DIABETIC RETINOPATHY DIAGNOSIS USING IMAGE PROCESSING AND MACHINE LEARNING TECHNIQUES

S.Jawahar[1], C.Vishnu[2], J.G Vijay Praveen[3], S. Karthic Kumar[4]

[1]Assistant Professor, Department of Computer Science and Applications, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu.

[2,3,4,5]Student, Department of Computer Science and Applications, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu.

Mail id: [1]shivamjawahar@gmail.com,
[2]vishnuchandran6998@gmail.com,[3]vijayguru1422000@gmail.com,
[4]karthickumar823@gmail.com

Abstract: Diabetic Retinopathy is a disease in human eye that affect people with diabetics. DR will lead to damage in retina of eye and may eventually cause completed blindness. Early detection of this disease is essential to avoid blindness. Thereare sufficient treatment for Diabetic Retinopathy are available through it requires early analysis and regular monitoring of diabetic patients. There are many physical tests like visual acuity test, pupil dilation and optical coherence tomography are used to identify Diabetic Retinopathy but they are time consuming. The main objective of this paper is to give decision about the occupation of Diabetic Retinopathy in Diabetes patients by implementing the machine learning classifier algorithm on feature derived from the results of different retinal images. It will produce accurate result for the presence of Diabetic Retinopathy disease in Diabetic patients based on prediction. Decision making for predicting the existences of Diabetic Retinopathy in diabetic patients is performed by using Logical Recursion algorithm.

Keywords: Diabetic Retinopathy (DR), learning classifier algorithm, Logical Recursion algorithm.

1. INTRODUCTION

Diabetics is a chronic disease that take places when the pancreas failed to produce enough insulin in the body. Diabetics cause the damage of retina in eye. Those effect is known as Diabetic Retinopathy. It is caused by damage to the blood vessels in the tissue at the back of the eye known as retina. Retinal blood vessels entering the retina from the optic circles are damaged which results in the loss of vision. In the initial stage of Diabetic there will be no change in vision, but with lack of diabetics control may cause in complete loss of vision. Diabetic Retinopathy occurs about 1.8 million from 37 million people. Diabetic Retinopathy can be classified as Non-Proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR). Based on the presence of features on the retina, the results are identified. Early stage of Diabetic Retinopathy is considered as Non-Proliferative Diabetic Retinopathy. They are identified by the signs of micro aneurysm, exudates and hemorrhage. Micro aneurysms are small red spots presence in the retina due to capillary swellings. Exudates are white or yellow drops caused leaking fluid from the capillary.

Hemorrhage are red spots caused due to rupturing of capillaries and micro aneurysms. The final stage of diabetic retinopathy is called as Proliferative Diabetic Retinopathy. They are characterized by enormous amount of growth of new blood vessels in the retina. Those leakage of blood vessels will lead to the loss of sights for the diabetic patients. Currently, the detection of DR is a tedious, time consuming and manual process that requires a trained clinician to consider and figure out the digital color fundus photographs of the retina.

2. RELATED WORK

The number of people with diabetic retinopathy is increasing day by day. It is predicted that the number will increases from 126.6 million to 191.0 million by 2030 and the vision-threatening diabetic retinopathy (VTDR) will also increase from 37.3 million to 56.3 million, if doesn't take any proper action. In spite of growing evidence recording the effectiveness of routine Diabetic Retinopathy screening and early medication, it is commonly leads to poor visual functioning and represents the leading cause of blindness. Most of the time it has been ignored in health care and in many low-income countries because of insufficient medical service. As there is inadequate ways to detect about diabetic retinopathy, we will build a system which will give prediction about diabetic retinopathy. Thus, we decided to use Machine Learning Algorithms for the prediction of this disease.

As improvements have taken place, a ton of studies and research have obtained on programmed persistence of diabetic retinopathy applying various process and increasingly exact procedures. The following papers were studied and deeply resolve in order to develop our project. The papers which are mentioned below are published as a part of renowned journals and were truly helpful in understanding the idea of Diabetic Retinopathy.

3. PROPOSED SYSTEM

The paper presents an automatic approach which accelerate the detection and classification of the diabetic retinopathy disease. A low power microscope named ophthalmoscope or the fundus camera is attached with a digital camera and captures the inner part of the eye which consists of retina, optic circle, macula and the blood vessels. A modified digital back unit (color video camera) is attached to the fundus camera to convert the fundus image into a digital image.

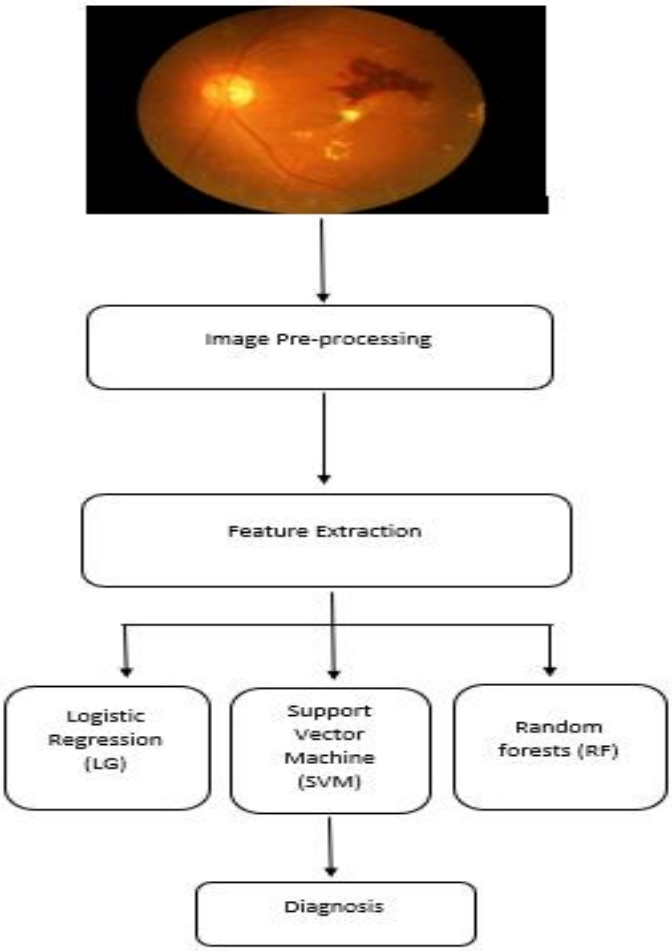

Fig. 1: Proposed Model

The images are usually obtained from the posterior pole's view including the optic circle and macula. The propounded model deals with numerous image processing techniques and machine learning methods to segment the fundus input image properly. The proposed system classifies the image into four stages:

1. Mild non proliferative retinopathy - Small areas of balloon like swelling in the retina's tiny blood vessels which is considered as microaneurysms will be occur at this earlier stage of the disease. These microaneurysms may leak fluid into the retina. Those leakage may lead to swelling of the macula.

2. Moderate non-proliferative retinopathy - As the evolution of disease, blood vessels that nourish the retina may swell and distort. They may also lose their ability to transfer the blood. Both actions will lead to the characteristic changes to the appearance of the retina.

3. Severe non-proliferative retinopathy – In this stage, many more blood vessels are blocked depriving blood supply to areas of the retina. These fields may secrete growth factors that signal the retina to develop new blood vessels.

4. Proliferative diabetic retinopathy (PDR) - At this final and advanced stage, growth factors secreted by the retina produce the generation of new blood vessels, which grow along with the inner side of the retina and into the vitreous gel, the fluid which is fills in the eyeball. The new blood vessels are weak, which makes them more likely to leak and bleed. PDR besides scar tissue can also contract and cause retinal detachment that pulling away of the retina from underlying tissue, like wallpaper peeling away from a wall. Retinal detachment can lead to permanent vision loss.

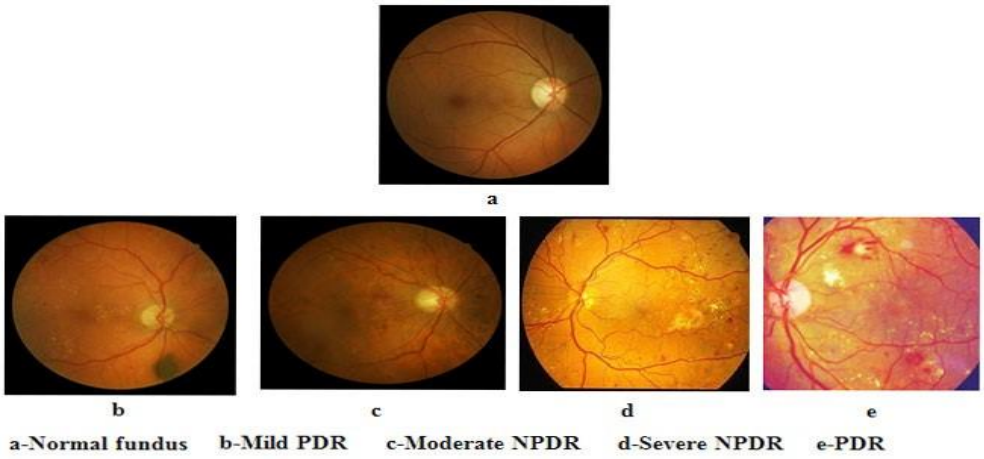

Fig. 2: Stages of Diabetic Retinopathy

Optic circle: Optic circle is the part of the posterior pole where the vasculature and retinal nerve axons enter and leave the eye. The Optic circle in active retinal image usually emerges as a bright yellowish and elliptical object marked by outer vessels. The presence of pathologic changes take place at the site.

Macula: The macula is an elliptical-shaped pigmented area near the center of the retina of the human eye. It has a width of around 5.5 mm. The fovea is placed near the center of the macula. It is a tiny hole that consists of the largest absorption of conoid cells. Hence it is responsible for the central, high-resolution, color vision that is possible in good light and this kind of vision will be affected if the macula is damaged.

Exudates: Exudates is fluid that leaks out of the blood vessels. Yellow flecks in the eye is considered as exudates. They are the lipid residues of serious leakage from impaired capillaries. The common cause is diabetes. They are also known as pus.

Microaneurysms: A microaneurysms is a small swelling that forms in the wall of tiny blood vessels. These small swellings may break and also blood to leak into the nearby tissues. They are sometimes found in the retina of eye. Increase in microaneurysms will lead to the increase in the risk of diabetic retinopathy as well.

Hemorrhages: Retinal hemorrhage is ooze from the blood vessels in the retina, inside the eye. The retina is the thin layer that lines the back of your eye. Medical conditions such as diabetes, high blood pressure, anemia, or leukemia cause eye problems such as macular degeneration, or a bulging of the blood vessels in the retina. They may have no symptoms. It may have a sudden or gradual loss of vision, ranging from mild to severe.

4. PROPOSED ALGORITHMS

There are so many algorithms for machine learning to detect the diabetic retinopathy, it is not possible to use all of them for analysis. For this research paper, we will be using three of them random forest (RF), logistic regression (LG) and support vector machine (SVM).

4.1 Random Forest

Random forest algorithm can be useful for both classification and the regression kind of problems. It is supervised classification algorithm which creates the forest with a number

of trees. In general, the most number of trees in the forest will results the more robust the forest looks like. It could be also said that the higher the number of trees in the forest gives the high accuracy of the results. There are many advantages in using random forest algorithms. The classifier will handle the missing values. It will also model the random forest classifier for categorical values. The over fitting problem will never come when we use the random forest algorithm in any classification problem. It can be used for feature engineering which means identifying the most important feature out of the available feature from the training dataset.

4.2 Support Vector Machine

The Support Vector Machine (SVM) is a linear classification method discovered in 1992 by Boser, Guyon, and Vapnik. A more formal definition is that a support vector machine constructs a hyper plane or set of hyper planes in a high or infinite-dimensional vector space, which can be used for classification, regression, or other tasks. It can be used to generate multiple separating hyperplanes such that the data is divided into segments and each segment contains only one kind of data. It is completely based on optimization and not based on statistics.

Support vector machine contains all the data that is essential for the classification problem, since even if all the other vectors are evacuated the solution will returns the same, the optimization problem that is used to find the optimal hyperplane and the decision function can be declared in dual models which depends only on dot products between vectors. SVM is utilized to classify fundus images into Normal, Non-Proliferative Diabetic Retinopathy and Proliferative Diabetic Retinopathy classes.

4.3 Logistic regression

Logistic regression is a supervised classification algorithm used to allow information to individual set of classes. It is a predictive analysis algorithm and it is based on the concept of probability. It is similar to linear regression model but the logistic regression uses a more complex cost function, this cost function can be defined as the sigmoid function. It is used for classification problems. It is also used a linear equation with independent predictors to predict a value. The predicted value can be somewhere between negative infinity to positive infinity.

5. DATASET

One of the major headache issues when developing machine learning or data science application is the data to be used, because it's the base that the application will be built on, so the data must be exact and large enough to develop a good model. To solve this common issue there are many websites that provide datasets for machine learning applications, they gather a large amount of data in different domains, organize and categorize them in a way that to be effective for the implementation in these applications.

In our project we have used a dataset that is obtained from the UCI (Unique Client Identifier) Machine Learning depository. This dataset contains features extracted from Messidor image set to predict whether an image contains signs of diabetic retinopathy disease or not. We have seen different types of datasets in kaggle, github and other websites which was used for different kind of projects based on diabetic retinopathy. As we wanted to work with detection of diabetic retinopathy, this dataset will be appropriate for our work as it has different types of characteristics. Outputs are represented in binary numbers. The value of "1" indicates that the patient has diabetic retinopathy and the value of "0" indicates absence of the disease in the diabetic patients.

The extracted data is split into two parts as training the data and testing the data. It is an important part in evaluating the data by using data mining techniques. Most of the time when we are dividing the data set into two different parts, most of the data is used for training, and only a smaller portion of the data is used for testing. We have split the dataset into two different parts. First part is used for training and another part is used for testing. The training set contains a known output and the model learns on this data in order to generate the accurate result. After the model has been processed in the training set, we have another method called testing, it tests the model by making predictions against the test set. Because the data in the testing set already consists of the known data that we need to predict, it is easy to resolve whether the model correct predict or not. We have also used 80% of our data for training and 20% for testing.

6. DETECTION

A clinician has classifies the presence of diabetic retinopathy in each image on a range of 0 to 4. In accordance with the following ranges:

0 - No DR

1 – Mild

2 – Moderate

3 - Severe

4 - Proliferative DR

6.1 Image preprocessing: Since the original images are adequately large and most of them incorporate adequately large significant black border. We started eliminating most of these black borders but before that, as we wanted square matrix images as the input for our system, the images were first resized to 3000 x 3000 by adding excess black borders and then resizing these images to 448 x 448. As if this does not taken correctly, we may have distorted FUNDUS images which may lose its initial circular shape.

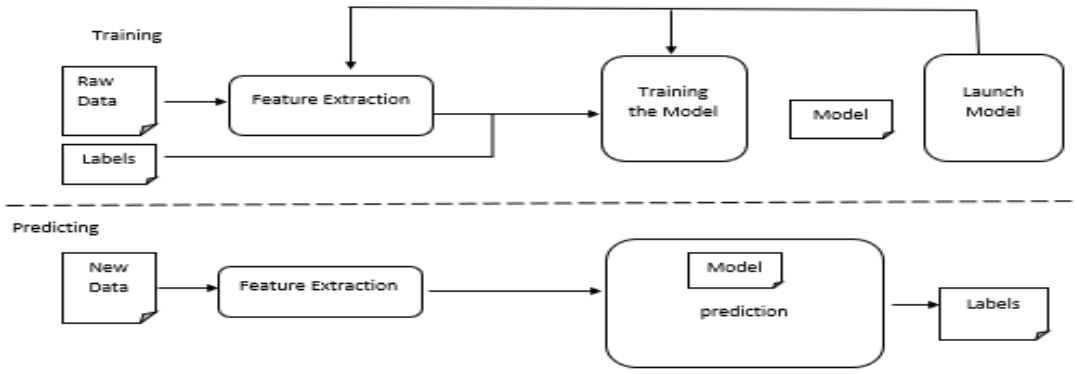

Fig. 3: Workflow of Diabetic Retinopathy Diagnosis

6.2 Cross validation: Cross validation is the step where the right parameters for the algorithm are selected. The problem of overfitting and underfitting is detected using cross validation. Generally, a machine learning problem has many input feature, so it is not possible to generate the data or the problems that might be arising. Using cross validation, such problems can be determined through the learning curves. The two main problems encountered are underfitting and overfitting.

6.3 Underfitting: Underfitting will occur when the algorithm cannot properly fit the training set. The learning curve produced is probably not complicated for the classification problems. Underfitting can be also called as high bias. To identify the presence of underfitting, the

learning curves wanted to be plotted. A learning curve with the training error and cross validation error also needs to be plotted. If both the training error and cross validation are high and there will be a tiny gap between the curves, it will be positively inferred that the algorithm has underfit the training set.

6.4 Overfitting: The easiest and most common method to remove over-fitting on image data is to artificially enlarge the dataset using label preserving transformations. We employed transformed images to be produced from the initial images with very little calculation. Overfitting occurs only when the algorithm fits the training set a bit too much and does poorly in the testing set. The algorithm fit the training set a bit too well, thus it was not able to visualize the dataset for unseen examples in the testing set. Hence overfitting can also be called as high variance.

6.5 Evaluation Metrics: The amount of accuracy for the network architecture is predicted by correctly classified Diabetic Retinopathy suffered images from the collection of images in the different dataset. Also, evaluate the algorithm which will be affected by overfitting or under-fitting could be visualized by plotting the training and validation loss. A whole objective is to minimize the cost function of the deep convolutional neural network outcomes necessarily reflected in the testing datasets.

In terms of diabetic retinopathy performance assessments, Specificity, Sensitivity and Accuracy are the essential parameters for determining the algorithms. Four parameters which will take place in measuring those performances. They are:

True Positive - Correctly identified Diabetic Retinopathy images.
True Negative - Correctly identified Non-Diabetic Retinopathy images.
False Positive - Number of Non-Diabetic Retinopathy images are identified wrongly as Diabetic Retinopathy images.
False Negative - Number of Diabetic Retinopathy images are identified wrongly as Non-Diabetic Retinopathy images.

At last, the Sensitivity, Specificity, and Accuracy are calculated for each fundus images available in the database.Sensitivity (true positive rate) measures that how likely the test is positive who has a diabetic retinopathy. Specificity (true negative rate) measures that how likely the test is negative who doesn't have the diabetic retinopathy. Positive predictive

value is also known as Precision. Accuracy measures both the diabetic and non-diabetic patients from the database.

7. TRAINING ACCURACY OF SVM ALGORITHM

SVM is a supervised learning algorithm which breaks down the model into hyperplanes and classifies the objects. SVM model can be utilized for class 2, class 3 or more problems, in our case we have class 3 problem and there are 3 classes are normal, NPDR and PDR. Training model is made for complete analysis of training data which tuned for extracting the features which can then be used for image classification. Features extracted from image are utilized by SVM for image classification. The image can be split into various se in terms of support vector classifiers. Kernel function is used to display data into hyperplane where hyperplane is detached between two classes. SVM can also perform non-linear classification by making use of non-linear kernel function. The non-linear kernel function maps data to higher dimensional space on which linear classifier can divide the input data.

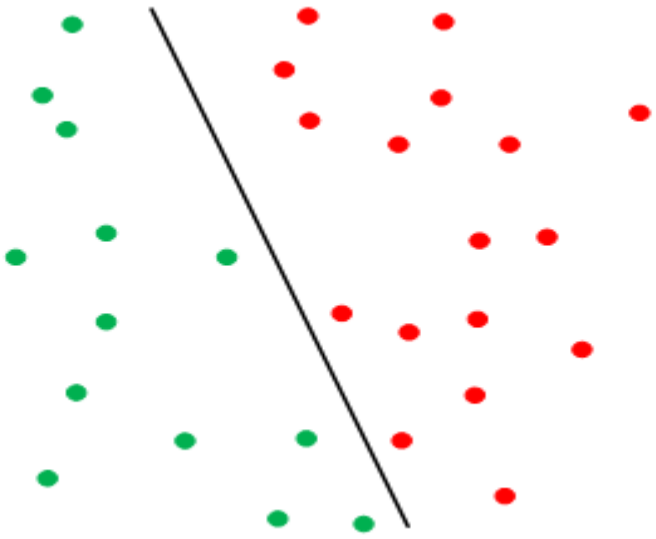

Fig.4: Architecture of SVM

For SVM algorithm we got training accuracy of 57.93%.We know Support Vector Machines a classifier that is defined using a separating Hyper plane between the classes. As SVM is capable of doing both classification and regression and also can capture much more complex relationships between data points we choose this algorithm.

7.1 TRAINING ACCURACY OF RANDOM FOREST ALGORITHM

The classifier will handle the missing values. It will also model the random forest classifier for categorical values. It can be used for feature engineering which means identifying the most important feature out of the available feature from the training dataset. For Random Forest our training accuracy is 66.09%. Random Forest can also be used for both classification and regression like SVM. Random forest gives an accuracy of 66.09% which also can be considered good for our work.

Displaying the result is when finally after all of the processing is done, the graphical user interface displays the output in text whether the person is affected with diabetic retinopathy or not. If affected, the both convey the symptoms also regarding the severity of the Diabetic Retinopathy is finally detected and displayed on to the Graphical User Interface respectively so that the patient can know about the situation of his sickness currently faced by the potential patient.

8. RESULTS

The outcomes of the project are the image processing techniques achieved on the input fundus image and this can be optionally disabled. In the result, we can see the graphical user interface displays the input image of the patient on its interface. From the plot of loss, we can see that the model has comparable performance on both training and testing datasets. If these parallel plots start to evacuate consistently, it might be a sign to stop training at an earlier stage. If the lines of train-test loss seem to gather to the same value and are closed at the end, then the classifier has high bias. Otherwise the lines are quite far apart, and then we have a low training set error but high validation error, then your classifier has too high variance. From these we can conclude that our train-test loss model training set loss is low and our test set error is not too high. So, from this it can be said that we have a good train-test accuracy model.

Since our model is very large (having fifteen convolutional layers and two dense layers), it is not easy to train such model on our personal laptops. Therefore, we tried two to simplify our model (using few layers only) and train it on a moderately very small dataset with even less dimension of each image. We also simplified our labels to only 2 classes (not having DR - class 0; having DR - class 1). In our research, the transformed images are developed in Python code on the CPU. The data augmentation consists of producing vertical and horizontal impressions with 50% probability.

9. CONCLUSION

We have implemented our entire model as an application on mobile phones, so that makes diabetic retinopathy diagnosis easier and time-saving. Till now, we have not tested the entire test data, therefore, our first target would be to obtain the same. Our main target would be to design such a detection system which is highly accurate and precise. This architecture has some setbacks such as an additional stage augmentation are needed for the images taken from a different camera with a different field of view. Also, our network architecture is complex and computation-intensive requiring high-level graphics processing unit to process the high-resolution images when the level of layers stacked more.

Currently in the scope of this paper, the processing time of diabetic retinopathy diagnosis is kept in mind, so as to deliver a graphical user interface to the user/patient, so that they can use their fundus images as input to the graphical user interface and thereby get to know whether they are suffering from diabetic retinopathy or not from a reliable source with faster processing time as well. Further upgrades later on might include the use of a superior estimation of something like a convolutional neural system which can help in arranging the images well than the present handled classifier individually. Apart from that, features like helpline or client manual ought to be given in the graphical User Interface, to be useful for the clients who probably won't be aware with creative progressions and utilization of the application.

REFERENCE

[1] Manisha Maliha, Ahmed Taeque and Sourav Saha Roy have employed a diabetic retinopathy using machine learning.

[2] R.Subhashini, T.N.R.Nithin and U.M.S.Koushik have presented a Diabetic retinopathy detection using image processing(GUI).

[3] VishakhaChandore and ShivamAsati have approached towards the concept of automatic detection of diabetic retinopathy using deep convolutional neural network.

[4] R.Priya and P.Aruna were presented a method for diagnosis of diabetic retinopathy using machine learning techniques.

[5] S. Jawahar and Dr. P. Sumathi," An Efficient K-mer counting and indexing method using Bloom filter for Biological DNA Sequences", Journal of Web Engineering, Vol.18, Issue.4, May 2019, ISSN: 1540-9589.

[6] S. Jawahar and Dr. P. Sumathi," A new method for detecting Fuzzy Tandem Repeats (FTR) using Levenshtein Distance for Biological data", International Journal of Research in Engineering, IT and Social Sciences, Vol.9, Issue.2, February 2019, ISSN: 2250-0588.

[7] S. Jawahar and Dr. P. Sumathi,"Fast and accurate identification of Short Tandem Repeats (STRs) using hash function in DNA sequences",International journal of Engineering and Technology (UAE), Vol.7, issue.4S2, December 2018, ISSN: 2277-3878.

[8] S. Jawahar and Dr. P. Sumathi, "The Risk Of Breast Cancer Associated With Brca1 And Brca2 Gene Genetic Mutation: A Review", International Journal of Management, IT and Engineering, Vol.8, Issue.8, August 2018, ISSN: 2249-0558.

[9] S. Jawahar and Dr. P. Sumathi, "An Efficient Contiguous Pattern Mining technique to predict mutations in breast cancer for DNA data sequences", International Journal of Pure and Applied Mathematics, Vol.119, No.15, July 2018, ISSN: 1314-3395.

[10] S. Jawahar and Dr. P. Sumathi, "A Survey on Sequential Generator Mining Algorithms", International Journal of Research in Engineering, IT and Social Sciences, Vol. 7, issue.12, December 2017, ISSN:2250-0588.

[11] Salman Sayed, Dr. Vandana Inamdar and Sangram Kapre developed a detection of diabetic retinopathy using image processing and machine learning.

[12] Khasanah, Sumardiyono, Phong Thanh Nguyenm, E.Laxmi Lydia and K.Shankar composed a exploration of retinopathy disease using machine learning methodology.