



Diabetic retinopathy detection through artificial intelligent techniques: a review and open issues

Uzair Ishtiaq^{1,2} · Sameem Abdul Kareem¹ · Erma Rahayu Mohd Faizal Abdullah¹ · Ghulam Mujtaba³ · Rashid Jahangir^{1,2} · Hafiz Yasir Ghafoor^{1,2}

Received: 14 September 2018 / Revised: 22 November 2018 / Accepted: 7 December 2018 /

Published online: 4 January 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Diabetic Retinopathy (DR) is the disease caused by uncontrolled diabetes that may lead to blindness among the patients. Due to the advancements in artificial intelligence, early detection of DR through an automated system is more beneficial over the manual detection. At present, there are several published studies on automated DR detection systems through machine learning or deep learning approaches. This study presents a review on DR detection techniques from five different aspects namely, datasets, image preprocessing techniques, machine learning-based approaches, deep learning-based approaches, and performance measures. Moreover, it also presents the authors' observation and significance of the review findings. Furthermore, we also discuss nine new research challenges in DR detection. After a rigorous selection process, 74 primary publications were selected from eight academic databases for this review. From the selected studies, it was observed that many public datasets are available in the field of DR detection. In image preprocessing techniques, contrast enhancement combined with green channel extraction contributed the most in classification accuracy. In features, shape-based, texture-based and statistical features were reported as the most discriminative in DR detection. The Artificial Neural Network was proven eminent classifier compared to other machine learning classifiers. In deep learning, Convolutional Neural Network outperformed compared to other deep learning networks. Finally, to measure the classification performance, accuracy, sensitivity, and specificity metrics were mostly employed. This review presents a comprehensive summary of DR detection techniques and will be proven useful for the community of scientists working in the field of automated DR detection techniques.

Keywords Diabetic retinopathy · Convolutional neural network · DIARETDB1 · Image preprocessing · Artificial neural network · Transfer learning

✉ Sameem Abdul Kareem
sameem@um.edu.my

1 Introduction

The retina is a spherical structure present at the inner side at the back of the eye. Its function is to process visual information through photoreceptors present in the eye known as rods and cones. At the center of the retina, there is a dark, round area known as macula. The central point of the macula called the fovea, is responsible for providing sharp vision. Like any other tissue in the body, the retinal tissue receives blood supply through the body's vasculature. The optic disc, consisting of the optic cup, is a structure where the optic nerve leaves the eye and is also the site for all the major arteries and veins to enter and exit the eye [42].

Diabetes is a very common disease all over the world. This disease often leads to blindness in patients between 20 to 74 years of age due to a condition caused by uncontrolled diabetes known as Diabetic Retinopathy (DR) [4]. Blood is carried to the layers of the retinal tissue by micro blood vessels that require a maintained blood sugar level and an unobstructed blood flow. If large amount of sugars, such as glucose or fructose, accumulate in the blood, these blood vessels start to breakdown due to improper distribution of oxygen to the cells. Structural abnormalities occur as a result of decreased metabolic rate which leads to DR. The major signs of DR include; microaneurysms, exudates, hemorrhages and swelling of the blood vessels in the eye. DR has two stages; non-proliferative DR (NPDR) and proliferative DR (PDR). The NPDR is further subdivided into three types depending on the severity; mild NPDR, moderate NPDR and severe NPDR [31]. This stage of DR occurs when the blood vessels in the eye rupture and fluid start to leak into the retina. As a result, the retina becomes swollen and wet. Signs such as microaneurysms, exudates and hemorrhages can be detected at this stage. Microaneurysms are the marks of early abnormality where the retinal vasculature widens, exudates are the lipid accumulations within the retina and hemorrhages are blood spots on the retina which happen due to the leakage of blood vessels. The PDR occurs when new abnormal blood vessels arise in various parts of the retina. This is a complicated stage of DR and may lead to blindness [23].

Detection of DR at an early stage is very important in order to save the patient's vision. It has been shown by many researchers that 90% of diabetic patients can be saved from DR through an early diagnosis [91]. Diagnosis of DR can be done either manually by an ophthalmologist or through an automated system. There are pros and cons to both of these methods of DR detection. The only benefit of manual detection is that, it does not require any computer assistance for the DR detection process, while it requires that the ophthalmologist must be an expert in the field. Sometimes the signs of DR at its initial phase are so small that even an expert ophthalmologist cannot recognize them properly. Due to the advancements in Artificial Intelligence (AI), early detection of the disease through an automated system is more likely and has more benefits over the manual DR detection system. The benefits may include; decreased amount of workload on the ophthalmologist and less chance of human error. Moreover, an automated system may be able to detect lesions and abnormalities with far greater ease and efficiency than that which is possible manually. Hence, automation of DR detection is essential. The DR automated systems can be developed either through machine learning approaches or through deep learning approaches.

In machine learning (ML) approaches, first the retinal images with signs of DR are collected, then these images are preprocessed using different image preprocessing techniques such as green channel extraction, contrast enhancement, cropping and resizing, etc.

Afterwards, discriminating features are extracted from the preprocessed images to construct a master feature vector (MFV). The constructed MFV is then fed as an input to machine learning algorithms to construct the DR classification model. Finally, the performance of the constructed model is evaluated using new test images. The limitations of ML approaches are that these approaches require expert-intervention to extract the discriminative region of interests (ROIs) as the features from the images. Thus, it requires enormous amount of time and human efforts to extract discriminative features for classification. Deep learning approaches are proposed in order to overcome the limitation associated with feature extraction in ML approaches.

Deep learning (DL) approaches simplify the process of feature extraction by automatically segmenting the ROIs from the training images. Moreover, these approaches provide end to end solution for the construction and evaluation of the classification model. In DL approaches, first, the DR images are collected. Afterwards, preprocessing techniques such as, contrast enhancement, illumination correction, and resizing are applied on the collected DR images to reduce the noisy features. These preprocessed images are then forwarded to the DL architecture to automatically extract the distinguishing features from the training images and to learn the classification rules from the extracted features and their weights. In DL learning, the features weights are optimized recursively to obtain the best features weight to classify the images more accurately. Finally, these optimized weights are tested on unlabeled images through a classification layer. A deep architecture requires a large amount of training data, therefore, it may not perform well with limited number of images. In addition, the DL approaches also require higher computational power and memory as compared to ML approaches. In general, DL approaches require a dedicated Graphical Processing Unit (GPU) to construct and evaluate the classification model.

As discussed above, both of these approaches have their own pros and cons. Hence, both approaches have been employed to develop an automated DR detection systems. In past years, several researchers have employed ML and DL approaches for the automation of DR. However, there are only a few review articles published in academic databases on DR detection. Review articles are important as they comprehensively summarize the existing work in a particular field. For instance, Sopharak, Uyyanonvara [82] and Joshi and Karule [40] presented a review article on various recent ML-based studies conducted specifically for microaneurysms and exudates respectively in DR detection. However, there are several studies that employed deep learning approaches for the detection of exudates using DR images which were not considered in their review studies. In addition, the authors did not consider the detection of any other lesions from DR images. Moreover, authors only reviewed ML algorithms but, some important aspects of ML, including datasets, preprocessing techniques and feature engineering approaches used for detection of DR were not covered in Joshi and Karule [40]. While Sopharak, Uyyanonvara [82] discussed only three publicly available datasets, three image preprocessing techniques, the algorithms used for microaneurysms segmentation and reported the comparative analysis of the earlier techniques. Amin, Sharif [7] and Jordan, Menolotto [39] also conducted the reviews on ML based approaches for diabetic retinopathy detection. The authors discussed the publicly available datasets in the field of DR detection. Moreover, the authors explored the available ML-based detection techniques for various DR lesions and presented the review of evaluation measures. However, no information regarding the preprocessing techniques used was discussed nor were DL-based approaches included in both of the reviews. Almotiri, Elleithy [6] conducted a survey on retinal vessel segmentation techniques for the detection of DR

and other eye related diseases. The authors provided an overview of fundus photography, preprocessing operations, rule-based and ML-based retinal vessel segmentation methods and performance evaluation of different methods. However, the authors did not consider the DL-based approaches in their review. Among all, only Rahimy [68] presented a review on DL-based DR detection techniques and included only a few studies. They also discussed other eye diseases, namely, age-related macular degeneration and glaucoma which are not included in our scope. The authors did not exclusively include some important aspects including, datasets, preprocessing techniques, information regarding DL architectures and the evaluation metrics. Moreover, the authors did not include the ML-based approaches for DR detection. Finally, all aforementioned reviews suffer from study selection biasness because the authors have not defined the inclusion and exclusion criteria for selecting the primary studies. Furthermore, the authors have not mentioned the range of publication years that cover the primary studies in their reviews. The existing reviews were too specific, either in terms of lesions or in terms of approaches (ML and DL). Therefore, there is a need for review articles that recapitulate the current approaches for the diagnosis of DR for current and future researcher communities in this discipline. Hence, to overcome the limitations of the aforementioned reviews, this paper provides a comprehensive review of both ML and DL approaches employed to automate DR detection published from 2013 to 2018 to cover the latest DR detection models constructed through ML or DL-based approaches. To cover the full range of existing primary studies, the descriptors like “Eye Disease”, “Retinopathy”, “Diabetic Retinopathy”, “Retinal Disorder”, “Retinal Disease”, “Retinal Abnormality”, “DR” and “Diabetic Eye Disease” were systematically applied on eight academic databases including, Web of Science, Scopus, IEEE Xplore, PubMed, MedLine, Science Direct, ACM and Springer Link. Due to the diversity and intricacy of applications of DR detection, a methodological decision analysis framework for the selection of articles was used. The target of this framework includes, (1) datasets, (2) preprocessing techniques, (3) DR detection using ML approaches, (4) DR detection using DL approaches and (5) performance measures. After the rigorous search criteria and inclusion and exclusion criteria, a total of 74 recent relevant primary studies were found from the eight databases. It is envisaged that this paper will help the researchers’ communities of DR detection discipline. The major research objectives of this study include:

1. Publicly available datasets in the field of DR detection and their availability.
2. Pre-processing techniques that can be applied on retinal images.
3. Widely used ML based techniques for DR detection, including features and classification algorithms.
4. Widely used DL based techniques for DR detection.
5. Performance measures that are employed to evaluate DR detection algorithms.
6. Future research directions and open issues that need to be addressed by the future researchers working in the field of DR detection.

The organization of this paper is as follows: Section 2 presents the research methods used for the selection of the primary studies. Section 3 explores and discusses the categorical review on DR detection and evaluates the results. Section 4 deliberates the discussion on the review. Section 5 presents some future research directions and open issues that need to be addressed for the future researchers. Section 6 concludes the paper.

2 Research methodology

The research methodology that was adapted for this review is shown in Fig. 1. As shown here, initially, the review targets were identified. Afterwards various search keywords were formulated to elicit the most relevant primary studies from the eight distinct academic databases. Next, nine different filters were applied to select the most relevant primary studies. Subsequently, we critically analyzed the selected articles on the following five aspects, namely, (1) dataset analysis, (2) preprocessing techniques analysis, (3) machine learning-based diabetic retinopathy detection, (4) deep learning-based diabetic retinopathy detection, (5) performance measures analysis. The filters that were applied to select the relevant publications that meet all the objectives and criteria of the review were identified. Finally, the current research challenges and future trends in the field of diabetic retinopathy detection using machine learning or deep learning are discussed.

2.1 Search and selection of articles

To search the relevant primary studies on the DR detection field, three different groups of keywords were formulated as shown in Table 1. The first group contains the keywords belonging to eye disease and DR. The second group contains keywords related to image classification. Finally, the third group contains the keywords that belong to either machine learning approaches or deep learning approaches. The ‘OR’ Boolean operator was used in

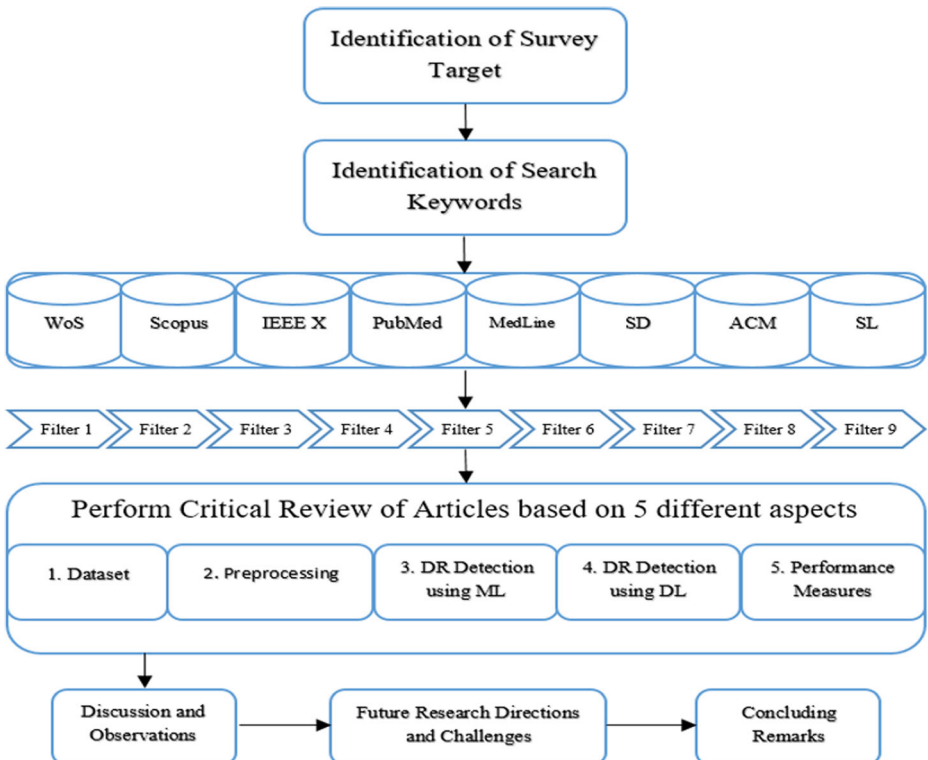


Fig. 1 Flowchart of research methodology

between the keywords that belong to the same group. In addition, ‘AND’ Boolean operator was used in between different groups to form the search query. This formulated search query was applied on eight different databases, namely, Web of Science, Scopus, IEEE Explore, PubMed, MedLine, Science Direct, ACM and Springer Link to search the relevant primary studies. In addition, nine different filters (namely, keywords-based filter, year-wise filter, article type filter, language filter, duplicates removal filter, article, abstract and keywords filter, full-text based filter, reference screening filter, and quality assessment filter) were applied to search the most relevant primary studies on DR detection. These filters (as shown in Fig. 2) are discussed in subsequent paragraphs. In addition, the search results of these filters are also shown in Table 2.

- *Filter 1:* Here, the search query was applied on article title, article abstract, and article keywords to find the associated primary studies. When Filter 1 was applied on the eight databases, we obtained 1777 articles in total, including 328 from Web of Science, 568 from Scopus, 17 from IEEE Explore, 152 from PubMed, 259 from Med Line, 15 from Science Direct, 64 from ACM and 374 from Springer Link.
- *Filter 2:* Here, the year filter was applied to select the primary studies published from 2013 to 2018. This filter retrieved in total 1115 articles out of 1777 articles.
- *Filter 3:* This filter retrieves only journal and conference articles from the pool of Filter 2 articles. When this filter was applied the number of articles were reduced to 979.
- *Filter 4:* Here, the language filter was applied to extract the articles that were only published in English language. This filter returned 969 articles.
- *Filter 5:* This filter removes the duplicate articles. As the articles may appear in more than one databases, therefore the duplicates were removed. This filter resulted in 684 articles.
- *Filter 6:* In this filter, the titles, abstracts and keywords of extracted 684 articles were rigorously read to check whether they fit the inclusion criteria (mentioned in Table 3), and dropped otherwise. This filter yielded overall 126 articles.
- *Filter 7:* The full text of all 126 extracted articles (filtered by filter 6) were downloaded and read critically to evaluate the suitability of those articles with our inclusion criteria and review objectives. This filter dropped 61 more articles. Thus, only 65 articles were retained.
- *Filter 8:* Here, the references of all 65 selected articles were scanned to see any relevant primary studies that suits our objectives and criteria. This filter found 9 more new articles. Hence, the number of selected articles increased from 65 to 74.

Table 1 Groups of formulated keywords to search relevant studies from academic databases

Group 1	Group 2	Group 3
“Eye Disease” OR “Retinopathy” OR “Diabetic Retinopathy” OR “Retinal Disorder” OR “Retinal Disease” OR “Retinal Abnormality” OR “DR” OR “Diabetic Eye Disease”	“Image Classification” OR “Image Segmentation” OR “Classification” OR “Segmentation” OR “Image Analysis” OR “Image Processing” OR “Medical Images” OR “Clinical Images” OR “Medical Image Classification” OR “Clinical Image Classification” OR “Fundus Images”	“Deep Learning” OR “*Neural Network” OR “CNN” OR “Convolutional Neural Network” OR “Machine Learning” OR “Transfer Learning” OR “AlexNet” OR “GoogleNet” OR “Supervised Machine Learning” OR “Unsupervised Machine Learning” OR “Semi-Supervised Machine Learning” OR “Active learning”

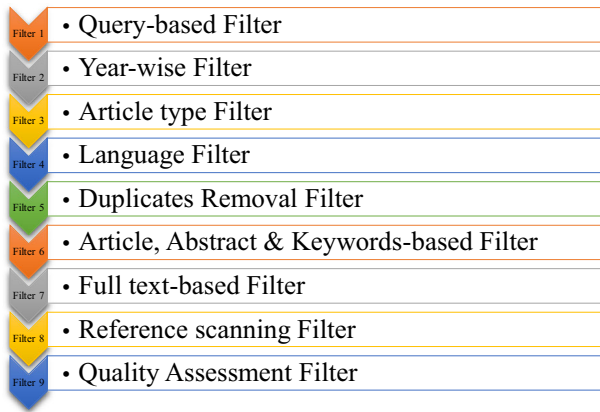


Fig. 2 Filters applied for the selection of articles

- *Filter 9:* The quality of these selected articles was analyzed in this Filter 9 and all of them were found to be suitable for this review. Finally, a total of 74 articles were selected after going through a rigorous articles selection process from eight renowned academic databases.

Figure 3a shows the distribution of the selected 74 articles with respect to academic databases. Among the 74 articles, 32 were selected from the Web of Science, 22 from Scopus, 4 from IEEE Explore, 6 from PubMed, 1 from Med Line, 1 from Science Direct, 4 were selected from the ACM and 4 from Springer link. Figure 3b shows the article-type wise distribution of selected articles. Among the 74 articles, 51 articles are journal articles and 23 articles are conference proceeding articles. Figure 3c shows the year-wise publication counts of the selected articles. The trend of publication can be seen through Fig. 3c which is fluctuating. The maximum number of articles were published in the year 2017 i.e. 37, followed by 13 in 2016, 8 in 2018 as well as in 2015, 7 and 1 in 2014 and 2013 respectively.

In all the 74 articles, the authors have either used Machine Learning techniques or Deep Learning techniques for the detection, classification or segmentation of Diabetic Retinopathy. Therefore, we classified the scope of the survey into nine major objectives (as shown in Table 4). There are a few studies that used the fusion of Machine Learning and Deep Learning techniques, these are categorized in a separate objective.

Table 2 Search and Filter Results from 8 Databases

Database	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7	Filter 8	Filter 9
WoS	328	242	236	236	684	126	65	74	74
Scopus	568	410	359	354					
IEEE Explore	17	17	17	17					
PubMed	152	110	110	108					
Medline	259	90	74	74					
Science Direct	15	15	15	15					
ACM	64	50	50	50					
Springer Link	374	181	118	115					
Total	1777	1115	979	969					

Table 3 List of inclusion criteria

S. No.	Inclusion Criteria
1	Deep Learning or Machine Learning is used.
2	Diabetic Retinopathy dataset is used.
3	Article is either journal article or conference article.
4	Article must be published in 'English' language.
5	Article must be published in between 2013 to 2018.

3 Review of diabetic retinopathy detection

This section reviews diabetic retinopathy detection techniques in the selected 74 studies from five different perspectives, namely, (1) datasets, (2) preprocessing techniques, (3) diabetic retinopathy detection through Machine Learning, (4) diabetic retinopathy detection through Deep Learning and (5) performance measures. Subsection 3.1 presents the review of different datasets used for diabetic retinopathy detection. Subsection 3.2 reviews various preprocessing techniques used for the detection of diabetic retinopathy. Subsection 3.3 presents a review on various machine learning techniques used to detect diabetic retinopathy (through novel image features, proposing new ML classifier, using unsupervised ML, proposing new segmentation method, applying rule-based algorithm and ML, applying existing feature extraction methods and ML classifiers). Subsection 3.4 indicates a review on diabetic retinopathy detection through deep learning which includes the usage of transfer learning or proposing a novel deep learning architecture. Finally, Subsection 3.5 gives a review on different performance metrics used for the evaluation of the performance of techniques for detecting diabetic retinopathy.

3.1 Diabetic retinopathy datasets

In the selected articles, most authors utilized publicly available datasets and others have used exclusive datasets which have been divided into training data and testing data. For instance, Yang, Li [100] used 22,795 images from a public dataset, namely Kaggle, for the detection of eye lesions and grading of DR which was divided into 21,995 images for training and 800 images for testing. Similarly, Tan, Fujita [87] used a total of 298 images from an exclusive dataset, CLOEPATRA, for the segmentation of eye lesions, in which 149 images were used for training and 149 for testing.

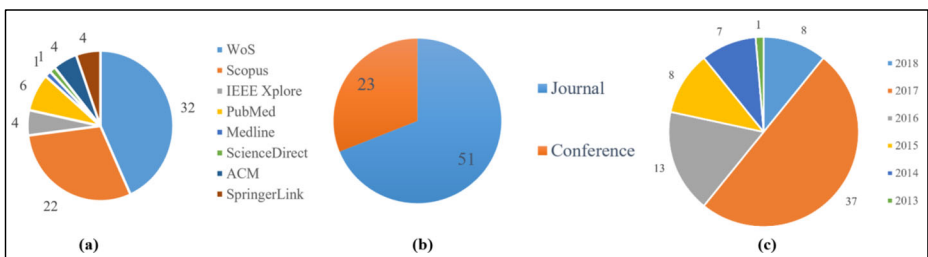


Fig. 3 Bibliometric characteristics of selected studies. ** Fig. 3a shows the distribution of selected studies as per academic databases. Figure 3b shows the distribution of selected studies as per article type. Figure 3c shows the distribution of selected studies from 2013 to 2018

Table 4 Typography of selected 74 studies based upon common objectives

S. No.	Objectives	References	Study Count
1	To employ transfer learning for Diabetic Retinopathy Detection	[3, 18, 30, 33, 46, 47, 50, 52, 60, 66, 71, 75, 86, 89]	14
2	To proposed new Deep Learning network for Diabetic Retinopathy Detection	[13, 19, 21, 22, 27, 29, 38, 62, 64, 65, 73, 87, 88, 93, 99, 100, 104]	17
3	To develop a Machine Learning model through novel image features	[5, 11, 14, 34, 41, 56, 59, 61, 69, 74, 79, 83, 90, 94, 96, 97]	16
4	To propose new Machine Learning classifier	[36]	1
5	To use unsupervised Machine Learning	[12, 44, 105, 106]	4
6	To propose new segmentation method	[25, 72]	2
7	To apply rule-based algorithm and Machine Learning to detect Diabetic Retinopathy	[98]	1
8	To apply existing feature extraction methods and machine learning classifiers to detect Diabetic Retinopathy	[8, 10, 15, 24, 26, 28, 35, 48, 49, 53–55, 63, 67, 81, 92]	16
9	To use a combination of Machine Learning and Deep Learning for Diabetic Retinopathy Detection	[1, 9, 57]	3

Table 5 shows a list of diabetic retinopathy datasets that were used in the selected academic articles. The table contains the dataset names, a very short description of each one of them, the total number of studies where these datasets were used by the authors of the selected articles, and the last column shows the references of the articles where these datasets were used. Table 6 contains the links of the publically available datasets in the selected academic articles. Based on our research, there are some datasets, which are publicly available and a few have been exclusively used by various researchers. These exclusive datasets were acquired at different hospitals with different fundus camera settings (available details in Table 6). Publicly available datasets can be considered as standard datasets as a number of studies and experiments are conducted on them, including, Retinopathy Online Challenge (ROC), E-Ophtha, Kaggle, DRIVE, STARE, DIARETDB, Messidor, etc.

Zhou, Wu [105] used the ROC dataset for unsupervised classification on DR images. ROC contains 100 fundus images, divided into two subsets, i.e. 50 for training and 50 for testing. The training images are provided with the ground-truths, whereas the testing images are not provided with the ground-truths. As the authors were unable to verify their results due to the unavailability of ground-truths, therefore, they only used the training images for training and testing (37 and 13 respectively).

E-Ophtha dataset consists of two sub-datasets, namely, E-Ophtha EX (82 images) and E-Ophtha MA (381 images). The former was mainly used for exudates detection and the latter was used for the detection of microaneurysms. Yu, Xiao [104] used E-Ophtha EX dataset to detect exudates using Convolutional Neural Network (CNN) while Orlando, Prokofyeva [57] used E-Ophtha MA dataset along with two other datasets (DIARETDB1 and Messidor) for red lesions detection.

To the best of our knowledge, Kaggle is the largest publicly available eye fundus dataset for the detection of diabetic retinopathy. Kaggle is provided by EyePACS, a platform for data exchange on eye-related ailments. EyePACS conducted competitions in 2015 with the aim of creating automated systems for the classification of five severity levels of DR (0 = No DR, 1 = Mild, 2 = Moderate, 3 = Severe and 4 = Proliferative DR). This dataset is imbalanced as there were more images with no DR than that of the other classes. Xu, Feng [99] used a subset from

Table 5 Datasets analysis

S. No.	Dataset	Description	No. of Studies	Reference
1	ROC (Retinopathy Online Challenge)	This dataset contains 100 images (50 images for training and the same for testing). The cameras used were Canon CR5-45NM, TopconNW 100 and NW 200.	7	[13, 19, 94, 96, 97, 105, 106]
2	E-Ophtha	This dataset was produced by OPHDIAT Tele-medical network. It is divided into two datasets, namely, E-Ophtha MA and E-Ophtha EX. Both consisting of 381 and 82 images respectively.	9	[19, 25, 44, 46, 57, 60, 66, 69, 104]
3	Kaggle	This dataset is provided by EyePACS for Diabetic Retinopathy detection challenge. It consists of 88,702 images in total (35,126 for training and 53,576 for testing).	14	[22, 27, 29, 30, 33, 38, 50, 62, 66, 71, 79, 89, 99, 100]
4	DRIVE	This dataset consists of 40 images (divided into training and testing dataset of 20 images each), captured in a diabetic retinopathy program in Holland. The camera used was Canon CR5 non-mydratric 3CCD with 45 degree FOV (Field of View).	8	[11, 21, 24, 52, 88, 92, 93, 97]
5	STARE	This dataset contains total 400 images. The camera used was Topcon TRV-50 fundus camera having 35 degrees FOV.	8	[11, 18, 52, 54, 72, 92, 93, 97]
6	DIARETDB0	This dataset contains total of 130 images (normal images = 20, images with signs of Diabetic Retinopathy = 110). The images were taken with fundus camera having 50 degrees FOV.	4	[49, 55, 56, 67]
7	DIARETDB1	This dataset contains total of 89 images (normal images = 5, images with at least mild signs of Diabetic Retinopathy = 84). The images were taken with fundus camera having 50 degrees FOV.	25	[1, 5, 12, 14, 19, 25, 30, 41, 44, 46, 49, 53, 54, 57, 59, 66, 67, 73, 74, 81, 83, 90, 94, 97, 98]
8	Messidor	This dataset contains 1200 images in total. These were taken by 3 ophthalmological departments in France. 400 images were taken without pupil dilation while 800 were taken with pupil dilation.	13	[1, 8, 15, 25, 28, 34, 46–48, 55, 57, 83, 89]
9	Messidor-2	This dataset contains 1748 images in total, taken with Topcon TRC NW6 non-mydratric fundus camera having 45 degrees FOV.	1	[3]
10	DRiDB	This dataset contains 50 images and it is available on request.	3	[63–65]
11	DR1	This dataset was produced by Department of Ophthalmology, Federal University of Sao Paulo	2	[47, 54]

Table 5 (continued)

S. No.	Dataset	Description	No. of Studies	Reference
12	DR2	(UNIFESP). It contains 234 images captured with TRX-50X, mydriatic camera having 45 degrees FOV. This dataset was also produced by Department of Ophthalmology, Federal University of Sao Paulo (UNIFESP). It contains 520 images captured with TRC-NW8, non-mydriatic camera having 45 degrees FOV.	1	[54]
13	ARIA	This dataset consists of 143 images. The camera used was Zeiss FF450+ fundus camera having 50 degrees FOV.	1	[9]
14	FAZ (Foveal Avascular Zone)	This dataset contains 60 images in total (25 normal images and 35 images of patients with Diabetic Retinopathy).	1	[1]
15	CHASE-DB1	This dataset contains total 28 images of 14 children (one image per eye). CHASE-DB1 is a part of Child Heart and Health Study in England (CHASE).	1	[52]
16	Tianjin Medical University Metabolic Diseases Hospital	This is a customized dataset containing 414 fundus images.	1	[97]
17	Moorfields Eye Hospital	This dataset contains 21,536 images of different ethnicities, collected at Moorfields Eye Hospital, London, including Kenya, Botswana, Mongolia, China, Saudi Arabia, Italy, Lithuania and Norway.	1	[94]
18	CLEOPATRA	There are 298 fundus images constituting CLEOPATRA dataset. The CLEOPATRA study was conducted in 15 hospitals in the United Kingdom for Diabetic Retinopathy detection.	1	[87]
19	Jichi Medical University	This dataset contains 9939 posterior pole fundus images of patients with diabetes. The camera used was AFC-230, NIDEK Co., Ltd., Aichi, Japan, with 45 degrees FOV.	1	[86]
20	Singapore National Diabetic Retinopathy Screening Program	This dataset was collected during Singapore National Diabetic Retinopathy Screening Program (SIDRP) between 2010 and 2013 in Singapore, which contains 197,085 retinal images.	1	[75]
21	Lotus Eye Care Hospital Coimbatore, India	This dataset contains 122 fundus images (normal images = 28, images with signs of diabetic retinopathy = 94). The images were taken with Cannon	3	[10, 35, 61]

Table 5 (continued)

S. No.	Dataset	Description	No. of Studies	Reference
22	Department of Ophthalmology, Kasturba Medical College, Manipal, India	non-mydratric Zeiss fundus camera having 90 degrees FOV. This dataset contains 340 images (normal images = 170, images with signs of diabetic retinopathy = 170). The images were taken with a non-mydratric retinal camera, namely, TOPCON.	1	[26]
23	HUPM, Cádiz, Spain	The fundus images in this dataset were taken from Hospital Puerta del Mar, Spain. It contains 250 images (normal images = 50, images with signs of diabetic retinopathy = 200)	1	[1]

Table 6 Availability links of datasets

S. No.	Dataset	Availability Link
1	ROC (Retinopathy Online Challenge)	http://www.webeye.ophth.uiowa.edu/ROC/
2	E-Ophtha	http://www.adcis.net/en/Download-Third-Party/E-Ophtha.html
3	Kaggle DR Dataset	https://www.kaggle.com/c/diabetic-retinopathy-detection/data
4	DRIVE	https://www.isi.uu.nl/Research/Databases/DRIVE/
5	STARE	http://www.cecac.clemson.edu/~ahoover/stare/
6	DIARETDB0	http://www.it.lut.fi/project/imageret/diaretdb0/
7	DIARETDB1	http://www.it.lut.fi/project/imageret/diaretdb1/index.html
8	Messidor	http://www.adcis.net/en/Download-Third-Party/Messidor.html
9	Messidor-2	http://www.latim.univ-brest.fr/indexfce0.html
10	DRiDB	https://www.ipg.fer.hr/ipg/resources/image_database
11	DR1	http://www.recod.ic.unicamp.br/site/asdr
12	DR2	http://www.recod.ic.unicamp.br/site/asdr
13	ARIA	https://www.eyecharity.weebly.com/aria_online.html
14	FAZ (Foveal Avascular Zone)	http://www.biosigdata.com/?download=colour-fundus-images-of-healthy-persons-patients-with-diabetic-retinopathy
15	CHASE-DB1	https://www.blogs.kingston.ac.uk/retinal/chasedb1/
16	Tianjin Medical University Metabolic Diseases Hospital	Authors have not made the data available publicly
17	Moorfields Eye Hospital	Authors have not made the data available publicly
18	CLEOPATRA	Authors have not made the data available publicly
19	Jichi Medical University	Authors have not made the data available publicly
20	Singapore National Diabetic Retinopathy Screening Program	Authors have not made the data available publicly
21	Lotus Eye Care Hospital Coimbatore	Authors have not made the data available publicly
22	Kasturba Medical College, Manipal, India	Authors have not made the data available publicly
23	HUPM, Cádiz, Spain	Authors have not made the data available publicly

Kaggle dataset for early detection of DR using Deep CNN. They utilized 1000 fundus images (out of around more than 88,000 fundus images), 800 for training their network and 200 for testing.

DIARETDB1, the most frequently used dataset, is also imbalanced in the sense that it comprises of a total 89 fundus images with 84 of which are with signs of mild DR while 5 are normal. In the selected publications, DIARETDB1 is used 25 times, either individually or in combination with other datasets. As an example, Somasundaram and Alli [81] designed a machine learning based ensemble classifier for early prediction of DR using DIARETDB1 with 60% of the data for training and 40% for testing. Moreover, Xiao, Yu [98] used 55 images from DIARETDB1 along with 35 images from an exclusive dataset for the detection of hemorrhages in retinal images using a rule-based and machine learning approach in which 70% and 30% of the dataset were used for training and testing respectively.

Some researchers have only used exclusive datasets and others have used a combination of exclusive dataset as well as publicly available datasets. Tan, Fujita [87] used an exclusive dataset, CLEOPATRA for the detection of different eye lesions through deep learning method, whereas Wang, Tang [94] used multiple datasets, including two publicly available dataset (ROC and DIARETDB1) and one exclusive dataset (Moorfields Eye Hospital London).

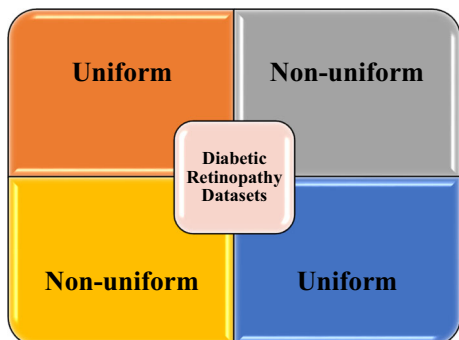
For the classification of DR into its severity levels, all the public datasets that were imbalanced and the authors have to apply different augmentations techniques or selecting almost equal amount of data for training and testing. This may result in the loss of important data and since they augment data, there is a chance that these systems are not accurate. Therefore, there is a need for a balanced dataset.

3.1.1 Dataset characteristics

Some diabetic retinopathy datasets are captured with the same device at a single site or at multiple sites, while others are captured with different devices at a single or at multiple sites. The devices used to capture the images can be broadly categorized into two namely, uniform devices and non-uniform devices. Similarly, the data sites can also be uniform or non-uniform (as shown in Fig. 4).

Uniform device & uniform site In this type of dataset, the images were captured with the same device at a single site or hospital. For instance, the exclusive dataset used by Takahashi, Tampo [86] from Jichi Medical University Japan contained 4709 images. These images were

Fig. 4 Data collection sites and devices



captured using a fundus camera, i.e. AFC-230; NIDEK Co., Ltd., Aichi, Japan from May 2011 to June 2015.

Uniform device & non-uniform site The images in this dataset were captured with the same device at multiple sites. For instance, Antal and Hajdu [8] used Messidor dataset. This dataset contains 1200 images taken by 3 ophthalmological departments in France, using Topcon TRC NW6 non-mydratic color video 3CCD camera. Out of 1200 images, 400 images were taken without pupil dilation and while the other 800 images were taken with pupil dilation, that is, one drop of Tropicamide at 0.5%.

Non-uniform device & uniform site The images in the datasets were captured with multiple devices at a single site. In the selected articles, there is not a single study that falls under this category.

Non-uniform device & non-uniform site The images in the datasets were captured with multiple devices at multiple sites. More than half of the selected articles fall under this category. This is because, they have either used a single diverse dataset or multiple datasets. For instance, the single dataset, containing 298 images used in [87], was gathered at 15 different sites and using 10 different types of cameras. On the other hand, the datasets used by Fraz, Jahangir [25] comprising of four publically available datasets, are of course, gathered at different places and captured with different cameras, compared to one another.

The number of studies lying in each category are shown in Table 7.

A total of 23 researches were involved in datasets where the images were captured with the same device at one particular site as shown in Table 7. Six (6) studies were on datasets in which the images were from the same device but captured at different sites or hospitals. No research has been conducted in which the images were captured with different devices but at the same site or hospital. A total of 40 investigations involved a dataset or multiple datasets where the images were captured with different devices and at different sites or hospitals.

3.2 Preprocessing techniques

Retinal images are preprocessed for the network for better visualization to extract more distinctive features once the image is clearer, bright and field of view is under consideration. Following is the brief description of the preprocessing techniques applied by the researchers of the selected academic articles.

Green Channel Extraction is applied on an RGB image to extract only the green band of the image as it provides more information on the insights of the image. Contrast enhancement is applied on the green channel of the image to improve its contrast. Illumination correction is usually applied after contrast enhancement, to increase the luminance and brightness of the

Table 7 Number of researches and types of dataset

	Uniform device	Non-Uniform devices
Uniform site	23	0
Non-Uniform sites	6	40

image. Gaussian filtering is the process of removing noise from the image and making it smoother. The image is resized into an image of lower resolution according to the appropriate network. Sometimes the researchers have to eliminate and mask optical disc and blood vessels so that they are not misclassified as diabetic retinopathy lesions. Augmentation is an artificial process of making equivalent amount of instances in different classes. Augmentation technique is used when there is imbalance in the images of different classes. Images are rotated, flipped, mirrored or translated to produce other instances of the selected images for a class where the number of DR images is lower than the other classes.

Augmentation also includes the processes of contrast scaling, enhancement, mirroring, shearing and so on. Quellec, Charrière [66] applied rotation in the range of 0 to 360, translation in the range of -10 pixels to 10 pixels, scaling in the range of 85% to 115%, contrast modification in the range of 60% to 167% and horizontal flipping.

Sometimes an RGB image is converted into a gray-scale image followed by further processing. Only 6 of the studies in our discussions utilized gray-scale conversion and all of these approaches were machine learning ones.

Most preprocessing techniques employ green channel extraction from the RGB image, as the green band shows maximum information as compared to the red and blue bands. Out of the 74 studies, 33 first extracted green channel and then did further preprocessing. For instance, Zhou, Wu [105] first extracted the green channel of the image for microaneurysms (MAs) detection, where the MAs exhibit better contrast from the background.

Contrast enhancement is also a popular preprocessing technique and is used in 30 studies. For instance, again taking the example of Zhou, Wu [105], the researchers enhanced the contrast on the extracted green channel by applying contrast limited adaptive histogram equalization (CLAHE) method so that the hidden features become visible.

Resizing the image according to the network is also a renowned preprocessing method, either using a machine learning technique or a deep learning technique, an image with high resolution is resized to conserve the computational power of the system. Resizing was carried out in 29 of the 74 selected academic studies. As an example, Doshi, Shenoy [22] scaled down the images to 512×512 pixels so that the training dataset is a standardized dataset. Similarly, Chudzik, Majumdar [19] resized the image in such a way that its width was equal to the smallest image width of the E-Ophtha dataset.

Most of the DR datasets contain images which have a black border, the researchers tend to segment the useless black border in order to concentrate on the region of interest (ROI).

Table 8 shows the various preprocessing techniques used in selected academic studies.

3.3 Diabetic retinopathy detection through machine learning approaches

This section discusses the ML approaches used for DR detection. The generic ML process for DR detection is shown in Fig. 5. As can be seen here, initially, the set of labeled images (a.k.a. training set) is collected for the construction of the DR detection model. This training set comprises of images belonging to different DR classes. Afterwards, several image preprocessing techniques are applied to remove unnecessary image features from the collected images. Later, various feature extraction techniques are employed to discover the most discriminative features from the preprocessed images. The outcome of this step is the master feature vector (MFV). This MFV is then fed as an input to ML algorithm to construct the DR detection model after learning the classification rules from the MFV. The performance of the constructed DR detection model is then evaluated on new unlabeled images (a.k.a. test set). Several

Table 8 Preprocessing techniques used in selected studies

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	References
✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	[53, 105, 106]
✓	✗	✓	✗	✓	✓	✗	✗	✗	✗	[92, 104]
✗	✓	✗	✗	✓	✗	✗	✓	✗	✓	[100]
✗	✗	✗	✗	✓	✗	✗	✓	✗	✓	[47, 62, 99]
✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	[9–11, 21, 74, 97]
✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	[98]
✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	[93, 94]
✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	[56, 90]
✗	✓	✗	✗	✓	✗	✗	✓	✗	✓	[89]
✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	[34, 87, 88]
✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	[33, 75, 86]
✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	[46, 55, 73, 83]
✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	[50, 79]
✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	[36, 72]
✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	[5, 15, 69]
✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	[67]
✗	✗	✓	✓	✓	✗	✗	✓	✗	✓	[13, 66]
✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	[64, 65]
✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	[25, 63]
✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	[18, 60]
✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	[49, 59]
✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	[57]
✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	[48]
✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	[41]
✗	✗	✗	✗	✓	✗	✓	✓	✗	✓	[38]
✓	✗	✗	✓	✗	✓	✓	✗	✓	✗	[35]
✗	✓	✓	✗	✓	✗	✗	✓	✗	✓	[27, 29, 30]
✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	[28]
✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	[24, 26]
✓	✓	✗	✗	✗	✗	✓	✓	✗	✗	[22]
✓	✗	✗	✗	✓	✗	✗	✓	✗	✓	[19]
✗	✓	✗	✗	✗	✓	✓	✗	✗	✓	[12, 14]

** P1 = Green Channel Extraction, P2 = Contrast Enhancement, P3 = Illumination Correction, P4 = Gaussian Filter, P5 = Resize, P6 = Optical Disc and Blood Vessels Removal, P7 = Histogram Equalization, P8 = Augmentation, P9 = Gray-scale Conversion, P10 = Crop for Region of Interest

performance metrics including accuracy, sensitivity, specificity, and area under the curve (AUC) are used to measure the classification performance.

In this review, 40 studies out of 74 employed machine learning approaches for the detection of diabetic retinopathy. Hence, in this section, we present the review of various features and ML algorithms that have been employed on DR image datasets to construct and evaluate the DR detection model. The subsequent subsections provide the details of these review aspects.

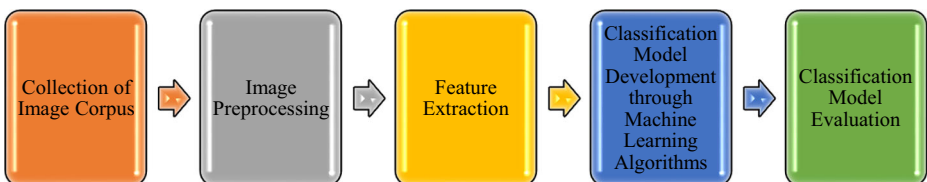


Fig. 5 Typical machine learning architecture

3.3.1 Features analysis

In the fields of image classification and machine learning, a feature is a measurable property of the overall activity that is under consideration. Generally, image processing and image segmentation are studied well in the literature. For instance, in [16], the authors segmented images based on full feature coverage, whereas, for the segmentation of medical images, the authors specifically proposed Active Contour Model (ACM) in [102]. Similarly, the authors applied region-based ACM through the relationship of neighboring pixels for the segmentation of medical images in [103]. In diabetic retinopathy detection algorithms, features may include whether the lesions (including hard exudates, soft exudates, microaneurysms, hemorrhages) are present in the image. The selection of features is an important task, so that the learning of the network becomes efficient and accurate. Different researchers of the selected academic articles have proposed and used different features to detect diabetic retinopathy. These features include, shape and structure-based features, color-based features, intensity features, statistical features and texture-based features. Furthermore, the researchers also combined hemorrhages and microaneurysms (HMA) as a single shape and structure-based features [78]. The subsequent paragraphs present the brief description of these features:

Shape and structure-based features These features include the shape and structure of various lesion of diabetic retinopathy, which includes hard and soft exudates, hemorrhages and microaneurysms. For instance, the shape-based features used by Zhou, Wu [105] for the detection of microaneurysms were area and perimeter, axis length, circularity and compactness.

Color-based features These features are based on the RGB colors of the image. For instance, Jaya, Dheeba [36] used 4 color-based features for the detection of hard exudates using color fundus images. They created color histograms using RGB color-space. These were, the normalized values of ‘R’ and ‘G’, extraction of luminance from color-space and the ‘Red-Green’ channel.

Intensity features Intensity is the pixel intensity expressed in the R, G and B planes. For instance, Bui, Maneerat [14] used intensity features for the detection of cotton-wool spots in DR images. Similarly, Kavitha and Palani [41] used intensity features by calculating maximum and minimum pixel intensities for the detection of hard and soft exudates.

Statistical features Statistical features are used for the statistical measurements from the pixels of a diabetic retinopathy image. For instance, Xiao, Yu [98] used statistical features along with color features to detect hemorrhages in retinal images. The statistical features used were mean, maximum, minimum, standard deviation.

Texture-based features These features give important information about the texture of the diabetic retinopathy images. Some researchers use gray-level intensities of the color fundus images as a feature through GLCM (Gray Level Co-occurrence Matrix). For instance, Vanithamani and Renee Christina [90] used four GLCM-based features including, entropy, cluster shade, dissimilarity and correlation. Similarly, Nijalingappa and Sandeep [55] used GLCM for textural feature extraction including, contrast, correlation, energy, homogeneity,

entropy, sum variance, difference variance, sum entropy, difference entropy, sum average, inverse difference moment.

In machine learning approaches, authors have employed different types of features in the selected studies, including shape, color, intensity, statistical and texture based features. The most frequently used combination of features is shape-based and statistical features. It has been used in 7 academic articles out of 40, followed by only statistical features (5 times). Individually, the most commonly used features in the researchers' community are again shape-based and statistical, both used 21 times in 40 articles. Texture-based features were used 15 times, color-based and intensity features both used 11 times. From this information, we can conclude that shape, statistical and texture based features are the most discriminative features in DR detection algorithms are can produce promising results for the researchers working in the field of machine learning based DR detection and classification.

Table 9 concludes 40 academic articles with their features applied using machine learning approaches.

3.3.2 Review of machine learning algorithms

In the selected primary studies belonging to ML approaches, authors have employed different kinds of machine learning algorithms to construct the DR classification model. Thus, this section is responsible for discussing the machine learning algorithms employed in selected primary studies. In general, eleven distinct machine learning algorithms were used in the selected primary studies. These are; Support Vector Machine (SVM), Random Forest (RF), k -Nearest Neighbor (k NN), Local Linear Discrimination Analysis (LLDA), Naïve Bayes (NB), Artificial Neural Networks (ANN), Decision Tree (DT), AdaBoost (AB), Self-adaptive Resource Allocation Network classifier (SRAN), Unsupervised Classifiers (UC) and Ensemble Classifiers. The details of these algorithms are discussed in subsequent sections. In addition, Table 10 shows study-wise usage of machine learning algorithms in each selected primary study. In Table 10, the first column represents the machine learning algorithms that were used

Table 9 Features used in selected studies

Shape	Color	Intensity	Statistical	Texture	References
✓	✓	✓	✗	✗	[105, 106]
✓	✗	✓	✗	✓	[8, 97]
✓	✓	✓	✓	✗	[96]
✗	✓	✗	✓	✗	[98]
✓	✗	✗	✓	✗	[34, 59, 61, 63, 79, 81, 94]
✓	✗	✗	✗	✗	[24, 53, 54, 83]
✗	✗	✗	✓	✓	[28, 41, 74]
✓	✓	✗	✓	✗	[15, 72]
✗	✓	✗	✗	✓	[36, 69]
✗	✗	✗	✓	✗	[11, 26, 44, 49, 67]
✗	✗	✗	✗	✓	[55, 56, 90]
✗	✗	✓	✗	✓	[35, 48]
✓	✓	✗	✗	✓	[25]
✗	✓	✓	✓	✗	[14]
✗	✓	✓	✗	✓	[12]
✓	✗	✗	✓	✓	[10]
✓	✗	✓	✗	✗	[5, 92]

by different researchers in the 40 selected studies, the second column shows the preeminent algorithm that outperformed other algorithms and obtained highest classification results. The third column displays those algorithms with which the preeminent technique is compared and lastly the performance of the preeminent technique is given. There were a few studies where the researchers have used only one algorithm and have not compared their machine learning classification techniques with any other classification algorithm. Therefore, the third column value is empty in such cases. Also, different performance measures have been employed in different studies for the classification purposes, so, the result of the obtained values have been shown with appropriate performance measure.

Support vector machine A Support Vector Machine (SVM) is basically a machine learning algorithm that is used for the purpose of classification. It draws a decision boundary (hyperplane) near the extreme points (support vectors) in the dataset. There are two classes A+ and A-, representing the closest distance to the positive and negative extreme points respectively. The plane that segregates A+ and A- classes is commonly known as the hyperplane, where one side of the plane consists of A+ class while the other contains A- class.

In several studies such as, [54, 69, 79, 83, 97] authors have employed SVM algorithms for the classification of different diabetic retinopathy lesions. In addition, the authors have reported better classification performance using the SVM.

Vanithamani and Renee Christina [90] compared the performance of SVM, SCG-BPN, and GRN algorithms to detect and classify exudates in retinopathy images. In experimental setup, authors have used the DIARETDB1 dataset comprised of 40 training images and 40 testing images. The experimental results showed that SVM algorithm obtained better classification performance compared to SCG-BPN and GRN algorithms. Jaya, Dheeba [36] used a Fuzzy SVM algorithm for the detection of hard exudates using an exclusive dataset of diabetic retinopathy images collected from different medical diagnostic centres. Their dataset consisted of 200 eye fundus images. They trained their machine learning network on 75 images and tested on all 200 images. According to the experimental results, Fuzzy SVM had better discriminating capability as compared to the traditional SVM. Carrera, Gonzalez [15] used two classification algorithms, namely, SVM and DT, to classify non-proliferative diabetic retinopathy into four classes. They used a subset of Messidor dataset comprising of 400 images and the results showed that SVM outperformed DT. Mahendran and Dhanasekaran [48] also used a subset of Messidor dataset consisting of 370 images to classify them according to their severity levels. In their experiments, they employed 150 images for training while 220 images were used for the purpose of testing. They compared the results obtained by SVM and PNN, where SVM algorithm showed better classification performance.

Wu, Xin [96] compared the performance of SVM, k NN and LLDA algorithms for microaneurysms detection using ROC dataset. Their experimental setup contained 50 images for training and 50 images for testing their algorithm. Their results proved that the performance of SVM algorithm was far better when compared with k NN and LLDA.

Random forest Random Forest (RF) is one of the most popular and powerful machine learning classification algorithm. It creates forests with decision trees (DT). Generally, the more trees in the forest, the more robust the projection is. To classify a new model based on the features, each tree gives a classification vote and the model is saved with the label of the tree. The forest chooses the class having the most number of votes. In other words, RF classification algorithm is similar to bagging technique. In RF, a subset

Table 10 Comparison of ML classifiers with performance of preeminent classifier

Study	Preeminent Classifier	Compared with	Performance
[105]	UC	–	0.6 FROC curve
[106]	UC	–	0.6 FROC curve
[97]	SVM	–	95% Accuracy
[98]	RF	–	93% Sensitivity
[96]	SVM	kNN, LLDA	0.286 Sensitivity
[94]	kNN	SVM, NB	0.464 ROC
[90]	SVM	SCG-BPN, GRN	96% Accuracy
[83]	SVM	–	0.973 AUC
[79]	SVM	–	1030 Mean Value
[74]	LMNN	SVM	88% Accuracy
[72]	PNN	RB, RBE, FFBPNN	100% Accuracy
[69]	SVM	–	92% Accuracy
[67]	DT, kNN	SVM	91% Accuracy
[63]	AB	–	75% Sensitivity
[61]	McNN	SRAN	97.6% Accuracy
[59]	PR	–	96% Accuracy
[56]	RBF	kNN	96.7% Accuracy
[55]	kNN	–	95% Accuracy
[54]	SVM	–	95% Accuracy
[53]	ED	–	89% Accuracy
[49]	Ensemble Classifier (HDT + FFNN)	HDT, LMNN	98% Accuracy
[48]	SVM	PNN	97.9% Accuracy
[44]	UC	–	89% Sensitivity
[41]	LMNN	–	90% Accuracy
[81]	BEC	–	87% Accuracy
[36]	Fuzzy SVM	SVM	93% Accuracy
[35]	HNN	–	97.3% Accuracy
[34]	MLP	–	93.8% Accuracy
[28]	ED	–	90.2% Accuracy
[26]	PNN	SVM	99% Accuracy
[25]	Ensemble Classifier (Bootstrap Decision Trees)	–	98.3% Accuracy
[15]	SVM	DT	80.4% Accuracy
[14]	ANN	–	85.5% Accuracy
[12]	UC	–	88.3% Sensitivity
[10]	McNN	SRAN, SVM	96.6% Accuracy
[8]	Ensemble Classifier (DT, kNN, AB, MLP, NB, RF, SVM, PR)	Single Best	90% Accuracy
[5]	ANN	–	94.4% Accuracy
[92]	LNN	SVM, MLP	94.4% Accuracy
[24]	MLP	–	95% Accuracy
[11]	Ensemble Classifier (FC, ANN, SVM)	FC, ANN, SVM	95.5% Accuracy

**UC = Unsupervised Classifier, SVM = Support Vector Machine, RF = Random Forest, kNN = k-Nearest Neighbor, LLDA = Local Linear Discrimination Analysis, NB = Naïve Bayes, SCG-BPN = Scaled Conjugate Gradient Back Propagation Network, GRN = Generalized Regression Neural Network, LMNN = Levenberg-Marquardt neural network, PNN = Probabilistic Neural Network, RB = Radial Basis fewer neurons, RBE = Radial Basis exact fit, FFBPNN = Feedforward Backpropagation Neural Network, DT = Decision Tree, AB = AdaBoost, McNN = Meta-cognitive Neural Network, SRAN = Self-adaptive Resource Allocation Network classifier, PR = Pattern Recognition network, RBF = Radial Basis Neural Network, ED = Expert-Driven Classification, HDT = Holoentropy enabled Decision Tree, FFNN = Feedforward Neural Network, BEC = Bagging Ensemble Classifier, HNN = Hopfield Neural Network, MLP = Multi-Layer Perceptron, ANN = Artificial Neural Network, LNN = Lattice Neural Network, FC = Fuzzy Classifier

of training set is created and a DT is constructed for each subset. While for the test set, all the DTs classify each input vector and finally the forest chooses the one having maximum votes.

From the selected 40 primary studies, RF classifier was employed only once. Xiao, Yu [98] used RF classifier for hemorrhage detection in the retinal images. They used 55 images from DIARETDB1 and 35 images from an exclusive dataset. For the training of the machine learning network, they used 70% of the total images and the rest of the 30% images were used for testing and classification with RF algorithm. The experimental results showed that the RF algorithm achieved good sensitivity.

***k*-nearest neighbor** *k*-Nearest Neighbor (*k*NN) algorithm is a fundamental and simplest machine learning classification algorithm. It classifies objects based on the nearest examples in feature space contained in the training set. “*k*” represents the number of nearest neighbors for the classifier to make its prediction.

*k*NN algorithm was used in several studies among the selected 40 machine learning articles. Nijalingappa and Sandeep [55] used *k*NN algorithm for the classification of diabetic retinopathy into its severity levels. In their experiments, 169 images from two public datasets were used including, Messidor and DIARETDB1 and an exclusive dataset. They trained their machine learning algorithm with 119 images and tested on the remaining 50 images. The classification results obtained with *k*NN algorithm are quite satisfactory. Wang, Tang [94] compared *k*NN, SVM and NB classification algorithms to localize microaneurysms in fundus images. For experimentation, they used three datasets including, ROC, DIARETDB1 and an exclusive diabetic retinopathy dataset collected from Moorfields Eye Hospital. Their experimental results showed that *k*NN algorithm outperformed the other two machine learning algorithms, namely SVM and NB. Similarly, Rahim, Jayne [67] compared the results of *k*NN, DT and SVM for microaneurysms detection in two publicly available diabetic retinopathy datasets, including, DIARETDB0 and DIARETDB1. The results obtained showed that *k*NN and DT performed better in comparison with SVM.

Local linear discrimination analysis Local Linear Discrimination Analysis is one of the most commonly used classification and dimensionality reduction technique. It can be used for multiclass discrimination. LLDA projects to a line which inherently preserves directions that is useful for data classification. It finds projection to a line, so that separation of samples from different classes is done.

LLDA was used only once in the selected primary machine learning studies. For microaneurysms detection, Wu, Xin [96] employed LLDA algorithm and compared the results with other two machine learning algorithms namely, SVM and *k*NN on the ROC dataset. According to their experimental results, LLDA algorithm was unable to perform well and SVM showed better accuracy when compared with LLDA and *k*NN.

Naïve bayes Naïve Bayes (NB) classification algorithm is a probability-based algorithm. It functions on numeric data and creates a probability model in the classes. For classification prediction, it only requires a small amount of numeric data. Thus, it is a fast and convenient classification algorithm.

NB algorithm was also employed only once in the 40 selected studies. Wang, Tang [94] compared three classification algorithms namely, NB, *k*NN and SVM for microaneurysms detection. They performed their experiments on two public and an exclusive dataset. In their experimental conditions, NB algorithm could not classify the microaneurysms images better than the other two classification algorithms and *k*NN was the most suitable classifier for their experiments.

Artificial neural networks Artificial Neural Network (ANN) usually comprises of three main layers, namely, input layer, hidden layer and output layer. The input and hidden layers contain many nodes and the output layer contains only one node. A neuron in a neural network is an activation node. The input layer forwards patterns to the hidden layer where actual processing is done. Random weights are assigned to the nodes in the hidden layer. The hidden layer is attached to the output layer which is responsible for the final output. It can be considered as a perceptron, which takes multiple inputs and produces one output.

Different varieties of ANN were used by different authors in the selected studies, which includes, Scaled Conjugate Gradient Back Propagation Network (SCG-BPN), Generalized Regression Neural Network (GRN), Levenberg-Marquardt neural network (LMNN), Probabilistic Neural Network (PNN), Radial Basis fewer neurons (RB), Radial Basis exact fit (RBE), Feedforward Backpropagation Neural Network (FFBPNN), Pattern Recognition network (PR), Meta-cognitive Neural Network (McNN), Radial Basis Neural Network (RBF), Hopfield Neural Network (HNN), Multi-Layer Perceptron (MLP) and Lattice Neural Network (LNN). Several authors have used a single ANN classification algorithm on diabetic retinopathy images and produced better results. In [5, 14, 24, 34, 35, 41, 59], authors have employed a single ANN algorithm and have reported this to be a better classification algorithm in the domain on diabetic retinopathy images classification. Shirbahadurkar, Mane [74] compared LMNN with SVM and proved the classification performance of LMNN algorithm to be better in their experimental setup. They used DIARETDB1 dataset and performed their experiments on lesions level. They trained their machine learning network with 652 lesions extracted from the dataset and tested with the same number of lesions. The experimental results showed that the LMNN algorithm outperformed SVM with better accuracy. Santhi, Manimegalai [72] and Ganesan, Martis [26] compared the PNN with multiple classifiers and after experimentation they stated that the performance of the PNN classification algorithm was better than others in their experimental setup.

Decision tree A Decision Tree (DT) is a simple algorithm that is mostly used in the classification problems. It has a flowchart-like structure, where the internal node is a representation of a test on an attribute, branch denotes an outcome of a test and the terminal node contains a class label. The top-most node in a tree is a root-node. In decision analysis, a DT can be used to represent the decisions. In DT, classification is done hierarchically, in a tree-like structure. One of the advantages of DT is that, it requires little or no data preparation. One disadvantage of DT is that, sometimes over-complex DTs are created, also known as overfitting. DTs can become unstable because a small variance in the data can produce a completely different tree. Thus, for performance improvement of DTs, multiple trees based structures were introduced, i.e. Random Forests (RFs).

Rahim, Jayne [67] compared the performance of DT, k NN and SVM to detect microaneurysms in retinopathy images using DIARETDB0 and DIARETDB1 datasets. 90% of the total images were used for training, while testing of these classification algorithms was performed on the remaining 10% of images. The authors reported that the DT and k NN algorithms outperformed with 91% accuracy and the same of SVM algorithm was 73%.

AdaBoost Adaptive Boosting (AdaBoost) is a general approach that can be applied to many statistical models. It works in a sequential manner and each tree is fitted on a modified version of the original dataset and finally creates a strong classifier.

In the selected primary studies, AdaBoost classification algorithm was used only once. Prentas and Loncaric [63] employed the AdaBoost algorithm for the detection of exudates in diabetic retinopathy images. In their experimental setup, they used disjoint sets of training and testing images from DRiDB dataset containing 50 images. The experimental results revealed that the AdaBoost obtained sensitivity of 75%.

Self-adaptive resource allocation network classifier As the name suggests, a Self-adaptive Resource Allocation Network classifier uses self-regularized phenomenon for choosing the training data and later on removes the redundant training data, thus, utilizing less memory and computational power. The selected samples having more information are then used for the training of the network.

The SRAN algorithm was employed twice in the selected primary machine learning studies but did not perform well when compared to the other classification algorithms. Ponnibala and Vijayachitra [61] and Bala and Vijayachitra [10] compared SRAN classification algorithm with McNN and SVM algorithms for the detection and classification of different eye related lesions. They used an exclusive dataset collected at the Lotus Eye Hospital, Coimbatore, India. Their experiments showed that the performance of SRAN was unsatisfactory in their experiments whereas, McNN brought them with better accuracies.

Unsupervised classifiers Usually when priori information is not available, then unsupervised classification is used. In such case, only the set of data and the features that belong to certain instances are known. In unsupervised classification, such a strategy is made to find groups of points in the feature space which is normally done with clustering approaches.

Unsupervised classification algorithms were employed several times in the selected studies. [Zhou, Wu [105], 61] employed unsupervised classification for microaneurysms detection using ROC dataset containing 100 images. In their experimentations, they used half the images for training and half for testing. The authors have reported that unsupervised classifiers achieved reasonable performance in their experimental results. Similarly, Kusakunniran, Wu [44] and Biyani, Patre [12] also used unsupervised classifiers for exudates detection in diabetic retinopathy images and reported that the unsupervised classification algorithms produced sensitivities of 89% and 88% respectively.

Ensemble classifiers Ensemble learning can be called group learning where multiple classification algorithms combine to form a model that is more accurate. It can be done in two ways, i.e. bagging and boosting. In bagging, several classification algorithms work in parallel and finally vote for the one which is more accurate. The one with majority votes becomes the final classifier. In boosting, various classification algorithms are employed in series. The weights of each model are adjusted based on the previous model. First, the data is divided into multiple parts, then one of them is validated using others and so on.

Ensemble classifiers were adopted by a few researchers in the selected articles. Mane, Jadhav [49] designed a classification algorithm which was an ensemble of HDT and FFNN. They compared its diabetic retinopathy images classification capability with HDT and LMNN individually and achieved an accuracy of 98% using two datasets, namely, DIARETDB0 and DIARETDB1. Fraz, Jahangir [25] designed an ensemble classifier with bootstrapped decision trees for the segmentation of exudates using 478 images from DIARETDB1, e-Ophtha Ex and Messidor datasets. In their experimental setup, they used 137 images for the training of machine learning network and 341 images for the testing of their ensemble-based classification

algorithm. They achieved 98% accuracy in their experimental results. Similarly, Somasundaram and Alli [81], Antal and Hajdu [8] and Barkana, Saricicek [11] also employed ensemble-based classification algorithms to classify diabetic retinopathy images and their experimental results produced accuracies of 95%, 87%, 90% respectively.

Figure 6 shows the frequency of preeminent classifiers across all 40 studies. As can be seen here, in most of the studies, ANN has shown the better classification results, followed by SVM and ensemble classifier. In addition, the tree-based classifiers including decision tree, random forest have been rarely used in the 40 studies. The Euclidean distance-based classifier including k NN also showed promising results in four selected studies. As can be seen here, we cannot conclude the single best classifier, as the classifier performed differently indifferent circumstances. Thus, it is suggested to compare the performance of various machine learning algorithms to evaluate which one performed best on collected images.

3.4 Review and analysis of diabetic retinopathy detection through deep learning approaches

This sections reviews the DL based approaches for DR detection. A generic DL based model for DR detection is shown in Fig. 7. Deep learning is nothing but an extension of machine learning with multilayers for feature extraction. The term “deep” refers to the depth of layers being involved in the DL architecture. Initially the dataset is collected that will be used for the training and testing of the DL architecture. Afterwards, some preprocessing techniques are applied on the dataset so that the network may learn the features from a clearer image. This preprocessed image is forwarded to the DL architecture for feature extraction and classification. The layers in the DL architecture use the output of the previous layer as its input, process it and pass it to the next layer. Finally, the last layer produces the diagnostic result.

Out of 74 studies, 31 selected studies used deep learning approaches for the detection of diabetic retinopathy while 3 approaches used a combination of machine learning and deep learning approaches.

3.4.1 Deep learning approaches employing transfer learning

According to the concept of transfer learning, the features learned using a deep learning architecture can be transferred to a similar deep learning architecture. 14 out of 31 studies have

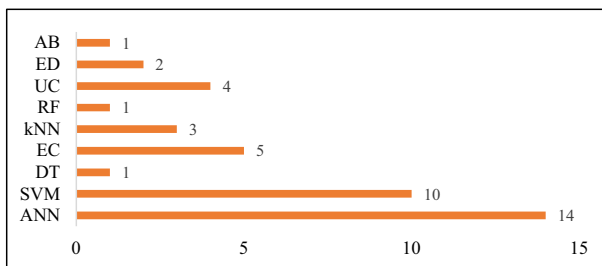


Fig. 6 Frequency of preeminent ML Algorithms used in the selected studies. ** AB = AdaBoost, ED = Expert-Driven classification, UC = Unsupervised Classification, FR = Random Forest, kNN = k-Nearest Neighbor, EC = Ensemble Classifier, DT = Decision Tree, SVM = Support Vector Machine, ANN = Artificial Neural Network



Fig. 7 A generic DL-based model for DR detection

used the transfer learning approach for the detection of diabetic retinopathy through deep learning.

Table 11 shows the list of researches which applied transfer learning for the detection of diabetic retinopathy. For instance, van Grinsven, van Ginneken [89] detected hemorrhages from color fundus images with the use of selective data sampling. They applied a 9 layered CNN which consisted of 5 convolution layers inspired by OxfordNet [76] where, in each convolution layer they used 32 filters of 3×3 size. Convolution layers were followed by ReLU and spatial max-pooling layers, a fully connected layer and lastly a softmax classification layer. They achieved a sensitivity of 91.90%, a specificity of 91.40% and AUC of 0.972 for hemorrhage detection using selective sampling on the Messidor dataset. Takahashi, Tampo [86] modified the GoogLeNet [84] to detect the stages of diabetic retinopathy stages. The modifications applied on the GoogLeNet were: first 5 accuracy layers were deleted, the crop size was 1272 pixels, and the batch size was reduced to 4. Using an exclusive dataset by Jichi Medical University, Japan, they trained 95% of the total data, tested on the remaining 5% data and achieved an accuracy of 81% and Prevalence And Bias-Adjusted Fleiss' Kappa (PABAK) value of 0.74. Shu Wei Ting, Yim-Lui Cheung [75] adapted 8 VGGNet [76] based CNN architectures to develop and validate a Deep Learning System (DLS) for the detection of diabetic retinopathy and other eye diseases including glaucoma and age-related macular degeneration (AMD). An exclusive dataset of multi-ethnicities was collected from 2010 to 2013 during the Singapore National Diabetic Retinopathy Screening Program (SIDRP) for this research. They achieved a sensitivity of 90.5%, specificity of 91.6% and AUC of 0.936 for referable diabetic retinopathy (results for glaucoma and AMD are not included due to the scope of the paper). Roy, Tennakoon [71] used a pre-trained CNN to estimate the severity level of diabetic retinopathy. The patches were fed into the CNN that was pre-trained on ImageNet [43]. They achieved a Kappa score (K-Sc) of 0.86 using a random forest classifier on Kaggle dataset. For the screening of diabetic retinopathy, Quellec, Charrière [66] used a 26 layered CNN adapted from o_O solution by Antony and Bruggemann which stood second in the Kaggle Diabetic Retinopathy Competition. The detection performance achieved by them was an AUC of 0.954 on the Kaggle dataset and 0.949 on e-optha dataset. Similarly, Gondal, Köhler [30] also adopted o_O solution for the localization of diabetic retinopathy lesions including, hemorrhages, microaneurysms, soft exudates and red small dots using the Kaggle dataset for training and DIARETDB1 dataset for evaluation. The modification they made on the o_O solution architecture was the replacement of final dense layers with Global Average Pooling (GAP) layer. For the classification of images as healthy or unhealthy, they achieved an AUC of 0.954 on DIARETDB1. Perdomo, Arevalo [60] explored LeNet architecture [45] with 8 layers including, input, convolution layer 1, max-pool layer 1, convolution layer 2, max-pool layer 2, fully connected layer 1, fully connected layer 2 and output layer. They used a softmax classifier in the output layer for the classification on diabetic retinopathy as healthy image or an image having exudates. The dataset used was e-optha and achieved a sensitivity of 99.8%, specificity of 99.6%, while the accuracy was 99.6%. Mansour [50] modified AlexNet [43] for

5 class diabetic retinopathy classification. The architecture consisted of an input layer, 5 convolutional layers, 2 fully connected layers and classification was done using the SVM classifier yielding a sensitivity of 100%, specificity of 93% and accuracy of 97.93% on the Kaggle dataset. For the classification of diabetic retinopathy images, Li, Pang [47] used different pre-trained CNN architectures, including AlexNet, GoogLeNet and several variations of the VGGNet architectures. They achieved a sensitivity of 86.03%, specificity of 97.11%, accuracy of 92.01%, and the AUC obtained was 0.9834 on the Messidor dataset using VGG-S architecture. Li, Zheng [46] modified U-Net [70] for exudates detection, where U-Net was actually designed for neuronal membranes segmentation. In the modified architecture, unpooling layers were used instead of deconvolutional layers of U-Net. They trained the network on e-optha EX dataset and tested on DIARETDB1, HEI-MED, Messidor and a subset of Kaggle dataset achieving AUC of 0.96 on DIARETDB1. Gulshan, Peng [33] used a variant of GoogLeNet, i.e. Inception-V3 to detect diabetic retinopathy in fundus images using Kaggle dataset and exclusive data collected from 3 Hospitals in India. For moderate or worse diabetic retinopathy, they achieved a sensitivity and specificity of 90.1% and 98.2% respectively. Choi, Yoo [18] classified retinal images into multiple categories (10 retinal diseases) while using a small dataset, i.e. STARE. They used transfer learning through a variant of VGGNet architecture, i.e. VGG-19 deep learning architecture with random forest classifier and achieved a sensitivity of 80.3%, specificity of 85.5% while the AUC was 0.903 for two class classification, i.e. normal and abnormal images. Abramoff, Lou [3] used a CNN architecture based on AlexNet with random forest classifier for the detection of diabetic retinopathy. They achieved sensitivity of 96.8%, specificity of 87.0%, negative predictive value was 99.0% and the AUC was 0.980 on Messidor-2 dataset. Mo and Zhang [52] applied transfer learning with the use of VGGNet architecture for the segmentation of retinal vessels and achieved sensitivity of 77.79%, specificity of 97.80%, accuracy of 95.21% and their results on AUC and Kappa score were 0.9782 and 0.7759 respectively.

Table 11 Studies employing transfer learning

Study	Method	Transfer Architecture	Results
[89]	CNN	OxfordNet	SE = 91.90%, SP = 91.40%, AUC = 0.972
[86]	CNN	GoogLeNet	PABAK = 0.74, Acc = 81%
[75]	CNN	VGGNet	SE = 90.5%, SP = 91.6%, AUC = 0.936
[71]	CNN	ImageNet	K-Sc = 0.86
[66]	CNN	<i>o_O</i> Solution	AUC = 0.954
[60]	CNN	LeNet	SE = 99.8%, SP = 99.6%, Acc = 99.6%
[50]	CNN	AlexNet	SE = 100%, SP = 93%, Acc = 97.93%
[47]	CNN	AlexNet, GoogLeNet, VGGNets	SE = 86.03%, SP = 97.11%, Acc = 92.01%, AUC = 0.9834
[46]	CNN	U-Net	AUC = 0.96
[33]	CNN	Inception-v3	SE = 90.1%, SP = 98.2%
[30]	CNN	<i>o_O</i> Solution	SE = 93.6%, SP = 97.6%, AUC = 0.954
[18]	CNN	VGGNet	SE = 80.3%, SP = 85.5%, AUC = 0.903
[3]	CNN	AlexNet	SE = 96.8%, SP = 87.0%, AUC = 0.980
[52]	CNN	VGGNet	SE = 77.79%, SP = 97.80%, Acc = 95.21%, AUC = 0.9782, K-Sc = 0.7759

** CNN = Convolutional Neural Network, SE = Sensitivity, SP = Specificity, AUC = Area Under Curve, PABAK = Prevalence And Bias-Adjusted Fleiss' Kappa, Acc = Accuracy, K-Sc = Kappa Score

3.4.2 Deep learning approaches using novel deep learning architectures

Out of 31 studies, 17 studies have produced novel deep learning architectures for the detection of diabetic retinopathy. Table 12 shows the studies where the authors have developed their novel DL architectures with the methods, number of layers in their DL architecture, the classifier that they employed and the results that they obtained through the novel DL architecture.

For diabetic retinopathy detection through novel deep learning architectures, different authors have presented their own architectures. There are distinct number of layers in the architecture and different classifiers have been used by the researchers. Most of the researchers have used Softmax classifier for the classification of diabetic retinopathy images. It has been used in 14 studies out of 17 selected studies, while in the rest of the 3 studies, the classifiers used were decision tree, pixel-wise classification and random forest, each of them was used only once. Yu, Xiao [104] proposed a 16 layers deep convolutional neural network architecture for exudates detection. They used CNN to identify exudates on pixel-wise classification through softmax classifier and achieved an accuracy of 91.92% and a sensitivity and specificity of 88.85% and 96% respectively. Yang, Li [100] used two CNN architectures, namely, local CNN and global CNN. The Local CNN was a 10 layer CNN used for the detection of lesions while the global CNN (26 layers) was used for the grading of diabetic retinopathy according to the standard [31]. They achieved a sensitivity of 95.90%, SP of 89.80% and the AUC was 0.9687. Xu, Feng

Table 12 Studies containing novel deep learning architectures

Study	Method	No. of Layers	Classifier	Results
[104]	CNN	16	Softmax	Acc = 91.92%, SE = 88.85%, SP = 96%
[100]	CNN	Local CNN = 10 Global CNN = 26	Softmax	SE = 95.90%, SP = 89.80%, AUC = 0.9687
[99]	CNN	16	Softmax	Acc = 94.5%
[87]	CNN	10	Softmax	SE = 87.58%, SP = 98.73%
[73]	SSAE	4	Softmax	SE = 91.16%, SP = 91.60%, Acc = 91.38%, AUC = 0.962
[64]	CNN	10	Softmax	SE = 77%, PPV = 77%, F-Sc = 77%
[65]	CNN	11	Softmax	SE = 78%, PPV = 78%, F-Sc = 78%
[62]	CNN	13	Softmax	SE = 95%, SP = 30%, Acc = 75%
[38]	CNN	17	Softmax	Acc = 75.70%
[29]	CNN	28	Softmax	K-Sc = 0.754, Prec = 88.20%, SE = 95%, Acc = 85%
[27]	CNN	6	Decision Tree	SE = 93%, SP = 87%, AUC = 0.94
[22]	CNN	29	Softmax	K-Sc = 0.3996
[19]	CNN	25	Pixel-wise	FROC = 0.562
[13]	CNN	9	Softmax	FROC = 0.221
[93]	CNN	6	RF	SE = 81.73%, SP = 97.33%, Acc = 97.67%, AUC = 0.9475
[88]	CNN	6	Softmax	SE = 75.37%, SP = 96.94%, Acc = 92.68%
[21]	CNN	10	Softmax	SE = 76.91%, SP = 98.01%, Acc = 95.30%, AUC = 0.9744

**CNN = Convolutional Neural Network, SSAE = Stacked Sparse Auto Encoder, Acc = Accuracy, SE = Sensitivity, SP = Specificity, AUC = Area Under Curve, PPV = Positive Predictive Value, F-Sc = F-Score, K-Sc = Kappa Score, Prec = Precision, FROC = Free-response Receiver Operating Characteristic

[99] presented a 16 layer deep architecture for early detection of diabetic retinopathy. They used softmax classifier and achieved an accuracy of 94.50%. Tan, Fujita [87] used a single CNN with 10 layers for simultaneous segmentation of different lesions including microaneurysms, hemorrhages and exudates. Using softmax classifier, they achieved sensitivities of 87.58%, 62.57% and 46.06% for exudates, hemorrhages and microaneurysms respectively. The specificities that they achieved were 98.73%, 98.93% and 97.99% for exudates, hemorrhages and microaneurysms respectively. Shan, Li [73] used a 4 layers deep Stacked Sparse Auto-Encoder (SSAE) architecture for microaneurysms detection. With the use of softmax classifier, they obtained a sensitivity of 91.16%, specificity of 91.60%, their accuracy was 91.38%, AUC was 0.962. [Prentašić and Lončarić [64], 32] used two deep architectures which were 10 and 11 layers deep respectively. The classifier used by them was softmax, which gave them sensitivities of 77% and 78%, PPV of 77% and 78% and F-Score of 77% and 78% respectively. Pratt, Coenen [62] proposed a CNN for the diagnosis of diabetic retinopathy and identification of its severity level. The classifier that was employed was softmax which attained a sensitivity of 95%, specificity of 30%, and accuracy of 75%. Jiang, Wu [38] used a 17 layer deep CNN based on the Caffe framework [37] for the two class classification of diabetic retinopathy using fundus images. They classified the images as either normal or having diabetic retinopathy through fully connected layer of the CNN architecture and achieved an accuracy of 75.7%. Ghosh, Ghosh [29] proposed a 28 layers deep CNN for two class and five class classification of diabetic retinopathy. They achieved a 95% accuracy for the two class and 85% accuracy for the five class classification using softmax classifier. Gargeya and Leng [27] classified the retinal fundus images into two classes, i.e. healthy or having diabetic retinopathy. They used a CNN for feature extraction and extracted 1024 features from the color retinal images, then initially classified them using softmax layer and decision tree classifier for final classification, achieving sensitivity of 93%, specificity of 87% and an AUC of 0.94. Doshi, Shenoy [22] presented a deep CNN based severity diagnosis of diabetic retinopathy. They used a 29 layers deep CNN and detected the five stages of diabetic retinopathy disease. The accuracy achieved on the ensemble of three CNNs was 0.3996 on kappa matrix. Chudzik, Majumdar [19] presented a 25 layered CNN for the detection of microaneurysms. The result achieved on the Free-response Receiver Operating Characteristic (FROC) metric was 0.562 using pixel-wise classification. Budak, Sengur [13] proposed a reinforcement sample learning algorithm through a 9 layered CNN for microaneurysms detection. The classifier they employed was softmax and the result achieved on the FROC was 0.221. Wang, Yin [93] used CNN for the segmentation of retinal blood vessels. The classifier employed by them was random forest and their results showed a sensitivity of 81.73%, sensitivity of 97.33%, accuracy of 97.67% and the AUC was 0.9475. Tan, Acharya [88] used a 7 layered single CNN with softmax classifier for automatic and simultaneous segmentation of blood vessels, fovea and optical disc. The sensitivities they achieved were 75.37%, 88.53%, 87.90% and the specificities were 96.94%, 99.14%, 99.27% for blood vessels, fovea and optical disc respectively. Their CNN achieved an average accuracy of 92.68% on the DRIVE dataset. For the segmentation of retinal vasculature, Dasgupta and Singh [21] presented a 10 layered CNN with softmax classifier. They tested their architecture on DRIVE dataset and achieved an accuracy of 95.33%, AUC of 0.9744, while the sensitivity and specificity were 76.91% and 98.01% respectively.

3.4.3 Combination of machine learning and deep learning approaches

In 3 out of the selected 74 studies, the authors have used a combination of deep learning and machine learning approaches. Table 13 shows the studies where the authors have applied a fusion of ML and DL based architectures for DR detection.

Orlando, Prokofyeva [57] used an ensemble of deep learning and machine learning approach for the detection of red lesions using three public datasets, namely, DIARETDB1, e-ophtha and Messidor. The features they extracted were based on the intensity and shape by applying transfer learned LeNet architecture [45] with 10 layers. They achieved a sensitivity of 97.21% and an AUC of 0.9347. Arunkumar and Karthigaikumar [9] used a Deep Belief Network (DBN) for the classification of diabetic retinal images. First, the deep features were extracted using Deep Belief Network (DBN) with three hidden layers, then these features were reduced using generalized regression neural network (GRNN) technique [20] and lastly, SVM was used for the classification of images. They achieved a sensitivity of 79.32%, specificity of 97.89% and the accuracy of 96.73% on publicly available ARIA dataset. Abbas, Fondon [1] developed a Deep Learning Neural Network (DLNN) to find out the severity level of diabetic retinopathy in images by learning Deep Visual Features (DVsFs). They used Dense Color Scale-Invariant Feature Transform (DCColor-SIFT) [2] and Gradient Location-Orientation Histogram (GLOH) [51] for feature extraction and then transformed features by using Principle Component Analysis (PCA). Afterwards, a three layered deep neural network was used to learn these features and finally an SVM classifier was used for the classification of diabetic retinopathy images into five severity levels, which includes, normal, moderate NPDR (Non-Proliferative Diabetic Retinopathy), Mild NPDR, Severe NPDR and PDR (Proliferative Diabetic Retinopathy). Their results show that, they achieved a sensitivity of 92.18%, specificity of 94.50% and AUC of 0.924 on three publicly available datasets (namely, DIARETDB1, Messidor, Foveal Avascular Zone) and one exclusive dataset (namely, Hospital Universitario Puerta del Mar, HUPM, Cádiz, Spain).

3.5 Review and analysis of performance measures

Table 14 shows the various performance measures used in selected studies. The detailed discussion of the performance metrics, namely, sensitivity, specificity, accuracy, area under curve, f-score, precision and positive predictive value can be found in [80]. Similarly, discussion on PABAK, Kappa Score and Youden's Index can be found in [17, 101] respectively. In the majority of the selected academic articles, the researchers have used sensitivity,

Table 13 Studies containing the combination of machine learning and deep learning

Study	Method	Features	No. of Layers	Classifier	Results
[57]	CNN	Intensity, Shape	10	RF	SE = 97.21%, AUC = 0.9347
[9]	DBN	DBN-based features extraction	3	SVM	SE = 79.32%, SP = 97.89%, Acc = 96.73%
[1]	CNN	DCColor-SIFT, GLOH	3	Softmax	SE = 92.18%, SP = 94.50%, AUC = 0.924

** CNN = Convolutional Neural Network, DBN = Deep Belief Network, RF = Random Forest, SVM = Support Vector Machine, SE = Sensitivity, SP = Specificity, AUC = Area Under Curve, Acc = Accuracy, DCColor-SIFT = Dense Color Scale-Invariant Feature Transform, GLOH = Gradient Location-Orientation Histogram

Table 14 Performance measures used in the selected articles

SE	SP	Acc	AUC	F-Sc	Prec	Pabak	K-Sc	PPV	Youden	References
X	X	X	✓	X	X	X	X	X	X	[13, 19, 46, 66, 83, 105, 106]
✓	✓	✓	X	✓	X	X	X	X	X	[104]
✓	✓	✓	✓	X	✓	X	X	X	X	[21, 100]
X	X	✓	X	X	X	X	X	X	X	[34, 38, 67, 69, 99]
✓	✓	✓	X	X	X	X	X	X	X	[5, 9, 14, 35, 41, 44, 48–50, 53, 55, 60–62, 72, 81, 88, 90, 92, 97]
✓	✓	X	X	X	X	X	X	X	X	[3, 12, 28, 33, 87, 98]
✓	X	X	X	X	X	X	X	X	X	[96]
✓	X	X	✓	X	X	X	X	X	X	[1, 8, 18, 27, 30, 75, 89, 94]
X	X	✓	X	X	X	✓	X	X	X	[86]
✓	✓	✓	X	X	✓	X	X	X	X	[74]
✓	✓	✓	✓	✓	✓	X	X	X	X	[73]
X	X	X	X	X	X	X	✓	X	X	[22, 71]
✓	X	X	X	✓	X	X	X	✓	X	[63–65]
✓	X	✓	X	X	✓	X	X	X	X	[59]
✓	X	X	✓	X	X	X	X	X	X	[57]
✓	✓	✓	X	X	X	X	X	✓	X	[10, 24, 26, 56]
✓	✓	✓	✓	X	X	X	X	X	X	[15, 47, 54, 93]
✓	✓	✓	✓	X	X	X	X	X	✓	[36]
✓	X	✓	X	X	✓	X	✓	X	X	[29]
✓	✓	✓	✓	X	X	X	X	✓	X	[25]
✓	✓	✓	✓	X	X	X	✓	X	X	[52]
X	X	✓	✓	X	X	X	X	X	X	[11]

** SE = Sensitivity, SP = Specificity, Acc = Accuracy, AUC = Area Under the receiver’s operating Curve, F-Sc = F-Score, Prec = Precision, Pabak = PABAK (Prevalence and Bias-Adjusted Kappa), K-Sc = Kappa Score, PPV = Positive Predictive Value, Youden = Youden’s Index

specificity, accuracy and area under the receiver’s operating curve (AUC) as their evaluation metrics to measure the classifier performance. The combination of performance measures that was found to be frequently used by the researchers was Sensitivity, Specificity and Accuracy. This combination was used 20 times out of total 74 studies, followed by Sensitivity, Specificity and AUC which was used 8 times. Some researchers have used Recall instead of Sensitivity, we have accommodated Recall under Sensitivity rather than using it as a different performance measure.

The performance measures that were frequently used by the researchers include, Sensitivity (57 times), Specificity (50 times), Accuracy (45 times), and AUC (27 times). Other performance measures that were not frequently used by the researchers were, F-Score (5 times), Precision (6 times), PABAK (1 time), Kappa Score (4 times), Positive Predictive Value (8 times), and Youden’s Index (1 time).

4 Discussion and observations

This study presented the review of ML-based approaches and DL-based approaches that were employed to develop automated DR detection systems. The primary studies were selected from eight well-known academic databases. Furthermore, nine different filters were employed to select the quality primary studies for this review and to reduce the study selection biasness. After the rigorous selection, a total of 74 primary studies were selected for this review. This

review discusses the selected papers based on five different aspects, namely, the datasets, image preprocessing techniques, ML-based DR detection approaches, DL-based DR detection approaches, and performance measures used to evaluate the performance of the classification model.

In the selected articles, researchers have either used exclusive datasets or publicly available datasets (see Table 5). In several studies some researchers have developed their own exclusive datasets (comprised of DR images) collected from any particular program or hospital. Our review revealed that several studies used exclusive dataset. However, such studies suffer from two major weaknesses. First, the developed classification model was based on single modality where the images were collected from one hospital and those images were taken from one particular device. Thus, the classification model trained from the collected images may not be applied on a wider scale. This is because, these days, there are so many medical devices for capturing the retinopathy images. Each device may have their own quality of images with different features and parameters. Thus, the collection multimodal DR images is suggested to develop the classification model where the DR images should be collected from different hospitals and the captured images should be captured from a variety of medical devices. Such multimodal dataset can produce more robust classification model and it can be deployed at a wider scale. Second, in many studies, the exclusive datasets contain a small number of DR images. Hence, the reported classification model may suffer from over-fitting or under-fitting. Thus, to overcome these limitations, many studies employed publicly available standard datasets. However, in such studies, there is still a need of more accurate DR detection systems that can yield better classification performance in terms of accuracy and speed.

Several datasets that were used for DR detection in selected primary studies suffer from data imbalance issue in which the images across the classes are not equally distributed. For instance, Ghosh, Ghosh [29] used the Kaggle dataset to detect DR and classify it into its stages. The Kaggle dataset is a publicly available dataset and it is provided with class labels for its training set only. The training set consists of 35,126 and test set contains 53,576 eye fundus images. The provided train data is imbalanced (as shown in Table 15) in a way that images with class label 0 are 25,810 (73.5%), class label 1 are 2443 (6.90%), class label 2 are 5292 (15.10%), class label 3 are 873 (2.50%) and class label 4 are 708 (2.00%). Therefore, the researchers who use such imbalanced data have to either downsize the train data (resulting in the loss of important data), or upsize the train data of any particular class using any image augmentation technique (resulting in biased classification results). Several studies have reported that image augmentation technique to deal with the data imbalance issue is suitable to achieve better classification performance [76, 84]. However, it is suggested that image augmentation is only used in the training set and not in the testing. This is because, if the test set (with no augmented images) yields a better classification accuracy, then, the classification results can be employed at a wider scale. Thus, it is highly recommended that a standardized

Table 15 Distribution of DR images in Kaggle dataset

Class Label	No. of Images	Percentage
0 – No DR	25,810	73.50%
1 – Mild DR	2443	6.90%
2 – Moderate DR	5292	15.10%
3 – Severe DR	873	2.50%
4 – Proliferative DR	708	2.00%

publicly available diabetic retinopathy dataset for classification of diabetic retinopathy into its stages is necessary. Furthermore, in case of data imbalance, the image augmentation should be done at the training stage but not in the testing.

The collected images may contain several noisy features that may not be result-oriented for the classification task. Thus, such noisy features should be removed by employing different image preprocessing techniques. For instance, resizing an image from the original high resolution to low resolution reduces the computational time. Similarly, cropping a retinal image can also result in reducing the noisy features that do not contribute to the classification task. Furthermore, contrast enhancement is also one of the useful image preprocessing techniques that can be used to enhance the contrast on the extracted region of interest and to unhide the useful and result-oriented features. Finally, as discussed earlier, the image augmentation image preprocessing technique can play a major role in enhancing the classification performance. Although, the image preprocessing techniques play a useful role in the improvement of classification performance. However, such techniques also suffer from limitations in some cases. For instance, augmentation provides different flavors of the same image thus making the system learn the same image. Therefore, there is a chance for the system to produce biased results (as it is already trained with the same set of images in its different forms) and not to work properly when encountered with real world diabetic retinopathy image. In addition, using image preprocessing techniques, there is a chance that some important information of diabetic retinopathy lesions including microaneurysms, hemorrhages, hard exudates and soft exudates may be lost. Furthermore, cropping may result in the loss of some important information from the portion of the image that is being cropped. Therefore, it is suggested to compare the performance of DR classification task with and without employing various image preprocessing techniques. For instance, according to the trend in the selected articles, while using deep learning approaches, a color fundus image does not need to be converted into gray-scale image. This is because, an RGB image holds important information regarding different lesions of diabetic retinopathy.

To detect the DR, researchers have either employed ML-based approaches or DL-based approaches. From the 74 selected studies, 40 studies have employed ML-based approaches and remaining 34 have employed DL-based approaches. In ML-based approaches, the most important task is the discovering of useful features that can be provided as an input to machine learning algorithms to produce the classification model. Thus, several researchers have employed a variety of features (namely, shaped-based features, texture-based features, statistical features, color and intensity-based features) to evaluate the classification performance to assess which features can yield the better classification performance. Several ML-based studies reported that the fusion of shape-based, texture-based, and statistical features are proven to be results-oriented and discriminative for DR detection classification model. Thus, it can be concluded that shape, statistical and texture based features are the most discriminative features in DR detection algorithms and can produce promising results for the researchers working in the field of machine learning based DR detection and classification. The possible reasons for importance of these features could be because in shape-based features we extract the shape and structure of various lesion of diabetic retinopathy including hard and soft exudates, hemorrhages and microaneurysms. Furthermore, in these features we compute the area and perimeter, axis length, circularity, compactness, etc. of any particular lesion related to DR image. In addition, texture-based features may be used extensively because these features give important information about the texture of the diabetic retinopathy images producing good results.

Finally, the statistical features such as, mean, median, mode include the summarized information of overall features, may also generate better classification results.

In machine learning approaches, authors have either used supervised or unsupervised machine learning algorithms to classify the DR images. Nevertheless, several studies have employed supervised machine learning algorithm compared to unsupervised approaches. This is because, the supervised machine learning algorithms are more accurate than unsupervised algorithms [32]. Moreover, the supervised machine learning algorithms learn the classification rules from existing labeled samples (a.k.a. training set). For instance, Kusakunniran, Wu [44] used unsupervised machine learning methods for the segmentation of hard exudates using two datasets, namely, e-optha EX and DIARETDB1. First, they extracted the expected initial seeds by using a supervised machine learning technique, i.e. MLP and then employed an unsupervised clustering approach, i.e. Graph Cut on the earlier extracted initial seeds for final segmentation. The results show that, they could only achieve a sensitivity of 89% on the DIARETDB1 dataset and 56% on thee-optha EX dataset. Similarly, Biyani, Patre [12] also used an unsupervised machine learning approach based on clustering, i.e. k -Means Clustering for the detection of exudates. The datasets used was also the DIARETDB1 and they only achieved a sensitivity of 88%. Therefore, it can be seen from these results that the unsupervised machine learning approaches were unable to detect DR lesions accurately when compared to the supervised machine learning approaches. For instance, Al-Jarrah and Shatnawi [5] used ANN, which is a supervised machine learning approach, for non-proliferative diabetic retinopathy symptoms detection. The results showed that they achieved an accuracy of around 97% on DIARETDB1 dataset. Similarly, Naqvi, Zafar [54] also utilized a supervised machine learning algorithm, i.e. SVM for the detection of hard exudates using four retinal images datasets, namely, DIARETDB1, DR1, DR2 and STARE. The authors reported that SVM produced outstanding results by accurately identifying 95% of the images in the test dataset. Hence, it can be seen that the supervised ML approaches yielded better classification results compared to unsupervised approaches. In most of the studies that employed supervised ML approaches, ANN obtained better classification results followed by SVM and the ensemble classifier. Nonetheless, in several studies, authors have employed many machine learning algorithms on extracted master feature vector to develop classification models and to compare which algorithm performs best on the used datasets. In addition, several studies have employed only one machine learning algorithm to develop a classification model. Nevertheless, as per no free lunch theorem [76], no single machine learning algorithm is suitable for all types of images. Therefore, one should evaluate the performance of several machine learning algorithms to see which one can yield the best results. For instance, Wu, Xin [96] evaluated the performance of SVM, k NN, and LLDA to classify DR images. The authors' experimental results showed that SVM performed better than k NN and LLDA.

Out of 74 studies, 34 studies utilized deep learning based approaches for the detection of DR. Out of 34 studies, 14 used transfer learning, 17 developed their own novel models and 3 studies used a fusion of deep learning and machine learning approaches. The most popular and commonly used deep learning architecture in the selected 34 studies was CNN which was used in 32 studies while SSAE and DBN were used once each. From this, we can infer that CNN is the most suitable deep neural network specifically for the detection of diabetic retinopathy and generally for the diagnosis of any other medical pathological images. For the detection of diabetic retinopathy where deep learning is used, we noticed that the deep architectures performed well when there were fewer number of classes. However, if the number of classes increased, the performance of the deep learning architectures was unsatisfactory. For instance,

Choi, Yoo [18] used a VGGNet-based architecture for categorizing STARE dataset into ten distinct retinal diseases (10 classes) and were able to obtain an accuracy of only 30.5%. For three classes, the accuracy of the same deep architecture was increased up to 72.8%. Similarly, Ghosh, Ghosh [29] developed a deep learning model for the detection and classification of DR. For a two class classification (no DR or DR), they achieved an accuracy of 95% whereas, for a five class classification (five severity levels of DR), the accuracy was dropped by 10%. Thus, there is a need for a suitable deep architecture that is capable to accurately distinguish multi-retinal diseases into multi-classes. The deep learning methods have shown substantial potential in the field of health-care and particularly in diabetic retinopathy detection. Still there are some limitations of its adoption in the field of diabetic retinopathy detection using screening software, as highlighted in [95].

Most of the researchers used sensitivity, specificity and accuracy to evaluate the performance of their classifier. This combination is mostly used in the DR detection using machine learning techniques. For instance, Kavitha and Palani [41] used a machine learning approach for the detection of hard and soft exudates. They used Sensitivity, Specificity and Accuracy as their performance measure. Their results showed that, they achieved sensitivity of 100%, specificity of 66.67% and the accuracy was 90%. Another frequently used combination by the researchers' community was sensitivity, specificity and AUC. This combination of evaluation is suitable in deep learning approaches where there are imbalanced classes of DR images and the authors have made the classes balanced through augmentation or by selecting almost equal amount of diabetic retinopathy images in different classes of training and testing sets. For instance, Gargeya and Leng [27] used a deep learning approach for the identification of diabetic retinopathy. They used sensitivity, specificity and AUC for the performance measure of their proposed approach and attained a sensitivity of 93%, specificity of 87% and an AUC of 94%.

5 Future research directions and challenges

This section presents various research problems that the researchers were unable to address in the previous studies on diabetic retinopathy detection. Thus, significant research is still needed to improve the performances of different techniques for the detection of diabetic retinopathy. The research challenges that need to be addressed are given below.

- (1) **Generic DL architecture for medical images:** In the field of deep learning, several transfer learning architectures (such as, AlexNet, GoogLeNet, LeNet, and VGGNet) for object identification are available to be retrained on new set of images such as, medical images. Nonetheless, these architectures are less suitable for medical images as far as classification performance is concerned. For instance, Choi, Yoo [18] employed VGGNet for DR detection using eye fundus images and achieved approximately 85.5% specificity. This is because, such transfer learning architectures are constructed for objects like animals, foods, etc. Hence, because of this, such architectures may be unsuitable for real-time medical images. Therefore, a possible research in future could be to apply a transfer learning architecture that has been trained on suitable medical images instead of objects, acting as a generic architecture, and subsequently retrained to produce better medical images classification accuracy.

- (2) **The integration of deep learning, cloud computing, and telemedicine:** In general, the rural areas suffer from the lack of human resources specially in medicine. Thus, in such cases, telemedicine can play a vital role to overcome this limitation. In future, deep learning, cloud computing, and telemedicine can be integrated to detect DR from eye fundus images. For instance, in rural areas, the patient can use his or her mobile phone attached with retinal camera to capture the eye fundus image. That image can be uploaded on cloud computing environment where the DR detection model (constructed through the machine learning or deep learning approach) can be deployed. This deployed model will then detect DR from the uploaded image and send the detection outcome and prescription back to the patient.
- (3) **Automated Eliciting the optimal values for DL Architectures:** Deep learning has achieved encouraging results generally in the field of computer vision and specifically in DR detection but the background of DL architectures is not well understood and is considered as a black-box. For example, many researchers fine-tune the parameters of existing deep learning algorithms such as, CNN or AlexNet to improve the classification performance. However, in many cases, the background of DL architectures is not well understood and is considered as a black-box. Thus, it is still difficult to find out the suitable structure and optimal values for the number of layers and nodes in different layers. Also specific domain knowledge is required for the selection of values for the learning rate, number of epochs, and strength of the regularizer. Therefore, in future, automated optimization techniques can be proposed to find the optimal values for different DL architectural components on different DR datasets and other medical images related datasets.
- (4) **Training with limited data:** A deep learning architecture usually requires large number of retinal images for training. It may not produce satisfactory results in terms of accuracy if the training set is small. There are two solutions for this problem. First, use of different augmentation techniques, including rotation, flipping, cropping, and colour casting. Second, use of weak learning algorithms to collect training data. Further investigations are needed for the generation of ampler training images, so that, the DL architecture could be trained with robustness and learn more distinctive features.
- (5) **Active Deep Learning (ADL):** The concept of ADL is very much similar to Active Learning (AL) concept of machine learning. The difference is that, in ADL a deep learning based classifier is used compared to a machine learning based classifier used in AL. The motivation behind the use of ADL is two ways. First, being a DL based approach, it is more scalable than traditional ML approaches. Second, being an AL based approach, it uses less labelled data compared to other DL approaches. ADL approach is suitable in cases when training data is big and this approach trains the network by using fewer training labels, not all. In case of DR detection, ADL can be used in such scenarios where there is big set of training data involved, for example, training through Kaggle dataset, etc.
- (6) **Stream-based selective sampling:** Mostly DR detection techniques work on pool-based data, i.e. the dataset is provided to the ML or DL based architecture which classifies the data with a classifier. A real-world application could be stream-based selection of DR images when there is huge amount of data that needs to be diagnosed with limited memory or processing capacity. This scenario is applicable in such areas with higher rate of DR and DR screening program needs to be setup with limited resources.

- (7) **The use of multi-modal images for construction of DR detection model:** There could be different types of images for the detection of different eye related diseases including diabetic retinopathy. The types of images may include, eye fundus images and optical coherence tomography (OCT). For DR detection, usually eye fundus images are used and for other eye related diseases, like, diabetic macular edema, OCTs are used. There is a clear research direction to produce such an architecture that is flexible enough and can handle eye fundus images as well as OCTs for the detection of diabetic retinopathy. In existing studies, most of the studies used single modality for the construction of DR detection model. However, in future, multi-modal DR detection model can be constructed to classify any type of DR images.
- (8) **Investigate the fusion of Patients' Metadata with DR images to construct the DR detection model:** Instead of relying only on the images related data, integrating patients' metadata with the images can bring perceptive interrelation of these two. Patients' metadata may include, age, duration of diabetes, genetic information, patient's family history, haemoglobin and other influential clinical factors regarding diabetes and diabetic retinopathy [77]. Thus, the fusion of this metadata with the information extracted from DR images can produce promising results. This is an open research direction and further investigations are needed in this domain.
- (9) **Development of stronger deep learning models:** Since deep learning have already produced very promising results in the field of medical imaging and diabetic retinopathy detection, it will be challenging to further improve and produce more powerful deep learning models. One solution could be increasing the computational power by increasing network size [76, 85] while considering the issue of overfitting. Another solution could be developing a specific object-based model rather than image-based model. For instance, if the researchers are interested to detect a particular eye lesion (e.g. haemorrhages only), then they would design such a deep learning model that trains on haemorrhages only and other lesions in which they are not interested may not be learned by the model. It is substantiated in [58] that the object-based detection is more beneficial than the image-based detection.

6 Conclusion

This paper presented a comprehensive review on automated diabetic retinopathy detection techniques. The primary studies were searched from eight different academic databases. After a rigorous study selection process 74 studies were selected for this review. Each study is discussed in this review from the perspective of five different aspects, namely, datasets, image preprocessing techniques, ML-based classification techniques, DL-based classification techniques, and performance measures. Several publicly available datasets with unique characteristics were found and reported in this review. Among these datasets, DIARETDB1, Kaggle, and Messidor were most commonly used for DR detection. Among image preprocessing techniques, green channel extraction combined with contrast enhancement obtained better results. In ML-based approaches, shape-based, texture-based, and statistical features yielded better results. In ML algorithms, most of the studies reported that ANN produced better classification performance followed by SVM, and ensemble classifier. In deep learning mostly CNN was used to automatically extract and classify the DR images. In the aspect of performance measure, most researchers used accuracy, sensitivity, specificity, and area under

the curve as metrics. The significance of the review findings was discussed in a separate discussion section. This review also presented nine new research challenges in the field of DR detection. We believe that this comprehensive review will provide a deep understanding of field of DR detection techniques and also it will provide the valuable insights to the researchers working in this domain. This review has two limitations. First, only papers published from January 2013 to March 2018 were included in this review. Second, although every precaution was taken, it may be possible that the authors missed some important keywords, and some other bibliographic databases which may have some useful articles. Thus, in future, the scope of the review can be enhanced to overcome these limitations. Finally, it is believed that this review will prove useful for scientists working in the area of automated DR detection.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Abbas Q et al (2017) Automatic recognition of severity level for diagnosis of diabetic retinopathy using deep visual features. *Med Biol Eng Comput* 55(11):1959–1974
2. Abdel-Hakim AE, Farag AA (2006) CSIFT: A SIFT descriptor with color invariant characteristics. *Comput Vision Pattern Recogn, 2006 IEEE Comput Soc Conf. IEEE*
3. Abramoff MD et al (2016) Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci* 57(13):5200–5206
4. Aiello LP et al (1998) Diabetic retinopathy. *Diabetes Care* 21(1):143–156
5. Al-Jarrah MA, Shatnawi H (2017) Non-proliferative diabetic retinopathy symptoms detection and classification using neural network. *J Med Eng Technol* 41(6):498–505
6. Almotiri J, Elleithy K, Elleithy A (2018) Retinal vessels segmentation techniques and algorithms: a survey. *Applied Sciences-Basel* 8(2):31
7. Amin J, Sharif M, Yasmin M (2016) A review on recent developments for detection of diabetic retinopathy. *Scientifica*: 20
8. Antal B, Hajdu A (2014) An ensemble-based system for automatic screening of diabetic retinopathy. *Knowl-Based Syst* 60:20–27
9. Arunkumar R, Karthigaikumar P (2017) Multi-retinal disease classification by reduced deep learning features. *Neural Comput & Applic* 28(2):329–334
10. Bala MP, Vijayachitra S (2014) Early detection and classification of microaneurysms in retinal fundus images using sequential learning methods. *Int J Biomed Eng Technol* 15(2):128–143
11. Barkana BD, Saricicek I, Yildirim B (2017) Performance analysis of descriptive statistical features in retinal vessel segmentation via fuzzy logic, ANN, SVM, and classifier fusion. *Knowl-Based Syst* 118:165–176
12. Biyani RS, Patre BM, IEEE (2016) A clustering approach for exudates detection in screening of diabetic retinopathy. 2016 International Conference on Signal and Information Processing. IEEE, New York
13. Budak U et al (2017) A novel microaneurysms detection approach based on convolutional neural networks with reinforcement sample learning algorithm. *Health Inform Sci Syst* 5:10
14. Bui T, et al (2017) Detection of cotton wool for diabetic retinopathy analysis using neural network. 2017 Ieee 10th International Workshop on Computational Intelligence and Applications. IEEE, New York, pp. 203–206
15. Carrera EV, Gonzalez A, Carrera R (2017) Automated detection of diabetic retinopathy using SVM
16. Chen X, He F, Yu H (2018) A matting method based on full feature coverage. *Multimed Tools Appl*: 1–29
17. Chen G et al (2009) Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. *BMC Med Res Methodol* 9:5–5
18. Choi JY et al (2017) Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database. *PLoS One* 12(11):16
19. Chudzik P et al (2018) Microaneurysm detection using fully convolutional neural networks. *Comput Methods Prog Biomed* 158:185–192
20. Cigizoglu HK, Alp M (2006) Generalized regression neural network in modelling river sediment yield. *Adv Eng Softw* 37(2):63–68

21. Dasgupta A, Singh S (2017) A fully convolutional neural network based structured prediction approach towards the retinal vessel segmentation
22. Doshi D, et al (2016) Diabetic retinopathy detection using deep convolutional neural networks. in 2016 International Conference on Computing, Analytics and Security Trends (CAST)
23. Fong DS et al (2004) Diabetic retinopathy. *Diabetes Care* 27(10):2540–2553
24. Franklin SW, Rajan SE (2014) Computerized screening of diabetic retinopathy employing blood vessel segmentation in retinal images. *Biocybernet Biomed Eng* 34(2):117–124
25. Fraz MM et al (2017) Multiscale segmentation of exudates in retinal images using contextual cues and ensemble classification. *Biomedical Signal Processing and Control* 35:50–62
26. Ganesan K et al (2014) Computer-aided diabetic retinopathy detection using trace transforms on digital fundus images. *Med Biol Eng Comput* 52(8):663–672
27. Gargeya R, Leng T (2017) Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 124(7):962–969
28. Gegundez-Arias ME et al (2017) A tool for automated diabetic retinopathy pre-screening based on retinal image computer analysis. *Comput Biol Med* 88(C):100–109
29. Ghosh R, Ghosh K, Maitra S (2017) Automatic detection and classification of diabetic retinopathy stages using CNN
30. Gondal WM et al (2017) Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. 2017 IEEE Int Conf Image Process (ICIP)
31. Group, E.T.D.R.S.R (1991) Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification: ETDRS report number 10. *Ophthalmology* 98(5): 786–806
32. Guerra L et al (2011) Comparison Between Supervised and Unsupervised Classifications of Neuronal Cell Types: A Case Study. *Dev Neurobiol* 71(1):71–82
33. Gulshan V et al (2016) Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama-J Am Med Assoc* 316(22):2402–2410
34. Hanůsková V et al (2013) Diabetic retinopathy screening by bright lesions extraction from fundus images. *J Electr Eng* 64(5):311–316
35. Hemanth DJ, Anitha J, Indumathy A (2016) Diabetic retinopathy diagnosis in retinal images using hopfield neural network. *IETE J Res* 62(6):893–900
36. Jaya T, Dheeba J, Singh NA (2015) Detection of hard exudates in colour fundus images using fuzzy support vector machine-based expert system. *J Digit Imaging* 28(6):761–768
37. Jia Y et al (2014) Caffe: Convolutional architecture for fast feature embedding. *Proc 22nd ACM Int Conference on Multimedia. ACM*
38. Jiang Y, Wu H, Dong J (2017) Automatic screening of diabetic retinopathy images with convolution neural network based on caffe framework. *Proc 1st Int Conf Med Health Inform 2017. ACM: Taichung City: 90–94*
39. Jordan KC et al (2017) A review of feature-based retinal image analysis. *Expert Rev Ophthalmol* 12(3): 207–220
40. Joshi S, Karule PT (2018) A review on exudates detection methods for diabetic retinopathy. *Biomed Pharmacother* 97:1454–1460
41. Kavitha M, Palani S (2014) Hierarchical classifier for soft and hard exudates detection of retinal fundus images. *J Intell Fuzzy Syst* 27(5):2511–2528
42. Kolb H (1995) Simple anatomy of the retina. In: Kolb H, Fernandez E, Nelson R (eds) *Webvision: the organization of the retina and visual system. University of Utah Health Sciences Center Copyright: (c) 2018 Webvision, Salt Lake City*
43. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*
44. Kusakunniran W et al (2018) Hard exudates segmentation based on learned initial seeds and iterative graph cut. *Comput Methods Prog Biomed* 158:173–183
45. LeCun Y et al (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324
46. Li G, Zheng S, Li X (2018) Exudate detection in fundus images via convolutional neural network: 193–202
47. Li X, et al (2017) Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. in 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)
48. Mahendran G, Dhanasekaran R (2015) Investigation of the severity level of diabetic retinopathy using supervised classifier algorithms. *Comput Electr Eng* 45:312–323

49. Mane VM, Jadhav DV, Shirbahadurkar SD (2017) Hybrid classifier and region-dependent integrated features for detection of diabetic retinopathy. *J Intell Fuzzy Syst* 32(4):2837–2844
50. Mansour RF (2018) Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. *Biomed Eng Lett* 8(1):41–57
51. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. *IEEE Trans Pattern Anal Mach Intell* 27(10):1615–1630
52. Mo J, Zhang L (2017) Multi-level deep supervised networks for retinal vessel segmentation. *Int J Comput Assist Radiol Surg* 12(12):2181–2193
53. Mumtaz R et al (2018) Automatic detection of retinal hemorrhages by exploiting image processing techniques for screening retinal diseases in diabetic patients. *Int J Diab Dev Countries* 38(1):80–87
54. Naqvi SAG, Zafar MF, ul Haq I (2015) Referral system for hard exudates in eye fundus. *Comput Biol Med* 64:217–235
55. Nijalingappa P, Sandeep B (2016) Machine learning approach for the identification of diabetes retinopathy and its stages
56. Omar M, Khelifi F, Tahir MA (2016) Detection and classification of retinal fundus images exudates using region based multiscale LBP texture approach
57. Orlando JI et al (2018) An ensemble deep learning based approach for red lesion detection in fundus images. *Comput Methods Prog Biomed* 153(C):115–127
58. Ouyang W, et al (2014) Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. *arXiv preprint arXiv:1409.3505*
59. Paing MP, Choomchuay S, Rapeeporn Yodprom MD (2017) Detection of lesions and classification of diabetic retinopathy using fundus images
60. Perdomo O, Arevalo J, Gonzalez FA (2017) Convolutional network to detect exudates in eye fundus images of diabetic subjects
61. Ponnibala M, Vijayachitra S (2014) A sequential learning method for detection and classification of exudates in retinal images to assess diabetic retinopathy. *J Biol Syst* 22(3):16
62. Pratt H et al (2016) Convolutional neural networks for diabetic retinopathy. *Proc Comput Sci* 90:200–205
63. Prentasic P, Loncaric S (2014) Weighted ensemble based automatic detection of exudates in fundus photographs. *Conf Proc IEEE Eng Med Biol Soc* 2014:138–141
64. Prentasić P, Lončarić S (2015) Detection of exudates in fundus photographs using convolutional neural networks. 2015 9th Int Sym Image Signal Process Anal (ISPA)
65. Prentasic P, Loncaric S (2016) Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion. *Comput Methods Prog Biomed* 137:281–292
66. Quellec G et al (2017) Deep image mining for diabetic retinopathy screening. *Med Image Anal* 39:178–193
67. Rahim SS et al (2016) Automatic detection of microaneurysms in colour fundus images for diabetic retinopathy screening. *Neural Comput Applic* 27(5):1149–1164
68. Rahimy E (2018) Deep learning applications in ophthalmology. *Curr Opin Ophthalmol* 29(3):254–260
69. Reshma Chand CP, Dheebea J (2015) Automatic detection of exudates in color fundus retinopathy images. *Ind J Sci Technol* 8(26)
70. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. *Int Conf Med Image Comput Computer-Assist Interv. Springer*
71. Roy P, et al (2017) A novel hybrid approach for severity assessment of diabetic retinopathy in colour fundus images. 2017 Ieee 14th Int Sym Biomed Imaging. IEEE, New York: 1078–1082
72. Santhi D et al (2016) Segmentation and classification of bright lesions to diagnose diabetic retinopathy in retinal images. *Biomed Engineering-Biomedizinische Technik* 61(4):443–453
73. Shan J, Li L, IEEE (2016) A deep learning method for microaneurysm detection in fundus images. 2016 Ieee First International Conference on Connected Health: Applications, Systems and Engineering Technologies. IEEE, New York, 357-358
74. Shirbahadurkar SD, Mane VM, Jadhav DV (2018) Early stage detection of diabetic retinopathy using an optimal feature set: 15–23
75. Shu Wei Ting D et al (2017) Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA: J Am Med Assoc* 318(22):2211–2223
76. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*
77. Simó-Servat O, Hernández C, Simó R (2013) Genetics in Diabetic Retinopathy: Current Concepts and New Insights. *Curr Genom* 14(5):289–299
78. Sinthanayothin C et al (2002) Automated detection of diabetic retinopathy on digital fundus images. *Diabet Med* 19(2):105–112

79. Sisodia DS, Nair S, Khobragade P (2017) Diabetic retinal fundus images: preprocessing and feature extraction for early detection of diabetic retinopathy. *Biomed Pharmacol J* 10(2):615–626
80. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 45(4):427–437
81. Somasundaram SK, Alli P (2017) A machine learning ensemble classifier for early prediction of diabetic retinopathy. *J Med Syst* 41(12):1–12
82. Sopharak A, Uyyanonvara B, Barman S (2013) Automated microaneurysm detection algorithms applied to diabetic retinopathy retinal images. *Maejo Int J Sci Technol* 7(2):294–314
83. Srivastava R et al (2017) Detecting retinal microaneurysms and hemorrhages with robustness to the presence of blood vessels. *Comput Methods Prog Biomed* 138:83–91
84. Szegedy C, et al (2015) Going deeper with convolutions. *CVPR*
85. Szegedy C, et al (2015) Going deeper with convolutions. *Proc IEEE Conf Comput Vision Pattern Recogn*
86. Takahashi H et al (2017) Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. *PLoS One* 12(6):e0179790
87. Tan JH et al (2017) Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network. *Inf Sci* 420:66–76
88. Tan JH et al (2017) Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network. *J Comput Sci* 20:70–79
89. van Grinsven M et al (2016) Fast convolutional neural network training using selective data sampling: application to hemorrhage detection in color fundus images. *IEEE Trans Med Imaging* 35(5):1273–1284
90. Vanithamani R, Renee Christina R (2018) Exudates in detection and classification of diabetic retinopathy: 252–261
91. Vashist P et al (2011) Role of early screening for diabetic retinopathy in patients with diabetes mellitus: an overview. *Ind J Commun Med: Off Publ Indian Assoc Prev Social Med* 36(4):247–252
92. Vega R et al (2015) Retinal vessel extraction using lattice neural networks with dendritic processing. *Comput Biol Med* 58:20–30
93. Wang SL et al (2015) Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. *Neurocomputing* 149:708–717
94. Wang S et al (2017) Localizing microaneurysms in fundus images through singular spectrum analysis. *IEEE Trans Biomed Eng* 64(5):990–1002
95. Wong TY, Bressler NM (2016) Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *Jama* 316(22):2366–2367
96. Wu JY, et al (2015) New hierarchical approach for microaneurysms detection with matched filter and machine learning. 2015 37th Annual International Conference of the Ieee Engineering in Medicine and Biology Society. *IEEE, New York*: 4322–4325
97. Xiao ZT et al (2017) Automatic non-proliferative diabetic retinopathy screening system based on color fundus image. *Biomed Eng Online* 16:19
98. Xiao D, et al (2017) Retinal hemorrhage detection by rule-based and machine learning approach
99. Xu KL, Feng DW, Mi HB (2017) Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. *Molecules* 22(12):7
100. Yang Y, et al (2017) Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks: 533–540
101. Youden WJ (1950) Index for rating diagnostic tests. *Cancer* 3(1):32–35
102. Yu H, He F, Pan Y (2018) A novel segmentation model for medical images with intensity inhomogeneity based on adaptive perturbation. *Multimed Tools Appl*: 1–20
103. Yu H, He F, Pan Y (2018) A novel region-based active contour model via local patch similarity measure for image segmentation. *Multimed Tools Appl*: 1–23
104. Yu S, Xiao D, Kanagasigam Y (2017) Exudate detection for diabetic retinopathy with convolutional neural networks
105. Zhou W et al (2017) Automatic microaneurysm detection using the sparse principal component analysis-based unsupervised classification method. *IEEE Access* 5:2563–2572
106. Zhou W et al (2017) Automatic microaneurysms detection based on multifeature fusion dictionary learning. *Comput Math Methods Med* 2017



Uzair Ishtiaq received his Master degree in Computer Science from National University of Computer & Emerging Sciences, FAST, Islamabad, Pakistan. He is currently enrolled as a Ph. D. student in the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. His research interests include image processing, medical image analysis and deep learning.



Sameem Abdul Kareem received her B. Sc. (Hons) in Mathematics from the University of Malaya, and M.Sc. in Computing from the University of Cardiff and Ph.D. from University of Malaya. She is currently a Professor at Artificial Intelligence Department, faculty of Computer Science and Information Technology, University of Malaya. Her current research interests include Artificial Intelligence in Medicine, Machine Learning, Data Analytics/Mining, Image Processing and Biomedical Informatics.



Erma Rahayu Mohd Faizal Abdullah received her B. Sc. (Hons) in Computer Science with specialization in Multimedia from the University of Malaya, and M.Sc. in Computing from Oita University and Ph.D. from University of Technology Mara. She is currently an Assistant Professor at Artificial Intelligence Department, faculty of Computer Science and Information Technology, University of Malaya. Her current research interests include Artificial Intelligence, Machine Learning, Data Mining and Biomedical Informatics.



Ghulam Mujtaba is a Ph. D. student in the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. He has published several articles indexed in well reputed databases. His field of research is online social networking, text mining, deep learning, and information retrieval.



Rashid Jahangir is a Ph.D. student at FCSIT, UM. He received his Master degree from University of New South Wales, UNSW, Sydney, Australia. He is working on digital signal processing and deep learning. The key research areas of his interest are: deep learning, pattern recognition, machine learning and data mining.



Hafiz Yasir Ghafoor is a Ph.D. student at FCSIT, UM. He received his Master degree from Islamia University Bahawalpur, Pakistan. He is working on text-mining and social media user profiling. The key research areas of his interest are: deep learning, social media trends, opinion mining and data mining.

Affiliations

Uzair Ishtiaq^{1,2} · Sameem Abdul Kareem¹ · Erma Rahayu Mohd Faizal Abdullah¹ · Ghulam Mujtaba³ · Rashid Jahangir^{1,2} · Hafiz Yasir Ghafoor^{1,2}

Uzair Ishtiaq
uzair@ciitvehari.edu.pk

Erma Rahayu Mohd Faizal Abdullah
erma@um.edu.my

Ghulam Mujtaba
mujtaba@iba-suk.edu.pk

Rashid Jahangir
rashidjahangir@ciitvehari.edu.pk

Hafiz Yasir Ghafoor
yasirghafoor@ciitvehari.edu.pk

¹ Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

² Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari, Pakistan

³ Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan