Division of Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling

Project Title

Chiara Caglieri

September 2022

Dissertation submitted in partial fulfillment for the degree of
Master of Science in Artificial Intelligence

September 2022

Abstract

Summary of the dissertation **within one page**.

This template starts the page numbering at the foot of this page. While you are printing drafts, you might find it useful to add the printing date and time into the footer – to help you, and your supervisor, tell which version is most current.

It is suggested that the abstract be structured as follows:

- Problem: What you tackled, and why this needed a solution
- Objectives: What you set out to achieve, and how this addressed the problem
- Methodology: How you went about solving the problem
- Achievements: What you managed to achieve, and how far it meets your objectives.

Attestation

I understand the nature of plagiarism, and I am aware of the University's policy on this.

I certify that this dissertation reports original work by me during my University project except for the following (*adjust according to the circumstances*):

- The technology review in Section 2.5 was largely taken from [17].
- The code discussed in Section 3.1 was created by Acme Corporation (`www.acme-corp.com/JavaExpert`) and was used in accordance with the licence supplied.
- The code discussed in Section 3.5 was written by my supervisor.
- The code discussed in Section 4.2 was developed by me during a vacation placement with the collaborating company. In addition, this used ideas I had already developed in my own time.

Signature: *(you must delete this, then sign and date this page)* **Date**

Acknowledgements

Acknowledge anyone that you wish to thank who has helped you in your work or supported you in any way: such as your supervisor, technical support staff, fellow students, external organisations or family. Acknowledge the source of any work that is not your own.

Contents

List of Figures

1 Introduction

1.1 Background

According to the most recent *World Health Organization (WHO)* report on vision [1], at least 2.2 billion people across the world are impacted by some vision impairment to a degree. Not only this, but many of these cases go undetected and therefore undergo no treatment or are addressed in an advanced stage of the disease. As pointed out by the *WHO* [1], certain subsets of the population are more likely to not receive appropriate care, due to factors like income, age or geographical location. The situation is also estimated to get worse over the next decade, due to the combination of population increase and the rise in insurgence of the diseases most commonly associated with vision loss.

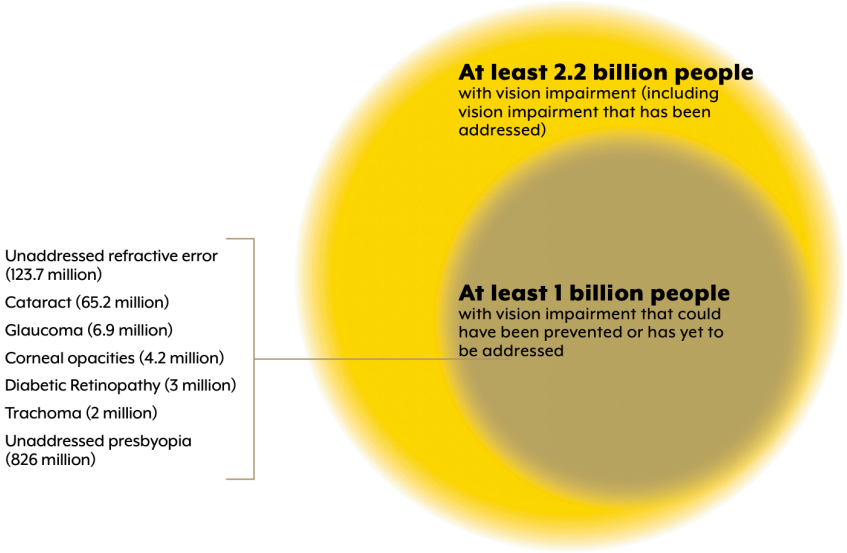


Figure 1.1: Proportion of cases of vision impairment that could have been prevented or addressed. Source: WHO [1]

As shown in *Figure 1.1*, one of the leading causes of vision impairment and blindness is *Diabetic Retinopathy*. This condition, which is a consequence of the insurgence of *Diabetes Mellitus*, is destined to increase in incidence due to the rise in number of diabetic patients, reaching 191 million by 2030 [2]. The next section expands on the medical definition of Diabetic Retinopathy, and illustrates the challenges related to the integration of automated detection systems in clinical settings.

1.1.1 Diabetic Retinopathy

Diabetic Retinopathy is defined as a complication of the insurgence of *Diabetes Mellitus*, and it is in many cases responsible for severe vision loss. Blood vessels in the eye suffer

ruptures caused by exceptionally high blood sugar levels, therefore preventing the retina from receiving an adequate blood supply to function properly. Since the retina is the structure within the eye responsible for the conversion of light into what we perceive as images, any damage to it will result in vision loss.
Early detection of Diabetic Retinopathy is crucial, since starting an appropriate treatment can prevent the condition from spreading and causing ulterior damage. As with Diabetes Mellitus, Retinopathy has no cure, but can be treated in a variety of ways, including through laser therapy and injections.

Diabetic Retinopathy is classed as a *microvascular disease* [3], and as such its diagnosis originates from the detection of a veriety of *microvascular lesions*. The lesions traditionally associated with Diabetic Retinopathy are thew following:

- *Microaneurysms*: small ruptures of blood vessels within the eye, usually detected through a dilated eye exam. Microaneurysms are considered to be the most prominent symptom of Retinopathy.
- *Hemorrhages*: can be the result of different complications, including vein occlusion caused by high blood sugar. Within the context of Diabetic Retinopathy, we look for *Intraretinal Hemorrhages*, which appear as dense, dark red, and sharply outlined [4].
- *Exudates*: lipid residues originating from leaking damaged capillaries.
- *Hard exudates*: exudates composed of extracellular lipid and usually found in the outer layer of the retina.

Figure 1.2 shows an example of these lesions:

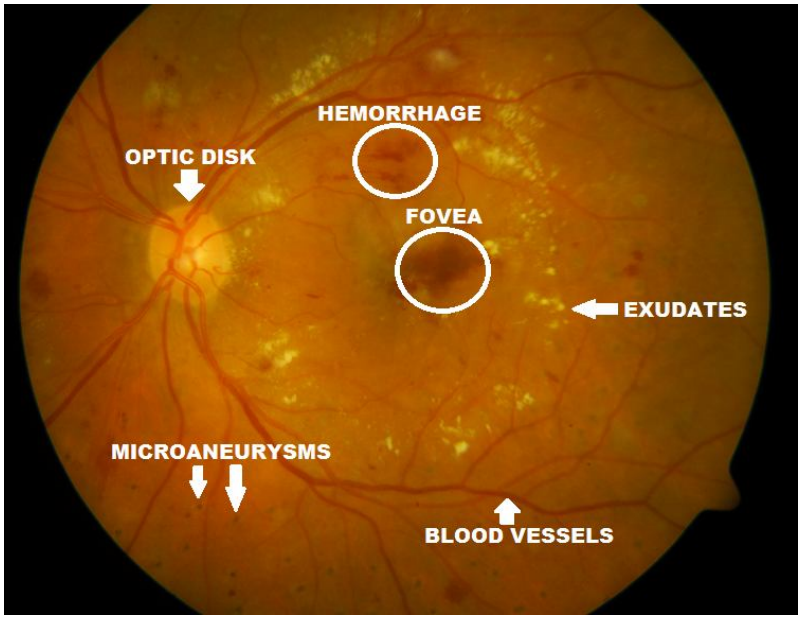


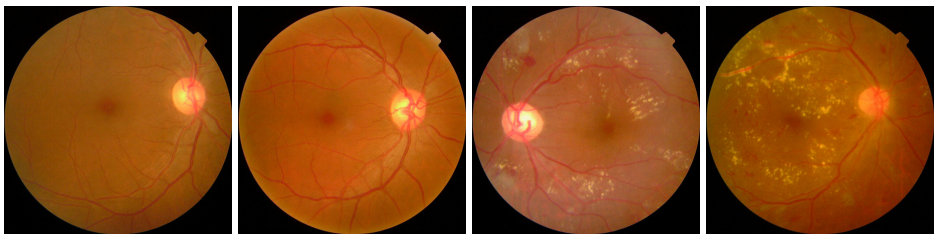
Figure 1.2: Different microvascular lesions associated with Diabetic Retinopathy. Source: [5]

Clinical grading for Diabetic Retinopathy can be provided according to different standards, of which the one proposed by the ETDRS [6] is the most popular. The grading scale used alongside the majority of public datasets resembles that of the ETDRS, although it does not distinguish between Non-High Risk Proliferative DR and High-Risk Proliferative DR. Such grading scale is the *International Clinical Diabetic Retinopathy Disease Severity Scale* (ICDRDSS) [7], and it recognizes Diabetic Retinopathy at five different stages:

- **No DR**: no fundus alterations attributed to diabetes.
- **Mild non-proliferative DR**: a few microaneurysms are present.
- **Moderate non-proliferative DR**: presence of microaneurysms, intraretinal hemorrhages, and non-severe venous beading.
- **Severe non-proliferative DR**: large hemorrhages, severe venous beading, or severe intraretinal abnormalities.
- **Proliferative DR**: neovascularization of the structures of the eye, including the retina and the optic disk. When Proliferative DR is diagnosed, it is also necessary to assess the presence of *Macular Edema*.

Such a scale will be used as a reference for automated disease grading.

Figure 1.3 shows an example of each stage.



(a) No DR (b) Mild DR (c) Moderate DR (d) Severe DR

(e) Proliferative DR

Figure 1.3: Stages of Diabetic Retinopathy. Source: [12]

1.2 Applications

Applications of automated systems for the detection of Diabetic Retinopathy are exclusively confined to a clinical setting. Although the introduction of a Machine Learning system in a medical context is sometimes a difficult process due to the high-risk nature

of the field of application itself, there are examples of systems that have been validated and approved for clinical use.

An example is *EyeArt®*, which obtained US FDA clearance, CE marking as a class 2A medical device in the European Union, and a Health Canada license [8]. The system can output either a "positive" or "negative" result, together with the appropriate disease grading. In the first case, the patient gets referred to an eye specialist, while in the second case a new check-up is scheduled after 9-12 months [9]. One of the main advantages of *EyeArt®* is that it is able to process large batches of images in a very short time, compared to the time required in case of manual human evaluation. Additionally, the system achieves high Specificity (91.1%) and Sensitivity (91.3%), as reported by Bhaskaranand et al. in their review study [9]. Two example output reports produced by *EyeArt®* are shown in *Figure 1.4*:

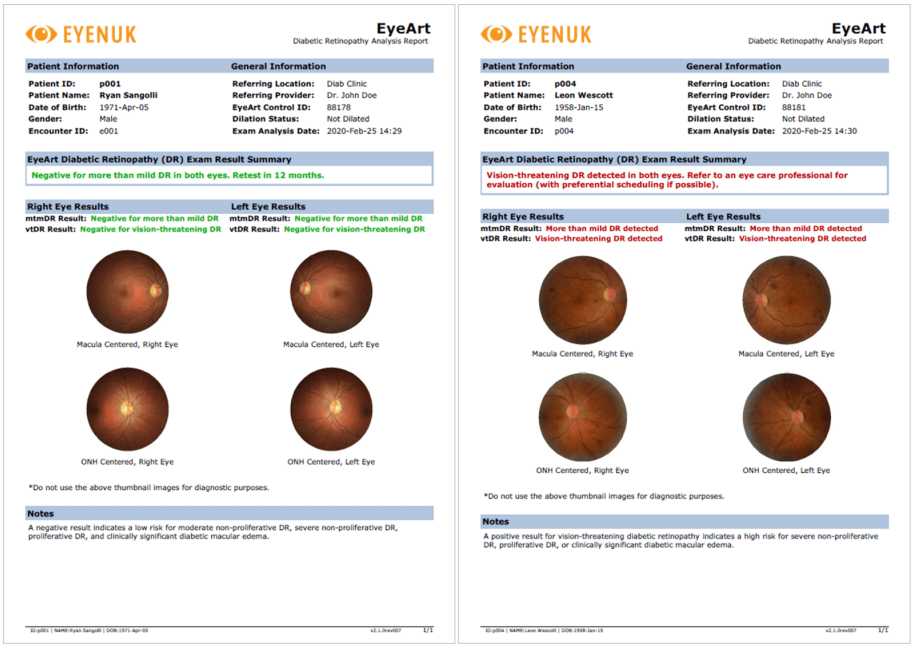


Figure 1.4: Example output reports produced by *EyeArt®*. Source: [8]

As it can be observed from the reports produced by *EyeArt®*, there is no information on which structures within the eye retina led the software to the advertised diagnosis. *EyeArt®* can therefore be used as an aid in the initial phases of Diabetic Retinopathy screening, and an in-depth medical examination is always required.

Another review study proposed by Tufail et al. [10] puts three commercial systems (*EyeArt®*, *Retmarker*, *iGradingM*) to the test to analyse their performance against that of human graders. As shown by the study, both *EyeArt®* and *Retmarker* achieved good Sensitivity and Specificity on all grades of Diabetic Retinopathy, and can therefore be considered as a valuable aid to the diagnosis process. However, the third system under evaluation proved to be unuseful, as it classified all samples are either "disease" or "ungradable".

1.3 Problem Definition

A major issue in the integration of Machine Learning algorithms into clinical environments is the frequent lack of explainability provided by the models. As discussed further in *Chapter 2*, many of the proposed methods for Diabetic Retinopathy detection achieve optimal performance, and have the potential to offer a substantial aid to medical professionals in the diagnosis process. However, in many cases their real-world employment is hindered by the lack of information on the salient features used by the model for the final classification of the sample, as the output is often just a single label indicating the predicted grade of the disease. Additionally, as pointed out by Muddamsetty et al. [26], previous attempts to integrate Machine Learning in clinical settings resulted in either over reliance or under reliance on the model employed. This is therefore a crucial point of attention when developing a new model and proposing its integration in a medical diagnosis process.

Interpretability in the context of Machine Learning is defined as the ability of the model *to explain or to present in understandable terms to a human* [11]. As Doshi-Velex et al. point out [11], providing explicitly interpretable methods is not always a necessity, as in many cases the knowledge that a model has undergone extensive testing and has been employed to solve similar tasks before is sufficient to instill trust in the end user. However, this is not the case for the problem at hand, as not only there is little to no evidence regarding ML models being validated for clinical diagnosis of Diabetic Retinopathy, but application of AI in the medical field is still in its early stages.
In order to address these issue, *Explainable AI* (XAI) techniques can be used to highlight the significant features that led the algorithm to pick a specific classification label, therefore allowing medical professionals to corroborate the diagnosis. Well-established techniques for interpretability will be discussed in *Chapter 2*.

In summary, there are several challenges that need to be address when developing an automated system for Diabetic Retinopathy detection:

- **Reliability**: the proposed model needs to achieve optimal accuracy, specificity and sensitivity values (see *Section 2.3*), as it should be able to correctly identify different stages of the disease from retinal fundus images.
- **Explainability**: the end user should be able to understand how and why the model selected a speciific disease class. This can be achieved by visually highlighting salient features in retinal images.
- **Flexibility**: the model should be able to analyse fundus images acquired with different devices and in different lighting conditions.

1.4 Scope and Objectives

The aim of this project is to review existing methods for Diabetic Retinopathy detection and grading, as well as analysing and applying model explainability techniques to such models. Additionally, a *ConvNext* model will be proposed and fine-tuned on the dataset of choice. Performance evaluation will be carried out by comparing to a benchmark of well-established models such as *ResNet50*. A summary of the objectives of the dissertation is presented below:

- Review **State-of-the-Art methods for Diabetic Retinopathy detection** and discuss on the achieved performance
- Define the **evaluation metrics** to be used for the analysis of the results.
- Identify the **datasets** to be used for training and testing.
- **Replicate existing implementations** to obtain a benchmark for performance evaluation.
- **Construct a new model** using a ConvNext Architecture and tune the hyperparameters to optimize performance.
- **Review Explainable AI (XAI) techniques** and **apply** them to the models to obtain a visualisation of salient features.
- Report on the **results** and identify future improvement possibilities.

1.5 Contributions

TO DO

1.6 Dissertation Organisation

The dissertation is organised in the following sections:

- Section 1: **Introduction**: an overview of Diabetic Retinopathy and retinal imaging analysis is provided. This section touches upon the medical context surrounding this project, and it discusses the objectives of the dissertation.
- Section 2: **State Of the Art**: a survey of current literature on Diabetic Retinopathy detection is provided. Both well-established and innovative methods are analysed, and an evaluation of current developments in Explainability is carried out.
- Section 3: **Methodology**: this section illustrates the process of implementing and/or replicating State Of the Art methods in Diabetic Retinopathy detection. Furthermore, it explains the implementation and integration of Explainability techniques.
- Section 4: **Results and Evaluation**: a review of the achieved results is provided in this section. Algorithm performance is analysed and compared against benchmarks, and visual results are presented.

- Section 5: **Conclusion and Future Work**: a final discussion is provided in this section, alongside an evaluation of possible future developments and improvements of the work.

2 State-of-the-Art

This section presents a review of the current State-of-the-Art methods for Diabetic Retinopathy detection and grading, and it gives an overview of recent developments in Explainable AI applied to the medical field. Evaluation metrics will be discussed and performance benchmarks will be provided. The final section of this chapter will also touch upon the different datasets available to the public.

2.1 Related Work

2.1.1 Preprocessing Techniques

A variety of preprocessing techniques are usually applied to retinal images before they are used for model training. The objective of this is to make structures within the retina more prominent for the classifier to identify. In addition, preprocessing techniques are used to reduce noise and enhance overall image quality.
In their review of preprocessing methods to be applied to fundus retinal images [16], Hernandez-Matas et al. illustrate a few techniques:

- *Green Channel Extraction*: visualising the three RGB channels separately shows that the red and blue channel carry more noise compared to the green channel, which seems to better retain feature information [17]. For this reason, preprocessing methods are generally applied exclusively to the green channel.
- *Contrast Normalisation*: illumination of the retina can be uneven due to a mix of factors, including acquisition techniques and eye shape. Contrast normalisation, which is usually applied solely to the green channel, aims to balance image contrast in each of the samples for a more even luminance. The most common type of equalization used for fundus images is *locally adaptive histogram equalization (CLAHE)* [16].
- *Noise Reduction*: noise in retinal fundus images mostly originates from acquisition devices and can vary in level within a single dataset.

2.1.2 Diabetic Retinopathy Detection

Several Deep Learning based approaches have been proposed in recent years for the detection and grading of Diabetic Retinopathy. A large part of the available literature focuses on the employment of Convolutional Neural Networks (CNNs) for classification of eye fundus images, which have been proven to yield optimal results. However, more complex and innovative solutions have been developed, together with methods to address low variability in the data, overfitting, image enhancement, and gradability.
Some popular approaches include:

- **Transfer Learning**: applying knowledge learned from a model trained on a larger dataset to the task at hand. Within the domain of Diabetic Retinopathy detection, this usually translates to loading the weights of a model trained on one of the variants of the *ImageNET* dataset. [18]
- **Ensemble Methods**: instead of employing one single model for classification, multiple base models can be combined to obtain a better performing one. [19]
- **Active Learning**: models are able to identify potentially valuable unlabelled samples and then interactively query the user for a ground truth. Active Learning can be useful when the dataset is too large to be manually labelled and therefore there is a need to establish an order of priority for the samples. [20]
- **Synergic Deep Learning**: a convolutional neural network can be used in conjunction with a Synergic Signal System, allowing the two to mutually learn image representation [**?**]. [22]
- **Multi-Task Learning**: attempts to solve multiple learning task at the same time by using similarities and differences among tasks. [19]

Convolutional Neural Networks (CNNs)

A simple classification model is proposed by Ting et al. [18], and it aims to employ the *Inception-v3* architecture for classification of preprocessed macula-centered retinal fundus images. In this case, the approach taken is that of transfer learning, in which the network is pre-instantiated with weights originating from training on a different dataset, *ImageNet* in this case. The final algorithm consisted of an ensemble of 10 networks trained on the same data and it was able to provide grading on 4 different levels of disease, including an indication of referrable Diabetic Macular Edema. The model was trained on the EyePACS dataset and validated on both EyePACS and Messidor-2, achieving an average sensitivity of 88.3%, and an average specificity of 98.6%.

The method proposed by Tymchenko et al. [19] makes use of an ImageNet-pretrained base model to initialize the Deep Convolutional Neural Network. The approach taken is that of Multi-Task Learning through the combined use of a Classification head, a Regression head and an Ordinal Regression head. Training is carried out across multiple stages and it uses different parameter settings and datasets at each stage. The CNN is pretrained on the Kaggle EyePACS dataset, as it is the largest available, while the main training is done on the IDRID and MESSIDOR datasets. Three different architectures (EfficientNet-B4, EfficientNet-B5, SE-ResNeXt50) are then conjoined in a 20 model ensemble for final scoring. Without Test-Time Augmentation, the final model achieved 98.6% Accuracy, 99.1% Sensitivity, 99.1% Specificity and a QWK score of 98.1%.

Qureshi et al. [20] propose an Active Learning approach, where a Convolutional Neural Network is employed to extract the salient features. An Active Learning technique called *Expected Gradient Length* is used to make the CNN label-efficient so that training time is

reduced. The CNN follows the structure of the popular *LeNet* architecture and it is trained on the EyePACS dataset. In the first stage, the model is able to identify highly informative samples The proposed model achieved an Accuracy of 98%, Sensitivity of 92.20%, and Specificity of 95.10%.

A different approach is proposed by Wang et al. [21], in which a *Regression Activation Map* (RAM) is added after the global averaging pooling (GAP) layer of the CNN. By adding such layer to the architecture, it is possible to identify the discriminative features that lead to the classification for a sample. A GAP layer is employed instead of using traditional fully-connected layers, in order to enable the RAM layer to function properly. Within the GAP layer, each feature map in the last layer of the CNN is transmuted to a scalar, while the RAM layer computes the weighted sum of the feature maps. The model was trained on the EyePACS dataset anch achieved a Quadratic Weighted Kappa score of 85%.

Shankar et al. [22] propose an approach based on *Synergic Deep Learning* (SDL). Their model is composed of three different elements: an input layer, a number of CNN components, and a Synergic Network at the end. After preprocessing, histogram-based segmentation is carried out in order to extract the salient regions from each sample. Subsequently, the samples are fed to the Synergic Network, which is in charge of performing the classification. This approach aims to address the problem of having samples acquired through a variety of devices by making use of multiple Deep Convolutional Neural Networks that are trained simultaneously and are able to learn mutually from each other. The role of the Synergic Network in the final stage is that of determining whether the k outputs from the k DCNNs are all samples from the same class [23].

Costa et al. [24] show an approach to Diabetic Retinopathy detection that exploits the *bag-of-visual-words* (BoVW) method. The proposed model is composed of two different neural networks that are able to work jointly. Following this approach, the images are first scanned for feature extraction, then a visual dictionary is learned by the model. Such dictionary is then used to create mid-level representation of the samples to be used for the final training of the classifier. While such steps are usually treated as separate problems, the architecture proposed by the authors aims to perform the four tasks jointly. The only evaluation metric provided for this model is the Area-Under-the-Curve (AUC) measure, which reached 90% for the Messidor dataset.

Vision Transformers (ViTs)

A more recent approach to image classification comes with the rise in popularity of **Vision Transformers**. A Transformer is a deep neural network which is heavily reliant on its self-attention mechanism [25]. This type of architecture is capable of achieving a similar or better performance compared to traditional CNNs, although the two are quite often used

in conjunction. Within the context of retinal image classification, however, the model of interest is **ViT** [35]. This specific architecture is regarded as a *Pure Transformer*, as it directly applies a Transformer to sequences of image patches.

The method proposed by Wu et al. [36] achieves an accuracy of 91.4%, specificity of 0.977, sensitivity of 0.926 and area under curve (AUC) of 0.986. The model (ViT) was trained to distinguish between the 5 traditional disease grades, and its structure is adapted from the work of *Dosovitskiy et al.* [35]. The retinal fundus images are split into non-overlapping patches and then converted into sequences through flattening and embedding operations. Such resulting sequences are then input to a series of multi-head attention layers. The model is trained on the EyePACS dataset, which undergoes preprocessing and data augmentation as usual.
Table 2.1 provides a summary of the methods analysed so far and provides an easy way to compare the reported performance.

Ref	Year	Dataset	Architecture	Rep. Accuracy	Rep. Sensitivity	Rep. Specificity	Rep. AUC	Rep. QWK
[18]	2017	EyePACS	Inception-v3	-	88.3%	98.6%	-	-
[19]	2020	EyePACS, IDRID, Messidor	-	98.6%	99.1%	99.1%	-	98.1%
[20]	2021	EyePACS	LeNet	98%	92.20%	95.10%	-	-
[21]	2018	EyePACS	-	-	-	-	-	85%
[22]	2020	Messidor	ResNet-50	99.28%	98.54%	99.38%	-	-
[24]	2017	Messidor	Conv BoVW	-	-	-	90%	-
[36]	2021	EyePACS	ViT	91.4%	92.6%	97.7%	98.8%	93.5%

Table 2.1: Summary of reviewed methods for DR Detection

2.1.3 Explainable AI

In recent years some work has been carried out towards building explainable Diabetic Retinopathy detection systems. This section provides a review of some of the available literature on the topic, both to in the form of a general introduction to Explainable AI (XAI), and its applications to the domain of interest.

As Muddamsetty et al. [26] explain, applying ML algorithms to a medical domain can result in either:

- **Over reliance** on the model: the predictions made by the model are considered to be correct and they are not properly corroborated by medical professionals.
- **Under reliance** on the model: there is scepticism towards the model performance, and correct predictions are ignored.

Explainable AI (XAI) techniques are therefore aimed at increasing trust in the model, so that application in a clinical settings can effectively take place, and they do so in different ways depending on the domain of interest. Three approaches to evaluate XAI techniques can be identified [11]: Application-Grounded, Human-Grounded, and Functionally-Grounded

evaluation. As pointed out by the Doshi et al. [11], Application-Grounded evaluation is the approach that addresses a real application involving human patients, and it is therefore the methodology adopted for explainability in the medical domain. Following this blueprint, outputs from the model have to be evaluated by a human domain-expert.

One of the most widely used algorithms for obtaining visual explanations from CNN classifiers is *Grad-CAM* [27]. This technique makes use of the gradients generated within a CNN execution on a sample target as they reach the final convolutional layer. This allows for the display of a heatmap which highlights the salient regions within the sample. The novelty introduced with *Grad-CAM* is that is it class-discriminative, meaning that is is able to highlight the correct features corresponding to the predicted class despite any possible class conflict. *Figure 2.3* shows an example of Grad-CAM as applied to fundus images. Technical details about this technique will be provided in *Chapter 3*.

Muddamsetty et al. propose *SIDU* (similarity difference and uniqueness) [28], a gradient-free method to visually identify salient regions in a sample image. Differently from the previously cited *Grad-CAM*, *SIDU* operates exclusively on the last activation layer of a Convolutional Neural Network. First, each feature activation map is converted to a feature activation mask through bi-linear interpolation and point-wise multiplication operations, then the algorithm computes probability prediction scores for both the activation mask and the original sample. This is needed in order to calculate a similarity score between the two, which will then indicate the ”importance” of the corresponding feature in the classification. Together with this, a ”uniqueness” score is computed, which gives an indication as to how different a certain region within the retina is from the rest. This follows the idea that the more peculiar the region, the more salient the corresponding lesion is. *Figure 2.1* shows an example of SIDU applied to retinal images for DR detection, and it compares the performance to that of *RISE*.

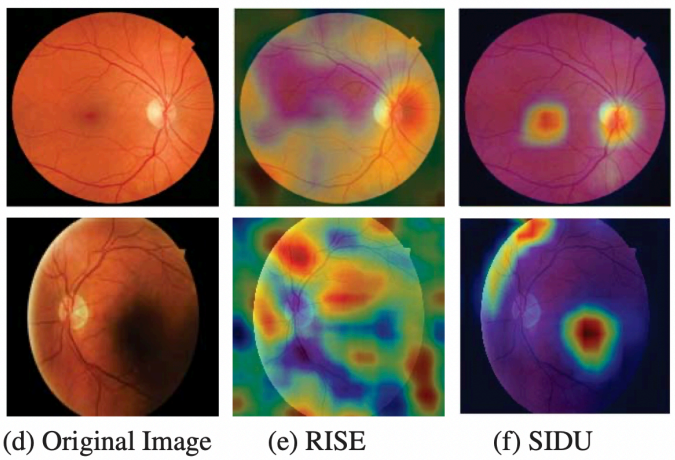


Figure 2.1: Lesion visualisation with SIDU [28]

Kind et al. [29] developed an interactive tool that allows medical professionals to visu-

alise and inspect areas of the retina deemed ”unhealthy” by the algorithm. The output offered by the model consists of a report indicating the type of lesions found in the retinal image, together with their location and visual bounding boxes that highlight them. The model proposed here is however unable to distinguish between different grades of Diabetic Retinopathy, as it is limited to identifying lesions within the retina. The opthalmologist is therefor in charge of the final diagnosis and eventual grading. *Figure 2.1* shows an example output for the model proposed, where numerous bounding boxes indicate the different types of lesions found.

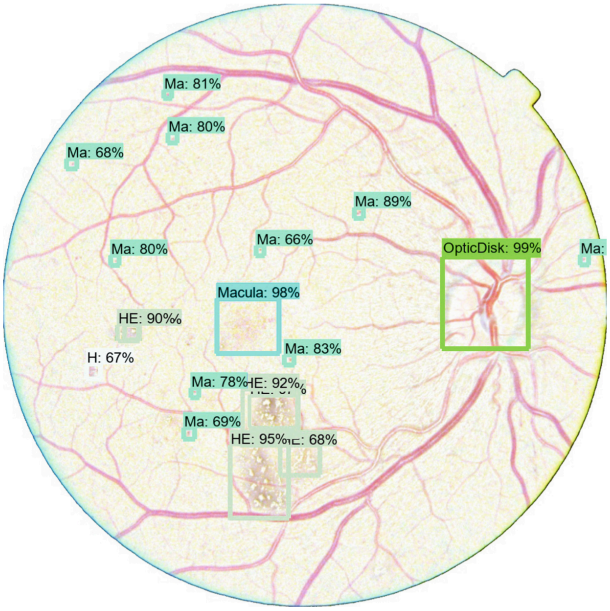
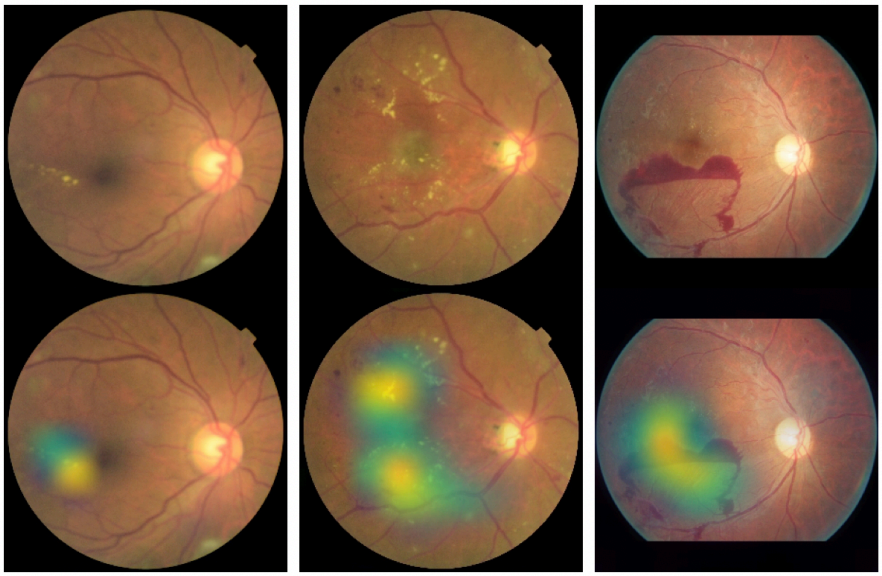


Figure 2.2: Classification output from the model proposed by Kind et al. [29]

Chetoui et al. [30] propose the use of EfficientNET to obtain a visually interpretable result, consisting of heatmaps superimpressed over the original images. To achieve such results, the author employed Grad-CAM together with fine-grained visualisations. The model achieved an AUC score of 98.4%, sensitivity of 91.7%, and specificity 98.9%. *Figure 2.2* shows an example of the output from the model, including the heatmaps obtained through the application of Grad-CAM to EfficientNet.



(a) Moderate (b) Severe (c) Proliferative

Figure 2.3: Classification output from the model proposed by Chetoui et al. [30]

Quellec et al. [31] propose *ExplAIn*, a new explanatory framework that is trained end-to-end, instead of being applied to the final layers of a CNN. The aim of this new approach is to make the classification process directly understandable by the end user, who will then ideally gain enough trust in the system for it to be integrated in a real-life clinical setting. Training images, which originate from the EyePACS and OPHDIAT datasets, undergo several preprocessing steps before being fed as input to *ExplAIn*. Data augmentation is performed online in the training process, and the final model achieves an accuracy of 74.76% and a QWK score of 86.73%. *Figure 2.3* shows an example output for *ExplAIn*. The color-coded maps next to the fundus images highlight the retinal structures that determine the diagnosis, with each color corresponding to a different type of lesion.

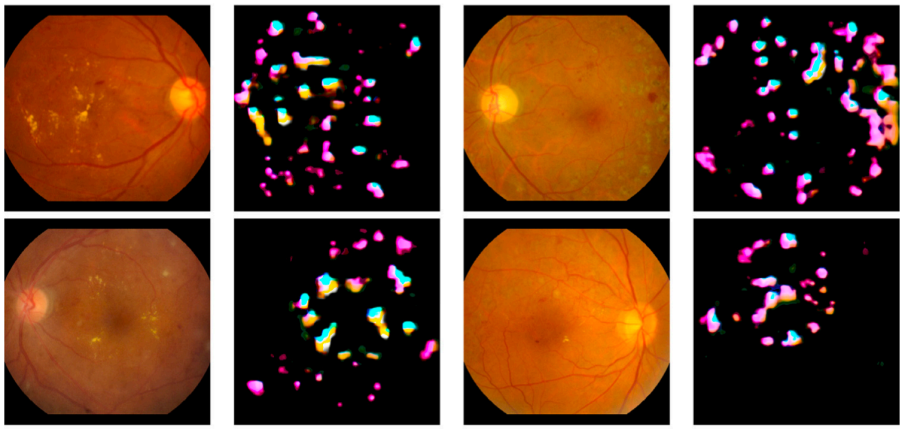


Figure 2.4: Classification output from *ExplAIn*. Source: [31]

2.2 Datasets

There are quite a few publicly available datasets which can be used to train DR detection models, some of which also provide grading for the disease and a risk index for developing Macular Edema. *Table 2.2* shows a summary of the characteristics of each dataset. In order for the grading information to be reliable, all images have been subject to evaluation by multiple opthalmologists.

The first two datasets worth of mention are the *Kaggle EyePACS* [13] and the *Kaggle APTOS 2019* [12] datasets, which were both assembled in the context of challenges held by Kaggle for the detection of Diabetic Retinopathy.

Both the Kaggle datasets provide high-quality data for the model to be trained on, with the only limitation being the absence of labels for the test sets. The *EyePACS* dataset is the largest available, with over 80.000 samples distributed across 5 different disease grading levels. Uniquely from the rest, this dataset provides specimens of both eyes for each patient. Additionally, different devices were used to capture the images, contributing to the variability of the dataset. This has the potential to enhance the flexibility of the final model, allowing it to perform well on a wide variety of inputs. In order to be used for training, this dataset requires some attentive preprocessing.

Similarly, the *Aptos 2019* dataset provides DR grading across 5 different levels. The dataset was put together by Aravind Eye Hospital in India, with the hope to aid detection of Diabetic Retinopathy in rural areas. Like in the EyePACS dataset, the images also present noise and variability due to the different devices used for acquisition, but this time only one fundus image is available for each patient.

The *Messidor* and *Messidor-2* datasets were put together by aggregating acquisitions from different sources within the Messidor Consortium in France, within the context of a project funded by the French Ministry of Research and Defense. Specifically, the images within the dataset come from three different opthalmologic departments, each using different equipment for the acquisition., which ensures variability in the data. Differently from the previous ones, the Messidor dataset only provides grading on four levels of disease, but it does provide the additional risk index for Macular Edema. However, grading on 5 levels was provided by a third party for the Messidor-2 dataset, alongside the risk index for ME.

The *DDR* dataset provides fundus images with image-level, pixel-level and bounding-box-level annotations [14]. Additionally tho the previous ones, this dataset also provides an annotation for images that were deemed ungradable.

Finally, the *ODIR* dataset comes from another Kaggle Competition and it is structured in a different way from the rest. Differently from the rest, it does not exclusively deal with Diabetic Retinopathy, but it contains sample images related to Glaucoma, Cataract, Age

related Macular Degeneration, Hypertension, Pathological Myopia, other than a general class of abnormal samples [15]. The dataset is rather small (about 5000 samples), but offers a good starting point for the development of a multi-disease classifier.

Ref	Dataset	No. Samples	Features
-	Kaggle EyePACS	> 80.000	Fundus images for DR detection (graded 0-4)
-	Kaggle APTOS 2019	5590	Fundus images for DR detection (graded 0-4)
-	MESSIDOR	1200	Fundus images for DR detection (graded 0-3) and assessment of ME risk (graded 0-2)
-	MESSIDOR-2	1748	Fundus images for DR detection (graded 0-4) and assessment of ME risk (graded 0-2)
-	DDR	12522	Fundus images for DR detection (graded 0-5)
-	ODIR	5000	Fundus images for detection of: Diabetes (D), Glaucoma (G), Cataract (C), Age related Macular Degeneration (A), Hypertension (H), Pathological Myopia (M), Other diseases/abnormalities (O)

Table 2.2: Publicly available datasets for DR Detection

2.3 Evaluation Metrics

A variety of different evaluation metrics are used in current literature to measure model performance. The following sections give a more detailed explanation of the most commonly used metrics for Diabetic Retinopathy detection models.

2.3.1 Accuracy

Accuracy is the most intuitive and commonly used metric. It measures the percentage of correct classifications made by the model. The corresponding formula is:

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

where:

- TP (True Positive Rate): number of positive observations classified as positive by the model
- TN (True Negative Rate): number of negative observations classified as negative by the model
- FP (False Positive Rate): number of negative observations classified as positive by the model
- FN (False Negative Rate): number of positive observations classified as negative by the model

2.3.2 Specificity

Specificity indicates the proportion of negatives that were predicted as positives by the classifier. This evaluation metric is especially relevant in clinical settings, since incorrectly labelling a patient as 'healthy' could constitute a major risk.

$$Specificity = \frac{TN}{(TN + FP)}$$

2.3.3 Sensitivity

Sensitivity offers information regarding how many positive instances were correctly identified by the model.

$$Sensitivity = \frac{TP}{(TP + FN)}$$

2.3.4 AUC Score

The Area Under the Curve Score, or AUC, represents the ability of the model to distinguish between classes. When this score is high, it indicates that the model has a high probability of correctly classifying each sample. Geometrically, the AUC Score is defined as the Area under the ROC curve, which can be visualised as the projection of the FP Rate against the TP Rate. *Figure 2.1* shows a visualisation of the ROC curve for different classifiers, from one that performs a "random" guess, to one with perfect performance.

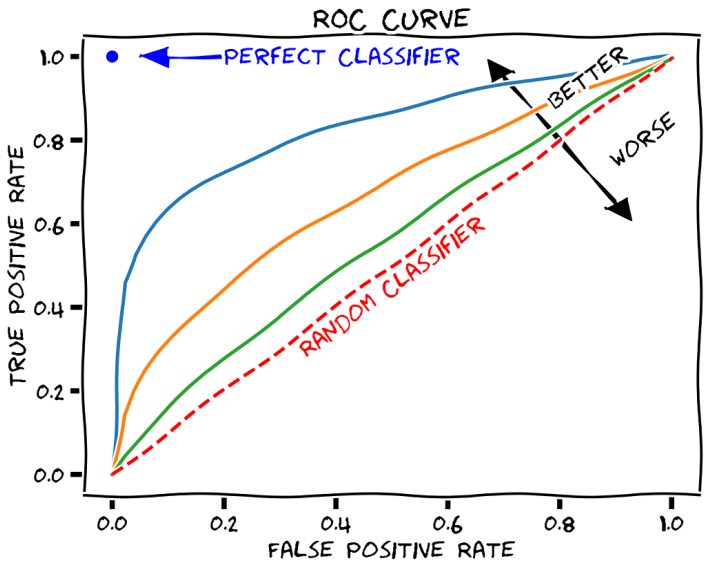


Figure 2.5: ROC Curve for classifier performance. Source: [32]

2.3.5 F-Score

The F-Score (or F1-Score) is a statistical measure defined as the harmonic mean of precision and recall. Mathematically, it is calculated as:

$$FScore = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

2.3.6 Quadratic Weighted Kappa (QWK)

The Quadratic Weighted Kappa (QWK) metric measures the degree to which two rating agree with each other. In the case of a classifier, it gives an indication as to how much better the model is performing compared to a random classifier [33].

It is calculated as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

where O is an NxN histogram matrix, such that $O_{i,j}$ equals the actual number of samples that received a prediction, and:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

where E is an NxN histogram matrix of expected values.

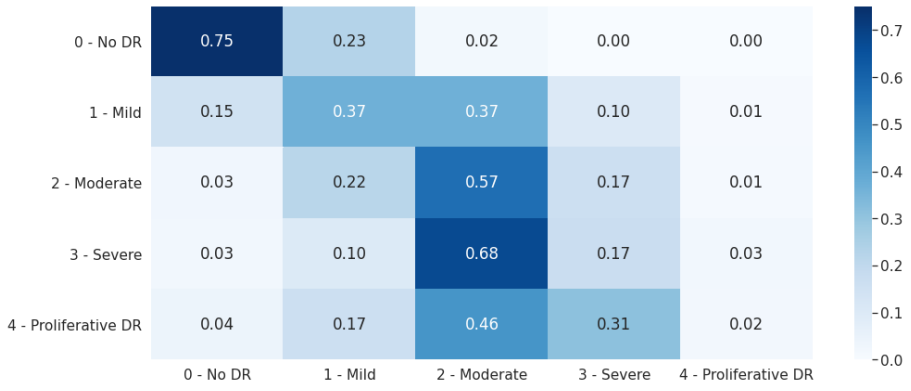
3 Technical Chapters (change this to something appropriate)

3.1 Model Replication

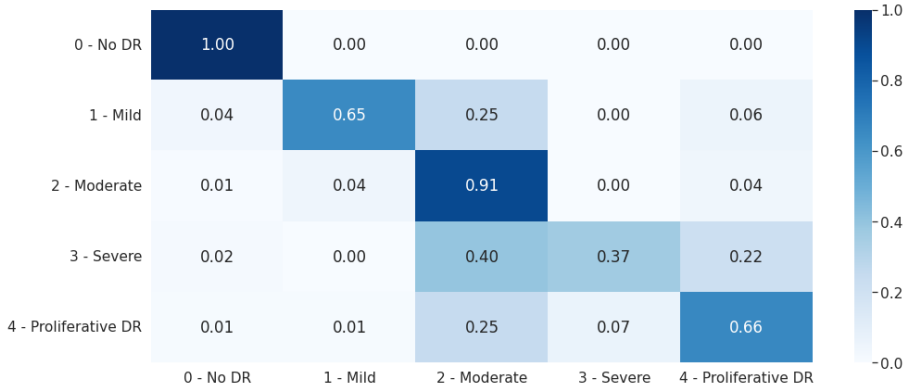
A few already available implementations have been replicated in order to corroborate the advertised results and to build a benchmark for evaluation. For efficiency purposes, the dataset used was the Kaggle APTOS 2019 for all the replicated models, instead of the more widely used EyePACS dataset. In most cases, alterations to the code have been made in order to fix dependencies, tailor the model to a different dataset than the one it was originally trained on, or adjust the training, validation, and test split. The alterations made to the original implementation will be specified, and a comparison of the results will follow. Wherever it was not already implemented, explainability methods were integrated to obtain a visual result of the application of the model to a sample.

3.1.1 ResNet 50

The CNN method proposed by Tymchenko et al. [19] was replicated, as implemented by *Debayan Mitra* and made available on GitHub [34]. The test accuracy score achieved was 0.872, which is considerably lower than the accuracy advertised by the authors, which is 0.928. However, the code implementation used for the replication advertises a Test Accuracy of 0.576.
The model was trained on the APTOS 2019 Dataset [12], although the test set was derived from the set of training images due to the lack of labels of the provided test set. This reduced the amount of samples in the training, validation, and test sets. Although in the original paper the main evaluation metric is the Kappa score, since the model is evaluated on the unlabeled test set, we only referred to the accuracy metric for comparison. Before training, the images go through a preprocessing stage which consists of cropping and resizing the samples. In addition, data augmentation was used to lower correlation between features and avoid overfitting.



(a) Original model



(b) Replicated model

Figure 3.1: Side by side confusion matrix for model evaluation

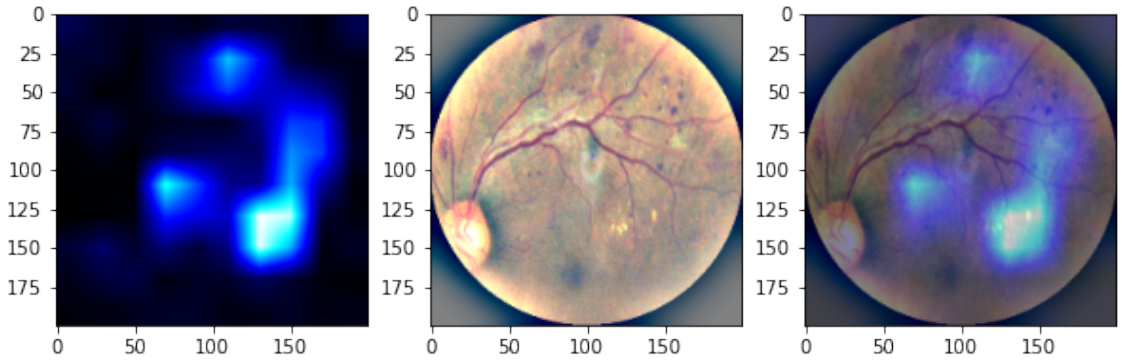


Figure 3.2: GradCAM applied to a sample image classified by the method proposed by [19], as implemented by [34]

3.2 Exploratory Data Analysis

3.3 Image Preprocessing

3.4 Model Definition

3.5 Hyperparameter Tuning

4 Conclusion

4.1 Summary

Summarise what you have achieved.

4.2 Evaluation

Stand back and evaluate what you have achieved and how well you have met the objectives. Evaluate your achievements against your objectives in section **??**. Demonstrate that you have tackled the project in a professional manner.

4.3 Future Work

Bibliography

[1] World Health Organization, *World report on vision*, Geneva, 2019

[2] D. Shu Wei Ting, G. Chui Ming Cheung, T.Y. Wong, *Diabetic retinopathy: global prevalence, major riskfactors, screening practices and public health challenges: a review*, Clinical and Experimental Ophthalmology, 44, pp. 260–277, 2016

[3] W. Wang, A.C. Lo, *Diabetic retinopathy: pathophysiology and treatments*, International journal of molecular sciences, 19(6), p.1816, 2018

[4] V.M. Kanukollu, S.S. Ahmad, *Retinal Hemorrhage* (Updated 2022 May 5), StatPearls Publishing, 2022, https://www.ncbi.nlm.nih.gov/books/NBK560777/

[5] S. Júnior, D. Welfer, *Automatic Detection of Microaneurysms and Hemorrhages in Color Eye Fundus Images*, International Journal of Computer Science and Information Technology, 5, pp.21-37, 2013

[6] Early Treatment Diabetic Retinopathy Study Research Group, *Grading diabetic retinopathy from stereoscopic color fundus photographs — an extension of the modified Airlie House classification: ETDRS report number 10*, Ophthalmology 98, 5, pp. 786-806, 1991

[7] L. Wu, P. Fernandez-Loaiza, J. Sauma, E. Hernandez-Bogantes, E. Masis, *Classification of diabetic retinopathy and diabetic macular edema*, World journal of diabetes, 2013

[8] K. Solanki, C. Ramachandra, S. Bhat, M. Bhaskaranand, M.G. Nittala, S.R. Sadda, *EyeArt: automated, high-throughput, image analysis for diabetic retinopathy screening. Investigative Ophthalmology Visual Science*, 56(7), pp.1429-1429, 2015, https://www.eyenuk.com/en/products/eyeart/, accessed 20/07/2022

[9] M. Bhaskaranand, C. Ramachandra, S. Bhat, J. Cuadros, M.G. Nittala, S.R. Sadda, K. Solanki, *The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes*, Diabetes technology therapeutics, 21(11), pp.635-643, 2019

[10] A. Tufail, C. Rudisill, C. Egan, V.V. Kapetanakis, S. Salas-Vega, C.G. Owen, A. Lee, V. Louw, J. Anderson, G. Liew, L. Bolter, *Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders*, Ophthalmology, 124(3), pp.343-351, 2017

[11] F. Doshi-Velez, B. Kim, *Considerations for evaluation and generalization in interpretable machine learning*, Explainable and interpretable models in computer vision and machine learning (pp. 3-17), Springer, Cham, 2018

[12] Aravind Eye Hospital, *APTOS 2019 blindness detection*, https://www.kaggle.com/c/aptos2019-blindness-detection, Accessed 10/06/2022

[13] EyePACS, *Diabetic Retinopathy Detection*, https://www.kaggle.com/c/diabetic-retinopathy-detection, Accessed 10/06/2022

[14] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, H. Kang, *Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening*, Information Sciences, 501, pp.511-522, 2019

[15] Shanggong Medical Technology Co., Ltd, *Ocular Disease Intelligent Recognition (ODIR)*, https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k, Accessed 10/06/2022

[16] C. Hernandez-Matas, A.A. Argyros, X. Zabulis, *Retinal image preprocessing, enhancement, and registration*, Computational Retinal Image Analysis, pp 59-57, 2019

[17] J.V. Soares, J.J. Leandro, R.M. Cesar, H.F. Jelinek, M.J. Cree, *Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification*, IEEE Transactions on medical Imaging, 25(9), pp.1214-1222, 2006

[18] D.S. Ting, C.Y. Cheung, G. Lim, G.S. Tan, N.D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I.Y. San Yeo, S.Y. Lee, E.Y. Wong, *Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes*, Jama, 2017

[19] B. Tymchenko, P. Marchenko, D. Spodarets, *Deep learning approach to diabetic retinopathy detection*, arXiv preprint, 2020

[20] I. Qureshi, J. Ma, Q. Abbas, *Diabetic retinopathy detection and stage classification in eye fundus images using active deep learning*, Multimedia Tools and Applications, 2021

[21] Z. Wang, J. Yang, *Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation*, Workshops at the thirty-second AAAI conference on artificial intelligence, 2018

[22] K. Shankar, A.R. Sait, D. Gupta, S.K. Lakshmanaprabu, A. Khanna, H.M. Pandey, *Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model*, Pattern Recognition Letters, 2020

[23] J. Zhang, Y. Xie, Q. Wu, Y. Xia, *Medical image classification using synergic deep learning*, Medical image analysis, 54, pp. 10-9, 2019

[24] P. Costa, A. Campilho, *Convolutional bag of words for diabetic retinopathy detection from eye fundus images*, IPSJ Transactions on Computer Vision and Applications, 2017

[25] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, *A survey on vision transformer*, IEEE transactions on pattern analysis and machine intelligence, 2022

[26] S.M. Muddamsetty, M.N. Jahromi, T.B. Moeslund, *Expert level evaluations for explainable AI (XAI) methods in the medical domain*, International Conference on Pattern Recognition, (pp. 35-46), Springer, Cham, 2021

[27] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, *Grad-cam: Visual explanations from deep networks via gradient-based localization*, Proceedings of the IEEE international conference on computer vision (pp. 618-626), 2017

[28] S.M. Muddamsetty, N.J. Mohammad, T.B. Moeslund, T.B., *SIDU: similarity difference and uniqueness method for explainable AI*, 2020 IEEE International Conference on Image Processing (ICIP) (pp. 3269-3273), IEEE, 2020

[29] A. Kind, G. Azzopardi, An explainable AI-based computer aided detection system for diabetic retinopathy using retinal fundus images, International Conference on Computer Analysis of Images and Patterns, (pp. 457-468), Springer, Cham, 2019

[30] M. Chetoui, M.A. Akhloufi, *Explainable diabetic retinopathy using EfficientNET*, 2020 42nd annual international conference of the IEEE engineering in Medicine Biology Society (EMBC) (pp. 1966-1969), IEEE, 2020

[31] G. Quellec, H. Al Hajj, M. Lamard, P.H. Conze, P. Massin, B. Cochener, *ExplAIn: Explanatory artificial intelligence for diabetic retinopathy diagnosis*, Medical Image Analysis, 72, p.102118, 2021

[32] Martin Thoma, *Receiver Operating Characteristic (ROC) curve with False Positive Rate and True Positive Rate*, https://commons.wikimedia.org/wiki/File:Roc-draft-xkcd-style.svg, accessed 22/07/2022

[33] G. Hongnan, *Understanding the Quadratic Weighted Kappa*, https://www.kaggle.com/code/reighns/understanding-the-quadratic-weighted-kappa, accessed 27/07/2022

[34] Debayan Mitra, *Blindness detection (Diabetic retinopathy) using Deep learning on Eye retina images*, https://towardsdatascience.com/blindness-detection-diabetic-retinopathy-using-deep-learning-on-eye-retina-images-baf20fcf409e, accessed 02/07/2022

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, S, J. Uszkoreit, *An image is worth 16x16 words: Transformers for image recognition at scale*, arXiv preprint arXiv:2010.11929, 2020

[36] J. Wu, R. Hu, Z. Xiao, J. Chen, J. Liu, *Vision Transformer-based recognition of diabetic retinopathy grade* Medical Physics, 48(12), pp.7850-7863, 2021

Appendix 1

You may have one or more appendices containing detail, bulky or reference material that is relevant though supplementary to the main text: perhaps additional specifications, tables or diagrams that would distract the reader if placed in the main part of the dissertation. Make sure that you place appropriate cross-references in the main text to direct the reader to the relevant appendices.

Note that you should **not** *include your program listings as an appendix or appendices.* You should submit one copy of such bulky text as a separate item, perhaps on a disk.

Appendix 2 – User guide

If you produced software that is intended for others to use, or that others may wish to extend/improve, then a user guide and an installation guide appendices are **essential**.

Appendix 3 – Installation guide

If you produced software that is intended for others to use, or that others may wish to extend/improve, then a user guide and an installation guide appendices are **essential**.