Turkish Journal of Computer and Mathematics Education Vol.12 No. 11 (2021), 3020- 3030
Research Article

Automated Diabetic Retinopathy Prediction using Machine Learning Classification Algorithms

Mrs. Pooja Rathi, Assistant Professor, Dept. of Computer Science,
St. Vincent Pallotti College, Raipur, India
E-mail: rathipooja.08@gmail.com

Dr. S. M. Ghosh, Professor, Department of Computer Science,
Sir C.V. Raman University, Kota, Bilaspur, India
E-mail: samghosh06@rediffmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Diabetes is the most widely spread disease which takes place due to increase in blood sugar and when the body stops producing insulin. Diabetic Retinopathy (DR) is an ailment which has become a key cause of vision destruction in diabetic patients. It's an eye disease which causes damage to retina. To protect a patient from complete blindness, prediction of DR in the early stage is necessary.

In this study, we focus on early prediction of diabetic retinopathy in an effective and affordable manner. We have considered Messidor image dataset for diabetic retinopathy which contains 1151 records. The ultimate aim of our research is to explore whether there is existence of diabetic retinopathy; by employing machine learning classification algorithms (Logistic Regression, K nearest neighbour, SVM, bagged trees) on the features that are extracted from outcome of various retinal images from the image dataset. As the data may contain outliers and noisy values, two types of data validation have been applied: cross validation and hold-out validation. For getting the best possible outcome, dimensionality reduction criteria using Principal Component Analysis has also been applied. In this research, the accuracy of logistic regression resulted as highest; giving 75.1% in cross validation and 82.6% in case of hold-out validation. Consequently, our findings imply that Logistic Regression is best suitable for DR prediction. Likewise, bagged trees have also turned up with 80% accuracy.

Keywords: Diabetic Retinopathy, Machine Learning, Classification, Logistic regression, SVM, KNN, Bagged trees

1. Introduction

Diabetes is a chronic, unceasing and a widely spread disease that takes place when the pancreas does not produce enough insulin. It arises when our body stops producing sufficient insulin. Insulin is one type of hormone and its function is to help the cells inside our body to use glucose in food. Because of the insufficient insulin production, the level of glucose increases in the body. This leads a person to diabetes. With the increase in diabetes duration, the entire body gets affected, including retina. Because of the high blood sugar, fluid seeps from the retinal blood vessels and damages the retina. This is called as Diabetic Retinopathy (DR). It is amongst the foremost common eye diseases and is a prime cause of vision destruction. Human eye contains light sensitive tissues at the rear side of the eye, also called as small blood vessels. Long duration of diabetes becomes harmful for these blood vessels. Leakages of blood and fluid on the retina yields many complications like microaneurysms, blurred vision, haemorrhages, fluctuating vision, hard exudates, cotton wool spots etc. Contingent on the different symptoms and features appearing on the retina, DR is categorized into two groups: Non-proliferative diabetic retinopathy (NPDR) and Proliferative diabetic retinopathy (PDR). NPDR is a more common form in the diabetic patients where the walls of the blood vessels in the retina weaken. NPDR can progress from mild to moderate to severe, as more blood vessels become blocked. PDR is a severe type of DR. [1][2].

2. Prediction of DR using classification algorithms

At present, prediction of DR is a manual process which expends a more time and also requires expert ophthalmologists who are able to analyse retinal fundus images. Moreover, the chances of delayed treatment, miscommunications, improper diagnosis etc are increased.

Diabetic retinopathy has become very widespread and escalating rapidly due to scarcity in medical services and human negligence. As per the present situation, if appropriate actions are not taken, it is presumed that the number of DR cases will grow from 126.6 million to 191 million by 2030 [4]. The key factor for leading vision problems in DM patients is due to poor blood sugar control and unawareness about retinal damages [36]. (Pooja Rathi P. S., 2020)

It is extremely essential to control this increase. This paper focuses on prediction of Diabetic Retinopathy using classification algorithms. Training datasets can be provided to classification algorithms such as Logistic

Regression, K-Nearest Neighbours, Bagged trees, Support Vector Machines etc to train the system and then these algorithms can be used for prediction by comparing the actual data and training data.

Data mining is the method of extracting and studying secret trends in data for categorization into beneficial information from various perspectives [3]. It is an empirical method for analysing data in order to identify clear trends and systematic associations between variables [6] (Pooja Rathi, 2017) . The study of systems that can learn from data is known as machine learning [6]. In machine learning algorithms, learning is based on computational methods applied directly on data. It does not rely on pre-programmed systems or models. The learning and outcome of the algorithms improves with the increase in the amount of data and samples involved in the learning process. The major foundation of machine learning is to deal with data representation and data generalisation [7]. Representation is deriving and evaluating functions from different data occurrences and generalisation is proficiency to function precisely on unobserved and new occurrences of data from the previous learning experiences. Using probability distribution techniques, training data are used by the algorithms to generate a general model that enables to produce some defined and accurate predictions in new instances of the same data. Generalisation performance is estimated with reference to the ability to reconstruct proven knowledge from newer cases[15].

Machine learning is broadly categorise into two types [35] (RAY, 2017)

- Supervised learning
- Unsupervised learning
- Reinforcement learning

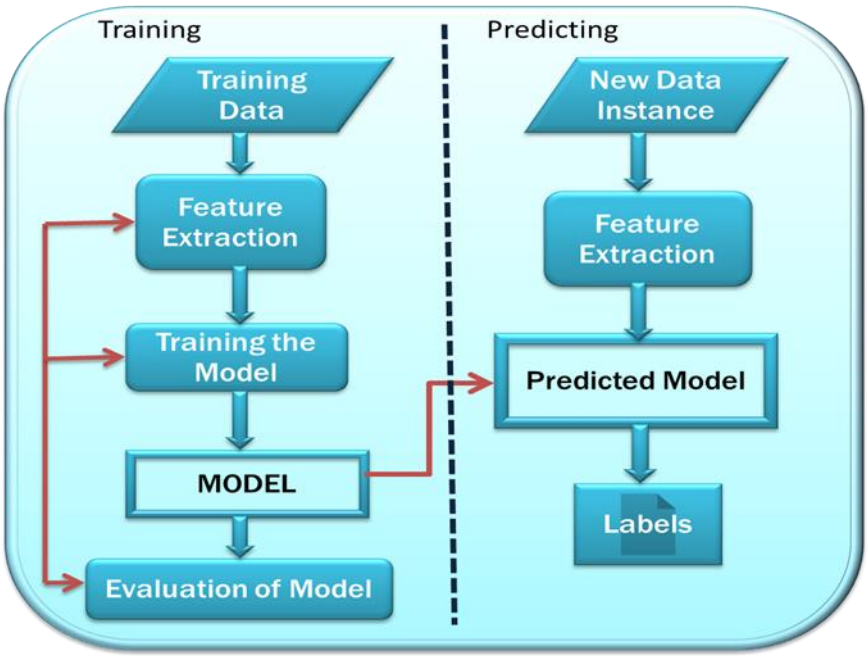

Figure 1: Workflow of Supervised Learning Model

Supervised learning is an approach to generate artificially intelligent systems where an algorithm has been trained on labelled input data paired with particular output. This system, also called a model, is trained to analyse the input data, detect the underlying patterns and relationships between the input data and output labels and finally, to produce the inferred function. The function has to predict the accurate outcome for any applicable input entity. This requires an algorithm to generalise from training data and yield accurate labelling results when working with unseen data [8].

3. Algorithms and Validations

3.1 Logistic Regression

Logistic Regression (LR) is a supervised Learning algorithm which is utilized for the classification problems. LR is a predictive analysis algorithm which follows the concept of likelihood and probability [11]. Given a set of independent variables, LR can be utilized for anticipating the categorical dependent variable.

To explain in simple words, the dependent variable is twofold in nature having information coded as either 1 (represents yes) or 0 (represents no)[10].

It is one of the handiest ML algorithms that may be used for diverse category problems together with junk mail detection, disease prediction and detection etc.

3.2 SVM

The objective of the SVM algorithm is to make the best line or choice limit that can isolate n-dimensional space into classes so we can without much of a stretch put the new information point in the right classification later on. This best choice limit is known as a hyperplane [13]. The point of the Support vector machine is to discover a hyperplane in a N-dimensional space (N — the number of features) that particularly characterizes the data points[14]. Objective is to track down a plane that has the maximum margin, i.e. the maximum distance between data points of both classes[18][19].

3.3 KNN

In KNN, classification is based on the concept of classifying the data points that are most similar in nature. Using the concept of "feature similarity", new data values will be compared with the similar types of values in the training set and then will be classified in the closely matching group. This means that KNN uses training dataset to make the predictions [16].

For every new instance, say P, prediction is performed by Predictions are made by looking for the whole training set for the K neighbours which are most similar instances to P and then recapitulating the output variable for those K instances. Distance Measures are used to reveal which of these K instances in the training dataset are most similar to a new input [15]. For real-valued input variables, Euclidean distance measure is widely used, however Manhattan distance, Minkowski distance methods are also in use.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

3.4 Bagged Tree

Bootstrap Aggregation is an ensemble method, a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model [23]. While prediction of the target values using ML techniques, the foremost causes of difference in actual and predicted values are variance, bias and Irreducible Error (noise). Ensemble helps to reduce these factors. Bagging is a concept wherein several decision tree classifiers collectively work which turn out to be better predictive execution than a single decision tree classifier [22]. In our study of predicting DR cases, bagging will be valuable as it will assist with making a numerous bootstrap tests; otherwise called base learners; and these samples will be consolidated for getting a resultant classifier which will result in more precise outcomes regarding prediction of DR and Non-DR patients. If we apply bagging to our DR dataset, the expected outcome can be achieved as this algorithm will help us in reducing the variance.

3.5 Cross Validation

Cross-validation is a procedure used for validating the model efficiency in which one part of dataset us used to train a model. When the model is trained and developed, the remaining part of dataset is used to evaluate this model to check the resulting outcome [25]. It is a method to check how a statistical model generalizes to an independent dataset. The general procedure is as follows [26]:

[1] Erratically arrange the dataset.

[2] Divide the dataset into k different groups

[3] For each group repeat the steps:

a. Randomly a test dataset (any group) is to be selected from these k groups

b. All other groups will be considered as a training dataset

c. Now, a model is to be applied

d. Evaluate this model on the test set that have been formed in (a)

e. Observe the evaluation accuracy and discard the model

[4] find out the efficiency of the model using the model evaluation scores

When the input features are more, data visualisation becomes difficult, owing to which problems detection, if any, becomes tough. Using learning curves in cross validation, these types of problems can be identified [15]. The two major problems that are encountered are underfitting and overfitting

3.5.1 Overfitting

Overfitting takes place when our ML algorithm used to build prediction models is very complex and it has over learned the underlying patterns in training data. Because of this, the model traps noise and erroneous values available in the dataset. These factors decrease the accuracy and efficiency of the algorithms intending the wrong results. The overfitted model has low bias and high variance [27].

3.5.2 Underfitting

Underfitting occurs when the algorithm used to build a prediction model is very simple and not able to capture complex patterns and trends from the training data. This results in failure to find the best fit of central outline in the data. In such cases, we will not get appropriate results in accuracy calculation on train as well as test data. This is underfitting. An underfitted model has high bias and low variance [28].

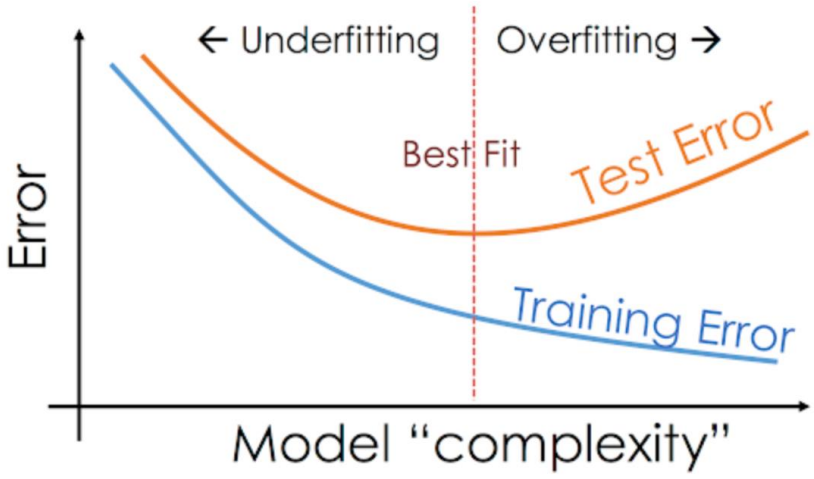

Figure 2: Curves showing Overfittng and Underfitting

4. Materials and Methods

4.1 Dataset

We have collected our dataset from UCI Machine Learning repository website. The dataset contains features extracted from the Messidor image set to predict whether an image has signs of diabetic retinopathy or not. Then features and labels of the dataset are identified [29]. The dataset is then partitioned into two sets, one for training in which majority of the information is utilized and the other one is testing. In training set 4 different class algorithms were equipped for the evaluation overall performance of the model. The algorithms we used are k-Nearest Neighbor, random forest, support vector machine and neural networks. After the framework has done with gaining knowledge from training datasets, more data is furnished without outputs. The last model generates the output by the use of the expertise it received from the algorithms on which it becomes trained. Finally, we get the accuracy of every set of rules and get to recognise which specific algorithm will provide us more precise outcomes for the prediction of diabetic retinopathy.

For the application of our work, we have collected a dataset from the UCI repository of machine learning which is available online.

The Messidor dataset has specially been developed for facilitating the researchers to explore the research in easy prediction of diabetic retinopathy. This dataset contains various features extracted from retinal fundus images using Messidor image set. These extracted features from a dataset that can be used for DR prediction.

Dataset contains altogether 20 columns where the first 19 fields are independent variables depicting different retinal features of a patient and the last column is a class column also called as output which indicates presence of Diabetic Retinopathy (1 for presence) or absence of DR (0 for absence). Total number of instances is 1151. The dataset has following features [29]

Table 1: Dataset at a glance [29] (Dua, 2017) (Balint Antal, April 2014)

Data Set Characteristics:	Multivariate	**Number of Instances:**	1151	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	20	**Date Donated**	2014-11-03
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	120264

Feature Information is as follows: (Dua, 2017) (Balint Antal, April 2014)

a) Quality :- The binary result of quality assessment. 0 = bad quality 1 = sufficient quality.

b) Pre-Screen :- The binary result of pre-screening, where 1 indicates severe retinal abnormality and 0 its lack.

c) nma.a – nma.f :- The results of MA detection. Each feature value stand for the number of MAs found at the confidence levels alpha = 0.5, . . . , 1, respectively.

d) nex.a – nex.h :- 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 different image sizes.

e) dd :- The Euclidean distance of the centre of the macula and the centre of the optic disc to provide important information regarding the patient's condition. This feature is also normalized with the diameter of the ROI.

f) dm :- The diameter of the optic disc.

g) amfm :- The binary result of the AM/FM-based classification.

h) class :- Class label. 1 = contains signs of DR (Accumulative label for the Messidor classes 1, 2, 3), 0 = no signs of DR.

We have calculated total count, minimum and maximum values, mean and standard deviation of all features in the dataset. We can use these different sorts of estimations dependent on the gathered information as an underlying advance towards creating surmising on the information. Standard Deviation is utilized to quantify the measure of changeability or scattering around a normal. Actually it is a proportion of instability. Dispersion is the contrast between the real and the normal value. The bigger this dispersion or variability is, the higher is the standard deviation. With the calculation of these measures, we are able to analyse the data in terms of noise, dispersion, variance etc.

Features →	quantity	Pre-screen	nma.a	nma.b	nma.c
Count	1151	1151	1151	1151	1151
Min	0	0	1	1	1
Max	1	1	151	132	120
Mean	0.9965	0.9183	38.4283	36.9096	35.1407
Std.Dev.	0.0589	0.2740	25.6209	24.1056	22.8054

Features →	nma.d	nma.e	nma.f	nex.a	nex.b
Count	1151	1151	1151	1151	1151
Min	1	1	1	0.349274	0
Max	105	97	89	403.93911	167.13143
Mean	32.2971	28.7472	21.1512	64.0967	23.0880
Std.Dev.	21.1148	19.5092	15.1016	58.4853	21.6027

Features →	nex.c	nex.d	nex.e	nex.f	nex.g
Count	1151	1151	1151	1151	1151
Min	0	0	0	0	0
Max	106.07009	59.76612	51.42321	20.098605	5.937799

Mean	8.7046	1.8365	0.5607	0.2123	0.0857
Std.Dev.	11.5676	3.9232	2.4841	1.0571	0.3987

Features →	nex.h	dd	dm	amfm	Class
Count	1151	1151	1151	1151	1151
Min	0	0.367762	0.057906	0	0
Max	3.086753	0.592217	0.219199	1	1
Mean	0.0372	0.5232	0.1084	0.3362	0.5308
Std.Dev.	0.1790	0.0281	0.0179	0.4726	0.4993

Figure 3: Statistical measures applied on data

5. Data visualization using Histograms:

For the better understanding of the dataset, we have applied the data visualization process on our data. In data visualization, data is graphically represented for analysis and to make data driven decisions. This process helped us to identify patterns, outliers and tendencies in our dataset. Histogram representation of each feature is depicted in the following figure. Histograms are the commonly used graphs to show frequency distributions [31]. With the help of this type of interpretation, we are able to graphically summarise the probability distribution of the continuous variables [30].

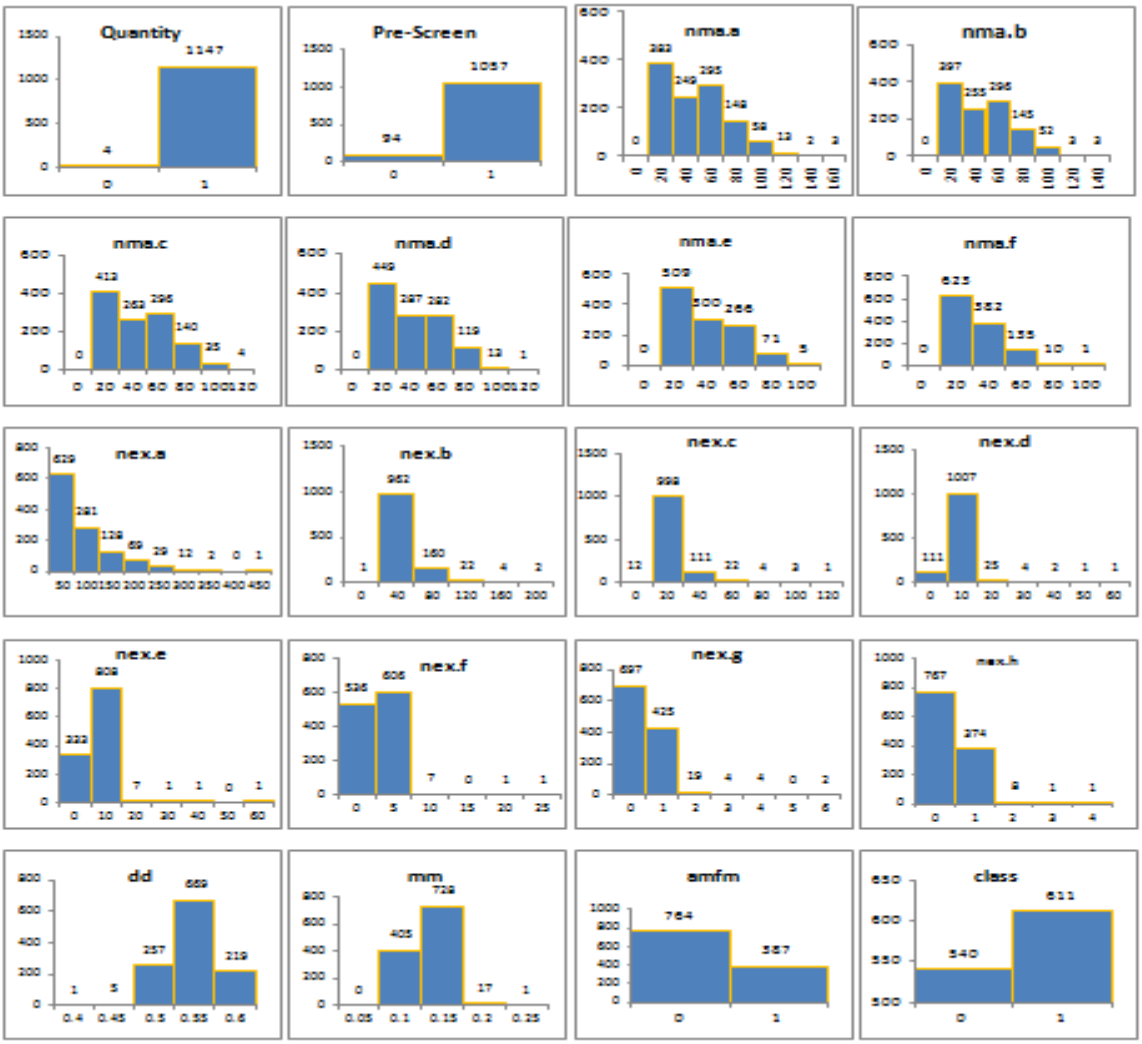

Figure 4 : Histogram representation of each feature

6. Experimental Setup

For the prediction of the diabetic retinopathy, we have applied five different classification algorithms on our dataset. For applying these algorithms, the dataset is seperated into a training set and testing set. To find out the best possible results, we have applied dimensionality reduction method on our dataset. Principal Component Analysis (PCA), is the most common method used for this.

Also data may contain some random noise which makes the pattern recognition difficult, may contain low frequency of some categorical variable, some incorrect values or may be few outliers. These types of irregularities will create difficulties to create a model. Data validation, although cannot directly find the problem, helps us to predict that there is some trouble with the consistency of the model. In order to find out the stability of the model, we have applied two types of data validations on our dataset.

Hold-out validation:

We chose to apply Hold-out validation as this will help us to partition our dataset into train and test sets. The model will be trained on training set and test dataset will be used to evaluate the performance of the model on the unseen data. We have used two different splits of the data as 10% hold out and 20% hold out where 10 % and 20 % data will be used as test set respectively and remaining will be used as training set [32][33].

Cross-validation

In cross validation, the dataset will randomly divide into 'k' groups. Amongst these 'k' groups, one group will be a test set and remaining will be training sets. Using these training sets, model will be trained and will further be analysed on the test set. This process will be repeated until each group has been used as test set. In our case we have applied 10-fold cross validation and 20-fold cross validation [32][33].

With these different validations, we are able to have better analysis of the model outcome, and in turn, better predictive model.

7. Algorithm Accuracies and Result Analysis:

7.1 For : Cross validation (10 fold and 20 fold)

PCA Applied for Number of Components to be 14, 16 and 18 on all algorithms

To predict the chances of diabetic retinopathy, we have selected various classification algorithms. These algorithms are logistic regression, support vector machine, K nearest neighbours and bagged trees. As explained in the machine learning algorithms section discussed above, these algorithms are feasible and suitable for our studies. In order to fit the best suitable model for our study as well as to prevent the data from underfitting and overfitting, we have applied two different types of validations on our dataset. These are: 'k' fold cross validation and Hold-out validation. Accuracies are computed by applying the algorithms with these validations and a comparative analysis has been discussed. In addition, for improving the algorithmic accuracy, the concept of dimensionality reduction by means of Principal Component Analysis has also been applied where the features are reduced to 18, 16 and 14. The accuracy outcomes for 10 fold and 20 fold cross validation are depicted in the accuracy table given below:

Table 2 : Accuracies with Cross Validation

Algorithm ⇩	**Validation ⇨**	**10 fold Cross Validation**	**20 fold Cross Validation**
	PCA ⇩	Accuracy	Accuracy
Logistic Regression	PCA 18	74.1	73.8
	PCA 16	75.1	74.1
	PCA 14	75.1	74.7
SVM	PCA 18	73	73.9
	PCA 16	74.2	74.1
	PCA 14	73.7	74.2
Fine KNN	PCA 18	62.9	61.7
	PCA 16	64.2	63.9
	PCA 14	64.1	64.3

Bagged tree	PCA 18	73	72.2
	PCA 16	74.2	73.1
	PCA 14	72.5	73.6

The above table depicts various algorithmic accuracies. It can be observed that Logistic regression when applied both with 10 fold and 20 fold cross validation with 14 principal components, gives highest accuracies as 75.1% and 74.7% respectively. The reason behind this is Logistic Regression is a characterization calculation used to discover the likelihood of occasion achievement or occasion disappointment (also called as success or failure) by estimating the exertion connection between the dependent variable (what we need to anticipate) and one or more independent variable (our features), by applying logistic function [34]. Since our dataset is linearly separable, it can construe model coefficients as pointer of feature significance. Likewise, SVM and bagged trees resulted in reasonable accuracies as 74.2 each, when applied with PCA 16.

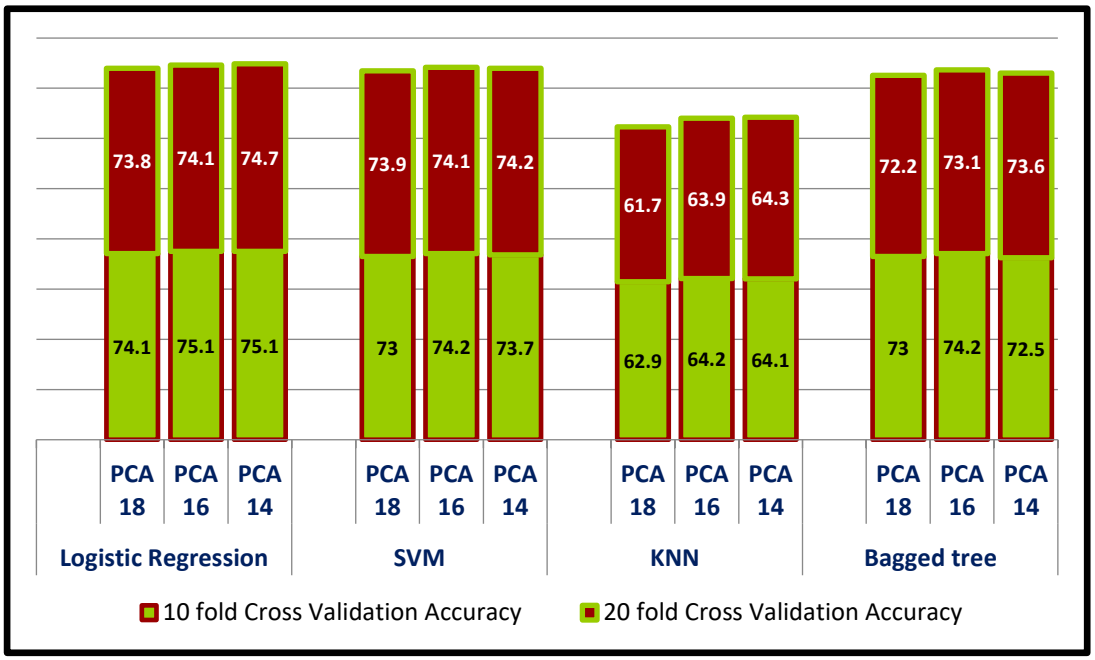

Figure 5: Graphical Representation: Accuracies with Cross Validation

7.2 For : Hold out validation (20% and 30% hold out)

PCA Applied for Number of Components to be 14, 16 and 18 on all algorithms

Since we have analysed the accuracies obtained after applying classification algorithms with cross validation on the dataset, now we have used hold out validation on our dataset. This process partitioned the dataset into train set and test set. The most common split is using 80% data as a training set and 20% as test a set. Using 70% data as a training set and 30% data as a test set is also in practice. In our experiment, we have used both the splits, for the better result analysis. As applied in cross validation above, the same structure of PCA is applied here also. The accuracy outcomes for 20% and 30% hold out validation are depicted in the accuracy table given below:

Table 3 : Accuracies with Hold-Out Validation

Algorithm ⇩	**Validation ⇨**	**10% hold out**	**20% hold out**
	PCA ⇩	Accuracy	Accuracy
Logistic Regression	PCA 18	81.7	75.7
	PCA 16	81.7	75.7
	PCA 14	82.6	76.1
SVM	PCA 18	80	70.4
	PCA 16	80.9	70.9
	PCA 14	81.7	73
Fine KNN	PCA 18	63.5	58.7
	PCA 16	69.6	66.1

	PCA 14	69.6	67.4
Bagged tree	PCA 18	78.3	77.8
	PCA 16	80	75.7
	PCA 14	79.1	76.1

When combined with the hold-out validation, the table above shows the algorithmic accuracies with different classification algorithms. The observation is that, with hold-out method, again, logistic regression has performed the best with 82.6% accuracy. Since this algorithm is discrete in nature, it worked well with our problem. Also, the performance of Support vector machine is good with 81.7% highest accuracy when applied with 14 principal components. When applied to our DR dataset, logistic regression gives a natural probabilistic view of class predictions. Bagged trees have also turned up with 80% accuracy which is helpful in decreasing the variance and eradicating the provocation of overfitting. In our case, the outcome of KNN is lowest.

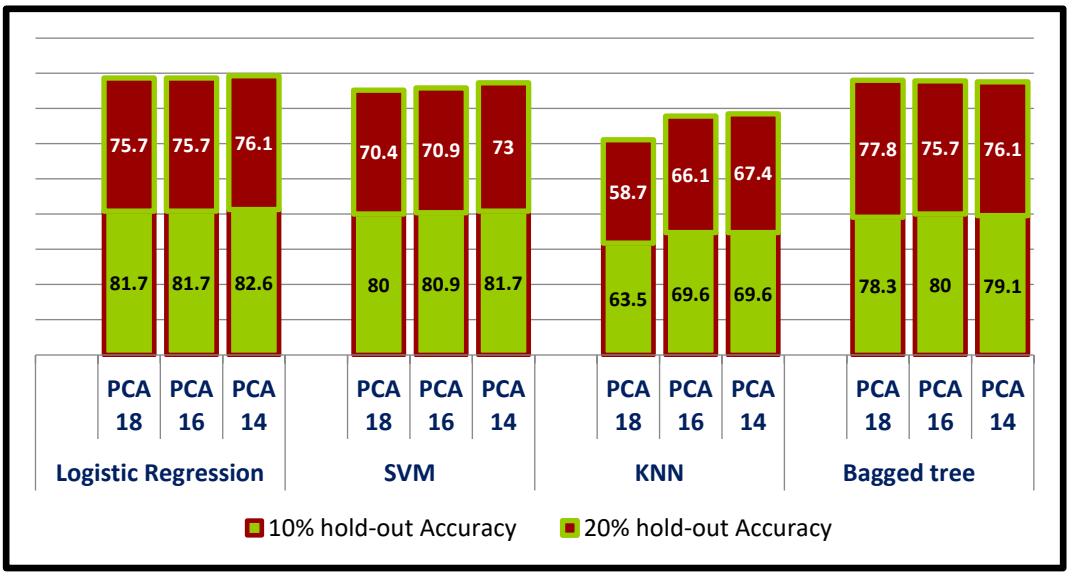

Figure 6: Graphical Representation: Accuracies with Cross Validation

Figure below shows collective analysis of all the algorithms with all validations and principal components

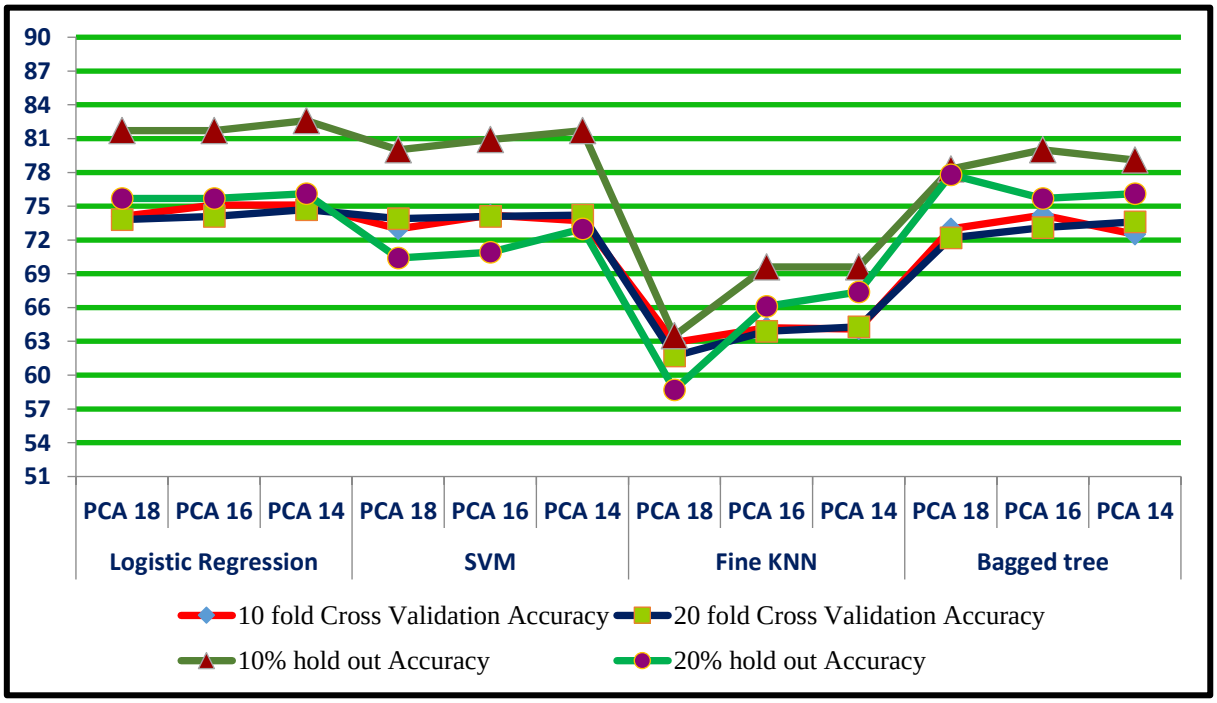

Figure 7: Collective analysis of all algorithms

8. Conclusion

In this study, we have used machine learning classification algorithms for the prediction of diabetic retinopathy using the Messidor dataset. This dataset contains features extracted from the Messidor image set for the prediction of signs of diabetic retinopathy. Four different classification algorithms have been applied on this dataset to attain the experimental results. As the data may contain noisy values, two types of data validation have been applied: cross validation and hold-out validation. This will help in reducing underfitting and overfitting of data and also in finding out a stable model. In this study, records of 1151 patients form our dataset. In cross validation, we have selected to apply 10 fold and 20 fold cross validation and in case of hold-out validation, we used 20% and 30% hold-out methods. For getting the best possible performance, dimensionality reduction criteria using Principal Component Analysis has also been applied. In this research, the accuracy of logistic regression resulted as highest; giving 75.1 in cross validation and 82.6% in case of hold-out validation. This work presented a method to effectively use and test distinct classification algorithms and try to build ensemble models that would outstrip distinct learners. This study also explores feature selection, extraction, data representation and ensemble selection problems and observes the outcome in specific period.

References

1. National-diabetes-and-diabetic-retinopathy-survey-2019, https://currentaffairs.gktoday. in/
2. Sara Cherchi, Alfonso Gigante, Maria Anna, Spanu, Pierpaolo Contini et al. "Sex-Gender Differences in Diabetic Retinopathy", Diabetology, 2020
3. Hayrettin Evirgen, Menduh Çerkezi, "Prediction and Diagnosis of Diabetic Retinopathy using Data Mining Technique", Turkish Online Journal of Science & Technology, Vol. 4 Issue 3, July 2014, pg.32-37
4. Yingfeng Zheng, Mingguang He, Nathan Congdon,The worldwide epidemic of diabetic retinopathy, Indian J Ophthalmol. 2012 Sep-Oct; 60(5): 428–431.
5. GauravSaxena, Dhirendra KumarVerma, AmitParaye, AlpanaRajan, AnilRawat; Improved and robust deep learning agent for preliminary detection of diabetic retinopathy using public datasets; Intelligence-Based Medicine, Volumes 3–4, December 2020, 100022
6. Mrs. Pooja Rathi, Dr. Anurag Sharma, " A Review Paper on Prediction of Diabetic Retinopathy using Data Mining Techniques", IJRIT, Vol 4, Issue 1,292-297, June-2017
7. Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M. et al. Deep learning applications and challenges in big data analytics. Journal of Big Data 2, 1 (2015). https://doi.org/10.1186/s40537-014-0007-7
8. Supervised learning by David Petersson, https://searchenterpriseai.techtarget.com /definition/supervised-learning
9. Dr.V.Ramesh1, R.Padmini2,"Risk Level Prediction System of Diabetic Retinopathy Using Classification Algorithms", IJSDR, Volume 2, Issue 6, June 2017
10. Machine Learning - Logistic Regression, https://www.tutorialspoint.com/
11. Introduction to Logistic Regression; https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148
12. Logistic Regression in Machine Learning; https://www.javatpoint.com/logistic-regression-in-machine-learning
13. Support Vector Machine Algorithm; https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
14. Support Vector Machine — Introduction to Machine Learning Algorithms SVM model from scratch; https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
15. Diabetic Retinopathy Detection Using Machine Learning By Maisha Maliha, Ahmed Tareque , Sourav Saha Roy
16. Wikipedia.http://en.wikipedia.org/wiki/Support_vector_machine.
17. Ben-Hur.A, Weston.J (2009) ."A User's Guide to Support Vector Machines". Data Mining Techniques for the Life Science.Humana Press. On Page(s): 223-239.
18. Akara S. ,BunyaritU.,SarahB.,Tom W.,Khine T. (2009)"Machine learning approach to automatic exudate detection in retinal images from diabetic patients" volume 57-issue 2.
19. Rajendra Acharya U.,E. Y. K. Ng, Kwan-Hoong Ng, Jasjit S. Suri (2012) "algorithms for the automated detection of diabetic retinopathy using digital fundus images" volume 36, Issue 1, pp

145–157
20. Varun G., Lily P., Mark C., "Development and validation of a deep learning Algorithm for Detection of Diabetic Retinopathy", December 2016.
21. Tiago T.G. "Machine Learning on the Diabetic Retinopathy Debrecen Dataset", knowledge-Based System60, 20-27. Published on June 25, 2016.
22. Bagged Trees: A Machine Learning Algorithm Every Data Scientist Needs, Robert Wood, https://towardsdatascience.com/bagged-trees-a-machine-learning-algorithm-every-data-scientist-needs-d8417ec2e0d9
23. Bagging and Random Forest Ensemble Algorithms for Machine Learning, by Jason Brownlee, https://machinelearningmastery.com/
24. Ensemble Learning — Bagging and Boosting, Jinde Shubham,https://becominghuman.ai/
25. Cross Validation in Machine Learning, https://www.geeksforgeeks.org/cross-validation-machine-learning/
26. Cross-Validation in Machine Learning, https://www.javatpoint.com/cross-validation-in-machine-learning
27. Overf) Overfitting and Underfitting in Machine Learning, https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning
28. Underf) Overfitting and Underfitting, ITBodhi, https://medium.com/@itbodhi
29. Data1) Balint Antal, Andras Hajdu: An ensemble-based system for automatic screening of diabetic retinopathy, Knowledge-Based Systems 60 (April 2014), 20-27.
30. Introductory Statistics, https://opentextbc.ca/introstatopenstax/chapter/histograms-frequency-polygons-and-time-series-graphs/
31. Know the "What, Where and How" of Histograms, https://www.cuemath.com/learn/histograms/
32. Hold-out vs. Cross-validation in Machine Learning, Eijaz Allibha, https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f
33. HOLDOUT CROSS-VALIDATION, by DataVedas | Jun 14, 2018 | Application in Python, Model Evaluation and Validation
34. https://machinelearning-blog.com/2018/04/23/logistic-regression-101/
35. Commonly used Machine Learning Algorithms (with Python and R Codes), Sunil Ray, September 9, 2017, https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
36. Rathi Pooja, Shrivastava Padmavati & Ghosh, S. (2020). Prediction of Diabetic Retinopathy Using Classification Techniques. Solid State Technology. 63. 9479.