International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 01 | Jan - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Diabetes Retinopathy Detection Using Hybrid Logistic regression

Narayan Singh Bhirsha
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad,India
21951a0435@iare.ac.in

K.Himani Krishna
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad,India
21951a0462@iare.ac.in

M.Vignesh
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad, India
21951a04Q1@iare.ac.in

Shamili Srimani Pendyala
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad,India
Shamilisrimani.pendyala@gmail.com

Abstract*—Diabetic Retinopathy (DR) results from damage to the retina, potentially leading to vision loss. DR often lacks early warning signs, gradually promoting the formation of new blood vessels at the back of the eye. This can lead to blood clots, bleeding, and distorted vision. Traditional methods have not achieved optimal classification accuracy. Consequently, this article focuses on the implementation of a hybrid logistic regression (HLR) based machine learning model for classifying DR. The process begins with histogram equalization to enhance the DR image region. Next, the segmentation of microaneurysms is conducted using image morphological operations. Features are then extracted using the gray level co-occurrence matrix (GLCM), which reveals the internal relationships associated with DR.The selection of features is conducted using the Gaussian Mixture Model (GMM). Subsequently, the HLR model is employed to execute the multi-class classification task. Simulation results indicate that the proposed method outperforms existing state-of-the-art approaches.

Keywords*— Diabetes Retinopathy, Gray level co-occurrence matrix, Gaussian mixture model, Hybrid logistic regression

I. INTRODUCTION

Diabetic Retinopathy (DR) has rapidly spread across the globe and is notably prevalent in the Indian subcontinent. The increasing incidence of this condition is contributing to a range of associated ailments. Key complications that arise from DR include diabetic neuropathy and diabetic nephropathy. The consequences of these diseases can be life-threatening, highlighting the importance of early

detection and proactive management to prevent further complications. If the condition is not diagnosed in timely manner, it can become quite challenging to address. At that stage, medical management may be the only viable option, as illustrated in the figure.

Diabetes mellitus (DM) is generally triggered by an imbalance in the body's insulin levels. It is classified into two types: type 1 and type 2. The exact cause of type 1 is often unknown and it can manifest in early childhood. Individuals with type 1 diabetes face significant challenges in managing their condition, typically requiring insulin administration. Consequently, many researchers have shifted their focus to type 2 diabetes, which predominantly affects middle-aged individuals. Current research primarily centers on understanding the symptoms and diagnosing the disease through various clinical parameters, leading to the development of appropriate medical guidelines and treatment strategies to address it .These medications must be taken for life, alongside the maintenance of a healthy lifestyle. Current circumstances have impacted individuals, leading to significant changes that have become a regular part of their daily routines.

II. LITERATURE SURVEY

Diabetic retinopathy (DR) is the most significant complication of diabetes. Early diagnosis through retinal image analysis helps prevent vision loss and blindness. The main goal is to automatically detect and classify DR using appropriate algorithms by providing retinal images as input. This survey discusses the basics of diabetes, its prevalence, complications and AI approaches for early detection of DR[1]. Machine learning algorithms have been widely used for DR detection due to their ability to learn from large datasets and improve their performance over time[2].

Several machine learning algorithms have been proposed for DR detection, including support vector machines (SVMs), random forests, and convolutional neural networks (CNNs). These algorithms can be trained using retinal images and their corresponding labels and can be used to classify the images into different stages of DR. One of the key challenges in DR detection is the availability of high-quality retinal images. Image quality can affect the performance of machine learning algorithms, so it is essential to develop algorithms that can handle noisy or low-quality images. Another problem is class imbalance, where the number of images in one class is significantly higher than the number of images in the other class. This can affect the performance of machine learning algorithms, so it is essential to develop algorithms that can handle class imbalance.

Several studies report high DR detection accuracy using machine learning algorithms. For example, a study used a CNN algorithm to detect DR and reported an accuracy of 95%. Another study used an SVM algorithm to detect DR and reported an accuracy of 92%[3].

III. PROPOSED SYSTEM

The machine learning-based approach for diabetic retinopathy (DR) classification has been developed to address the limitations of statistical methods, which often struggle to achieve high accuracy across different variations of DR images. Unlike disease progression methods, this machine learning approach focuses on feature descriptor-based recognition of DR. As illustrated in Figure 2, the proposed HLR deep learning model for classifying DR grades utilizes local feature descriptors to extract disease-specific features. Among the local descriptors discussed is the Gray Level Co-occurrence Matrix (GLCM). Subsequently, a Gaussian Mixture Model (GMM) is employed to select the relevant characteristics for consideration. Finally, the HLR model performs the multi-class classification task.

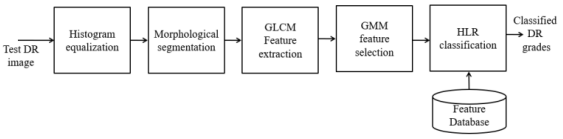

Fig 3.1 ***Block Diagram of proposed method***

A. Methodology

a. Image Enhancement

Histogram equalization means discovering the cumulative distribution function (CDF) for the data PDF (probability density function). Histograms can be modeled using a continuous process function than a discrete process function. Changes can be made using the given PDF image intensity levels Pr (r)[4]. If we consider a given image, the intensity levels are continuous amount; then it is normalized to the range [0,1]. After changes, the image of the photo will be enlarged dynamic range, increased contrast, and the output PDF will be uniform and treated as a CDF. The the methods discussed above can be called the histogram equalization method. In digital images, the intensity groups are different.

Fig 3.2 Histogram equalization images before and after

b.Segmentation

Erosion, opening, dilation, closure and skeletonization are some of the morphological ones procedures that can be performed. Automatic object recognition and machine vision are two areas in which they have a significant influence on DR image processing[6]. Solidification (thickening or growing) items in a binary DR image is done by a technique known as dilation[7]. It is represented by the matrices 0 and 1, which show how much this stiffening is regulated by structure (SE). When a structuring element is used in conjunction with a binary DR image, the process of erosion shrinks or thins the items in the binary photographic image.

Opening and Closing: In general, dilation and erosion are two key processes in an image processing that are widely used in many different configurations. It can be a single shot of DR be subjected to multiple dilations and/or erosions, each involving the use of the same or different structural elements. The interaction of these two concepts is what causes the opening of a closing morphological images, or a morphological opening can be considered as an erosion procedure followed by dilation surgery, with dilation surgery first. If the morphological opening of the DR image X by Y is represented by the symbol XoY, then this is the case the usual notation for this. X Y erosion followed by dilation of the obtained result using Y to close and open. Skeletonization is an alternative way of shrinking binary DR image objects with thin lines that may show significant data about the original in the form of subjects. Used to display important information about the shape of the original objects. It's comparable to thinning in that it stores a large amount of information about the internal structure of objects with a thickness of one pixel.

c. Feature Extraction

After the segmentation of retinal structures, the next step is feature extraction. Textural properties are obtained using a gray occurrence matrix (GLCM). Texture refers to a physical structure characterized by roughness or smoothness. Texture elements can be spectral, statistical or structural. The statistical features obtained are entropy, mean, third moment and standard deviation[8]. Mean: The mean value of the image (gray level) refers to the average of all pixel values ˆ Standard Deviation: The most commonly used metric tells how much variation was found from the average. Its low value indicates the degree to which the data points are related to the mean value and a high value indicated that the points are spread over a large area. ˆ Entropy: Entropy measures the error or disorder in grayscale image data. It measures the randomness of the texture in the image. ˆ Third moment: This function is a measure of skewness in the image histogram. GLCM: Second-order statistical features can be extracted using the GLCM matrix. The The Matlab function to calculate GLCM is used to calculate all the functions and the result is returned as a structure.

d. Gaussian mixture model feature selection

The extracted features are selected using an attribute selection filter. This filter implementation allows you to choose different search and rating methods. Linear discriminant analysis identifies a linear combination of characters that characterize or separate two or more classes and select the attributes in the proposed work. The primary concern here is estimation Probability density for a given amount of data. Previous decisions on this issue included parametric and non-parametric methods. Using parametric calculations, we obtain a specific baseline regression and apply these estimates to k estimate data values. The problem is that this distribution can be misleading. On conversely, when nonlinear methods are considered, model selection is not mandated; however, in these cases the data are estimated under the guidance of experts but the disadvantage is that he needs a lot of information to make effective decisions[9]. To overcome these limitations Gaussian mixture models are considered, which are assumed to be a semi-directed model and thus model the data taking into account the advantages of both methods.

e. Dataset

The data set For the current research is considered the standard DR database, i.e. DIARETDB1. This particular database contained a total of 89 color fundus images, of which 84 had mild non-proliferative signs. The remaining five were treated as usual and showed no signs of it The experts in this assessment declared DR. By setting the digital fundus camera, various images are taken with 50 degree fields. This data set can be called the "calibration level 1 file". images that can be used to evaluate the performances of various diagnostic methods and images may be comparable. Images are collected using a software tool used for image annotation with four experts they are experienced in medical technology and perform independent labeling. In general, that the doctor will be asked to mark areas related to bleeding, soft and hard exudates, and microaneurysms[10]. The considered database is public for DR detection comparison from digital images.

IV. IMPLEMENTATION

HARDWARE AND SOFTWARE SETUP

We conducted the experiments on a workstation with the following compute configuration:
CPU: Core i7 8700K,
Main memory: 64 GB DDR4,
40 GPU: Nvidia Titan X (Pascal).
Python 3.7.4 was used as the base language, and fastai [36] v1.4 was used along with Pytorch [40] as the deep learning frameworks. We also used Pandas, Numpy, Scikit-learn, and Matplotlib python packages in our experiments for data preprocessing, manipulation, cross-validation, etc.

4.1 Explanation of Key Elements

4.1.1 PYTHON

Guido Van Rossum designed Python, an interpreter, high-level, general-purpose programming language that was initially released in 1991.

The Python language has the following features:
• Simple to Understand and Python is a simple language to learn and use. It's a high level programming language that's helpful to developers.
• Expressive Language Python is more expressive than other languages, making it more intelligible and readable.
• Python is a interpreted language, which means the interpreter runs the code line by line. This makes debugging simple, making it suitable for novices.
• Language that is cross-platform Python can operate on a variety of platforms, including Windows, Linux, UNIX, and Macintosh, among others.Python is a portable language.
• Open Source and Free Software Python is a free programming language that can be downloaded from a secure internet address. It's also possible to get the source code. As a result, there is an open supply.
• Object-Oriented Programming Python aids the development of object-oriented languages and concepts such as classes and gadgets.
• Adaptable It means that other languages, such as C/C++, may be used to put the code together, and as a result, it can be utilized in our Python code.

4.2 Library Files

4.2.1 NUMPY:

NumPy is a Python library that includes support for enormous, multi-dimensional arrays and matrices, as well as a large set of high-level mathematical functions that may be applied to those arrays. Numeric, the forerunner to NumPy, was designed by Jim Hugunin with help from a number of other people. Travis Oliphant built NumPy in 2005 by heavily modifying 46 Numeric and combining features from the competitor Numarray. NumPy is a large open source software project with numerous contributors. The Python programming language was not initially intended for numerical computing, but it quickly caught the attention of the clinical and engineering communities, prompting the formation of a special interest group known as matrix-sig in 1995 with the goal of defining an array computing bundle. Guido van Rossum, a Python clothier and maintainer, was one of its contributors, adding changes to Python's grammar (especially the indexing syntax) to make array computation easier. Jim Fulton completes a matrix package, which is later modified by Jim Hugunin to create Numeric, also known as Numerical Python extensions or NumPy. Hugunin, a PhD student at MIT, joined the Corporation for National Research Initiatives (CNRI) to work on JPython in 1997, leaving the maintainer position to Paul Dubois of Lawrence Livermore National Laboratory (LLNL). NumPy is a non-optimizing byte code interpreter that targets the CPython Python reference implementation. Algorithms created for this version of Python are frequently much slower than their compiled counterparts. NumPy tackles the slowness issue in part by providing multidimensional arrays and capabilities and operators that work appropriately on arrays, which necessitates rewriting some code, generally inner loops, in order to use NumPy. NumPy arrays are used to store and perform information in the Python bindings of the widely used computer vision library OpenCV. Because images with more than one channel are really stored as 3-dimensional arrays, indexing, slicing, and protecting with separate arrays are all highly eco-friendly ways to access specific pixels in a picture. The NumPy array, which is used as a standard statistical structure in OpenCV for pictures, extracted feature factors, kernel filtering, and other tasks, greatly simplifies the development process and debugging.

4.2.2 PANDAS

Pandas is an open-source, BSD-certified library for the Python programming language that provides high-overall performance, easy-to-use statistics systems, and data analysis tools. A panda is a Python module that provides quick, flexible and expressive facts structures for 47 working with "relational" or "labelled" data in a clean and understandable manner. Its goal is to become the most important high-level building element for undertaking realistic; realworld international records evaluations in Python. It also has the larger goal of being the most powerful and versatile open source data analysis/manipulation device available in any language. It is already well on its way to achieving this goal.
A panda is well-suited to a wide range of statistical applications: • Tabular statistics containing columns of varying types, such as those seen in a SQL database or an Excel spread sheet.
• Row and column labels for arbitrary matrix information (homogeneously typed or heterogeneous)
• Any observational/statistical statistics set of any kind. To be placed into a panda's facts form, the data does not need to be tagged in any way. Pandas' two basic statistics systems, Series (1-dimensional) and DataFrame (2-dimensional), address the vast majority of common use cases in finance, information, social science, and a wide range of engineering disciplines. A panda is built on top of NumPy and is intended to work in conjunction with a variety of different third-party libraries in scientific computing.

4.2.3 MATPLOT LIB:

Matplotlib is a Python 2D plotting toolkit that generates book-quality figures in a variety of hardcopy codecs and interactive contexts. Matplotlib is a Python library that may be used in scripts, the Python and IPython shells, the Jupyter notebook, web applications servers, and four

graphical user interface toolkits. Matplotlib aims to make both smooth and difficult tasks feasible. With just a few lines of code, you can create graphs, histograms, electrical spectra, bar charts, error charts, scatter plots, and more. See the pattern plots and thumbnail galleries for samples. The pyplot package provides a MATLAB-like interface for convenient plotting. Through an object-oriented interface or a set of methods common to MATLAB users, you have complete 48 control over line styles, font houses, axis houses, and so on for the electricity consumer.

4.2.4 SEABORN:

Seaborn is a data visualisation package for Python that is mostly based on matplotlib. It provides a high-level interface for creating visually beautiful and useful statistics graphs. It is a Python module for creating statistical visuals. It's based on matplotlib, and it's tightly integrated with panda's data systems. The goal of Seaborn is to make visualisation a major aspect of information exploration and understanding. Its dataset-oriented charting features operate on facts frames and arrays holding whole datasets, doing the necessary semantic mapping and statistical aggregation internally to provide relevant charts.

4.2.5 Scikit-learn:

To put it another way, sci-kit study is a free software system studying library for Python. It includes support vector machines, random forests, gradient boosting, k-method, and DBSCAN, among other categorization, regression, and clustering algorithms, and is designed to work with the Python numerical and clinical libraries NumPy and SciPy. Scikit-research was created in 2007 as a Google summers of code initiative by David Cornopean. Matthieu Brucher joined the challenge and began using it in thesis paintings. INRIA was considered in 2010, and the initial public release (v0.1 beta) was released in late January 2010. The project presently has over 30 active participants and has received financial support from INRIA, Google, Tiny Clues, and the Python Software Foundation. In most cases, solving a problem involves looking at a collection of n samples of facts and then attempting to predict attributes of unknown facts. If a sample has more than one kind, such as a multi-dimensional entry (also known as multivariate information), it is said to have many characteristics or capabilities.

V. RESULTS AND DISCUSSION

In this project we are detecting presents of diabetes and its stages and to implement this project we are using Convolution Neural Network (CNN) Algorithm). To train CNN we are using an IDRID (Indian diabetes Retinopathy Diabetes) image which consists of 5 different types of diabetes disease images. Those 5 types of diseases are

Fig.5.1 Upload the dataset

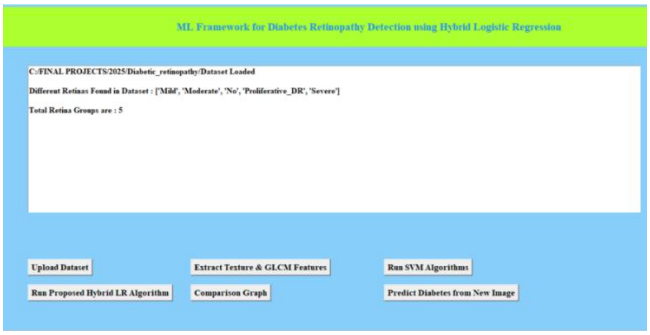

Fig.5.2 Dataset

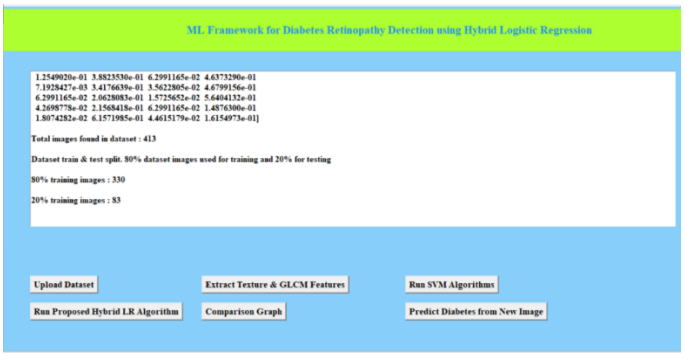

Fig.5.3 Preprocessing dataset

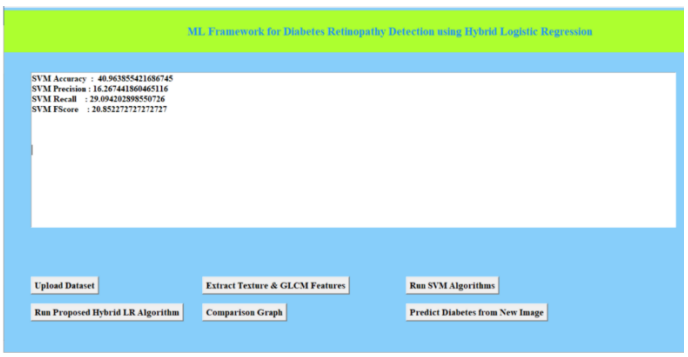

Fig.5.4 Run SVM algorithm

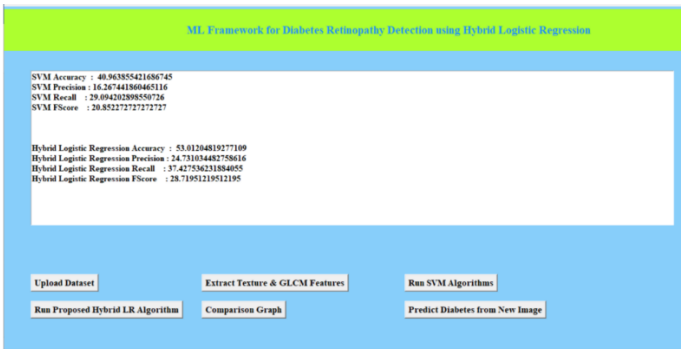

Fig.5.5 Run proposed hybrid LR algorithm

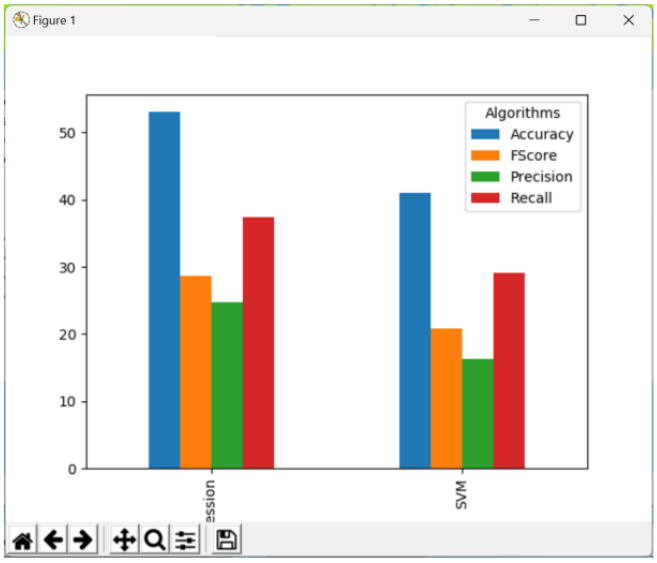

Fig.5.6 Comparison graph

Fig.5.7 Retina predicted as: NO

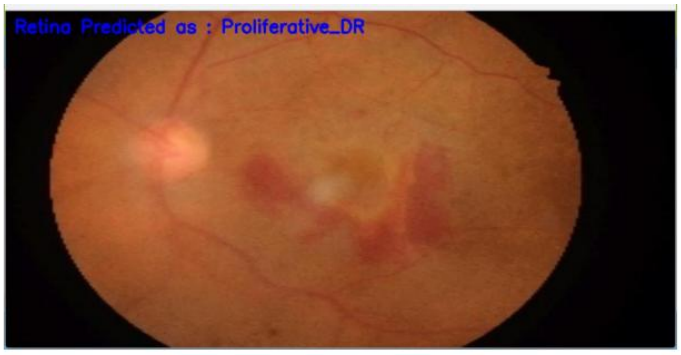

Fig.5.8 Retina predicted as: Proliferative_DR

VI. CONCLUSION

DR is a disorder associated with the consequence of fundus problems. This leads to blindness if not identified and treated at appropriate stages. The proposed work is automated method for multi-class DR classification using HLR. The input image is preprocessed and segment using various morphological operations and thresholding techniques for segmentation MA, bleeding and blood vessels. Statistical features are extracted from these images Attribute-based feature selection is used to select optimal features. These functions are embedded in the HLR classifier. The proposed method prevails over earlier methods described in the literature in terms of increased accuracy. As a future improvement, more images from heterogeneous database and better image preprocessing techniques can be improved accuracy.

VII. FUTURE SCOPE

The future scope of machine learning detection of diabetic retinopathy is huge and promising. Recent advances in deep learning have enabled the development of accurate and efficient systems for detecting diabetic retinopathy from retinal fundus photographs. The development of explainable AI models can provide insight into the decision-making process of machine learning models, which can improve trust and acceptance in clinical settings.

VIII. REFERENCES

[1] D. Mellitus, "Diagnosis and classification of diabetes mellitus," Diabetes Care, vol. 37, no. 1, pp. S81–S90, 2014. [Online]. Available: https://care.diabetesjournals.org/content/37/Supplement_1/S81

[2] E. S. Shin, C. M. Sorenson, and N. Sheibani, "Diabetes and retinal vascular dysfunction," J. Ophthalmic Vis. Res., vol. 9, no. 3, pp. 362–373, Sep. 2014.

[3] X. Zhang, J. B. Saaddine, C.-F. Chou, M. F. Cotch, Y. J. Cheng, L. S. Geiss, E. W. Gregg, A. L. Albright, B. E. K. Klein, and R. Klein, "Prevalence of diabetic retinopathy in the United States, 2005-2008," JAMA, vol. 304, no. 6, pp. 649–656, Aug. 2010.

[4] J. Yau, S. Rogers, and R. Kawasaki, "Global prevalence and major risk factors of diabetic retinopathy," Diabetes Care, vol. 35, no. 3, pp. 556–564, 2012.

[5] S. D. Solomon, E. Chew, E. J. Duh, L. Sobrin, J. K. Sun, B. L. VanderBeek, C. C. Wykoff, and T. W. Gardner, "Diabetic retinopathy: A position statement by the American diabetes association," Diabetes Care, vol. 40, no. 3, pp. 412–418, Mar. 2017. [Online]. Available: https://care.diabetesjournals. Org/content/40/3/412

[6] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, p. 2402, Dec. 2016. [Online]. Available: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2016.17216

[7] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," Ophthalmology, vol. 124, no. 7, pp. 962–969, Jul. 2017.

[8] M. M. Islam, H.-C. Yang, T. N. Poly, W.-S. Jian, and Y.-C. (Jack) Li, "Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis," Comput. Methods Programs Biomed. vol. 191, Jul. 2020, Art. No. 105320. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260719311010

[9] A. Bora, S. Balasubramanian, B. Babenko, S. Virmani, S. Venugopalan, A. Mitani, G. de Oliveira Marinho, J. Cuadros, P. Ruamviboonsuk, G. S. Corrado, L. Peng, D. R. Webster, A. V. Varadarajan, N. Hammel, Y. Liu, and P. Bavishi, "Predicting the risk of developing diabetic retinopathy using deep learning," Lancet Digit. Health, vol. 3, no. 1, pp. e10–e19, Jan. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2589750020302508 51

[10] S. Musleh, T. Alam, A. Bouzerdoum, S. B. Belhaouari, and H. Baali, "Identification of potential risk factors of diabetes for the qatari population," in Proc. IEEE Int. Conf. Informat., IoT, Enabling Technol. (ICIoT), Feb. 2020, pp. 243–246.