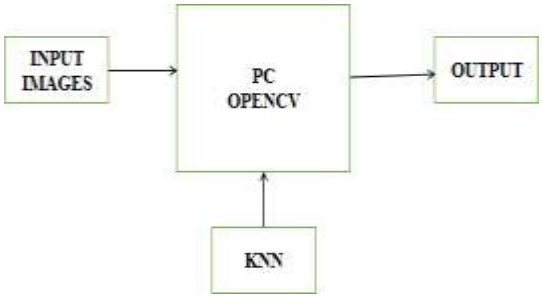
Segmenting colors from different types for retinopathy and classifying it using feature extraction, perimeter finding and train the machine using KNN algorithm. Testing the accuracy of a disease using confusion matrix and classify the disease with above 97% accuracy.

A novel architecture of convolutional neural networks (CNN) is proposed to recognize the existence and severity of ROP. The architecture is composed of a feature extract sub-network, followed by a feature aggregate operator to bind features from variable images in an examination. The prediction is accomplished using a second sub-network with the aggregated features as inputs. Max and mean aggregate operators are explored based on the architecture. Several ImageNet pertained networks are tested in the study, including VGG-16, Inception- V2, and ResNet50. The proposed architecture is verified with a large dataset of 2668 examinations of the fungus in infants. The experimental results demonstrate that the Inception-V2 with the max aggregate operator in module 2 is a proper network architecture for the recognition of the existence and severity of ROP. Compared with the mean aggregate operator, the max has better classification accuracy and convergence speed. Meanwhile, a patient's multiple examinations in train, validation, and test datasets has little impact on model's performance, mainly because the characteristics of the eyes of the premature infants are varied overtime.

3.1 System architecture

The input images with certain features, with all of those in the training data set, the model comes to a decision by judging the image with the severity of KNN and produce training image features

3.2 Disadvantages of existing system



- CNN gives many frameworks.
- Only predict the stages of retinopathy of prematurity
- Processor should be high configuration
- CNN has problem of over fitting and its mostly computationally expensive because it has to take a large database set for training

4. IMPLEMENTATION

4.1Image Segmentation

Segmentation is the process of assigning a label to every pixel. In other words, the segmentation is partitioning a digital image into multiple segments "pixels." The goal of segmentation is to simplify the representation of an image into something that is more meaningful and easier to analyze. Whereas the result of image segmentation is a set of segments that cover the entire image, or a set of contours extracted from the image. The simplest method of image segmentation is called the threshold method. This method is based on threshold value to turn a gray scale image into a binary image. During this process, every pixel in an image is called as object pixel if the value is greater than the threshold value and it is named as background pixel if the value is lower than the threshold value. An object pixel is being given a "1" value while the background pixel is given the "0" value. After which a binary image is being created with all the object and background pixels.

4.2 Feature Extraction

Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.

When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called feature selection.[1] The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data

4.3 Dataset Preparation:

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question.

In the open data discipline, data set is the unit to measure the information released in a public open data repository. The European Open Data portal aggregates more than half

a million data sets.[2] In this field other definitions have been proposed,[3] but currently there is not an official one. Some other issues (real-time data sources,[4] non-relational data sets, etc.) increases the difficulty to reach a consensus about it.

4.4 Import the Dataset

A lot of datasets come in CSV formats. We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program) and read it using a method called read_csv which can be found in the library called pandas.

After inspecting our dataset carefully, we are going to create a matrix of features in our dataset (X) and create a dependent vector (Y) with their respective observations. To read the columns, we will use iloc of pandas (used to fix the indexes for selection) which takes two parameters — [row selection, column selection].

4.5 Taking Care of Missing Data in Dataset

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data.

The library that we are going to use for the task is called Scikit Learn preprocessing. It contains a class called Imputer which will help us take care of the missing data.

From sklearn .preprocessing import Imputer

A lot of the times the next step is to create an object of the same class to call the functions that are in that class. We will call our object imputer. The Imputer class can take a few parameters —
i. missing values — We can either give it an integer or "NaN" for it to find the missing values.
ii. strategy — we will find the average so we will set it to mean. We can also set it to median or most_frequent (for mode)as necessary.
iii. axis — we can either assign it 0 or 1, 0 to impute along columns and 1 to impute along rows.
imputer = Imputer(missing_values = "NaN", strategy = "mean", axis = 0)

4.6 Encoding Categorical Data

Sometimes our data is in qualitative form, that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data.

This is an example of categorical data. In the first column, the data is in text form. We can see that there are five categories — Very, Somewhat, Not very, Not at all, Not sure — and hence the name categorical data.

So the way we do it, we will import the scikit library that we previously used. There's a class in the library called LabelEncoder which we will use for the task from sklearn.
Preprocessing import LabelEncoder
As I have mentioned before, the next step is usually to create an object of that class. We will call our object labelencoder_X.
labelencoder_X = LabelEncoder()

4.7 Splitting the Dataset into Training Set and Test Set

Now we need to split our dataset into two sets — a Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set.

4.8 Feature Scaling

It is a method used to standardize the range of independent variables or features of data.
But why is it necessary? A lot of machine learning models are based on Euclidean distance. For example, the values in one column (x) is much higher than the value in another column (y), (x2-x1) squared will give a far greater value than (y2-y1) squared. So clearly, one square difference dominates over the other square difference. In the machine learning equations, the square difference with the lower value in comparison to the far greater value will almost be treated as if it does not exist. We do not want that to happen. That is why it is necessary to transform all our variables into the same scale. There are several ways of scaling the data. One way is called Standardization which may be used.

4.9 K-Nearest Neighbor Algorithm in Python

k-Nearest Neighbors, or KNN for short, is one of the simplest machine learning algorithms and is used in a wide array of institutions. KNN is a non-parametric, lazy learning algorithm. When we say a technique is non-parametric, it means that it does not make any assumptions about the underlying data. In other words, it makes its selection based off of the proximity to other data points regardless of what feature the numerical values represent. Being a lazy learning algorithm implies that there is little to no training phase. Therefore, we can immediately classify new data points as they present themselves.

4.10 K- NEAREST NEIGHBOUR

knn=KNeighborsClassifier(n_neighbors=5,
metric='euclidean')
knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)

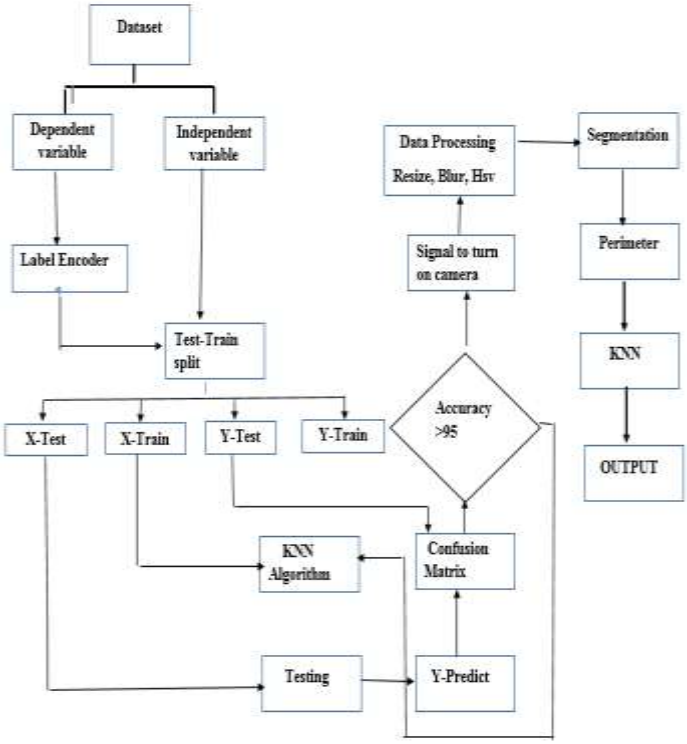
KNN Algorithm

Load the data
Initialize K to your chosen number of neighbors
3. for each example in the data
3.1 Calculate the distance between the query example and the current example from the data.
3.2 Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels
8. If classification, return the mode of the K labels

Advantages

The algorithm is simple and easy to implement.
There's no need to build a model, tune several parameters, or make additional assumptions.
The algorithm is versatile. It can be used for classification, regression, and search (as we will see in the next section).

4.11 Flowchart :



5. RESULTS AND DISCUSSION

In this paper, we have used K-nearest neighbor algorithm with accuracy of 97% to classify the different types of retinopathy using the concept of machine learning.

Feature extracted from the image with the help of Huge Saturated Value .This is followed by differentiating the colors present in the image. Finally the Retinopathy is classified .

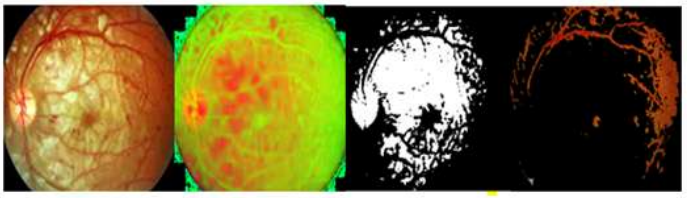


Fig-2: Feature extraction of diabetic image

6. FUTURE PERSPECTIVE:

Future treatment will probably involve early vitreous sampling followed by an injection of multiple combined medications tailored to the individual's vitreous signature. Use of risk profiles will facilitate improvements in risk factor control, reduce incidence and progression of DR, and reduce costs. In addition, manipulation of pharmacokinetic drug properties may allow for a longer intravitreal half-life and less need for frequent injections. As the duration and

efficacy of intravitreal drugs improve, the use of treatment modalities that destroy retinal tissue, such as focal laser photocoagulation or pan-retinal photocoagulation, may no longer be necessary.

7. CONCLUSION

We have tried to construct an ensemble to predict if a patient has retinopathy using features from retinal photos. After training and testing the model the accuracy we get is quite similar. Datasets is providing higher accuracy rate for predicting retinopathy. Despite the shortcomings in reaching good performance results, this work provided a means to make use and test multiple machine learning algorithms and try to arrive to ensemble models that would outperform individual learners. It also allowed exploring a little feature selection, feature generation, parameter selection and ensemble selection problems and experiences the constraints in computation time when looking for possible candidate models in high combinatorial spaces, even for a small dataset as the one used. The structure of our research has been built in such a way that with proper dataset and minor alternation it can work to classify the disease in any number of categories.

REFERENCES

[1] B. C. Chu, and I. Y. Wong, "Incidence and risk factors for retinopathy of prematurity in multiple gestations: A chinese population study,"
Medicine, vol. 94, no. 18, pp. 185–191, 2015

[2] A. Gschlieer, E. Stifter, T. Neumayer, E. Moser, A. Papp, N. Pircher, G. Dorner, S. Egger, N. Vukojevic, and I. Oberacher-Velten, "Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity," American Journal of Ophthalmology, vol. 160, no. 3, pp. 553–560, 2016.

[3] D. E. Worrall, C. M. Wilson, and G. J. Brostow, "Automated retinopathy of prematurity case detection with convolutional neural networks," in International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, 2016, pp. 68–76.

[4] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fifine tuning," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.