PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
الجمهورية الجزائرية الديمقراطية الشعبية
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
وزارة التعليم العالي و البحث العلمي

University of Djilali BOUNAAMA Khemis Miliana
جامعة الجيلالي بونعامة خميس مليانة
Faculty of Science and Technology
كلية العلوم والتكنولوجيا
Department of Mathematics and Computer Science
قسم الرياضيات والإعلام الآلي

Master's Thesis

For the attainment of the Master's degree in Computer Science

Option: Software Engineering and Distributed Systems

Preprocessing of Retinal Images Combined with Deep Learning for Diabetic Retinopathy Classification

Realized by:
Imane Mezouaghi

Thesis Supervisor:
Mr. O. Boukadoum

Date of Defense: 2023. 07. 20, in front of the Thesis Defense Committee:

Mr. N. Azzouza:	President
Mr. A. Khalfi:	Examiner
Mr. O. Boukadoum:	Supervisor

Academic year: 2022/2023.

Dedications

“

To my mother, thank you for your unwavering love, guidance, and sacrifices. Your nurturing nature and endless support have shaped me into the person I am today.

To my father, thank you for your strength, wisdom, and constant encouragement. Your belief in me has given me the confidence to chase my dreams and overcome any obstacles that come my way.

*To my sister **Nabila**, thank you for being my confidante, my best friend. Your presence brings joy and laughter into my life, and I cherish the bond we share.*

To my brothers, thank you for your camaraderie, protection, and shared experiences. Our bond is unbreakable, and I am grateful for the memories we have created together.

To my friends, thank you for being my pillars of support, my source of laughter, and my chosen family. Your presence has brought light and positivity into my life, and I am grateful for the countless memories we have made together.

With heartfelt gratitude,

”

- Imane

Acknolodgments

First and foremost, I would like to thank God for His blessings, guidance, and unwavering support throughout this journey. His divine grace has provided me with strength, wisdom, and perseverance to overcome challenges and achieve this milestone.

I would like to thank my supervisor, **Mr. Boukadoum Omar**, for their guidance, support, and valuable insights throughout the research process. Their expertise and encouragement have been instrumental in shaping this thesis.

I am also grateful to the faculty members of University DJilali Bounaama for their valuable teachings and mentorship during my academic journey. Their dedication to education and research has been truly inspiring.

I am grateful to the members of the jury for their time and expertise in evaluating this thesis.

Lastly, I would like to express my gratitude to all the individuals, colleagues, and research peers who have provided assistance, feedback, and constructive discussions throughout this thesis.

This research would not have been possible without the collective efforts and contributions of all these individuals, and for that, I am sincerely thankful.

Abstract

Diabetic retinopathy (DR) is a prevalent medical condition that affects individuals with diabetes mellitus, leading to potential vision loss or even blindness if left untreated. Accurate and timely classification of DR is crucial for effective diagnosis and intervention.

This thesis explores the application of deep learning techniques, specifically the Vision Transformer (ViT) model, for the classification of DR. The study investigates both binary and multi-class classification tasks, aiming to develop robust models capable of accurately distinguishing between different stages of DR. The thesis encompasses comprehensive data preprocessing, model training, and evaluation processes to achieve optimal performance.

The results obtained exhibit excellent classification accuracy, demonstrating the potential of deep learning in enhancing DR diagnosis. The findings highlight the significance of leveraging ViT models to improve the accuracy and efficiency of DR classification, offering promising advancements in medical image analysis. This research contributes to the field by providing valuable insights into the use of deep learning for DR classification and paves the way for future developments in automated DR diagnosis systems.

In this project, we have utilized the Vision Transformer (ViT) architecture with various parameters to classify diabetic retinopathy on two different datasets with diffrent structurs.

Keywords : Healthcare, Health care Diabetes, Diabetic retinopathy, Artificial intelligence, Machine learning, Deep learning.

Résumé

La rétinopathie diabétique (DR) est une condition médicale prévalente qui affecte les personnes atteintes de diabète sucré, pouvant entraîner une perte de vision potentielle, voire la cécité en l'absence de traitement. Une classification précise et rapide de la DR est essentielle pour un diagnostic et une intervention efficaces.

Cette thèse explore l'application des techniques d'apprentissage profond, en particulier le modèle Vision Transformer (ViT), pour la classification de la DR. L'étude examine à la fois les tâches de classification binaire et multi-classes, dans le but de développer des modèles robustes capables de distinguer avec précision les différents stades de la DR. La thèse englobe un prétraitement complet des données, ainsi que des processus de formation et d'évaluation du modèle pour obtenir des performances optimales.

Les résultats obtenus montrent une excellente précision de classification, démontrant le potentiel de l'apprentissage profond pour améliorer le diagnostic de la DR. Les conclusions mettent en évidence l'importance de l'utilisation des modèles ViT pour améliorer la précision et l'efficacité de la classification de la DR, offrant ainsi des avancées prometteuses dans l'analyse d'images médicales. Cette recherche contribue au domaine en fournissant des informations précieuses sur l'utilisation de l'apprentissage profond pour la classification de la DR et ouvre la voie à de futures avancées dans les systèmes de diagnostic automatisés de la DR.

Dans ce projet, nous avons utilisé l'architecture Vision Transformer (ViT) avec différents paramètres pour classer la rétinopathie diabétique sur deux ensembles de données distincts et de structures différentes

---

Mots-clés : Santé, Diabète, Rétinopathie diabétique, Intelligence artificielle, Apprentissage automatique, Apprentissage profond.

ملخص

التشوه الشبكي السكري (DR) هو حالة طبية شائعة تؤثر على الأشخاص الذين يعانون من مرض السكري، مما يؤدي إلى فقدان الرؤية المحتمل أو حتى العمى إذا لم يتم علاجه بشكل فعال. تعتبر التصنيف الدقيق والفوري للتشوه الشبكي السكري أمرًا حاسمًا للتشخيص والتدخل الفعال.

تستكشف هذه الرسالة تطبيق تقنيات التعلم العميق، وتحديدًا نموذج Transformer Vision (ViT)، لتصنيف التشوه الشبكي السكري. تتناول الدراسة المهام المتعددة الفئات والفئات الثنائية، بهدف تطوير نماذج قوية قادرة على تمييز المراحل المختلفة للتشوه الشبكي السكري بدقة. تتضمن الرسالة معالجة شاملة للبيانات، وتدريب النماذج، وعمليات التقييم لتحقيق أفضل أداء.

تظهر النتائج تحقيق دقة تصنيف ممتازة، مما يدل على إمكانية تعزيز تشخيص التشوه الشبكي السكري بواسطة التعلم العميق. تسلط الاستنتاجات الضوء على أهمية استغلال نماذج ViT لتحسين دقة وكفاءة تصنيف التشوه الشبكي السكري، وتقدم تقدمًا واعدًا في تحليل الصور الطبية. تسهم هذه البحث في المجال عن طريق توفير رؤى قيمة حول استخدام التعلم العميق لتصنيف التشوه الشبكي السكري، وتمهد الطريق للتطورات المستقبلية في أنظمة التشخيص الآلي للتشوه الشبكي السكري.

في هذا المشروع، استخدمنا نموذج Transformer Vision (ViT) مع مجموعة متنوعة من المعاملات مختلفة لتصنيف التشوه الشبكي السكري باستخدام مجموعتي بيانات مختلفتين وذات هياكل مختلفة.

كلمات مفتاحية :

الرعاية الصحية، السكري، التشوه الشبكي السكري، الذكاء الاصطناعي، التعلم الآلي، التعلم العميق.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

IA	*Intelligence Artificielle*
ML	*Machine Learning*
DL	*Deep Learning*
ViT	*Vision Transformer*
HRF	Hemorrhages, Exudates, and Microaneurysms
ANN	*Artificial Neural Networks*
MLP	*Multilayer Perceptron*
CNN	*Convolutional Neural Network*
SVM	*Support Vector Machine*
Adam	Adaptive Moment estimation
Adaboost	Deep Learning Approach for Advanced Binary Object Segmentation Technique
RGB	Red, Green and Blue

General introduction

Contexte

In the present era, the advancement of computers and their remarkable computational capabilities has led to the implementation of various algorithms and learning techniques aimed at automating tasks and aiding human decision-making processes. These techniques find application in numerous domains, including online shopping, web search, cybersecurity, healthcare, and more. Particularly in the healthcare sector, significant advancements have been made through the automation of tasks using machine learning and deep learning approaches Davenport T 2019.

In this context, several systems have been developed to assist clinicians in disease diagnosis and decision-making processes [38]. Notably, a considerable emphasis has been placed on the extraction of clinical information from medical images using machine learning and deep learning methods.

Among the most prevalent diseases globally, diabetes affects approximately 425 million individuals “LE DIABÈTE DANS LE MONDE” 2021. Diabetic retinopathy, one of its complications, is a significant concern.

Problem statemnt

”How can deep learning techniques be effectively utilized to classify different stages of Diabetic Retinopathy (DR) from retinal images with high accuracy and reliability?”

This problem encompasses the development and optimization of deep learning models that can accurately classify DR stages, taking into account factors such as image quality, lesion complexity, and the need for efficient feature extraction. The goal is to provide an automated and reliable solution for DR classification, assisting healthcare professionals in early detection and effective management of the disease.

Objective of the work

The objective of this work is to develop a deep learning-based system for accurately classifying different stages of Diabetic Retinopathy (DR) using retinal images. By leveraging advanced deep learning techniques, the goal is to improve the accuracy and efficiency of

DR diagnosis. The specific objectives of this work include:

1. collecting and preprocessing a large dataset of retinal images.
2. Designing and implementing deep learning models, specifically utilizing Vision Transformer (ViT), for DR classification.
3. Training and optimizing the models using the collected dataset to achieve high accuracy in DR classification.
4. evaluating the performance the developed models.
5. comparing performance of the developed models with existing approaches, and providing insights for further research in this field.

The ultimate aim is to contribute to the development of computer-aided diagnosis systems that enable early detection and intervention for better patient outcomes.

Organization of the Thesis:

After this general introduction, the dissertation is structured as follows:

The first chapter **Context medical: Diabetic retinopathy** provides an introduction to the field of DR classification and its significance in healthcare. It discusses the increasing prevalence of diabetes mellitus and the associated risks of DR.It covers the anatomy of the eye and the medical classification of DR. The chapter emphasizes the importance of early detection and accurate classification for effective treatment and management of DR.

Chapter 2 **Preprocessing Mecial Images** delves into the characterization of medical images and the importance of preprocessing techniques in enhancing their quality. It explores the unique characteristics of medical images. The chapter further discusses the challenges associated with medical image processing and the employed preprocessing techniques to overcome it.

The third chapter **Deep learning** provides an overview of deep learning techniques and their applications in medical image analysis. It discusses the key components of deep learning.

The fourth chapter **Implementation and results** presents the result of our work, It explores the process of model training, including dataset preparation, data augmentation, and hyperparameter tuning. and provides a comparative section to show the novelties of this work compared to previous ones

in this last chapter **Conclusion & Prespectives**, we summarize and review our ideas and results and giving some perspectives.

Chapter 1

contexte medical: diabetic retinopathy

1.1 Introduction

This chapter offers crucial medical background information for comprehending diabetic retinopathy (DR).It describes diabetes mellitus as a chronic disease with inadequate glucose control that results in hyperglycemia. The eye's anatomy and physiology are therefore the main focus, with an emphasis on the eye's function in vision.Additionally, it gives a general overview of contemporary techniques for diagnosing and screening diabetic retinopathy, including optical coherence tomography (OCT).The clinical characteristics of diabetic retinopathy are also discussed, including how it is divided into four stages (no DR, mild, moderate, severe nonproliferative, and proliferative).It also discusses the increased risk of various health problems as well as the social and emotional effects of diabetic retinopathy.

Understanding the medical context of diabetic retinopathy is crucial for developing effective strategies to prevent, diagnose, and manage this sight-threat.

1.2 Diabetes and the Eye

1.2.1 Diabetes mellitus

Diabetes mellitus (DM) is a chronic health condition characterized by poor glucose regulation leading to hyperglycemia. Normal blood glucose levels depend (among many other important metabolic effects) on the efficient use of the peptide hormone insulin, which triggers the uptake of glucose into the cellular space.
Insulin is not used efficiently in diabetics. This is either due to underproduction in type 1 diabetes (T1DM) or peripheral tissue resistance to insulin effects in type 2 diabetes (T2DM). Long-term and/or uncontrolled elevation of blood glucose levels is a major risk factor for DM sequelae such as cardiovascular disease, neurological damage, renal disease, and eye disease including retinopathy Lazzaro 2021a

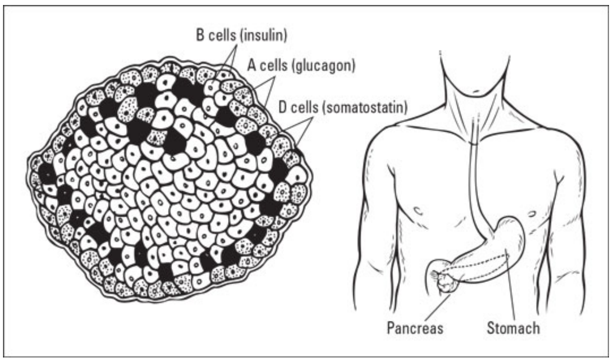


Figure 1.1: Diabetes pancreas

1.2.2 Anatomy and Physiology of the Eye

The eyes are a component of the visual system that enables vision, as well as the acquisition, processing, and photoresponse of visual information by living beings. The eyes pick up light, which causes it to be converted into electrochemical impulses in neurons.connections.Willoughby et al. 2010 There are three layers distinguished in the eye organ: in the outer region (the protective layer): cornea and selca, the middle layer (the vascular layer): uvea, and finally receptor layer: the retina.

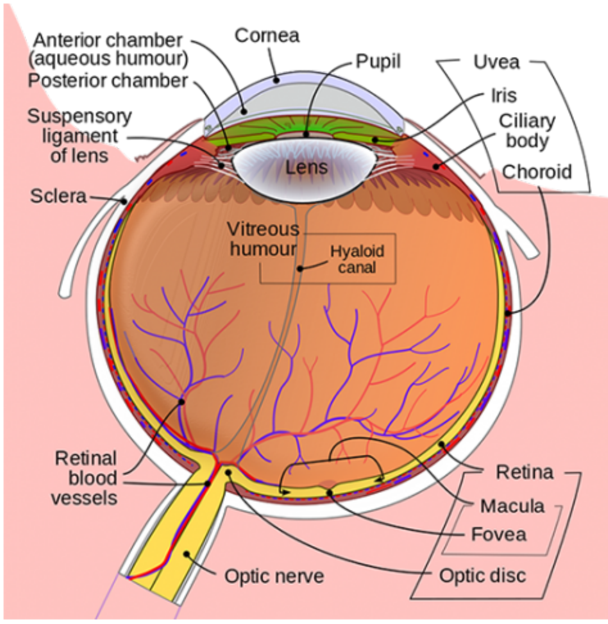


Figure 1.2: Schematic$_{d}iagram_{o}f_{t}he_{h}uman_{e}ye$

- Light enters the eye through the cornea, the clear, curved layer in front of the iris and pupil.

- The cornea serves as a protective covering for the front of the eye and also helps focus light on the retina at the back of the eye.Ophthalmology 2007a

- After passing through the cornea, light travels through the pupil (the black dot in the middle of the eye).The iris—the circular, colored area of the eye that surrounds the pupil—controls the amount of light that enters the eye.Ophthalmology 2007b

- Thus, the pupil dilates and constricts like the aperture of a camera lens as the amount of light in the immediate surroundings changes. The size of the pupil is controlled by the action of the pupillary sphincter muscle and dilator muscle.Willoughby et al. 2010

- Behind the iris sits the lens. By changing its shape, the lens focuses light onto the retina.Ophthalmology 2007c

- **Retina:** The light-sensitive layer of tissue that lines the back of the eye, The retina contains the cells that sense light (photoreceptors) and the blood vessels that nourish them. sclera, and cornea in the outer region surrounds the retina and protects it. Retina serves a function comparable to a camera film because it contains highly specialized receptors which process what we see and send impulses through the optic nerve to the brain to create visual perception.Ophthalmology 2007b

- The macula, a small region of the retina with millions of densely populated cone-shaped photoreceptors, is the most sensitive section of the retina. The macula's dense cone population gives the visual image detail.Willoughby et al. 2010

- The optic nerve and other cranial nerves, which carry impulses from the retina to the brain, travel through the orbit, the bony cavity that surrounds the eyeball.Ophthalmology 2007b

- The blood vessels within the retina itself that carry oxygen and nutrients are called arteries. Superior and inferior branches emerge from the central retinal artery, which passes through the optic nerve to enter the eye.

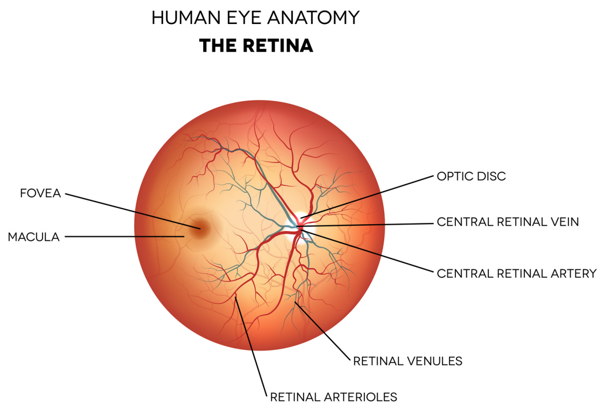


Figure 1.3: Retina anatomy

1.3 Overview of the current approaches for diabetic retinopathy screening and diagnosis

1.3.1 Diabetic retinopathy

Diabetic retinopathy is the leading cause of new-onset blindness in those aged 20 to 74. Almost every T1D patient and more than 60The retina's blood supply is cut off due to a blockage of the tiny blood capillaries that nourish it. As a result, the eye attempts to form new blood vessels. However, these new blood vessels do not form properly and can result in increased vascular permeability. This permeability allows edema and hemorrhage to spread into the retina or vitreous fluid. Lazzaro 2021b

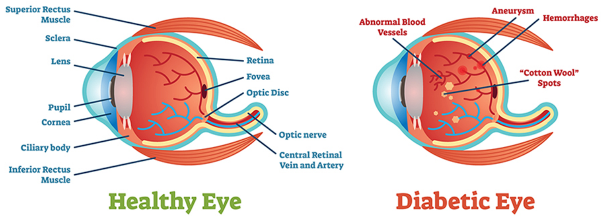


Figure 1.4: Diffrence between diabtic Eye and Healthy Eye

1.3.2 screening of DR

Individuals with diabetes should be properly screened for signs of DR and the progression stages of DR due to the asymptomatic nature of DR. Screening guidelines recommend an annual dilated fundus examination for T2 and T1 diabetes at the earliest time point at which intervention could reduce morbidity. People with obvious DR require more frequent follow-ups to determine if treatment is indicated, which necessitates more exams as disease severity worsens. Screening for and treating DR on time reduces the likelihood of developing severe vision loss by up to 94% and is also extremely cost-effective. Lazzaro 2021c According to studies, there is a lack of committing to guidelines, with nearly half of the people with diabetes not receiving an eye exam regularly.Kumari et al. 2022

1.3.3 diagnosis of DR

Retinal imaging :

Figure 1.5: retinal imaging

Retinal imaging helps optometrists to better analyze the health of the eye by photographing the back of the eye specifically the retina with a high-resolution camera. Ophthalmic medical technology has witnessed rapid-fire development in recent years, new technologies have appeared that are not only able to assist in a better understanding of the disease and give early diagnosis but also to advance the treatments and therapy. an optometrist can inspect the eye using a variety of techniques including what is called OCT

Optical coherence tomography (OCT):

Figure 1.6: NIDEK RS 3000 Advance Optical Coherence Tomography

is a solution for physicians and clients in this matter, OCT has revolutionized ophthalmology rapidly, the statistics say that in 2012 more than 30 million OCT exams took place, this technology studies diabetic retinopathy in more depth and clarifies more details in all areas of the disease, it examination highlights the retinal regions making it possible to follow its evolution, determine what caused the vision problem, decide the treatment method...etc

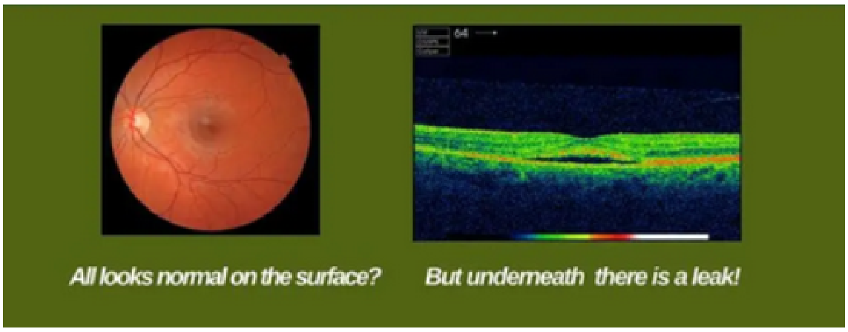


Figure 1.7: OCT Retinal imaging

1.4 Review of the clinical features of diabetic retinopathy

1.4.1 Diabetic Retinopathy Classification

In addition to stage No DR, The four diabetic retinopathy stages are classified as mild, moderate, severe nonproliferative, and proliferative.

NO DR:
At this stage, we detect the absence of microaneurysms, exudates, hemorrhages, and other indicators of retinal blood vessel injury. This, of course, can only be approved and witnessed with screening, which is why it is vital to screen for diabetic retinopathy even

in people who do not have DR or show no indicators of it, particularly those in the initial years after being diabetic.

Figure 1.8: healthy fundus image

MILD:
This stage is characterized by balloon-like swellings in particular zones of the retinal blood vessels called microaneurysms. Even though this stage rarely influences vision or requires treatment, it does show a tall hazard of diabetic harm and disease progression. These patients ought to have a widened eye examination every 12 months. There's a 5'u hazard that mild DR will advance to PDR within 1 year. Ayman S. El-Baz n.d.(a) Ayman S. El-Baz n.d.(b)

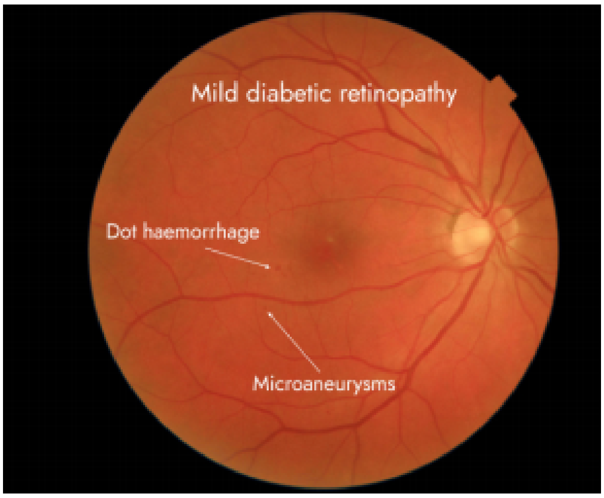


Figure 1.9: Mild DR

Moderate:
The next advanced stage of DR is the presence of numerous small bulges in the blood

vessels known as microaneurysms (as indicated in 1.10) and retinal hemorrhages in the retina. Hemorrhages are the blood spots that form on the retina as a result of blood leaking from blood vessels that have been damaged. These blotches are transient and vanish with time. Despite the fact that this stage is not life-threatening, it does induce pericyte loss in the blood vessels, which the patient can feel, but no further therapy can be started until the next test. Exudates that could leak fluid into the retina may also be visible to patients. Ayman S. El-Baz n.d.(b)

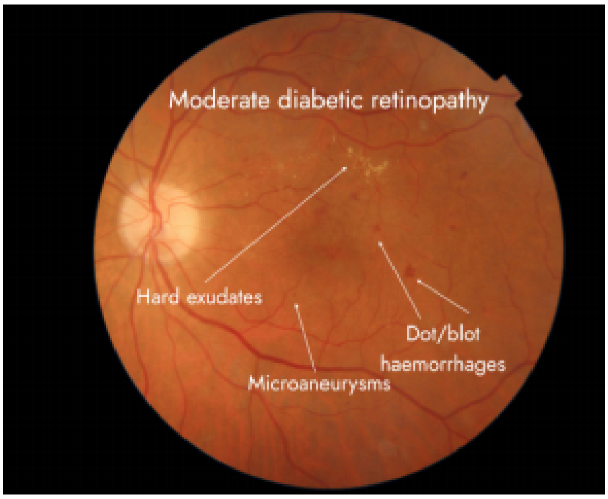


Figure 1.10: Moderate DR

Severe

During this stage Deep round hemorrhages develop as additional blood vessels are blocked, causing pericyte damage to the blood vessels. Widespread retinal hemorrhages and NFL hemorrhages are also present . Numerous layers of the retina are affected by diffuse bleeding known as widespread retinal hemorrhages, which point to a serious disruption of retinal blood vessels. NFL hemorrhages, on the other hand, appear as tiny, flame- or splinter-shaped bleeding regions within the retina's nerve fiber layer. Due to the significant vascular involvement and increased blood vessel fragility in severe DR, as seen in 1.11, small new vessels trap inside the retina.Ayman S. El-Baz n.d.(c)

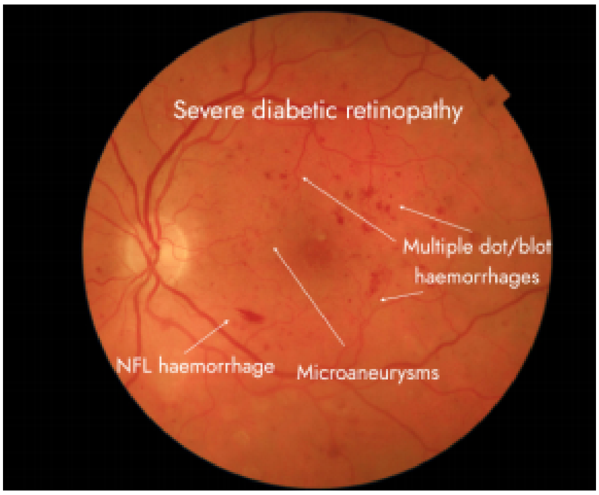


Figure 1.11: Severe DR

Proliferative

The retina is nourished by blood vessel blockages, which cause the retinal blood vascularization—a disorder known as neovasularization—by stimulating the creation of new blood vessels. As a result, severe NPDR progresses into PDR and causes the retina to sprout new blood vessels, usually with new vessels near the disc (NVD) or elsewhere (NVE). Because these newly formed blood vessels are frail, blood or fluid can seep into the vitreous portion of the retina, causing clouded vision and the disease vitreous hemorrhages. The leaking blood vessels can pollute the vitreous region, leading to blindness or severe vision loss, which causes the retina to deform or pull inward as seen in 4.28. Ayman S. El-Baz n.d.(a)

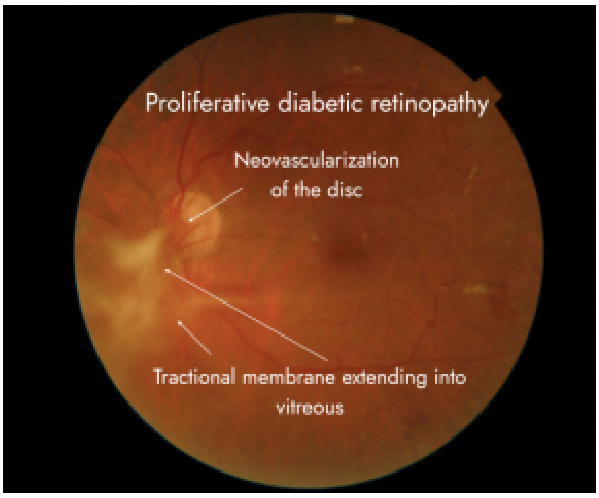


Figure 1.12: Proliferative DR

1.4.2 treatment options

The most crucial aspect of treatment is keeping diabetes under control. Controlling diabetes in the early stages of diabetic retinopathy can help prevent eye problems from developing. When diabetes is advanced and vision is impaired or threatened, keeping diabetes under control can help prevent the situation from worsening. The major treatments

for diabetic retinopathy that is threatening or impairing vision are:
1) Laser treatment: used to treat the formation of new blood vessels at the back of the eye (retina) in patients with proliferative diabetic retinopathy, as well as to stabilize some cases of maculopathy.
2) Eye injections: used to treat severe maculopathy that is threatening vision.
3) Steroid eye implants: to treat severe maculopathy when eye injections are ineffective.
4) eye surgery: If laser treatment is not possible due to extensive retinopathy, eye surgery may be performed to remove blood or scar tissue from the eye.

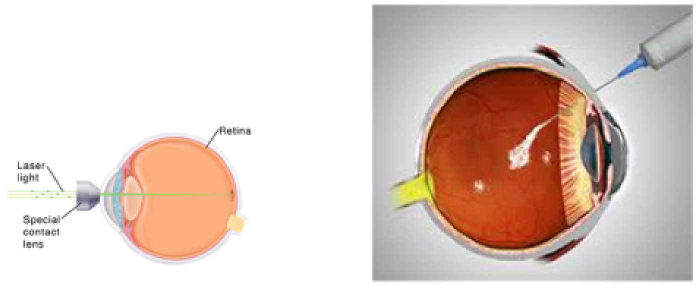


Figure 1.13: DR treatment operations

1.4.3 DR challenges

Despite breakthroughs in screening and diagnosis, DR is still the biggest cause of blindness globally. Current screening approaches have drawbacks, such as a lack of access to expert treatment and inconsistency in screening regimens, which can result in missing or delayed diagnosis. There are difficulties in accurately diagnosing and grading DR, such as inter-observer variability and imaging method limitations. Continuous research and innovation are required to enhance the screening, diagnosis, and treatment of this devastating condition.

1.5 The Social and emotional impact of diabetic retinopathy

Diabetes retinopathy not only causes vision loss, but it is also a significant cause of decreased life quality and a variety of other health issues.

- Several researchers have examined the link between DR and psychosocial consequences, Depression, anxiety, and social isolation are among the consequences of DR.

- add to that the importance of frequent eye exams, as well as the potential cost and hardship of treatment.

- Furthermore, it has been established that DR increases the risk of additional health concerns, including damage to the kidneys, nerves, and cardiovascular system.

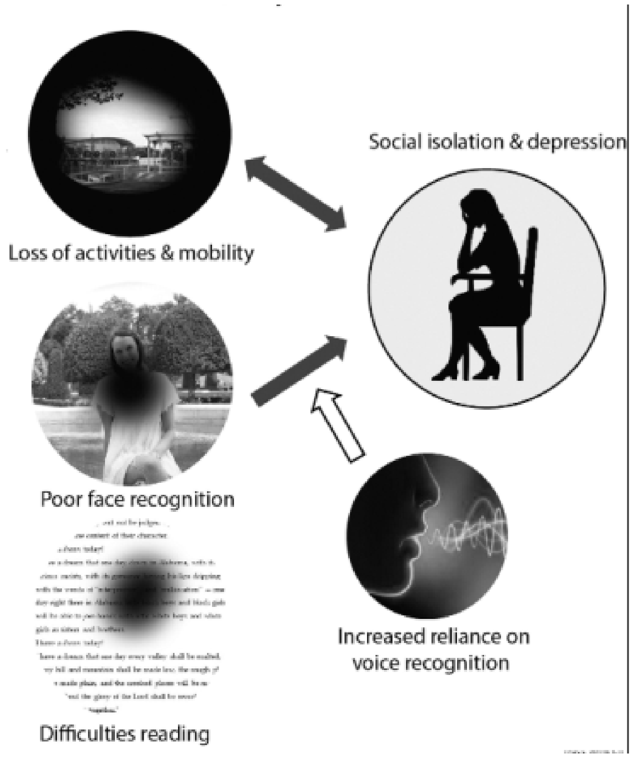


Figure 1.14: The social and psychological impact of DR

1.6 Conclusion

In conclusion, this chapter provides a comprehensive introduction to the medical context of diabetic retinopathy. In addition to emphasizing the connection between diabetes mellitus and the eyes, it also covers the anatomy and physiology of the eye, current methods for DR screening and diagnosis, clinical characteristics, and available treatments, as well as the social and emotional effects of this condition. For successful prevention, detection, and management of this sight-threatening consequence of diabetes mellitus, it is essential to comprehend the medical context of diabetic retinopathy

Chapter 2

Peprocessing medical images

2.1 Introduction

This chapter discusses the importance of preprocessing in medical image analysis and the challenges and limitations of retinal image preprocessing. It also explores the characteristics of medical image data and provides an overview of general preprocessing techniques. The chapter begins by highlighting the need for specialized preparation methods to overcome difficulties such as contrast, uneven illumination, and others in medical image preprocessing. The goal of preprocessing is to improve image quality overall, leading to more informative photos that aid in accurate diagnosis. The chapter then discusses the characteristics of medical image data, such as noise, artifacts, contrast ..etc Understanding these characteristics are essential for effective preprocessing. Additionally, it explains and specifies what are the challenges and limitations involved. General preprocessing techniques are introduced, focusing on noise removal and contrast enhancement in retinal images. Various types of noise are described, along with filtering techniques. Contrast enhancement techniques are discussed, These techniques aim to improve the visual quality and perception of images by increasing the contrast between distinct objects or regions. By employing appropriate preprocessing techniques, medical image analysis can overcome challenges, improve image quality, and facilitate accurate interpretation and diagnosis.

2.2 Preprocessing Medical Images

2.2.1 Purpose of data preprocessing

Real-world data is regularly fragmented, contradictory, inconsistent, and likely to contain many errors. Data preprocessing is the concept of changing the raw data into a clean data set that is suitable to be fed into a Model before executing it to an algorithm.
This process includes data cleaning, normalization, transformation, feature extraction, and selection.
In the case of CT images, a potential sequence of pre-processing steps may include denoising, interpolation, registration, organ windowing followed by normalization, and potentially zero-padding to improve the performance of the deep learning model.

2.2.2 Overview of Medical Image Preprocessing

Medical imaging evolved quickly to serve a critical role in medicine today, assisting in disease diagnosis and therapy. Ultrasonography, x-rays, mammography, computed tomography (CT scans), and nuclear medicine are all examples of medical imaging technology. The term "medical image preprocessing" refers to a broad range of methods used to improve the accuracy, dependability, and readability of images produced by various medical imaging modalities. Without a doubt, the topic of medical image preprocessing should include images of the retina. These images can be obtained utilizing modalities such as fundus photography or optical coherence tomography (OCT).

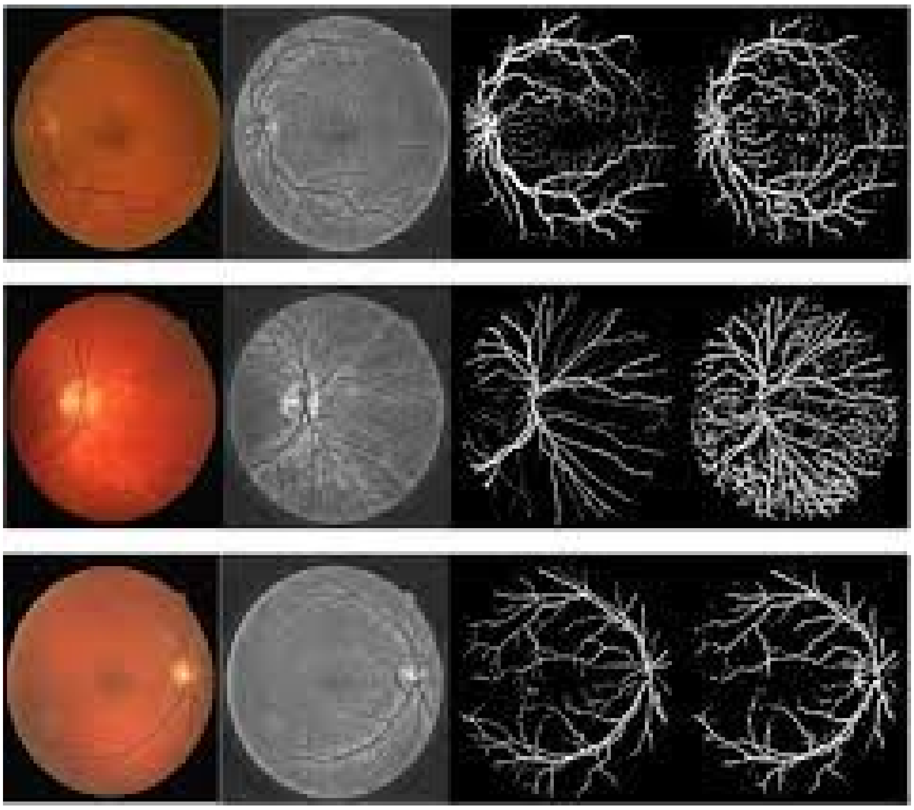


Figure 2.1: some preprocessing steps for Retina images

However, it is crucial to recognize that, similar to other medical images, retina scans can be impacted by a number of problems that might lower their quality and interpretation. These problems can vary in image quality due to things like ocular media opacities or patient cooperation, as well as noise, artifacts, low contrast, uneven illumination, and others. Therefore, Specialized preparation methods must be developed to overcome these difficulties.

2.2.3 Importance of Preprocessing in Medical Image Analysis

The fundamental goal of medical image preprocessing is to solve these issues mentioned earlier and improve image quality overall. It is now possible to minimize noise, improve image contrast, eliminate artifacts, rectify geometric distortions, and normalize image intensities using preprocessing techniques. This results in crisper, more informative images, which can aid in accurate diagnosis, quantitative analysis, and computer-aided decision-making.

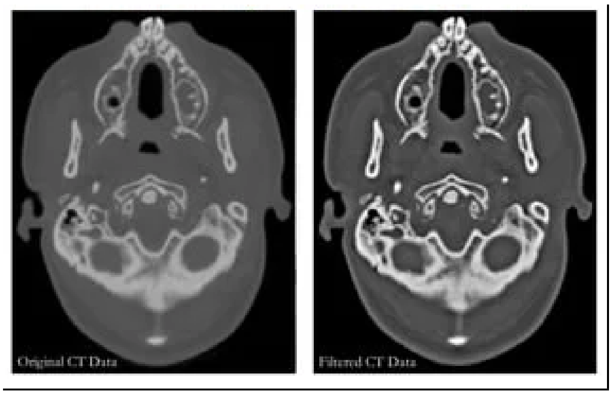


Figure 2.2: preprocessing medical images

2.2.4 challenges and limitations of retinal image preprocessing

The nature of medical imaging and the complexity of the human body present numerous challenges in pre-processing medical images. Among the extra issues are:

1 Heterogeneity of Imaging Modalities: Medical imaging technologies include X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and positron emission tomography (PET). Each modality has unique properties, imaging principles, and difficulties. Preprocessing algorithms must account for modalities' variances in order to maximize image quality.

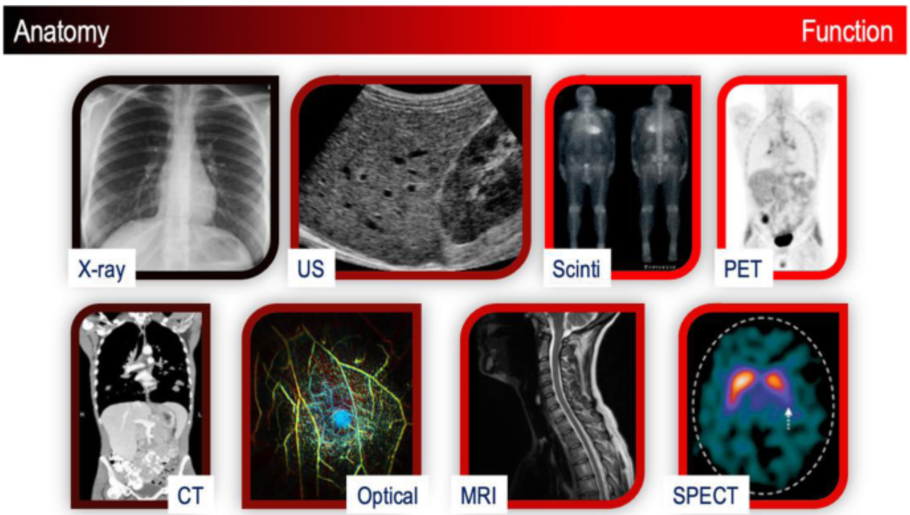


Figure 2.3: Medical Imaging system modalities

2 Limited Labeled Data Availability: Due to the demand for expert knowledge and time-consuming human annotations, obtaining well-annotated medical picture datasets may be difficult. This limitation can hamper preprocessing algorithm development and testing.

3 Time and Resource Constraints: Preprocessing procedures should be economical and computationally possible in a therapeutic context where time is important. To enable quick diagnosis and treatment decisions, real-time or near real-time processing is frequently necessary.

4 Explainability and interpretability: Preprocessing procedures should be visible and interpretable, particularly in the medical field where decisions affect patient care. The ability to explain how preprocessing processes affect the final image and subsequent analysis is crucial for building trust and ensuring that these techniques are used in clinical practice.

2.3 Characteristics of Medical Image Data

The internal structures and functions of the human body are typically not visible to the naked eye. However, through the use of various technologies, it becomes possible to generate images that allow medical professionals to examine the body, diagnose abnormalities, and guide treatment procedures. These medical images serve as windows into the body. It is important to note that no single image can reveal all aspects of the human body. Different medical imaging methods offer insights into different characteristics. The quality and visibility of structures within the images can vary significantly depending on factors such as the imaging equipment's capabilities, the operator's expertise, and considerations regarding patient radiation exposure and imaging time.

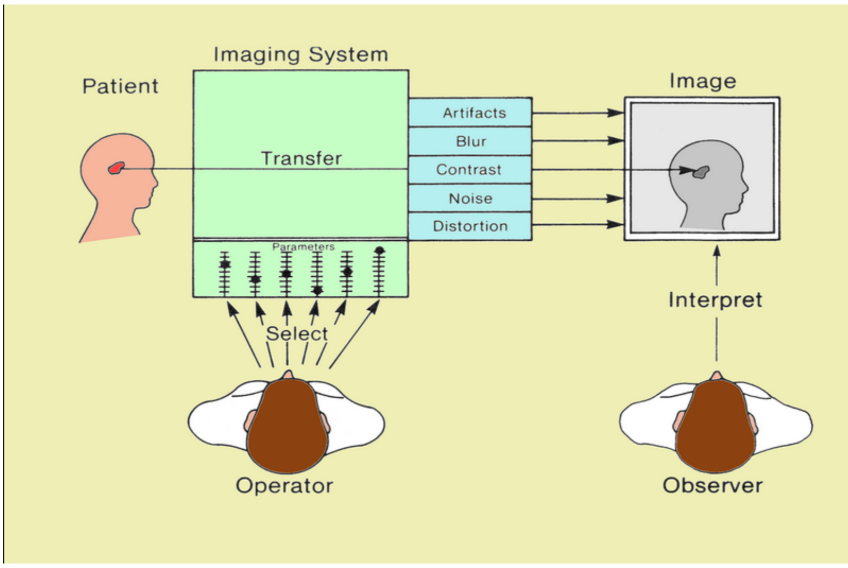


Figure 2.4: Medicl imaging system

1 Pixel: Each pixel in picture data correlates to a discrete value that reflects a specific point in the image. The visual representation of the image is formed by the arrangement and combination of pixels. In the context of medical images, Pixels are critical for capturing and portraying the fine details of anatomical structures and pathological findings. The spatial arrangement of pixels in medical imaging tells us about the shape, size, and relationships between various regions of interest within the body.

2 Spatial resolution: Image data is composed of a grid of pixels (in 2D) or voxels (in 3D), each of which represents a discrete value associated with a specific point in the image. The spatial resolution of the image determines the level of detail acquired, with higher resolution yielding finer spatial representation. In the case of medical images, great spatial resolution is frequently required to capture small anatomical details and effectively detect pathology.

3 Color or Grayscale: Image data can be stored in either color or grayscale format. Color images are made up of numerous channels (such as red, green, and blue) that work together to produce a wide range of colors. Grayscale images, on the other hand, consist of a single channel that represents various shades of gray. This attribute allows for the visualization of different tissue types and contrasts within medical images.

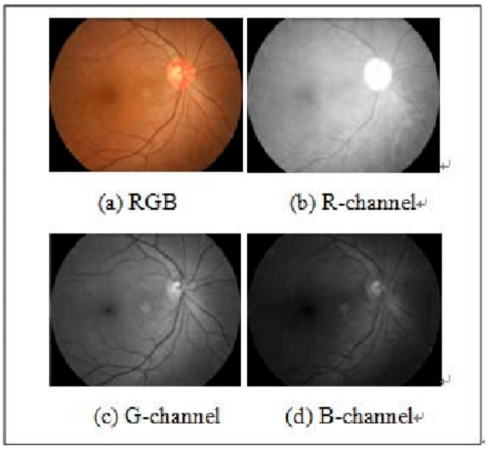


Figure 2.5: RGB Color channels

4 Image Format: Images can be saved in a variety of formats, including JPEG, PNG, TIFF, and DICOM(Digital Imaging and Communications in Medicine). Each format has its own set of advantages, compression methods, and metadata specifications, thus choosing the right format for a given application is critical.

5 Image Size and File Size: The resolution, color depth, and compression used determine the size of an image file. Larger photos with higher resolution and color depth have larger file sizes, which might affect storage and processing times.

6 Brightness: The brightness or intensity of an image refers to the overall lightness or darkness of the image. It plays an important role in image interpretation and analysis. In medical imaging, including retinal imaging, it is crucial to adjust the brightness levels carefully to ensure clear visualization of anatomical structures and abnormalities. By appropriately adjusting the brightness, subtle features like retinal vessels, lesions, or the optic disc can be enhanced.

7 Illumination: Proper illumination is crucial for high-quality and accurate interpretation of medical images, including retina images. Illumination refers to the distribution and intensity of light within an image. Variations in illumination can occur due to different factors, leading to inconsistencies in image quality. Uneven illumination poses challenges in accurately identifying and analyzing anatomical structures.

8 Dynamic range: The dynamic range of an image relates to the range of intensity values that it can capture. A broader dynamic range provides for more differentiation between different levels of brightness or color within an image. An adequate dynamic range is critical in medical imaging to effectively portray tiny alterations in tissue characteristics or pathological changes.

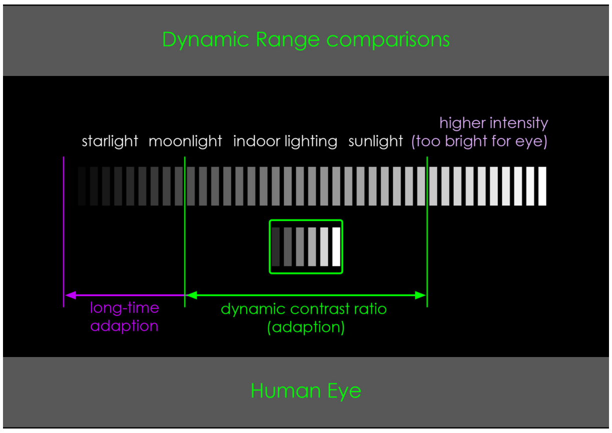
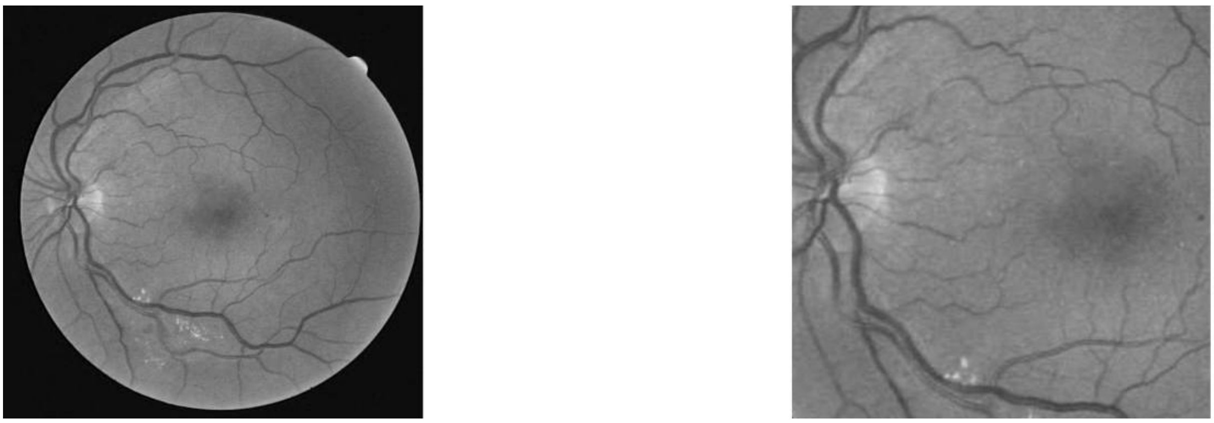


Figure 2.6: Dynamic range

9 Noise: Image data may contain many types of noise, such as random changes or interference, which can distort or decrease image quality. Noise reduction techniques are frequently used to increase the signal-to-noise ratio and image quality.



(a) Noisy image (b) A fragment of a Noisy image

Figure 2.7: Noise on retina images

10 Articrafts: The majority of imaging techniques can produce visual characteristics that do not represent a body structure or object. These are artifacts of images. In many cases, an artifact does not affect object visibility or diagnostic accuracy. Artifacts can have an impact on image interpretation and may necessitate the use of specific procedures to minimize or remove them.

11 Contrast: Contrast denotes distinction. Contrast in an image might take the form of contrasting shades of gray, light intensities, or colors. The most essential aspect of an image is its contrast. Only if an object within the body has adequate physical contrast relative to surrounding tissue will it be seen in an image. However, visual contrast that exceeds what is required for adequate object visibility is rarely useful and, in many circumstances, detrimental. When contrast is given a value, it relates to the difference between two distinct spots or areas in an image.

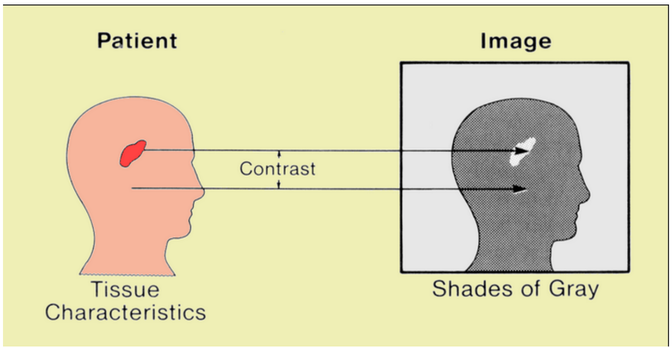


Figure 2.8

12 Blur: Blur is defined as a lack of sharpness or clarity in a picture, which results in a loss of small details and smooth transitions between pixels. It is a prevalent phenomenon in images and can be caused by a variety of circumstances, including motion blur, optical aberrations, or out-of-focus capture. Image blur may be particularly difficult in medical imaging because it might hide crucial anatomical structures or disease characteristics.

Figure 2.9: Blurry retina image

13 Histogram: The histogram of an image represents the distribution of pixel intensities across the image. It provides valuable information about the image's contrast, brightness, and the presence of specific structures or abnormalities. Analyzing the histogram can help identify image characteristics, such as the range of intensities, the presence of outliers. The histogram analysis of a retina image can help identify regions of interest, assess the presence of lesions or anomalies, and guide further processing steps.

14 Distortion: A medical image should not only show interior body objects, but also provide an accurate representation of their size, shape, and relative positions. These three factors can, however, be distorted by an imaging method.

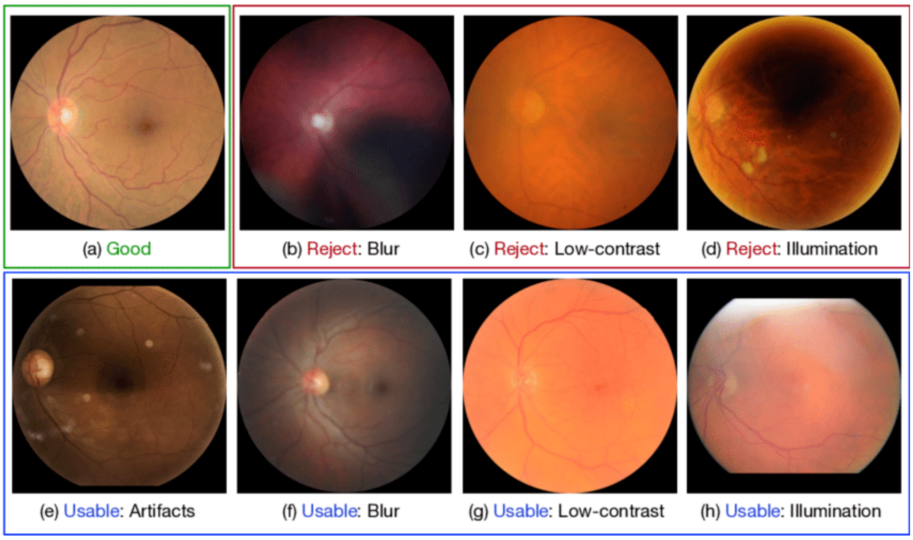


Figure 2.10

2.4 General Preprocessing Techniques

2.4.1 Noise Removal Techniques on Retinal Images

Types of Noise

A. Gaussian noise Gaussian noise, often called geometric noise, has a normal distribution probability density function. In other words, the noise has a Gaussian distribution of possible values. Gaussian noise is appropriately defined as noise with a Gaussian amplitude sharing.

B. Salt and Pepper Noise Salt and pepper noise is caused by sudden, abrupt changes in the picture signal and manifests as sporadic white or black dots all over the image. Black pixels show up in bright areas in salt and pepper noise, whereas bright pixels show up in dark areas. Dead pixels, issues with the analog-to-digital converter, and transmission bit errors can all cause this type of noise.

C. Speckle Noise Speckle noise has an impact on all core facets of logical imaging, especially clinical ultrasound imaging. The cause is due to the sound processing of backscatter signals from various distributed targets. Basic scatter signals result in speckle noise. In medical literature, speckle noise is referred to as texture, and it may include diagnostic information. It could be less ideal for visual interpretation to smooth out the roughness.

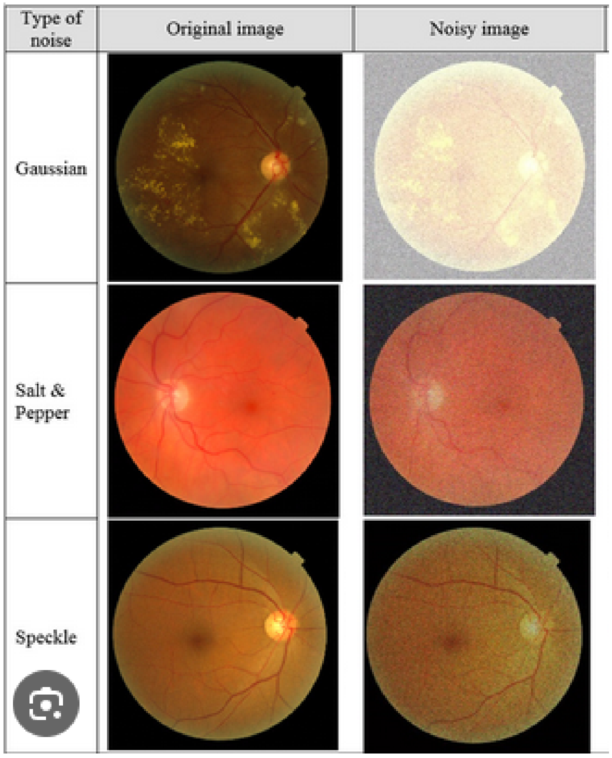


Figure 2.11: specific type of noise on a retina image

Filtering techniques

Gaussian Filtering: The best time domain filters are considered to be Gaussian filters. It is a non-uniform lowpass filter type. Gaussian filters are ones that have a Gaussian function, and these filters have a Gaussian impulse response. The quickest possible grouping delay is present. A Gaussian filter's main objective is to lessen distortion in the lowest and highest signals. The result of hiding a picture with a Gaussian capacity in image processing is a Gaussian obscure, also known as Gaussian smoothing.

The Gaussian filter is represented by the following equation: $G(x, y) = 1\frac{}{2\pi\sigma^2 \cdot \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)}$

where G(x, y) denotes the Gaussian kernel at position (x, y) and represents the standard deviation of the Gaussian distribution.

Median Filtering Median filtering is a non-linear filtering technique that replaces each pixel's value within its immediate vicinity. It successfully eliminates salt-and-pepper noise while keeping the image's borders and features.

The median filtering operation can be defined as:

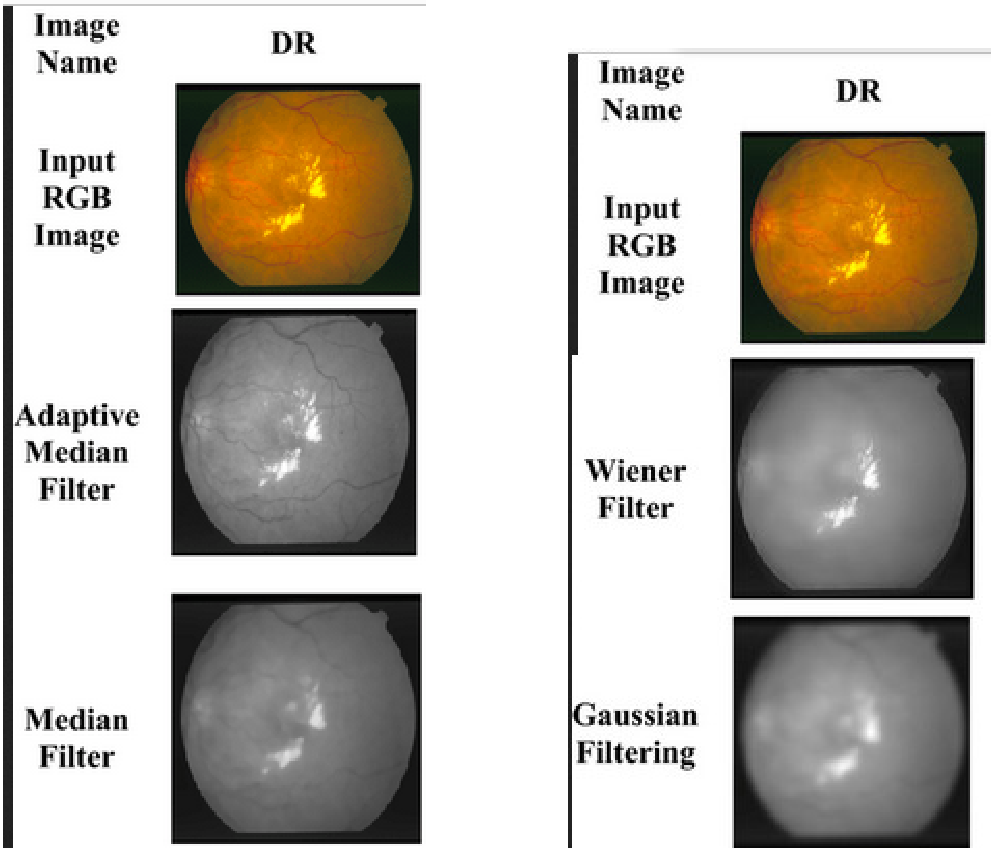
$$\mathrm{I}_{\text{filtered}}(x, y) = \text{median}\left(I_{\text{neighborhood}}\right)$$

where $I_{\text{filtered}(x,y)}$is the filtered pixel value at position (x, y), and $I_{\text{neighborhood}}$ represents the pixel values in the neighborhood of (x, y).

Wiener Filter By calculating the power spectra of both the noise and the signal, the Wiener filtering process calculates the original, noise-free image. It effectively reduces additive white Gaussian noise (AWGN).
The Wiener filter computes the estimated signal using the following equation:

$$\begin{aligned}
H(u, v) &= \frac{R_f(u, v)}{R_f(u, v) + K} \\
F(u, v) &= H(u, v) \cdot G(u, v) \\
f(x, y) &= \mathcal{F}^{-1}(F(u, v))
\end{aligned}$$

where $H(u, v)$ is the Wiener filter transfer function, $R_f(u, v)$ is the power spectrum of the noisy image, K is a regularization constant, $G(u, v)$ is the Fourier transform of the noisy image, $f(x, y)$ represents the denoised image, and $F(u, v)$ represents the Fourier transform of the denoised image.



(a) (b)

Figure 2.12: Preprocessed Images with different types of filtering techniques

Max and Min Filter: In the max and min filters, each pixel in an image is given a new value based on whether its surrounding area has the highest or lowest value. The neighborhood represents the form of the filter.

Non-local Means Filtering: This technique utilizes the similarity between image patches to denoise an image. It takes into account non-local similarities across the entire image to remove noise while preserving image details.
The non-local means filter calculates the denoised pixel value using weighted averaging of similar patches: f(x) = Σ w(x, y) * g(y) where f(x) represents the denoised pixel value at position x, g(y) denotes the pixel value at position y in a similar patch, and w(x, y) represents the weight assigned to each patch based on their similarity.

2.4.2 Contrast enhancement techniques

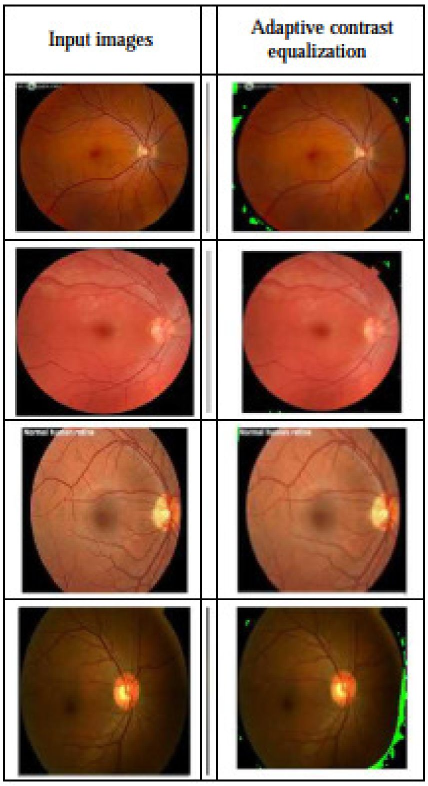
Contrast enhancement techniques are a group of image processing technologies that are used to improve the visual quality and perception of a picture by increasing the contrast between distinct objects or regions within the image. These strategies seek to improve detail recognition, increase the exposure of critical characteristics, and make images more aesthetically attractive and informative.

1. **Histogram equalization:**

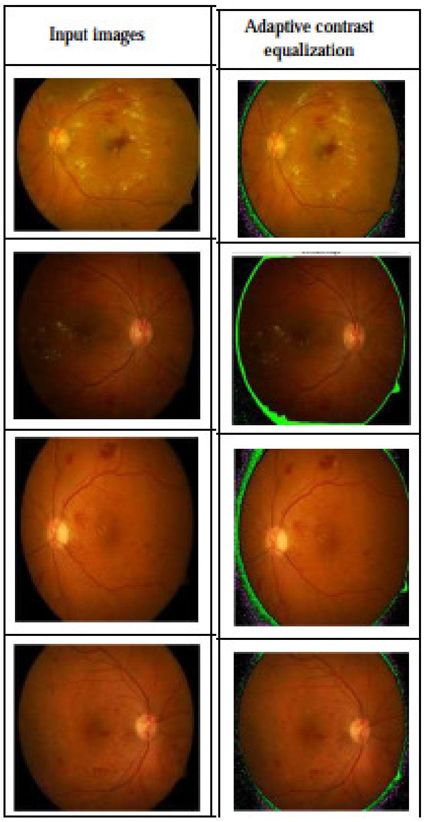
Histogram equalization is One of the most used color-normalizing approaches. Changing the range of pixel intensities yields the required intensity histogram. Histogram equalization is achieved by utilizing nonlinear transfer functions to map the pixel values of input and output pictures. It redistributes the pixel intensity values in an image to enhance contrast by stretching the intensity range. It can be applied globally or locally to specific regions of an image see 2.14. Ayman S. El-Baz n.d.(d)

2. **Adaptive Histogram Equalization (AHE)**

Adaptive contrast equalization can be used to improve the low-contrast sections of an image and sharpen the features. AHE is used to improve the clarity of retinal images by dividing the image into smaller regions and applying histogram equalization independently to each region. This allows AHE to enhance the contrast in specific areas while preserving local details and reducing the risk of over-amplifying noise or artifacts. It is also used to detect retinal pathologies by improving the visibility of important features such as blood vessels, lesions, or abnormalities see 2.13b[Ayman S. El-Baz n.d.(e)



(a) Preprocessed output for normal images.



(b) Preprocessed output for abnormal images.

Figure 2.13: Applying AHE on retina images

3. **Contrast limited adaptive histogram equalization (CLAHE)**

Contrast-constrained adaptive histogram equalization is a contrast enhancement method used in image processing to improve picture contrast. Contrast-restricted adaptive histogram equalization (CLAHE) outperforms adaptive histogram equalization in that it does not overamplify noise in relatively homogenous areas of a picture. CLAHE's contrast limiting overcomes this issue, making it suitable for local contrast enhancement in image areas. Ayman S. El-Baz n.d.(e)The 5 steps in CLAHE are mentioned below:

- Divide the image into tiny regions.
- Decide the mapping functions of the local histogram.
- Choose the clipping point of the histogram.
- Apply the function to every region.
- Reduce the noise by the background subtraction method.

CLAHE (Contrast Limited Adaptive Histogram Equalization) is particularly beneficial for enhancing retinal images due to its local nature, which improves contrast selectively in different regions. This preservation of local details is crucial for analyzing retinal features. see 2.14 Karel Zuiderveld 1994

4. Contrast stretching

Contrast stretching (often called normalization) is a simple image enhancement technique that attempts to improve the contrast in an image by 'stretching' the range of intensity values it contains to span a desired range of values, e.g. the full range of pixel values that the image type concerned allows. It differs from histogram equalization in that it can only apply a linear scaling function to the image pixel values. As a result, the 'enhancement' is less harsh.
Contrast stretching is used in retinal image analysis to enhance visibility by expanding the range of pixel intensities. It improves image quality and facilitates accurate analysis and diagnosis of retinal diseases see 2.14.

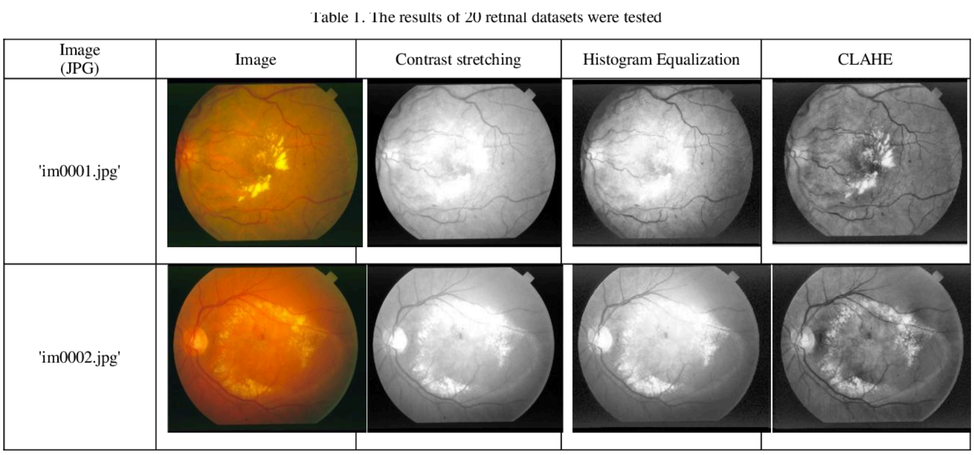


Table 1. The results of 20 retinal datasets were tested

Image (JPG)	Image	Contrast stretching	Histogram Equalization	CLAHE
'im0001.jpg'				
'im0002.jpg'				

Figure 2.14: Applying contrast enhancement techniques on retina

2.5 Conclusion

In conclusion, preprocessing techniques have a significant impact on improving the quality and reliability of medical images, specifically in retinal image analysis. By addressing issues like noise reduction and contrast enhancement, these techniques enhance image clarity and enable more accurate interpretation and diagnosis. Despite challenges such as varying imaging modalities and limited labeled data, preprocessing plays a crucial role in assisting medical professionals in making informed decisions and improving the overall quality of medical imaging.

Chapter 3

Deep learning

3.1 Introduction

Artificial intelligence (AI) has rapidly transformed various industries, including healthcare, finance, and technology. In particular, the field of machine learning (ML) has played a crucial role in advancing AI applications. ML algorithms have shown remarkable capabilities in tasks such as classification, regression, and pattern recognition. However, with the emergence of deep learning (DL), a subset of ML, there has been a significant shift in the landscape of AI research and development. This chapter aims to explore the advancements and advantages of deep learning over traditional machine learning approaches. While both ML and DL involve training models on data to make predictions, DL leverages artificial neural networks with multiple layers to learn complex representations of the data. This allows DL models to capture intricate relationships and dependencies within the data, leading to more accurate and robust predictions. Furthermore, this chapter will delve into two prominent components of deep learning: the transformer model and the Vision Transformer (ViT). The transformer model, originally designed for natural language processing tasks, has demonstrated its efficacy in various domains by employing self-attention mechanisms. Similarly, the ViT model has introduced a novel approach to image analysis by applying self-attention to image patches instead of traditional convolutional operations. By understanding the advancements and potential of DL, particularly the transformer model and ViT, researchers and practitioners can harness the power of deep learning to tackle complex problems in AI. This chapter will provide insights into the key concepts, benefits, and applications of these models, showcasing their transformative impact on the field of artificial intelligence.

3.2 Artificial intelligence and health care

Artificial intelligence (AI) is a type of intelligence technology that has been artificially created by humans to replicate human behavior. This artificial intelligence is integrated with a computer system known as an AI system, which eventually functions as a ”thinking machine.”

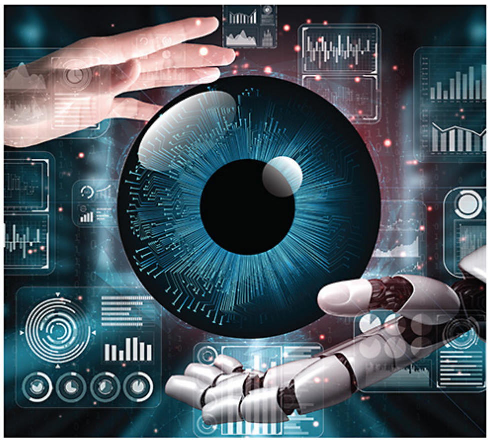


Figure 3.1: AI and Eyecare

AI, including subsets like deep learning (DL) and machine learning (ML), has revolutionized healthcare. It improves diagnostics, treatments, and research outcomes by utilizing computational methods and statistical analyses. AI surpasses traditional empirical studies in accuracy, providing valuable insights for disease prognostics. This technology enables precise diagnostics and personalized treatments, benefiting patients and advancing healthcare. Ankur Saxena n.d.

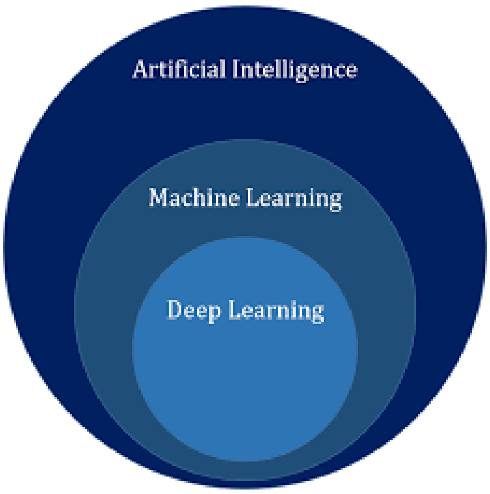


Figure 3.2: Artificial Intelligence

3.3 Machine learning

In essence, machine learning is a type of applied statistics that places more of an emphasis on using computers to statistically estimate complex functions than on demonstrating confidence intervals around these functions;

With the help of machine learning, we can complete tasks that are too challenging for fixed programs created and written by humans. Machine learning is intriguing from a scientific and philosophical perspective because learning more about it requires learning more about the concepts that underlie intelligence.

Machine learning can be used to solve many different types of tasks. The following are some of the most popular machine-learning tasks: classification, regression, etc. Bengio et al. 2016

3.3.1 Machine learning algorithms

An algorithm that can learn from data and is given access to an entire dataset is called a machine learning algorithm. A dataset is a grouping of numerous examples, which themselves are groups of features. The majority of machine learning algorithms fall into the supervised learning and unsupervised learning categories;

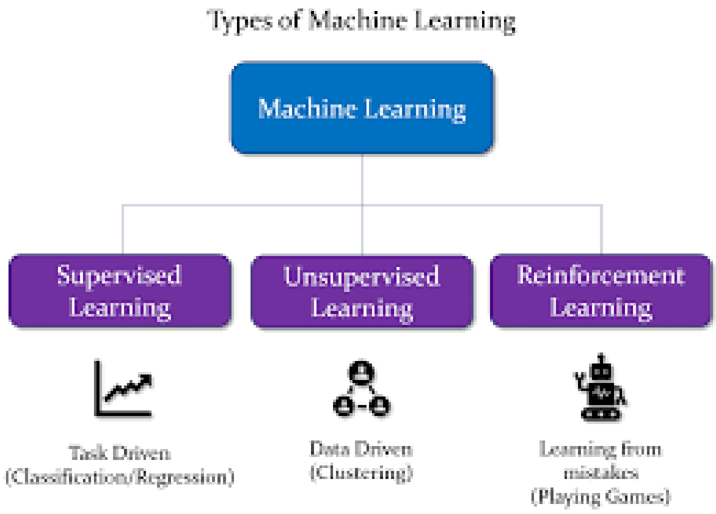


Figure 3.3: Machine learning types

Supervised Learning Algorithms :

Supervised learning algorithms are designed to learn from datasets where each example is accompanied by both input features and a corresponding label or target. The term ”supervised learning” stems from the notion that the machine learning system receives guidance on how to achieve the target value (represented by the label) from a teacher or instructor. It is worth noting that the phrase remains applicable even if the training set labels were obtained automatically or through alternative means.

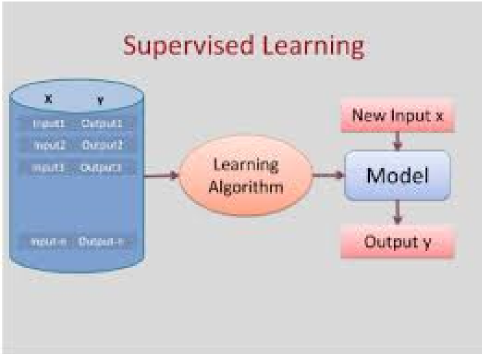


Figure 3.4: Supervised learning

Unsupervised Learning Algorithms :
In unsupervised learning, the algorithm explores a dataset containing a multitude of features and uncovers the meaningful attributes and structure it possesses.
In simpler terms, without the presence of a teacher or instructor, unsupervised learning algorithms autonomously learn to interpret and make sense of the data.
Other unsupervised learning algorithms fulfill distinct roles, such as clustering, which partitions the dataset into clusters comprising related examples. This allows the algorithm to group together instances that share similarities, even without prior knowledge of class labels or target values. Bengio et al. 2016

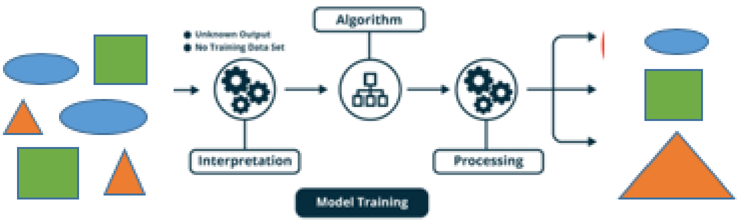


Figure 3.5: Unsupervised learning

Reinforcement Learning Algorithms :
The third category of machine learning algorithms is reinforcement learning, which focuses on determining the appropriate course of action to take in a circumstance in order to maximize the proper outcome. The decisions are taken in order. The algorithm's output can be positive or negative at any point along the road to total outcomes. Thus, the overall outcome is the sum of all positive and negative outcomes along the path. The purpose of the algorithm is to identify the best approach to maximize the outcome. Ankur Saxena n.d.

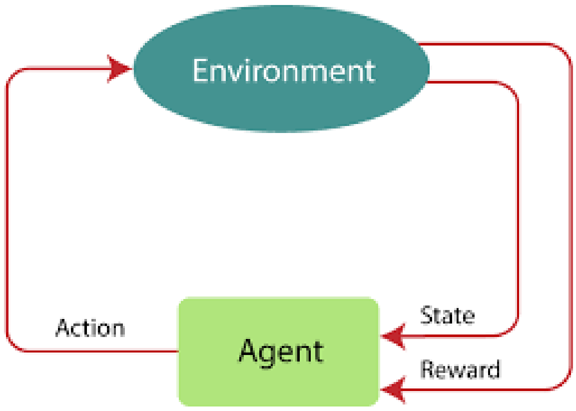


Figure 3.6: Reinforcement learning

We frequently measure the model's accuracy to assess a machine learning algorithm's capabilities for tasks like classification. Simply put, accuracy is the percentage of examples for which the model generates the desired result.
By measuring the error rate, we can also get equivalent information. the percentage of examples for which the model fails to produce the desired results. Bengio et al. 2016

3.4 Introduction to Deep Learning:

Deep learning is a specific kind of machine learning that involves complex learning processes using artificial neural networks. These networks consist of multiple layers that transform input data into meaningful information for performing prediction tasks. Deep learning allows machines to learn and improve their performance through data processing.

In deep learning, the learning process occurs within the layers of the neural network. Each layer contains units that process input data and pass the transformed information to the subsequent layer. These units perform computations using adjustable parameters called weights and biases. During the training process, these weights and biases are iteratively adjusted to optimize the model's performance.

Deep learning models excel at learning complex representations of the input data. Through the hierarchical structure of the neural network, each layer knows to extract and refine features at different levels of abstraction. This enables the model to capture intricate relationships and dependencies within the data, leading to more accurate predictions.

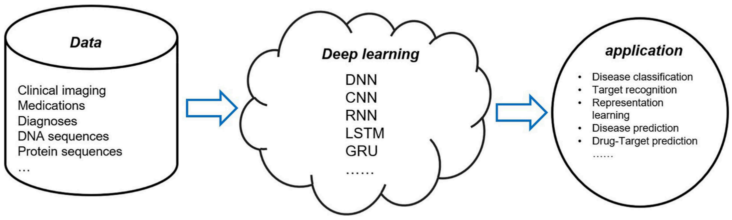


Figure 3.7: Deep learning simplified process

One of the critical advantages of deep learning is its ability to learn from large amounts of data. Deep learning models can discover meaningful patterns and correlations by processing extensive datasets, allowing them to generalize well to new, unseen data. This aspect of self-learning is a fundamental characteristic of deep learning, as the model adapts its internal representations based on the data it encounters.

Deep learning has demonstrated significant success in various fields, including computer vision, natural language processing, and more. Its capability to automatically learn representations and process large-scale datasets has propelled advancements in tasks such as image classification, object detection, speech recognition, and many other applications.

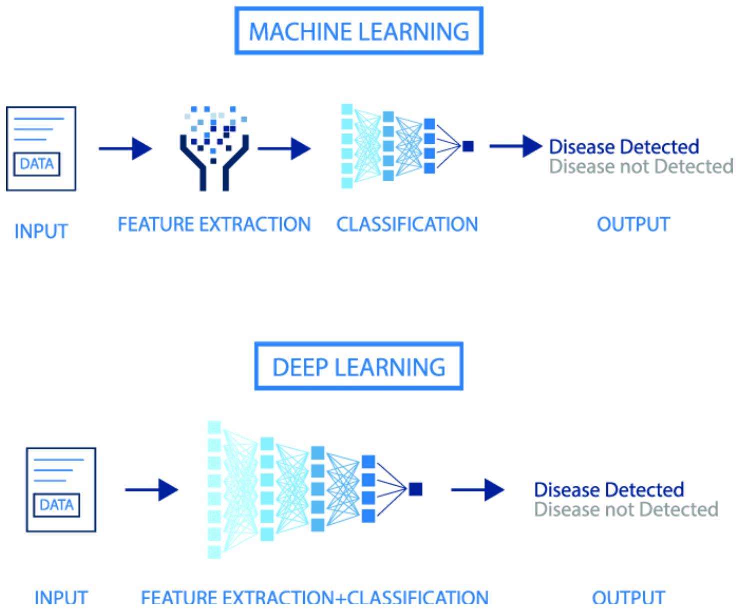


Figure 3.8: Diffrence between ML and DL

3.5 Fundamentals of Neural Networks:

Neural networks are a class of machine learning models inspired by the human brain's structure and functioning. They are widely used for tasks such as image and speech recognition, natural language processing, and many others. A neural network consists of interconnected layers (also called nodes or units) that work together to process and analyze data.

Here's a brief overview of the basic components of neural networks:

Neuron:
This is a basic building block of a NN. It takes weighted values, performs the mathematical calculation, and produces output. It is also called a unit, node, or perceptron.

Input:
This is the data/values passed to the neurons.

Weights:
These values explain the strength (degree of importance) of the connection between any two neurons.

Bias:
is a constant value added to the sum of the product between input values and respective weights. It is used to accelerate or delay the activation of a given node.

Activation function:
is a function used to introduce the non-linearity phenomenon into the NN system. This property will allow the network to learn more complex patterns.

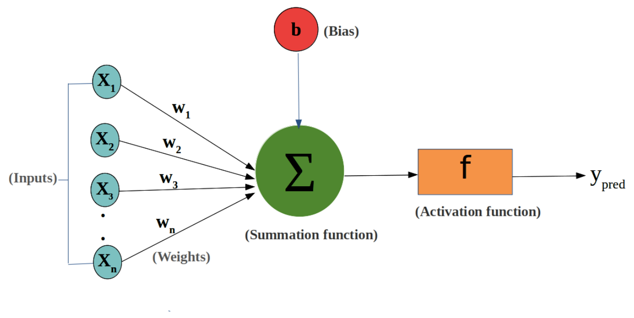


Figure 3.9: activation function technique

3.6 Artificial Neural Network :

An artificial neural network (ANN) is a computational network inspired by biological neural networks, which are the intricate networks of neurons in human brains. The ANN nodes are said to be trained to behave like actual neurons, hence they are artificial neurons. Manaswi 2018

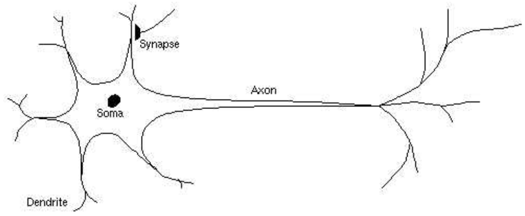


Figure 3.10: Neuron

A typical neural network is made up of a large number of artificial neurons called units that are organized into three layers: input, hidden, and output . Manaswi 2018

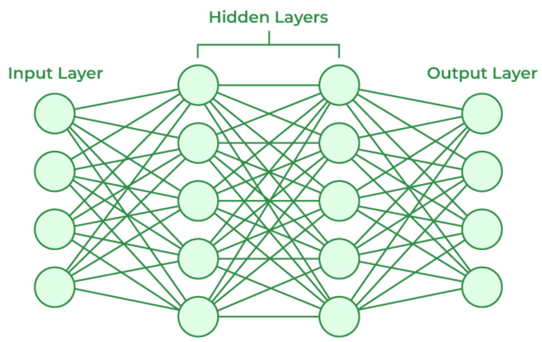


Figure 3.11: Artificial Neural Network Architecture

each neuron in the hidden layer is fully connected to every neuron in the previous input layer and the next output layer. A neural network learns by iteratively modifying the weights and biases in each layer to achieve the best results.

3.6.1 Evaluation metrics

evaluation metrics for classification models are defined as quantitative measures used to assess the performance and effectiveness of these models. These metrics provide objective insights into the model's predictive accuracy, precision, recall, and other relevant factors. Vujović 2021 Some commonly used metrics highlighted in the article include:

Accuracy: It measures the overall correctness of the model's predictions by calculating the ratio of correctly predicted instances to the total number of instances.Vujović 2021

Precision: It quantifies the proportion of correctly predicted positive instances out of all instances predicted as positive. Precision focuses on the accuracy of positive predictions. Vujović 2021

Recall (Sensitivity): It calculates the proportion of correctly predicted positive instances out of all actual positive instances. Recall measures the model's ability to identify positive instances correctly.Vujović 2021

F1 Score: It combines precision and recall into a single metric by calculating their harmonic mean. F1 Score provides a balanced measure between precision and recall.Vujović 2021

Specificity: It calculates the proportion of correctly predicted negative instances out of all actual negative instances. Specificity is particularly relevant in binary classification problems with imbalanced classes.Vujović 2021

Area Under Curve (AUC): It represents the area under the receiver operating characteristic (ROC) curve and provides a measure of the model's ability to distinguish between classes. A higher AUC value indicates better discrimination between positive and negative instances.Vujović 2021

Confusion Matrix: It is a tabular representation that summarizes the model's predictions against the actual class labels, displaying true positives, true negatives, false positives, and false negatives.Vujović 2021

These metrics help evaluate the performance of classification models by assessing their accuracy, precision, recall, F1 score, specificity, and discriminatory ability.

3.7 Convolutional Neural Networks (CNNs):

A convolutional neural network (CNN) is a deep, feed-forward artificial neural network that maintains the hierarchical structure by learning internal feature representations and generalizing the features in common image problems such as object recognition and other computer vision problems. Manaswi 2018

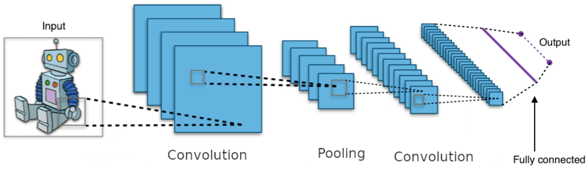


Figure 3.12: Convolutional Neural Network

3.7.1 CNN features

Convolutional layers:

These layers help the network identify small patterns in the input data, such as edges or textures.
it performs a mathematical operation called convolution. It takes an input image and applies a set of filters to detect patterns and features. The filters slide across the image, multiplying their values with the corresponding pixels in the input. The results are summed up to create a new output feature map. This process is repeated for each filter, capturing different patterns in the image.

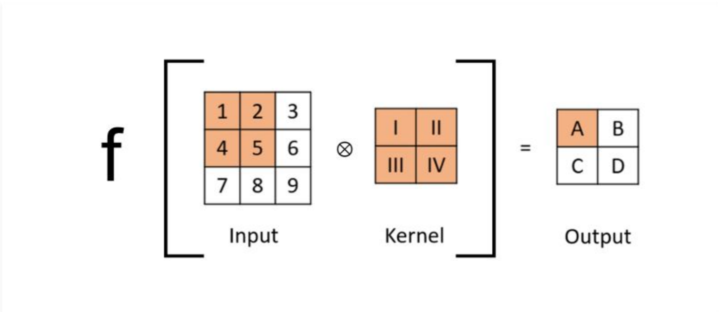


Figure 3.13: Illustration for how a convolutional layer operates.

pooling layers:

These layers make the network focus on the most important features while reducing the amount of information to process. Similar to the convolutional layer, the pooling operation sweeps a filter across the entire input, but the difference is that this filter does not have any weights. Instead, the kernel applies an aggregation function to the values within the receptive field, populating the output array. There are two main types of pooling:

Max pooling: As the filter moves across the input, it selects the pixel with the maximum value to send to the output array. As an aside, this approach tends to be used more often compared to average pooling.

Average pooling: As the filter moves across the input, it calculates the average value within the receptive field to send to the output array.

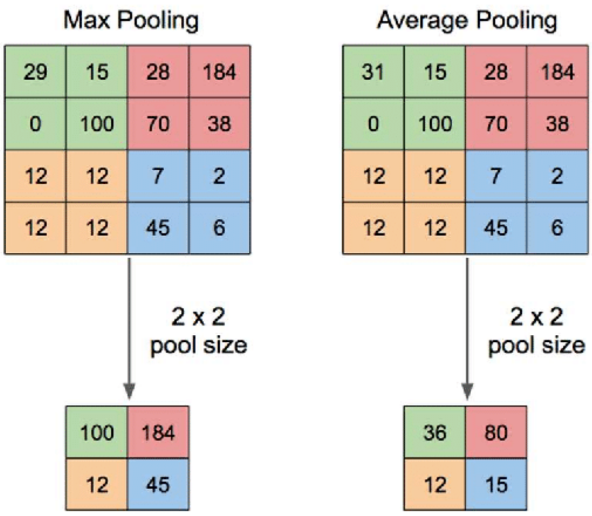


Figure 3.14: Illustration of Max Pooling and Average Pooling

Fully Connected Layers:

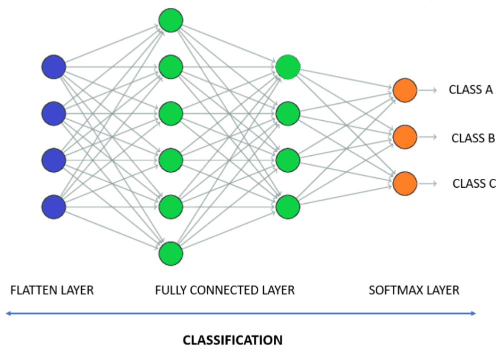


Figure 3.15: Hidden layer

These layers learn more complex patterns and make predictions based on the features extracted by earlier layers. the pixel values of the input image are not directly connected to the output layer in partially connected layers. However, in the fully-connected layer, each node in the output layer connects directly to a node in the previous layer.

Mathematically, the output of the fully connected layer is calculated by multiplying the input values by a weight matrix, adding a bias term, and applying an activation function. The weights and biases are learned during training to make accurate predictions.

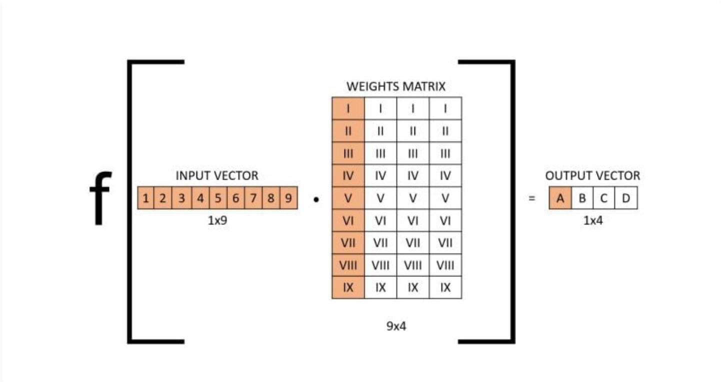


Figure 3.16: Illustration representing the fully connected layer

This layer performs the task of classification based on the features extracted through the previous layers and their different filters.

Activation Functions:

These functions help the network understand which features are important and which can be ignored, While convolutional and pooling layers tend to use ReLu functions, FC layers usually leverage a softmax activation function to classify inputs appropriately, producing a probability from 0 to 1.

3.7.2 CNN Modular Architectures

A CNN is a feed-forward deep neural network design that consists of several convolutional layers followed by a pooling layer, activation function, and optional batch normalization. It also includes fully connected layers. As an image travels through the network, it shrinks due to max pooling. The final layer predicts the class probabilities. Manaswi 2018

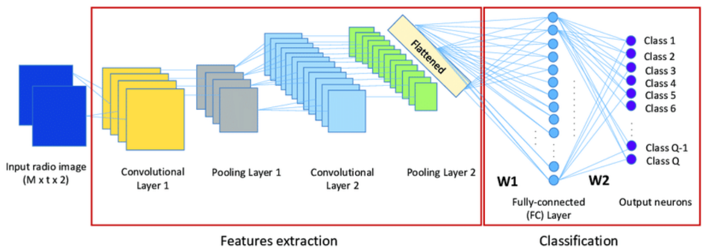


Figure 3.17: CNN architecture

3.7.3 Applications of CNNs in image classification

Convolutional neural networks (CNNs) are used for image classification tasks, such as recognizing objects in images. These networks are powerful tools in the field of artificial

intelligence (AI) known as computer vision. Computer vision involves extracting meaningful information from visual inputs, like images or videos, to enable systems to make informed decisions.
CNNs find various applications in different domains:

Marketing: Social media platforms use CNNs to suggest people's identities in photos, making it easier to tag friends in photo albums.

Healthcare: CNNs are incorporated into radiology technology to assist doctors in accurately identifying cancerous tumors within healthy anatomy.

Retail: Some e-commerce platforms use visual search powered by CNNs to recommend items that complement a customer's existing wardrobe.

Automotive: Although fully driverless cars are still being developed, CNN technology is already improving driver and passenger safety in vehicles. For example, CNNs can help with features like detecting and staying within lane lines.

3.8 Transformer Architecture:

The Transformer architecture is widely used in natural language processing (NLP). It replaces traditional layers with self-attention and feed-forward networks. Self-attention captures word relationships, while positional encoding handles word order. The Transformer has an encoder and a decoder with multiple layers. It uses residual connections, layer normalization, and autoregressive decoding for training and inference. The Transformer is effective for various NLP tasks due to its parallelizability and ability to capture long-range dependencies.

The vision Transformer (ViT) is introduced as an alternative to convolutional neural networks (CNNs) for image recognition, It divides images into patches, converts them into token embeddings, and uses self-attention in the Transformer encoder to capture global dependencies. Pretraining with contrastive learning helps learn meaningful visual representations. [An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," published by Dosovitskiy et al. 2021

3.8.1 key components and features of the ViT

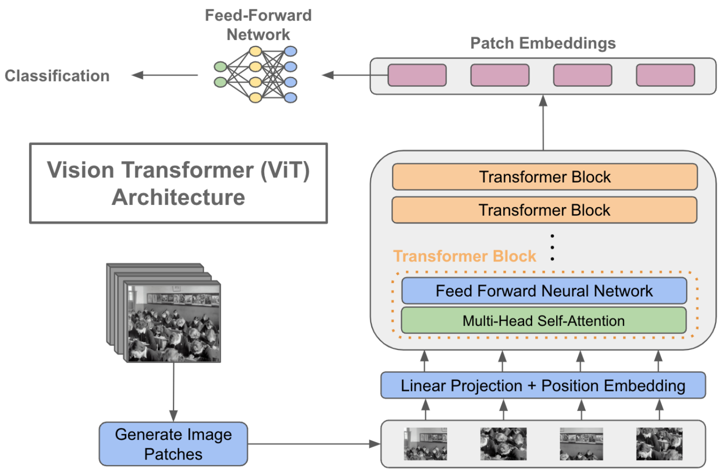


Figure 3.18: ViT architecture

- **Self-Attention Mechanism:** Self-attention is a key component of Transformers that allows the model to weigh the importance of different elements in a sequence by attending to all other elements. In the case of Vision Transformers, self-attention is utilized to capture dependencies between image patches, enabling the model to understand the global context of the image.

- **Patch Embeddings:** In Vision Transformers, the input image is divided into fixed-size patches, which are then linearly projected into a sequence of token embeddings. These patch embeddings serve as the input to the Transformer encoder and enable the model to process image information.

- **Pretraining:** Pretraining involves training a model on a large-scale dataset with a self-supervised or unsupervised learning objective. In the case of ViTs, the model is pre-trained on a dataset using a variant of contrastive learning, where the model learns to distinguish between similar and dissimilar image patches. Pretraining helps the ViT model learn useful visual representations before fine-tuning specific downstream tasks.

- **Fine-tuning:** After pretraining, the ViT model is fine-tuned on a specific image-related task, such as image classification. Fine-tuning involves training the model further on a labeled dataset for the target task, often using supervised learning objectives such as cross-entropy loss.

3.8.2 ViTs architecture

Input Processing: The input image is divided into fixed-size patches, and each patch is converted into a patch embedding. This allows the model to process the image in a sequence-based manner.

Transformer Encoder: The patch embeddings are processed by a stack of Transformer encoder layers. Each encoder layer includes self-attention mechanisms that capture dependencies between image patches and enable the model to understand the global context of the image. The encoder layers also include MLP layers that perform non-linear transformations on the patch embeddings, refining the learned representations.

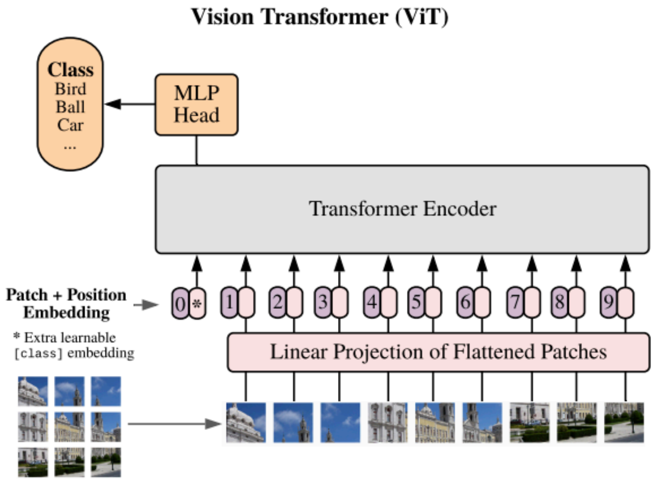


Figure 3.19: Vision transformer

Classification Head: A classification head is added on top of the Transformer encoder. It typically consists of fully connected layers (MLPs) that take the sequence of embeddings and produce the final output logits. Activation functions, such as softmax, are applied to the logits to obtain predicted probabilities for each class in image classification tasks.

Pretraining and Fine-Tuning: The ViT architecture often undergo a two-step training process. In the pretraining phase, the model is trained on a large-scale dataset using contrastive learning. This helps the model learn meaningful visual representations by distinguishing between similar and dissimilar image patches. After pretraining, the model is fine-tuned on a specific image-related task, such as image classification, by further training it on a labeled dataset using supervised learning objectives like a cross-entropy loss. Fine-tuning allows the model to adapt its learned representations to the target task, improving performance.

3.9 Advantages and Limitations of ViTs:

Vision Transformers (ViTs) possess both strengths and weaknesses, similar to other models. They bring forth certain advantages but are also accompanied by challenges and limitations that must be taken into account during their application.

- **Challenges:**[An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" by Dosovitskiy et al. 2021

  1. **Computational requirements:** ViTs typically require more computational resources compared to CNNs due to the self-attention mechanism. Training and inference times can be longer, making ViTs less suitable for real-time applications or resource-constrained environments.
  2. **Data dependencies:** ViTs often rely on large-scale labeled datasets for pre-training, making data collection and curation challenging. Acquiring diverse and representative datasets is crucial for ViTs to learn effective visual representations.

- **Limitations:**

  1. **Attention to local details:** ViTs may face challenges in precisely capturing fine-grained local details compared to CNNs. The global focus on self-attention may not give as much emphasis to specific localized features, potentially affecting performance in tasks that heavily rely on such details.
  2. **Interpretability:** Although attention maps in ViTs provide interpretability, understanding the precise reasoning behind the model's decision-making can still be challenging. Interpretability is an ongoing research area for ViTs.

Despite these limitations, the advancements of ViTs in image classification demonstrate their potential as powerful alternatives to CNNs, particularly for tasks that benefit from capturing global dependencies and complex relationships within images. Ongoing research aims to address these challenges and enhance the capabilities of ViTs for a wider range of applications.

3.10 Application of ViTs in Diabetic Retinopathy Classification:

Studies have demonstrated the effectiveness of Vision Transformers (ViTs) in medical image classification tasks, although specific research on their application in diabetic retinopathy (DR) classification is limited. However, similar medical imaging studies have showcased the potential of ViTs. For example, ViTs have been successfully employed in classifying skin lesions Yonis Gulzar 2022 and identifying breast cancer Komorowski et al. 2023. While these studies focus on different medical conditions, they highlight the versatility and efficacy of ViTs in medical image analysis. Applying ViTs to DR classification holds promise for improving the accuracy and reliability of detecting DR-related abnormalities in retinal images.

3.11 Litterature review

Diabetic retinopathy (DR) classification is a crucial task for effective diagnosis and treatment of this prevalent eye disease. Several studies have investigated different methodologies to improve DR classification. These approaches encompass data mining, transfer learning, and other techniques, aiming to enhance early detection and diagnosis in the field of ophthalmology. In this literature review, we summarize key findings from various studies, focusing on methodologies, preprocessing techniques, model architectures, and performance evaluation metrics.

Authors: T. Jemima Jebaseeli 2018proposed a sustainable IoT technique for diagnosing diabetic retinopathy. Their system involved preprocessing, two-step segmentation using Kirsch's template and Fuzzy c-means clustering, and optic disc removal with Region-Based Active Contour. Performance was evaluated against ground truth images, achieving 99.58% accuracy, 72.51% sensitivity, and 99.83% specificity. Testing was conducted on 581 publicly accessible images from DRIVE, STARE, REVIEW, and HRF datasets.

Hayrettin Evirgen n.d. developed an automated system for diabetic retinopathy (DR) diagnosis using fundus images. The system focused on binary classification and employed data mining techniques. Nine image features were extracted from the Eye clinic of the Sakarya University Educational and Research Hospital databases. These features were used as input to a Naïve Bayes classifier, which predicted the presence of DR (0 for non-diabetic retinopathy, 1 for diabetic retinopathy). The new dataset, including the predicted user's state (10 features), was evaluated using 10-fold cross-validation. The system achieved an 89% accuracy in grading DR.

Houby 2021utilized the VGG-16 model, a pre-trained CNN, for transfer learning. The model was built using a publicly available Kaggle dataset. The preprocessing involved resizing the images to 224x224 pixels, cropping the retinal fundus image area to remove the background, and applying augmentation to address class imbalance. The augmented images were rotated by angles of 90°, 180°, and 270° and flipped. The preprocessed retina images were passed through VGG-16 with different convolutional layers based on the number of classes. The model achieved a maximum accuracy of 73.7% in classifying DR severity stages using five classes.

Akhilesh Kumar Gangwar 2021 applied transfer learning using a pre-trained hybrid Inception-ResNet model. The proposed method involved preprocessing the input color retina images from datasets APTOS2019 and Messidor-1, applying Inception-ResNet to them, augmenting the training dataset, and modifying the model architecture. The modified model included a custom CNN block, global max-pooling, fully connected layers, and a softmax layer for prediction. The model achieved an accuracy of 72.33% on 300 test images from Messidor-1 and 82.18% on 550 images from the APTOS dataset.

Chetoui et al. 2020introduces a novel approach for detecting diabetic retinopathy (DR) using the EfficientNet deep learning architecture. The researchers achieved state-of-theart results on two datasets, EyePACS and APTOS 2019, with an Area Under Curve (AUC) of 0.984 for referable diabetic retinopathy (RDR) and 0.990 for vision-threatening DR on EyePACS. On the APTOS 2019 dataset, they achieved an AUC of 0.966 for referable DR and 0.998 for vision-threatening DR. The paper also includes the development of an explainability algorithm, enhancing the interpretability of the proposed approach.

NEM. et al. 2019 compares different transfer learning architectures for detecting diabetic retinopathy (DR). The study found that the AlexNet model achieved the highest testing accuracy of 97.9% with dataset APTOS. This indicates the effectiveness of transfer learning models in improving the accuracy of DR diagnosis. The research suggests that utilizing pre-trained models can enhance the detection of DR, leading to better medical interventions for patients.

Taher et al. 2019 presents a computer-aided diagnosis tool based on convolutional neural networks (CNN) for classifying fundus images into the five stages of diabetic retinopathy (DR). The proposed CNN architecture includes preprocessing and convolutional layers followed by fully connected layers. By leveraging transfer learning on a large dataset and applying it to the APTOS 2019 Kaggle DR dataset, the tool achieves a testing accuracy of 77%

The reviewed studies highlight diverse methodologies for diabetic retinopathy (DR) classification. These approaches, such as data mining and transfer learning, aim to enhance accuracy and efficiency in DR diagnosis. Preprocessing, customized models, and performance evaluation contribute to improving early detection and diagnosis of DR, benefiting patients in ophthalmology. Further research in these areas holds promise for advancing DR classification and improving patient outcomes.

3.12 Conclusion

In conclusion, deep learning techniques, including the transformer model and Vision Transformer (ViT), have surpassed traditional machine learning approaches. The transformer model's self-attention mechanism has revolutionized language processing and proven effective in computer vision tasks by applying self-attention to image patches. ViT, a specialized implementation of the transformer model for image analysis, has outperformed traditional convolutional neural networks (CNNs) in certain cases. These advancements in deep learning, driven by the transformer and ViT models, have improved performance and demonstrated the power of self-attention and learning from large-scale datasets. Incorporating these techniques will continue to advance artificial intelligence in various fields.

Chapter 4

Implementation and results.

4.1 Introduction

this chapter focuses on the implementation and results of the classification of Diabetic Retinopathy (DR) using Deep Learning (DL). In this chapter, we present the implementation details and results, highlighting the preprocessing steps, model architectures, training procedures, and evaluation metrics used in the classification system.

Throughout this chapter, we demonstrate the steps taken to implement the DR classification system. we start from presenting the tools, libraries and framworks that we used. We discuss the preprocessing techniques employed, These techniques enhance the quality and suitability of the datasets for DR classification. We also describe the data augmentation techniques applied, which increase the diversity of training samples and address class imbalance issues. For the binary classification task, we utilize the Vision Transformer (ViT) model and optimize it using the Adam optimizer. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and AUC. Furthermore, we explore the fine-tuning of pre-trained ViT models for multi-class classification of DR. The chapter concludes with the presentation of the performance metrics, including test loss, precision, recall, accuracy, and AUC, for both the binary and multi-class classification tasks.

4.2 Frameworks , tools and libraries

Google Colab: Google Colab provides a cloud-based environment that allows researchers to execute and collaborate on code. It offers the convenience of running code directly in the browser without the need for local setup. Google Colab's notable feature is the provision of GPUs and TPUs, which significantly accelerate deep learning model training. This resource facilitated efficient experimentation and collaboration by providing access to powerful computing resources.

Figure 4.1: Collab logo

Kaggle: Data science platform for diverse datasets and community engagement Kaggle serves as a comprehensive data science platform with a vast collection of datasets across various domains. It played a crucial role in this research by providing access to diverse diabetic retinopathy datasets, allowing researchers to obtain labeled images for training and evaluation. Kaggle's competitions feature provided a platform for benchmarking the developed model against other data scientists. Additionally, Kaggle's community engagement aspect enabled researchers to participate in discussions, seek guidance, and share knowledge with fellow data scientists.

Figure 4.2: Kaggle logo

TensorFlow: TensorFlow, an open-source deep learning framework, was instrumental in developing the deep learning model for diabetic retinopathy classification. Its extensive set of tools, libraries, and functionalities facilitated the entire model development process. TensorFlow's flexibility allowed researchers to experiment with different model architectures, including traditional Convolutional Neural Networks (CNNs) and newer approaches like Vision Transformers (ViTs). The framework's scalability ensured efficient training and deployment of the deep learning model on various hardware configurations.

Figure 4.3: Tensorflow logo

Keras: Keras, a high-level API built on top of TensorFlow, simplified the implementation of the deep learning model. Its user-friendly interface and abstraction of low-level implementation details accelerated the model development process. Researchers could design and implement the model architecture using a concise and intuitive syntax. Keras provided convenient functions for data preprocessing, model evaluation, and optimization, allowing researchers to streamline these tasks. By leveraging Keras, researchers could focus more on model architecture design and experimentation.

Figure 4.4: Keras logo

Python: Python served as the language of choice for deep learning in this research. Its versatility, extensive ecosystem of libraries, and clean syntax made it a popular choice among data scientists. Python libraries such as NumPy and Pandas facilitated efficient data manipulation, allowing researchers to preprocess the diabetic retinopathy dataset effectively. Visualization and analysis of model performance were made possible through libraries like Matplotlib, Seaborn, and Plotly. Python's popularity and active community support ensured access to a wealth of resources and knowledge for deep learning endeavors.

Figure 4.5: Python logo

- **Efficient Data Manipulation with Libraries like NumPy and Pandas:** Python libraries such as NumPy and Pandas provide powerful tools for efficient data manipulation and analysis. NumPy offers multidimensional array objects and functions for numerical computations, while Pandas provides data structures and operations for handling structured data. These libraries were used to preprocess and manipulate the diabetic retinopathy dataset.

Figure 4.6: Numpy logo

- **Visualization and Analysis Using Matplotlib and Other Tools0:** Python's rich ecosystem includes libraries like Matplotlib, which enable researchers to visualize data, model performance, and training progress. Matplotlib provides a wide range of plotting functions and customization options for generating informative visualizations. Python also offers libraries like Seaborn and Plotly, which provide additional visualization capabilities to enhance the analysis of the deep learning results.

Figure 4.7: Matplotlib logo

4.3 System Design

To classify new diabetic retinopathy images using ViT, the process involves acquiring the dataset, preprocessing it, and splitting it into training and test sets. The ViT model is then trained on the dataset to learn to classify the images. Finally, the model is evaluated using the test set to assess its performance.

4.3.1 Dataset

Data Collection

We have used two datasets

(a) The first dataset **fundus_images** was acquired from Department of Ophthalmology, Hospital of Clínicas, Faculty of Medical Sciences (FCM), National University of Asunción (UNA), Paraguay., Paraguay.

(b) The second dataset the APTOS2019 dataset "Asia Pacific Tele-Ophthalmology Society (APTOS)" 2019 from the Kaggle 2019 competition.

Data Description

(a) The first dataset: The images set were of colored high-resolution retina images obtained using a Visucam 500 camera from the Zeiss brand and was classified by expert ophthalmologists. This database contains 396 color fundus images, and originally had seven categories, We proposed this structure (see figure 3.4)to replace the original (see figure 3.2) one due to the lack of data in the some classes (Mild -Severe - Proliferate) and we will be classifying 2 classes instead of 7:
i. No DR signs: The patient's retina is healthy and has no diabetic retinopathy (24 images).
ii. PDR: The patient's retina has diabetic retinopathy (372 images).

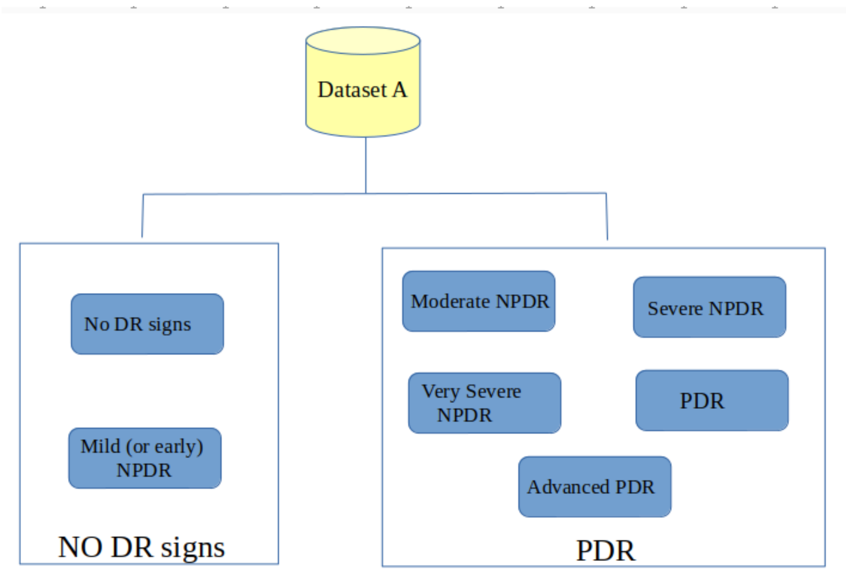


Figure 4.8: Medicl imaging system

Hence, the task is a binary classification problem.

(b) The second dataset:Aptos Similar to the first dataset, provided a large set of colored high-resolution retina images that were gathered from multiple clinics using different types of cameras and are of different sizes and appearances. The images were classified by the clinician according to the severity of DR into five levels as follows:

- 0 - No DR
- 1 - Mild
- 2 - Moderate
- 3 - Severe
- 4 - Proliferative DR

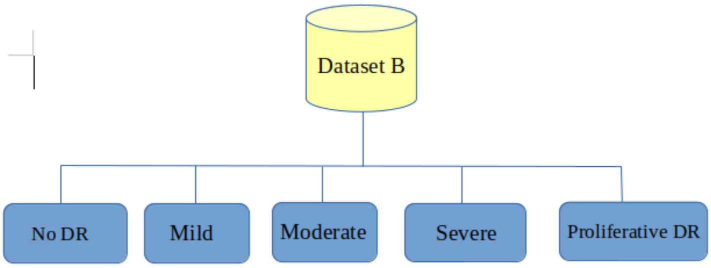


Figure 4.9: Medical imaging system

Hence, the task is a multiple-classification problem.

Data composition

(a) The first data set: (a)

CLASS	**N° of Images**
No DR signs	20
Mild(OR early)NPDR	4
Moderate NPDR	58
Severe NDPR	92
Very severe NPDR	83
PDR	70
Advanced PDR	69

Table 4.1: Original Dataset A Summary

CLASS	**N° of Images**
No DR signs	24
PDR	372

Table 4.2: New Dataset A Summary

(b) The second dataset: Aptos

Class	N° of Images
NoDR	1683
MILD	1744
mod	1797
SEVERE	1616
PD	1565

Table 4.3: Dataset Aptos Summar

4.4 Preprocessing

Both datasets (a) and Data Aptos underwent the following preprocessing steps.

1. Greenscale: The green channel extraction is performed in both datasets to focus on the green color information in the image. It can be useful in applications where green is of particular interest or when the green channel provides better visibility and less sensitivity to lighting conditions.
then The green images were converted to grayscale. This conversion simplifies the data representation by removing color information while retaining important structural features of the green channel.
the resulting image will appear in grayscale while retaining the original green channel values.

2. Resizing: The retinal images in both Datasets were resized to a standardized resolution, ensuring uniformity in image dimensions, it recommended resizing medical images to a square shape, the chosen size was 224 X 224

3. Median Filter: A median filter was applied to both Datasets to reduce noise and smooth the images, enhancing image quality.

4. Normalization: Pixel normalization was performed on both Datasets, scaling the pixel values to a standardized range, such as [0, 1].

5. Contrast-Limited Adaptive Histogram Equalization (CLAHE): CLAHE was utilized on both Datasets to improve local contrast and enhance the visibility of important features.

6. Rotation: To increase the diversity of data samples in Dataset (a), we applied rotation-based data augmentation. Each image was rotated five times using different angles, creating variations of the original images. This technique enhances the dataset by introducing different orientations and perspectives, improving the model's ability to handle various real-world scenarios.

6. Class sampling(fliping): In order to address the class imbalance issue in the "No DR signs" class, which had a relatively low number of samples in datase (a), we applied class sampling. Specifically, we employed a technique called flipping, where we horizontally mirrored the available data within that class. We performed 15 flips in total, effectively increasing the number of samples and helping to balance the dataset for improved model training and performance.

By applying these preprocessing steps to each dataset, the quality, and suitability of the datasets for DR classification were improved. These steps standardized the data, reduced noise, enhanced image quality, and increased the diversity of training samples. These preprocessing techniques played a crucial role in preparing the datasets for subsequent model development and evaluation.

Preprocessing Step	Dataset (a)	Dataset (b)
Greenscale Conversion	Applied	Applied
Resizing	Applied	Applied
Median Filter	Applied	Applied
Normalization	Applied	Applied
CLAHE	Applied	Applied
Rotation	Applied	Not Applied
Class Sampling (Flipping)	Applied	Not Applied

Table 4.4: Preprocessing Steps for Datasets (a) and (b)

4.4.1 Dataset Statistics after Preprocessing

Dataset (a):

Here are the statistics for the new dataset after applying the preprocessing steps as seen in The Dataset will be splitted into 3 subsets: training and validation, testing subset

Class	Number of Images
PDR	2232
No DR signs	2160

Figure 4.10: Class Distribution

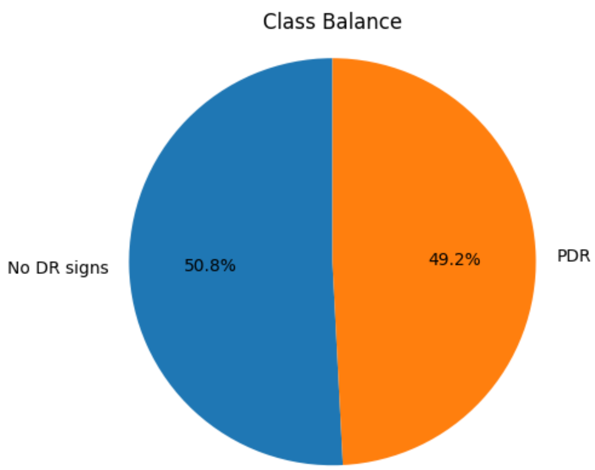


Figure 4.11: Class Distribution Chart

Dataset	Number of Images
Training Set	2810
Validation Set	703
Test Set	879

Table 4.5: Dataset Statistics

Dataset Aptos:

The Aptos dataset has been split into training, validation, and test sets. The following table displays the number of images used in each set

Dataset	Number of Images
Training	4601
Validation	913
Test	1681

Class	Number of Images
NoDR	1683
MILD	1744
mod	1797
SEVERE	1616
PD	1565

4.5 Binary classification using ViT

In this part, we investigate the first approach for training our model, which involves utilizing the Vision Transformer (ViT) architecture and optimizing it using the Adam optimizer, the dataset utilized here is the dataset (a)

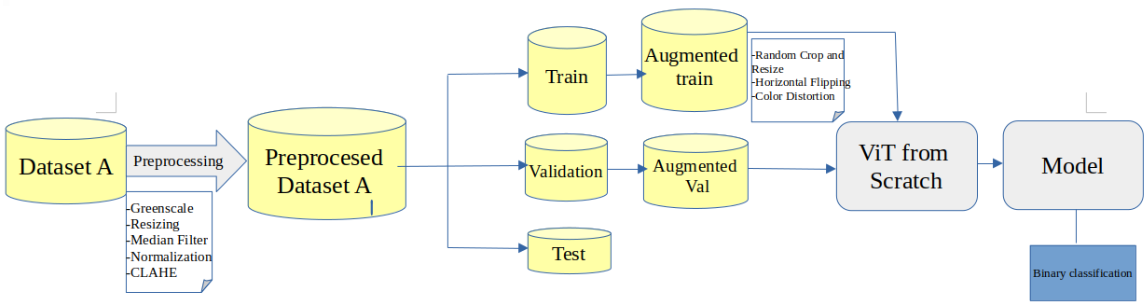


Figure 4.12: Design of system 1

As shown i the figure 4.12 The system 1 starts firstly by getting the dataset A, do preproccessing on the dataset, split dataset,and augment this dataset with certain techniques then we feed the ViT model with the augmented dataset, by the end we will have a model which can classify new diabetic retinopathy images to two classes, No DR & PDR.

4.5.1 Data augmentation

in our model, data augmentation is implemented using the transforms module from the PyTorch library. The transforms module provides a set of common image transformations that can be applied to the training dataset.
The following data augmentation techniques are applied :

Random Crop and Resize: This technique randomly crops and resizes the input image. By doing so, it allows the model to learn features from different regions of the image and handle variations in object sizes.

Horizontal Flipping: This technique randomly flips the input image horizontally. By introducing horizontal flips, the model learns features that are invariant to such flips, improving its ability to recognize objects regardless of their orientation.

Color Distortion: This technique randomly adjusts the brightness, contrast, saturation, and hue of the input image. It introduces variations in the image's color distribution, making the model more robust to changes in lighting conditions.

By applying these data augmentation techniques, we create variations of the original images, resulting in an augmented dataset with a greater number of samples compared to the original dataset. Data augmentation helps mitigate the risk of overfitting and improves the model's ability to generalize well to unseen data. It also enhances the model's robustness by exposing it to a wider range of variations and increasing its ability to handle diverse real-world scenarios.

4.5.2 Model Architecture

Self-Attention Mechanism: our model includes the MultiHeadAttention layer, which is a key component of the self-attention mechanism. It allows the model to attend to different elements in the sequence (image patches) to capture dependencies between them.

Patch Embeddings: The model we created includes the Patches layer, which divides the augmented image into fixed-size patches. These patches are then linearly projected into a sequence of token embeddings using the PatchEncoder layer. This step enables the model to process image information, Each patch represents a local region of the image and captures localized informatio

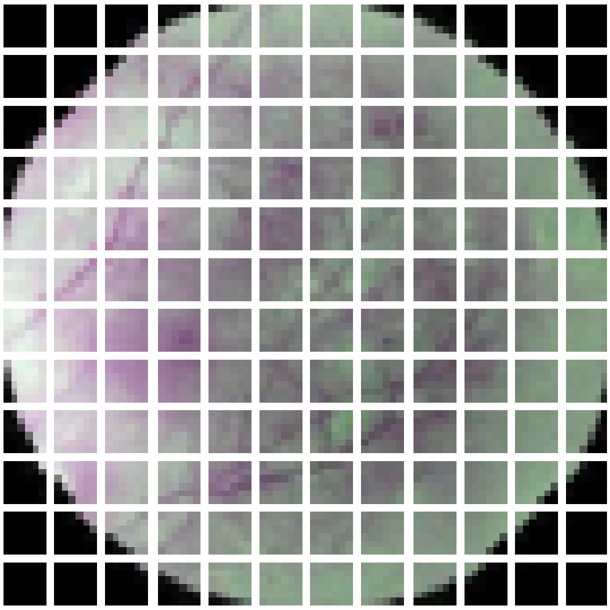


Figure 4.13: our model Retina image patched embeded

Patch encoder layer: next, The model defines a PatchEncoder layer that processes the sequence of patches using a Transformer encoder-like approach. Each patch is passed through a dense layer, followed by positional embeddings. The positional embeddings help the model capture spatial relationships between patches. The output of the PatchEncoder layer is a sequence of encoded patches, where each patch is represented by a vector in a higher-dimensional space.

Transformer layer: The encoded patches are then fed into a transformer layer, which consists of multi-head self-attention and feed-forward network operations. The self-attention mechanism allows the model to capture global dependencies and relationships between the patches.

Classification layers: The model defines an MLP layer that applies a multi-layer perceptron (MLP) to the output of the Transformer network. The MLP consists of multiple dense layers with dropout regularization between them. It aims to further refine the learned representations and extract high-level features. The final output layer of the model is a dense layer with sigmoid activation, which produces a binary classification output.

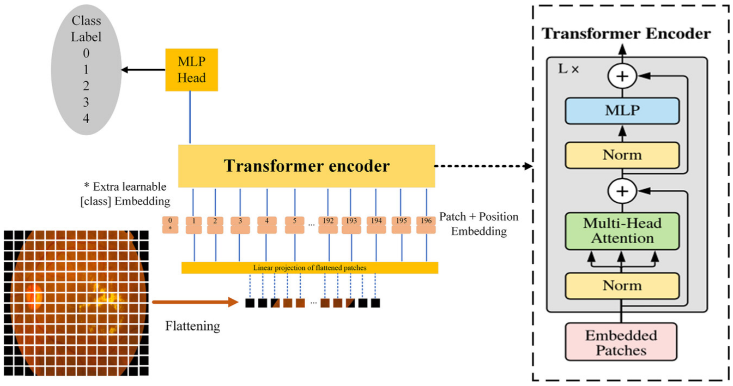


Figure 4.14: ViT architecture

4.5.3 Training and evaluation

Model 1 In this first structure, we propose the following set of parameters in the table 4.6

Parameter	Value
Learning Rate	0.0001
Weight Decay	0.0001
Epochs	100
Number of Patches	$(\text{Size/Patch Size})^2$
Projection Dimensions	64
Number of Heads	4
Hidden Units	[128, 64]
Output Units	[2048, 1024]
Optimizer	AdamW
Loss Function	Binary Crossentropy
Metrics	Precision, Recall, Accuracy, AUC

Table 4.6: Parameters Used in the Model

Once the models are trained, we evaluate their performances using metrics such as accuracy, precision, recall, F1-score, and the confusion matrix. These metrics provide insights into the models' abilities to accurately classify DR cases.

experiment	train-acc	train-loss	val-acc	val-loss	image-size	PATCH-SIZE
1	0.9768	0.2361	0.9777	0.2385	72	6
2	0.9777	0.1343	0.9685	0.1439	80	16
3	0.9726	0.1344	0.9406	0.1985	128	32

Table 4.7: Performace results and Hyperparameters of model 1

The model was evaluated on images of different sizes (72, 80, and 128 pixels) and patch sizes (6, 16, and 32 pixels).

The model achieved high training and validation accuracy, indicating its ability to accurately classify images. The training and validation losses are relatively low, indicating that the model has a good fit to the data.

Overall, the model demonstrates strong performance and shows sensitivity to variations in image and patch sizes, In addition to the provided evaluation results, additional performance metrics can be used to assess the model's performance provided in table 4.7 .

experiment	recision	AUC	Recall
1	0.9916	0.9953	0.9753
2	0.9780	0.9957	0.9753
3	0.9854	0.9919	0.9260

Table 4.8: Additional Performance Metrics results for model1

We created visualizations of accuracy and loss plots to easily analyze the performance of the models and observe trends in accuracy and loss over the training period.

- A plot of training accuracy and validation accuracy over 100 epochs (first experiment) :

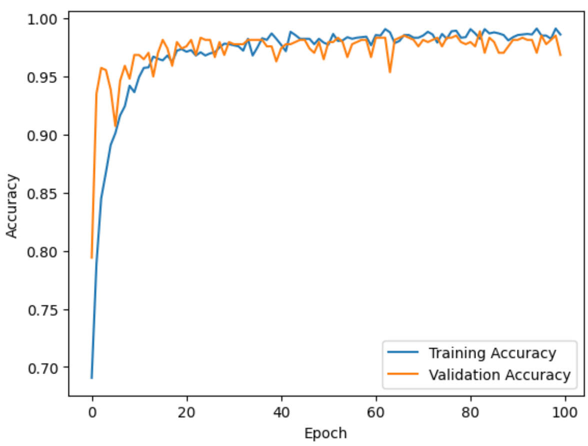


Figure 4.15: The model Accuracy

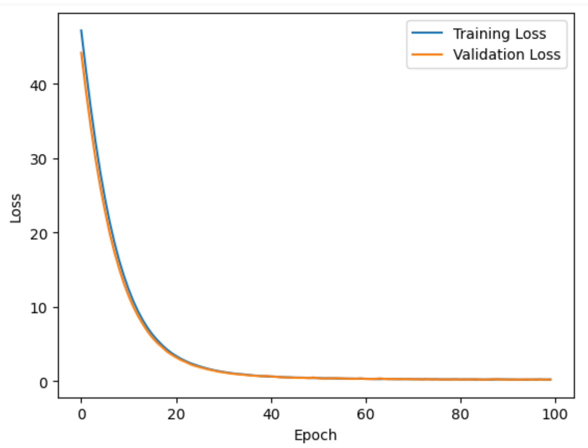


Figure 4.16: The model Loss

- Second experiment:

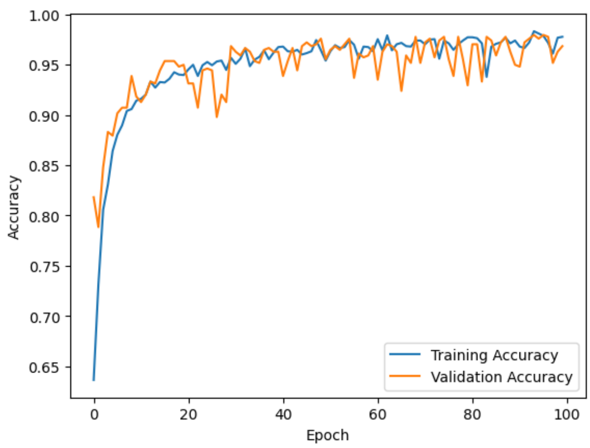


Figure 4.17: The model Accuracy

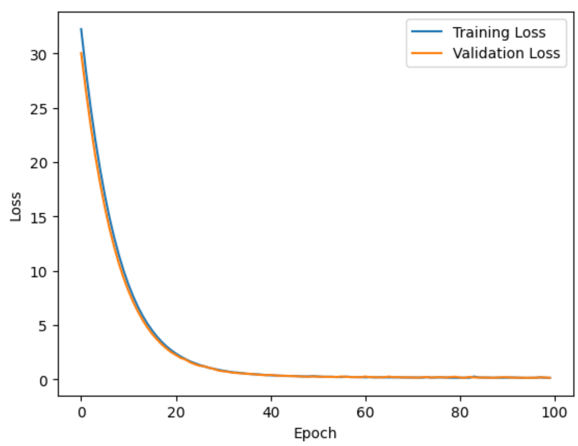


Figure 4.18: The model Loss

- Third experiment:

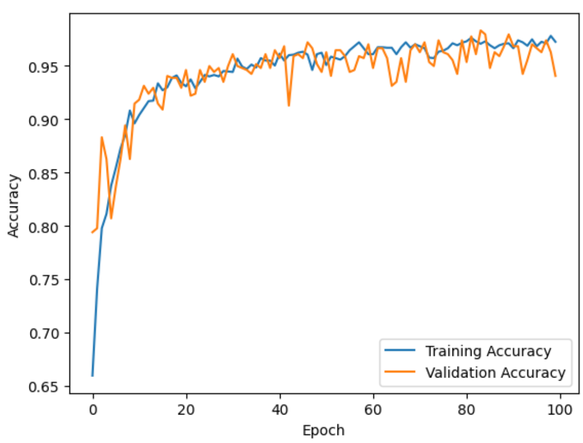


Figure 4.19: The model Accuracy

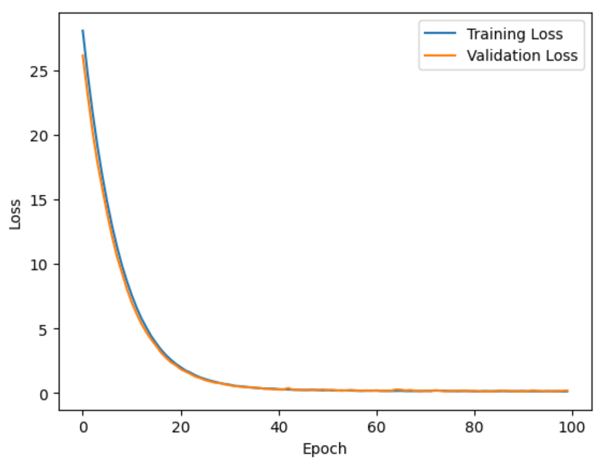


Figure 4.20: The model Loss

Model 2 in the second structure, The ViT (Vision Transformer) model was trained using the same parameters as the previous model, AdaBoost is introduced as the classifier instead of the activation function. The AdaBoost algorithm is a boosting technique that combines multiple weak classifiers to create a strong classifier.

In the implementation, the AdaBoostClassifier is utilized to train the model. This involves selecting a Decision Tree classifier as the base estimator, creating the AdaBoost model, and training it using the provided data. The AdaBoost algorithm combines multiple weak classifiers (Decision Trees) to form a stronger ensemble model. During training, the model iteratively adjusts the instance weights to focus on misclassified instances, improving its overall performance. By following this process, the AdaBoostClassifier is trained and ready for evaluation and prediction tasks, offering an effective solution for the classification problem at hand. here are the main parameters used to initiate the Adaabost algorithm as seen in table 4.8

To assess the performance of AdaBoost classifiers, we have experimented with various base estimators as shown in table 4.9. The experiment yielded consistent results, showing

Parameter	Value
`base_estimator`	`DecisionTreeClassifier(max_depth=1)`
`n_estimators`	50
`learning_rate`	1.0
`algorithm`	`SAMME.R`
`random_state`	42

Table 4.9: Parameters of the AdaBoostClassifier

Table 4.10: Accuracy Results for AdaBoost with Three Base Estimators

Base Estimator	Accuracy (%)
Decision Tree Classifier	99
Random Forest Classifier	99
SVC	99

a high accuracy of 99% across all three base estimators used in AdaBoost.

4.5.4 Testing model

During this phase, we will utilize the third subset to evaluate the performance of our ViT model(model 1) on new images from the dataset and assess its effectiveness.

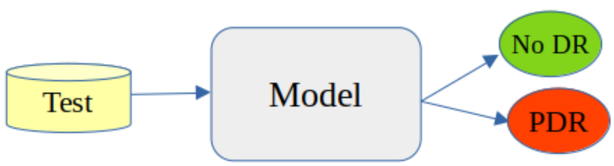


Figure 4.21: system for DR detection

4.5.5 Results and discussion

After testing our model on the third subset:

- Metrics results on test subset
- **Confusion matrix** The confusion matrix is a table that is often used to describe the performance of a classification model.

expriment	Test Loss	Test Precision	Test Recall	Test Accuracy	Test AUC
1	0.2322	0.9720	0.9858	0.9732	0.9976
2	0.1628	0.9764	0.9810	0.9732	0.9907
3	0.1697	0.9854	0.95744	0.9643	0.9894

Table 4.11: Test Metrics for model 1

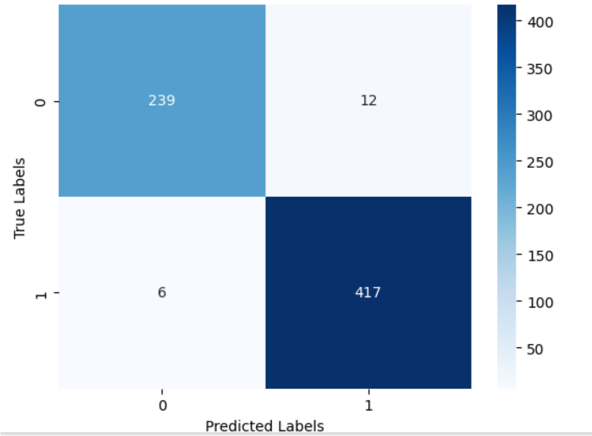


Figure 4.22: Confusion matrix first experiment

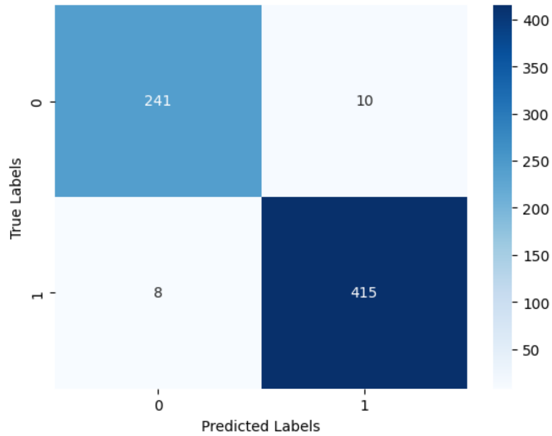


Figure 4.23: Confusion matrix second experiment

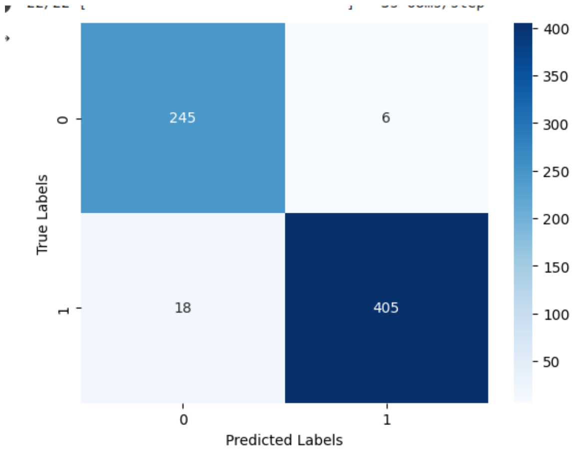


Figure 4.24: confusion matrix for third experiment

Experiment 1:

NO DR test folder: Out of 251 images, our model correctly predicted 239 images.

PDR test folder: Out of 423 images, our model correctly predicted 417 images.

Experiment 2:

NO DR test folder: Out of 251 images, our model correctly predicted 241 images.

PDR test folder: Out of 423 images, our model correctly predicted 415 images.

Experiment 3:

NO DR test folder: Out of 251 images, our model correctly predicted 245 images.

PDR test folder: Out of 423 images, our model correctly predicted 405 images.

These results indicate that our model achieved relatively high accuracy in predicting the presence or absence of diabetic retinopathy. It demonstrated consistent performance across the different experiments, accurately classifying a significant majority of images in both the NO DR and PDR test folders.

4.6 Multiclass classification using ViT

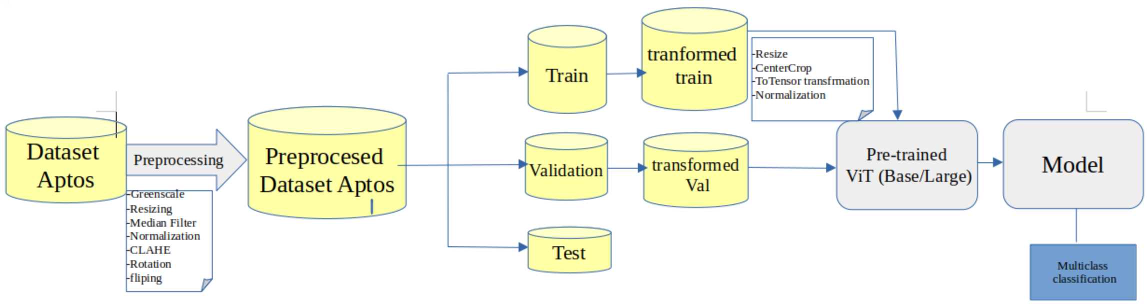


Figure 4.25: Design of system 2

In Figure 4.17, System 2 begins by obtaining the Aptos dataset. The dataset undergoes preprocessing, operation. It is then split into training, validation and testing subsets. The transformed dataset is fed into a pre-trained Vision Transformer (ViT) model. The ViT model has been trained on a large-scale dataset. we will use two pretrained ViT models At the end of the process, the system produces a trained ViT model capable of accurately classifying new diabetic retinopathy images into five classes, No DR & Mild & mod & Severe & PDR.

4.6.1 Data transformations:

a series of data transformations are applied to the input data. These transformations include:

Resize Transformation: The input data, typically images, are resized to a resolution of 128x128 pixels. This step ensures consistency in the size of the images for further processing.

CenterCrop Transformation: After resizing, the images are then cropped to a square shape with dimensions of 224x224 pixels. This center cropping technique allows for the extraction of the most important region or focal point of each image.

ToTensor Transformation: The preprocessed images are converted into PyTorch tensors. Tensors are the primary data structure used in PyTorch for efficient computation and compatibility with deep learning models.

Normalize Transformation: Finally, the tensor data is normalized by subtracting the mean values [0.485, 0.456, 0.406] and dividing by the standard deviation values [0.229, 0.224, 0.225] for each channel. This normalization process standardizes the input data, aiding in better convergence during model training.

By sequentially applying these transformations, the input data is processed and formatted to meet the requirements of chosen model.

4.6.2 Implementation Details:

For the multi-class classification of Diabetic Retinopathy (DR), we utilized two specific pre-trained Vision Transformer (ViT) models: ”ViT-base” and ”ViT-large.” Both models have undergone pre-training on large-scale datasets, allowing them to learn meaningful visual representations. These pre-trained models serve as powerful feature extractors for our DR classification task.

4.6.3 Fine-tuning the Pre-trained Models:

The process of fine-tuning involves adapting the pre-trained ViT models to the multi-class DR classification task. Initially, we keep the pre-trained layers frozen in both the ViT-base and ViT-large models to retain the valuable visual representations learned during pre-training. However, to make the models suitable for our specific classification task, we replace the original classification heads with new linear layers. These new layers are randomly initialized and trained alongside the pre-trained layers. Through this process, both models can learn task-specific features and make accurate predictions for different DR classes.

4.6.4 Training and Evaluation:

During the training phase, we fine-tune both the ViT-base and ViT-large models using our DR dataset. We specify hyperparameters such as the learning rate, batch size, number of training epochs, and weight decay for each model. The models are trained by optimizing their parameters using an optimizer such as AdamW and minimizing the Cross-Entropy Loss. This allows both models to adapt to the specific characteristics of the DR classification task.

preTrained	patches	Tr loss	Tr Acc	Val Loss	Val Acc	Prec	Recall	F1 Sc
ViTLarge	16	0.0174	98.97	0.0370	98.80	0.9881	0.9880	0.9879
	8							
ViTBase	16	0.0179	98.80	0.0423	98.47	0.9848	0.9847	0.9847
	8	0.0166	98.95	0.0521	98.14	0.9820	0.9814	0.9814

Table 4.12: Model Performance Metrics (Epoch 30)

We generated visualizations of accuracy and loss plots to facilitate the analysis of the models' performance and to observe trends in accuracy and loss throughout the training process.

- plots for accuracy and loss for both training and validation (ViT base 8 pretrained)

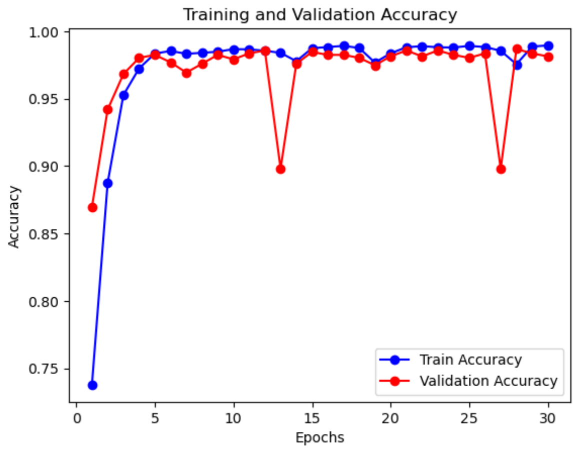


Figure 4.26: The model Accuracy

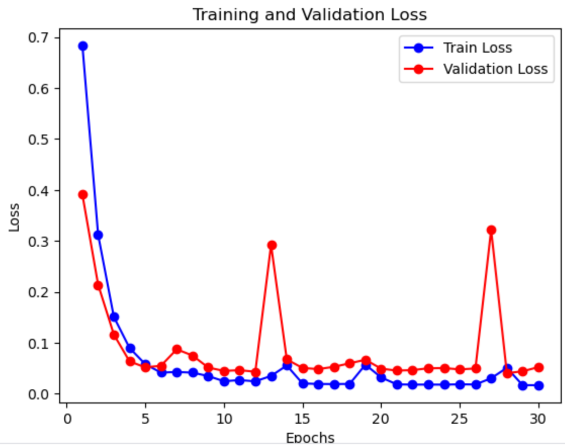


Figure 4.27: The model Loss

- ViT base 16 pretrained

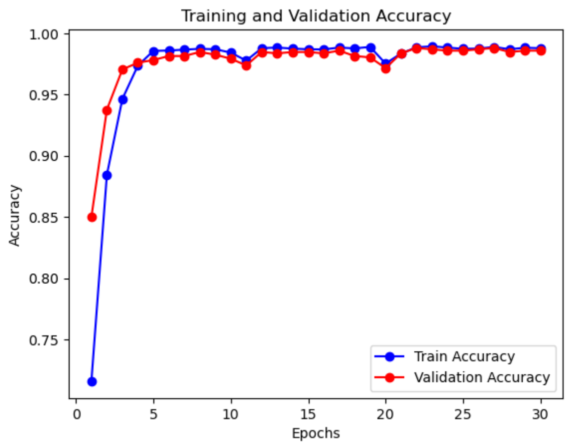


Figure 4.28: The model Accuracy

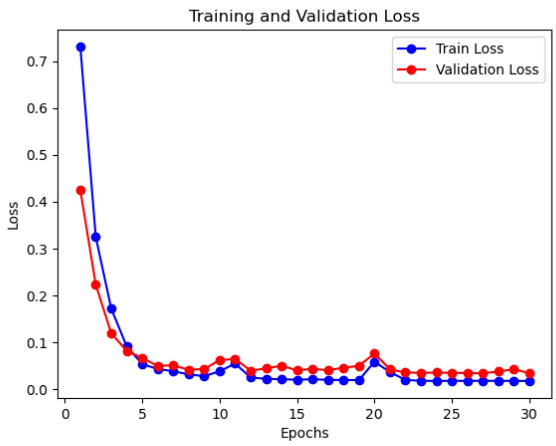


Figure 4.29: The model Loss

- ViT large 16 pretrained

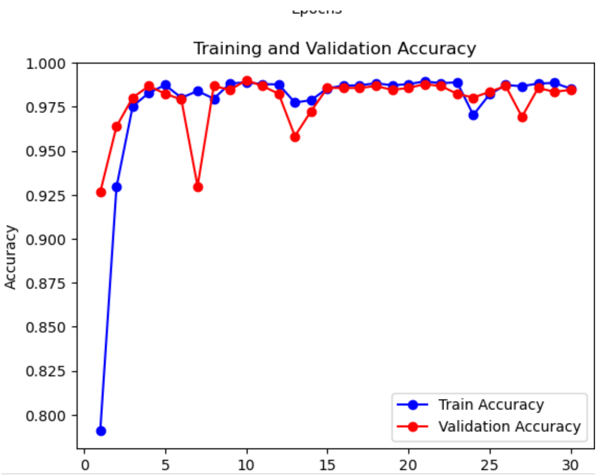


Figure 4.30: The model Accuracy

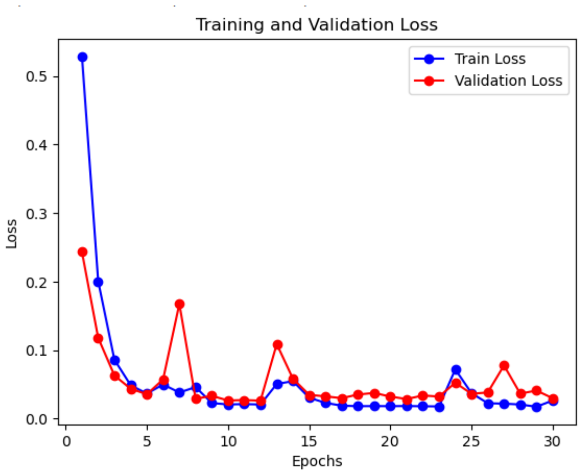


Figure 4.31: The model Loss

4.6.5 Testing model

a third subset was utilized to evaluate the performance of our model on new images

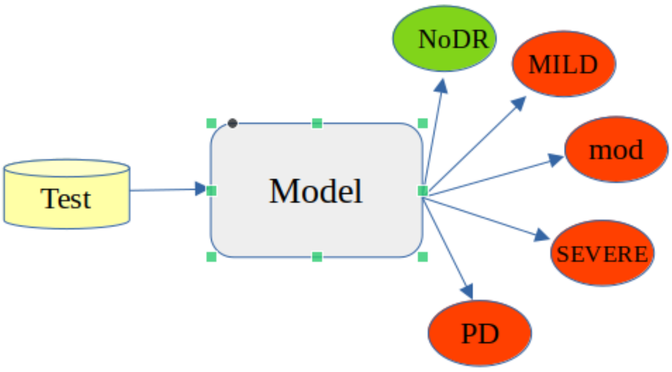


Figure 4.32: DR detection

4.6.6 Results and discussions

After testing our model on the third subset:

– Metrics results on test subset

model	Test Accuracy
VitLarge16	98.93
VitLarge8	
ViTBase16	98.75%
ViTBase8	98.39%

Table 4.13: Test Metrics for model

– **Confusion matrix**

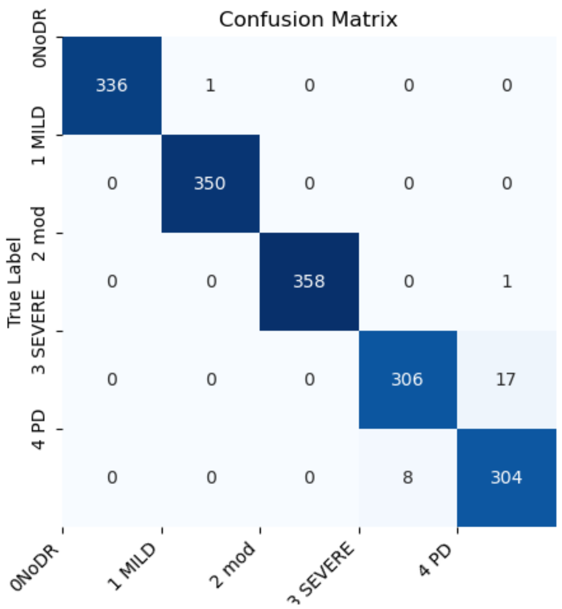


Figure 4.33: Confusion matrix: VitBase8

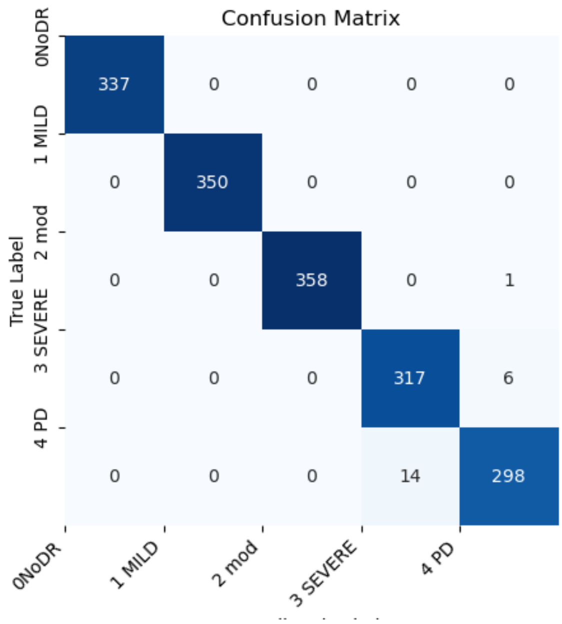


Figure 4.34: Confusion matrix: ViTBase16

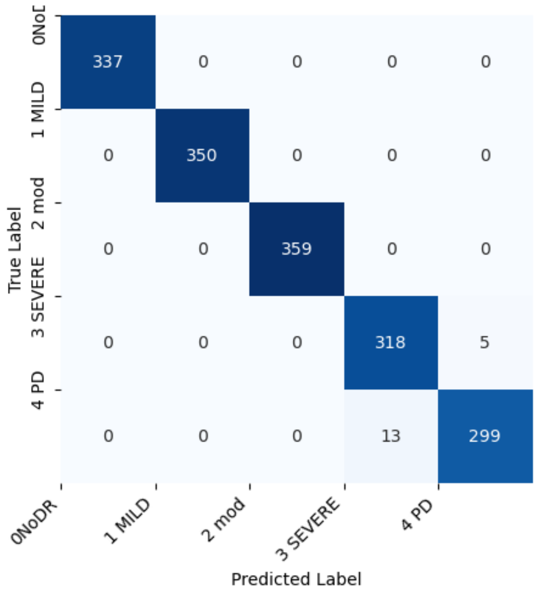


Figure 4.35: confusion matrix ViTLarge16

model ViT base8 pretrained
NO DR test folder: Out of 337 images, our model correctly predicted 336 images.
MILD test folder: the model predicted all 350 images correctly.
mod test folder: Out of 359 images, our model correctly predicted 358 images.
severe test folder: out of 323 images, our model correctly predicted 306 images.
PD test folder: out of 312 images, our model correctly predicted 304 images.

Model ViT base16 pretrained
NO DR test folder: Out of 312 images, our model correctly predicted 298 images.
MILD test folder: the model predicted all 350 images correctly.
mod test folder: Out of 359 images, our model correctly predicted 358 images.
severe test folder: out of 323 images, our model correctly predicted 317 images.
PD test folder:Out of 312 images, our model correctly predicted 298 images.

Model ViT large16 pretrained
the test folders: O DR, Mild and Mod predicted correctly al images
severe test folder: out of 323 images our model predicted correctly 318 images.
PD test folder: out of 312 images our model predicted 299 images

The pretrained ViT models (base8, base16, and large16) did a great job predicting the severity of retinal diseases in different test folders. The base8 model had high accuracy, while the base16 model had slightly lower accuracy. The large16 model performed exceptionally well, predicting all images correctly in some test folders.

4.7 Comparison of our work with previous works

In this section, we will compare our work with previous studies based validation accuracy.

Since we have modified the dataset (a)and changed it structure we will not compare its results with others works results. In deep learning, we compare results only when we use the same dataset

Model	Accuracy (%)
ViTbase8	98.39
ViTbase16	98.75
BiTlarge8	-
ViTlarge16	98.93
Previous Work 1NEM. et al. 2019	97.9
Previous Work 2 (Akhilesh Kumar Gangwar 2021)	82.18
Previous Work 3 Taher et al. 2019	77

Table 4.14: Comparison of Models with Previous Works on APTOS Dataset

Table 4.14 provides a comparison of our ViT models for multiclassification task in terms of their accuracy on the APTOS dataset. The accuracy values, presented as percentages, represent the performance of each model in classifying the dataset.

The ViTbase8 model achieved an accuracy of 98.39%, while the ViTbase16 model achieved a slightly higher accuracy of 98.75%. The BiTlarge16 model also performed well with an accuracy of 98.57%. Unfortunately.

the table includes the accuracy results from previous works for comparison. Previous Work 1 achieved an accuracy of 97.9%, indicating a slightly lower performance compared to the ViT models. Previous Work 2 obtained an accuracy of 82.18%, while Previous Work 3 achieved an accuracy of 77%. These values highlight the superiority of the ViT models in terms of accuracy on the APTOS dataset.

4.8 Conclusion

In summary, this chapter presented the implementation details and results of the Diabetic Retinopathy (DR) classification using Deep Learning (DL). Various pre-processing steps and data augmentation techniques were applied to enhance the datasets and improve model performance. The Vision Transformer (ViT) model showed promising results in accurately classifying DR cases. The findings highlight the effectiveness of DL-based approaches and provide insights for future research.

Conclusion and perspectives

4.9 Conclusion

The application of Deep Learning (DL) in healthcare has garnered significant attention from researchers, offering promising possibilities in various domains. Diabetic retinopathy is a disease caused due by complications of diabetes mellitus which can lead to blindness.

Throughout this dissertation, we have explored the utilization of Vision Transformers (ViT) for the detection and classification of diabetic retinopathy levels. Our research involved a comprehensive analysis of existing works and architectures in the field of image classification and diabetic retinopathy detection. Drawing from this knowledge, we proposed a variant architecture leveraging ViT to enhance the performance of the models. Two distinct datasets, SA 2021 and "Asia Pacific Tele-Ophthalmology Society (APTOS)" 2019, were employed for model training and validation and testing.

The experimental results revealed the effectiveness of Vision Transformers (ViT) in medical diagnosis tasks. Our proposed models demonstrated favorable performance in terms of multiclass classification using "Asia Pacific Tele-Ophthalmology Society (APTOS)" 2019, as well as binary classificationSA 2021 tasks. However, we acknowledge that further exploration of preprocessing techniques and other advanced deep learning methodologies was limited due to hardware and time constraints.

Looking ahead, this work paves the way for several future research avenues. Firstly, the improvement of the model's performance could be achieved by leveraging the potential of a third dataset. Additionally, exploring diverse deep learning techniques and preprocessing methods holds promise for enhancing the overall accuracy and robustness of the diabetic retinopathy classification system.

In conclusion, this study showcases the potential of Vision Transformers (ViT) in addressing the challenges of diabetic retinopathy classification. By building upon the existing knowledge and continuing further investigations, we can unlock new opportunities for accurate diagnosis and treatment in the field of healthcare.

Perspectives

Looking ahead, there are several perspectives for future research and development. Firstly, the refinement of the proposed ViT architecture can be explored by fine-tuning hyperparameters and conducting additional experiments to optimize its performance. Secondly, investigating the impact of different preprocessing techniques and data augmentation strategies could potentially enhance the model's robustness and generalization capabilities. Furthermore, exploring the interpretability of the ViT model can provide insights into the decision-making process and increase trust in its predictions.

These perspectives open up new avenues for advancing the field of diabetic retinopathy detection using Vision Transformer models. By addressing these challenges and exploring novel methodologies, we can contribute to the continuous improvement of automated diagnostic systems for better healthcare outcomes."

Bibliographie

Akhilesh Kumar Gangwar, Vadlamani Ravi (2021). "Diabetic Retinopathy Detection Using Transfer Learning and Deep Learning." en. In.

Ankur Saxena, Shivani Chandra (n.d.). *Artificial Intelligence and Machine Learning in Healthcare*. spriger.

"Asia Pacific Tele-Ophthalmology Society (APTOS)" (2019). en. In.

Authors: T. Jemima Jebaseeli C. Anand Deva Durai, J (2018). "IOT Based Sustainable Diabetic Retinopathy Diagnosis System." en. In: *Sustainable Computing: Informatics and System* 3.

Ayman S. El-Baz, Jasjit Suri (n.d.[a]).

— (n.d.[b]).

— (n.d.[c]).

— (n.d.[d]).

— (n.d.[e]).

Bengio, Ian Goodfellow Yoshua and Aaron Courville (2016). *Adaptive Computation and Machine Learning*. MIT Press.

Chetoui, M. and M. A. Akhloufi (2020). "Explainable Diabetic Retinopathy using EfficientNET." In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*.

Davenport T, Kalakota R. (2019). "The potential for artificial intelligence in healthcare." In: *Future Healthc J.*

Dosovitskiy, Alexey et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." en. In: *arXiv preprint arXiv:2010.11929*.

Hayrettin Evirgen, Menduh Çerkezi (n.d.). "Prediction and Diagnosis of Diabetic Retinopathy using Data Mining Technique." en. In: ().

Houby, Enas M.F. El (2021). "Using transfer learning for diabetic retinopathy stage classification." en. In: *Applied Computing and Informatics*.

Karel Zuiderveld, Paul S Heckbert (1994). *Graphics Gems IV*. Elsevier.

Komorowski, piotr, Hubert Baniecki, and Przemysław Biecek (2023). "Towards Evaluating Explanations of Vision Transformers for Medical Imaging." en. In.

Kumari, Somya et al. (2022). "Selfie fundus imaging for diabetic retinopathy screening." en. In: *Eye* 36.3, p. 6.

Lazzaro, Douglas R. (2021a). *Diabetes and the Eye*. Bentham Science Publishers Pte. Ltd. Singapore.

— (2021b). *Diabetes and the Eye*. Bentham Science Publishers Pte. Ltd. Singapore.

— (2021c). *Diabetes and the Eye*. Bentham Science Publishers Pte. Ltd. Singapore.

"LE DIABÈTE DANS LE MONDE" (2021). In.

Manaswi, Navin Kumar (2018). *Deep Learning with Applications Using Python.* apress.

NEM., Khalifa et al. (2019). "Deep Transfer Learning Models for Medical Diabetic Retinopathy Detection." In: *Acta Inform Med.*

Ophthalmology, Aravind Eye Hospitals Postgraduate Institute of (2007a). *Anatomy and Physiology of the Eye.* Graphico Associate Press.

— (2007b). *Anatomy and Physiology of the Eye.* Graphico Associate Press.

— (2007c). *Anatomy and Physiology of the Eye.* Graphico Associate Press.

SA, Castillo Benítez VE. Castro Matto I. Mello Román JC. Vázquez Noguera JL. García-Torres M. Ayala J. Pinto-Roa DP. Gardel-Sotomayor PE. Facon J. Grillo (2021). "Dataset from fundus images for the study of diabetic retinopathy." en. In.

Taher, O. Dekhil A. Naglah M. Shaban M. Ghazal F. and A. Elbaz (2019). *Deep Learning Based Method for Computer Aided Diagnosis of Diabetic Retinopathy.*

Vujović, Željko Đ. (2021). "Classification Model Evaluation Metrics." In: *International Journal of Advanced Computer Science and Applications,*

Willoughby, Colin E et al. (2010). "Anatomy and physiology of the human eye: effects of mucopolysaccharidoses disease on structure and function – a review." en. In: *Clinical & Experimental Ophthalmology* 38.3, p. 10.

Yonis Gulzar, Sumeer Ahmad Khan (2022). "Skin Lesion Segmentation Based on Vision Transformers and Convolutional Neural Networks—A Comparative Study." en. In.