Automated Diabetic Retinopathy Prediction Using Machine Learning Classifiers

Pooja Rathi[1,2]
1. Research Scholar, Rungta College of Engg. & Technology, Bhilai, India
2. Assistant Professor, St. Vincent Pallotti College, Raipur, India
Sourabh Rungta[3]
3. Professor, Department of Computer Science Engineering,
Rungta College of Engg. & Technology, Bhilai, India
S. M. Ghosh[4]
4. Professor, Department of Computer Science,
Sir C.V. Raman University, Kota, Bilaspur, India

Abstract:

Diabetes is the most common disease, and it develops when the body stops making insulin and blood sugar levels rise. One of the main conditions causing diabetic individuals to lose their vision is diabetic retinopathy (DR). It is an eye condition that affects the retina. The early detection of DR is essential for preventing total blindness in patients.This work concentrates on a cost-effective and early diagnosis of diabetic retinopathy. The study took into account the 1151-record Messidor imaging collection for diabetic retinopathy. By using machine learning classification algorithms (Logistic Regression, K Nearest Neighbor, SVM, bagged trees) on the features that are extracted from the results of various retinal images from the image dataset, the ultimate goal of this research is to investigate whether there is an existence of diabetic retinopathy. Cross validation and hold-out validation have both been used to validate the data because it may contain outliers and noisy numbers. Principal Component Analysis-based dimensionality reduction criteria have also been used to achieve the best results. In this study, logistic regression had the highest accuracy, scoring 77.2% in cross validation and 83.8% in the case of hold-out validation. The best method for DR prediction, according to the findings, is Logistic Regression.Likewise, bagged trees have also turned up with 80% accuracy.

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

DOI Number: 10.48047/NQ.2022.20.20.NQ109232 **NeuroQuantology2022;20(20): 2303-2317**

Introduction

Diabetes develops when the pancreas does not create enough insulin, and it is a persistent, widespread condition. It develops when our body is unable to produce enough insulin. One type of hormone, insulin, aids the body's cells in utilizing the glucose included in food. The body produces less insulin, which causes the level of glucose to rise. Diabetes results from this. The entire body, including the retina, is impacted by an increase in diabetes duration. High blood sugar causes fluid to leak from retinal blood vessels, damaging the retina. The term for this is diabetic retinopathy (DR). It is one of the most prevalent eye conditions and a leading cause of vision loss. The back of the human eye has light-sensitive tissues,

eISSN 1303-5150

commonly known as tiny blood vessels. Diabetes that lasts a long time damages these blood vessels. Numerous issues, including microaneurysms, impaired vision, haemorrhages, fluctuating vision, hard exudates, cotton wool patches, etc., can result from blood and fluid leaks on the retina. DR is divided into two classes based on the many signs and characteristics of the retina: Non-proliferative diabetic retinopathy (NPDR) and Proliferative diabetic retinopathy (PDR). When a diabetic patient develops NPDR, the walls of the blood vessels in the retina deteriorate. As additional blood vessels are blocked, the severity of NPDR can increase from mild to moderate to severe. PDR is a serious variety of DR. [1][2].

1. Prediction of DR using classification algorithms

Currently, predicting DR requires expert ophthalmologists who can analyze retinal fundus images and is a manual process that takes extra time. Additionally, there is a higher likelihood of delayed treatment, misunderstandings, incorrect diagnoses, etc.

Due to a lack of medical resources and carelessness on the part of individuals, diabetic retinopathy has spread widely and is getting worse quickly. According to the current condition, it is predicted that by 2030, there would be 191 million DR cases, up from the current 126.6 million [4]. Poor blood sugar control and ignorance of retinal damage are the main causes of vision issues in DM patients [36].

It is crucial to keep this increase under control. This study uses classification algorithms to forecast diabetic retinopathy. To train the system, classification algorithms like Logistic Regression, K-Nearest Neighbors, Bagged trees, Support Vector Machines, etc. can be given training datasets. Then, by comparing the real data with the training data, these algorithms can be used to make predictions.

Data mining is a technique for locating and analysing hidden patterns in data so that it can be categorised into useful information from a variety of viewpoints [3]. It is an empirical strategy for examining data in order to uncover clear trends and consistent correlations between variables [6]. Machine learning is the study of computer systems that can learn from data [6]. Learning is based on computational techniques that are applied directly to data in machine learning algorithms.

It does not rely on models or pre-programmed systems. With more data and samples used in the learning process, the algorithms learn better and produce better results. Data representation and data generalisation are the main tenets of machine learning [7]. Generalization is the ability to perform precisely on new and unobserved instances of data from the previous learning experiences, while representation is the process of deriving and evaluating functions from various data occurrences. The algorithms create a general model from training data using probability distribution techniques, which allows them to make precise predictions from new instances of the same data. The effectiveness of generalisation is measured in terms of the capacity to extrapolate established knowledge from more recent cases[15].

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

eISSN 1303-5150

- Supervised learning
- Unsupervised learning
- Reinforcement learning

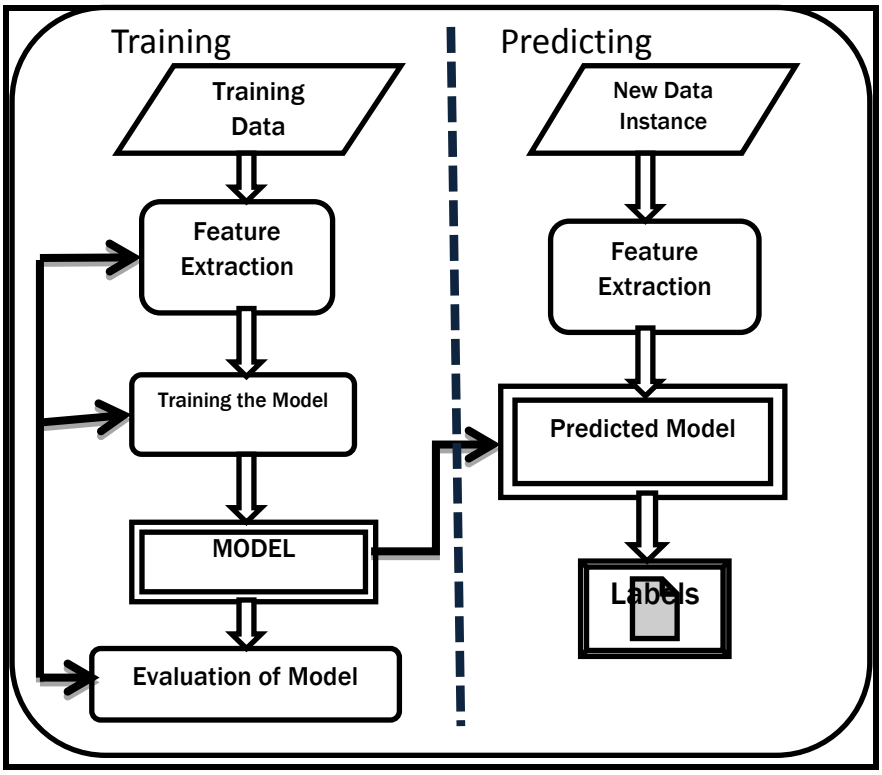


Figure 1: Workflow of Supervised Learning Model

A strategy to creating artificially intelligent systems called supervised learning uses an algorithm that has been trained on labelled input data combined with specific output. This system—also known as a model—is trained to analyse the input data, find the underlying patterns and connections between the input data and output labels, and then generate the inferred function. The function must correctly forecast the result for any relevant input entity. This necessitates the use of an algorithm that can generalise from training data and produce precise labelling results when applied to unexplored data [8].

2. Algorithms and Validations

3.1 Logistic Regression

For classification issues, the supervised learning algorithm logistic regression (LR) is used. LR is a predictive analysis algorithm that adheres to the ideas of probability and likelihood [11]. LR can be used to foresee the categorical dependent variable given a set of independent variables.To put it simply, the dependent variable has two components, each of which is coded as either 1 (which represents yes) or 0 (which represents no)[10].It is one of the most practical ML algorithms that can be applied to a variety of category problems, including the detection and prediction of diseases, junk mail, and other issues.

3.2 SVM

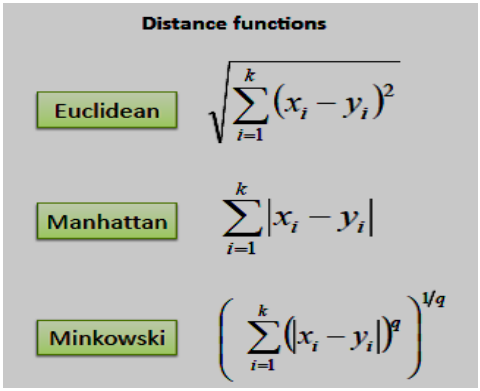
eISSN 1303-5150

The goal of the SVM method is to create the optimal line or option limit that can categorise n-dimensional space, allowing us to easily classify any additional information points in the future. A hyperplane is the name given to this best-option limit [13]. Finding a hyperplane in an N-dimensional space (N being the number of features) that specifically describes the data points is the goal of a support vector machine[14]. The goal is to locate a plane with the largest margin, or the greatest separation between data points belonging to both classes[18][19].

3.3 KNN

KNN bases classification on the idea of grouping data points that share the most characteristics. New data values will be compared with similar sorts of values in the training set using the notion of "feature similarity," and will then be classified in the group that closely resembles them. This indicates that KNN makes predictions using training data [16].

Every new occurrence, let's say P, is predicted by By searching the entire training set for the K neighbours that are the most similar P examples, and then recapitulating the output variable for those K instances, predictions are formed. To determine which of these K instances in the training dataset is most comparable to a new input, distance measures are used [15]. Euclidean distance measurement is frequently used for input variables with real values, while Manhattan distance and Minkowski distance techniques are also in use.



3.4 Bagged Tree

Bootstrap Aggregation is an ensemble method that aggregates predictions from various machine learning algorithms to get predictions that are more precise than those from any single model [23]. The main sources of discrepancy between actual and anticipated values when employing ML approaches include variation, bias, and Irreducible Error (noise). These factors are lessened by the ensemble. According to the idea of bagging, several decision tree classifiers working together produce better prediction results than a single decision tree classifier [22]. Bagging will be helpful in this study of predicting DR cases because it will help with the creation of numerous bootstrap tests, also known as base learners. These samples will then be combined to create a resultant classifier, which will produce more accurate results regarding the prediction of DR and Non-DR patients. The predicted result can be obtained if bagging is applied to this DR dataset because this method will assist us in lowering the variance.

eISSN 1303-5150

3.5 Cross Validation

When one portion of a dataset is used to train a model, cross-validation is a process used to verify the model's efficacy. The remaining portion of the dataset is utilised to evaluate the model once it has been trained and improved in order to verify the results [25]. It's a technique for determining how well a statistical model generalises to a different dataset. Following is the general process [26]:

[1] Arrange the dataset erratically.
[2] Assign each group in the dataset k.
[3] Repeat the procedure for every group:

- A test dataset (any group) is chosen at random from these k groups.
- All other groups will be used as a training dataset in the next step, which is to apply the model.
- Assess this model using the test set that was created in (a)
- Check the accuracy of the evaluation and omit the model.

[4] Use the model assessment scores to determine the model's effectiveness.

When there are more input features, data visualisation is challenging, making it challenging to identify problems, should there be any. These kinds of issues can be found using learning curves in cross validation [15]. Underfitting and overfitting are the two main issues that arise.

3.5.1 Overfitting

When the machine learning algorithm, which is used to create prediction models, is extremely complicated and it has overlearned the fundamental patterns in training data, overfitting occurs. As a result, the model captures noise and incorrect values included in the dataset. These elements reduce the efficacy and accuracy of algorithms that aim for incorrect outputs. Low bias and large variance characterise the overfitted model [27].

3.5.2 Underfitting

When the algorithm employed to create a prediction model is very basic and unable to extract intricate patterns and trends from the training data, underfitting occurs. As a result, the best fit of the core contour in the data cannot be found. In such circumstances, accuracy calculations on both train and test data will not produce the expected results. This is underfitting. High bias and low variance characterise an underfitted model.

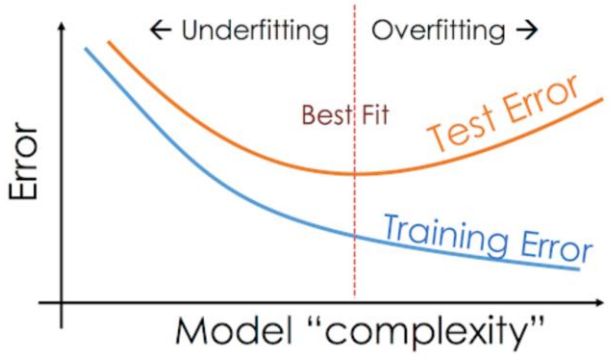


Figure 2: Curves showing Overfittng and Underfitting

eISSN 1303-5150

3. Materials and Methods

4.1 Dataset

The dataset was obtained from the website of the UCI Machine Learning Repository. To determine if an image contains symptoms of diabetic retinopathy or not, the dataset contains features that were taken from the Messidor image set. The dataset's characteristics and labels are then determined [29]. After then, the dataset is divided into two sets: one for training, where most of the data is used, and one for testing. Four alternative class algorithms were put up in the training set to evaluate the model's overall performance. The study employed k-Nearest Neighbor, random forest, support vector machines, and neural networks as prediction techniques. More data is provided without outputs once the framework is finished learning from training datasets. The final model produces the results using the knowledge it learned from the methods that it was trained on. Finally, it is learnt how accurate each set of rules is and may identify which particular algorithm will provide the results for the prediction of diabetic retinopathy that are more accurate.

The Messidor dataset was created specifically to aid academics in exploring their work on the simple prediction of diabetic retinopathy. Using the Messidor image set, numerous features were derived from retinal fundus pictures and included in this dataset. These dataset features that can be utilised to predict DR were extracted. The dataset has a total of 20 columns, the last of which is a class column also known as an output that indicates the presence of diabetic retinopathy (1 for presence) or absence of DR. The first 19 columns of the dataset are independent variables that show various retinal aspects of a patient (0 for absence). There are 1151 instances overall. The dataset has the features listed below. [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.

All of the features in the dataset have had their total count, minimum and maximum values, means, and standard deviation determined. These many types of estimates that are based on the information gathered can be used as a foundational step toward developing information-based estimations. For the purpose of quantifying changeability or scattering around a normal, standard deviation is used. In reality, it is an amount of instability. The difference between the real and the expected value is known as dispersion. The standard deviation increases as the dispersion or variability increases. Data can be analysedin terms of noise, dispersion, variance, etc. by calculating these metrics.

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

4. Data visualization using Histograms:

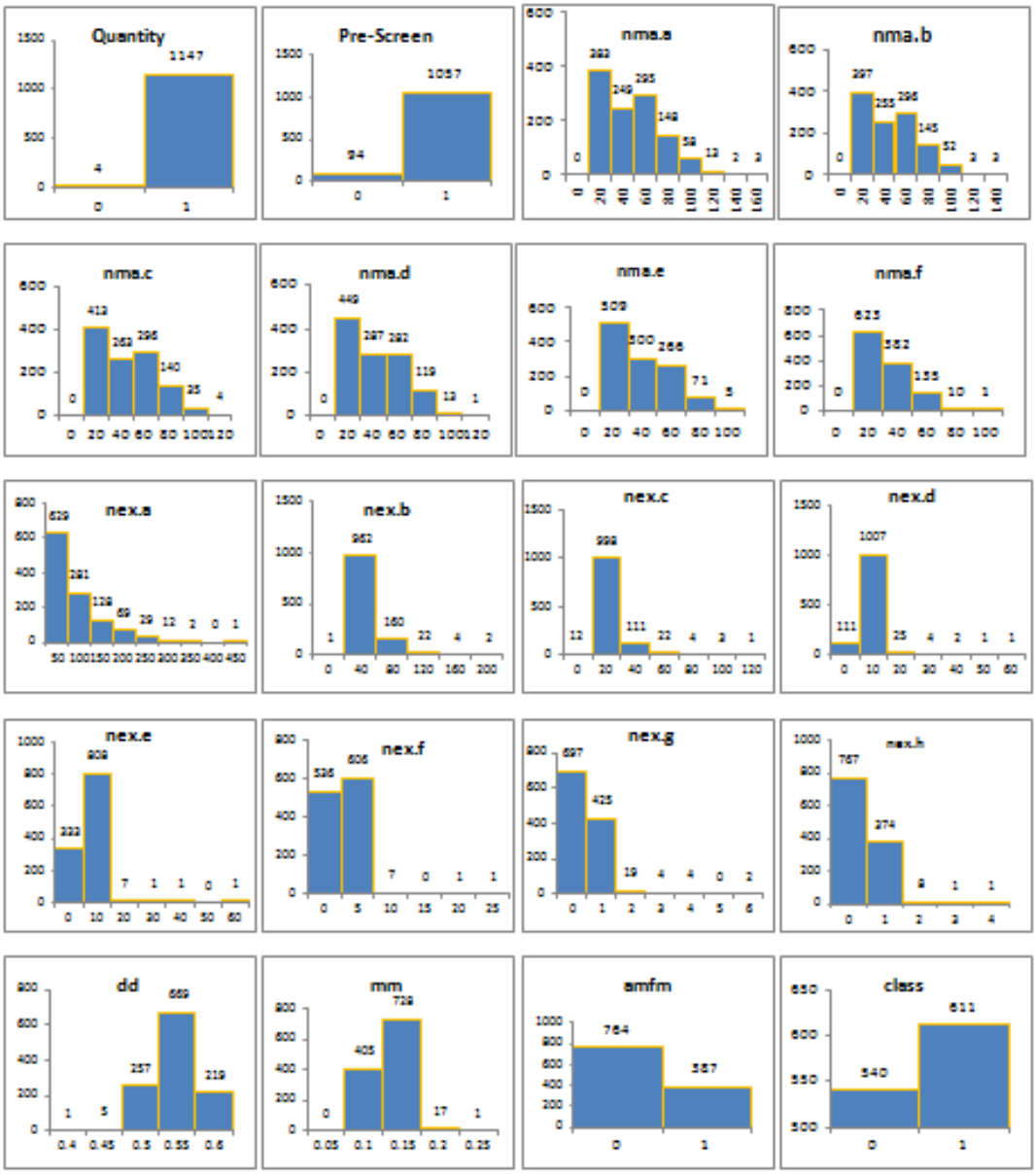


Figure 4 : Histogram representation of each feature

5. Experimental Setup

On this dataset, five different classification algorithms have been appliedfor the prediction of diabetic retinopathy. The dataset is divided into a training set and a testing set for the purposes of applying these methods.Dimensionalityreduction method has been used on the dataset to get the best results possible. The most popular technique for doing this is principal component analysis, or PCA.

Additionally, data may have a small number of outliers, low frequency of a category variable, inaccurate values, or other factors that make it difficult to recognise patterns. The creation of a model will be challenging due to these kinds of inconsistencies. Although data validation cannot directly identify the issue, it aids in the ability to foresee that there may be some issues with the model's consistency. In order to find out the stability of the model, two types of data validations have been appliedon this dataset.

Hold-out validation:

In order to divide the dataset into train and test sets, hold-out validation is used. The model will be trained using a training set, and a test dataset will be used to gauge how well it performs on data that has not yet been seen. Data is divided into two different groups: 15% holdout and 25% holdout. Of these, 15% and 25% will be utilised as test sets, respectively, while the remaining data will be used as a training set [32][33].

Cross-validation

The dataset will be randomly divided into 'k' groups for cross validation. One of these "k" groups will be a test set, while the other "k" groups will be training sets. These training sets will be used to train the model, which will then be further examined on the test set. Each group will be utilised as a test set once more after this process has been repeated. 15-fold cross validation and 25-fold cross validation are applied in thisstudy [32][33].

Betteranalyse of the model outcome is possible and, in turn, stronger prediction models can be built,with the help of these different validations.

6. Algorithm Accuracies and Result Analysis:

7.1 For : Cross validation (15 fold and 25 fold)

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

This research has used a variety of classification techniques to forecast the likelihood of diabetic retinopathy. These methods include bagged trees, logistic regression, support vector machines, and K closest neighbours. These methods are practical and appropriate for these studies, as stated in the section on machine learning algorithms that was covered above. Here two distinct forms of validations on dataset are used in order to construct the most appropriate model for this research and avoid under- or overfitting the data. They are hold-out validation and 'k' fold cross validation. Applying the algorithms with these validations results in the computation of accuracy, and a comparative analysis has been discussed. Additionally, the idea of dimensionality reduction through Principal Component Analysis has also been used, where the features are reduced to 18, 16, and 14. This improves algorithmic accuracy. The accuracy table below shows the results for

eISSN 1303-5150

accuracy for 15 fold and 25 fold cross validation:

TABLE 2 : Accuracies with Cross Validation

Algorithm ⇩	**Validation ⇨**	**15 fold Cross Validation**	**25 fold Cross Validation**
	PCA ⇩	Accuracy	Accuracy
Logistic Regression	PCA 18	75.2	74.9
	PCA 16	75.6	75.1
	PCA 14	77.2	76.5
SVM	PCA 18	74.1	74.8
	PCA 16	75.3	74.6
	PCA 14	74.2	74.9
KNN	PCA 18	66.7	65.4
	PCA 16	65.3	64.4
	PCA 14	64.8	64.5
Bagged tree	PCA 18	75.1	75.0
	PCA 16	75.2	74.1
	PCA 14	74.5	73.9

The table above shows different algorithmic accuracy levels. It can be seen that using 14 primary components and 15 fold and 25 fold cross validation using logistic regression yields the best accuracies of 77.2% and 76.5%, respectively. This is due to the fact that logistic regression uses the logistic function to estimate the exertion relationship between the dependent variable (what we must expect) and one or more independent variables (our features), allowing us to determine the likelihood of occasion success or occasion disappointment (also known as success or failure) [34]. This dataset can interpret model coefficients as a pointer of feature significance because it is linearly separable. When combined with PCA 16, SVM and bagged trees produced reasonable accuracies of 75.3 and 75.2respy.

eISSN 1303-5150

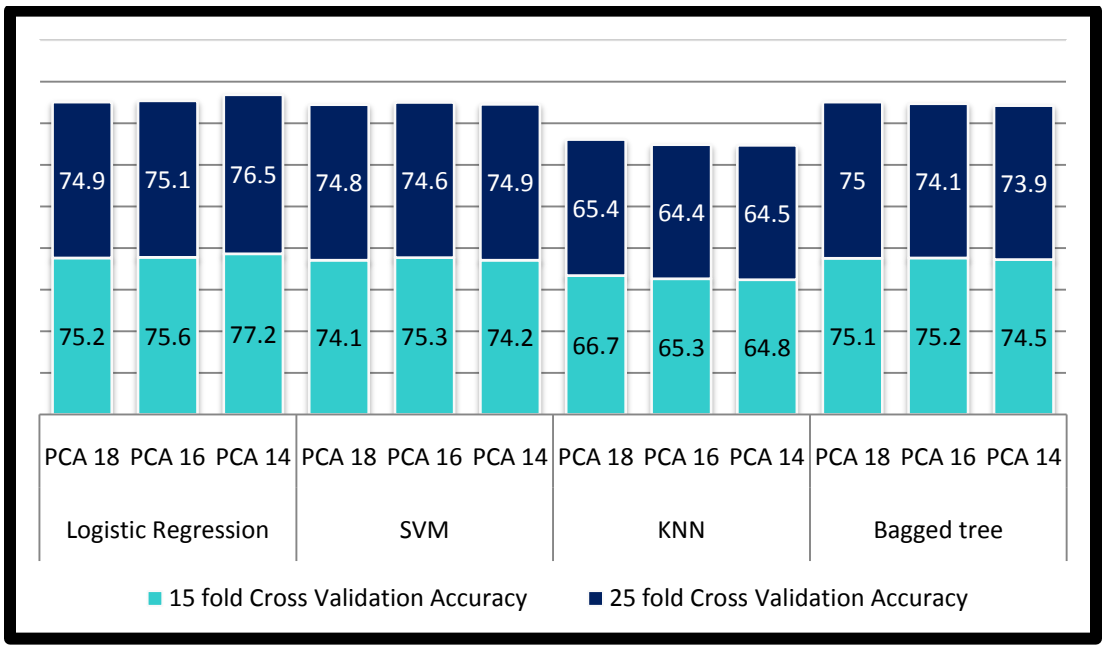


Figure 5: Graphical Representation: Accuracies with Cross Validation

7.2 For : Hold out validation (15% and 25% hold out)

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

The study employed hold out validation on the dataset since it has examined the accuracy results from applying classification algorithms using cross validation on the dataset. The dataset was split up into a train set and a test set by this technique. Utilizing 80% of the data as a training set and 20% as a test set is the most typical split. It is also common practise to use 70% of the data as a training set and 30% as a test set. This research employed both divides in the experiment in order to analyse the results more effectively. The same PCA structure that was used earlier for cross validation is also used here. The accuracy table shown below shows the results for 15% and 25% hold out validation.

TABLE 3 : Accuracies with Hold-Out Validation

Algorithm ⇩	**Validation ⇨**	**15% hold out**	**25% hold out**
	PCA ⇩	Accuracy	Accuracy
Logistic Regression	PCA 18	82.4	77.1
	PCA 16	82.3	76.9
	PCA 14	83.8	79.1
SVM	PCA 18	80.3	72.5
	PCA 16	81.3	73.0

	PCA 14	83.2	73.4
KNN	PCA 18	64.6	59.9
	PCA 16	70.4	68.3
	PCA 14	71.5	68.5
Bagged tree	PCA 18	79.6	78.1
	PCA 16	81.3	77.2
	PCA 14	79.9	77.7

The table above displays the algorithmic accuracy for various categorization algorithms when paired with hold-out validation. The finding indicates that logistic regression once more outperformed all other methods with 83.8% accuracy when using the hold-out strategy. This algorithm was successful with the present issue because it is discrete in nature. Additionally, when used with 14 primary components, the performance of the support vector machine is good, with the greatest accuracy reaching 83.2%. Logistic regression provides a natural probabilistic perspective of class predictions when used on our DR dataset. Additionally, 81.3% accurate bagged trees have been found, which helps to reduce variance and eliminate the overfitting provocateur.

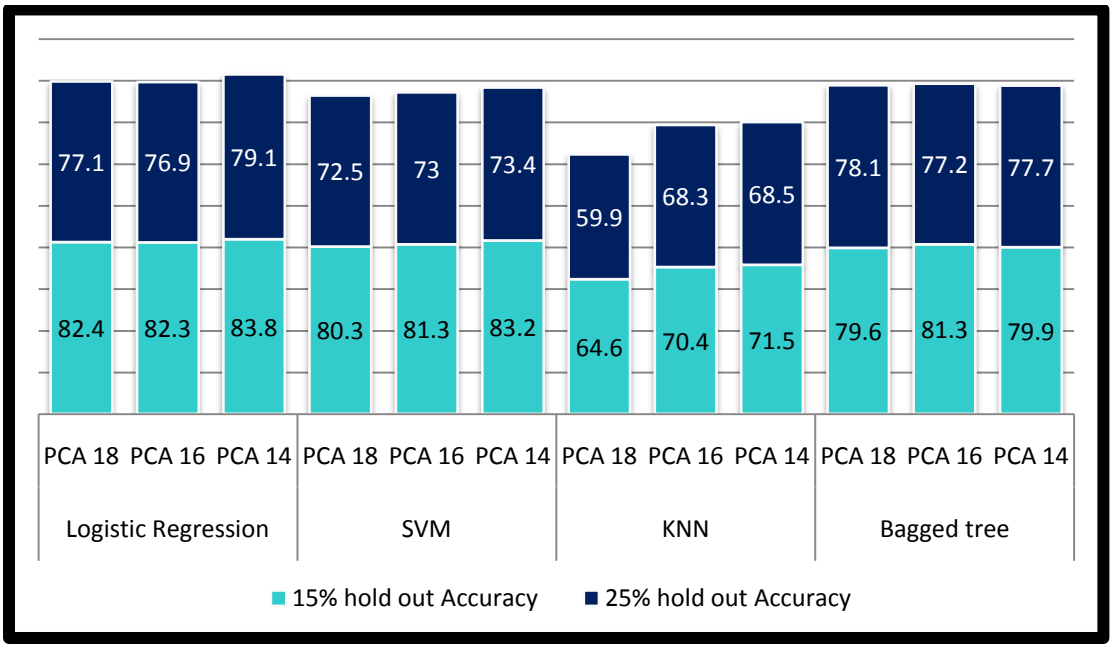


Figure 6: Graphical Representation: Accuracies with Cross Validation

7. Conclusion

In this study, Messidor dataset is used and machine learning classification methods are applied to predict diabetic retinopathy. This dataset includes features for the prediction of diabetic retinopathy symptoms that were taken from the Messidor picture set. To obtain the experimental results, four different categorization algorithms were applied to the dataset. Cross validation and

eISSN 1303-5150

hold-out validation, two types of data validation, have been used since the data could contain noisy values. This will aid in lowering under- and over-fitting of the data as well as in identifying a stable model. The dataset for this study is made up of the records of 1151 patients. The study utilises15 fold and 25 fold cross validation for hold-out validation, and 15% and 25% hold-out methods for cross validation. Principal Component Analysis has also been used to reduce dimensionality in order to achieve the greatest results. In this study, logistic regression had the highest accuracy, with cross validation results of 77.2% and hold-out validation results of 83.8%. The approach for using and testing various classification algorithms to generate ensemble models that would outperform various learners was provided in this paper. This study also investigates challenges with feature selection, extraction, data representation, and ensemble selection while observing the results over a predetermined time frame.

8. 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] HayrettinEvirgen, MenduhÇerkezi, "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

eISSN 1303-5150

[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, JindeShubham,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.