**PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA MINISTRY HIGHER EDUCATION AND SCIENTIFIC RESEARCH KASDI MERBAH UNIVERSITY OUARGLA**

**Faculty of New Technologies of Information and Communication Department of Electronics and Telecommunication**


**Final Dissertation**

**Thesis for Master Degree in**
**Automatic and Systems**

**Thesis**

# Detection of Glaucoma on Fundus Images Using Deep Learning

**Submitted by:**

ABDELDJAOUAD SOHEYB & OUGUIS YAKOUB

**Defense Committee:**

| | | | |
|---|---|---|---|
| Dr. F. Charif | Prof | President | UKM Ouargla |
| Dr. A. Abimouloud | MCA. | Examiner | UKM Ouargla |
| Dr. A. Benchabane | MCA | Supervisor | UKM Ouargla |

**Academic year:** 2024/2025

# Dedication

## Abdeldjaouad Soheyb

I would like to dedicate this project to my family, who have been my constant source of love, support and inspiration throughout my academic path. Your unwavering encouragement and belief in my abilities have been invaluable, and I am grateful for the sacrifices you have made to help me pursue my dreams.

I would also like to express our gratitude to my professors, whose guidance and expertise have challenged me to grow and develop as a student. Your dedication to teaching and commitment to excellence have been a source of motivation and inspiration, and I are honoured to have had the opportunity to learn from you.

## Ouguis Yakoub

I dedicate this thesis to my beloved family, whose unwavering support, encouragement, and sacrifices have been the foundation of my academic journey.

To my parents, for their unconditional love and belief in my potential, even in moments when I doubted myself, and I dedicate it to my dear sister, who has departed this world.

To my teachers and mentors, for their guidance, inspiration, and wisdom that helped shape both my intellect and character.

And to all those who continue to strive for knowledge and understanding may this work serve as a small contribution to that noble pursuit.

# Acknowledgments

We would like to acknowledge some of our classmates and friends, whose camaraderie and support have made this dream become true. Their encouragement, collaboration and friendship have helped us to overcome hindrances and celebrate achievements, and we are grateful for the memories we have shared.

To our family and professors. Thank you for being a part of our educational background and for helping us to achieve our goals, and our standing at this honourable place is the best example for that.

Finally and without any slight of doubt, this project is a testament to the power of collaboration and perseverance, and we are proud to share it with you.

## Abstract

Detecting glaucoma at an early stage is crucial in preventing irreversible vision loss. Deep learning (DL), especially models like Convolutional Neural Networks (CNNs) and Vision Transformers (ViT), has demonstrated significant potential in medical image analysis. This study introduces a glaucoma detection system that leverages deep learning for fast, accurate, and reliable diagnosis using retinal fundus images. Three models are used in this system: ResNet50, Vision Transformer (ViT), and a hybrid model combining both.


**Keywords:** Glaucoma detection, Deep Learning (DL), Convolutional Neural Networks (CNNs), Vision Transformers (ViT), Retinal fundus images, ResNet50, Hybrid model



## Résumé

La détection précoce du glaucome est essentielle pour prévenir la perte de vision irréversible. L'apprentissage profond (Deep Learning), en particulier les modèles tels que les réseaux de neurones convolutifs (CNN) et les Vision Transformers (ViT), a montré un fort potentiel dans l'analyse d'images médicales. Cette étude propose un système de détection du glaucome basé sur l'apprentissage profond, permettant un diagnostic rapide, précis et fiable à partir d'images du fond d'oeil. Trois modèles sont utilisés dans ce système : ResNet50, Vision Transformer (ViT), et un modèle hybride combinant les deux.


**Mot-clé:** Détection du glaucome, Apprentissage profond (Deep Learning), Réseaux de neurones convolutifs (CNN), Vision Transformers (ViT), Images du fond d'oeil, ResNet50, Modèle hybride



**الملخص**

يُعد الكشف المبكر عن الغلوكوما أمرًا أساسيًا للوقاية من فقدان البصر غير القابل للعكس. أظهر التعلم العميق، وخاصة النماذج مثل الشبكات العصبية الالتفافية (CNN) ومحوّلات الرؤية(Vision Transformers - ViT) ، قدرة قوية في تحليل الصور الطبية. تقترح هذه الدراسة نظامًا للكشف عن الغلوكوما قائمًا على التعلم العميق، مما يتيح تشخيصًا سريعًا ودقيقًا وموثوقًا انطلاقًا من صور قاع العين. يعتمد هذا النظام على ثلاثة نماذجResNet50 :، محوّل الرؤية(ViT) ، ونموذج هجين يجمع بين الاثنين.



الكلمات المفتاحية: الكشف عن الغلوكوما، التعلم العميق، الشبكات العصبية الالتفافية(CNN) ، محوّلات الرؤية(ViT) ، صور قاع العين، ResNet50، نموذج هجين.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **DL** | Deep Learning |
| **ML** | Machine Learning |
| **NN** | Neural Network |
| **ANN** | Artificial Neural Networks |
| **CNN** | Convolutional Neural Networks |
| **CAD** | Computer-Aided Diagnosis |
| **RGCs** | Retinal Ganglion Cells |
| **FF** | Fundus Fluorescein (used in Zeiss FF retinal camera) |
| **IOP** | Intraocular Pressure |
| **VCDR** | Vertical Cup to Disc Ratio |
| **VCD** | Vertical Cup Diameter |
| **VDD** | Vertical Disc Diameter |
| **CCT** | Central Corneal Thickness |
| **RNFL** | Retinal Nerve Fiber Layer |
| **ViT** | Vision Transformer |
| **MLP** | Multilayer Perceptron |
| **Acc** | Accuracy |
| **SN** | Sensitivity |
| **P** | Precision |
| **ACRIMA** | Annotated Color Retinal Images for Automatic Glaucoma Detection |
| **DRISHTI-GS** | Drishti Glaucoma Screening |
| **LAG** | Labelled Axial Glaucoma dataset |
| **WHO** | World Health Organization |
| **ISBI** | IEEE International Symposium on Biomedical Imaging |

# General Introduction

Artificial Intelligence (AI) is a vital branch of computer science with a wide range of research applications. Its primary goal is to create intelligent systems capable of processing and learning from data to make informed decisions. At the core of AI lies machine learning (ML), which plays a pivotal role in various domains, particularly in medical diagnosis. Deep learning, a powerful subset of ML, emphasizes learning data representations through multiple levels of abstraction. Among deep learning architectures, Convolutional Neural Networks (CNNs) have proven especially effective in the field of image analysis and classification.

Glaucoma is a progressive and irreversible eye disease that leads to optic nerve damage and is one of the leading causes of blindness worldwide. Early detection is essential to prevent vision loss, yet diagnosing glaucoma can be challenging due to its asymptomatic nature in the early stages. Glaucoma often manifests without noticeable symptoms until significant damage has occurred, making it crucial to develop computer-aided diagnostic (CAD) systems that can support early and accurate detection.

This study implements a CAD system using advanced deep learning models such as Vision Transformer (ViT), ResNet-50, and a hybrid approach combining both. The models are trained and evaluated using three publicly available retinal fundus image datasets: ACRIMA, DRISHTI-GS, and LAG. These datasets contain labelled images of glaucomatous and non-glaucomatous eyes and provide a robust foundation for developing automated glaucoma detection methods.

Our work is organised into three main chapters: In the first chapter, we present general background information on glaucoma, its various types, diagnostic parameters such as intraocular pressure and cup-to-disc ratio, and the major challenges in early detection. In the second chapter, we detail the methodology used in our study, including a description of the datasets, the architecture of the models applied (ViT, ResNet-50, and Hybrid), and the training and evaluation process. In the third chapter, we present the results of our experiments using the three models on the three datasets. We evaluate the models using performance metrics such as accuracy, precision, and sensitivity, and compare their effectiveness in detecting glaucoma. Through this thesis, we aim to demonstrate the potential of deep learning-based CAD systems in supporting early and accurate glaucoma diagnosis, ultimately contributing to better clinical outcomes and reduced rates of vision impairment.

## CHAPTER 1

# Overview on glaucoma's disease

# 1.1 Introduction

Glaucoma is a group of optic neuropathies.It is characterized by progressive degeneration of retinal ganglion cells (RGCs).It shows the leading cause of irreversible blindness in the developed world.Typical signs with gradual loss of peripheral (side) vision that is followed by progressive loss of central vision.In case there is no treatment ,glaucoma can develop to whole blindness. The fact that the eye damage is slow and painless, only half of the carriers recognize the disease , and unrecoverable harm most of the time takes place long before diagnose [1].

In 2013, people aged 40-80 years and suffer from glaucoma was estimated to be 64,3 million-this number would reach to 76.0 million in 2020 and 111.8 million by 2040 [2]. It is also estimated that approximately 1.3 billion people live with some form of vision impairment. Another study projected that nearly 1.3 billion people live with some form of vision impairment. Glaucoma optic neuropathy is considered as the fourth major cause of vision impairment by the World Health Orgonisation (WHO). In 2015,people who went blind because of this disease would be approxatively three million [3].

Patients with glaucoma will gradually lose their peripheral (side) vision and may miss objects to the side. In this situation, they tend to be looking through an obscure pathway or tunnel. As it is illustrated in Figure 1.1 , it could be seen the difference between people with a normal vision and people with a glaucoma [4].

Figure 1.1: Illustration of decreased side vision due to a gradual progression of glaucoma [3].

Glaucoma, as a chronic eye disease, usually appears when fluid builds up intraocularly. That causes the intraocular pressure which damages the eye's optic nerve. Even the increasing level of aqueous humor provokes the intraocular pressure. That chemical reaction is also another reason for glaucoma. In a non-damaged eye, the two both amount of liquid: the produced and discharged by the eye have to be equal to maintain a balance. However, in glaucoma the liquid remains stagnant inside the eye. That increases stress on this latter and resulting damage in the optic nerve which is indispensable for brain-eye communication. The continuous Increase pressure leads to harmful destruction of optic nerve and probably ends up with irreversible blindness [5].

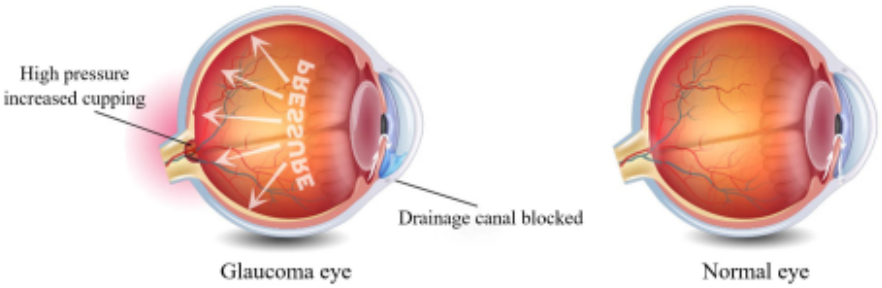



Figure 1.2: Glaucoma eye vs. healthy one [6].

As shown in Fig .1.2 The eye produces aqueous humor or vitreous fluid that continually supplies the ganglion cells of the eye. This liquid keeps on moving through the eyeball and the overflow is evacuated from the angel of the eye.

In case of the amount of aqueous level is higher than the healthy eye, eventually the person is at the gate of blindness/loses the sight/his or her sight. As a result, the blockade drainage canal plays a negative role in the existence of much liquid in the eye and increases pressure.

## 1.2 Types of glaucoma

### 1.2.1 Open-angle glaucoma

Open-angle glaucoma is the most common form of the disease. The drainage angle formed by the cornea and iris remains open, but the trabecular meshwork is partially blocked. This causes pressure in the eye to gradually increase. The increase in pressure damages the optic nerve. Blockage of the trabecular meshwork slows the drainage of aqueous humor out of the eye. This causes a buildup of fluid in the anterior chamber and gradual increase of intraocular pressure [7].

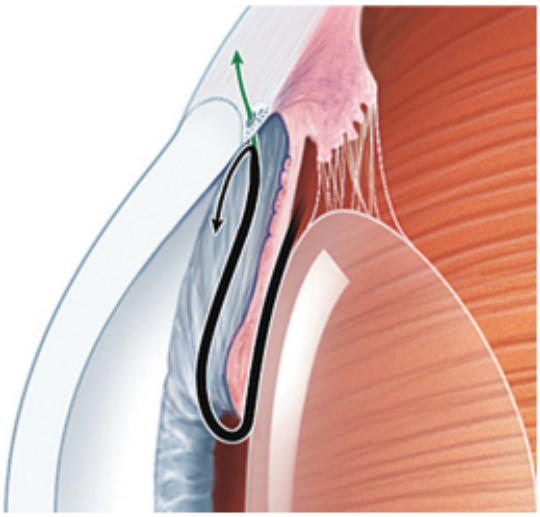


Figure 1.3: Open-angle Glaucoma eye [7].

### 1.2.2 Angle-closure glaucoma

Angle-closure glaucoma, also called closed-angle glaucoma, occurs when the iris bulges forward to narrow or block the drainage angle formed by the cornea and iris, making it more difficult for aqueous humor to drain away. With the acute form, the angle formed by the cornea and the iris closes completely, blocking aqueous humor from draining out of the eye. This causes a sudden, rapid increase in intraocular pressure [7].

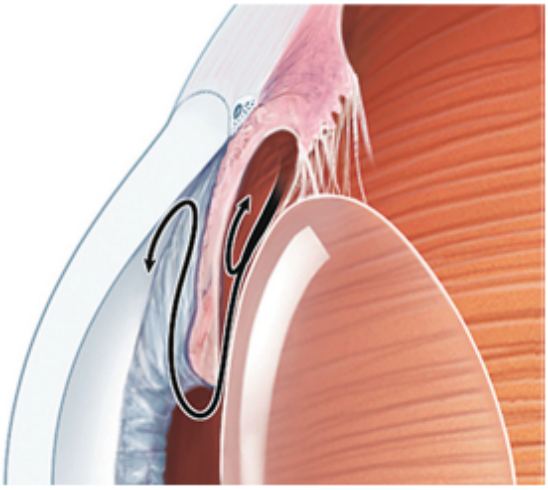


Figure 1.4: closed-angle Glaucoma eye [7].

This angle (in the Figure 1.4) is the pathway for aqueous humor to reach the trabecular meshwork. In a chronic form of this disease, the iris forms scar tissue over the meshwork, permanently closing off parts of the drainage pathway. As fluid backs up, pressure within the eye increases. Though not as prevalent as the open-angle form, angle-closure glaucoma is more common among farsighted individuals, who tend to

have smaller eyes, and among certain racial or ethnic groups, including Asians. As you get older, your lens becomes larger, pushing the iris forward and narrowing the angle [7].

### 1.2.3 Normal-tension

In normal-tension glaucoma, also known as low-tension glaucoma, your optic nerve becomes damaged even though your eye pressure is within the normal range. Damage to the optic nerve occurs in a pattern consistent with glaucoma. Normal-tension glaucoma is a poorly understood, though not uncommon, form of the disease. Why damage occurs is unknown. It's possible you may have a sensitive or fragile optic nerve, in which even normal amounts of pressure can damage it. Or your optic nerve may be receiving a reduced amount of blood. Limited blood flow may be associated with atherosclerosis the buildup of fatty deposits (plaques) in the arteries or other conditions that may impair circulation, such as abnormally low blood pressure [7].

### 1.2.4 Secondary glaucomas

Glaucoma is referred to as secondary when it's a complication of another medical condition. (Glaucoma is a primary condition when the cause is unknown.) Secondary glaucoma may stem from a variety of diseases, medications, physical injuries or trauma, and eye inflammation or abnormalities. Infrequently, eye surgery can cause secondary glaucoma. Examples of secondary glaucoma include pseudoexfoliative glaucoma, which occurs when small particles produced in the eye accumulate and block the trabecular meshwork, and neovascular glaucoma, associated with diabetes [7].

## 1.3 Signs and symptoms

Acute angle-closure glaucoma, you'll recall, results from a sudden and rapid increase in internal eye pressure. An attack often happens in situations when your pupils become dilated, such as at night in a room with dim lighting.

Emergency signs and symptoms include:

- Blurred vision.
- Halos around lights.
- Reddening of the eye.
- Severe headache or eye pain.

- Nausea and vomiting.

If any of these signs or symptoms occur, seek immediate medical attention. Permanent vision loss can happen within hours of the attack [7].

On the other hand, for open -angle glaucoma to reach an advanced stage only on the condition that it becomes symptomatic. If visual defects are present, they usually do not lie in the same part of the fields of the two eyes and are therefore replaced by binocular vision. Consequently, people with open-angle glaucoma report most of the time no symptoms [8],and a lot of do not know that they have the condition [9]. One-third of patients already have condition in an advanced or late stage in at least one eye at the time of diagnosis [10]. According to Gramer et al's report, 10-20% of patients were already incapable to drive a vehicle at the time presentation at the clinic because of binocular visual field defects. [11].

## 1.4 Glaucoma's disease diagnosis

There are several parameters to be considered in order to differentiate the healthy and glaucomatous eyes. Major parameters are as follows:

### 1.4.1 Intraocular Pressure

In non-damaged eye, central chamber of the eye is the producer of the aqueous fluid. This fluid moves around the lens, penetrates the drainage meshwork and then comes out of the eye. As illustrated in the figure 1.5 (a). If drainage meshwork is obstructed, fluid remains still inside the eye, as shown in figure 1.5(b). Thus, fluid level becomes higher inside the eye on one side. On the other side, the pressure increases inside the eye and damages appear at the level of optic nerve or retinal nerve fiber layers and in turn loss of vision, in this case, is inevitable. Tonometry, is the device that is used for measuring the pressure inside the eye. The unit of measurement of this device is millimeters of mercury (mmHg)-it is similar to blood pressure tests. (See the figure 1.6). In normal eye, IOP is usually between 10-20 mmHg. If IOP gets higher inside the eye, it may increase glaucoma. Meanwhile, in some cases, there are normal people may have high IOP, but they are not categorized as patients of glaucoma [12]. The following formula is used to calculate IOP of an eye.

$$IOP = [\frac{aqueos\ fluid\ formation\ rate}{out\ flow\ rate}] + venous\ pressure \tag{1.1}$$

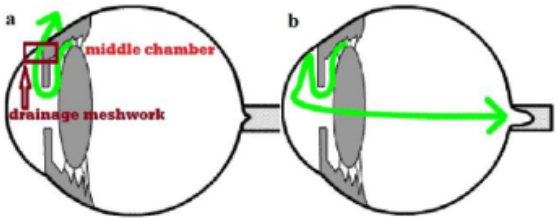



Figure 1.5: Process of retina (a) in healthy eye and (b) in glaucoma eye [13].

Figure 1.6: Tonometry measures intraocular pressure [7].

### 1.4.2 Vertical Cup to Disc Ratio

In the human eye, Inner portion of the optic nerve is called optic disk, as shown in figure 1.7.

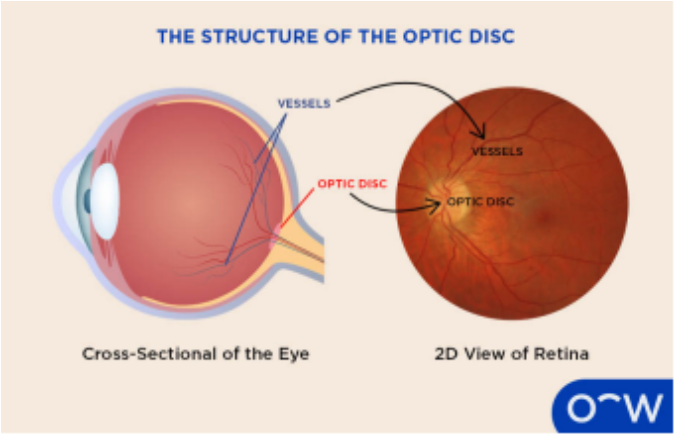



Figure 1.7: the structure of the optic disc [14].

The optic disk has an inner potion named the optic cup. Disk of the optic nerve is always larger than the cup in normal and glaucomatous eyes. The big difference between the two mentioned eye types regarding optic disk and cup is that disk size may be the same in both, but cup size is often larger in the presence of glaucoma eye compared to normal eye. Glaucoma can, therefore, be diagnosed by measuring the Vertical Cup to Disk Ratio (VCDR).Vertical cup to disk ratio is defined as the ratio of vertical cup diameter (VCD ) to vertical disk diameter (VDD).In normal eye, VCDR is always less or equal to 0.5 [15]. When an eye affected by glaucoma, the VCDR is greater than 0.5 This leads to loss of RNFL.

Ophthalmoscopy or Fundus camera is a device used to estimate VCDR. There are different brand fundus cameras available in the market like Zeiss FF retinal camera, portable Nidek NM-100, Canon fundus camera…etc. This kind of cameras is used to snap the eye images of patients. These taken images are called fundus eye images. After that, the estimation of the vertical cup diameter and the vertical disk diameter are based on the extraction of cup and disk areas from fundus eye images as indicated in the figure 1.8.Lastly, VCDR is calculated by the below formula [13].

$$VCDR = \frac{VCD}{VDD} \tag{1.2}$$

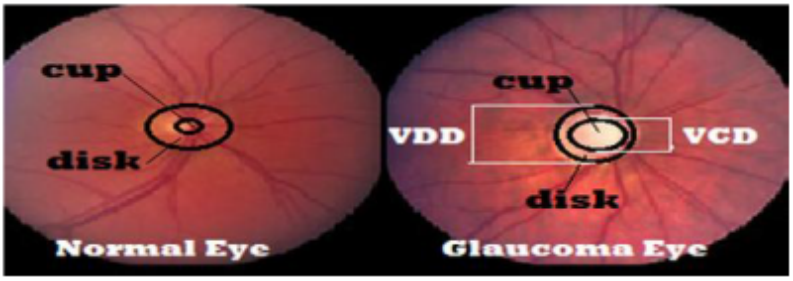



Figure 1.8: Healthy eye with small cup and glaucoma eye with larger cup [13].

### 1.4.3 Central Cornea Thickness

The cornea is a highly sensitive, transparent, dome-shaped layer that covers the front portion of the eye [16]. The thickness of the central cornea is a crucial factor in diagnosing and managing glaucoma [17]. According to the Ocular Hypertension Treatment Study, there is a connection between central corneal thickness (CCT) and intraocular pressure (IOP). Individuals with thicker CCT may show a normal IOP reading (10–20 mmHg), while those with thinner CCT might have a higher IOP readings when measured by tonometry (10–20 mmHg). This suggests that elevated IOP or reduced CCT can be indicators of glaucoma [18], whereas lower IOP or increased CCT generally points to a healthy eye.

Currently, the thickness of the central cornea is measured using ultrasonic pachymetry and non-contact devices like Orbscan II and SP3000P. In healthy eyes, CCT typically ranges from 0.50 to 0.54 mm [19, 20], while in eyes with glaucoma, it is usually less than 0.50 mm.

Figure 1.9: A patient taking a Central Cornea Thickness test by SP3000P instrument [21].

### 1.4.4 Visual Field

Visual field of a human eye is the total area covered by the eye in which objects are visible from peripheral or side vision. Visual field differs from person to another. There is a technique, in this point, called a visual field test or perimetry test measured by Goldmann Perimetry device. This common device is used by a very considerable number of ophthalmologists in order to detect, monitor and examine the glaucoma disease. It can also diagnose other diseases . The Goldmann Perimeter Device Working Principle:

- First, patient sits and looks inside the perimeter
- Light is flashed on one particular spot of eye.
- Next, a button is pressed by patient once he sees the flash.
- Device records both spots in which patient pressed the button and spots in which patient did not press the button.
- Finally, the vision loss areas are shown by the device. These are the areas in which patient did not press the button.

In healthy eye, vision loss areas are nearly non-existent. However, in glaucoma, vision loss areas are remarkable. In such a case, visual field test facilitates the detection of glaucoma disease at an early stage [13].

Figure 1.10: A patient taking a visual field test [22].

## 1.5 Challenges in the Diagnosis of Glaucoma

Glaucoma, a progressive and multifactorial optic neurodegenerative disease, still poses significant challenges in both diagnosis and management and remains a perpetual enigma [23].

### 1.5.1 Challenges to Early Diagnosis

#### Asymptomatic nature

Usually, there are not any signs produced by glaucoma in its early stage, In glaucoma , the gradual loss of side and central vision goes unnoticed because the primary cortical region of the brain ( the visual cortex )"fills" in the missing parts. As a result, diagnosis and treatment are delayed until glaucoma reaches a moderate or severe stage with significant vision loss [24].

#### Variation in the appearance of the optic disc

To diagnose glaucoma in its earliest stages, which is considered as the biggest challenge that doctors face, is the significant variation in the emergence of the optic disc and peripapillary region found in the normal eye. Another challenge that doctors face, and it is summarized in some anomalous optic discs that are too complicated or impossible to be set apart from glaucomatous discs. Eyes with such optic disc features may have normal visual fields , which are virtually undistinguishable from glaucomatous field defects [25].

### 1.5.2 Population screening limitations

The issue of glaucoma is contraversial. Glaucoma screening is expensive and may not detect many new cases. Targeting high-risk groups or adding checks routine exams, like driving tests, could improve early detection and reduce costs [26].

This challenge is particularly evident in developing nations, where less than 10% of those with glaucoma are detected. In such settings, population□based screening is even more impractical due to the relatively low prevalence of the disease, limited resources, and significant logistic barriers [27].

### 1.5.3 Measurement of intraocular pressure

Although intraocular pressure (IOP) is a key factor in diagnosing glaucoma, it has low sensitivity for detection, as many patients show structural and functional damage from glaucoma even when their IOP levels are within the normal range (below 21 mm Hg) [28, 29, 30, 31, 32, 33] Due to this limitation, greater focus has shifted toward examining the optic disc and conducting visual field tests to more accurately diagnose glaucoma.

## 1.6 Conclusion

This chapter has provided an overview of glaucoma detection, highlighting the complexities and challenges inherent in identifying this sight-threatening disease in its early stages. By examining various types of glaucoma, key diagnostic parameters, and the limitations of current screening methods, this work underscores the need for improved strategies to facilitate timely and accurate diagnosis.

## CHAPTER 2

# An Overview on Artificial Intelligence (AI)

## 2.1 Introduction

With the rising world of big data, the world has become forced to develop effective strategies that help us make informed decisions. All professional fields have become extremely complex due to the growing demand to find the most efficient ways to extract value from the existing data. With the advent of a technology called artificial intelligence, it has enabled it to keep pace with the naturally growing big data movement [34].

Artificial Intelligence (AI) refers to the capability of machines to learn and make decisions in a way that mimics human intelligence. As a prominent branch of computer science, AI focuses on simulating intelligent behaviour in computer systems. It enables machines to perform tasks that typically require human cognition, such as speech recognition, visual perception, decision-making, and language translation. AI has significantly impacted a wide range of consumer products and has driven major advancements in fields such as physics and healthcare [34].

## 2.2 Machine learning (ML)

Machine Learning (ML) is a subfield of Artificial Intelligence (AI), all machine learning approaches are part of AI,while not all AI techniques involve machine learning. ML focuses on developing and utilising algorithms to create generalised models capable of identifying patterns and making accurate predictions. These algorithms are typically grounded in mathematical and statistical optimisation techniques [34].

According to Arthur Samuel's 1959 definition, machine learning is the process of feeding data into a computer system in a way that enables it to learn and perform tasks in the future without being explicitly programmed or requiring additional input. In layman's terms, this means that computers will be able to develop their own "mind" and will be able to "think" for themselves [35].

### 2.2.1 Types of machine learning

Machine learning algorithms are trained by utilising labelled, unlabelled or hybrid datasets. Consequently, the nature of the requisite data for a particular task dictates the archetype of machine learning model that is formulated. Consequently, four fundamental categories of machine learning have been established developed as [36], [37], [38], [39] :

## Supervised Learning

Supervised learning is a machine learning approach used to learn a function from labeled training data. This training data is composed of input-output pairs, where the inputs are typically represented as vectors and the outputs are the corresponding desired values. The predicted output can be a continuous value known (as regression), or a discrete class label, known (as classification). The objective of a supervised learning model is to accurately predict the output for any new, valid input after after being trained on a sufficient number of examples (i.e. pairs of input and target output) [40].

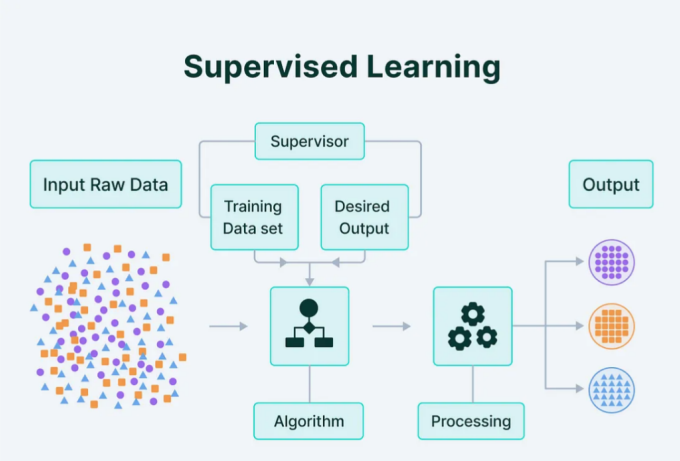



Figure 2.1: Supervised Learning [41].

## Unsupervised Learning

Unsupervised learning involves training a machine using only input data, without any known output labels, using only input samples or labels, it identifies underlying patterns in data and creates its own data clusters. The most common types of unsupervised learning algorithms are clustering and association analysis. This method is particularly useful for discovering hidden patterns in datasets, such as in recommendation systems for online stores that often employ clustering techniques [38], [42].

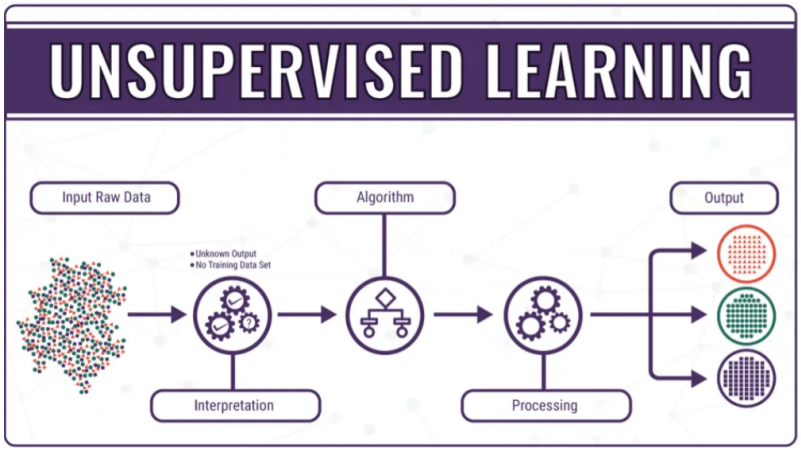



Figure 2.2: Unsupervised Learning [43].

## Semi-Supervised Learning

Semi-Supervised Learning (SSL) is a powerful technique that combines supervised and unsupervised learning to improve learning accuracy [44]. In this case, the input data is jumbled and consists of a mixture of labeled and unlabeled cases, with no discernible pattern. The model must learn the structures that are required to not just organize the data, but also to generate predictions based on the information. Classification and regression difficulties are two examples of this type of problem. Extensions are algorithms that are used as examples [35].

## Reinforcement Learning

Reinforcement Learning (RL) is a distinct type of machine learning that allows machines or software agents to determine the best actions in order to reach desired goals [45]. Instead of following predefined instructions, RL uses a trial-and-error method, where the agent learns by receiving feedback that reinforces a good actions and discourages the poor ones. Over time, as the agent gains more experience, it improves its decision-making skills to better accomplish its goals [46], In this process, the agent interacts with a changing environment, learning to reach its objective through rewards and punishments, without direct instruction from a teacher [42].

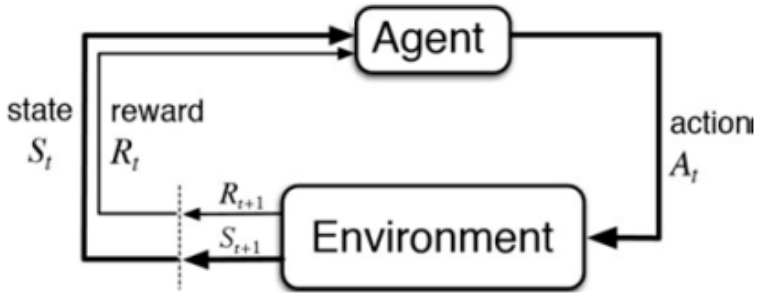



Figure 2.3: Reinforcement Learning [44].

### 2.2.2 Neural Networks (NN)

A neural network (NN) is a subset of machine learning that mimics the human brain designed to recognize patterns and make decisions based upon data. Neural networks consist of layers of nodes, sometimes called "neurons," that are connected in a manner similar to neural networks in the human brain. Each connection between nodes has an associated weight that is modified during learning. The broad objective of the neural network is to result in minimal error at its predictions or classification; this would call for tuning these weights based on the given data [47].

#### 1. Architecture

The architecture generally comprises an input layer, one or more hidden layers, and an output layer that constitutes the neural network: the input layer takes the raw data, followed by the hidden layers for processing, and eventually an output. In training, it is fed with a large dataset, and connections are varied to enhance the performance of the network in making good predictions. This process is similar to how the human brain strengthens or weakens synaptic connections depending on learning and experience [47].

- **Input layer**: is the first layer of the network, responsible for receiving raw data or features. Each neuron in this layer corresponds to one feature of the input data.
- **Hidden layers**: Hidden layers are situated between the input and output layers of a neural network and are referred to as "hidden" because their inputs and outputs are not directly observable. These layers enable the model to learn and identify patterns within the data by processing the inputs and adjusting internal parameters (weights) during training. This learning process allows the network to build increasingly abstract representations of the data, improving its predictive performance.
- **Output layer**: The output layer produces the final result of the neural network's operation. Depending on the task, this could be, for example, identified word categories in a text classification task or

detected objects in an image recognition task.

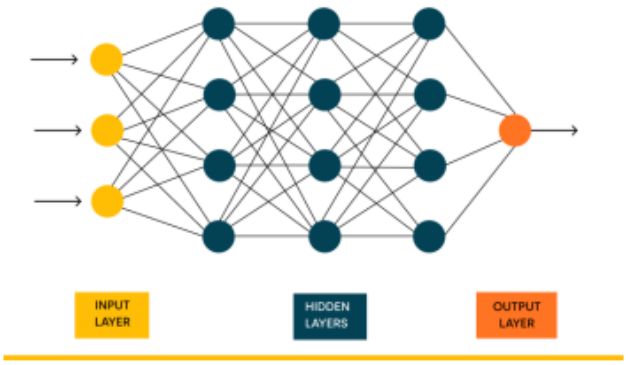



Figure 2.4: Artificial Neural Network (ANN) [48].

## 2. Learning and Training

- **Forward Propagation**: Forward propagation is the process where a neural network transforms input data into predictions or outputs. It can be considered as as the "thinking" phase of a neural network,when shown an input (like an image or text), forward propagation is how the network processes that information through its layers to produce a result. In technical terms, it's the sequential calculation that moves data from the input layer, through hidden layers, and finally to the output layer. During this journey, the data is transformed by weighted connections and activation functions, allowing the network to capture complex patterns [49].

- **The loss function**: The loss function, also known as the error function, is a vital element in machine learning that measures the difference between the model's predicted outputs and the actual target values. The computed value (the loss), indicates how accurately the model is performing, lower loss values signify more accurate predictions.

  The resulting value, the loss, reflects the accuracy of the model's predictions. During training, a learning algorithm such as the backpropagation algorithm uses the gradient of the loss function with respect to the model's parameters to adjust these parameters and minimise the loss, effectively improving the model's performance on the dataset [50].

- **Backpropagation**: Backpropagation is a method used to train neural networks where the model learns from its mistakes. It works by measuring how wrong the output is and then adjust the weights and the biases step by step to make better predictions next time. It is based on gradient descent and updates weights by minimising the error between predicted and actual output. Training of backpropagation consists of three stages [51]:

  - Forward propagation of input data.

- Backward propagation of error.

- Updating weights to reduce the error.

### 3. Activation Function

Activation functions are essential components of neural networks, they are a mathematical functions used in neural networks to determine whether a neuron should be activated or not. It introduces non-linearity into the model, allowing neural networks to learn complex patterns beyond simple linear relationships. [52].

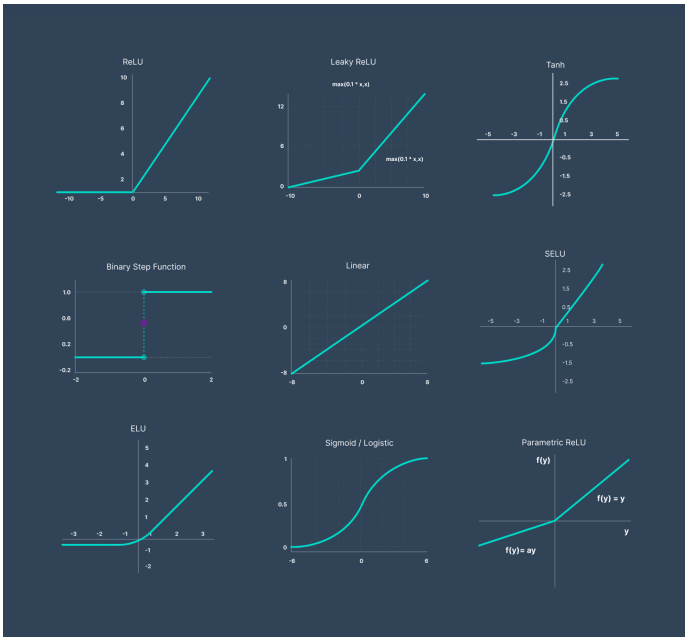



Figure 2.5: Some activation function [53].

## 2.3 Deep learning (DL)

Deep learning is a subset of artificial intelligence that focuses on learning from large datasets [54][55]. It utilises multiple layers of neural networks, composed of complex structures and nonlinear transformations, to extract high-level features from data [56]. with the ongoing development of big data and Computational power, deep learning has witnessed a signficant increase of applications in various fields, including image processing [57], natural language processing [58], speech recognition [59], and data mining [60] …etc, it has significantly outperformed traditional machine learning techniques and domain-specific algorithms.

### 2.3.1 Convolutional neural network (CNN)

Convolutional Neural Networks (CNNs) are advanced artificial intelligence models built on multi-layer neural network architectures, designed to identify, recognize, classify, detect, and segment objects within images. In fact, ConvNet (CNN) are widely used as discriminative deep learning architectures capable of learning directly from raw input data, without the need for manual feature extraction (human intervention) [61–63].

This network is widely applied in areas such as visual recognition, medical image analysis, image segmentation, natural language processing (NLP), and various other fields, as it is specifically built to handle different types of 2D structures [61][62][63]. It outperforms traditional networks by automatically detecting important features from the input data, eliminating the need for manual intervention.

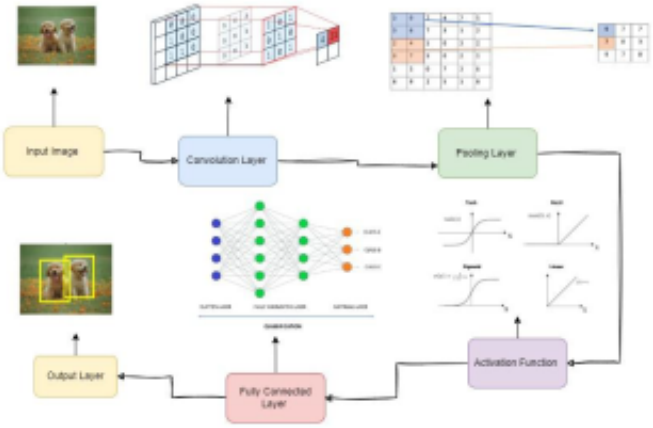



Figure 2.6: Conceptual model of CNN [64].

#### CNN Components

A CNN is typically composed of four types of layers:

- **Convolutional layers:**

  The convolutional layer is a fundamental part of the CNN architecture. It is a set of kernels (filters), that are applied to the input data to extract meaningful features. Each kernel has a defined width, height, and set of weights, which are initially initialised randomly but progressively become more informed through training. In other words, the convolutional layer generates a feature map by combining the input image (represented as an N-dimensional matrix) with these kernels [61][62][63].

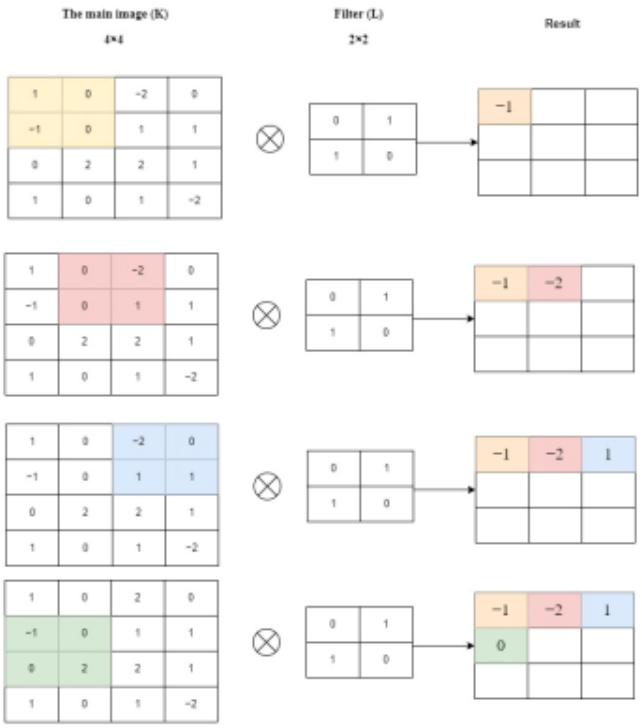



Figure 2.7: convolutional layer [64].

- **Pooling (sub-sampling) layer**:

  The pooling layer, also known as the down-sampling layer, serves to reduce the spatial dimensions of feature maps while maintaining the most critical information. This is achieved by sliding a filter across the input and applying a pooling operation, such as max, min, or average pooling. Among these, max-pooling is the most widely used method in existing research.

  The key purpose of pooling is to lower the computational burden on subsequent layers through down-sampling, it may be comparable to reducing the resolution. Importantly, pooling does not change the number of filters. In max-pooling, the image is segmented into rectangular regions, and the highest value within each region is selected. A $2 \times 2$ window is one of the most commonly used configurations for max-pooling [64].

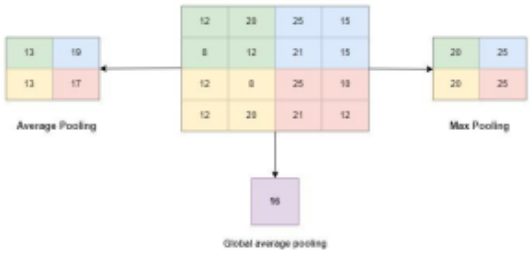



Figure 2.8: Pooling layer [64].

- **Non-Linearity (Function of Activation)**:

The primary role of an activation function in any neural network model is to transform the input into an output. This input is typically derived by computing the weighted sum of a neuron's inputs with an added bias (if present). In other words, the activation function determines whether a neuron should activate, or "fire" for a given input by producing an appropriate output signal [65].

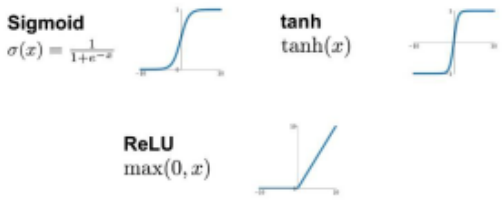



Figure 2.9: Some non-linear Function of Activation [64].

- **Fully Connected Layer (FC layer)**:

  In most CNN architectures designed for classification tasks, the final section typically consists of fully connected (FC) layers. In these layers, every neuron is linked to all neurons in the preceding layer. The last of these fully connected layers acts as the output layer, serving as the classifier for the CNN model. Fully connected layers are a type of feed-forward artificial neural network (ANN) and are based on the structure of traditional multi-layer perceptrons (MLPs). These layers receive input from the last convolutional or pooling layer in the form of feature maps. These feature maps are flattened into a single vector, which is then passed through the fully connected layers to produce the final output of the CNN, as illustrated in Figure 2.10 [65].

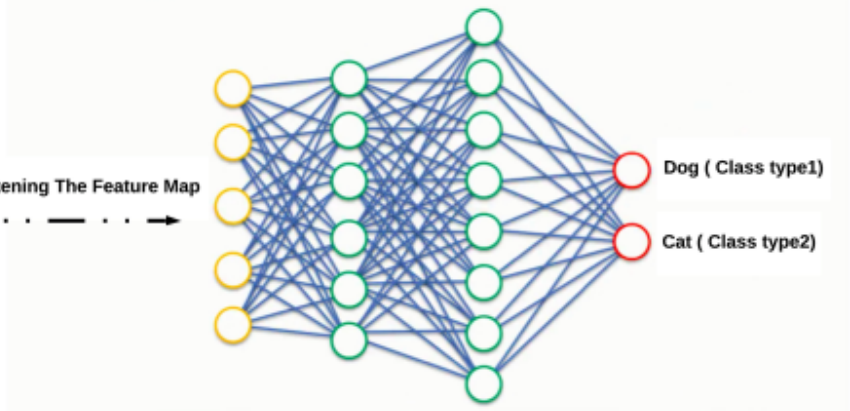



Figure 2.10: The architecture of Fully Connected Layers [65].

## Training process of Convolutional Neural Network

Training a CNN model involves using a dataset composed of images along with their corresponding labels, which may include classes, bounding boxes, and masks. The training process involves backpropagation, a

method that calculates the error based on the output of the previous layer. This error is then used to adjust the weights of the neurons in that layer [66]. As illustrated in Figure 2.11, new weights are applied to compute the error and update the old weights. This process is repeated layer by layer, moving backward through the network until it reaches the input layer. Each activation unit sums all incoming signals, including the bias term, and applies an activation function to generate the output. The network then evaluates the cost function and propagates the error backward, updating the weights iteratively until cost is minimised [64].

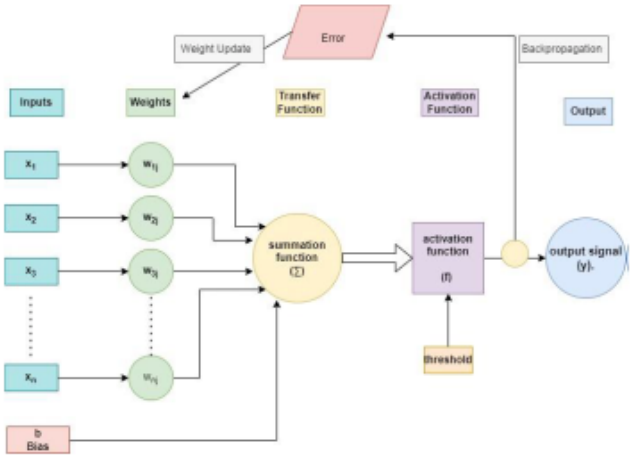



Figure 2.11: Forward and backpropagation in hidden CNN layers [64].

## Most Popular Convolutional Neural Networks

**LeNet Network:** In 1998, LeCun et al. introduced the LeNet-5 model, designed for the digital classification of diffrent people. At the time, it outperformed all existing approaches [67]. Notably, this model marked the first use of backpropagation to train a convolutional neural network (CNN). LeNet-5 is regarded as a foundational model in deep learning, paving the way and inspiring numerous architectures that followed.

The LeNet-5 network consists of 7 layers and comprises around 60,000 parameters. As illustrated in Figure 2.12, the architecture is divided into two main parts: the convolutional section and the fully connected (FC) section. The convolutional section is built from repeated units, where each unit includes a convolutional layer (Conv) followed by a max-pooling layer (Pool). The FC section contains three fully connected layers with 120, 84, and 10 neurons, respectively. The model uses the sigmoid activation function throughout and applies a softmax classifier at the output layer. Before entering the FC section, the output from the convolutional part is flattened so that each feature map in the mini-batch is converted into a vector of length equal to channels × height × width [68].

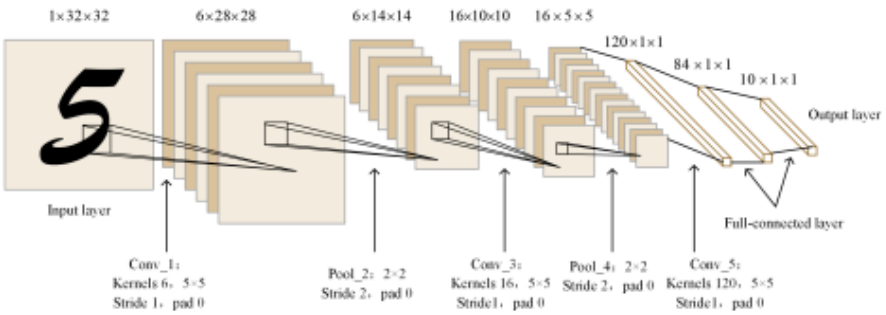



Figure 2.12: The architecture of the LeNet-5 network [64].

**AlexNet Network:** In 2012, AlexNet, developed by Krizhevsky et al, made a significant impact [69]. his network won the ILSVR competition in 2012 by a large margin. AlexNet consists of 8 layers, including five convolutional layers and two fully connected layers designed for feature learning. Max-pooling is applied after the first, second, and fifth convolutional layers. The network contains a total of 630 million connections, 60 million parameters, and 650,000 neurons. AlexNet was the first to showcase the power of deep learning in computer vision tasks [70]. It also utilized a single convolutional layer with large kernels sized 7 x 7 and 11 x 11 [64].

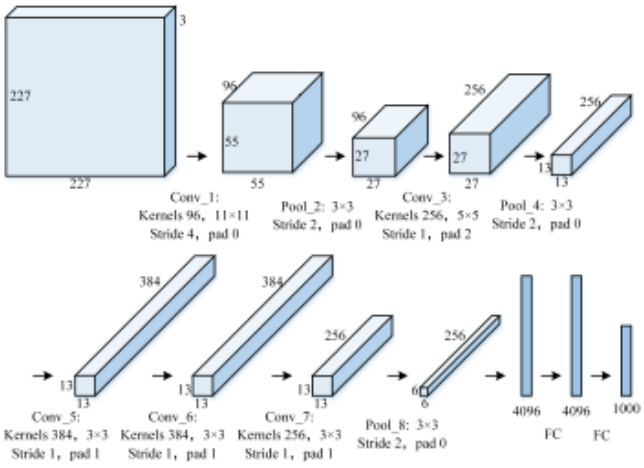



Figure 2.13: The architecture of the AlexNet network [68].

**ResNet 50:** ResNet50 is a deep convolutional neural network consisting of 50 layers and approximately 26 million parameters. It was developed by Microsoft and introduced in 2015 [70]. The network's structure can be extended to as many as 152 layers. It utilises relatively few pooling layers but incorporates a high number of downsampling, which enhances the efficiency of forward propagation. This design achieved outstanding image recognition results at that time and demonstrated the effectiveness of using residual connections [71, 72].

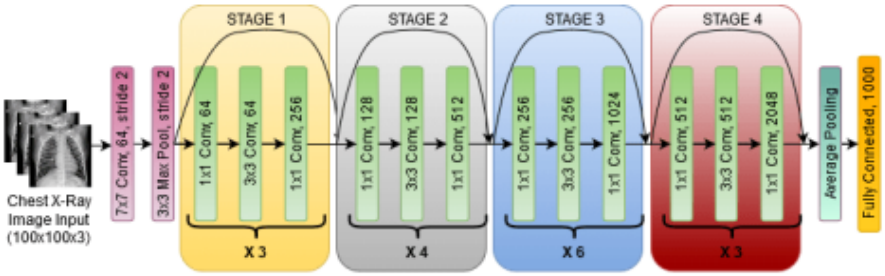



Figure 2.14: Resnet50 architecture with convolutional blocks of different filter sizes and max pooling [73].

**DenseNet 201:** DenseNet (Dense Convolutional Network) was introduced by Gao Huang et al in (2017), denseNet-201 is a deep convolutional neural network with 201 layers that belongs to the DenseNet family, which features dense connections between layers. In denseNet-201, each layer receives input from all previous layers, allowing efficient feature reuse, improved gradient flow, and reduced redundancy. This structure enables the model to achieve high accuracy with fewer parameters compared to traditional CNNs. DenseNet-201 is especially effective in tasks like image classification and object detection, thanks to its depth and efficient design [74].

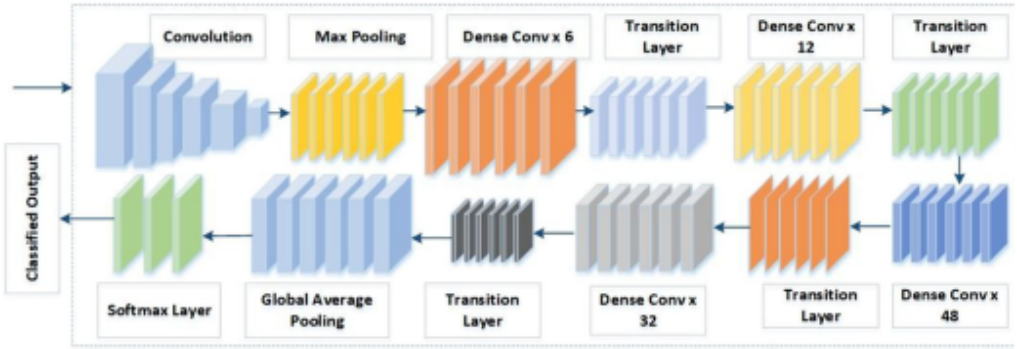



Figure 2.15: Detailed architecture of an Efficient DenseNet-201 [75].

**VGG 19:** VGG19 is a deep convolutional neural network developed by the Visual Geometry Group at Oxford, consisting of 19 layers (16 convolutional and 3 fully connected layers) (see Figure 2.16) [76]. and its input is a 224 × 224 size, 3-channel image with its average RGB value subtracted. The convolutional layers have a small 3 × 3 slot size with 1 pixel padding and stride. The array has 5 max-pooling layers with 2×2 core size and 2 pixel stride [77].

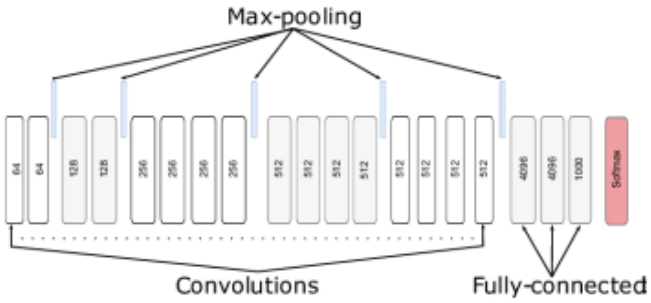



Figure 2.16: VGG-19 model architecture [78].

## 2.4 Vision Transformer (ViT)

The Vision Transformer (ViT) is a natural generalization of the transformer model proposed by Vaswani et al. in 2017 and modifying it to perform on visual information. As opposed to the classical convolutional neural networks (CNNs) which mainly use convolutions to extract local spatial information, Vision Transformers use self-attention mechanism to capture global introduction relation over the image patches. This novel positioning has state-of-the-art results in a million computer vision submission categories such as image classification, object detection and segmentation [79].

## 2.5 Fundamental Concepts in ViTs

The basic architecture of the converter is illustrated in Figure 2.17. First, the input image is divided into patches, then normalized and transformed into lower-dimensional linear vectors (Patch embeddings). Positional embeddings and class tokens are subsequently added to these embeddings which are then fed to the encoder block of the converter to generate the class labels. The encoder block consists of several components such as the multi-head attention (MSA) layer, feed-forward neural network (FFN) , normalization layers, and residual connections.Finally, The final prediction is reached through the final header, which can be an MLP layer, or a decoder block. Each of these components is discussed in detail in the following subsections [80].

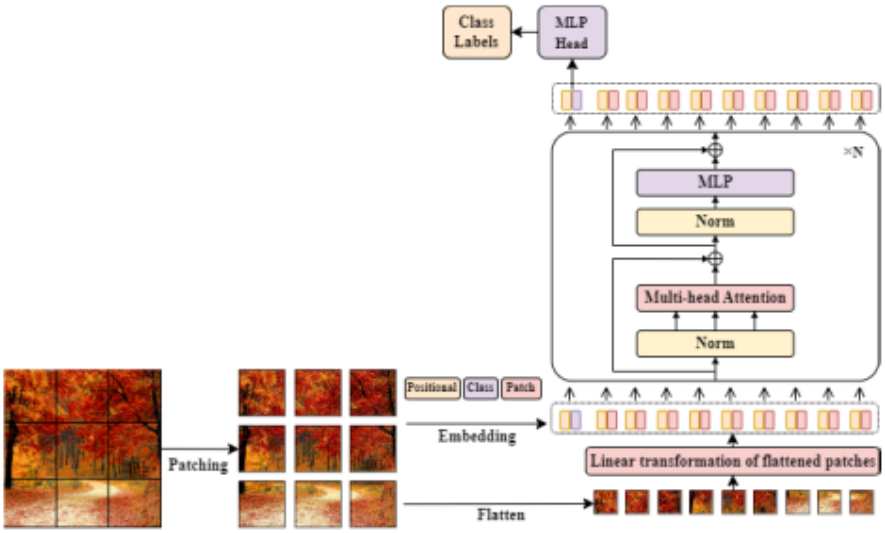



Figure 2.17: Detail architecture of Vision Transformer (ViT) [80].

### 2.5.1 Patch Splitting

The input image, typically with dimensions $H \times W \times C$ (height, width, and channels), is divided into fixed-size patches. For example, an input image of size $224 \times 224$ can be divided into $16 \times 16$ patches, non-overlapping, which results in $\frac{224}{16} \times \frac{224}{16} = 14 \times 14 = 196$ patches [79].

### 2.5.2 Patch Flattening

Each patch is flattened into a one-dimensional vector (1D) vector. For example a patch of size $P \times P \times C$ (e.g., $16 \times 16 \times 3$) is transformed into a vector of size $P^2 \times C$,resulting in 196 patch vectors for an image [79].

### 2.5.3 Patch embedding

Patch embedding is a fundamental step in Vision Transformer (ViT) architecture. It transforms image patches into vector representations, this allows the model to treat images as token sequences using a transformer-based approach 14. The input image is split into fixed-size, non-overlapping patches, each of which is flattened into a one-dimensional (1D) vector, and then mapped to a higher-dimensional feature space using a linear projection with *D* embedding dimensions (Equation 1). This mechanism enables ViT to capture long-range dependencies across different patches, leading to strong performance on various image-related tasks.

$$X_{\text{patch}}^{N \times D} = R(I_{\text{image}}^{A \times B \times C})$$

$I_{\text{image}}$is the input image, and its size is $A \times B \times C$. $R(\cdot)$ . It is the reshaping function to produce $N$ number of patches $N$ number of patches "$X_{\text{patch}}$" with size $D$, and $N = \frac{A}{P} \times \frac{B}{P}, \quad D = P \times P \times C$ where $P$ is the patch size and $C$ is the number of channels [80].

## Positional embedding

Positional encoding is used by Vision Transformers (ViTs) to embed positional information into the input sequence and preserve it across the network. This encoding allows the model to capture the sequential relationships between patches by integrating position embeddings directly into the patch embeddings. Since the developement of ViTs, various positional embedding methods have been proposed to effectively model sequential data 21. These methods can be broadly classified into three main categories [80]:

- **Absolute Position Embedding (APE):**

  Absolute Positional Encoding (APE) is used to integrate positional embeddings into the patch embeddings before the encoder blocks.

$$X = X_{patch} + X_{pos} \tag{2.1}$$

  where, the Transformer's input is represented by $X$ ,$X_{patch}$ represents patch embeddings, and $X_{pos}$ is the learnable position embeddings. Both $X_{patch}$ and $X_{pos}$ have dimensions $(N + 1) \times D$, where $D$ represents the dimension of an embedding. It is possible to train corresponding to positional embeddings of a single or two sets that can be learned [80].

- **Relative Position Embedding (RPE):**

  The Relative Position Embedding (RPE) technique is primarily used to incorporate information related to relative position into the attention module. This technique is based on the idea that the spatial relationships between patches carry more weight than their absolute positions. To compute the RPE value, a lookup table is used, which is based on learnable parameters. The lookup process is determined by the relative distance between patches. Although the RPE technique is extendable to sequences of varying lengths, it may increase training and testing time [80].

- **Convolution Position Embedding (CPE):**

  The Convolutional Position Embeddings (CPE) method takes into account the 2D nature of the input sequences. 2D convolution is employed to gather position information using zero-padding to take advantage of the 2D nature. Convolutional Position Embeddings (CPE) can be used to incorporate positional data at different stages of the Vision Transformer. The CPE can be introduced between two

encoder layers, or to the Feed-Forward Network (FFN), or specifically to the self-attention modules [80].

### 2.5.4 Attention Mechanism

The self-attention mechanism plays a vital part in the Vision Transformer (ViT) architecture due to its capability to express the relationships between the entities in a sequence. It decides the relevance of each item to all others by representing each entity based on global contextual information, allowing the model to capture their interactions. In this process, the self-attention module transforms the input sequence into three distinct embedding spaces: query, key, and value. The self-attention module takes sets of key-value pairs along with the corresponding query vectors as input and computes output vectors by taking a weighted sum of the values. The weights are derived using a scoring function, followed by the application of the softmax operator (Equation 7).

$$Attention(Q, K, V) = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{2.2}$$

where, $Q$, $V$, and $K_T$ are query, value, and transposed key matrix, respectively. $\sqrt{d_k}$ is the caling factor, and $d_k$ is dimensions of the key matrix [80].

#### Multi-Head Self-Attention (MSA)

A single-head self-attention module has limited capacity, often focusing on only a few positions while potentially overlooking other important ones. This limitation is addressed by Multi-Head Self-Attention (MSA), and the purpose is to use parallel stacking to enhance the effectiveness of the self-attention layer's. The MSA can detect a variety of intricate interactions among the different sequence components by assigning various representation subspaces (query, key, and value ) to the attention layers. MSA consists of several self-attention blocks. Each self-attention block has learnable weight matrices for the query, key, and value sub- spaces. The outputs from these parallel heads are then concatenated and projected to the output space using a learnable parameter $W^o$. The mathematical representation of the attention process is given below:

$$MSA(Q, k, V) = Concat(head_1, head_2, head_3, \cdots, head_n) \cdot W^o \tag{2.3}$$

$$head_i = Attention(Q_i, K_i, V_i) \quad \text{where} \quad i = 1, 2..., h \tag{2.4}$$

Self-attention has the key advantage of being able to compute filters dynamically for each input sequence. unlike convolutional filters that are usually fixed. This means self-attention can adapt based on the specific context of the input data. It is also flexible when dealing with varying numbers of input points or different

arrangements, making it well-suited for irregular data. In contrast, traditional convolutional methods are less flexible and require inputs to have a grid-like format, such as 2D images. Self-attention is a strong technique for modelling sequential data and has been successful in natural language processing tasks [80].

### 2.5.5 Transformer layers

The Vision Transformer (ViT) encoder is composed of multiple layers designed to process the input sequence effectively. These layers include components such as layer normalisation, residual connections, a feed-forward neural network (FFN), and a Multi-Head Self-Attention (MSA) mechanism. These components (layers) are structured into a unified block, which is repeated multiple times throughout the encoder to enable the model to learn the complex representations of the input sequence [80].

#### Residual connection

The sub-layers within the encoder/decoder block employ a residual link to enhance performance and facilitate better information flow. The original input positional embedding is added to the output vector of MSA, as additional information. Then the residual connection is followed by a layer-normalisation operation (Equation 10) [80].

$$X_{output} = LayerNorm(X + Attention(X)) \tag{2.5}$$

#### Normalization layer

Layer normalisation can be performed in several ways; one of them is pre-layer normalisation (Pre-LN), which is often used and the normalisation layer is placed prior to the MSA or FFN and within the residual connection. Other normalisation methods, such as batch normalisation, have been proposed to improve the training of transformer models, however they may be less effective due to variations in feature values [80].

#### Feed-forward network

The transformer-specific Feed-Forward Network (FFN) is used to extract more complex features from the input data. It consists of two fully connected layers and a nonlinear activation function in between the layers, such as GELU, as shown in (Equation 11). An FFN is used after the self-attention module within every encoder block. The hidden layer in the FFN typically has a dimensionality of 2048.

$$FFN(X) = b2 + W2 * \sigma(b1 + W1 * X) \tag{2.6}$$

In Eq. 11,the non-linear activation function GELU is denoted by $\sigma$. The network weights are represented by W1 and W2, while b1 and b2 denote the layer-specific bias terms [80].

### 2.5.6 Classification Token (CLS Token)

In Vision Transformers, a unique classification token, known as the CLS token, is added to the start of the input sequence. This token plays a key role by aggregating information from all image patches as it passes through the transformer layers. Through the self-attention mechanism, the CLS token learns to capture a comprehensive representation of the entire image by interacting with all the patches. Once the data passes through the transformer layers, the CLS token is extracted and forwarded to a classification head to produce the final prediction [79].

### 2.5.7 MLP Head (Classification Head)

After the transformer encoders process the sequence of patches and the CLS token, the output corresponding to the CLS token is used for classification. The output of the CLS token is fed into an MLP, typically consisting of one or two fully connected layers. A softmax layer is applied at the end of the MLP for classification tasks, predicting the image's label [79].

## 2.6 Classification metrics

The performance of an Artificial Neural Network (ANN) is influenced by numerous parameters, and changes in these parameters can significantly impact the overall design and effectiveness of the algorithm. Due to the vast number of possible parameter combinations, the potential configurations for an ANN are virtually limitless. Therefore, understanding how one ANN configuration performs in comparison to another is crucial for achieving optimal performance based on specific classification needs. Performance metrics are used as objective mathematical tools that evaluate how well each algorithm performs based on the results generated by each ANN [81]. Specifically, in this case, evaluation metrics help measure and summarise the performance of a trained classifier when it is tested on unseen data [82].

### 2.6.1 Confusion Matrix

A confusion matrix is an l x l array, that compares a classifier's predicted outputs with the actual class labels of the input data. In this section, the real classification labels are represented along the columns, while the predicted labels generated by the algorithm are placed along the rows. Figure 2.18 illustrates a confusion matrix for a binary classification problem. It includes the following components: the set of actual labels cr*i* = P,N and the set of predicted labels cp*i* = P,N, where P stands for "Positive" and N for "Negative". The letters T and F indicate "True" and "False" respectively. The interpretation of these label combinations is provided in the following lines, and their corresponding mathematical formulations are presented in Equations 1 to 4 [81].

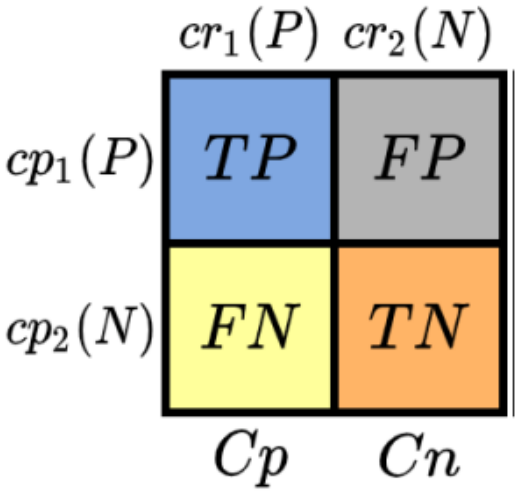



Figure 2.18: Confusion matrix for a binary classification [81].

- **TP (True Positives)** represents the number of elements that belong to the class P and were classified correctly.
- **TN (True Negatives)** is the number of elements that do not belong to the P class and that were correctly classified.
- **FP (False Positives)** denotes the number of elements that refer to class N but were labeled as class P.
- **FN (False Negatives)** expresses the number of elements that pertain to the class P but were labeled as N.

$$TP = \sum_{i=1}^{l} tpi \quad \text{where} \quad tpi = \begin{cases} 0 & \text{if } cri \neq cpi \\ 1 & \text{if } cri = cpi \end{cases}$$

$$TN = \sum_{i=1}^{l} tni \quad \text{where} \quad tni = \begin{cases} 0 & \text{if } cri = cpi \\ 1 & \text{if } cri \neq cpi \end{cases}$$

$$FP = \sum_{i=1}^{l} fpi \quad \text{where} \quad fpi \begin{cases} 1 & \text{if } cri = N \bigotimes cpi = P \\ 0 & \text{otherwise} \end{cases}$$

$$FN = \sum_{i=1}^{l} fni \quad \text{where} \quad fni \begin{cases} 1 & \text{if } cri = P \bigotimes cpi = N \\ 0 & \text{if otherwise} \end{cases}$$

Cp is the number of elements that really belong to the class P(TP +FN) and Cn is the sum of all the elements that really belong to the class N(FP + TN)

A confusion matrix helps you see how well a model is working by showing correct and incorrect predictions. It also helps calculate key measures like accuracy, precision, and recall, which give a better idea of performance, especially when the data is imbalanced.

**Accuracy:**

The accuracy parameter (Equation 1) is the most used metric in classifications. Represents the overall effectiveness of the algorithm in correctly categorizing the input elements. Establishes what percentage of the total elements were properly classified [81].

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

**Precision:**

The Equation 2 calculates the precision of a classifier. This metric gives the percentage from P predictions made by the classifier, are correct [81].

$$Presision = \frac{TP}{TP + FP} \tag{2.8}$$

**Recall (Sensitivity):**

Recall is a metric that focuses on observing the percentage of the P class elements are correctly labeled by the classifier (Equation 3) [81].

$$Recall = \frac{TP}{TP + FN} \tag{2.9}$$

**F1-Score:**

Equation 8 shows the mathematical expression to calculate this evaluation metric. F1 score gives a weighted average between Precision and Recall. F1 score comes from F-score 13, but with $\beta = 1$ to give the same importance recall and precision. Except for cases where one of these two variables is really more important, F1-score is most widely used [81].

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.10}$$

**Specificity:**

Specificity is another important metric in the evaluation of classification models, particularly in binary classification. It measures the ability of a model to correctly identify negative instances. Specificity is also known as the True Negative Rate. Formula is given by [81]:

$$Specificity = \frac{TN}{TN + FP} \tag{2.11}$$

## 2.7 Conclusion

This chapter discussed artificial intelligence and its different breadth, niches, specifically machine and deep learning techniques. We talked about different varieties of machine learning including supervised, unsupervised, semi-supervised and reinforcement learning and how each addresses the problem of analyzing data and making decisions. It also discussed neural networks and deep learning architectures, including convolutional neural networks (CNNs), and presented the Vision Transformer (ViT) as a recent method for image-based tasks.

# CHAPTER 3

# Results and discussions

# 3.1 Introduction

This chapter presents the evaluation of a deep learning framework for glaucoma detection using fundus images. Three model configurations ViT, ResNet-50, and a hybrid of both are tested on the ACRIMA, DR-ISHTI_GS, and LAG datasets. The models are assessed using accuracy, precision, and sensitivity. The aim is to compare their performance and demonstrate the effectiveness of feature fusion in enhancing classification accuracy for automated glaucoma detection.

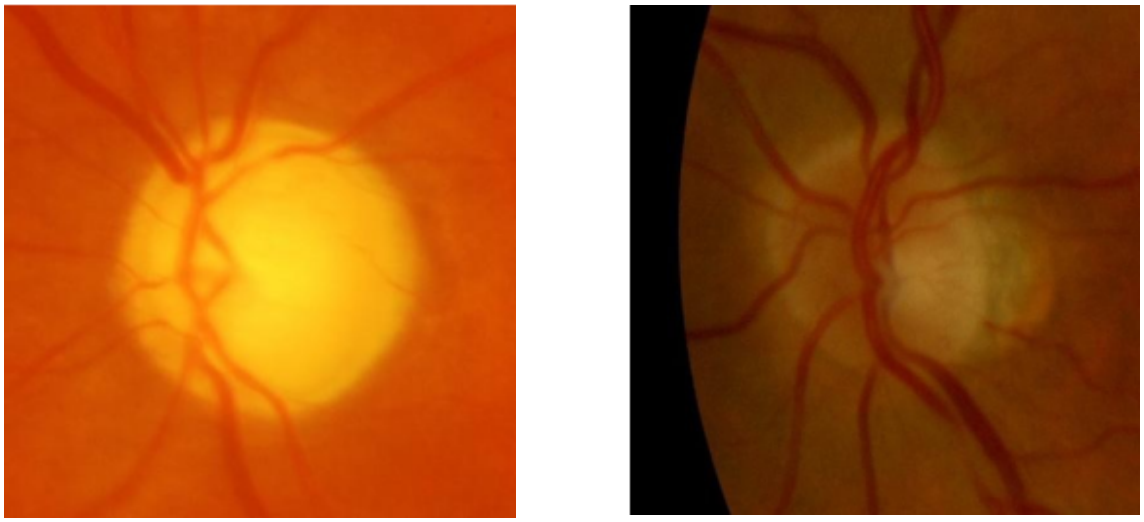
# 3.2 Methodology

## 3.2.1 Dataset Description

The data used in this study was token from the Kaggle repository.

### ACRIMA

The ACRIMA dataset is a publicly available dataset commonly used in glaucoma research, especially for training and evaluating machine learning models for glaucoma detection using fundus images (retinal images) [83], it is specifically designed to facilitate the development of automated systems for glaucoma diagnosis. This dataset comprises a total of 705 retinal images, including 309 images diagnosed with glaucoma and 396 normal cases [84] .



(a) Glaucomatous eye

(b) Healthy eye

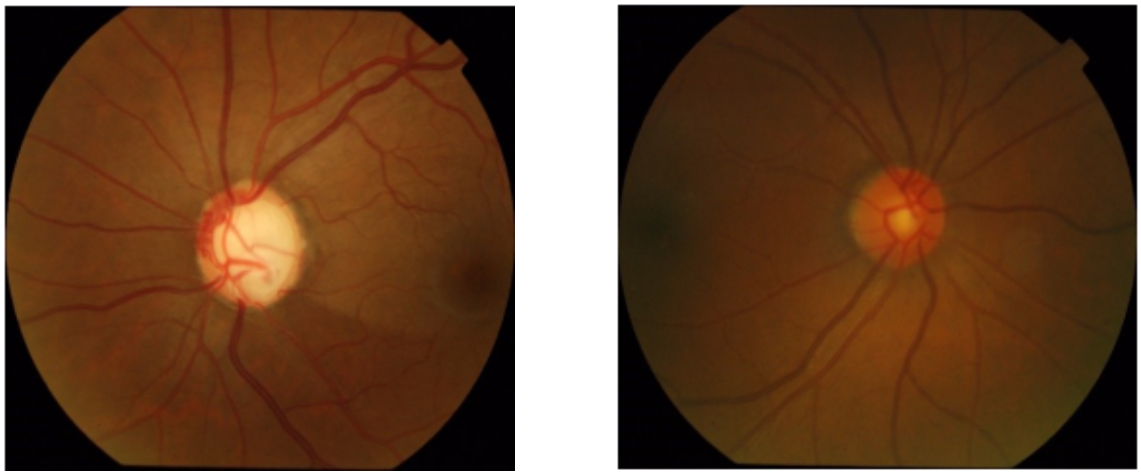
Figure 3.1: Retinal images from the ACRIMA dataset [84].

## DRISHTI_GS

The DRISHTI_GS dataset is a publicly available retinal fundus image dataset designed primarily for glaucoma detection and optic disc/optic cup segmentation. It contains high-resolution color fundus images , and includes both glaucomatous and non-glaucomatous cases, making it a valuable resource for developing and evaluating computer-aided diagnostic systems for glaucoma screening. The dataset was created to support research in medical image analysis, especially for tasks like [85]:

- Optic disc and cup segmentation.
- Cup-to-disc ratio computation.
- Glaucoma classification.

This dataset comprises a total of 101 retinal images, including 70 images diagnosed with glaucoma and 31 normal cases [86].

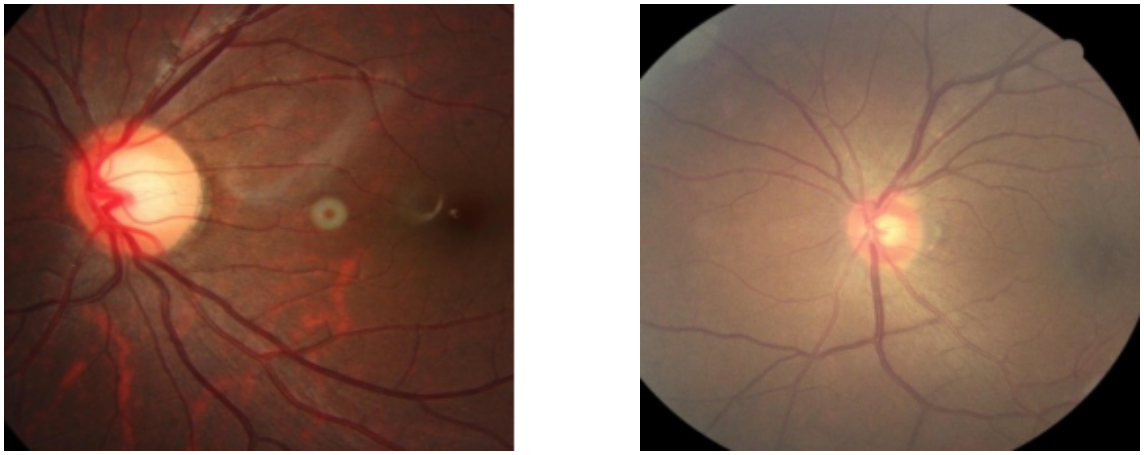


(a) Glaucomatous eye (b) Healthy eye

Figure 3.2: Retinal images from the DRISHTI_GS dataset [86].

## LAG

The LAG (Labelled Axial Glaucoma) dataset is a publicly available collection of retinal fundus images designed specifically for the development and evaluation of automated glaucoma detection systems. It was created to aid in early diagnosis and research in ophthalmology using machine learning and deep learning techniques [87]. The LAG dataset consists of 4854 retinal images, including 1711 images diagnosed with glaucoma and 3143 normal cases, also these images divided into 3883 training and 971 testing images [88].



(a) Glaucomatous eye (b) Healthy eye

Figure 3.3: Retinal images from the LAG dataset [88].

The table below shows details of the datasets taken from kaggle:

Table 3.1: Kaggle's repository dataset details

| Dataset | Glaucoma | | No Glaucoma | | Total |
|---|---|---|---|---|---|
| | Test | Training | Test | Training | |
| Acrima | 79 | 317 | 62 | 247 | 705 |
| DRISHTI_GS | 38 | 32 | 13 | 18 | 101 |
| LAG | 342 | 1369 | 629 | 2514 | 4854 |

### 3.2.2 Proposed method

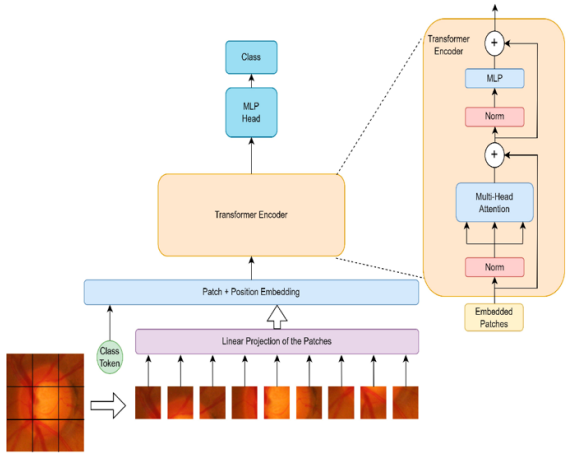
In this work, we propose a deep learning-based classification framework for glaucoma detection using feature fusion from multiple pre-trained models. The approach consists of two stages: feature extraction and classification. Features (datasets) are extracted using Vision Transformer (ViT) and ResNet-50 models in three configurations: ViT alone, ResNet-50 alone, and a combination of both. These features are input to a Multilayer Perceptron (MLP) classifier composed of fully connected layers with 128 neurons, a ReLU activation function to introduce non-linearity, and a dropoutLayer with a dropout rate of 0.1 to reduce overfitting, the output is then passed to a final fully Connected Layer matching the number of target classes, followed by a softmax layer for prediction. The model is trained using the Adam optimiser over 350 epochs with a learning rate of 0.0001. Performance is evaluated using accuracy, precision, and sensitivity. This comparative setup highlights the strengths of each model and the advantages of feature fusion for improved glaucoma detection. Figure 3.4 presents the overall system architecture incorporating ViT, ResNet 50 and the combined of them, outlining the general framework of the design.

- Data Input: retinal fundus images (ACRIMA, DRISHTI_GS, LAG)
- Feature Extraction:

  ACRIMA: Use pre-trained ViT then, ResNet 50 then, the hybrid (ViT and ResNet 50) to extract features from ACRIMA images.

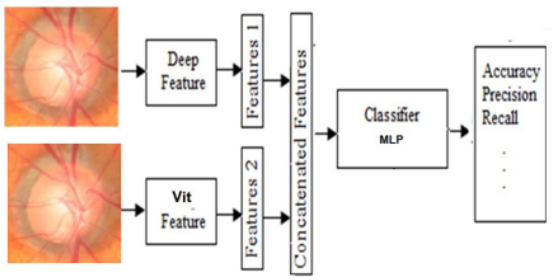
  DRISHTI_GS: Use pre-trained ViT then, ResNet 50 then, the hybrid (ViT and ResNet 50) to extract features from DRISHTI_GS images.

  LAG: Use pre-trained ViT then, ResNet 50 then, the hybrid (ViT and ResNet 50) to extract features from LAG images.
- Classifier: Feed the extracted features into a fully connected layers classifier.
- Output: The output of the classifier indicates the prediction for glaucoma's disease diagnosis.




(a)


(b)




(c)

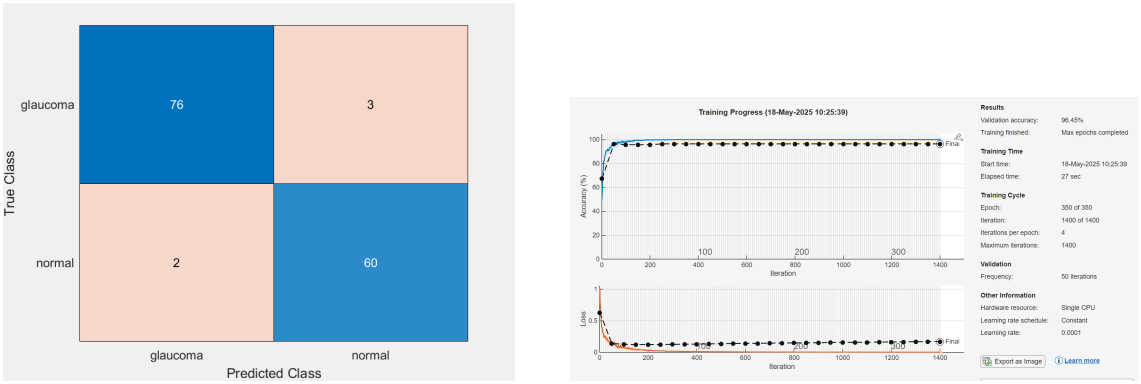
Figure 3.4: (a) ViT model (b) ResNet 50 (c) hybrid model (ViT and ResNet 50).

The performance of the proposed method was evaluated using three key metrics: accuracy (Acc), sensitivity (SN), also referred to as recall (R), and precision (P).

## 3.3 Case Study 1: Results and discussion for ACRIMA

### 3.3.1 ViT model

The ViT model was evaluated in this work, and the corresponding results are presented in the table below. As observed, the model achieved a max accuracy of 96.45%, with further performance details illustrated in the confusion matrix and shown in Figure 3.5.

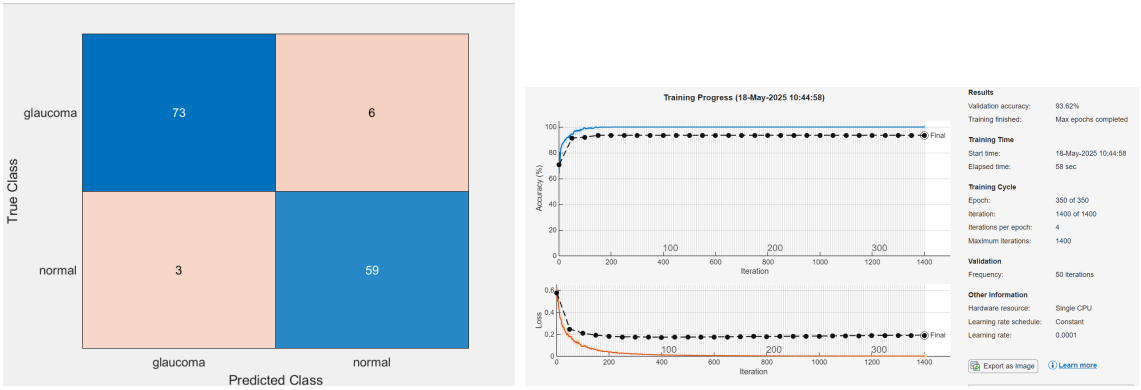



(a) (b)

Figure 3.5: (a) Confusion matrix of test (b) Training process of ViT.

### 3.3.2 ResNet 50

In this project, the ResNet–50 model was also tested. The results, outlined in the table below, show that it achieved an accuracy of 93.62% .

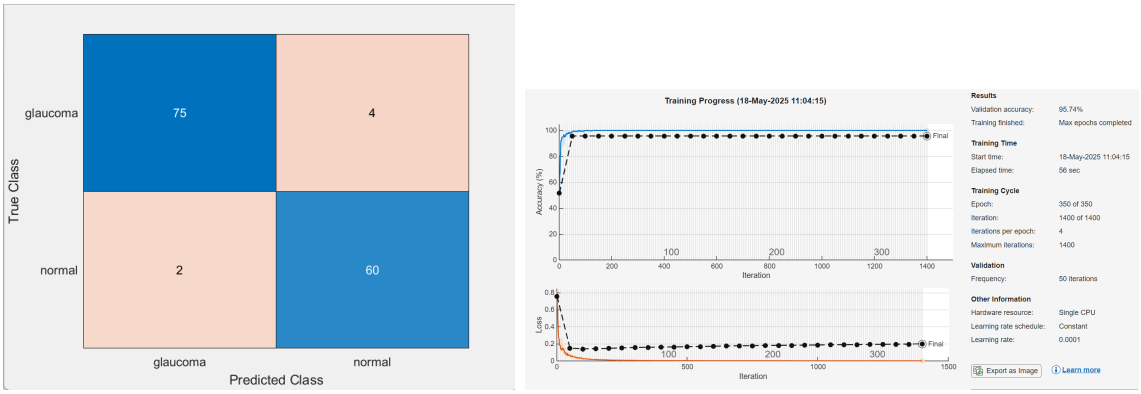



(a) (b)

Figure 3.6: (a) Confusion matrix of test (b) Training process of ResNet 50.

### 3.3.3 ViT and ResNet 50 (Hybrid)

ViT and ResNet 50 combined achieved an accuracy of 95.74%, Which is relatively less than ViT's accuracy, and is biger compared to the results obtained from ResNet. Figure 3.7 presents the training loss graph along with the corresponding confusion matrix.



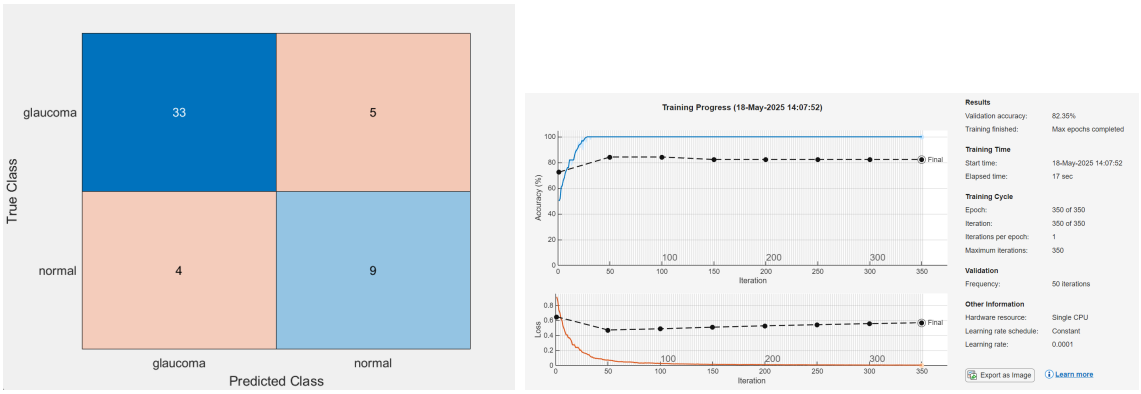

(a) (b)

Figure 3.7: (a) Confusion matrix of test (b) Training process of ViT and ResNet combined.

## 3.4 Case Study 2: Results and discussion for DRISHTI_GS

### 3.4.1 ViT model

The application of ViT on DRISHTI_GS data yielded an accuracy of 82,35%.

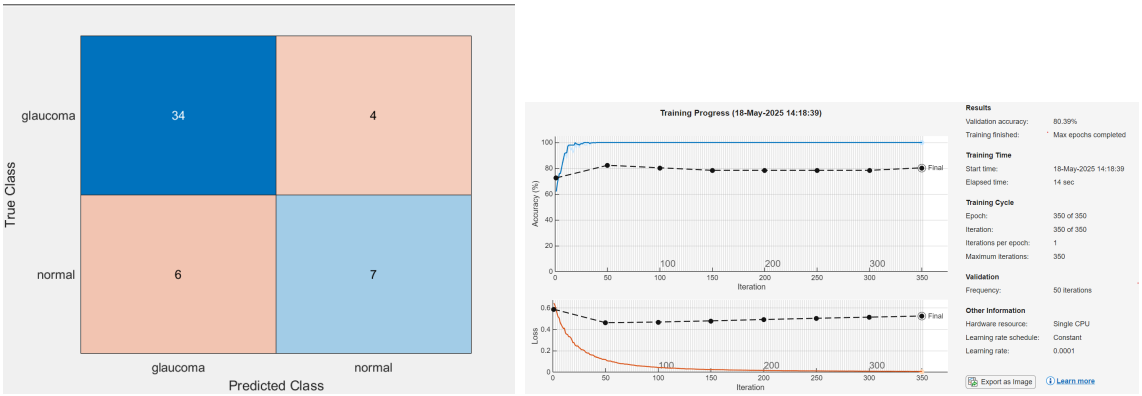



(a) (b)

Figure 3.8: (a) Confusion matrix of test (b) Training process of ViT.

### 3.4.2 ResNet 50

Using ResNet 50 on the DRISHTI_GS data did not result in any performance gain compared to ViT, with an accuracy of 80.39%. On the contrary, ViT was better.

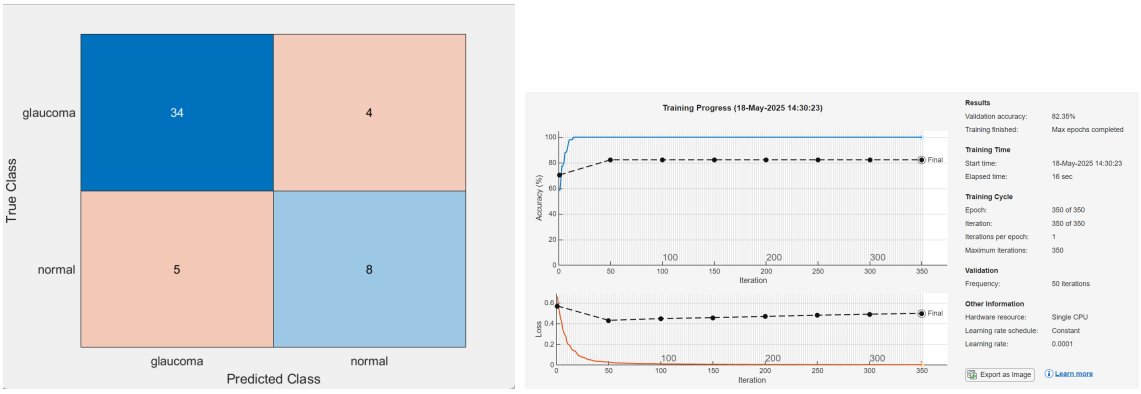



(a) (b)

Figure 3.9: (a) Confusion matrix of test (b) Training process of ResNet 50.

### 3.4.3 ViT and ResNet 50 (Hybrid)

ViT and ResNet (combined) achieved an accuracy of 82.35% on the DRISHTI_GS data, that result indicates that the accuracy achieved by this model matches that of ViT model, and better than the ResNet 50 model



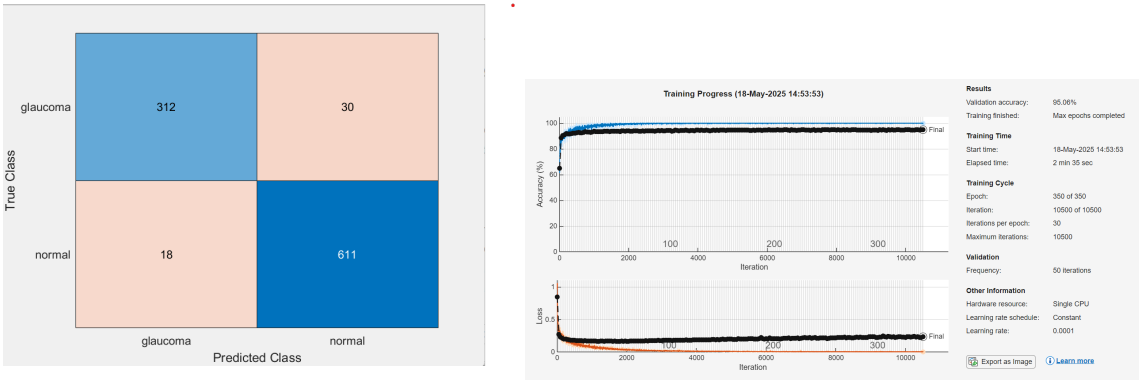

(a) (b)

Figure 3.10: (a) Confusion matrix of test (b) Training process of ViT and ResNet combined.

# 3.5 Case Study 3: Results and discussion for LAG

## 3.5.1 ViT model

When ViT was applied to LAG data, it achieved an accuracy of 95.06%, as illustrated in Figure 3.11 below :

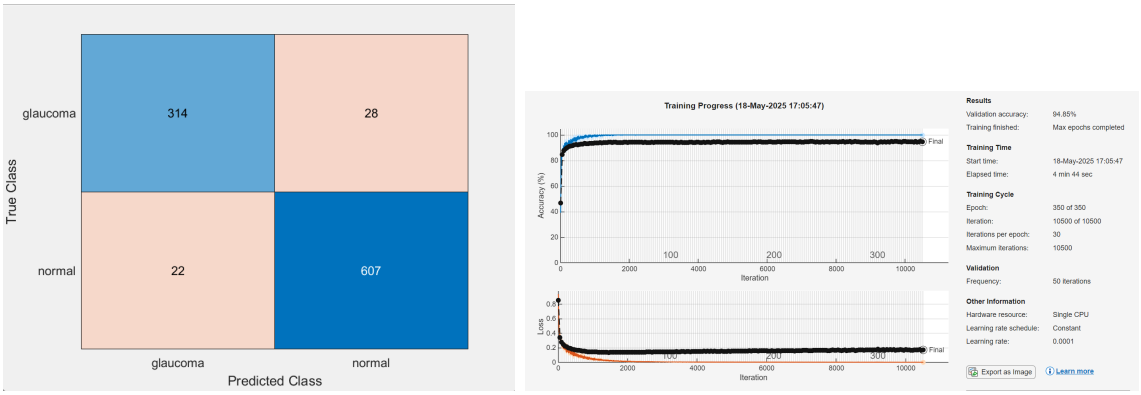



(a)

Figure 3.11: (a) Confusion matrix of test (b) Training process of ViT.

## 3.5.2 ResNet 50

ResNet-50 achieved an accuracy of 94.85% on LAG data, which is less than ViT's value, but negligible value in performance in comparison with , as depicted in Figure 3.12 below.

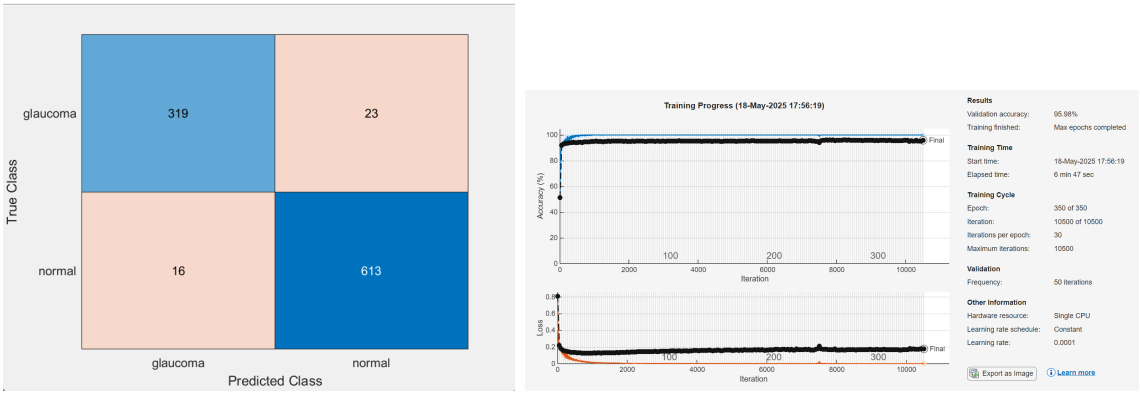



(a) (b)

Figure 3.12: (a) Confusion matrix of test (b) Training process of ResNet 50.

### 3.5.3 ViT and ResNet 50 (Hybrid)

ViT and ResNet 50 (combined) proved to be the most accurate pre-trained model in this dataset (LAG), achieving an accuracy of 95.98%, as illustrated by the figures at the bottom.




(a) (b)

Figure 3.13: (a) Confusion matrix of test (b) Training process of ViT and ResNet combined.

## 3.6 Performance Comparison of ViT, ResNet-50, and Hybrid Models Across Datasets

Table 3.2: Comparison of different Cases studies .

| METHOD | DATASET | ACCURACY % | PRECISION % | SENSITIVITY % |
|---|---|---|---|---|
| ViT | ACRIMA | 96.45 | 96.20 | 97.44 |
| ResNet 50 | | 93.62 | 92.41 | 96.05 |
| Hybrid | | 95.74 | 94.94 | 97.40 |
| ViT | DRISHTI_GS | 82.35 | 86.84 | 89.29 |
| ResNet 50 | | 80.39 | 92.11 | 85 |
| Hybrid | | 82.35 | 89.47 | 88.57 |
| ViT | LAG | 95.06 | 91.52 | 94.80 |
| ResNet 50 | | 94.85 | 92.69 | 93.45 |
| Hybrid | | 95.98 | 94.44 | 95.22 |

The table (3.2) presents a comparative evaluation of three different methods Vision Transformer (ViT), ResNet-50, and their Hybrid fusion across three glaucoma datasets:
ACRIMA, DRISHTI_GS, and LAG. The evaluation metrics include accuracy, precision, and sensitivity.

**1. ACRIMA dataset:**

- ViT outperforms the other models, achieving the highest accuracy (96.45%), precision (96.20%), and sensitivity (97.44%).

- Hybrid model follows closely with 95.74% accuracy and nearly comparable precision (94.94%) and sensitivity (97.40%).
- ResNet-50 lags slightly behind, scoring 93.62% accuracy, 92.41% precision, and 96.05% sensitivity.
- Conclusion: ViT excels in all metrics, showing its strength in global feature representation for this dataset.

**2. DRISHTI_GS dataset:**

- ViT and Hybrid models both record an accuracy of 82.35%. However, the Hybrid approach demonstrates better precision (89.47%) and sensitivity (88.57%) compared to ViT (86.84% and 89.29%, respectively).
- ResNet-50 shows the lowest accuracy (80.39%) and sensitivity (85%), although it has the highest precision (92.11%) among the three.
- Conclusion: The Hybrid model offers a balanced performance on this dataset, especially in precision and sensitivity, while ViT and ResNet-50 show trade-offs between metrics.

**3. LAG Dataset:**

- The Hybrid model again achieves the best overall performance, with the highest accuracy (95.98%), precision (94.44%), and sensitivity (95.22%).
- ViT performs slightly lower, achieving 95.06% accuracy, 91.52% precision, and 94.80% sensitivity.
- ResNet-50 scores the lowest among the three, with 94.85% accuracy, 92.69% precision, and 93.45% sensitivity.
- Conclusion: The Hybrid model offers superior classification on the LAG dataset, benefitting from the complementary strengths of both ViT and ResNet-50.

**Overall Insights:**

- ViT performs best on the ACRIMA dataset.
- Hybrid models consistently provide balanced and strong results across all datasets, especially for LAG and DRISHTI_GS.
- ResNet-50 generally shows competitive precision but falls behind in overall accuracy and sensitivity.

These results demonstrate the benefit of feature fusion in the Hybrid approach, which combines the global contextual learning of ViT and the local hierarchical features of ResNet-50 to achieve improved performance in glaucoma detection.

## 3.7 Conclusion

In this chapter, a deep learning-based methodology for glaucoma detection using fundus images was presented. The approach involved the use of Vision Transformer (ViT), ResNet-50, and a hybrid fusion of both models for feature extraction, followed by classification using a Multilayer Perceptron. The models were evaluated across three publicly available datasets: ACRIMA, DRISHTI_GS, and LAG, using accuracy, precision, and sensitivity as performance metrics.

The experimental results demonstrated that the ViT model consistently achieved high accuracy and sensitivity, especially excelling on the ACRIMA dataset. Meanwhile, the ResNet-50 model, although effective, often lagged behind in accuracy and sensitivity but showed strong precision in some cases. Most notably, the hybrid model, which combined the strengths of both ViT and ResNet-50, provided the most balanced and robust performance across all datasets.

These findings underscore the value of combining global and local feature extraction methods through feature fusion, which enhances the diagnostic capability of the system. Overall, the hybrid model emerged as the most promising approach for automated glaucoma detection,

# General Conclusion

This research presents an automated approach to glaucoma detection by leveraging deep learning models applied to retinal fundus images. The proposed method supports early diagnosis of glaucoma, a critical goal in preventing irreversible vision loss. By utilizing state-of-the-art neural network architectures Vision Transformer (ViT), ResNet-50, and a hybrid of both the system offers a powerful, scalable, and non-invasive diagnostic aid.

The study evaluates model performance across three benchmark datasets: ACRIMA, DRISHTI-GS, and LAG. While each model showed strengths on specific datasets, the hybrid model demonstrated consistently strong performance across all three, validating the benefits of combining local feature extraction (ResNet-50) with global context modeling (ViT). Notably, ViT alone achieved the highest accuracy on the ACRIMA dataset, while the hybrid model outperformed others on the LAG dataset, confirming the importance of feature fusion for robust classification.

The results affirm that all primary and secondary objectives of this research were met. The approach proves effective in enhancing the reliability of glaucoma detection, which could significantly benefit clinical workflows and support early screening in under-resourced or remote settings.

In future work, expanding the datasets with more diverse and annotated images could further improve model accuracy and reduce overfitting. Additionally, incorporating multimodal data such as patient history or intraocular pressure readings could enrich the system and enable more comprehensive glaucoma assessment. This work lays the foundation for accessible, AI-driven tools that can aid ophthalmologists and support large-scale glaucoma screening.

# References


[1] Seth R Flaxman et al. “Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis”. In: *The Lancet Global Health* 5.12 (2017), e1221–e1234.

[2] Yih-Chung Tham et al. “Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis”. In: *Ophthalmology* 121.11 (2014), pp. 2081–2090.

[3] Mohammad Norouzifard. “Computer Vision and Machine Learning for Glaucoma Detection”. PhD thesis. Doctoral dissertation, Auckland University of Technology, 2020.

[4] Andrés Yesid Díaz Pinto. “Machine learning for glaucoma assessment using fundus images”. In: *Universitat Politècnica de València* (2019).

[5] Anum A Salam et al. “Automated detection of glaucoma using structural and non structural features”. In: *Springerplus* 5 (2016), pp. 1–21.

[6] Laser Eye Surgery Hub. *A Comprehensive Guide to Glaucoma: Normal Versus Glaucoma Schematic Figure*. 2025. URL: `https://www.lasereyesurgeryhub.co.uk/a-comprehensive-guide-to-glaucoma/` (visited on 04/22/2025).

[7] Sophie J Bakri. *Mayo Clinic Guide to Better Vision: Preventing and treating disease to save your eyesight*. Simon and Schuster, 2022.

[8] David P Crabb et al. “How does glaucoma look?: patient perception of visual field loss”. In: *Ophthalmology* 120.6 (2013), pp. 1120–1126.

[9] Ko Eun Kim et al. “Prevalence, awareness, and risk factors of primary open-angle glaucoma: Korea National Health and Nutrition Examination Survey 2008–2011”. In: *Ophthalmology* 123.3 (2016), pp. 532–541.

[10] Anders Heijl, Boel Bengtsson, and Sigridur Erla Oskarsdottir. “Prevalence and severity of undetected manifest glaucoma: results from the early manifest glaucoma trial screening”. In: *Ophthalmology* 120.8 (2013), pp. 1541–1545.

[11] Gwendolyn Gramer and Eugen Gramer. “Stage of visual field loss and age at diagnosis in 1988 patients with different glaucomas: implications for glaucoma screening and driving ability”. In: *International Ophthalmology* 38 (2018), pp. 429–441.

[12] The New York Times. *Glaucoma - Health Guide*. http://www.nytimes.com/health/guides/disease/glaucoma/print.html. Accessed: 2025-04-22. 2025.

[13] TC Manjunath, Dharmanna Lamani, Ranjan Kumar HS, et al. "Different clinical parameters to diagnose glaucoma disease: a review". In: *International Journal of Computer Applications* 115.23 (2015).

[14] Oscar Wylee. *Optic Disc Anatomy*. https://www.oscarwylee.com.au/glasses/eye/anatomy/optic-disc. Accessed: 2025-04-22. 2025.

[15] Brian Chon, Mary Qiu, and Shan C Lin. "Myopia and glaucoma in the South Korean population". In: *Investigative ophthalmology & visual science* 54.10 (2013), pp. 6570–6577.

[16] St. Luke's Eye Institute. *Cornea Anatomy*. http://www.stlukeseye.com/anatomy/cornea.html. Accessed: 2025-04-22. 2025.

[17] Tsung-Ho Ou, IC Lai, and Mei-Ching Teng. "Comparison of central corneal thickness measurements by ultrasonic pachymetry, Orbscan II, and SP3000P in eyes with glaucoma or glaucoma suspect". In: *Chang Gung Med J* 35.3 (2012), pp. 255–62.

[18] Mae O Gordon et al. "The Ocular Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma". In: *Archives of ophthalmology* 120.6 (2002), pp. 714–720.

[19] James D Brandt. "Corneal thickness in glaucoma screening, diagnosis, and management". In: *Current opinion in ophthalmology* 15.2 (2004), pp. 85–89.

[20] J Roger and MA Buckley. *HonFCOptom,—The cornea Differentiating sightthreatening from non-sight-threatening eye disease□, continuing education and training*. 2006.

[21] Medlikim. *SP 3000P User Manual*. https://www.medlikim.com/wp-content/uploads/2019/11/SP_3000P.pdf. Accessed: 2025-04-22. 2019.

[22] Health Jade. *Visual Field Test*. https://healthjade.net/visual-field-test/. Accessed: 2025-04-23. 2025.

[23] Karanjit S Kooner, Dominic M Choo, and Priya Mekala. *Meeting Challenges in the Diagnosis and Treatment of Glaucoma*. 2024.

[24] Samantha Sze-Yee Lee and David A Mackey. "Glaucoma–risk factors and current challenges in the diagnosis of a leading cause of visual impairment". In: *Maturitas* 163 (2022), pp. 15–22.

[25] Review of Ophthalmology. *Diagnosing Early Glaucoma: Pearls and Pitfalls*. https://www.reviewofophthalmology.com/article/diagnosing-early-glaucoma-pearls-and-pitfalls. Accessed: 2025-04-23. 2025.

[26] Paolo Bettin and Federico Di Matteo. "Glaucoma: present challenges and future trends". In: *Ophthalmic research* 50.4 (2013), pp. 197–208.

[27] R Krishnadas. “The many challenges in automated glaucoma diagnosis based on fundus imaging”. In: *Indian Journal of Ophthalmology* 69.10 (2021), pp. 2566–2567.

[28] Kirsti Grødum, Anders Heijl, and Bo Bengtsson. “A comparison of glaucoma patients identified through mass screening and in routine clinical practice”. In: *Acta Ophthalmologica Scandinavica* 80.6 (2002), pp. 627–631.

[29] James M Tielsch et al. “A population-based evaluation of glaucoma screening: the Baltimore Eye Survey”. In: *American journal of epidemiology* 134.10 (1991), pp. 1102–1110.

[30] C Jackson et al. “Screening for glaucoma in a Brisbane general practice—the role of tonometry”. In: *Australian and New Zealand journal of ophthalmology* 23.3 (1995), pp. 173–178.

[31] Steven L Mansberger et al. “Causes of visual impairment and common eye problems in Northwest American Indians and Alaska Natives”. In: *American journal of public health* 95.5 (2005), pp. 881–886.

[32] WE Sponsel. “Tonometry in question: can visual screening tests play a more decisive role in glaucoma diagnosis and management?” In: *Survey of Ophthalmology* 33 (1989), pp. 291–300.

[33] WE Sponsel. “Tonometry in question: can visual screening tests play a more decisive role in glaucoma diagnosis and management?” In: *Survey of Ophthalmology* 33 (1989), pp. 291–300.

[34] William Brown. *Artificial Intelligence and Machine Learning: Ultimate Guide on the Technologies Impacting Our Daily Life, Including a Deep Analysis into Finance, Medicine, and Business Revolution*. Paperback edition, 154 pages. London, UK: Sagittarius Publishing, 2021. ISBN: 9781802524635.

[35] R. Mendez. *Machine Learning For Beginners: A Complete Overview for Understand How to Build Artificial Intelligence Through Data Science*. Independently published, 2020.

[36] John Paul Mueller and Luca Massaron. *Machine learning for dummies*. John Wiley & Sons, 2021.

[37] Sergios Theodoridis. *Machine learning: a Bayesian and optimization perspective*. Academic press, 2015.

[38] Nada Hussain Ali, Matheel Emaduldeen Abdulmunem, and Akbas Ezaldeen Ali. “Learning evolution: A survey”. In: *Iraqi Journal of Science* (2021), pp. 4978–4987.

[39] Ankit Ram et al. “Short Review on Machine Learning and its Application.” In: *Special Education* 1.43 (2022).

[40] Taiwo Oladipupo Ayodele. “Machine learning overview”. In: *New Advances in Machine Learning* 2.9-18 (2010), p. 16.

[41] V7 Labs. *Supervised vs Unsupervised Learning: Key Differences*. Accessed: 2025-07-03. 2023. URL: `https://www.v7labs.com/blog/supervised-vs-unsupervised-learning`.

[42] N d Sandhya and KR Charanjeet. “A review on machine learning techniques”. In: *International Journal on Recent and Innovation Trends in Computing and Communication* 4.3 (2016), pp. 451–458.

[43] Aditi Prerna. *Unsupervised Learning*. Accessed: 2025-07-03. 2020. URL: `https://aditi22prerna.medium.com/unsupervised-learning-a24caf362e79`.

[44] Amer FAH Alnuaimi and Tasnim HK Albaldawi. “An overview of machine learning classification techniques”. In: *BIO Web of Conferences*. Vol. 97. EDP Sciences. 2024, p. 00133.

[45] Vanshika Rastogi et al. “Machine learning algorithms: Overview”. In: *International Journal of Advanced Research in Engineering and Technology* 11.9 (2020).

[46] P. Ariwala. *9 Real-World Problems that can be Solved by Machine Learning*. Accessed: 2025-04-27. Maruti Techlabs. Oct. 2023. URL: `https://marutitech.com/problems-solved-machine-learning/`.

[47] Pushkar Nikam. *Neural Networks and the Human Brain: A Comparative Study of Imitation and Innovation in AI*. `https://doi.org/10.6084/m9.figshare.28076066`. Preprint. 2024.

[48] GeeksforGeeks. *Introduction to Neural Network*. Accessed: 2025-04-30. n.d.

[49] DataCamp. *Forward Propagation in Neural Networks: A Complete Guide*. Accessed: 2025-04-30. n.d.

[50] DataCamp. *Loss Functions in Machine Learning Explained*. Accessed: 2025-04-30. n.d.

[51] GeeksforGeeks. *Backpropagation in Neural Network*. Accessed: 2025-04-30. n.d.

[52] DataCamp. *Introduction to Activation Functions in Neural Networks*. Accessed: 2025-04-30. n.d.

[53] S. Rahman. *Activation Functions — Neural Networks*. Towards Data Science. June 2019.

[54] Adam A Alli and Muhammad Mahbub Alam. “SecOFF-FCIoT: Machine learning based secure offloading in Fog-Cloud of things for smart city applications”. In: *Internet of Things* 7 (2019), p. 100070.

[55] Ahmed Dawoud, Seyed Shahristani, and Chun Raun. “Deep learning and software-defined networks: Towards secure IoT architecture”. In: *Internet of Things* 3 (2018), pp. 82–89.

[56] Shuo Ouyang et al. “Communication optimization strategies for distributed deep neural network training: A survey”. In: *Journal of parallel and distributed computing* 149 (2021), pp. 52–65.

[57] Jiarui Fang et al. “RedSync: reducing synchronization bandwidth for distributed deep learning training system”. In: *Journal of Parallel and Distributed Computing* 133 (2019), pp. 30–39.

[58] Danyang Xiao et al. “EGC: Entropy-based gradient compression for distributed deep learning”. In: *Information Sciences* 548 (2021), pp. 118–134.

[59] Tien-Dung Cao et al. “A federated deep learning framework for privacy preservation and communication efficiency”. In: *Journal of Systems Architecture* 124 (2022), p. 102413.

[60] Jinrong Guo et al. “Tail: an automated and lightweight gradient compression framework for distributed deep learning”. In: *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE. 2020, pp. 1–6.

[61] Louis Bouchard. *Introduction to Convolutional Neural Networks (CNNs) | The Most Popular Deep Learning Architecture*. Medium, Accessed: 2025-05-02. 2025.

[62] Jayanth Koushik. “Understanding convolutional neural networks”. In: *arXiv preprint arXiv:1605.09081* (2016).

[63] Nebojša Bačanin Džakula et al. “Convolutional neural network layers and architectures”. In: *Sinteza 2019-International Scientific Conference on Information Technology and Data Related Research*. Singidunum University. 2019, pp. 445–451.

[64] Mohammad Mustafa Taye. “Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions”. In: *Computation* 11.3 (2023), p. 52.

[65] Anirudha Ghosh et al. “Fundamental concepts of convolutional neural network”. In: *Recent trends and advances in artificial intelligence and Internet of Things* (2020), pp. 519–567.

[66] Andrew Y Ng. “Feature selection, L 1 vs. L 2 regularization, and rotational invariance”. In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 78.

[67] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (2002), pp. 2278–2324.

[68] Leiyu Chen et al. “Review of image classification algorithms based on convolutional neural networks”. In: *Remote Sensing* 13.22 (2021), p. 4712.

[69] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[70] A Victor Ikechukwu et al. “ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images”. In: *Global Transitions Proceedings* 2.2 (2021), pp. 375–381.

[71] Chien-Yao Wang et al. “CSPNet: A new backbone that can enhance learning capability of CNN”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 390–391.

[72] Sangdoo Yun et al. “Cutmix: Regularization strategy to train strong classifiers with localizable features”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6023–6032.

[73] Moseli Mots’ oehli and Kyungim Baek. “Deep active learning in the presence of label noise: A survey”. In: *arXiv preprint arXiv:2302.11075* (2023).

[74] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[75] Rabbia Mahum et al. “A novel framework for potato leaf disease detection using an efficient deep learning model”. In: *Human and Ecological Risk Assessment: An International Journal* 29.2 (2023), pp. 303–326.

[76] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[77] Monika Bansal et al. “Transfer learning for image classification using VGG19: Caltech-101 image data set”. In: *Journal of ambient intelligence and humanized computing* (2023), pp. 1–12.

[78] Kamal Kamal and EZ-ZAHRAOUY Hamid. “A comparison between the VGG16, VGG19 and ResNet50 architecture frameworks for classification of normal and CLAHE processed medical images”. In: *Research Square* (2023).

[79] GeeksforGeeks. *Vision Transformer (ViT) Architecture*. Accessed: 2025-05-24. Jan. 2025. URL: `https://www.geeksforgeeks.org/vision-transformer-vit-architecture/`.

[80] Asifullah Khan et al. “A survey of the vision transformers and their CNN-transformer based variants”. In: *Artificial Intelligence Review* 56.Suppl 3 (2023), pp. 2917–2970.

[81] Alfonso Ramos-Michel et al. “Image Classification with Convolutional Neural Networks”. In: *Metaheuristics in Machine Learning: Theory and Applications*. Springer, 2021, pp. 445–473.

[82] Mohammad Hossin and Md Nasir Sulaiman. “A review on evaluation metrics for data classification evaluations”. In: *International journal of data mining & knowledge management process* 5.2 (2015), p. 1.

[83] Andres Diaz-Pinto et al. “CNNs for automatic glaucoma assessment using fundus images: an extensive validation”. In: *Biomedical engineering online* 18 (2019), pp. 1–19.

[84] Toaha Rahman Ratul. *ACRIMA Dataset*. `https://www.kaggle.com/datasets/toaharahmanratul/acrima-dataset`. Accessed: 2025-05-24. 2023.

[85] Jayanthi Sivaswamy et al. “Drishti-GS: Retinal Image Dataset for Optic Nerve Head (ONH) Segmentation”. In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. 2014, pp. 53–56. DOI: `10.1109/ISBI.2014.6867807`. URL: `https://doi.org/10.1109/ISBI.2014.6867807`.

[86] Lokesh Sai Pureddi. *Drishti-GS - RETINA DATASET FOR ONH SEGMENTATION*. `https://www.kaggle.com/datasets/lokeshsaipureddi/drishtigs-retina-dataset-for-onh-segmentation`. Accessed: 2025-05-24. 2023.

[87] Liu Li et al. “Attention Based Glaucoma Detection: A Large-scale Database and CNN Model”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10571–10580. DOI: 10.1109/CVPR.2019.01082. URL: https://doi.org/10.1109/CVPR.2019.01082.

[88] Kaggle Contributor. *LAG - Labelled Axial Glaucoma Dataset*. https://www.kaggle.com/datasets/kaggle-contributor/lag-dataset. Accessed: 2025-05-24. 2023.