Diabetic Retinopathy Binary Image Classification Using Pyspark

Bina Kotiyal
Department of Computer Science,
Gurukula Kangri (Deemed to be University), Haridwar, Uttarakhand, India.
Corresponding author: kotiyalbina@gmail.com

Heman Pathak
Department of Computer Science,
Gurukula Kangri (Deemed to be University), Haridwar, Uttarakhand, India.
E-mail: hpathak@gkv.ac.in

(Received on March 14, 2022; Accepted on June 18, 2022)

Abstract

Diabetic Retinopathy is a significant complication of diabetes, caused by a high blood sugar level, which damages the retina. In its earliest stages, diabetic retinopathy is asymptomatic and can lead to blindness if not discovered and treated promptly. As a result, there is a need for a reliable screening method. According to studies, this problem affects a large section of the population, and it is thus linked to Big Data. There are several obstacles and issues with Big Data, but Deep Learning is providing solutions to these issues. As a result, academics are extremely interested in Big Data with Deep Learning. It has been our goal in this study to employ effective preprocessing and Deep Learning approaches to accomplish binary classification of Diabetic Retinopathy. The experiment is done out using a dataset from Kaggle that was collected from India. The peculiarity of the paper is that the work is implemented on the Spark platform, and the performance of three models, InceptionV3, Xception, and VGG19 with the Logistic Regression classifier is compared. The accuracy of the models is used as a comparison criterion. Based on the results of the trial, the accuracy of InceptionV3 is 95 percent, the accuracy of Xception is 92.50 percent, and the accuracy of VGG19 is 89.94 percent. Consequently, InceptionV3 outperforms the other two models.

Keywords- Deep learning, Big data, Spark, Diabetic retinopathy, Image preprocessing.

1. Introduction

In this rapidly developing digital world Deep Learning (DL) and Big Data are seeking the enormous attention of researchers. Big Data stands for massive data on which it is arduous to apply the accustomed tools for managing, analyzing and generating meaningful data (Jan et al., 2019; Kotiyal et al., 2013; Singh et al., 2019). Due to the exponential growth in the digital world by means of volume, formats and shapes, it becomes especially important to master this massive data as per the utilities of the system. The prominence of technology-based company preserves data in Exabyte to name a few are Google, Facebook, and YouTube etc. Moreover, the traditional tools are not capable for managing it (Wilamowski et al., 2016). Therefore, the various companies have advanced outputs for the simulations, interpretation of data, monitoring, experimentation and several other requirements of business.

Big Data has some challenges associated to its characteristics such as the quality and size of data (Rehman et al., 2016). The Convolutional Neural Networks (CNNs) are employed to handle such kind of data. The need for large amounts of data to ensure the convergence of their training algorithms to achieve good accuracies is a common constraint of CNNs. However, some applications simply do not have enough data to train these models due to the various constraints such as security, cost, data acquisition etc. To overcome this limitation, some works employ fine-tuning strategies to transfer knowledge from one problem to the next. This means that rather than just randomly setting the weights of a CNN, the training procedure employs a CNN with previously trained weights to hasten the training algorithm's convergence

(Zavarez et al., 2017). One of the extensively used practices for amplifying the size of the training data is through data augmentation method (Gopalakrishnan et al., 2017). In this procedure the data is transformed or altered for generating more data for the training part. However, it is arduous to find the validity of the newly generated data. Existing research shows that noise and anomalies in the training datasets decreases the performance of the system significantly (Ding et al., 2017; Zavarez et al., 2017). Thus the solution to the Big Data problems are DL algorithms (Takam et al., 2020; Wilamowski et al., 2016).

In contradiction to more traditional machine learning, DL has the benefit of conceivably giving a result to the problems encountered in heavy quantities of data analysis and training problems in the input data. The implementation of conventional data analysis approaches makes the medical data enormous, intricate, and complex to interpret. Hence, DL offers a surpassing solution in the accumulation of valuable knowledge from such aggregate medical data (Jakhar & Hooda, 2018). To be specific, it assists in instinctively excerpting hidden data descriptions from high amounts of apart data. Due to this, it becomes an important mechanism for Big Data Analytics. The hierarchical training and removal of various levels of difficulty, data abstractions in DL present a specific grade of clarification for Big Data Analytics jobs, mainly for examining heavy data. DL algorithms are one such hopeful approach of research at higher levels of concept into the automatic extraction of difficult data descriptions (Hamilton et al., 2018). Such algorithms produce a stratified, learning through hierarchical structure and rendering data, where lower-level features represent the higher-level features. However, the pre-processing of data plays a crucial role in generating the valuable insights. If this phase is not handled properly, it can lead to the poor performance. To name a few challenges associated to pre-processing phase are Data augmentation. It is an approach to unnaturally amplify the datasets and increase the performance of the algorithm. It also reduces image over fitting.

Diabetic Retinopathy (DR) is one such application of BD where DL techniques are employed. DR is the dilemma pertaining to loss of vision that is globally taking place (Raman et al., 2019). Therefore, necessary steps should be taken for preventing this disease. DR produces four stages, stage 1 is Mild non-proliferative retinopathy (mild NPR), stage 2 is Moderate non-proliferative retinopathy (moderate NPR), stage 3 is Severe non-proliferative retinopathy (severe NPR) and the last stage is Proliferative DR (PDR). Stage 1 is a very early phase that concerns the event of micro aneurysms, in stage 2 the blood transportation ability is lost by the blood vessels due to the distortion and swelling in the blood vessels that nurture the retina, in stage 3 the new blood vessels are grown due to the impediment in blood flow through the various blood vessels and at last in stage 4, which is the high-level stage, where the increase traits discharged by the retina stimulate the generation of the extra blood vessels, rising simultaneously within the integument of the retina into a number of vitrified gel, satisfying the eye. Afresh-formed blood vessels are brittle thus more often result in bleed and leak. Furthermore, the associated defect tissue may shrink hence causing retinal detachment, driving to permanent eyesight loss (Dutta et al., 2018). Automated methods for DR examinations are essential to solving these problems (Ashikur et al., 2020). While DL for binary classification generally has achieved better results, multi-level classification results are less effective, particularly for initial stage disease. DL is a sub-field of machine learning (Raman et al., 2019). Machine learning problems generally fall under two categories. For the first category we can implement a supervised algorithm for finding a relation between the data input and output, and later use the relation for estimating the outputs from an input which does not have an output. Some common supervised machine learning algorithms are linear regression, and logistic regression (Ksiazek et al., 2021). Contrary, to this the second set of problems can be solved using an unsupervised algorithm. Here the algorithm learns on its own and in most cases pairs up similar data entries under one label. These algorithms are mostly used for obtaining clusters of similar kind inside a dataset. Some common unsupervised learning algorithms are KNN (k-nearest neighbor), and K-means clustering. The major difference between the two categories of algorithms is the implementation on a dataset. Supervised algorithms are always used for classification, and prediction problem statements that have a labeled data set. However, unsupervised algorithms are used for clustering problem statements that have an unlabeled dataset. We have considered the problem of binary classification in our paper using Pyspark i.e. Python using Apache Spark. Apache spark is a great demand platform independent and an open source libraries for handling huge data. The distributed nature of spark and parallelization of data are the boon to the spark. Spark with Spark MLib is provided for performing the machine learning on the datasets such as dimension reduction, regression, classification, clustering etc (Gupta et al., 2017). The researchers have focused on the various applications of machine learning and their effective usage but the work done in Spark MLib is still in its infancy stage. Apache spark provides a better solution to the problem of Big Data analytics through advanced computational infrastructures and generating the results in an efficient manner while considering the time factor also. The use of Data Frame (DF) in spark version 2.3.0 shows the image support for extracting the features through the DL techniques. We propose deep learning approach using the apache spark platform for identifying and classifying the retinopathy images and protecting the vision of the patient.

Our contributions in the present paper are as follows:

- The DL techniques and Big Data tool studied with respect to DR.
- The literature related to the various other dataset of DR are studied and investigated for their performances.
- The Indian Diabetic Retinopathy dataset (IDRID) is used for performing the experiment. The pre-processing of the IDRID dataset is done for removing the unnecessary part of the image (cropping image), resizing and converting grayscale.
- Finally, the DL framework is integrated in pyspark for performing the binary classification on the IDRID dataset using Logistic Regression (LR).

The paper is organized as follows section 2 throw light on DL Techniques in Big Data, section 3 related work is presented concerning the DR using DL, section 4 present the methodology adopted, section 5 exhibits the experimental work and ultimately, section 6 puts the conclusion.

2. Deep Learning Techniques and Big Data Tool for Diabetic Retinopathy

The metabolic disorder transpired due to high level of sugar in the blood known as Diabetic. This diabetic is responsible for causing the eye deficiencies know as DR (Dutta et al., 2018). DR plays a significant role in bad eye vision. Therefore, detection of RD on an early stage can be helpful in protecting the vision of a patient (Alyoubi et al., 2020; Bhimavarapu & Battineni, 2022; Pires et al., 2019; Shankar et al., 2020). Although there are many new techniques developed for classification of DR but still the research is in its infancy. Researchers have employed the DL techniques for effectively classifying the problem of retinal fundus image lesions. The various classes of the fundus images are micro-aneurysms, soft exudates, optic disk, hard exudates, etc. developed by the researchers for diagnosing the problem in the retinal lesions. A huge population is affected by Diabetes as stated by the report of the World Health Organisation (WHO) (Kumar et al., 2020). The increase in the rate of diabetes will be double to 4.4% by 2030. Research finding also shows that 347 million populations suffer from diabetes in 2014. Research finding also indicates that high diabetes can be the root cause of DR which can be cured if identified at an early stage (Luo et al., 2020). The International Diabetes Federation generated a report showing that 151 million adults are suffering from diabetes and they also prognosticate that by 2035, the persons suffering from diabetes will be 6000 lakhs, 7000 lakhs by 2045 and studies also found that the major age group affected from DR are between 20-79 years showing problem related to the low vision or loss of vision (Vocaturo & Zumpano, 2020). The AI is helpful in the field of medical as it can drastically reduce the time of doctors to diagnose the patient and also result in enhancing the medical treatment. DL algorithms select high-level, difficult concepts as data descriptions by a hierarchical training method. A key advantage of DL is the interpretation and learning of large amounts of apart data, making it a worthy piece of equipment for Big Data Analytics such that fresh data is mostly not labelled and categorized (Najafabadi et al., 2015).

Experimental inspects have proved that data descriptions gathered through piling up non-linear characteristic extractors like DL frequently generate more reliable results, such as classification modelling are improved, through the generative probabilistic designs the more reliable variety of samples are generated and the invariant section of data descriptions. DL has generated exceptional results in several machine learning utilization, including natural language processing, computer vision and speech recognition.

DL algorithms practice to automatically select a complex description of fresh data from the tremendous measure. DL algorithms are chiefly triggered by the area of unnatural or Artificial Intelligence (AI). Its working is based on the human brain and can do the decision-making by observing, analyzing and learning things for remarkably complex dilemmas. Literature about these complex difficulties has been a fundamental urge after DL algorithms. Harvesting the vital information from complex system using DL plays a vital role in providing solutions to Big Data (Ding et al., 2017). DL is practiced in diverse applications of Big Data for instance in text analytics, computer vision and speech recognition (Hamilton et al., 2018; Zhang et al., 2018).

According to the IDC (International Data Corporation) report the employment of Big Data tools increased by 39% (Gantz et al., 2012). To obtain better results with fast training, the researchers are inspecting Spark with DL. The spark has the advantages of processing the data in ram and its distributed parallel processing quality (Bharill et al., 2016; Hamilton et al., 2018; Mavridis & Karatza, 2017).

The all-in-one solution to the Big Data problem is merging the DL with Spark. The combination of DL with apache spark provides a good solution in reducing the training time as well as giving more accurate results than the DL alone (JayaLakshmi & KrishnaKishore, 2018; Lee et al., 2018; Sahlsten et al., 2019; Vocaturo & Zumpano, 2020). Hyperparameter tuning (Shankar et al., 2020) and large scale predictions are two main advantages of using DL with spark (Venkatesan et al., 2019).

DL and CNNs have performed lofty accuracy with respect to the classical methods for classification, pattern recognition etc. in previous years (Bisht & Gupta, 2020; Pires et al., 2019). Within the last few years, DL is being used widely in detection and classification of DR (Bisht & Gupta, 2021; Gao et al., 2019). It is capable of successfully learning the characteristics of raw data even when a large number of heterogeneous sources are integrated. Many DL-based techniques are available, including sparse coding, restricted Boltzmann Machines, CNNs and auto encoders. In comparison to machine learning methods, these methods perform better as the number of training data increases due to the increase in learned features. Furthermore, DL methods did not necessitate hand-crafted feature extraction (Mateen et al., 2020). In medical image analysis, CNNs are used in greater numbers than other methods, and are extremely effective (Takam et al., 2020). The CNN architecture consists of three principal layers (Gupta et al., 2019; Sun, 2019), convolution layers (CONV), pool layers and fully connected layers (FC). Depending on the author's vision, the number of layers, size and filters in the CNN vary. The structure of CNN architecture is made of a set of distinct layers. Different filters convolve an image in the CONV layers to extract the features. Typically, the pooling layer is applied after the CONV layer to reduce the dimension of feature maps. Average pooling and max pooling are typically adopted for pooling strategies. An FC layer describes the complete image of the input as a compact feature. Some of the CNN architectures available on the ImageNet datasets are InceptionV3, VGG16 and AlexNet and each offers different possible functionalities. Some studies use pre-trained models for transfer learning and some go the other route and build their CNNs from scratch. In general, the process of detecting and classifying DR images using DL begins with the collection of the dataset and the application of the necessary pre-processing to improve and enhance the images.

According to author (Aljunid & Manjaiah, 2021; Sisodia et al., 2017) pre-processing is the steps taken in organizing the data efficiently. The healthcare data is present in different formats and is suffering from the problem of noise, quality data, dimensionality etc. Pre-processing the data is an essential part of any data. The pre-processing of data helps in generating the quality data (Pitaloka et al., 2017). Data augmentation is mostly used pre-processing techniques with CNNs, it increases the size of dataset by applying many transformations to the actual datasets (Araujo et al., 2020; Tabik et al., 2017).

3. Related Work

In the field of DR, lots of work is being done by researchers, subject to the area of research and zone of interest. The past work in the realm of medical sciences and machine learning shows that researchers have suggested and executed various machine learning methods. However, implementing the DL algorithm to the Big Data tool is still lacking with respect to DR is perturbed. We are the first to implement and examine the DR Dataset through the DL skeleton using the Big Data tool spark and classifying them into the various stages.

The loss of vision is globally taking place and one of the factors behind it is DR (Wang et al., 2020). Therefore, necessary steps should be taken for preventing this disease. The author has used the IDRID dataset for performing the classification on it. They used the VGG19 model along with the LR, Support Vector Machine (SVM), Random Forest (RF) and Neural Network (NN) for the classification of the DR into five classes (Gupta et al., 2019).

The authors have developed a customized DL algorithm for detecting the DR automatically. This algorithm is capable of processing the fundus images. It can classify the images into relevant sections such as DR or no DR. Therefore, proceeding the genuine cases to the ophthalmologist (Gargeya & Leng, 2017).

The authors, Wu & Hu (2019) applied transfer learning using the pre-trained patterns InceptionV3, VGG19 and Resnet50 for the classification of DR images. They have classified the dataset into 5 DR classes (No, Mild, Moderate, Severe, Proliferative). However, class imbalanced was associated with the data. Preprocessing of the data was done by means of augmentation so that the images can be enhanced and the problem of classes can be balanced for generating results.

The authors have not only worked on the multiclass classification problem in DR but also overcome the problem of class imbalanced problem. A pipeline is formed, from image preprocessing to image classification. Guassian and Resnet18 are used in the model. However, other approaches of handling the imbalanced classes are not discussed (Sallam et al., 2020).

The author focused on image preprocessing leveraging various filter mechanisms to boost the image's features. Another method explored by them in this study was extracting statistical features from images. The information extracted from a resized image of 2000X2000, since high resolution allows for better exploration. A problem with image training can emerge as a result of the image resize factor; due to a complete lack of computational capability of systems, thus leading to feature loss factor. Fuzzy C-means clustering (FCM) was employed to determine the cluster levels pertaining to the training data that lead to improved training accuracy. The three methods were employed Feed Forward, CNN and Deep Neural Network (DNN) for classifying and predicting the labels. However, DNN has outperformed than the other models. It is 89.6% for the training and 86.3 % for testing (Dutta et al., 2018).

Table 1. Previous work done on DR.

References	DL Techniques for DR Dataset and Models	Contributions	Result	Advantage	Limitations
Gupta et al. (2019)	IDRID VGG19, (NN, SVM, RF, LR)	To overcome the problem of small dataset they have used the DL with Transfer Learning.	The accuracy of the four classifiers is above 90%.	Even on the small dataset the system has achieved good accuracy.	Worked on very less dataset
Gargeya & Leng (2017)	EyePACS, MESSIDOR 2, E-Ophtha, Customized DL	Design an automated tool based on DL for detection of DR. It can be run on a simple computer	AUC = 0.97	Preprocessing is also done.	-
Wu & Hu (2019)	From Kaggle VGG19, Resnet50, InceptionV3	Managed the intricacy of class imbalance	InceptionV3 performance was best.	Technique applied to manage intricacy of class imbalanced	Results were not good
Sallam et al. (2020)	From Kaggle Preprocessing and Resnet18	Removed black boundary, resized the images, size of dataset reduced from 35GB to 1GB, contrast enhancement, augmentation of data is done in Preprocessing phase	Accuracy = 69.4%	Handled the problem of class imbalanced	Trained on Resnet 18 only
Qummar et al. (2019)	Five Deep CNN Models (Resnet50, InceptionV3, Xception, Dense121, Dense169)	Preprocessing of the dataset is done. Imbalanced classes are handled Different models are used with changes in the last layer.	All the classes are separately measured.	The organization for distinct grades of DR	The data is passed to the model separately.
Dutta et al. (2018)	From Kaggle Feed Forward –NN, DNN, CNN	Preprocessing is focused and implemented through Fuzzy C-means clustering Finally comparison of all the models are done	DNN Accuracy is 86.3%	It can classify the various levels of severity accurately.	For higher accuracy, more data can be trained on GPU system.
Sarki et al. (2021)	Messidor, Messidor-2, DRISHTI-GS, and Retinal Dataset from GitHub	Image Enhancement, Segmentation and Augmentation	Modified CNN accuracy is 100% for Glaucoma, for DR 93.33 and for DME (Diabetic Macular Edema) 91.43	The mild symptoms can be classified very accurately.	Only limited-to-moderate data set sizes were employed
Saranya et al. (2022)	From Kaggle DRIVE and STARE Modified CNN is used with VGG-16 architecture	Preprocessing, Vessel Segmentation, Removal of extra features	Achieved accuracy 96%, 95% in both the datasets	Detection of neovascularization	Evaluated only for Non-Proliferative DR and Proliferative DR

Fundus Images of 35126 are taken from the kaggle dataset, dimensions of 3888 × 2951. The measure for the five classes is based on the severity of DR. The images are perfectly imbalanced. The author has worked on the problem of class imbalance to improve the classification of the system. The preprocessing step consists of resizing of images to 786X512, performed up-sampling (done through augmentation) and down-sampling (some of the instances are discarded). Finally, the preprocessed images are passed through the pre-trained models like Dense121, InceptionV3, Resnet50, Xception, and Dense169. The last layers of the models are modified through 5X1 dimension. The author has not only considered accuracy to monitor the achievement of the design but the other measures such as precision, recall etc. are also taken into consideration (Qummar et al., 2019).

The authors worked on the early stage classification and identification of mild and normal images of diabetic eye diseases. They employed the traditional preprocessing techniques on the fundus images like Image enhancement, fine tune, data balance and feature enhancement. Various datasets are applied to check the working of the modified CNN. The maximum accuracy was 100% obtained using the proposed model with the traditional image processing approaches (Sarki et al., 2021).

The authors have done the work on the classification by No Proliferative DR and Proliferative DR through the blood vessel segmentation and neovascularization. In the preprocessing techniques, custom features were extracted and later on segmentation is used for identifying the blood vessels in the dataset, which resulted in maximum 96% accuracy (Saranya et al., 2022). Table 1 shows the related work done on DR.

4. Methodology

In this paper the DL is integrated with pyspark. Pyspark is a Big Data tool for processing the IDRID dataset. The standalone machines are not capable of processing Big Data due to its massive size. Therefore, due to the restrictions of the resources, we are using the databricks community edition for performing our experiment. It is giving us the platform that supports the DL library. The machine used for this experiment is 16GB RAM. It is providing us with 15.3GB RAM, 2 Cores and 1DBU. It supports the scala as well as python. The use of the python language for coding is known as Pyspark (spark using python language). The spark version spark 2.4.3 is used in this experiment. Maven source is used to install the libraries of DL. Figure1 shows our proposed framework for the DL pipeline with DR. We used the Databricks Environment for our analysis. It's an analytics platform built on Apache Spark. Databricks is a joint ecosystem that allows customers to altogether their computational methods and train machine learning designs through their entire life process (Benbrahim et al., 2020). We built a cluster in this platform to run our model as a series of commands. The scalable DL methods are adapted in the area of medical imaging, precisely classifying the presence DR images and no DR images from IDRID dataset consisting of 516 images (training =413 images, testing = 103). The 0 and 1 stand for the classification of the images, where 1 stands for no DR and 0 stands for DR of any category like mild, moderate etc. The solution is capable of automatically classifying and can process thousands of images instantly. It likely mitigates the necessity for the resource-intensive old-fashioned examination. However, due to the restrictions of the resources a small size of the dataset is taken for the experimental purpose. Figure 1 shows the suggested skeleton for the analysis of no DR/DR with Pyspark.

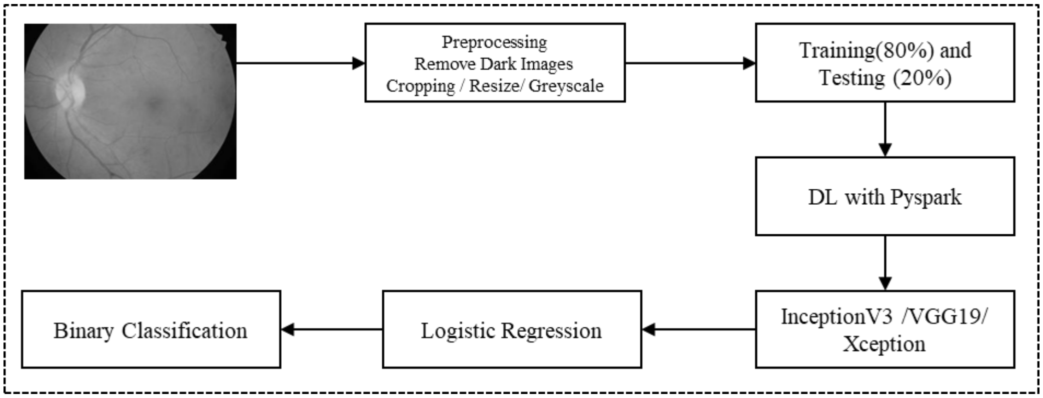

Figure 1. Proposed framework for the classification of no DR/DR with Pyspark.

4.1 Preprocessing

In the preprocessing phase the images are cleaned for generating better results. First of all the dark images are removed from the dataset as it cannot provide the sufficient information to the machine for generating the results. Then cropped the images and removed the unwanted portion of the image so that the unnecessary processing of the machine can be reduced and our results can be enhanced. For this we have used the *PILLOW* library in python language using the Jupiter notebook in anaconda environment. Moreover, the IDRID dataset suffers from the problem of class imbalanced. In order to overcome the problem of class imbalanced the augmentation of the images can be done. Through the augmentation techniques (Gargeya & Leng, 2017) the size of various classes can be brought to same level. The images can be horizontally flipped, vertically flipped and rotated to various degrees. However, it is not required in our case because we are doing the binary classification and due to the limitation of resources very less records are taken into consideration. The other preprocessing techniques include contrast enhancement (Sarki et al., 2021)and image segmentation (Das et al., 2021). These techniques are not used in our preprocessing phase and thus out of the scope of this paper.

4.2 Classification

The classification of the dataset is done through the LR. As we have already discussed earlier, a LR classifier has been used in our research. Here we will take a closer look at how a LR classifier works also the computations that occur behind the scenes. LR has gained a massive fan base of its own in the last two decades on account of its strong affinity for calculating the probabilities of certain events in a given problem statement. The main idea behind a LR classifier is to find the probabilities of an event's happening or vice versa. In an ideal world, we can think of the problem of tossing a coin for obtaining a head or tails. Here the probability is 0.5 for both events as long as we are taking a fair coin for the toss.

$$P(Y = head|X) = P(Y = tail|X) = 0.5 \tag{1}$$

Unlike the above illustration, our LR classifier will always give us probabilities of an event's happening or vice versa between 0 and 1. It is imperative to note here that the sum of probabilities of an event happening and vice versa will always give us 1.

$$P(X = happening|Y) + P(X = not\ happening|Y) = 1 \tag{2}$$

Similar to a linear regression classifier the first operation performed is the multiplication of respective weights (w) with the variables (x) from dataset. A vital point to note here that x is a vector of shape nx1 where n is the number of entries in our dataset. Upon multiplying the above two, we use a sigmoid function that converts each value in a range of 0 and 1.

$$S = \sum d_i = w_i x_i \tag{3}$$

In next step we pass the output S as z as sigmoid function, the main formulae for converting a value in a range of 0 and 1, which is utterly useful when dealing with probabilities. Using the sigmoid function is as follows.

$$\emptyset(z) = \frac{1}{1+ e^{-z}} \tag{4}$$

Figure 2 shows the curve for sigmoid function and helps us to understand how this function clipping the value between 0 and 1. Furthermore, where *e* stands for exponent and *z* is the variable for which sigmoid is to be determined.

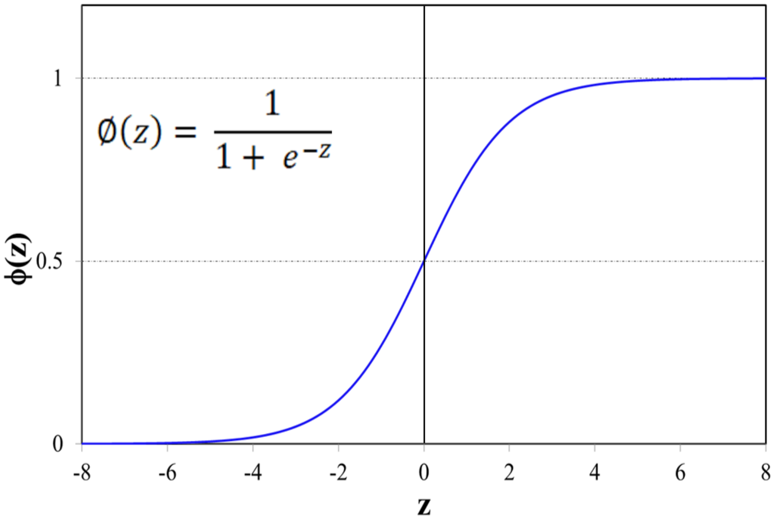

Figure 2. Sigmoid function of LR.

Where t is a function consisting of our features $(x_j)$and their corresponding weights (w_j) in a linear form shown below.

$$t = x_0 + x_1w_1 + \cdots + x_kw_k \tag{5}$$

At last the LR classifier follows the "Maximum Likelihood Principle" by applying simple formulae stated below.

$$argmax_y \prod_i P(X_i = x_i|Y = y)P(Y = y) \tag{6}$$

Where output y is between (0,1), $P(y_i|x_i)$ is the probability which is equal to $\emptyset(z)$ from equation (4)

For the coin flip problem that we noticed, equation (6) will look as follows

$$argmax_{y\in\{+1,-1\}} \sum_i \log P(x_i \mid y) + \log P(y) \tag{7}$$

At last, we will take a look at how everything works. For the label '1', we expect our classifier to predict a value closer to 1 i.e., greater than 0.5. And for the label '0', we expect our classifier to predict a value closer to 0 i.e., smaller than 0.5.

Some of the measures considered are discussed below used as evaluation criteria to quantitatively assess the proposed model.

Accuracy

The correctness or accuracy can be revealed in terms of positive and negative classes.

$$Accuracy = \frac{(TP+ TN)}{(TP+TN+FP+FN)} \tag{8}$$

where, TP (True Positives) is the amount of rightly labeled instances of the class under consideration. TN (True Negatives) is the amount of rightly labeled cases of the residue of the classes, FP (False Positives) is the amount of not rightly classified cases of the residue and FN (False Negatives) is the amount of not rightly classified cases of the class under consideration.

Recall/Sensitivity

It is the division of a model correctly distinguishing True Positives.

$$\text{Recall} = \frac{TP}{(TP+FN)} \qquad (9)$$

Precision

Precision is the ratio between the True Positives and all the Positives.

$$Precision = \frac{TP}{(TP+FP)} \qquad (10)$$

F1 – Score

F1 Score strength a better measure to use when we want to endeavor stability between Precision and Recall, and uneven distribution of class is there. The value ranges from 0 to 1, where the best value is 1 and the worst is 0.

$$F1 - Score = 2\ X\ (Precision\ X\ Recall)/(Precision + Recall) \qquad (11)$$

ROC (Receiver Operating Curve)

It is used for the problems of binary division. The multi-classification problem can also be handled through it. It is a likelihood arch that sketches the True Positive Rate (TPR) against False Positive Rate (FPR) at various outset values.

AUC Score

It is the measure of the strength of a classifier to discriminate among classes and is employed as a review of the ROC curve. The higher the AUC rate indicates the more reliable the model and so on.

5. Experiment

This section discusses the various steps conducted to perform the experiment starting with the dataset to the pipeline adopted for the processing and binary classification of the dataset into normal and DR.

5.1 Dataset

The dataset is downloaded from the kaggle.com website for the Indian DR Images and it is easily available. The original size is of image is 40 KB with dimensions 4288 X 2848 (width, height), and 300 dpi. Actual images are 516 (training = 413, testing = 103) and suffering from the problem of class imbalanced. The experiment is conducted on 192 images dataset chosen as 80% of training and 20% of testing dataset. Augmentation is not required in our condition as we are doing binary classification. The dark images are difficult to visualize therefore, the lighting condition effects from the dataset are discarded. After that we have cropped the unwanted part of the images as it can increase or burden up the processing time of the machine. In this experimental setup, the size of the image is reduced to 256 X 256 with 96 dpi. The high resolution or colour images give better results no doubt but it increases the computation time of the system or burden up the processing time which again is a challenge for the standalone system therefore the images are converted into grayscale for the fast processing of the system. Figure 3 shows no DR image and Figure 4 shows the various stages of retinopathy (Stage 1-4). Although, we are using the Databricks Community Edition which giving us support for the DL pipeline but the challenge associated with it is the computational resources. It is giving us the image of the cluster where the number of systems is working to achieve a common task. However, it is processing on our system only and thus not capable of processing the massive data with the free edition.

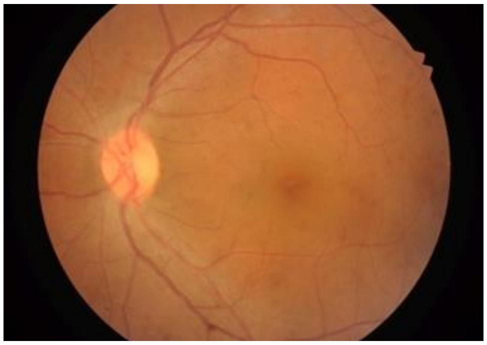

Figure 3. No retinopathy.

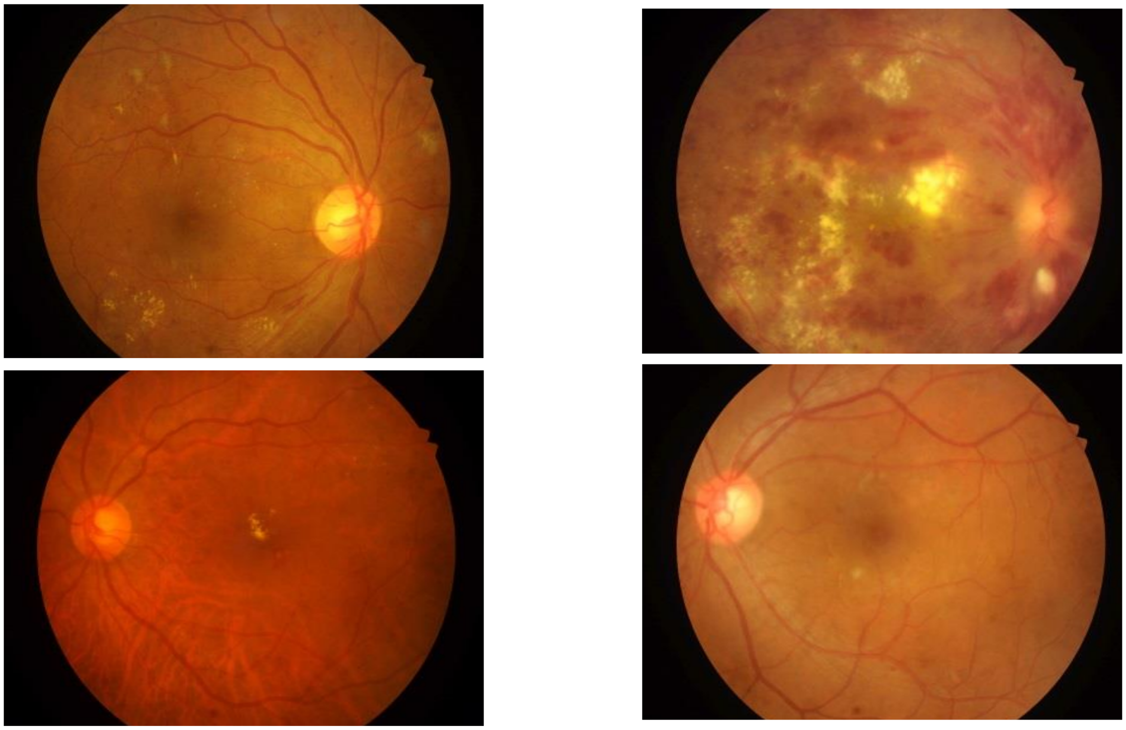

Figure 4. Different stages of retinopathy.

In this paper, we have applied DL Pipelines on Apache Spark, which allows quick transfer learning. The pre-trained Convolutional Neural Network model InceptionV3, Xception and VGG19 are used with the LR (LR) for the classification of DR images. The DL library is a publicly available outline for DL. Apache spark provides a better solution to the problem of Big Data analytics through advanced computational infrastructures and generating the results in an efficient manner while considering the time factor also (Assefi et al., 2017). The use of Data Frame (DF) in spark version 2.3.0 shows the image support for extracting the features through the DL techniques (Téllez-Velázquez & Cruz-Barbosa, 2019). We proposed DL approach using the apache spark platform for identifying and classifying the retinopathy images and protecting the vision of the patient.

In the Databricks Runtime environment, the DL pipelining can be established through the Maven library. To examine the outcome we have used the LR, LR is a statistical technique employed on machine learning to examine the independent features that mark an outcome. The Xception, VGG19 and InceptionV3 are the models, which are pre-trained on the Imagenet chosen for the processing of the images and are used for processing of Big Data images. Finally, we applied LR to investigate the independent features that characterize an outcome. For this experiment we have opted two types of images i.e. No DR and DR.

5.2 Image Cropping and Resizing

Removing the extra part of the images is known as cropping. We have used the PILLOW library from the python using the jupyter notebook through the anaconda environment. The original size of the images is 397 KB and the size on the disk is 400 KB. The dimensions are 4288 X 2848 (width X height) with the horizontal and vertical resolution of 300 dpi. After cropping and converting the images the dimensions are reduced to 256 X 256. The cropping and resizing of the image is shown in Figure 5.

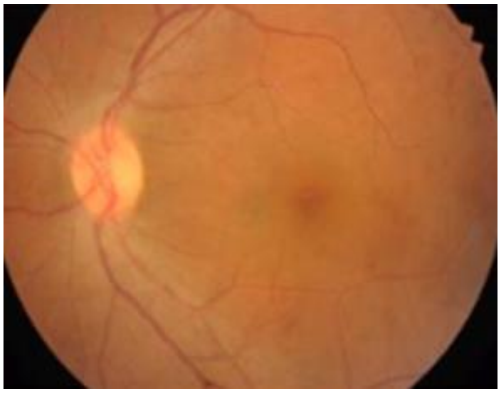

Figure 5. Cropping and resizing of image.

Then the images are converted into the grayscale. Figure 6 shows the image in grayscale. The collected processed datasets were split as two sets that are training and testing. We considered 80% of the dataset as training and the rest 20% as testing. After the preprocessing phase, the images are loaded to the proposed model using Databricks File System (DBFS). It's a distributed file device that let us store data for queries within Databricks and make it available across clusters. The data remains in storage after the termination of the cluster and can be accessed without any permit. After loading the images, the images are passed to the Spark Data Frame. The images are stored in the File System of Databricks. Before passing the dataset to the DL pipeline, the images are labeled as one for normal and label zero for DR.

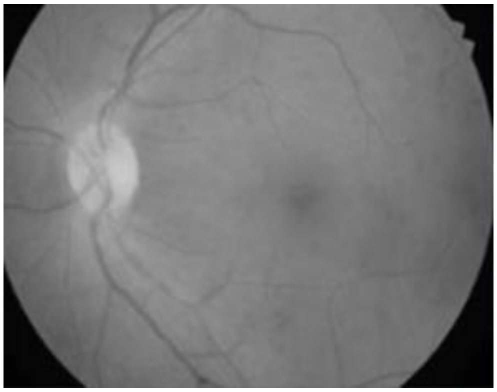

Figure 6. Image converted into grayscale.

5.3 Results and Discussion

A process established on deep transfer learning employing CNN based InceptionV3, Xception and VGG19 were proposed for the detection of no DR and DR via the Apache Spark. In this paper, we have performed a comparative study on the above three models. The empirical results have shown that among all the three models InceptionV3 has outperformed and shown the highest accuracy of 95% then Xception with accuracy of 92.5% and at last VGG19 with accuracy of 90%. The other measures are also tested such as F1-Score, Accuracy and ROC AUC score and are shown in Table 2.

Table 2. Overview of the performance measures of the models.

Model Name	**Accuracy**	**F1-Score**	**ROC AUC Score**
InceptionV3	95.00	95.00	94.98
Xception	92.50	92.50	92.60
VGG19	89.94	90.00	90.47

From the overview of the Table 2, InceptionV3 is best classifying the binary classes. The accuracy, F1-score and ROC AUC score are also very close to each other, which mean that our classes are correctly classified. Although we achieved an AUC of 0.94 on IDRID dataset in InceptionV3, our algorithm strived to differentiate between normal and initial phases of DR. The missing cases are fine microaneurysms. Microaneurysm small appearance makes it difficult to identify even for human graders and thus poses a significant limitation in early and accurate detection of the DR in the coming systems.

We assume that blending of manual features, targeting particular aspects of microaneurysms for identification of mild DR, with the sturdy potential of DL systems to describe accurately all other grades of DR without trouble from the brightness and capturing artifacts, will yield more robust results in future early DR detection studies. The confusion matrix can give better idea about the correctly classified no DR/DR. Figure 7 shows the accuracy of the three models. According to the displayed results, InceptionV3 has achieved 95 percent accuracy, Xception has achieved 92 percent accuracy, and VGG19 has achieved 89 percent accuracy. A Confusion matrix is a matrix of NXN used for assessing the performance of a classification model. Here N, indicates the target class numbers and showing the predicted outcomes whether they pertain to the actual class. It gives an entire view of the types of errors made by the model and also shows the performance of the classification model (Kotiyal et al., 2014).

The confusion matrix for the IDRID dataset is shown for the different models. In the confusion matrix, the horizontal row shows the predicted values and the vertical row shows the actual values. We have taken 192 images for the DR and no DR. We have only compared the performance of the models based on no DR and DR. The literature surveys conducted in this paper are not based on the Apache Spark framework. We are the first to perform the DR problem in this framework using the DL pipeline. Therefore, we performed comparative measures among the models for solving the problem of binary classification considering the Accuracy, F1-Score and ROC AUC score. Figure 8 shows the confusion matrix VGG19, Figure 9 shows for InceptionV3and Figure 10 shows for Xception. Such that Figure 8 shows that out of the 40 total images, 17+19 is correctly identified (DR or No DR) and 4 are wrongly classified, allowing us to draw this conclusion.

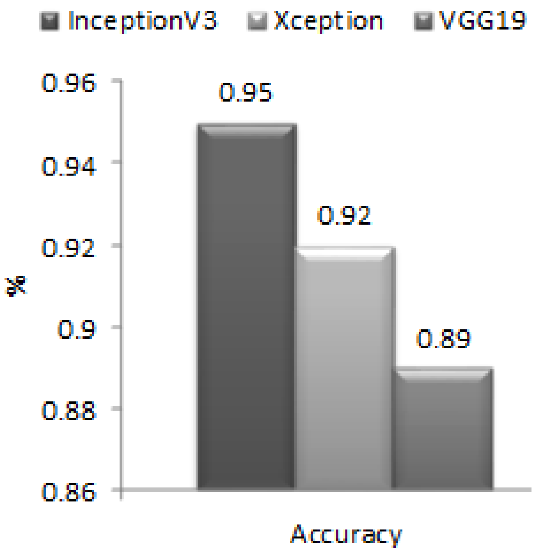

Figure 7. Accuracy of the Models.

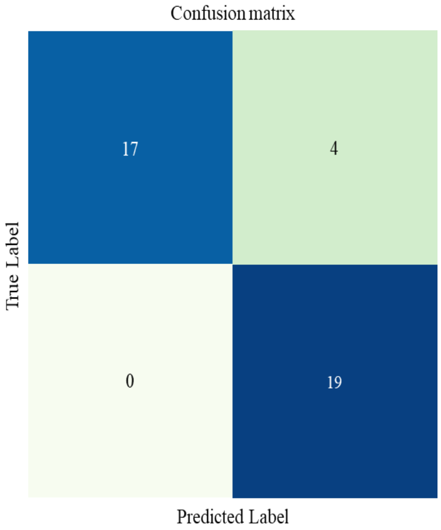

Figure 8. Confusion Matrix for VGG19.

Similarly, it is implemented for Figure 9 and Figure 10.

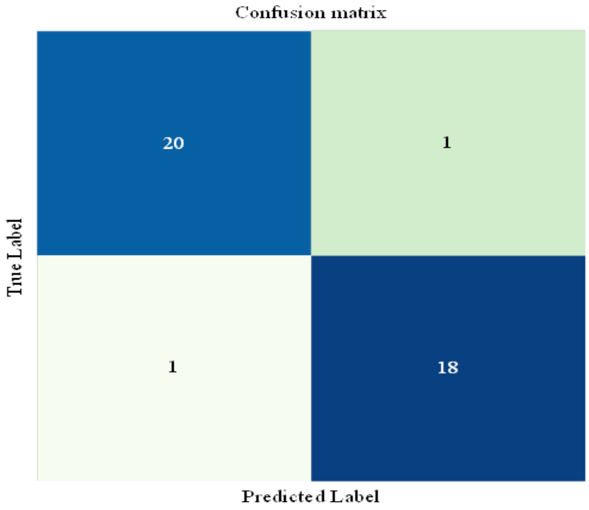

Figure 9. Confusion Matrix for InceptionV3.

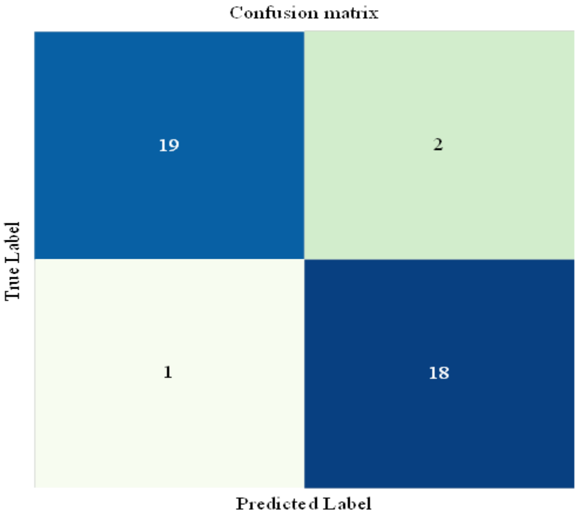

Figure 10. Confusion Matrix for Xception.

6. Conclusion

DR plays a significant role in bad eye vision. Therefore, detection of DR on an early stage can be helpful in protecting the vision of a patient. A huge population is affected by Diabetes as stated by the report of the World Health Organisation (WHO). The AI is helpful in the field of medical as it can drastically reduce the time of doctors to diagnose the patient and also result in enhancing the medical treatment. The manual examination of DR is a time-consuming and challenging process and needs deeply trained experts whereas; DL algorithms are one such hopeful approach of research at higher levels of concept into the automatic extraction of difficult data descriptions. Such algorithms produce a stratified, learning through hierarchical structure and rendering data. This research proposed a novel pipeline for the binary classification of DR by employing the techniques of DL. Before passing the IDRID dataset to the pipeline the preprocessing of the dataset is done. The InceptionV3, Xception and VGG19 are examined here for their performance. The parameters for the performance taken into consideration are accuracy, f1-score and AUC score. It has been observed that InceptionV3 best performed for the problem of DR images. It has given the accuracy of 95% whereas the accuracy of Xception is 92.50% and accuracy of VGG19 89.94%. However, the increase in the dataset will increase the accuracy even better than this. At last, our model can be used for a massive number of datasets. Due to the limitation of resources, we have applied to the small dataset. Implementation of our proposed pipeline for the DR through computer-aided techniques could diminish the rate of vision loss and thus results in clinical management. We are the first to implement the problem of DR to pyspark employing the DL techniques. However, in the past literature, it is employed using only the DL techniques. In future work, the model can use by employing different classifiers to classify the problem and the preprocessing phase can be enhanced for better results.

Conflict of Interest

The authors confirm that this article contents have no conflict of interest.

Acknowledgement

No funding is provided by the institution or any grant is taken for carrying out this research.

References

Aljunid, M.F., & Manjaiah, D.H. (2021). Data management, analytics and innovation. In *Proceedings of ICDMAI* (Vol. 70). http://link.springer.com/10.1007/978-981-13-1402-5%0Ahttps://link.springer.com/10.1007/978-981-16-2934-1.

Alyoubi, W.L., Shalash, W.M., & Abulkhair, M.F. (2020). Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, *20*, 100377. https://doi.org/10.1016/j.imu.2020.100377.

Araujo, T., Aresta, G., Mendonca, L., Penas, S., Maia, C., Carneiro, A., Mendonca, A.M., & Campilho, A. (2020). Data augmentation for improving proliferative diabetic retinopathy detection in eye fundus images. *IEEE Access*, *8*, 182462-182474. https://doi.org/10.1109/access.2020.3028960.

Ashikur, M., Arifur, M., & Ahmed, J. (2020). Automated detection of diabetic retinopathy using deep residual learning. *International Journal of Computer Applications*, *177*(42), 25-32. https://doi.org/10.5120/ijca2020919927.

Assefi, M., Behravesh, E., Liu, G., & Tafti, A.P. (2017). Big data machine learning using apache spark MLlib. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 3492-3498. https://doi.org/10.1109/BigData.2017.8258338.

Benbrahim, H., Hachimi, H., & Amine, A. (2020). Deep transfer learning with apache spark to detect COVID-19 in chest X-ray images. *Romanian Journal of Information Science and Technology*, *23*(April), S117-S129.

Bharill, N., Tiwari, A., & Malviya, A. (2016). Fuzzy based scalable clustering algorithms for handling big data using apache spark. *IEEE Transactions on Big Data*, *2*(4), 339-352. https://doi.org/10.1109/tbdata.2016.2622288.

Bhimavarapu, U., & Battineni, G. (2022). Automatic microaneurysms detection for early diagnosis of diabetic retinopathy using improved discrete particle swarm optimization. *Journal of Personalized Medicine*, *12*(2), 317. https://doi.org/10.3390/jpm12020317.

Bisht, M., & Gupta, R. (2020). Multiclass recognition of offline handwritten devanagari characters using CNN. *International Journal of Mathematical, Engineering and Management Sciences*, *5*(6), 1429-1439. https://doi.org/10.33889/IJMEMS.2020.5.6.106.

Bisht, M., & Gupta, R. (2021). Fine-tuned pre-trained model for script recognition. *International Journal of Mathematical, Engineering and Management Sciences*, *6*(5), 1237-1314. https://doi.org/10.33889/IJMEMS.2021.6.5.078.

Das, S., Kharbanda, K., Suchetha, M., Raman, R., & Edwin Dhas, D. (2021). Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy. *Biomedical Signal Processing and Control*, *68*(March), 102600. https://doi.org/10.1016/j.bspc.2021.102600.

Ding, J., Li, X., & Gudivada, V.N. (2017). Augmentation and evaluation of training data for deep learning. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-January*, 2603-2611. https://doi.org/10.1109/BigData.2017.8258220.

Dutta, S., Manideep, B.C.S., Basha, S.M., Caytiles, R.D., & Iyengar, N.C.S.N. (2018). Classification of diabetic retinopathy images by using deep learning models. *International Journal of Grid and Distributed Computing*, *11*(1), 89-106. https://doi.org/10.14257/ijgdc.2018.11.1.09.

Gantz, B.J., Reinsel, D., & Shadows, B.D. (2012). Big data , bigger digital shadow s , and biggest grow th in the far east executive summary: a universe of opportunities and challenges. *Idc*, *2007*(December 2012), 1-16.

Gao, J., Leung, C., & Miao, C. (2019). Diabetic retinopathy classification using an efficient convolutional neural network. *Proceedings - 2019 IEEE International Conference on Agents, ICA 2019*, 80-85. https://doi.org/10.1109/AGENTS.2019.8929191.

Gargeya, R., & Leng, T. (2017). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, *124*(7), 962-969. https://doi.org/10.1016/j.ophtha.2017.02.008.

Gopalakrishnan, K., Khaitan, S.K., Choudhary, A., & Agrawal, A. (2017). Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. *Construction and Building Materials*, *157*, 322-330. https://doi.org/10.1016/j.conbuildmat.2017.09.110.

Gupta, S., Panwar, A., Goel, S., Mittal, A., Nijhawan, R., & Singh, A.K. (2019). Classification of lesions in retinal fundus images for diabetic retinopathy using transfer learning. *Proceedings - 2019 International Conference on Information Technology, ICIT 2019*, 342-347. https://doi.org/10.1109/ICIT48102.2019.00067.

Gupta Thakur, H.K., Shrivastava, R., Kumar, P., & Nag, S. (2017). A big data analysis framework using apache spark and deep learning. *IEEE International Conference on Data Mining Workshops, ICDMW, 2017-Novem*(1), 9-16. https://doi.org/10.1109/ICDMW.2017.9.

Hamilton, M., Raghunathan, S., Annavajhala, A., Kirsanov, D., De Leon, E., Barzilay, E., Matiach, I., Davison, J., Busch, M., Oprescu, M., Sur, R., Astala, R., Wen, T., & Park, C.Y. (2018). Flexible and scalable deep learning with MML spark. *ArXiv*, *1*, 1-12.

Jakhar, K., & Hooda, N. (2018). Big data deep learning framework using keras: A case study of pneumonia prediction. *2018 4th International Conference on Computing Communication and Automation*, *ICCCA 2018*, 1-5. https://doi.org/10.1109/CCAA.2018.8777571.

Jan, B., Farman, H., Khan, M., Imran, M., Islam, I.U., Ahmad, A., Ali, S., & Jeon, G. (2019). Deep learning in big data Analytics: A comparative study. *Computers and Electrical Engineering*, *75*, 275-287. https://doi.org/10.1016/j.compeleceng.2017.12.009.

JayaLakshmi, A.N.M., & KrishnaKishore, K.V. (2018). Performance evaluation of DNN with other machine learning techniques in a cluster using Apache Spark and MLlib. *Journal of King Saud University - Computer and Information Sciences*, 1-9. https://doi.org/10.1016/j.jksuci.2018.09.022.

Kotiyal, B., Kumar, A., Pant, B., & Goudar, R.H. (2014). Classification technique for improving user access on web log data. *Advances in Intelligent Systems and Computing*, *243*, 1089-1097. https://doi.org/10.1007/978-81-322-1665-0.

Kotiyal, B., Kumar, A., Pant, B., & Goudar, R.H. (2013). Big data: Mining of log file through Hadoop. *International Conference on Human Computer Interactions, ICHCI 2013*, 1-7. https://doi.org/10.1109/ICHCI-IEEE.2013.6887797.

Ksiazek, W., Gandor, M., & Plawiak, P. (2021). Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma. *Computers in Biology and Medicine*, *134*, 1-13. https://doi.org/10.1016/j.compbiomed.2021.104431.

Kumar, G., Chatterjee, S.K., & Chattopadhyay, C. (2020). Drdnet: diagnosis of diabetic retinopathy using capsule network (Workshop Paper). *Proceedings - 2020 IEEE 6th International Conference on Multimedia Big Data, BigMM 2020*, 379-385. https://doi.org/10.1109/BigMM50055.2020.00065.

Lee, S., Kim, H., Park, J., Jang, J., Jeong, C.S., & Yoon, S. (2018). TensorLightning: A traffic-efficient distributed deep learning on commodity spark clusters. *IEEE Access*, *6*, 27671-27680. https://doi.org/10.1109/ACCESS.2018.2842103.

Luo, Y., Pan, J., Fan, S., Du, Z., & Zhang, G. (2020). Retinal image classification by self-supervised fuzzy clustering network. *IEEE Access*, *8*, 92352-92362. https://doi.org/10.1109/ACCESS.2020.2994047.

Mateen, M., Wen, J., Hassan, M., Nasrullah, N., Sun, S., & Hayat, S. (2020). Automatic detection of diabetic retinopathy: a review on datasets, methods and evaluation metrics. *IEEE Access*, *8*, 48784-48811. https://doi.org/10.1109/ACCESS.2020.2980055.

Mavridis, I., & Karatza, H. (2017). Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark. *Journal of Systems and Software*, *125*, 133-151. https://doi.org/10.1016/j.jss.2016.11.037.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1-21. https://doi.org/10.1186/s40537-014-0007-7.

Pires, R., Avila, S., Wainer, J., Valle, E., Abramoff, M.D., & Rocha, A. (2019). A data-driven approach to referable diabetic retinopathy detection. *Artificial Intelligence in Medicine*, *96*(March), 93-106. https://doi.org/10.1016/j.artmed.2019.03.009.

Pitaloka, D.A., Wulandari, A., Basaruddin, T., & Liliana, D.Y. (2017). Enhancing CNN with preprocessing stage in automatic emotion recognition. *Procedia Computer Science*, *116*, 523-529. https://doi.org/10.1016/j.procs.2017.10.038.

Qummar, S., Khan, F.G., Shah, S., Khan, A., Shamshirband, S., Rehman, Z.U., Khan, I.A., & Jadoon, W. (2019). A deep learning ensemble approach for diabetic retinopathy detection. *IEEE Access*, *7*, 150530-150539. https://doi.org/10.1109/ACCESS.2019.2947484.

Raman, R., Srinivasan, S., Virmani, S., Sivaprasad, S., Rao, C., & Rajalakshmi, R. (2019). Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. *Eye (Basingstoke)*, *33*(1), 97-109. https://doi.org/10.1038/s41433-018-0269-y.

Rehman, M.H., Liew, C.S., Abbas, A., Jayaraman, P.P., Wah, T.Y., & Khan, S.U. (2016). Big data reduction methods: a survey. *Data Science and Engineering*, *1*(4), 265-284. https://doi.org/10.1007/s41019-016-0022-0.

Sahlsten, J., Jaskari, J., Kivinen, J., Turunen, L., Jaanio, E., Hietala, K., & Kaski, K. (2019). Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Scientific Reports*, *9*(1), 1-11. https://doi.org/10.1038/s41598-019-47181-w.

Sallam, M.S., Asnawi, A.L., & Olanrewaju, R.F. (2020). Diabetic retinopathy grading using resnet convolutional neural network. *2020 IEEE Conference on Big Data and Analytics, ICBDA 2020*, 73-78. https://doi.org/10.1109/ICBDA50157.2020.9289822.

Saranya, P., Prabakaran, S., Kumar, R., & Das, E. (2022). Blood vessel segmentation in retinal fundus images for proliferative diabetic retinopathy screening using deep learning. *Visual Computer*, *38*(3), 977-992. https://doi.org/10.1007/s00371-021-02062-0.

Sarki, R., Ahmed, K., Wang, H., Zhang, Y., Ma, J., & Wang, K. (2021). Image preprocessing in classification and identification of diabetic eye diseases. *Data Science and Engineering*, *6*(4), 455-471. https://doi.org/10.1007/s41019-021-00167-z.

Shankar, K., Zhang, Y., Liu, Y., Wu, L., & Chen, C.H. (2020). Hyperparameter tuning deep learning for diabetic retinopathy fundus image classification. *IEEE Access*, *8*, 118164-118173. https://doi.org/10.1109/ACCESS.2020.3005152.

Singh, N., Singh, D.P., & Pant, B. (2019). ACOCA: Ant colony optimization based clustering algorithm for big data preprocessing. *International Journal of Mathematical, Engineering and Management Sciences*, *4*(5), 1239-1250. https://doi.org/10.33889/IJMEMS.2019.4.5-098.

Sisodia, D.S., Nair, S., & Khobragade, P. (2017). Diabetic retinal fundus images: Preprocessing and feature extraction for early detection of Diabetic Retinopathy. *Biomedical and Pharmacology Journal*, *10*(2), 615-626. https://doi.org/10.13005/bpj/1148.

Sun, Y. (2019). The neural network of one-dimensional convolution-an example of the diagnosis of diabetic retinopathy. *IEEE Access*, *7*, 69657-69666. https://doi.org/10.1109/ACCESS.2019.2916922.

Tabik, S., Peralta, D., Herrera-Poyatos, A., & Herrera, F. (2017). A snapshot of image Pre-Processing for convolutional neural networks: Case study of MNIST. *International Journal of Computational Intelligence Systems*, *10*(1), 555-568. https://doi.org/10.2991/ijcis.2017.10.1.38.

Takam, C.A., Samba, O., Tchagna Kouanou, A., & Tchiotsop, D. (2020). Spark Architecture for deep learning-based dose optimization in medical imaging. *Informatics in Medicine Unlocked*, *19*, 1-13. https://doi.org/10.1016/j.imu.2020.100335.

Téllez-Velázquez, A., & Cruz-Barbosa, R. (2019). A Spark image processing toolkit. *Concurrency Computation*, *31*(17), 1-11. https://doi.org/10.1002/cpe.5283.

Venkatesan, N.J., Nam, C.S., & Shin, D.R. (2019). Deep learning frameworks on apache spark: a review. *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)*, *36*(2), 164-177. https://doi.org/10.1080/02564602.2018.1440975.

Vocaturo, E., & Zumpano, E. (2020). The contribution of AI in the detection of the Diabetic Retinopathy. *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, 1516-1519. https://doi.org/10.1109/BIBM49941.2020.9313541.

Wang, S., Wang, X., Hu, Y., Shen, Y., Yang, Z., Gan, M., & Lei, B. (2020). Diabetic retinopathy diagnosis using multichannel generative adversarial network with semisupervision. *IEEE Transactions on Automation Science and Engineering*, 1-12. https://doi.org/10.1109/TASE.2020.2981637.

Wilamowski, B.M., Wu, B., & Korniak, J. (2016). Big data and deep learning. *INES 2016 - 20th Jubilee IEEE International Conference on Intelligent Engineering Systems, Proceedings*, *2015*, 11-16. https://doi.org/10.1109/INES.2016.7555103.

Wu, Y., & Hu, Z. (2019). Recognition of diabetic retinopathy based on transfer learning. *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics*, *ICCCBDA 2019*, 398-401. https://doi.org/10.1109/ICCCBDA.2019.8725801.

Zavarez, M.V., Berriel, R.F., & Oliveira-Santos, T. (2017). Cross-database facial expression recognition based on fine-tuned deep convolutional network. *Proceedings - 30th Conference on Graphics, Patterns and Images, SIBGRAPI 2017*, *October*, 405-412. https://doi.org/10.1109/SIBGRAPI.2017.60.

Zhang, Q., Yang, L.T., Chen, Z., & Li, P. (2018). A survey on deep learning for big data. *Information Fusion*, *42*(August 2017), 146-157. https://doi.org/10.1016/j.inffus.2017.10.006.