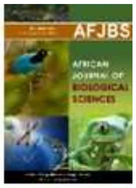
https://doi.org/10.33472/AFJBS.6.7.2024.2382-2391

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com



Research Paper Open Access

A Robust Diabetic Retinopathy Prediction System using Model based Feature Extraction and Machine Learning Algorithms

[1]*S.Suman Das [2]Dr. B. Ramakrishna [3]TulasiRaju Netala
[4]Dr. K .Sreerama Murthi [5]Dr. Guntha Karthik [6]T. Sandeep

[1,2,3]Department of Computer Science and Engineering, Swarnandhra College of Engineering and Technology (Autonomous), Andhra Pradesh, India.
[4]Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Hyderabad, Telangana, India
[5,6]Stanley College of Engineering and Technology for Women (Autonomous),Hyderabad, Telangana, India
ssumanme@gmail.com drbramakrishna.cse@swarnandhra.ac.in ntulasiraju.cse@swarnandhra.ac.in
drsreeram1203@gmail.com gunthakarthik@rocketmail.com tsandeep@stanley.edu.in

Volume 6, Issue 7, Jun 2024

Received: 09 March 2024

Accepted: 10 April 2024

Published: 20 June 2024
doi: 10.33472/AFJBS.6.7.2024.2382-2391

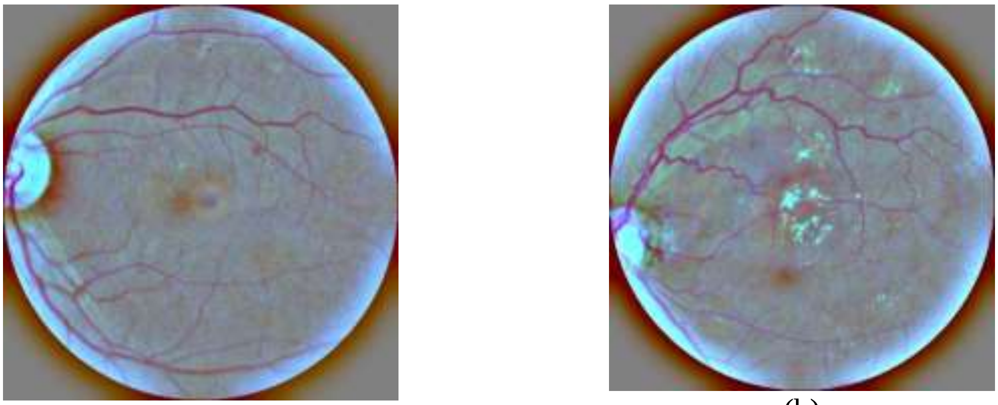
Abstract
In this paper, a robust diabetic retinopathy prediction system using model-based feature extraction strategy and machine learning algorithms has been proposed. The proposed system uses linear support vector classifier for feature extraction. To identify the optimal features, the optimal model of the classifier is identified by using nature inspired optimization algorithms – differential evolution, genetic algorithm and particle swarm optimization. The optimal features are then used to train the following algorithms – K Nearest Neighbours, Decision Tree, Logistic Regression, Random Forest and a Voted Classifier built from an ensemble of Decision Tree, Logistic Regression and Random Forest. The experimentations were conducted using Aptos 2019 datasets. The results of the experimentations show that Random Forest and Voted Classifier are optimal in predicting DR. The proposed system has also out performed most of the systems for DR prediction in the literature.

Keywords: Diabetic Retinopathy, Feature Extraction, Machine Learning Algorithms, Differential Evolution, Genetic Algorithm, Particle Swarm Optimization.

1. Introduction

Diabetic Retinopathy (DR) is one of the major causes of blindness among people suffering from diabetic mellitus (RR Bourne ea, 2013). Most of the time, it so happens that people tend to visit ophthalmologist only when their eye sights are impacted, which may be due to advanced stage of DR. DR can be treated effectively if people report for treatment at early stage of DR. Automating the screening process of DR will help in detecting it at an early stage.

DR is a condition resulting from damage of blood vessels in the retina due to high blood sugar levels (Atwany et al, 2022). Some of the damages caused are swelling and leaking of blood vessels and proliferation of abnormal new blood vessels in the retina. DR exhibits no symptoms in its early stages. But as it progresses, it might cause blurring of vision, poor vision during night hours, colour fading and under extreme conditions loss of vision. Hence automated detection of DR can help in a great way to manage and treat DR.



(a) (b)

Figure 1 (a) Normal retinal image (b) DR affected retinal image.

As the presence of DR is done by screening the retina for symptoms as shown in Figure 1, any autonomous detection system must have image processing capabilities. Hence most of the solutions in the literature involve deep learning architectures which are good at image classification (Archana et al, 2024). But deep learning architectures are of high complexity and requires extensive resources for its operation. Building DR prediction systems using basic machine learning algorithms with lesser complexity will help in wider usage of such systems.

The objective of this paper:

1. Experiment basic machine learning algorithms capability to predict DR.
2. Identify an optimal feature selection strategy for classifiers to be built.
3. Identify the optimal classifiers for DR prediction problem.

The paper is organized as follows: Section 1 Introduction – introduce what DR is? Give a brief about the DR prediction problem and list the objectives of this paper. Section 2 Literature Review – review DR prediction systems in the literature. Section 3 Proposed System – details of the DR prediction system architecture and modules. Section 4 Experimentation and Results – details of the experiments conducted and analysis of the results. Section 5 Conclusion and Future Directions – conclusion of the paper and directions for future research.

2. Literature Review

In this section, a brief review of some of the state-of-the-art diabetic retinopathy systems is given. In (Mohanty et al, 2023), the performances of two DR prediction systems are studied. One of the systems is a hybrid system formed by the combination of VGG16 and XGBoost Classifier. The other system was built using DenseNet121 network. The imbalance in the image dataset used to experiment with, has been subjected to balancing techniques before training the model. It was found that the DenseNet121 exhibited superior performance compared to other similar systems.

In (Sarobin et al, 2022), three DL models have been experimented with for DR prediction – CNN, Hybrid CNN with ResNet and Hybrid CNN with DenseNet2.1. As ResNet and DenseNet 2.1 are pretrained models, and hence the proposed uses transfer learning for training the models. The models were built by extracting relevant features and training the models using the extracted features. It was found that the Hybrid CNN with DenseNet 2.1 outperformed other similar approaches in predicting DR.

In (Kurup et al, 2021), a pretrained DL model, Inception-V3 has been used for DR prediction. The classifier model has been built using transfer learning. The proposed system was able classify DR data with considerable accuracy compared to similar systems in the literature. In (Gangwar and Ravi, 2020), a pretrained DL model, Inception-ResNet-v2 model is used. This model is further enhanced by adding CNN layers on top. The proposed system uses transfer learning for training the model and was able to achieve reasonable performance in classifying DR cases.

In (Saranya et al, 2022), a DenseNet based deep learning model has been used for DR prediction. The system employs noise reduction techniques to make the system robust to noise. The system was able to achieve good classification accuracy compared to similar system then in literature. In (Sanjana et al, 2021), the usage of transfer learning models for DR prediction has been studied. The experimentation involved models such as Xception, InceptionResNetV2, MobileNetV2, DenseNet121, and NASNetMobile. The results of the experimentation indicated the superiority of these models in DR prediction.

In (Kumar and Karthikeyan, 2021), pretrained models based on Attention based Networks, CNN and Multi-layered perceptron has been experimented with. The models - EfficientNet, ResNet, Swin-Transformer, Vision-Transformer (ViT) and MLP-Mixer were trained using annotated retinal image datasets. On testing the model’s performance, it was found that transformer-based models outperformed others. Among the transformer-based models, Swin-Transformer yielded the best result.

In (Khan and Okatan, 2023), a pretrained CNN model has been employed for detecting DR defect. The performance of the classifier is further enhanced using stacking meta learning technique. The classifier was able to exhibit superior performance under lack of availability of sufficient pre-processed data sets. The review of the approaches reveals that most of the systems built have employed pre-trained models based on deep learning algorithms with high complexity. In order to build systems for common use there is a need for building system with lesser complexity without compromising the performance to a large level.

3. Proposed System

The architecture of the proposed system is shown in the Figure 2.

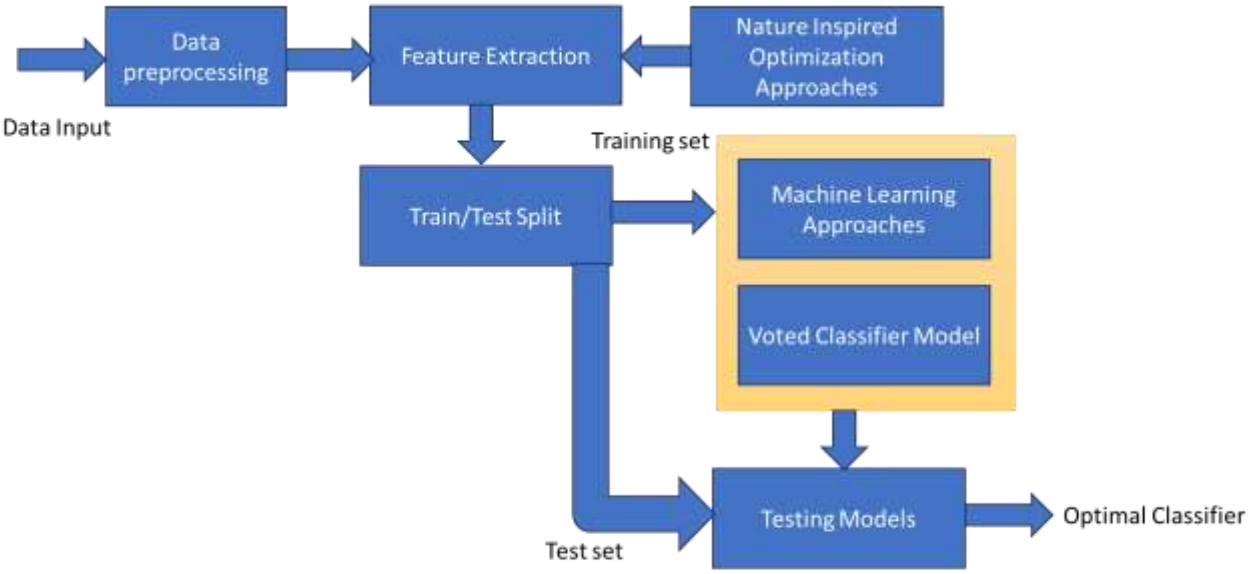


Figure 2 Proposed system architecture

As shown in the Figure, the input data is pre-processed for providing as input to the classifiers algorithms to build the machine learning model. This is followed by the feature extraction stage where the optimal features for categorization are identified. The dataset with the optimal features is then used to train the ML algorithms. The trained models are then tested for performance and the optimal model is chosen for predicting diabetic retinopathy.

3.1 Dataset

To build and evaluate the proposed system, the APTOS 2019 dataset (Karthik et al, 2019) has been used. The dataset comprises of 3662 retinal images as training data and 1928 retinal images as test data. Since this data was part of a competition held by Kaggle in 2019, the test data is not labelled. So, to build and evaluate the proposed system only the training dataset has been used. These images were taken under varied photographic conditions and hence of diverse quality (low – high). The resolution of these images varies from 1440 × 960 pixels to 5184 × 3456 pixels. The images have been categorized into five grades as shown in Table 1.

Table 1 Grades of DR in the dataset

Grade	Description
0	No DR
1	Mild case
2	Moderate case
3	Severe case
4	Proliferative case

3.2 Data Preprocessing

As discussed in the previous section, the images in the dataset are of varied resolution and quality. In order to process the images with uniformity and lesser complexity, the images are read as grey scale images and resized to 128 x 128 pixels. To provide the data as inputs to the machine learning algorithms, it is flattened to form image vectors of size 16,384 pixels i.e. 128 x 128 each. After resizing and flattening, there are around 3662 image vectors corresponding to the 3662 images with 16,384 features representing the pixel values of the images at the respective positions.

The pixel values in the image vectors varies from 0 – 128. To avoid higher valued features from dominating the learning process, the feature values are normalized between the range -1 to 1. Since the aim of this work is to predict DR, the labels of the images with 0 value i.e. No DR is retained as it is, while the labels with values 1, 2 and 3 i.e. different grades of DR are converted to just 1 to indicate DR.

3.3 Feature Extraction

The pre-processed dataset now contains 3662 image vectors with 16,384 features. Since the amount of data required to build a model that generalizes well increases exponential with the number of features used (curse of dimensionality), feature extraction is employed to select the most optimal features for the classification process. In the proposed system, model-based feature selection is employed to achieve dimensionality reduction. In model-based feature extraction, a machine learning algorithm is used for building a classifier using a subset of the dataset. The classifier model is tuned for optimal performance and the features that contributed the most for building this optimal model is chosen as the optimal features.

In the proposed system, linear support vector classifier (lsvc) is used for feature selection. In order to conduct the experiments to build and evaluate the classifier model, the dataset is divided into training, validation and test sets in the ratio 70%:15%:15% respectively. The training dataset and validation dataset is used in the feature extraction process. The lsvc is trained using the training dataset and its performance is evaluated using the validation dataset. The accuracy of the classifier is used as the metric to evaluate the performance the classifier. The tuning of the classifier is achieved by varying the values assigned to the C parameter of lsvc. The C parameter controls the strength of regularization of lsvc.

In order to find the optimal value for C, three nature inspired optimization algorithm has been employed.

1. Differential Evolution (DE) algorithm (Storn and Price, 1997)
2. Genetic Algorithm (GA) (Sastry et al, 2005), and
3. Particle Swarm Optimization algorithm (Kennedy and Eberhart, 1995)

The optimal classifier model is one which produces the best performance in classifying the validation dataset. The C value of this model is taken as the optimal value and the features contributing to the performance of the optimal model is chosen as the optimal features.

Machine Learning (ML) Model for DR prediction

To identify a robust ML model for DR prediction, experiments were conducted with the following machine learning algorithms – K Nearest Neighbours (KNN) (Mucherino et al, 2009), Decision Tree (DT) (Fürnkranz, 2011), Logistic Regression (LGR) (Bisong, 2019), Random Forest (RF) (Breiman, 2001), Naïve Bayes (NB) (Webb, 2011) and Voted Ensemble Classifier (VC) formed from DT, LGR and RF. The training and testing data containing only the optimal features identified in the feature extraction stage are used for building and evaluating the classifier models for DR prediction. Based on the performances of the classifiers, the best classifier for DR prediction is identified.

4. Experimentations and Results

The first set experimentations were conducted to identify the optimal feature set using lsvc and nature inspired optimization algorithms – DE, GA and PSO. The results of these experimentations are shown in the Table 2.

Table 2 Feature extraction results

Optimization Algorithm	Optimal C Value	# Features Identified
DE	0.40791201	2068
GA	0.26788858	1396
PSO	2.65063521	6968

The next set of experimentations were conducted to identify the performance of the machine learning algorithms when trained using the optimal feature sets identified by the optimization algorithms. The results of the experimentations are shown in the Tables. It can be seen from Table 3 that RF and VC has achieved an accuracy of 92.36% outperforming the other algorithms. The results are visualized in Figure 3.

Table 3 Performance for C = 0.40791201, #Features = 2068

ML Algorithm	Accuracy%
KNN	80.18
DT	88
LGR	90.18
RF	92.36
NB	76.36
VC	92.36

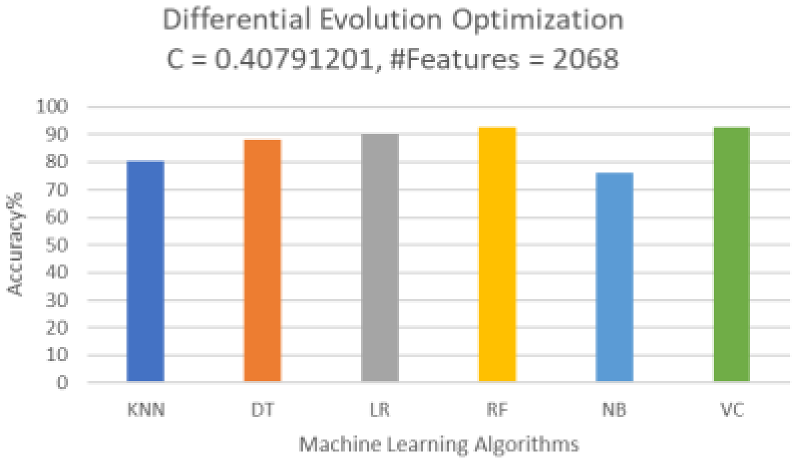


Figure 3 Performance for C = 0.40791201, #Features = 2068

From Table 4 it can be seen that both RF and VC again outperformed other algorithms by achieving an accuracy of 92.18%. The results ae visualized in Figure 4.

Table 4 Performance for C = 0.26788858, #Features = 1396

ML Algorithm	Accuracy%
KNN	78.54
DT	87.81
LGR	90.36
RF	92.18
NB	74.72
VC	92.18

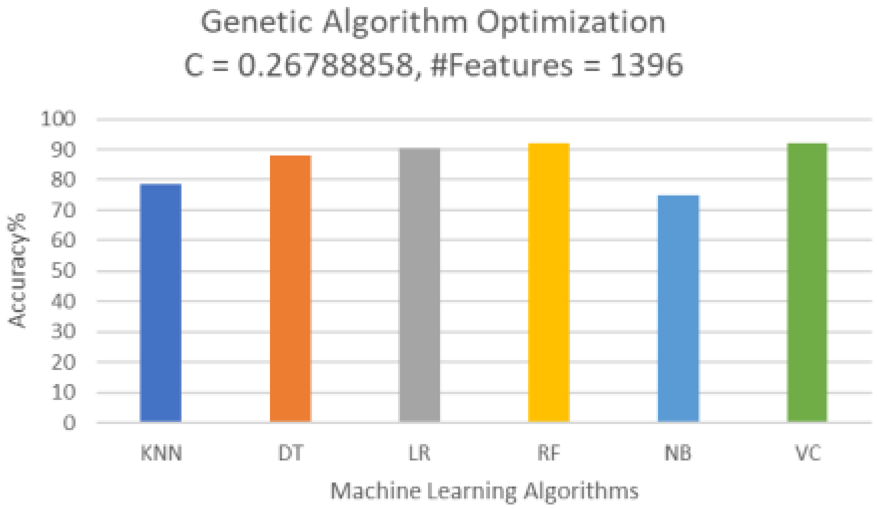


Figure 4 Performance for C = 0.26788858, #Features = 1396

It can be seen from Table 5, that RF has outperformed others by achieving 92.9% accuracy very closely followed by VC with an accuracy of 92.54%. The results are visualized in Figure 5.

Table 5 Performance for C = 2.65063521, #Features = 6968

ML Algorithm	Accuracy%
KNN	78
DT	87.45
LGR	91.45
RF	92.90
NB	74.18
VC	92.54

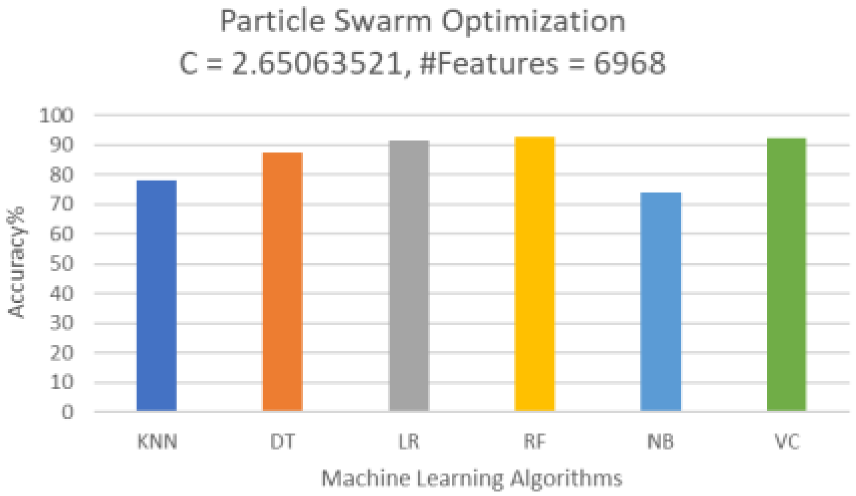


Figure 5 Performance for C = 2.65063521, #Features = 6968

The optimal performances for the different number of optimal features identified is summarized in the Table 6.

Table 6 Optimal Classifier models

Optimization Algorithm	Optimal Features Identified	Optimal ML Algorithm	Accuracy%
DE	2068	VC & RF	92.36
GA	1396	VC & RF	92.18
PSO	6968	RF	92.90

As seen in Table 6, in terms of accuracy the RF algorithm was able to achieve an accuracy of 92.9% with 6968 features identified by PSO. In terms of number of features, the GA was able select the minimum number of features – 1396. The RF and VC algorithms were able to achieve an accuracy of 92.18% using the said features. This is followed by DE with 2068 features identified. Again VC & RF algorithms were able to achieve an accuracy of 92.36%. It can also be seen that the accuracy achieved by the algorithms for the different number of features are very close to one another.

The Table 7 shows the performance of the proposed system relative to similar systems in the literature.

Table 7 Comparative performance

System	Accuracy%
DenseNet 121 (Mohanty, et al, 2023)	97.30
Hybrid Model (Mohanty, et al, 2023)	80
Meta Learning and Deep Learning Techniques (Khan and Okatan 2023)	93
Inception V3 (Kurup, et al, 2021)	82
Inception Res Net V2 (Gangwar and Ravi, 2020)	82.18
DenseNet based Deep Learning Model (Saranya et al. 2022)	83
Transfer Learning Models (Sanjana et al. 2021)	86.25
CNN, Transformer and MLP based Architectures (Kumar and Karthikeyan 2021)	86.4
Deep hybrid architectures (Lahmar and Idri 2022)	89
Single-modality and joint fusion deep learning (El-Ateif and Idri 2022)	90.7
Proposed System	92.90

It can be seen from the Table 7 that the proposed system clearly out performs most of the systems in terms of the accuracy achieved. This despite the fact that the proposed system has been built using simple machine learning algorithms compared to the others listed. The systems that have better performance than the proposed system has been built with deep learning models with much higher complexity and higher number of features.

5. Conclusion and Future Directions

The paper has proposed successfully an intelligent DR prediction system utilizing basic machine learning models with far less complexity than deep learning models. It has also come up with a model-based feature selection method optimized by nature inspired optimization algorithms. From the experimentations conducted and their results, it has been identified that

Random Forest and a Voted Classifier Ensemble comprising of Decision Tree, Logistic Regression and Random Forest were able predict DR with considerable accuracy.

The optimization algorithms were operated in minimal configuration in the current experimentations. In future, various configurations can be tried out to find the optimal recommendations of the algorithms. Also in the current experimentations, the optimization has been applied only in the feature selection process, in future it could be applied to hyper tuning the ML algorithm's parameters.

References

1 Archana Senapati, Hrudaya Kumar Tripathy, Vandana Sharma, Amir H. Gandomi, Artificial intelligence for diabetic retinopathy detection: A systematic review, Informatics in Medicine Unlocked, Volume 45, 2024, 101445, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2024.101445.

2 Atwany, M.Z.; Sahyoun, A.H.; Yaqub, M. Deep learning techniques for diabetic retinopathy classification: A survey. IEEE Access 2022, 10, 28642–28655.w

3 Bisong, E. (2019). Logistic Regression. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_20

4 Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

5 Fürnkranz, J. (2011). Decision Tree. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_204

6 Gangwar, A.K.; Ravi, V. Diabetic Retinopathy Detection Using Transfer Learning and Deep Learning. In Evolution in Computational Intelligence—Frontiers in Intelligent Computing: Theory and Applications (FICTA); Advances in Intelligent Systems and Computing; Springer: Singapore, 2020; Volume 1176, pp. 679–689

7 J. Kennedy and R. Eberhart, "Particle swarm optimization," Proceedings of ICNN'95 - International Conference on Neural Networks, Perth, WA, Australia, 1995, pp. 1942-1948 vol.4, doi: 10.1109/ICNN.1995.488968.

8 Karthik, Maggie, Sohier Dane. (2019). APTOS 2019 Blindness Detection. Kaggle. https://kaggle.com/competitions/aptos2019-blindness-detection

9 Khan, M. A., & Okatan, A. Diabetic Retinopathy Detection Using Meta Learning and Deep Learning Techniques. EURAS - Journal of Engineering and Applied Sciences, - Volume 3 Issue 2 - August - 2023 (85 - 101).

10 Kumar, N.S.; Karthikeyan, B.R. Diabetic Retinopathy Detection using CNN, Transformer and MLP based Architectures. In Proceedings of the 2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Hualien City, Taiwan, 16–19 November 2021; pp. 1–2.

11 Kurup, G.; Jothi, J.A.A.; Kanadath, A. Diabetic Retinopathy Detection and Classification using Pretrained Inception-v3. In Proceedings of the IEEE International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 29–30 October 2021; pp. 1–6.

12 Mohanty, C., Mahapatra, S., Acharya, B., Kokkoras, F., Gerogiannis, V. C., Karamitsos, I., & Kanavos, A. (2023). Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy. Sensors, 23(12), 5726.

13 Mucherino, A., Papajorgji, P.J., Pardalos, P.M. (2009). k-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications, vol 34. Springer, New York, NY. https://doi.org/10.1007/978-0-387-88615-2_4

14 R., Y.; Sarobin, M.V.R.; Panjanathan, R.; Jasmine, S.G.; Anbarasi, L.J. Diabetic Retinopathy Classification Using CNN and Hybrid Deep Convolutional Neural Networks. Symmetry 2022, 14, 1932.

15 RR Bourne ea. 2013. Causes of vision loss worldwide, 1990-2010: a systematic analysis. In: Projections of global mortality and burden of disease from 2002 to 2030;. Vol. 1, Lancet Glob Health; pp. 339–349.

16 Sanjana, S.; Shadin, N.S.; Farzana, M. Automated Diabetic Retinopathy Detection Using Transfer Learning Models. In Proceedings of the 2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), Mirpur, Dhaka, 18–20 November 2021; pp. 1–6.

17 Saranya, P.; Devi, S.K.; Bharanidharan, B. Detection of Diabetic Retinopathy in Retinal Fundus Images using DenseNet based Deep Learning Model. In Proceedings of the 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 10–11 March 2022; pp. 268–272.

18 Sastry, K., Goldberg, D., Kendall, G. (2005). Genetic Algorithms. In: Burke, E.K., Kendall, G. (eds) Search Methodologies. Springer, Boston, MA. https://doi.org/10.1007/0-387-28356-0_4

19 Storn, R., Price, K. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. Journal of Global Optimization 11, 341–359 (1997). https://doi.org/10.1023/A:1008202821328

20 Webb, G.I. (2011). Naïve Bayes. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_576