Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Experiments in the automatic detection of optic disc abnormalities using ‘code-free’ software in practical ophthalmology

By
Dr Ciara O’Byrne
M.B., B.Ch., B.A.O. (Hons), MCh

A thesis presented to Trinity College, University of Dublin for the degree of
DOCTOR IN MEDICINE (M.D.)

February 2026

Supervised by
Professor Lorraine Cassidy

Table of Contents

Declaration, online access and the General Data Protection Regulation

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

_____

Ciara O'Byrne
Date: 31.08.2024

Declaration and Statement of Plagiarism

I have read and understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar. I have also read and understood the guide, and completed the 'Read Steady Write' Tutorial on avoiding plagiarism, located at https://libguides.tcd.ie/academic-integrity/ready-steady-write.

I declare that this work is my own and that I have clearly identified any areas where there was contribution from other participants on my team in Acknowledgements and I have made no use of sources or materials in this thesis that have not been openly and fully acknowledged in the text. My research involved human participants but as there was no active participation and all patient data was de-identified, informed consent was exempted. All research was conducted in alignment with the approved data governance regulations in order to ensure patient data security throughout.

_____

Ciara O'Byrne
Date: 31.08.2024

Acknowledgements

I would like to acknowledge my supervisor, Professor Lorraine Cassidy, for her continuous support and patience throughout the past few years. I would also like to acknowledge Professor Pearse Keane for envisaging the study and for his expertise, guidance and support. I would like to thank Dr Caroline Kilduff for her assistance with data collection and Dr Fares Antaki for his expertise and advice on the use of code-free deep learning. In terms of data scientist support, I would like to thank Timing Liu and Dominic Williamson for their assistance in processing the raw dataset, checking for errors and converting the images into the correct format necessary for code-free deep learning model development. I would like to thank Moorfields Eye Hospital and the team at Softwire for ethical approval, de-identification of the dataset, processing and upload to the Moorfields Google Cloud and for providing me with the required computer access needed to carry out this research. Thank you to all the patients and staff at Moorfields Eye Hospital for their continuous cooperation and support with the use of de-identified patient data for research purposes. Finally, I would like to thank my parents John and Fiona, and my fiancé Callum, for their unending support, encouragement and motivation.

List of Figures and Tables

Figures

Figure 8. Precision-recall curve and precision-recall at confidence threshold 0.5 for model one for all labels

Figure 9. Model one precision-recall curve for papilloedema label at 0.5 confidence threshold

Figure 10. Precision-recall curve and precision-recall at confidence threshold 0.5 for model three for all labels

Figure 11. Model three precision-recall curve for papilloedema label at 0.5 confidence threshold

Figure 12. External validation dataset 1

Figure 13. External dataset 2

Tables

Table 15. Model sixteen external validation performance by image

List of Abbreviations

AI	Artificial intelligence
AUC	Area under the curve
AUROC	Area under the receiver operating curve
AUPRC	Area under the precision recall curve
AutoML	Automated machine learning
BMI	Body mass index
BONSAI	Brain and Optic Nerve Study Artificial Intelligence
CFDL	Code-free deep learning
CFP	Colour fundus photograph
CNN	Convolutional neural network
CSV	Comma separated values
DICOM	Digital Imaging and Communications in Medicine
FDA	Food and Drug Administration
GPU	Graphics Processing Unit
ICG	Indocyanine green
IIH	Idiopathic intracranial hypertension
OCT	Optical coherence tomography

PNG	Portable Networks Graphics
NHS	National Health Service
RIS	Research information systems
ROP	Retinopathy of prematurity

Abstract

Background to the problem

The accurate diagnosis of optic disc swelling remains a challenge for non-ophthalmic clinicians and community opticians resulting in a large volume of referrals for tertiary review [1]. At present, the demand on Ophthalmology services is increasing considerably and it is now the busiest outpatient specialty within the National Health Service, United Kingdom [2]. In 2020, Milea et al. developed a bespoke deep learning model to differentiate between papilloedema, pseudopapilloedema and normal optic discs using colour fundus photographs with impressive results [3]. Unfortunately, bespoke deep learning depends upon highly advanced technical expertise as well as significant computational and financial resources. Code-free deep learning allows users to develop deep learning models without any coding expertise. It is also largely cloud-based removing the need for expensive computing resources. The aim of this thesis was to investigate the performance of code-free deep learning in the binary classification of 1) the presence or absence of papilloedema and, 2) normal versus abnormal optic discs.

The method

A systematic review of the clinical literature was performed to identify all records investigating the application of deep learning to the defined aims of this study. A bespoke dataset of optical coherence tomography images and colour fundus photographs was curated from the Moorfields Eye Hospital NHS Foundation Trust electronic patient healthcare record. A second dataset was identified from a publicly-available repository and downloaded for additional code-free model development and training. Following completion of model training, all models achieving comparable or superior performance to those cited within the clinical literature were subject to external validation.

Results

The systematic review identified seven records meeting the inclusion criteria that investigated the application of deep learning to optic disc abnormalities. A variety of deep learning approaches used within the literature were identified, however, none used code-free deep learning. A dataset of 6980 optical coherence tomography images and 6982 colour fundus photographs was curated from the electronic patient record including normal optic disc, abnormal optic disc and papilloedema labels. A second dataset was identified and downloaded for model training. Sixteen code-free deep learning models were trained in total by a clinician

with no coding expertise: fourteen models were trained using the bespoke curated dataset and two were trained using the publicly-available dataset downloaded from the online repository. Four code-free deep learning models achieved performances comparable or superior to the bespoke deep learning models previously described. External validation of these models demonstrated poor performances at all tasks.

Conclusion

The application of bespoke deep learning for papilloedema and/or optic disc abnormality identification has demonstrated impressive results within the clinical literature. Based on the results of this thesis, code-free deep learning is not presently a feasible alternative to bespoke deep learning for implementation into the direct patient pathway. Code-free deep learning instead may serve a role within the clinical research environment or as a tool for medical education.

Lay Abstract

The optic nerve transmits visual information from the eye to the brain. The optic disc is the part of the nerve at the back of the eye that can be examined by a clinician with the appropriate equipment. If there is increased pressure within the skull, this disc can become swollen. Optic disc examination is difficult and requires equipment such as an ophthalmoscope or lenses and a slit lamp. It is also a skill that is often poorly taught in medical schools. This means that many doctors are under-confident at examining the optic discs and often neglect this part of the clinical assessment. Accurate optic disc examination can also be challenging for opticians and community optometrists. Papilloedema describes the swelling of the optic discs secondary to raised pressure inside the skull. It can be caused by anything that increases pressure around the brain such as infection, bleeding, trauma, blockage of drainage channels or a brain tumour. Regardless of the specific cause, papilloedema generally indicates a serious problem and requires urgent assessment and investigation to exclude anything fatal. However, the definitive diagnosis of papilloedema by examination alone can be difficult. There are many other conditions that can lead to a similar appearance of swollen optic discs but do not require the same urgency of investigation. Nonetheless, a large proportion of these patients are referred urgently to specialist eye doctors for further investigation which places huge strain on services that are already under-resourced and under-staffed.

Deep learning is a type of artificial intelligence which has demonstrated impressive results across a variety of areas within healthcare. Despite excellent results, deep learning remains largely inaccessible to many clinicians and researchers. The development of these models requires significant financial resources for computing power, dataset development, and to enlist a data scientist(s) with artificial intelligence expertise who may otherwise be financially incentivised towards private industries. Unfortunately, these resources are often unattainable to hospitals, universities and small research groups. Code-free deep learning is a type of deep learning that allows users to develop deep learning models without coding expertise. It is available on many commercial platforms that are relatively cheap to use. There have been a number of articles published indicating that the performance of these code-free deep learning models is comparable to the deep learning models developed by computer scientists, herein known as bespoke deep learning models. This has generated significant interest and code-free deep learning has been proposed as a solution to the democratisation of artificial intelligence.

This thesis aimed to explore the application of code-free deep learning to the classification of papilloedema, other optic disc abnormalities and normal optic discs. This required 1) the identification of datasets to train the models, 2) a close assessment of the available articles published on the same topic, 3) code-free deep learning model development and 4) an examination of its reliability when tested with images previously unseen.

Two datasets are described in this thesis. The first dataset was created for the purpose of this research from the electronic patient healthcare record and included two different types of scans of the back of the eye. The second dataset was identified from an online database which is available to the public and free to download. This dataset contained one type of scan. Both datasets contained images of papilloedema, other optic disc abnormalities and normal optic discs. A comprehensive search was performed to identify all articles published that reported on deep learning models trained to investigate for papilloedema or optic disc abnormalities. This identified only bespoke deep learning models with no reported code-free deep learning models. Sixteen code-free deep learning models were developed and trained by a clinician without coding expertise. Four of these code-free deep learning models demonstrated comparable performance to those identified in the article search. However, these models performed poorly when they were tested on images that they had not been trained on.

The conclusion of this thesis is that code-free deep learning is not an appropriate substitution for bespoke deep learning at present. With future improvements this may change, however, at present, it is better suited for areas such as medical education or clinical research where it cannot impact the direct patient pathway.

Aims and Hypothesis of the Project

Hypothesis

The hypothesis of this project is that code-free deep learning demonstrates comparable or superior performance to bespoke deep learning when trained in the following tasks:

1. The binary classification of papilloedema versus no papilloedema
2. The binary classification of normal optic discs versus abnormal optic discs

Aims

1. Curation of a bespoke dataset from the Moorfields Eye Hospital NHS Foundation Trust electronic patient healthcare record to include both colour fundus photographs and optical coherence tomography images of 1) normal optic discs, 2) pseudopapilloedema, and 3) papilloedema
2. Identification of bespoke deep learning models trained for the same tasks published in the clinical literature and analysis of bespoke deep learning model performance
3. Development of code-free deep learning models in each imaging modality and for each task
4. Appraisal of code-free deep learning model performance and comparison to bespoke deep learning models reported on in the literature

Value of Research

There are many meaningful conclusions to be derived from the research presented in this thesis. Primarily, I have devised a method for training and evaluating machine learning systems for diagnostic techniques in ophthalmology. This has, perhaps, a broader application for image analysis in diagnostic medicine.

- Code-free deep learning models with relatively good performance can be designed and developed by a clinician without any coding expertise. This is significant as it implies that with future developments and reiterations of the products that provide code-free deep learning, it may be possible to develop models that have comparable performance to those developed by data scientists. This would be impactful in many ways:
    - Clinicians could become leaders within healthcare and artificial intelligence in the design and development of deep learning models particularly suited to the needs of their patients.
    - Lower resource areas would not be dependent on expensive resources and highly specialised artificial intelligence data scientists. Code-free deep learning models could be developed locally for the target population to enhance and expand pre-existing scarce resources e.g. in the form of screening programmes.

- This thesis indicates that code-free deep learning in its current state is not fit for implementation into the direct patient pathway as a substitute to bespoke deep learning. This is insightful as there are a number of publications within the clinical literature that propose code-free deep learning to be comparable to bespoke deep learning. It is possible that they were trained on datasets that were highly reiterated and revised as opposed to the real-world dataset (Dataset I) presented in this study that is reflective of a large multi-ethnic population.

- The development and curation of Dataset I, which is a very large diverse collection of labelled images in two different imaging modalities, represents a valuable resource that can be applied to future research projects.

- Code-free deep learning can serve as a useful proof-of-concept tool. By facilitating a clinician with no coding expertise to develop deep learning models, a deeper

understanding of the process around model development is acquired. This could be harnessed and used within medical schools or postgraduate teaching programmes when providing education on clinical artificial intelligence. This, in turn, could allow clinicians and researchers to develop their own code-free deep learning models for a clinical case of their choosing to test its performance before applying for investment and resources for a larger, more costly bespoke deep learning project.

Chapter 1: Introduction

1.1 Background to the problem

The demands on healthcare are increasing at an unprecedented rate in the face of an ageing population and the rising burden of chronic diseases and obesity. The COVID-19 pandemic exposed systemic weaknesses in healthcare systems globally with residual effects reflected today in lengthy waiting lists, strained resources and long-term health effects. Patients are becoming more autonomous in their healthcare plans which, though largely positive, has potentially led to an increase in investigative testing. This leads to significant administrative burdens that are challenging but remain necessary, in part due to rising medicolegal challenges within healthcare. Therefore, considerable pressure is being placed on clinicians and those working in the healthcare services with burnout reports at an all time high[4].

1.1.1 Ageing population and chronic diseases

Birth rates are declining and life expectancies are increasing. In Ireland, the Central Statistics Office reported a 20% decrease in births registered in 2022 compared to 2012 (annual birth rate of 11.3 per 1000 population compared to 15.7 per 1000 population in 2012) [5]. Similarly, in the United Kingdom, the birth rate was recorded at 10.2 births per 1000 people in 2020 compared to 12.9 in 2010 [6]. The effects of this are very concerning - we are facing an increasingly ageing population with a reduced available workforce. In 2020, Cristea et al. published an analysis of the impact of population ageing on European Union labour markets. The authors reported that the ageing population with associated health demands and increased medical costs will be at an imbalance with the workforce and that governments will “need to cope with increasing expenditures and declining revenues.” The authors advocate that an emphasis on the adoption of new technologies and the integration of datasets on age groups and conditions may facilitate the growing demands on services [7].

A growing elderly population will mean an increase in the burden of age-related chronic diseases. Many of these chronic diseases require regular follow-up and intervention to prevent further complications and sequelae. Jones and Dolsten discuss the challenges faced by the US healthcare systems in terms of a "healthcare paradox." The authors discuss the ageing population in the context of major limitations within the healthcare systems, notably the reduced workforce and capacity for these increased demands. The high prevalence of chronic diseases in this population requires specialised and coordinated care across a variety of settings. If this is not provided, patients face increased morbidity and mortality with reduced quality of life. The limitations in the healthcare services could mean that if demands continue to escalate, the system will be ill-equipped for the needs of its population [4].

Age-related macular degeneration, for example, is a degenerative disease of the retinal layers characterised by the presence of any of the following: drusen, pigmentary abnormalities, macular neovascularisation or geographic atrophy. It is the leading cause of legal blindness in the developed world and those with the neovascular type may require regular intravitreal injections in order to reduce the risk of vision loss [8,9]. In 2020, a systematic review and meta-analysis on the prevalence and incidence of age-related macular degeneration in Europe reported a prevalence of early or intermediate disease in 25% of those older than 60 years. The authors also predicted that there would be a 15% increase in any form of age-related macular degeneration within the European Union by 2050 [10]. This represents a concerning number of people with either a potential reduction in function due to loss of vision and/or a vast increase in service demands on healthcare systems that are already struggling. Data published by the National Treatment Purchase Fund (NTPF) in September 2023 indicated that there were 8777 people on the Ophthalmology adult waiting list, making it one of the busiest specialties in Ireland [11]. Similarly, the National Health Service (NHS) reported 7.5 million ophthalmology outpatient attendances in England in 2021/2022 with ophthalmology accounting for nearly 10% of all outpatient appointments. The United Kingdom estimated economic impact of vision loss is £28 billion[12].

However, it is not just the elderly that are placing demands on healthcare services. It is estimated that multiple chronic conditions affect one in three adults globally [13]. In 2020, it was reported that chronic disease affected 50% of the US population and accounted for 86% of healthcare expenditure [14]. In their narrative on the global burden of multiple chronic conditions, Hajat and Stein list ischaemic heart disease, stroke, lung disease, depression, diabetes, and

back and neck pain as some of the top conditions contributing to mortality and morbidity in high-income countries. The International Diabetes Federation reported in 2021 that there were 537 million adults living with diabetes globally with a USD health expenditure increase of 316% in the past 15 years [15].

The implications of these multiple chronic conditions result both in individual patient challenges, such as reduced quality of life, associated expenses, impact on role in the workplace and the burden of medication adherence, as well as challenges related to the system as a whole, in particular expenditure and resources. It is also expected that as the proportion of younger adults affected by multiple chronic conditions get older, these burdens will continue to increase. The authors propose that technology may provide a potential solution to some of these burdens by improving adherence, supporting lower resourced areas and through the provision of personalised care [16]. Jones and Dolsten also call attention to newer environmental issues that will affect our future health and longevity namely "obesity, processed food intake, microbiome changes, climate change, pandemics and pollution." This emphasises the need for robust healthcare systems equipped to deal with the potentially unforeseeable impacts these triggers may have [4].

1.1.2 COVID-19

1.1.2.1 Morbidity and secondary effects causing burden to healthcare systems

Chang et al. (2022) reported that "the outbreak of COVID-19 has affected almost all countries and caused significant loss of life, health and economic output" [17]. The Coronavirus disease 2019 (COVID-19) pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), placed unprecedented strain on healthcare systems around the world. The World Health Organisation (WHO) reported that at least 3 million people died worldwide in 2020 alone [18]. This excludes the millions of people infected by the disease, many of whom have been left with residual effects following recovery from the acute phase of the disease. In 2021, Lopez-Leon et al. reported on the results of a large systematic review and meta-analysis on the long-term effects of COVID-19. They found at least one residual effect beyond the acute phase of two weeks in 80% of the patients in their cohort with confirmed COVID-19. The authors report on 55 effects some of which include fatigue, lung dysfunction, headaches and neurological disorders as well as psychiatric disorders including anxiety, insomnia and dementia [19]. The strain on services during the pandemic also affected access for those with chronic diseases and

regular care needs. Olmastroni et al. (2023) commented on the effects of the COVID-19 pandemic on patients with chronic illnesses who depend upon regular medical input and therapy in order to effectively control their disease and reduce the risk of further sequelae [20].

1.1.2.2 Acute redistribution of resources during pandemic and secondary effects

Arsenault et al. comment that the duty of a healthcare system during a crisis such as the pandemic is to respond by providing acute services while also continuing ongoing essential health services. The authors discuss how declining healthcare use during the pandemic could have lasting effects on morbidity of the populations. The authors reported on the results of the QuEST network study which aimed to measure the resilience of healthcare systems during the COVID-19 pandemic through the sharing of national healthcare system data from a variety of countries. The authors demonstrated a decrease in outpatient attendances across all countries with the biggest impact observed in preventive care such as childhood vaccinations and screening programmes. Chronic disease appointments for diabetes and hypertension were reduced by 20% in six of the ten countries examined in the study. The authors conclude that their findings demonstrate a breakdown in the resilience of these healthcare systems when faced with acute stress and emphasise the need for ongoing preventive care and chronic disease management in such situations [21].

1.1.2.3 Economic effects of COVID-19 to healthcare systems

COVID-19 also had significant economic implications. In 2021, the International Monetary Fund estimated a $22 trillion cumulative output loss from 2020 to 2025 [22]. As discussed, the sheer magnitude of the pandemic highlighted any deficiencies in healthcare systems, notably infrastructures, staffing and resources. In 2022, Chang et al. published a review on the determinants of COVID-19 morbidity and mortality across countries looking at a variety of factors including technology, happiness, GDP, corruption, tourism and happiness. In terms of technology, the authors focused on “1) investment in emerging technology, 2) 4G mobile network coverage, and 3) use of the virtual social network.” The authors emphasise the importance of these three elements in the setting of a pandemic. Technology facilitates the monitoring and prediction of disease activity and also assists with the public health response in the form of contact tracing, awareness campaigns and data sharing between institutions to aid with ongoing decisions and management. Interestingly, Chang et al. found that technology was

associated with a reduced number of COVID-19 deaths. The authors postulated that this was due to the more effective public health response and improved technology-driven diagnostics allowing patients with COVID-19 to access medical input earlier, thereby reducing the risk of severe complications [17].

1.1.3 Limited resources and clinician burn-out

Jones and Dolsten predict that by 2030 there will be a deficit of 1.2 million registered nurses and 121,900 clinicians within the United States of America. The reasons for this are multifactorial but include increased workload, clinician pressures leading to burnout, uneven distribution of limited resources and, as discussed, an imbalance between the ageing workforce and sufficient supply of labour. The current systems are not fit for purpose and this is placing immense strain on healthcare workers often faced with the difficult situation of rationing care in the context of overcrowding, limited resources and cancellations [4]. Burnout can be defined as a syndrome of emotional exhaustion, depersonalization and a reduced sense of personal accomplishment in relation to a person's work [23]. Though exacerbated by the pandemic [24], the levels of clinician burnout were already rising [25]. West et al. reported a prevalence in burnout symptoms of greater than 50% among physicians in the United States. These estimates are significantly higher than other careers even taking into account similar work hours and other pressures. The main stressors implicated include excessive workloads, inadequate resources, large administrative and bureaucratic burdens and changing patient-doctor relationships. Technology has played two opposing roles within the literature - poorly designed technology can greatly increase the clinician's workload and stress levels whereas integrated digital services that enhance the clinical workflow can alleviate some of these pressures [262728]. Medicolegal considerations have also added to clinician anxiety [29] and administrative burden.

1.2 Papilloedema and its diagnostic challenges

1.2.1 Definition and epidemiology

Papilloedema is defined as optic disc swelling or oedema secondary to raised intracranial pressure [30]. It typically presents with blurred vision and may be accompanied by field defects. It also can present with features of raised intracranial pressure, such as vomiting, early morning headaches and fatigue [31].

The incidence of papilloedema varies depending on age and gender. Obese women of childbearing age are most commonly affected with incidence ranging from 11.9-19.3/100,000. In contrast, incidence rates in the general population range from 0.9-2.2/100,000 depending on the country and obesity levels [30]. The incidence of papilloedema in those with space-occupying lesions varies considerably depending on factors such as aetiology, location and chronicity. However, as intracranial tumours in children more frequently affect the posterior fossa with potential obstructive hydrocephalus, papilloedema is more common in this cohort [32].

1.2.2 Pathophysiology

The precise pathophysiology of papilloedema has been debated within the clinical literature [33] but, in essence, occurs when the Monro-Kellie doctrine is disrupted. The Monro-Kellie doctrine is a principal stating that the combined volumes of neuronal tissue, blood and cerebrospinal fluid (CSF) within the rigid cranial vault is constant [34]. An increase in one component must be off-set by a decrease in another to maintain homeostatic intracranial pressure. Xie et al describe the cerebrospinal fluid dynamics, axoplasmic flow stasis and secondary vascular changes in terms of their contribution towards the pathogenesis. Furthermore, the specific pathology behind axoplasmic flow stasis has been a subject of contention with the two leading mechanisms proposed to be (1) mechanical theory, versus (2) ischaemic theory [35].

Cerebrospinal fluid dynamics

The Monro-Kellie doctrine is one of intracranial homeostasis that states there is a fixed volume of blood, brain and cerebrospinal fluid and that an increase or decrease in one of these constituents results in an opposite reciprocal change in the others. Cerebrospinal fluid is continuously produced at a rate of 0.37mL/min by the arachnoid villi maintaining a stable volume within the brain. It is primarily produced by the two lateral ventricles where it travels via the interventricular foramina of Monro into the third ventricle, through the cerebral aqueduct into the fourth ventricle exiting into the subarachnoid space through the foramina of Magendie and Luschka [3336]. The failure of reciprocal change results in raised intracranial pressure [3734]. A number of different mechanisms can disrupt this pressure-volume relationship e.g., an increase in total amount of intracranial tissue, an increase in the production of cerebrospinal fluid or decrease in the cerebrospinal fluid either by reduced outflow or decreased absorption [30,34].

Axoplasmic flow stasis

The retinal ganglion cells transit from the eye through the lamina cribrosa, a rigid mesh-like disc at the optic nerve head, to form the optic nerve. The optic nerve sheath is continuous with the subarachnoid space. Axoplasmic flow is the metabolic exchange of nutrients between the cell bodies and axon terminals. Ordinarily, the intraocular pressure is higher than the optic nerve sheath pressure allowing for efficient axoplasmic flow. Therefore, if there is raised intracranial pressure, the pressure in the optic nerve sheath will increase. If this increase is greater than the intraocular pressure, the pressure gradient across the lamina cribrosa is affected and axoplasmic flow stasis can occur leading to optic nerve oedema visible on fundoscopy as papilloedema [3230,35]. The precise mechanism behind why this leads to flow stasis is described by two leading theories: the mechanical theory and the ischaemic theory.

The mechanical theory

The mechanical theory is based on the hypothesis that raised intracranial pressure leads to direct compression of axons. As discussed, this interferes with axoplasmic transport and causes accumulation at the lamina cribrosa. The effects of this are most evident at Bruch's membrane resulting in prelaminar optic nerve head oedema. Arterial compromise within the intralaminar optic nerve does not occur and the axoplasmic impairment is solely due to direct axonal compression [3538] [32].

The ischaemic theory

The second theory is known as the ischaemic theory which proposes that axoplasmic flow stasis occurs due to the effects of raised intracranial pressure on arterial microcirculation. The short posterior ciliary arteries are a network of arteries derived from the ophthalmic artery that supply the optic nerve head. A decrease in oxygen and glucose supply as a result of reduced perfusion affects the energy-dependent process of axoplasmic flow. This leads to stasis, buildup of cerebrospinal fluid metabolites, hypo-oxygenation and distension of the optic nerve axons [39] [32] [38] [35].

It must be noted, however, that raised intracranial pressure does not always correlate with optic nerve oedema. Due to natural anatomical variants within the population, the available space within the optic nerve sheath for cerebrospinal fluid flow varies and may lead to very early or very late manifestation of papilloedema. Regardless of the mechanism of flow stasis, papilloedema is associated with a number of clinical features. Although visual acuity can initially be unaffected due to relative sparing of the papillomacular bundle, symptoms tend to occur as oedema progresses. It must also be observed that the absence of papilloedema does not rule out raised intracranial pressure. In the case of chronic optic nerve atrophy, optic nerve oedema may not occur. In addition, Donaldson & Margolin reported that in certain cases of optic nerve atrophy, oedema may not occur even in the setting of large intracranial tumours [32].

It has been reported that sixty-eight percent of patients suffer from transient visual obscurations which are believed to be associated with periodic ischaemic episodes. One of the earliest visual field defects is enlargement of the physiological blind spot which may be due to the presence of peripapillary subretinal fluid. Subretinal fluid at the macula can also lead to reduced visual acuity as can macular haemorrhages and choroidal folds [40]. The mechanism behind the absence of spontaneous venous pulsations in patients with raised intracranial pressure has not been defined. Schirmer and Hedges propose the absence to be secondary to the central retinal vein path within the subarachnoid space [40]. Compression of venules within the retinal nerve fibre layer from axoplasmic stasis results in optic disc microaneurysms, central retinal vein distension and, in some cases, central retinal vein occlusion and retinal nerve fibre layer haemorrhages.

Prolonged optic nerve head oedema can cause irreversible damage to the retinal nerve fibre layer, leading to arcuate shaped visual field defects and optic neuropathy [41] [3540].

1.2.3 Aetiology of papilloedema

Papilloedema can arise secondary to any cause of raised intracranial pressure, some of which can be life-threatening. The most common cause of papilloedema is idiopathic intracranial hypertension (IIH) [42]. Idiopathic intracranial hypertension is a diagnosis of exclusion featuring raised intracranial pressure and papilloedema in a patient with no other identifiable cause. However, as some causes can be life-threatening, this must be considered a diagnosis of exclusion.

In 2020, Crum et al. carried out a retrospective population-based cross-sectional study of patients treated for papilloedema at outpatient eye clinics in Olmsted County, Minnesota, using the Rochester Epidemiology Project collecting data from January 1990 to December 2014. They concluded that most patients (67%) presenting to the eye clinic with papilloedema without a previously known cause were found to have idiopathic intracranial hypertension. These patients were more likely to present with headaches and had statistically higher body mass index (BMI). The most common causes of papilloedema without idiopathic intracranial hypertension were intracranial tumour, intracranial haemorrhage, and cerebral venous sinus thrombosis [42]. For this reason, idiopathic intracranial hypertension must be a diagnosis of exclusion following the investigation of any potentially life-threatening aetiologies. Herein, the more common aetiologies associated with papilloedema will be discussed in further detail.

1.2.3.1 Idiopathic intracranial hypertension (IIH)

As discussed, idiopathic intracranial hypertension is the most common cause of papilloedema. In 1955, it was described as benign intracranial hypertension by Foley [43]. However, without prompt treatment of the papilloedema, damage to the optic disc can occur which can result in permanent visual impairment leading to the most recent description of idiopathic intracranial hypertension.

Idiopathic intracranial hypertension typically affects women of childbearing age. It occurs most commonly in raised BMI and prevalence is increasing in line with increasing levels of obesity worldwide. Incidence has been reported to be 3-5/100,000 in females of reproductive age

increasing to 19/100,000 in those 20% or more over ideal body weight [44]. In fact, Crum et al. reported that 87% of patients presenting with papilloedema of unknown cause had idiopathic intracranial hypertension and that 95% of young obese women presenting with papilloedema had idiopathic intracranial hypertension [42]. Furthermore, obesity has been shown to correlate with symptom severity with odds of vision loss rising with every increase in body mass index [45].

The exact pathogenesis of idiopathic intracranial hypertension is not clear but it is believed to occur as a result of impaired CSF dynamics and venous sinus pressure [41]. Furthermore, the strong association between obesity and idiopathic intracranial hypertension has also not been defined. It has been postulated that it may be related to effects of central obesity on cardiac output and central venous pressures, however, the prevalence of idiopathic intracranial hypertension is not reflected in obese males or in pregnancy. Hormonal factors have been implicated which may explain the prevalence in women of childbearing age and in drug-induced idiopathic intracranial hypertension as can be seen in those using oral contraceptive pills[46]. Other hypotheses include the metabolic effects of adipose tissue on cerebrospinal fluid homeostasis, or reduced cerebral venous outflow [47].

There are a number of diagnostic measures, however, the most commonly utilised is the modified Dandy Criteria for diagnosis of idiopathic intracranial hypertension. This specifies that the patient must present with features of raised intracranial pressure such as headache, vomiting, transient visual obscurations and papilloedema, have a demonstrably high opening pressure >25cmH20 on lumbar puncture, neuroimaging that out rules any other known cause for raised intracranial pressure and no localising neurological deficits. This excludes sixth nerve palsies which can present in any patient with raised intracranial pressure and can be unilateral or bilateral. In 2014, Wall et al. published the results of the Idiopathic Intracranial Hypertension Treatment Trial. This was a randomised controlled trial which demonstrated the most common symptom to be headache followed by transient visual obscurations, back pain and pulsatile tinnitus. Visual changes ranged from visual loss, affecting 32% of patients, to visual field changes, most commonly arcuate defects with enlarged blind spot [48]. Specific radiological features reported in patients with idiopathic intracranial hypertension include empty sella syndrome, lateral sinus collapse, flattened sclera and unfolded optic nerve sheaths [4449,50].

Weight loss has been shown to be an effective first line management approach in patients with idiopathic intracranial hypertension. The mainstay pharmaceutical measure is the use of

acetazolamide. Associated conditions such as obesity, polycystic ovarian disease, Addison's disease and hypoparathyroidism should also be managed [47]. Refractory cases may require more invasive approaches such as optic nerve sheath fenestration, cerebrospinal fluid diversion and venous sinus stenting. Optic nerve sheath fenestration is a surgical intervention whereby small incisions are made in the meninges surrounding the optic nerve to allow for the release of cerebrospinal fluid, thus relieving elevated intracranial pressure. Recent systematic reviews and meta-analyses have demonstrated that optic nerve sheath fenestration significantly improves visual outcome, with visual acuity improvement rates ranging from 34.5% to 41.1% and visual field improvement rates from 69.4% to 76.3%. Early surgical intervention is associated with better visual outcomes, particularly in preventing advanced ischemic damage [51] [52] [53]. Due to the risk of optic nerve atrophy with chronic papilloedema, this condition must be followed up for a number of years until pressures have decreased to a normal level.

1.2.3.2 Space-occupying lesions

Benign and malignant intracranial tumours can lead to raised intracranial pressure due to mass effect. In the case of benign slow-growing tumours, compensatory sequelae can occur initially with an absence of optic nerve swelling. Location of the tumour is also important in the development of papilloedema with those located posteriorly more likely to cause ventricular obstruction and subsequent optic disc swelling than supratentorial tumours [30]. More rarely, spinal tumours can lead to an increase in intracranial pressure with subsequent papilloedema. The pathogenesis of this is postulated to be due to impaired arachnoid granulation absorption of cerebrospinal fluid as a result of blockage by tumour-produced cerebrospinal fluid proteins or bleeding from tumours, direct compression of the cerebellum and foramen magnum cerebrospinal fluid obstruction via direct compression in the case of upper cervical tumours [30].

1.2.3.3 Cerebral venous sinus thrombosis

Cerebral venous sinus thrombosis is a rare venous thromboembolic disease which can be associated with high mortality if not diagnosed in a timely manner [54]. It classically affects younger female patients, with a mean age reported to be 35 years and a risk ratio of 2:1 female to male. Other risk factors include genetic thrombophilias, malignancy, combined oral contraceptive pill, pregnancy, certain autoimmune conditions such as systemic lupus erythematosus, Behçet disease and vasculitis [55]. The incidence of cerebral venous sinus thrombosis has not been clearly defined but ranges from 2 to 15.7 annual cases per million within the clinical literature [55].

The most common vessels implicated in cerebral venous sinus thrombosis are the superior sagittal and transverse sinuses but the internal jugular veins and cortical veins can also be affected. Multi-vessel disease is also common, affecting up to two thirds of patients [55]. Raised intracranial pressure ensues when the increased venous pressure affects the venular and capillary pressure [56]. Patients often present with non-specific symptoms which can result in a delay in diagnosis and subsequent morbidity and mortality [54]. The most commonly reported symptoms are headache, seizures, paresis, papilloedema and mental state changes [57]. Papilloedema affects approximately 32-48% of adult patients with cerebral venous sinus thrombosis and confers a poorer prognosis in these patients when present [54]. Other ocular manifestations such as proptosis, orbital pain, chemosis and cranial nerve palsies can occur in cavernous sinus thrombosis [55]. Visual deficits could previously be attributed to delayed diagnosis, however, time to diagnosis has improved [58]. Liu et al. reported the results of a retrospective observational case series examining sixty five patients with diagnosed cerebral venous sinus thrombosis and documented papilloedema. The authors reported high levels of advanced papilloedema in their cohort of patients with Frisén grade 3 or worse affecting more than half of all patients. These patients had a significantly higher risk of vision loss and visual field loss with one fifth of patients showing progression of optic disc swelling during the course of their treatment [54].

Neuroimaging must be performed urgently in all patients presenting with features suspicious of cerebral venous sinus thrombosis. As discussed, the initial form of neuroimaging in patients with suspected intracranial disease is a computed tomography scan of the head. This may demonstrate the ‘dense triangle sign’ in superior sagittal sinus thrombosis or ‘dense cord sign’ for cortical or deep vein thrombosis. Contrast is also often added to increase the sensitivity for diagnosis of thrombosis whereby the ‘empty delta sign’ can be visualised in some cases [55]. Magnetic resonance imaging is the imaging modality of choice in these patients including magnetic resonance angiography and venography. The patient should also be thoroughly investigated for any of the known risk factors associated with cerebral venous sinus thrombosis [59]. Patients are managed first-line with anticoagulation but thrombolysis and endovascular treatment can also be initiated where appropriate. Furthermore, patients with papilloedema should receive serial ophthalmic evaluation with further input guided on an individual patient basis [54].

1.2.3.4 Cerebral haemorrhage and trauma

Papilloedema can occur, though less commonly, in subarachnoid and intraparenchymal haemorrhages either through cerebrospinal fluid outflow obstruction at the level of the ventricles or prevention of absorption. Papilloedema can also be seen in patients with subdural haemorrhage, particularly in acute stages [30]. Papilloedema has also been reported following head trauma. Hypothesised reasons include raised intracranial pressure secondary to intracranial haemorrhage, or in the absence of haemorrhage, impaired cerebrospinal fluid absorption, or cerebral oedema [30].

1.2.3.5 Meningitis

Papilloedema in meningitis is uncommon, particularly in the case of viral or aseptic meningitis where prevalence has been reported as little as 2% [60]. It typically occurs secondary to cerebral oedema, arachnoid inflammation leading to reduced cerebrospinal fluid absorption or obstructive hydrocephalus or mass effect from infiltrations in the case of cryptococcal meningitis or tuberculous meningitis, where papilloedema has been reported in twenty five percent of cases [6130].

1.2.3.6 Hydrocephalus

Obstructive hydrocephalus with subsequent raised intracranial pressure can occur as a result of space-occupying lesions, intracranial haemorrhages, infection or congenital aqueduct stenosis [30].

1.2.3.7 Malignant hypertension

Malignant hypertension is defined by a significant elevation in clinical blood pressure associated with papilloedema. As well as papilloedema, fundus examination demonstrates venous engorgement, absent venous pulsations, optic disc and/or peripapillary haemorrhage and macular exudates. The sequelae of malignant hypertension can be fatal leading to complications including cerebrovascular and cardiovascular infarctions, pulmonary oedema, acute kidney injury and intracerebral haemorrhage [62,63].

1.2.4 Pseudopapilloedema

Pseudopapilloedema is often mistaken for true papilloedema and can lead to unwarranted tests and investigations (Figure 1). It typically refers to congenitally anomalous discs, vitreopapillary traction or optic disc drusen. Pseudopapilloedema can result in a raised appearance of the papillary and/or peripapillary tissues often leading to the incorrect diagnosis of papilloedema [64] [35].

Pseudopapilloedema most commonly indicates buried disc drusen [65–69] [70]. Optic disc drusen have been reported to be present in approximately two percent of the Caucasian population and 0.4% of the paediatric population [71]. They are hyaline deposits that are variably calcified [64]. Optic disc drusen can be buried or visible on the optic nerve surface but tend to migrate superficially with age. They are believed to be autosomal dominant so family history must play an important role in the clinical assessment. Congenital anomalies can include absence of a physiological cup, elevated disc margins or abnormal vascular branching. Spontaneous venous pulsations can typically be visualised and eye examination is otherwise normal [71].

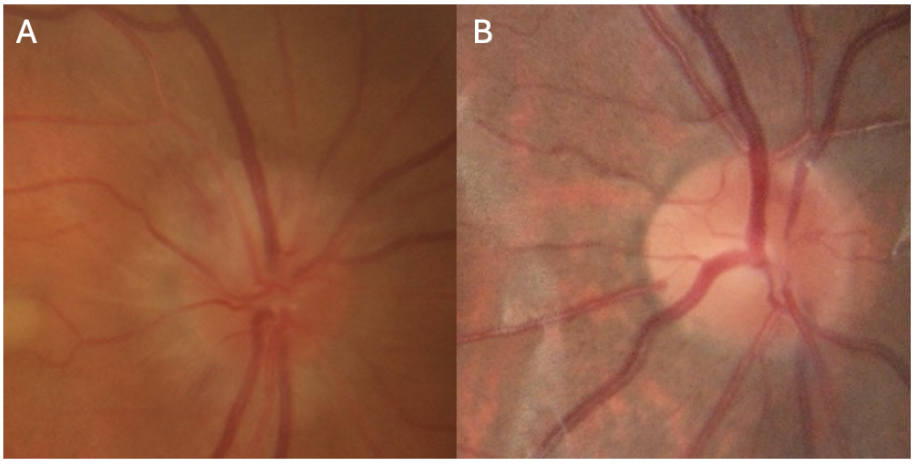

Figure 1. Papilloedema (A) versus pseudopapilloedema (B); *adapted from https://www.kaggle.com/datasets/shashwatwork/identification-of-pseudopapilledema?resource=download*

1.2.5 Diagnosis

Clinical examination

The clinical assessment of a patient with suspected papilloedema must be comprehensive. The clinician must ascertain from the clinical history whether there are any features suspicious for raised intracranial pressure including morning headaches, nausea, vomiting and pulsatile tinnitus. Features of a sixth nerve palsy should be assessed for including binocular horizontal diplopia and a new esotropia. Visual acuity must be recorded and the clinician should assess colour vision, which can be affected in more advanced disease. Typically, visual acuity is relatively unaffected in early papilloedema unless any previously discussed sequelae have occurred such as macular oedema, subretinal fluid or choroidal folds [71]. Careful examination of both optic discs with fundoscopy is essential and can be classified using the Frisén scale, a scoring system whereby stage 0 is normal optic disc and stage 5 indicates severe papilloedema [35]. A key factor to be conscious of when performing the clinical examination is that the optic nerve disc area is frequently smaller than average in cases of optic neuropathy, especially when using optical coherence tomography.

Formal visual field testing should also be carried out with defects such as enlarged blind spot, arcuate defects and generalised peripheral field constriction commonly reported [41].

Ocular imaging

Ocular imaging plays an important role in the investigation of suspected papilloedema.

Ultrasound

Ocular ultrasound has been shown to have a high sensitivity and specificity in the identification of anterior optic nerve sheath dilatation but cannot provide visualisation of the posterior portion risking unidentified posterior swelling [72]. Ultrasonography can be useful in the case of suspected disc drusen demonstrating optic disc calcification and acoustic shadowing. The clinician can also perform the “30 degree test” which shows a reduction in size of the intraorbital nerve with eccentric gaze as the oedema becomes compressed [71].

Fundus photographs

Colour fundus photographs are now readily accessible and quick to perform which is particularly favourable when assessing young children. Colour fundus photographs can also play an important role in the follow up care of these patients allowing clinicians to compare disc appearance with the previous visit or share images with colleagues for second opinions. Nonetheless, the use of colour fundus photographs in isolation when confirming the presence or absence of papilloedema has been shown to be poorly specific and should be used in conjunction with the rest of the clinical assessment [73].

Optical Coherence Tomography (OCT)

Sibony et al. (2021) performed a literature review with the aim of describing a practical-office-based set of tools using spectral domain optical coherence tomography in the diagnosis and monitoring of papilloedema, optic disc oedema and pseudopapilloedema. They found that although it is not a substitute for a complete history and careful examination, spectral domain optical coherence tomography is a convenient adjuvant test that aids in the diagnosis. It is particularly helpful in monitoring changes over the course of time and distinguishing early papilloedema from buried drusen which may appear as hyporeflective lesions [74 71].

Fundus Fluorescein Angiography (FFA)

In 2017, Chang et al. performed a prospective observational study to identify the most accurate diagnostic imaging modality for classifying paediatric eyes as papilloedema or pseudopapilloedema. Fluorescein angiography was found to have the highest accuracy (97%) for distinguishing between papilloedema or pseudopapilloedema. Fundus fluorescein angiography can demonstrate optic disc leakage in the case of papilloedema whereby optic disc drusen may show nodular staining or no hyperfluorescence at all if the drusen are buried [75] [64].

Neuroimaging

Any patient presenting with clinically concerning features of papilloedema must undergo urgent neuroimaging to exclude life-threatening conditions such as space-occupying lesions, hydrocephalus, meningeal abnormalities or cerebral venous sinus thrombosis. Neuroimaging can also demonstrate features of idiopathic intracranial hypertension including empty sella, transverse sinus stenosis and posterior scleral flattening [41]. Computed tomography of the head with computed tomography venogram are typically performed initially due to easier availability on an urgent basis. The imaging method of choice is magnetic resonance imaging of the head and orbits with contrast and magnetic resonance venogram with Mollan et al. recommending fat suppression sequences for enhanced visualisation of the intraorbital optic nerve [64].

Laboratory and lumbar puncture assessment

After urgent neuroimaging to exclude any intracranial space-occupying lesions, lumbar puncture should be performed to determine both the opening pressure and to ensure no cerebrospinal fluid abnormalities such as infection or malignancy. Elevated opening pressures >250mm H20 in adults are a defining sign of idiopathic intracranial hypertension. Blood samples may also be taken to exclude anaemia, certain renal or endocrine disorders, hypervitaminosis A and serum calcium and glucose abnormalities [4164].

Diagnostic challenges

The clinical diagnosis of papilloedema is challenging. Mollan et al. discuss the difficulties in distinguishing papilloedema from pseudopapilloedema and advise senior ophthalmological or neuro-ophthalmic input at an early stage in suspected papilloedema [64]. This process can cause significant anxiety to the patient and, in the case of paediatric referrals, the parents.

Furthermore, performing fundoscopy on very young children can be challenging which results in a stressful situation also for the clinician involved. Accurate diagnosis of pseudopapilloedema in the initial stage can help to avoid this stress and reduce the need for further investigations [71].

Challenges of fundoscopy

Mackay et al. described fundoscopy as a 'dying art.' Previously the mainstay method in which to view the fundus, challenges such as poor view and limited experience have made it unpopular in recent years. The ability to obtain high resolution colour fundus photographs have provided an accessible way in which to acquire a detailed view of the fundus which can be stored within the patient's electronic health record for future comparisons. Clinicians can magnify and measure various elements within the image and, if maintaining appropriate data governance regulations, colour fundus photographs can be shared with other clinicians for second opinions and specialised expertise [76].

The Teaching Ophthalmoscopy to Medical Students (TOTeMS) was a prospective randomised study that assessed accuracy and preferences of 119 first-year medical students learning to examine the fundus using human volunteers, a simulator and ocular fundus photographs. The majority of students (77%) reported a preference for fundus photography over fundoscopy and also performed more accurately with the images. The authors followed up the same cohort one year later reporting the same preference for fundus photographs. The authors also found that student use of ophthalmoscopy while performing clinical examinations was <10% with reasons including lack of confidence and discouragement from supervisor [77 78].

In 2016, Fisayo et al. reported a 39.5% overdiagnosis of idiopathic intracranial hypertension in patients referred to a neuro-ophthalmic clinic. The authors outlined that when faced with a patient with headache, 20% of clinicians did not perform fundoscopy and 44% incorrectly diagnosed papilloedema on fundus examination. The authors note that their results are consistent with the FOTO-ED I study in which only 12% of clinicians performed fundoscopy when examining patients for headache in the emergency department. Potential reasons cited by the authors include easy access to neuro-imaging, poor training in fundus examination with resultant lack of confidence in interpreting findings [79]. Similarly, Ang et al. found that 56.9% of junior house officers did not routinely examine the ocular fundus with the main obstacles being lack of available equipment, time pressures, poor training and limited confidence [80]. Nonetheless, this is a shocking and alarming statistic, and highlights the danger of loss of

clinical skills that clinicians now show especially with an overemphasis and an overreliance on imaging.

FOTO-ED was a series of studies examining fundus examination in the emergency department. In the initial phase of the study, Bruce et al. reported that emergency physicians performed direct ophthalmoscopy in 14% of patients presenting to the ED with complaints warranting fundus examination. The authors demonstrated that the use of non-mydriatic fundus photography resulted in the detection of more than 80% of pertinent findings, previously missed during clinical evaluation. The authors cite similar reasons as are cited within the clinical literature for poor use of fundoscopy including limited training [81,82], infrequent pupillary dilation in the ED, suboptimal identification of relevant ophthalmic findings [80,83] and time pressures on clinicians. The authors advocate the use of fundus photography to provide clinicians with a clear and large field of view. Images can also be performed by technicians, further reducing the burden on clinician time [84]. The authors subsequently assessed the quality of fundus photography obtained by nursing staff in the emergency department and reported high-quality images using a non-mydriatic camera in 85% of patients [85]. Phase II of the FOTO-ED study then examined diagnostic accuracy of emergency department physicians when presented with ocular fundus photographs. The authors reported that 46% of those clinicians that had identified none of the relevant ophthalmic findings with only direct ophthalmoscopy available were able to identify the relevant findings. The authors caution, however, that specialist ophthalmic input should not be replaced where required and that colour fundus photographs should primarily serve as a support tool for emergency clinicians in triage decisions. The authors conclude that non-mydriatic ocular fundus photographs were used more frequently than direct ophthalmoscopy by emergency department clinicians with a higher sensitivity for pertinent ophthalmic findings than direct ophthalmoscopy [86].

Underconfidence in papilloedema diagnosis

In February 2012, an optometrist in the United Kingdom was convicted of gross negligence manslaughter for a missed diagnosis of papilloedema in an 8-year-old boy. This has had a resounding effect on optometric and ophthalmic practice with a resultant over-diagnosis of papilloedema in community centres requiring subsequent urgent tertiary centre referral for specialist opinion. This has resulted in an immense increase in the burdens on specialist

ophthalmic centres but has also led to significant stress and anxiety for patients and often unnecessary tests and investigations. In 2018, Poostchi et al. performed a retrospective review of all computed tomography and magnetic resonance imaging head requests made over a three year period referencing papilloedema. The authors reported an increase from 0.85% to 1.1% after August 2016, most notable in the Emergency Department. [87]

Blanch et al. reported very high false positive rates of papilloedema in community testing based on fundus photography images which they reported to be 21% for a community sample of 12-14 year olds which subsequently decreased to 7% for a hospital sample of images within the eye clinic. The authors also reported a high sensitivity for papilloedema among emergency medicine clinicians when fundoscopy was performed in a targeted approach and taken into context with the clinical history. The authors emphasise that their results should discourage the use of fundus imaging as a screening measure given the potential harm the high rates of false positives could cause both in terms of patient anxiety, over-investigation and burden on healthcare services [73].

Once a patient is referred into a tertiary centre with suspected papilloedema, they are subject to a number of further investigations which can often be invasive or associated with high levels of radiation. In the case of young children who would not tolerate these investigations, general anaesthetic or sedation may be required further increasing a risk of complication. In their prospective analysis of suspected papilloedema paediatric referrals to an ophthalmology clinic, Kovarik et al. reported the presence of papilloedema in two of the 34 patients with the remainder diagnosed with either pseudopapilloedema, normal variant or unrelated aetiology. Many of these patients had been subjected to unnecessary and invasive investigations. The authors also note that while headache was a common symptom in most of the cohort, only the patients with true papilloedema reported headache-features typical for raised intracranial pressure [70]. These features include headaches that are constant and typically worse in the morning, when lying flat or performing the Valsalva manoeuvre. They can be associated with nausea, photophobia, pulsatile tinnitus and visual symptoms.

1.3 Code-free deep learning as a solution

1.3.1 Deep learning

Artificial Intelligence (AI), and more specifically deep learning, has been the subject of much interest in recent years as a potential solution to the various problems facing healthcare at present.

1.3.1.1 What is deep learning?

Alan Turing is considered a founder of computer science having proposed the 'Turing Test' in 1950, which "examines a machine's ability to show intelligence indistinguishable from that of human beings"[88,89]. The actual term 'Artificial intelligence' was first coined in 1956 at the Dartmouth Summer workshop which was held by John McCarthy, Assistant Professor of Mathematics at Dartmouth and other pioneers in the area including Marvin Minsky, Nathaniel Rochester and Claude Shannon. Experts in a variety of fields including mathematics, psychology, electrical engineering, and biology were invited to discuss various approaches to the progression of artificial intelligence [90]. This paved the way for many of the founding concepts in Artificial Intelligence. Throughout those initial decades, artificial intelligence experienced 'winters' and 'summers' whereby a new concept would be proposed generating hype and investment but following disappointing results, interest would fall until the proposal of a new concept. Some of those initial concepts included Logic Theorist, Expert Systems, Bayesian networks and Artificial Neural Networks [88].

The landscape would change drastically in the early 2000s when significant advancements in computing power in the form of Graphic Processing Units (GPUs) coincided with a rapid increase in the generation of data [91]. The power of harnessing such data was recognised by artificial intelligence researcher, Fei-Fei Li, who coordinated the production of ImageNet, a dataset reported to contain approximately two million images in 2009. Each image was annotated with a corresponding WordNet noun [92]. This led to the launch of the 2010 ImageNet Large Scale Visual Recognition Challenge. This was a competition that assessed 1) binary image classification performance, and 2) object recognition with the goal of encouraging ongoing advances in the field of artificial intelligence. The winners in the initial two years made

use of vector machines which showed modest results. [93]. However, in 2012, AlexNet was submitted demonstrating record-breaking results. AlexNet was a deep learning model trained using supervised learning. [94] Deep learning, first described by Geoffrey Hinton in 2006 [95], would quickly become the next “biggest thing”.

Deep learning is a subtype of artificial intelligence. Artificial intelligence (AI) is the capacity of computers or other machines to exhibit or simulate intelligent behaviour; the field of study concerned with this. In later use also: software used to perform tasks or produce output previously thought to require human intelligence, esp by using machine learning to extrapolate from large collections of data. Also as a count noun: an instance of this type of software; a (notional entity exhibiting such intelligence. Machine learning is the capacity of computers to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and infer from patterns in data; the field of artificial intelligence concerned with this.

Deep learning is a type of machine learning considered to be in some way more dynamic or complete than others; esp. machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively more features from data. A Convolutional Neural Network (CNN) is a type of deep feedforward neural network designed for analyzing grid-like data (e.g. images, audio, time-series). It automatically learns hierarchical feature detectors - from simple edges in early layers to complex objects in deeper ones. Convolutional neural networks are the architecture that was used in the deep learning model, AlexNet, previously discussed. Following this performance, Convolutional neural networks are used in nearly all recognition and detection tasks [96]. They are particularly suited to clinical data and have shown impressive results in a number of areas including dermatology [97,98], cardiology [99], radiology [100], ophthalmology [101] and others [102,103].

1.3.1.2 Challenges of dataset curation for deep learning

The raw input data forms an integral part of the process and deep learning models depend upon robust datasets in order to learn and output high quality predictions. Healthcare data is abundantly available ranging from advanced radiological imaging such as magnetic resonance imaging or computed tomography scans to those more available in the community including chest x-rays or optical coherence tomography scans. More recently, smartphone photographs

have progressed in quality allowing patients to upload images to the appropriate place for assessment. All of these modalities generate significant amounts of valuable data with the potential to greatly enhance the future of healthcare if applied appropriately with deep learning. Nonetheless, this data is often inaccessible for a variety of reasons. Firstly, healthcare data is highly sensitive and must be stringently protected. Anonymisation of the data may be required prior to usage and governance barriers can be difficult to navigate in order to obtain the appropriate ethical approvals. Secondly, data can often be difficult to access if it is not easily transferred from where it is stored, if it is not computationally compliant or if it is of variable quality. Images should ideally be converted to the same format that is used in routine clinical practice [104]. Thirdly, dataset curation requires significant resources. The upload and storage of healthcare data in a secure location is expensive and may require specialised input. Labelling of datasets is timely and requires a team with expertise in the area to ensure the labels are highly accurate. In the case of healthcare datasets, the optimum team for this task would typically be clinicians in their specialty field of medicine. A labelling protocol should be designed in advance and the team should perform rigorous quality checks to ensure accuracy. However, identifying busy clinicians with the time to devote to dataset curation can be challenging and costly. Therefore, hospitals and clinical research centres with access to healthcare data often do not have the financial resources to appropriately utilise this data in a meaningful way.

Publicly available imaging datasets are a valuable resource allowing researchers to bypass the challenges of dataset curation by downloading the appropriate dataset for their use case. A use case describes the selected application for the deep learning model, for example a clinical use case could be a deep learning model trained to distinguish between malignant and benign skin lesions. Publicly-available datasets exist within a number of areas, particularly in ophthalmology and feature heavily within the clinical literature [105]. However, these datasets have their limitations. Thirunavukarasu et al. (2023) outline the importance in assessing the accuracy of labels within publicly available datasets. Furthermore, the dataset may have been designed for a different project and researchers must examine it to ensure it is appropriate for their specific use case [104]. Researchers must also establish whether the dataset is representative of their intended population in order to avoid bias and poorer performance for groups under-represented in these datasets [106].

Dataset curation is not the only challenge facing researchers and clinicians in the development of deep learning models. These models require considerable computing power, often in the form

of graphics processing units (GPU), as well as data scientists highly specialised in the field of deep learning. The demand for these experts is not limited to the field of healthcare research and they are often financially incentivised towards the major technological and business firms. Herein, deep learning models designed in this way will be referred to as bespoke deep learning models.

1.3.2 Code-free deep learning

Deep learning, also known as bespoke or custom deep learning, depends upon considerable resources as outlined. For this reason, automated machine learning (AutoML) started to gain attention. AutoML is defined as the process of automating part of the design and development of machine learning models [107]. It can reduce resource requirements in the development of machine learning models by assisting with dataset management, model selection and tuning, and hyperparameter optimisation. Methods include Auto-WEKA, Auto-SkLearn, TPOT, and AlphaD3M [107,108]. Nonetheless, autoML is a tool of *assistance* and some expertise in machine learning is still required.

Code-free deep learning, also known as automated deep learning, is a further subtype of automated machine learning. Code-free deep learning automates the entire process of network architecture selection, pre-processing and optimisation of hyperparameters to allow those without coding expertise to design and develop deep learning models. The model uses supervised learning when provided with a labelled training dataset to infer patterns and produce predictions. Code-free deep learning has been described as a potential solution to the democratisation of artificial intelligence by allowing those without the budget to fund expensive deep learning projects to devise and develop their own systems. [109]. For this reason, it has gained considerable attention from the clinical and research communities which is visibly reflected in the clinical literature [110–114].

Neural Architecture Search, introduced by Zoph et al. in 2016, uses a recurrent neural network to generate neural network architectures which is what forms the basis of code-free deep

learning algorithms. The models are trained using reinforcement learning whereby the model learns by trial and error [115]. Code-free deep learning is now available on a number of different commercial platforms including Google Cloud AutoML Vertex AI, Microsoft Custom Azure, Apple ML, Amazon Rekognition and Huawei.

1.3.2.1 The process

Model development using code-free deep learning involves data upload, data visualisation and model evaluation. Although specific features may vary between platforms, the core components remain comparable. There is a simple graphical user interface that can be navigated without any need for coding expertise including processes such as data upload and model training. The type of code-free deep learning model that can be developed depends upon the specific platform but options include image classification, segmentation and object detection and tabular data models.

Most code-free deep learning platforms are cloud-based with the exception of Apple Create ML which uses the local hardware. The advantages of cloud-based software are a reduction in resources required in terms of local computing power but also a secure network in which to store sensitive data, such as patient imaging. Cloud-based approaches are typically subject to industry-standard encryption and compliance audits making them a more appropriate choice when storing sensitive data provided the correct governance procedures are followed [104].

The feature differences between various platforms have been explored within the clinical literature. In 2021, Korot et al. explored the diagnostic accuracy and features of six code-free deep learning platforms: Amazon Rekognition Custom Labels, Apple Create ML, Clarifai Train, Google Cloud AutoML Vision, MedicMind Deep Learning Training Platform and Microsoft Azure Custom Vision [110]. Subsequently, in 2023, Santomartino et al. reported on the usability of code-free deep learning platforms assessing aspects such as the feasibility of data labelling, upload and model training, external testing and costs [111].

1. Data labelling

Code-free deep learning does not remove the task of curating high quality and robust datasets and this remains a critical part of the process. High performing models are dependent on accurate datasets reflective of the target population and appropriate for the clinical use case while adhering to governance considerations. Santomartino et al. were unable to perform any code-free modifications of the datasets on any platforms explored [111].

2. Data upload

The graphical user interface of each platform uses simple icons or drag and drop features to allow data to be uploaded which must be in pre-labelled folders for Amazon, Clarifai, MedicMind and Microsoft. Nonetheless, this loses feasibility when very large datasets are to be uploaded for Amazon, Clarifai, Google AutoML Vertex AI and Microsoft. In this instance, the dataset can be uploaded to a storage bucket if using Amazon or Google. Google and MedicMind facilitate the upload of comma separated value files to be uploaded via the graphical user interface which, in the case of Google, link it to the specific cloud bucket containing the dataset. An advantage of comma separated value files is that the researcher can perform local data manipulation prior to upload. For upload of large datasets with comma separated value files with Google, some simple coding is required. However, speed of upload is improved when utilising the coding approach in comparison to the graphical user interface. Korot et al. were unable to process large datasets using Clarifai and MedicMind [110]. Santomartino et al. reported that though there were no limitations in uploading their largest dataset (n=112120 images) via the graphical user interface using the MedicMind platform, the dataset ultimately failed to upload due to system crashing. As coding upload is not an option, this significantly limits the usability of MedicMind when working with large datasets. Apple allowed for the dataset to be uploaded entirely code-free, however, this uses the local computer hardware and data is labelled based on local folders [111]. As discussed, this may not be preferable when using sensitive data that would be more securely-stored on the cloud.

3. Model training

Korot et al. assessed each platform on image classification tasks using two modalities of medical imaging optical coherence tomography images and colour fundus photographs from publicly available datasets. In order to assess the overall deep learning model performance, the

authors calculated mean F1 scores for all model-dataset pairs for each platform. Results showed large variation between platforms: Amazon, 93.9 (5.4); Apple, 72.0 (13.6); Clarifai, 74.2 (7.1); Google, 92.0 (5.4); MedicMind, 90.7 (9.6); Microsoft, 88.6 (5.3) [110]. Santomartino explored image classification, object detection and segmentation. Santomartino et al. reported that no segmentation model was successfully trained. Binary image classification is the most robust task across platforms based. Nonetheless, the authors were unable to successfully train a single-label classification model using Apple due to crashes. Multilabel classification models could be trained using Google, Amazon, Microsoft and Clarifai. The authors report that an object detection model was successful only on the Google platform. [111]

4. External validation

External validation is an essential part of the model development pipeline. This is an area where many platforms remain suboptimal. Only Google and MedicMind allow external validation of models. Similar to the upload of large datasets, both Korot et al. and Santomartino et al. were unable to perform external validation using MedicMind. Google allows for minimal datasets to be uploaded for external validation via the GUI and requires coding expertise in order to perform batch prediction with large external datasets [110,111].

5. Additional platform features

Korot et al. also explored features considered to enhance reproducibility and model explainability. Amazon, Apple, Google and MedicMind allow the dataset split to be manually performed. Apple allows data augmentation and Google and MedicMind allow for datasets to be uploaded via comma separated value files which provides researchers with the option to manipulate data locally. MedicMind offers saliency maps. Saliency maps, or heat maps, visually represent the locations within an image the model has focused on [116]. For this reason, they are considered an explainability tool within the field of deep learning although the actual usability of them is mixed according to the clinical literature [117]. Threshold adjustment can be carried out with Clarifai and Google. This is an important element to consider as sensitivity or specificity priorities may vary depending on the clinical use case. Confusion matrices are also an important measure that are provided by Apple, Clarifai, Google and MedicMind in order to truly assess model performance [110].

6. Cost

Though the resources required to develop a code-free deep learning model are considerably less than bespoke, all available platforms, at present, are commercial and have associated costs. Nonetheless, they offer free credit and tier options. Santomartino et al. report that the code-free deep learning models they developed with Apple and MedicMind were free and that the other platforms cost a maximum of $100 per model. The authors report in descending order the costliest aspects to be model training, image storage, application transactions and batch predictions [111]. Korot et al. acknowledge that free tiers are often insufficient when utilising large datasets which can lead to high costs when using Amazon, Google and Microsoft which charge per node hour and Clarifai which charges per image number. The capacity with Google to select an edge model rather than cloud-based could be more cost-effective [110]. Edge computing allows the model to be downloaded to a local device and can be run offline [118].

1.3.2.2 Applications of code-free deep learning (CFDL)

Image classification in ophthalmology

One of the fundamental papers on code-free deep learning was published in 2019 by Faes et al. examining the feasibility of "automated deep learning design for medical image classification by health-care professionals. The authors demonstrated that two ophthalmologists with no coding expertise could develop code-free deep learning models using "five publicly available datasets of retinal fundus images, optical coherence tomography images, dermatological skin lesion images and chest x-ray images" with comparable results to bespoke deep learning models published in the literature [119]. Although this specific paper includes a variety of specialties, ophthalmology has been an exemplary specialty in the applications of code-free deep learning. As previously discussed, Korot et al. evaluated six commercially available CFDL platforms at the task of image classification. This was carried out using two retinal fundus photograph datasets and two optical coherence tomography datasets [110].

In 2020, Kim et al. curated a dataset of 783 ultra-widefield indocyanine green (ICG) angiography images and developed two image classifier code free deep learning models using Google Cloud AutoML. In this paper, the results of the code-free deep learning model were compared against

ophthalmic residents and retinal specialists rather than bespoke deep learning models. Interestingly, one of the models trained actually demonstrated better precision and accuracy than the retina specialists with comparable recall and specificity [120]. Korot et al. (2021) also published the results of a code-free deep learning model trained to predict reported sex from a dataset of 84, 743 retinal fundus photographs from the UK Biobank dataset [121]. This had previously been demonstrated by Poplin et al. using bespoke deep learning to accurately make a variety of predictions from fundus photographs including cardiovascular risk factors, age and gender [122]. Poplin et al. reported an area under the receiver operating curve (AUROC) of 0.97 for their model which was comparable to the code-free deep learning model designed by Korot et al. performing with an AUROC of 0.93 and requiring no coding expertise for development. Notably, the code-free deep learning model performance was worse on external validation with a dataset containing foveal pathology in 42.9% of the images. In images without foveal pathology, the code-free deep learning model accuracy was reported to be within 1.5% of the Biobank validation set. The authors postulate that the poor performance in eyes with foveal disease indicates that the fovea plays an important role in the prediction of sex. This was supported by saliency maps generated by both the authors and by Poplin et al [121].

Humer et al. designed a code-free deep learning for detection and evaluation of posterior capsule opacification. Posterior capsule opacification is the most common complication after cataract surgery with intraocular lens implantation [123]. The authors used a dataset of 179 images with varying severity of posterior capsule opacification. A binary code-free deep learning model was trained to classify the images as significant or non-significant. The authors reported good performance in the classification of clinically significant posterior capsule opacification when the model was externally validated reporting area under the curve (AUC) of 0.9661 (95% CI 0.93 to 0.99), sensitivity of 0.84 and specificity of 0.92. The authors propose a potential use case for this code-free deep learning classifier to be as a decision support tool in the exclusion of posterior capsule opacification, particularly if included in a smartphone application [112].

Retinopathy of prematurity (ROP) is another area within ophthalmology particularly suited to the application of deep learning. Retinopathy of prematurity is a condition occurring in newborn babies associated with abnormal vasculature development in the retina. If left undetected and untreated, it can lead to permanent sight loss. Unfortunately, as it is challenging to diagnose, it depends upon the availability of highly expertised clinicians. Diagnosis is typically made either by clinical examination or by utilising wide angle contact fundus cameras [124]. Acknowledging the

'scarcity' of experts available to diagnose retinopathy of prematurity, Wagner et al. designed and developed both custom-engineered and code-free deep learning models for the detection of plus disease in retinopathy of prematurity. The authors demonstrated a comparably high performance in the classification of healthy versus pre-plus or plus disease for the bespoke model (AUC 0.986) and the code-free deep learning model (AUC 0.989) which was subsequently reflected on external validation and on comparison of the performance of senior paediatric ophthalmologists at the same task. Notably, the code-free deep learning model demonstrated poorer performance when distinguishing pre-plus disease from healthy or plus disease [125].

More recently, Jacoba and colleagues published the results of a code-free deep learning model for diabetic retinopathy classification using handheld retinal images. A dataset of 17829 de-identified retinal images with diabetes was curated and Google Cloud AutoML was subsequently used to train an image classification model. External validation with a publicly available dataset of 3662 retinal images demonstrated an AUPRC of 0.987. The United States Food and Drug Administration (FDA) has approved two autonomous deep learning models for the detection of diabetic retinopathy [126]. However, as discussed, the development of these models requires huge resources. Based on the results of this code-free deep learning model, the authors propose the significant value it could offer, particularly in lower-resourced areas [113].

Image classification in other specialties

Code-free deep learning has also been applied to use cases across a variety of other clinical specialties. As discussed, the landmark paper by Faes et al. in 2019 did not just use ophthalmic imaging but also explored the application of code-free deep learning in dermatology and radiology. The authors trained a code-free deep learning model to classify between seven categories of skin lesions which had an AUPRC of 0.93, sensitivity of 91% and positive predictive value of 91%. A binary classification model was trained to distinguish between normal versus pneumonia on paediatric chest x-rays with the model demonstrating a sensitivity of 97% and a specificity of 100% with an AUPRC of 1. Furthermore, one model trained in the classification of chest x-rays showed comparable performance to bespoke deep learning models published in the clinical literature [119]. Code-free deep learning chest x-ray classification was also explored by Borkowski et al. using Microsoft CustomVision with impressive results, supported by external validation [127].

Due to the vast generation of imaging data, histopathology is also very suited to the application of deep learning. One of the earliest applications of code-free deep learning in histopathology came from Borkowski et al. in 2019 with their comparison between Google AutoML Vision versus Apple Create ML in a series of binary code-free deep learning models trained in the differentiation of a variety of lung and colon pathologies [128]. Zeng et al. explored the use of code-free deep learning for the identification of invasive ductal carcinoma using Google Cloud AutoML using a publicly available histopathology dataset. The authors reported an AUPRC of 91.6% on initial evaluation with the model scoring a balanced accuracy of 84.6% on external testing [129]. Koga et al. used Google Cloud AutoML in the image classification of three different types of tauopathies with good results [130]. More recently, Jungo & Hewer compared Google AutoML and Microsoft Custom Vision in the classification of central nervous system tissue samples using haematoxylin and eosin-stained images [131]. One of the latest applications of code-free deep learning in the clinical literature is in the task of pill recognition. Ashraf et al. curated a dataset of 26880 training images, 1440 validation and 1440 testing images of thirty commonly dispensed solid oral dosage pills at three hospitals participating in the study. They subsequently used Microsoft Azure Custom Vision to train a pill recognition model with 98.7% precision and 95.1% recall. Notably, the authors downloaded their model to run offline on a smartphone application with a precision of 86.50% and recall of 75% [132].

Video Classification

There is a paucity of evidence within the literature on the use of code-free deep learning video classification. In 2022, Touma et al. developed a CFDL model to differentiate between cataract surgery phases. The model was trained using two publicly available datasets of surgical videos on Google Cloud AutoML Video Classification with external validation performed from an alternative dataset. The authors demonstrated good performances with an AUPRC of 0.855, sensitivity of 81.0% and recall of 77.1% [133]. In 2024, Touma et al. utilised optical coherence tomography macular videos to train code-free deep learning models to detect retinal pathology. The video classification performance was excellent with the authors reporting an AUPRC of 0.984. This code-free deep learning model was subsequently compared against a custom-engineered deep learning model created using transformers, an approach known to have excellent results in the field of video classification. The AUPRC of the custom model was 0.9097. Nonetheless, for both code-free deep learning and bespoke deep learning, image classification models performance was superior to video classification in all cases [114].

Tabular data

Tabular data, also known as structured data, is based on a tabular format of columns and rows. The application of deep learning has huge potential benefits considering the widespread adoption of electronic healthcare records. At present, there have been limited applications of code-free deep learning to tabular data. In 2020, Antaki et al. used MATLAB to develop a predictive model for postoperative proliferative vitreoretinopathy using clinical data from the electronic health record. The authors reported comparable results in the code-free deep learning model to a custom-built model trained on the same dataset. Nonetheless, the authors acknowledge that data scientist input was required for dataset preparation prior to training the code-free deep learning model [134]. In 2022, Abbas et al. published the results of a code-free deep learning model trained to predict visual acuity outcomes in patients with neovascular age-related macular degeneration using Google AutoML Tables. The model showed high specificity and had results comparable to a bespoke deep learning model trained for the same task [135].

Object detection

Wang et al. developed a code-free deep learning object detection model on Google AutoML for thyroid nodule identification and risk stratification. The model achieved an outstanding performance of 100% in object detection of nodules. Malignancy risk stratification was slightly poorer with a sensitivity of 73.9% and specificity of 70.8%. Nonetheless, when the artificial intelligence model was incorporated into the workflow as a clinical support tool, radiologists performance improved with sensitivity increasing from 52.1% and specificity increasing from 65.2% to 53.6% and 83.3%, respectively [136]. Unadkat et al. (2022) applied Google AutoML Vision object detection to the detection of surgical instruments using a dataset of 31,443 images acquired from endoscopic endonasal surgical videos. This model was subsequently compared to two bespoke object detection models trained using the same dataset. The authors reported mean average precision for the code-free deep learning system to be 0.708 while the bespoke models achieved average precisions of 0.669 and 0.527 [137]. Furthermore, as previously discussed in detail, Santomartino et al. explored the feasibility of object detection across a variety of commercial code-free deep learning platforms. Despite availability on Amazon, Apple, Google, and Microsoft, the authors only successfully created an object detection model using Google Cloud AutoML with poor performance [111].

1.4 Deep learning and papilloedema

In 2020, Milea et al. published the results of a deep learning model developed to identify optic nerve abnormalities from colour fundus photographs. The authors curated their dataset from retrospectively collected digital colour fundus photographs obtaining 14,342 photographs for training and validation and a subsequent 1505 photographs for external validation. The training dataset was collected from 19 sites spanning 11 countries and the external validation dataset came from 5 other centres in 5 countries. All eyes had been pharmacologically dilated prior to photographs being taken. The photographs were labelled by neuro-ophthalmologists who incorporated the clinical evaluation and ancillary testing of the patient at time of imaging in order to maintain high levels of accuracy. An image could only be labelled as papilloedema if there was confirmed intracranial hypertension. Other disc abnormalities included optic neuropathies, optic disc drusen, optic atrophy and congenital abnormalities. Patients from which the normal optic disc photographs came from were also thoroughly assessed by ophthalmologists and excluded if they had any coexisting ocular conditions such as retinal disorders or glaucoma. The authors demonstrated impressive results. The model trained to distinguish between normal and abnormal optic discs achieved an AUC of 0.99 in the validation dataset. Similarly, the validation dataset model demonstrated an AUC of 0.99 in the prediction of the presence or absence of papilloedema. External testing further established the robust performance of the deep learning models with an AUC of 0.96 to 0.99 for the normal compared to abnormal optic disc model and an AUC of 0.93 to 0.98 for distinguishing optic discs with papilloedema from optic discs without papilloedema. The authors reported that their model achieved a sensitivity of 96.4% and specificity of 84.7% for the detection of papilloedema. Despite these results, the authors describe some limitations to their work including the challenges of accurately labelled images, the fact that the images were taken after pharmacologically dilation of pupils which is not necessarily reflective of general practice, and the class imbalance among groups such as lesser representation of some optic disc abnormalities [3].

In addition, it must be acknowledged that considerable resources were required in order to develop this deep learning model. Dataset curation spanned a number of countries and years and model design and development was dependent upon highly specialised data science experts. Herein, this thesis aims to explore the application of code-free deep learning to develop image classification models to distinguish between normal optic discs, papilloedema and other disc abnormalities.

Chapter 2: Methods

The methodology for this research comprises three parts:

1. Identification and collection of two distinct datasets
2. Systematic review of the literature to identify any previous publications describing deep learning and optic disc abnormalities
3. Code-free deep learning model development using both datasets
4. External validation of the models

2.1 Ethics and governance

This research involves human participants and received research and development approval from the Institutional Review Board and the United Kingdom Health Research Authority (20/HRA/2158) “Advanced Statistical Modeling of Multimodal Data of Genetic and Acquired Retinal Disease” Integrated Research Application System project ID: 281957. In order to adhere to the data governance regulations, there was no active participation of human subjects. Only de-identified data was used which was collected retrospectively. The de-identifications were validated by Moorfields Eye Hospital NHS Foundation Trust. Therefore, informed consent was exempted. This study was conducted in accordance with the principles of the Declaration of Helsinki.

2.2 Dataset and participants

Two datasets are utilised in this study:

1. Dataset I: a retrospective dataset curated from Moorfields Eye Hospital NHS Foundation Trust
2. Dataset II: a publicly-available dataset obtained online which is freely available to download

2.2.1 Dataset I: Moorfields Eye Hospital

Dataset I was curated from retrospective data from Moorfields Eye Hospital NHS Foundation Trust. This required identification of appropriate sources within the Electronic Healthcare Record such as data from Neuro-ophthalmology clinics and Accident and Emergency attendances followed by rigorous and methodological data collection. The dataset was then de-identified in accordance with data governance approvals before being uploaded to the Moorfields Eye Hospital Cloud to allow for model development.

2.2.1.1 Participants and meta-data

Data was collectively retrospectively from patients attending Moorfields Eye Hospital NHS Foundation Trust between the 2nd April 2008 and the 8th August 2022. Moorfields Eye Hospital NHS Foundation Trust incorporates a primary central site with an additional nine district sites within London in the United Kingdom. Patients were included if they had attended an appointment and underwent optical coherence tomography and/or colour fundus photography imaging within the time period specified. The following meta-data was collected at time of curation: patient identification number, letter date, date of imaging examination, fixation, hardware manufacturer, hardware code name, image ID, laterality, protocol and time. Specific gender, ethnicity and socioeconomic deprivation data were not collected.

2.2.1.2 Dataset development

2.2.1.2.1 Labelling protocol

A labelling protocol was formulated by clinicians with varying experience in ophthalmology, ranging from one to six years, to assist with the application of labels to colour fundus photographs. This was based on the Modified Frisén Scale for Papilloedema and outlined clear features for each grade of papilloedema. This grading tool describes five different grades. Each grade is described in detail in accordance with the clinical features ranging from Grade 0 which describes *normal optic disc* to Grade V describing *severe papilloedema* with complete loss of normal optic disc landmarks. Grade I indicates *very early papilloedema* showing ‘c shaped’ disc oedema with sparing of the temporal margins. Grade II shows blurring and oedema of the entire disc margin and correlates with *early papilloedema.* Grade III indicates *moderate papilloedema*

showing inferior vessel obscuration while all major vessels become obscured by oedema in Grade IV *marked papilloedema* [138]. Ultimately, despite detailed description within the labelling protocol, images were labelled in a binary manner describing either the presence or absence of papilloedema regardless of the Grade applied.

Other optic disc abnormalities were also defined within the labelling protocol including non-glaucomatous optic atrophy of any cause, optic disc oedema secondary to acute anterior optic neuropathies, optic nerve head drusen and congenitally anomalous optic nerves. The labelling protocol specified that normal optic discs required a vertical cup/disc ratio of <0.5; a neuroretinal rim that was pinkish in appearance with equal thickness throughout or thickest inferiorly and thinning temporally; did not feature any sharp displacement of the retinal vessels at the point of transverse across the disc surface; and an intact nerve fibre layer. Of note, disc size and refractive error were not taken into account which is a limitation to the data. The labelling protocol also stipulated that any image labelled as papilloedema required confirmed evidence of raised intracranial pressure either in the form of lumbar puncture with an opening cerebrospinal fluid pressure of >250mm or positive neuroimaging.

2.2.1.2.2 Dataset curation

A dataset was curated retrospectively from the electronic healthcare record, OpenEyes, of Moorfields Eye Hospital NHS Foundation Trust. A record search of all patient letters within OpenEyes (ToukanLabs, London, UK) was performed utilising the incorporated tool, Magic xpa (Magic Technologies, California, USA). Magic xpa Application Platform is a low code application integrated into the Electronic Healthcare Record containing a number of features, primarily a search function which was used in this research to identify appropriate patient records. A record search was performed using the following keywords: ‘papilloedema,’ or ‘papilledema’ and ‘OCT.’ The letters yielded including those search terms were then manually reviewed to assess for the presence or absence of papilloedema and to identify presence of confirmed raised intracranial pressure. Following manual filtration of the results, the images were assigned the following labels: papilloedema; pseudopapilloedema; normal optic disc; other disc abnormality; suspected papilloedema; and previous papilloedema.

1. Papilloedema

Images could only be labelled as ‘Papilloedema’ if they met the clinical description of swollen optic disc appearance as described in the labelling protocol and if there was documented evidence confirming raised intracranial pressure at time of imaging. Raised intracranial pressure could be confirmed either by positive neuroimaging or by raised opening pressure (>250mHg) on lumbar puncture examination. As discussed, the diagnosis of ‘Papilloedema’ was initially taken from within the clinical letters, determined by clinicians with varying experience and expertise. However, as the label of ‘Papilloedema’ was strictly only applied with confirmed evidence of raised intracranial pressure, further adjudication on the label was not performed and the expertise level of the clinician applying the diagnosis in the first place was not recorded.

2. Pseudopapilloedema

‘Pseudopapilloedema’ was applied to an optic disc featuring any of the common associations as described in the Introduction. This included optic disc drusen and congenital abnormalities. The images within this cohort were assessed over multiple visits to ensure no evolution of symptoms and to confirm there was no documented clinical symptoms or evidence of raised intracranial pressure.

3. Normal optic disc

Optic discs were deemed ‘Normal’ if there were no abnormal clinical features and this remained stable over multiple consecutive visits.

4. Other disc abnormality

‘Other Disc Abnormality’ included the following pathologies: vitreopapillary traction, orbital apex congestion, optic atrophy, glaucomatous disc changes, inherited maculopathies, neuroretinitis, congenital fibrosis and drug-induced optic neuropathy. These other disc abnormalities were based on the documented diagnosis as labelled within the clinical letters. Closer analysis of the specific abnormalities, e.g., specific glaucomatous optic disc changes, were considered outside the scope of feasibility for this study.

5. Suspected papilloedema

The “Suspected Papilloedema” label was applied if patients had no confirmed evidence of raised intracranial pressure.

6. Previous papilloedema

“Previous Papilloedema" was applied to those that had a history of confirmed papilloedema but confirmed normal intracranial pressure at time of imaging, despite the optic disc appearance.

The labels used within this study described optic disc appearance as ‘Papilloedema,’ ‘Pseudopapilloedema,’ or ‘Normal.’ ‘Suspected Papilloedema’ and ‘Previous Papilloedema’ images were excluded from analysis as they were not in keeping with the clinical use case. As there is less propensity for diagnostic confusion between the described conditions and papilloedema, the ‘Other Disc Abnormality’ was also subsequently excluded from the final dataset.

A targeted database search was used to identify all patients listed within the Electronic Healthcare Record who had undergone disc-centred optical coherence tomography when attending the eye casualty. This search was cross-referenced against the MagicXPA search to ensure all identifiable cases of papilloedema had been included. As patients with papilloedema often had multiple visits with imaging from various different dates, only the first attendance from each was included in the dataset to avoid data leakage and to ensure images were not included labelled as papilloedema after the intracranial pressure had normalised with management.

Optical coherence tomography images were extracted from the 2nd April 2008 until the 23rd May 2022. Colour fundus photographs were collected from the 2nd April 2008 until the 8th August 2022. Optical coherence tomography images were only included if they were disc-centred. Any macula-centred optical coherence tomography images were excluded from the dataset, even if the optic disc was visible. Colour fundus photograph inclusion criteria specified disk centred, optic nerve head or peripapillary retina fixation. The identified images were then filtered to include only rectangular volume and disc centred ART volume optical coherence tomography images.

The data collected came from two distinct hardware manufacturers: TopCon (Topcon, Inc.) and Heidelberg (Heidelberg Engineering, Heidelberg, Germancy) in the following models: 3DOCT-1000MK2, 3DOCT-2000, 3DOCT-2000SA, FD-OCT, Spectralis HRA + OCT, Spectralis OCT, Triton and Triton plus. Although analysis of the image hardware was not considered pertinent to this research, the decision was made to document the source of each image at time of data collection to allow for future work including analysis of model performance between different hardwares. The images were collected in Digital Imaging and Communications in Medicine (DICOM) format.

2.2.1.2.3 Dataset pre-processing and preparation

Pre-processing

Upload to cloud

Following curation of the dataset, the spreadsheet containing patient identification numbers with image identification numbers was exported to a third party company (Softwire, London, United Kingdom) to be fully anonymised in order to comply with data governance approvals. Clinical metadata was not exported. Following anonymisation, the dataset was uploaded by the third party company to a storage bucket within the Moorfields Eye Hospital Google Cloud Platform [139] to an account connected to the local institution (Moorfields Eye Hospital) from where the dataset was curated. Security features were imposed to disallow any download or sharing of the dataset without explicit permissions. A spreadsheet containing the anonymised image title with corresponding label in numerical format was also provided:

- -1 assigned to Accident and Emergency “Normal” images
- 0 assigned to “Normal” Neuro-ophthalmology clinic images
- 1 assigned to “Papilloedema” images
- 2 assigned to “Pseudopapilloedema” images
- 3 assigned to “Other abnormalities”

Image processing

In order to meet Google Cloud AutoML Vertex AI compatibility requirements, the dataset was processed from DICOM to Portable Network Graphics (PNG) format requiring data scientist

support. The dataset was subsequently uploaded to a second separate Google Cloud storage bucket. For each image modality, datasets were provided in a variety of forms to ascertain the role image form plays in model performance.

Colour fundus photographs were uploaded in two forms:

1. High resolution images
2. 512x512 square resized images

Optical coherence tomography images were uploaded as

1. Transverse coronal scans (enface) images: Enface images were formatted and provided in both square resized 224x224 images and square resized 512x512 images
2. Middle b scan squared images: Middle b scan images were provided in the following formats:
    a. full resolution
    b. resized 224x224
    c. resized 512x512
    d. square full resolution
    e. square resized 224x224
    f. square resized 512x512.

Dataset selection was made with the aim of maximising training quality. Both colour fundus photograph datasets were selected for use. For optical coherence tomography model development, the largest size of 512x512 was selected for both enface and middle b scan optical coherence tomography images. For consistency, the square format for each was selected to allow for comparison.

Preparation

Using the spreadsheet generated following anonymisation of the images, comma separated values files were prepared for each model to be trained using Microsoft Excel on the local computer. Comma separated values files are a text file format whereby text data is organised in a string separated by commas. For use within Google Cloud AutoML Vertex AI, comma separated values files can be formatted with either two columns or three columns. In the case of

three columns, the third column would designate the dataset split i.e., the proportion of images each designated to the training, validation and test sets. As Google Cloud AutoML Vertex AI also can perform automatic data split allocating 80% of images to the training set, 10% of images to the validation set and 10% of images to the test set, the comma separated values files were formatted to contain two columns. Column A specified the image name file path within the Google Cloud Storage Bucket and column B specified the image label.

2.2.2 Dataset II: Publicly-available dataset

The repository, Kaggle.com, was used to identify the publicly-available dataset used in this study. Kaggle.com is an online platform whereby datasets can be published and shared in order to facilitate collaboration and research. A dataset of 1369 disc-centred colour fundus photographs entitled “Identification of Pseudopapilloedema” was identified consisting of 779 images labelled “Normal,” 295 images labelled “Papilloedema,” and 295 images labelled “Pseudopapilloedema.”

2.2.2.1 Participants and meta-data

The dataset ‘Identification of Pseudopapilloedema’ was downloaded from the Center for Open Science [140]. It was created on the 1st August 2018. There was no available metadata provided with the dataset. Furthermore, there was no available information on the labelling protocol used or how papilloedema was confirmed in those images containing that label. Therefore, it was not possible to determine the level of experience and expertise of the clinician finalising the diagnosis. This would clearly impact the quality of dataset labelling and is a limitation to the use of this dataset.

2.2.2.2 Dataset preparation and processing

The dataset was downloaded into folders on the local computer into files entitled “Normal,” “Pseudopapilloedema” and “Papilloedema.” It was subsequently uploaded onto a new Moorfields Google Cloud storage bucket entitled Kaggle using the Terminal application on MacOS. This required some basic coding ability but was completed using Google Cloud Documentation [141] and with no data scientist input.

The downloaded files from the dataset were contained in separate folders for each label. The file names were exported into a spreadsheet with the appropriate corresponding label and the three folders were subsequently merged. Python 3 (Python 3.12.4) was downloaded onto the local computer. Following this, Google Cloud Command Line Interface was installed by downloading the appropriate macOS package, macOS 64-bit (ARM64, Apple M1 silicon), and extracting the file to the Home directory on the local computer. The provided script from within the Google Cloud Documentation entitled “Install the gcloud CLI” was run using Terminal to both install and initialise the product [141]. Google Cloud Platform account credentials were then used to link the local computer to the specific Moorfields Eye Hospital, Google Cloud Platform account and the specific project from within that account was selected. Google Cloud Platform allows datasets to be uploaded directly from the local computer using the Graphical User Interface, however, this is limited to small datasets and is time consuming. For this reason, basic coding was used to upload the downloaded Kaggle dataset to the Moorfields Eye Hospital Google Cloud Platform account. Similarly, this did not require data scientist support. A new storage bucket was created within the project on the Cloud entitled “Kaggle” and a parallel, multi-threaded command was utilised within the Terminal application on the local computer to upload the downloaded “Identification of Pseudopapilloedema” dataset to the new cloud storage bucket [142]. All of the above was performed by a clinician with no coding expertise with the assistance of Google Cloud Platform documentation as referenced.

As with dataset preparation performed for the Moorfields Eye Hospital dataset, comma separated values files were then formatted to contain only two columns: column A to specify the image name file path within the Google Cloud Storage Bucket and column B specifying the image label. Dataset split was performed automatically by Google Cloud AutoML Vertex AI at time of dataset upload to each model.

2.3 Systematic review of the clinical literature

Following curation of the datasets, a comprehensive review of the literature was conducted in order to identify any deep learning models trained to classify optic disc abnormalities.

2.3.1 Search strategy

A search was performed using combinations of the following terms: ‘papilloedema,’ ‘papilledoema,’ ‘pseudopapilloedema,’ ‘pseudopapilloedema,’ and ‘deep learning’. The primary search strategy was created in MEDLINE using medical subject headings (MeSH in PubMed), natural language and abbreviations of relevant terms and adapted for use in the other search platforms. Boolean operators (AND, OR, NOT) were also employed to refine the search results. There were no language or restrictions applied. Publication date was restricted to literature published after 2012 to incorporate systems utilising current approaches deep learning as opposed to those in practice prior to the launch of AlexNet [94].

Inclusion criteria included literature describing any form of deep learning system with a focus on optic disc abnormalities, specifically papilloedema or pseudopapilloedema. Exclusion criteria included articles that did not utilise artificial intelligence, literature that examined optic disc abnormalities not relevant to the use case of this paper including glaucomatous discs or abnormalities secondary to retinal pathology, descriptive publications including systematic or literature reviews, correspondence and editorials and articles not in full text. Full inclusion and exclusion criteria can be visualised in Table 1.

The searches were carried out using the Cochrane Library (CENTRAL), Embase (Emtree), PubMed (MEDLINE), and Scopus for publications meeting the inclusion criteria. Any eligible publications were manually reviewed to assess for relevant literature within the references. A search of the grey literature was also performed examining the following for relevant publications: NLM National Information Center on Health Services Research and Health Care Technology (NICHSR) including both Health Services Research Projects and Health Services/Sciences Search Resources (HSRR); OpenDOAR; Bielefeld Base; Lenus; arxiv (ARXIV) and RIAN.

Table 1. Inclusion and Exclusion Criteria	
Inclusion	**Exclusion**
Deep learning system (any approach)	No use of artificial intelligence
Publication date later than 2012	Publications prior to 2012
Optic disc abnormalities: papilloedema pseudopapilloedema	Optic disc abnormalities secondary to retinal or glaucomatous pathology
Binary classification outcome measures	Assessment of optic disc structure
Detailed information provided on dataset	Non-binary assessment of optic disc abnormalities such as degree of swelling or severity
Original research	Descriptive articles, correspondence, editorials, full text not available

2.3.2 Data extraction

- EndNote 20.0(C) was utilised as the reference manager for the purposes of this literature review. A new EndNote library was created to allow filing of references via the Groups feature. The search findings from each database were downloaded as Research Information Systems (RIS) files and opened in the EndNote library. The “Check duplicates” tool and manual review allowed removal of duplicate papers. The titles and abstracts of all remaining publications were then examined. The EndNote library was used to organise publications into the following groups: ‘Inclusion’; ‘Exclusion - non-AI’; and ‘Exclusion - incorrect topic.’ A full text review was then performed on all publications within the ‘Inclusion’ group to confirm relevance and assess against inclusion and exclusion criteria. A standardised data extraction form was formatted using Microsoft Excel and key information from each publication was collected. This included author(s), publication year, title, dataset size, dataset modality, main outcome measure, dataset labels and artificial intelligence model(s) performance. The selected data for extraction can be reviewed in Table 2.

Table 2. Standardised data extraction form for systematic review	
	Data for extraction
1	Author(s) and year
2	Title of record
3	Primary outcome measure(s)
4	Dataset modality
5	Dataset size
6	Dataset labels
7	Deep learning model(s) performance

2.3.3 Data synthesis

As this literature review was performed for the goal of model benchmarking, meta-analysis of the data was not performed. A structured narrative synthesis approach was used as the method of data synthesis in order to systematically review and examine the findings. A detailed summary of each publication was recorded as outlined in the information collected during data extraction. Following this, each article was categorised into those utilising colour fundus photographs, those utilising optical coherence tomography images, or articles using both.

2.4 Model development

All binary models were trained using Google Cloud Platform AutoML Vertex AI [143] by a clinician with no coding expertise. The models were cloud-based and trained to perform binary image classification. The single-label classification for each imaging modality was based on two clinical use cases:

1. The presence or absence of papilloedema
2. Presence or absence of any optic disc abnormality, i.e. normal versus abnormal optic discs.

AutoML Vertex AI automatically performs dataset split allocating 80% of images for training, 10% validation and 10% testing. In order to avoid overfitting, early stopping was automatically applied when training each model. Overfitting occurs when a deep learning model becomes too aligned to its training data, i.e., it can make excellent predictions on the dataset it was trained on but then cannot generalise and make accurate predictions when presented with a new dataset. It can happen for a variety of reasons including a dataset that is too small without enough representative samples, or training a model for too long on a relatively small dataset [144].

A series of code-free deep learning models were developed using four different image formats from within the curated dataset: high resolution colour fundus photographs, square re-sized 512x512 colour fundus photographs, enface optical coherence tomography images and middle b scan square optical coherence tomography images. A summary of the code-free deep learning model development can be visualised in Figure 2. An example of code-free model development using Google Cloud AutoML process can be visualised in Figures 3-6.

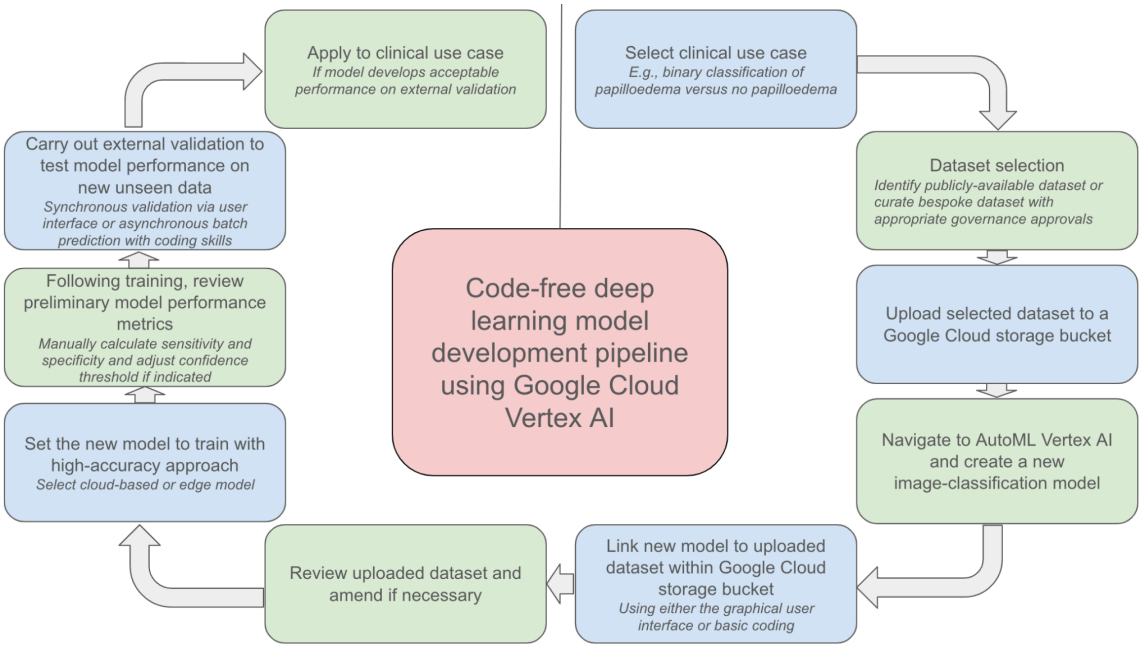

Figure 2. The code-free deep learning model development pipeline for Google Cloud Platform AutoML Vertex AI

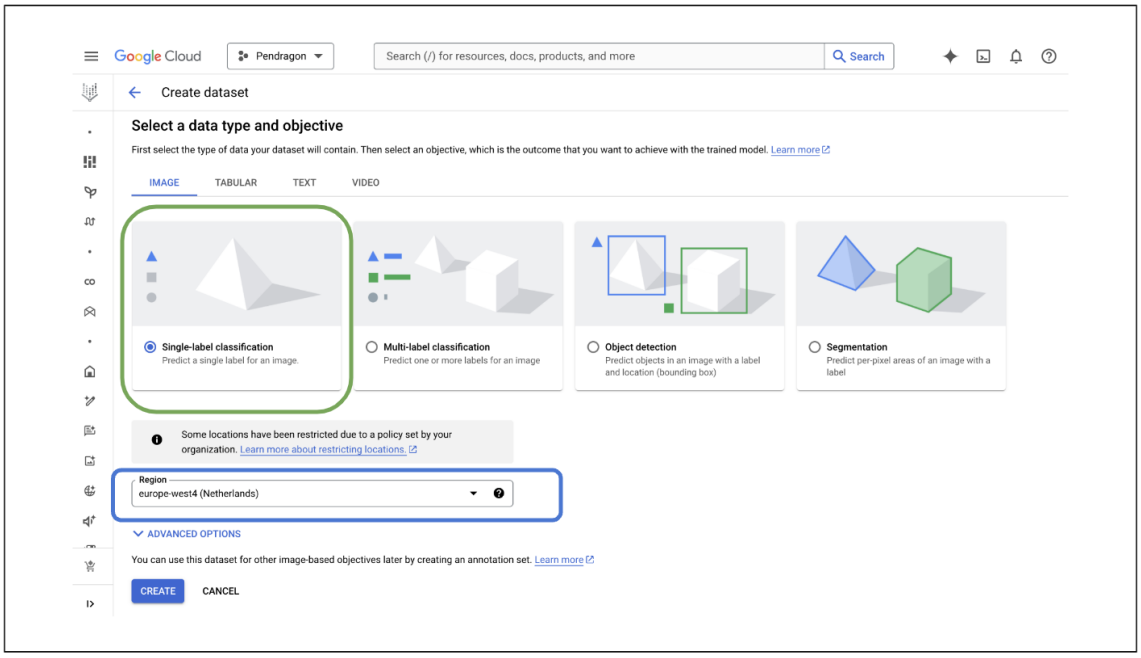

Figure 3. Step one: Creating a code-free deep learning model on Google Cloud AutoML Vertex AI first requires selection of dataset type (see green box for single-label classification utilised in this research) followed by selection of location where the model will be stored (blue box for europe west4). (Source: Google Cloud Platform).

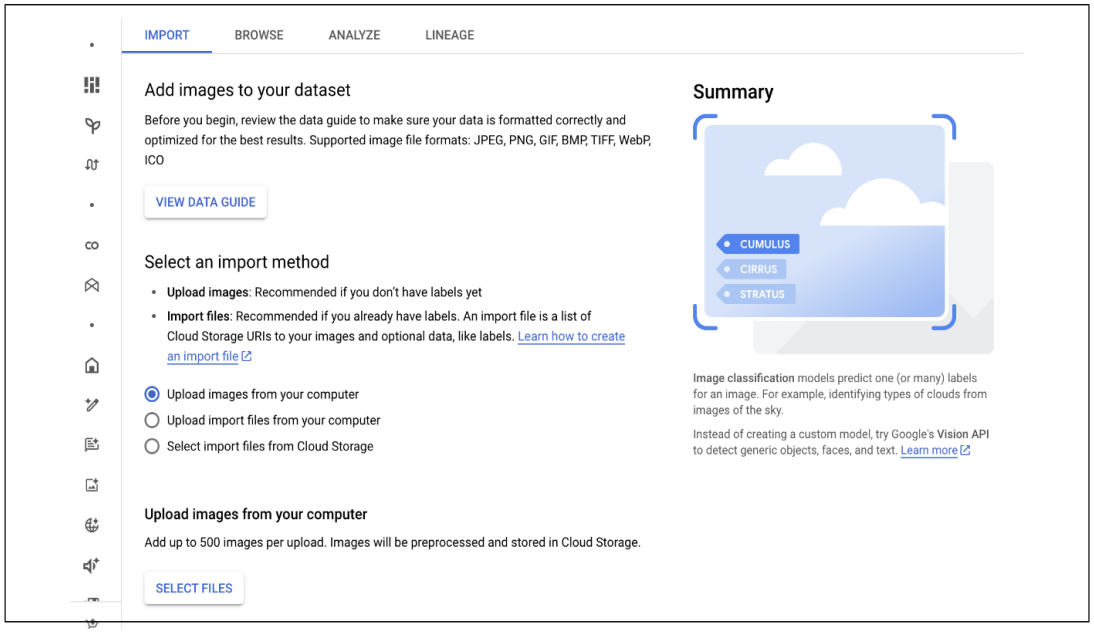

Figure 4. Step two: Upload images to the code-free deep learning model. This can be done either by direct upload from the computer or by selecting a dataset pre-uploaded onto Google Storage (Source: Google Cloud Platform).

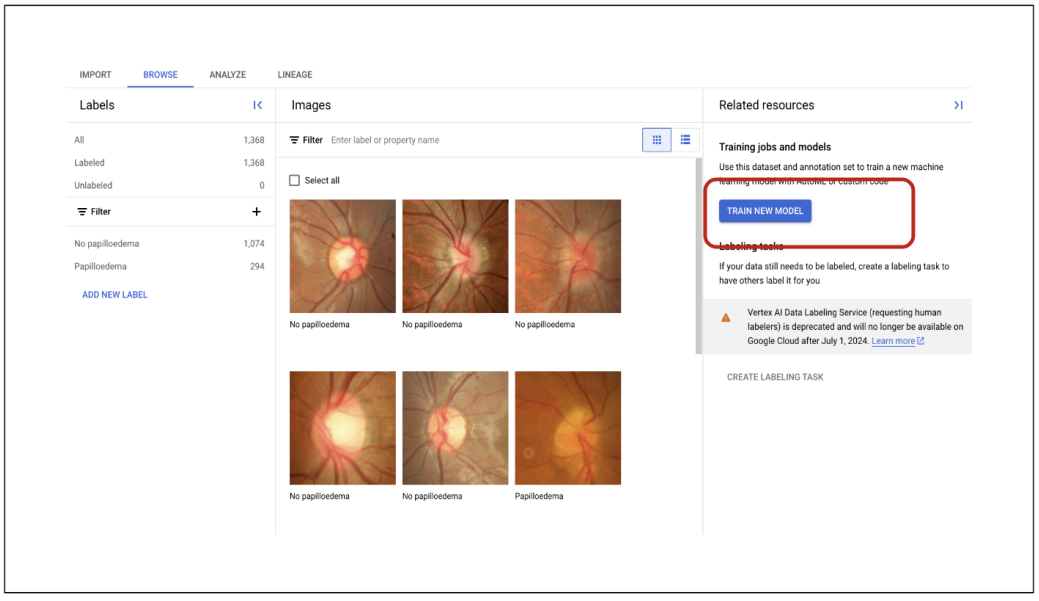

Figure 5. Step three: The dataset and the number of images per label can be visualised following upload to ensure there is an acceptable balance between labels. Following this, 'train new model' can be selected and the appropriate settings can be chosen (Source: Google Cloud Platform).

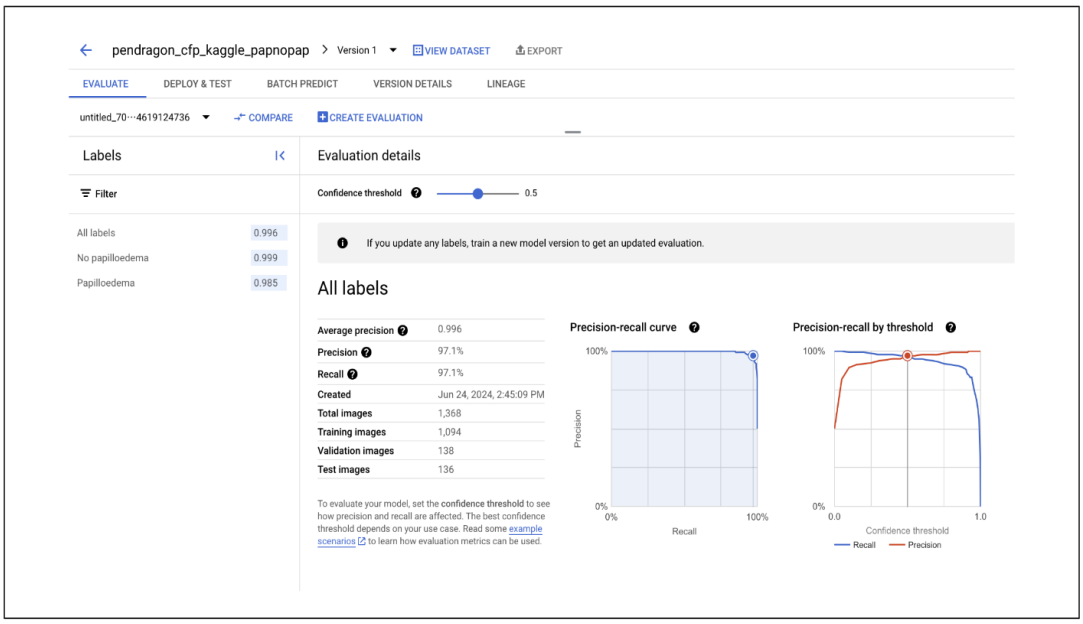

Figure 6. Step four: Following training of the code-free deep learning model, the preliminary results can be visualised in the Evaluation tab. The performance measures are provided as well as precision-recall curves and precision-recall by threshold. The confidence threshold can also be adjusted at this point (Source: Google Cloud Platform).

2.4.1 Dataset I: Moorfields Eye Hospital model training

A comma separated values file was uploaded to the Google Cloud storage bucket containing labels ‘papilloedema’ or ‘no papilloedema’ and linking the appropriate file path and label to its corresponding image within the storage bucket. Google Cloud AutoML Vertex AI was then used for development of the code-free deep learning models.

Code-free deep learning models were developed within Google Cloud AutoML Vertex AI by selecting the Create dataset using the graphical user interface.
Data type and objective were then chosen. The selected data was image and the objective was text classification (single-label) described to predict a single label for an image. Google Cloud AutoML Vertex AI is currently only compatible within two regions: us-central1 (Iowa) and europe-west4 (Netherlands). In order to remain as locally as possible (United Kingdom), europe-west4 (Netherlands) was selected as the geographical region whereby the data would be handled and stored.

Herein, the development of each code-free deep learning model using Dataset I will be described based on the file format used for training. The models will be described in accordance to their clinical use case with those trained to classify between ‘Papilloedema versus no papilloedema’ discussed first followed by the models trained in the classification of ‘Normal versus abnormal optic discs.’

2.4.1.1 Papilloedema versus no papilloedema

Model one: High resolution colour fundus photographs for the classification of papilloedema versus no papilloedema

Model one was created to distinguish between papilloedema or no papilloedema using high resolution colour fundus photographs using the Moorfields Eye Hospital dataset. As described, the code-free deep learning model was trained within Datasets of Google Cloud AutoML Vertex AI. A comma separated values file was formatted specifically to contain file path names linking the labels of ‘papilloedema’ or ‘no papilloedema’ to the Google Cloud storage bucket of high resolution colour fundus photographs from the curated dataset. The comma separated values file was uploaded to the Google Cloud storage bucket. Following creation of the image-classification dataset within Google Cloud AutoML Vertex AI, images were added by linking the dataset to the formatted comma separated values file within the Google Cloud storage bucket. This was all performed using the Graphical User Interface. Once the images had been uploaded to the dataset, labels could be reviewed and altered if necessary. ‘Train new model’ was then selected via the Graphical User Interface. The model training method selected was AutoML and the model was chosen to be cloud-based in order to facilitate deployment for online predictions. Higher accuracy was selected with an estimated latency of 200ms-300ms. Maximum node hours were selected based on the recommended minimum requirement as provided by Google Cloud AutoML Vertex AI. Model training was then initiated and email notification was enabled to alert once model training was complete.

Model two: Square re-sized 512x512 colour fundus photographs for the classification of papilloedema versus no papilloedema

Model two was developed using the square re-sized 512x512 colour fundus photographs to distinguish between papilloedema or no papilloedema. A comma separated values file was

formatted to contain file paths linking to the square resized 512x512 colour fundus photographs within the Google Cloud storage bucket specifying the labels ‘papilloedema’ or ‘no papilloedema.’ Following upload of the comma separated values file to the bucket, an image-classification dataset was selected using Google Cloud AutoML Vertex AI and the formatted comma separated values file was imported. The new model was trained using the AutoML method with higher accuracy and a cloud-based approach. Maximum node hours were selected based on the recommended minimum time. Model training was commenced with notifications enabled to alert on training completion.

Model three: Optical coherence tomography enface 512x512 images for the classification of papilloedema versus no papilloedema

Model three was developed using optical coherence tomography enface 512x512 images to distinguish between the presence or absence of papilloedema. As described, a comma separated values file was formatted to link the labels of ‘papilloedema’ and ‘no papilloedema’ to the corresponding 512x512 enface optical coherence tomography images within the Google Cloud storage bucket. The model was developed and trained with Google Cloud AutoML Vertex AI using the same parameters described, namely AutoML as the selected approach using cloud-based computing and maximum latency. The minimum required nodes per hour were selected as the maximum based on the Google Cloud recommendation.

Model four: Optical coherence tomography middle b scan square 512x512 images for the classification of papilloedema versus no papilloedema

Model four utilised middle b scan square 512 optical coherence tomography scans to distinguish between papilloedema or no papilloedema. A specific comma separated values file was formatted linking the corresponding labels to the file path names within the Google Cloud storage bucket. A new dataset was created using AutoML Vertex AI and the specifically formatted comma separated values file was imported. Following review of the uploaded images, the model was trained using a cloud-based AutoML with higher accuracy and the minimum node requirements. The notification feature was selected to alert once training was completed.

2.4.1.1 Normal versus abnormal

Four code-free deep learning models were subsequently developed and trained to distinguish between normal optic discs versus abnormal optic discs.

Model five: High resolution colour fundus photographs for the classification of normal optic discs versus abnormal optic discs

Model five utilised high resolution colour fundus photographs for the distinction between normal or abnormal optic discs. A comma separated values file was created and firstly all papilloedema and other disc abnormalities were combined into a single category of ‘abnormal optic discs’ to allow for binary classification. This comma separated values file was then formatted to link the new corresponding labels to the file path names of the high resolution colour fundus photographs within the Google Cloud storage bucket. A new model was created on AutoML Vertex AI and the comma separated values file was imported with the images visualised to ensure there were no errors. The model was trained using an AutoML cloud-based approach with maximum latency and minimal node requirements. Notifications were enabled to alert on completion of training.

Model six: Square resized 512x512 colour fundus photographs for the classification of normal optic discs versus abnormal optic discs

Model six was trained using square resized 512x512 colour fundus photographs to classify images as normal optic discs or abnormal optic discs. The comma separated values file was formatted to link the labels of ‘normal optic disc’ and ‘abnormal optic disc’ to the corresponding file path name within the Google Cloud storage bucket. A new model was selected within AutoML Vertex AI and trained using maximum latency and an AutoML cloud-based approach. Minimal node requirements were selected based on the Google Cloud recommendation and notifications were enabled to alert on completion of training.

Model seven: Optical coherence tomography enface 512x512 images for the classification of normal optic discs versus abnormal optic discs

Model seven was an image classification model trained to identify normal versus abnormal optic discs using optical coherence tomography scans formatted to 512x512 enface files. The model was developed by formatting a new comma separated values file to link the labels of “normal optic disc” and “abnormal optic disc” to the corresponding file path name within the Google Cloud storage bucket. A new model was developed using AutoML Vertex AI with the selected dataset imported and the model was trained using an AutoML cloud-based approach with maximal latency and minimum nodes per hour.

Model eight: Optical coherence tomography middle b scan square 512x512 images for the classification of normal optic discs versus abnormal optic discs

Model eight utilised middle b scan square 512x512 optical coherence tomography scans to identify normal optic discs from abnormal optic discs. A new comma separated values file was formatted to link the described labels to the appropriate file path name within the Google Cloud storage bucket. A new code-free deep learning model was developed using AutoML Vertex AI with the optical coherence tomography middle b scan square 512x512 images imported and reviewed prior to training. The model was subsequently trained using AutoML with cloud-based computing, maximum accuracy and the minimum nodes required for training as recommended by Google Cloud.

2.4.2 Random Under Sampling

In order to address class imbalance, random undersampling was performed. Data imbalance occurs with highly skewed data when the number of labels from one class (majority class) is significantly higher than another (minority class). This is problematic as the deep learning model can then preferentially predict the majority class with the assumption that it is more likely to be the correct prediction rather than performing feature analysis of the image it is presented with [145]. For this reason, random undersampling was performed whereby the majority class is reduced down at a selected ratio to the minority class.

The publicly-available data mining software Waikato Environment for Knowledge Analysis software was downloaded from the University of Waikato in New Zealand website [146,147] The

relevant comma separated values files were re-formatted into attribute-relation file format (.arff) files using the TextEdit application on MacBook Pro to allow upload onto the Waikato Environment for Knowledge Analysis software. Random undersampling was performed for each file at a ratio of 2:1 for the majority class. The software allows for any ratio to be selected, however, the lower ratio of 2:1 was selected with the aim of maximally addressing any class imbalance issues within the dataset. This did not require any coding expertise and was carried out by the same clinician developing the models with some assistance from open-access online resources including YouTube [148]. Following completion of random undersampling on the Waikato Environment for Knowledge Analysis software, the new attribute-relation file format (.arff) files were downloaded, converted back into compatible comma separated values files and uploaded to the Cloud Storage Bucket.

2.4.2.1 Random undersampling: Papilloedema versus no papilloedema

Model nine: High resolution colour fundus photographs for the classification of papilloedema versus no papilloedema following random undersampling

Model nine was developed to distinguish between papilloedema versus no papilloedema using high resolution colour fundus photographs from the random undersampling dataset at a ratio of 2:1. A new comma separated values file was formatted to link the new dataset following random undersampling to the corresponding images within the Google Cloud storage bucket. A new code-free deep learning model was developed using AutoML Vertex AI following import and review of the smaller dataset of high resolution colour fundus photographs. A cloud-based model was selected for maximum accuracy with the minimum nodes required for training chosen.

Model ten: Optical coherence tomography enface 512x512 images for the classification of papilloedema versus no papilloedema following random undersampling

Model ten was developed to distinguish papilloedema versus no papilloedema using enface optical coherence tomography images from the random undersampling dataset. A new comma separated values file was formatted to link the randomly undersampled dataset to the corresponding optical coherence tomography images in the Google Cloud storage bucket. A new model was selected and the imported dataset was reviewed. The model was then trained using cloud-based computing with maximum accuracy and the minimum required nodes.

Model eleven: Optical coherence tomography middle b scan square images for the classification of papilloedema versus no papilloedema following random undersampling

Model eleven utilised middle b scan square optical coherence tomography images following random undersampling to distinguish between papilloedema versus no papilloedema. The new comma separated values file was formatted to link the middle b scan square optical coherence tomography images to the corresponding dataset within the Google Cloud storage bucket and a new model was trained using cloud-based computing with maximum accuracy and the minimum recommended nodes.

2.4.2.2 Random undersampling: Normal versus abnormal

Model twelve: High resolution colour fundus photographs for the classification of normal optic discs versus abnormal optic discs following random undersampling

Model twelve utilised high resolution colour fundus photographs following random undersampling to identify normal versus abnormal optic discs. A new comma separated values file linking the randomly undersampled dataset with the corresponding labels of normal or abnormal optic disc to the correct images within the Google Cloud storage bucket. A new model was trained with cloud-computing. The minimum recommended nodes were used and the maximum accuracy.

Model thirteen: Optical coherence tomography enface 512x512 images for the classification of normal optic discs versus abnormal optic discs following random undersampling

Model thirteen utilised enface optical coherence tomography scans to distinguish between normal and abnormal optic discs following random undersampling. A new comma separated values file was created to link the optical coherence tomography images following random undersampling to the appropriate labels and corresponding images within the storage bucket. A new AutoML model was selected and following upload of the dataset, it was trained with highest accuracy and cloud-based computing. The minimum node requirements were selected as recommended by the Google Cloud Platform.

Model fourteen: Optical coherence tomography middle b scan square 512x512 images for the classification of normal optic discs versus abnormal optic discs following random undersampling

Model fourteen utilised mid b square optical coherence tomography scans to distinguish between normal and abnormal optic discs following random undersampling. A new comma separated values file was formatted for the optical coherence tomography images following random undersampling and linked with the corresponding label and image within the storage bucket. An AutoML cloud-based model was trained with maximum latency and minimum nodes per hour and notifications were enabled to alert on completion of training.

2.4.3 Public dataset

The publicly available dataset was uploaded to the Google Cloud storage bucket and two code-free deep learning models were subsequently developed.

Model fifteen: Publicly-available colour fundus photographs for the classification of papilloedema versus no papilloedema

Model fifteen was developed using the Kaggle dataset of colour fundus photographs to identify between papilloedema or no papilloedema. A comma separated values file was formatted linking the provided image label to the corresponding image uploaded to the Google Cloud storage bucket. An AutoML image classification model was selected for cloud-based computing with maximum accuracy and minimum nodes per hour.

Model sixteen: Publicly-available colour fundus photographs for the classification of normal optic discs versus abnormal optic discs

Model sixteen was developed using the Kaggle dataset of colour fundus photographs to identify between normal versus abnormal optic discs. A new comma separated values file was formatted to include the labels normal or abnormal optic disc and to link the file path to the appropriate image within the Google Cloud storage bucket. An AutoML model was selected and following import of the dataset, it was trained using cloud computing and maximum accuracy. The minimum recommended nodes per hour were selected and email notifications were enabled to alert on completion of training.

2.5 External Validation

External validation was considered an important final step in the development of code-free deep learning models for the purpose of this thesis. External validation describes the process of testing the performance of a model when provided with new different images to those used in the training, validation and test datasets. External validation is an important step to test the reliability and generalisability of a deep learning model. For example, a model may demonstrate very strong performances on the dataset it is trained on but perform poorly if presented with data taken from a different patient population, imaging system or protocol. It is particularly important in clinical artificial intelligence that a deep learning model is confirmed to be robust when presented with a wide variety of datasets for its given task in order to avoid bias and poor accuracy if deployed to direct clinical care [149].

Based on poor preliminary results, the decision was made not to perform external validation on all code-free deep learning models. The results were defined as poor as the performance was considerably worse than the bespoke deep learning models reported on in the literature. This applied to models one to eight, ten, eleven, thirteen and fourteen.

2.5.1 Dataset I: External Validation

External validation was performed on models nine and twelve using a selection of images from Dataset II. These models had been trained using Dataset I meaning the images selected from Dataset II were previously unseen to the model. As discussed, the metadata on how the labels were applied to Dataset II was not available which is a limitation to the study. As each image was manually uploaded, a smaller number was selected than would be optimal. This emphasises the need for batch prediction which externally validates the model on large quantities of unfamiliar data and will be discussed in further detail later.

Firstly, the models were deployed within the Google Cloud Vertex AI Model Registry tab. Coding expertise was not required. ‘Deploy and test’ was selected using the graphical user interface. This allows the user to create and name an endpoint whereby the model can be deployed selecting the location and number of nodes to be used. Once the model was active, it could be tested by directly uploading images via the user interface and a prediction was provided

demonstrating the model confidence at each given prediction. External validation was performed on models nine and twelve using fifteen images selected from Dataset II including five normal optic disc colour fundus photographs, five pseudopapilloedema colour fundus photographs and five pseudopapilloedema colour fundus photographs.

2.5.2 Dataset II: External Validation

For data governance reasons, Dataset I could not be downloaded to the local computer. Therefore, alternative images were identified for the external validation of models trained using Dataset II. Models fifteen and sixteen (trained using Dataset II) were deployed within the Google Cloud Vertex AI Model Registry tab using the graphical user interface. Ten colour fundus photographs were identified from a Google Images search. The source of each image was identified and its label was confirmed. The external validation dataset curated for Dataset II included three normal optic disc colour fundus photographs, four pseudopapilloedema colour fundus photographs and three colour fundus photographs of papilloedema.

2.5.3 Batch prediction

It is acknowledged that batch prediction is an essential step for the robust evaluation of deep learning models. Batch predictions allow for large datasets to be uploaded to the deep learning model in order to comprehensively examine how the model performs on previously unseen data. The data is processed offline asynchronously and predictions are returned once the external validation is complete [150]. However, coding expertise is required and a significant amount of additional data scientist support would have been necessary. This is beyond the scope of this research thesis which aims to investigate the development of code-free deep learning models without the need for coding expertise.

2.6 Statistical analysis

Performance for each model was statistically evaluated using a series of metrics provided by Google Cloud AutoML Vision calculated at a confidence threshold of 0.5. Threshold was additionally lowered to 0.25 for each model with the corresponding performance metrics evaluated. The lower threshold was selected in accordance with the potential use case for this research as a screening tool.

Area under the precision recall curve (AUPRC) was provided for each model at each threshold. This is a metric commonly used in machine learning for binary classification models to demonstrate a general summary of algorithm performance whereby a higher score indicates a better performance. Precision was calculated for each model at each threshold. This is a measure of the amount of true positive predictions among all positive predictions identified by the algorithm and is also known as the positive predictive value. Recall was provided for each model at each threshold. A term used widely in machine learning, recall, more commonly known as sensitivity or the true positive rate, describes the proportion of true positive predictions from all of the true positive predictions within the dataset [151].

Google Cloud AutoML Vertex AI also provides confusion matrices for each model with predicted code-free deep learning model labels cross-tabulated against the ground truth label. Utilising the performance metrics provided by Google Cloud AutoML Vertex AI, contingency tables were produced including true-positive, false-positive, true-negative, and false-negative results to calculate specificity for each model at each threshold. Sensitivity was calculated using the formula *[a/(a+c)] x 100* and specificity was calculated using the formulate *[d/(d+b)] x 100* whereby a = true positives, b = false positives, c = false negatives and d = true negatives.

Chapter 3: Results

The results will be described as follows:

1. Results of the systematic review of the literature
2. Dataset and participants
    a. Dataset I: Moorfields Eye Hospital
    b. Dataset II: Kaggle
3. Model development
4. External validation

3.1 Systematic review of the literature

3.1.1 Search strategy

The final search was updated on the 28th July 2024 of the following electronic databases: Cochrane CENTRAL, Embase (Emtree), PubMed (MEDLINE), Scopus and Web of Science using key words ‘deep learning,’ ‘papilloedema,’ and ‘pseudopapilloedema.’ This search yielded 329 records as demonstrated in Table 3. A search of the grey literature was also performed using Lenus, OpenDoar, Rian, Base search and Arxiv identifying five additional papers for review which can be visualised in Table 4.

The search results were downloaded in Research Information System (RIS) file format and uploaded onto EndNote 20.0 ©. The “Check Duplicates” EndNote 20.0 © tool was used to remove 11 duplicate records. Manual review yielded an additional 4 duplicate papers. This resulted in 319 records for title review. Each record was classified as relevant or irrelevant based on the title of the paper. Eighty four records were excluded as there was no use of artificial intelligence. A further 198 records were excluded as they did not specifically examine the use of deep learning for optic disc abnormalities, specifically papilloedema and/or pseudopapilloedema. This resulted in 33 records for abstract review. Two records were excluded as the full text was unavailable and another one record was excluded as it did not provide any pertinent information relating to details of the dataset or deep learning algorithm

used. Six records were excluded as the pathology assessed was retinal or glaucoma-focused rather than papilloedema and/or pseudopapilloedema. Two records were excluded as they were descriptive articles rather than original research. Eleven records were excluded as the goal of model performance was to look at structural abnormalities within the optic nerve (including optic disc laterality, image quality assessment, degree of disc swelling and degree of optic disc tilt) as opposed to binary classification of the specified abnormalities. Two records were excluded as they examined papilloedema severity but not binary classification. Two records were excluded as the aim of the study was to compare deep learning performance against clinician performance which was beyond the scope of this research. Following this process, there were 7 records remaining that met the full inclusion criteria. The Prisma flow diagram for the systematic review can be visualised in Figure 7.

Table 3. Database search strategy

Database	**Search strategy**
Cochrane CENTRAL (n=3)	(‘Deep Learning/exp’ [MeSH]’ OR ‘deep learning’) AND ((‘Papilledema/exp [MESH]’) OR (‘papilloedema’) OR (‘papilledema’) OR ‘(pseudopapilloedema’))
Embase (n=9)	(‘Deep AND ‘learning’/exp) AND (‘papilloedema’ OR ‘papilledema’ OR pseudopapilloedema’ OR ‘pseudopapilledema’)
PubMed (MEDLINE) (n=78)	(("Deep Learning"[MeSH]) AND "Papilledema"[MeSH]) OR "Pseudopapilledema" [Supplementary Concept]
Scopus (n=239)	TITLE-ABS-KEY (deep AND learning) AND (papilloedema OR papilledema OR pseudopapilloedema OR pseudopapilledema)

Table 4: Grey literature search strategy	
Grey literature	
Base search (n=2) - excluded not specific to papilloedema	“Papilloedema” and “deep learning”
Lenus, OpenDoar , RIAN (n=0)	
Arxiv (n=3) https://arxiv.org/abs/2112.09970 https://arxiv.org/abs/1907.09449	Abstract = papilledema AND abstract = deep learning

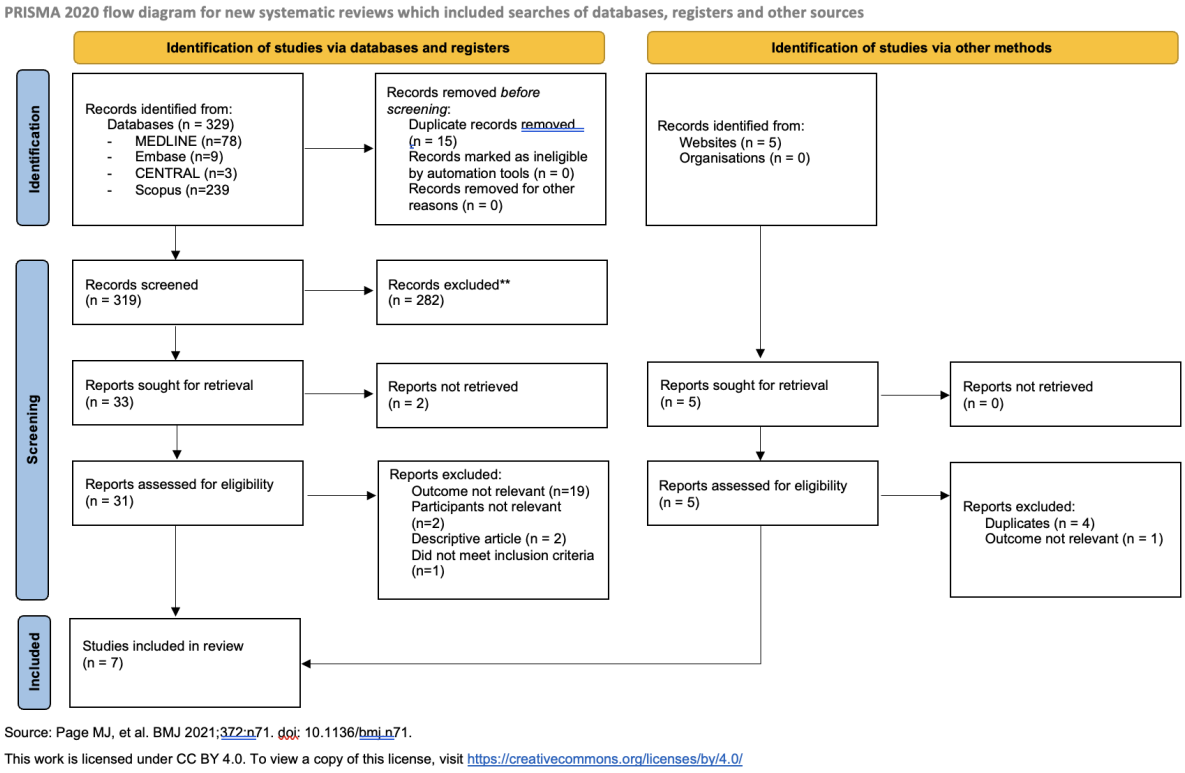

Figure 7. Prisma flow diagram for systematic review updated on the 28th June 2024

3.1.2 Findings

An overview of the findings from this systematic review can be visualised in Table 5 which corresponds to the data selected for extraction in the Methods (Table 2).

Table 5. Summary of systematic review

Author(s) (year)	**Title**	**Classification aim**	**Dataset modality**	**Dataset size**	**Dataset labels**	**Performance**
Avramidis et al. (2022)	Automating detection of papilledema in pediatric fundus images with explainable machine learning	1. Papilloedema versus pseudopapilloedema	Colour fundus photographs	331	Papilloedema (n=149) Pseudopapilloedema (n=182)	1. AUC 78.3%, accuracy 80%
Biousse et al. (2024)	Application of a deep learning system to detect papilledema on nonmydriatic ocular fundus photographs in an emergency department	1. Papilloedema versus none 2. Abnormal versus normal optic disc	Colour fundus photographs	1608	Papilloedema (n=50) Other disc abnormalities (n=180) Normal (n=1378)	1. AUC 0.97, sensitivity 84%, specificity 98.9% 2. AUC 0.92, sensitivity 75.6%, specificity 89.6%
Diener (2023)	Discriminating healthy optic discs and visible optic disc drusen on fundus autofluorescence and colour fundus photography using deep learning - a pilot study	1. Pseudopapilloedema versus normal optic disc	Colour fundus photographs	160	Pseudopapilloedema (n=80) Normal (n=80)	1. Accuracy 92.5%, sensitivity 85%, specificity 100%
Katarzyn	Residual attention	1. Pseudopa	Optical	2339	Pseudopapill	1. Accuracy

a (2024)	network for distinction between visible optic disc drusen and healthy optic discs	pilloedema versus normal optic disc	coherence tomography images		oedema (n=1023) Normal (n=1316)	87.54%, precision 97.16%, recall 90.49%
Lin (2024)	The BONSAI deep learning system can accurately identify paediatric papilledema on standard ocular fundus photographs	1. Papilloedema versus none 2. Abnormal versus normal optic disc	Colour fundus photographs	898	Papilloedema (n=254) Other disc abnormalities (n=86) Normal (n=558)	1. AUC 0.99, sensitivity 98%, specificity 94.1% 2. AUC 0.99, sensitivity 87.3%, specificity 98.5%
Liu (2021)	Detection of optic disc abnormalities in color fundus photographs using deep learning	1. Abnormal versus normal optic disc	Colour fundus photographs	944	Abnormal (n=364) Normal (n=580)	1. AU-ROC 0.87, sensitivity 90%, specificity 69%
Milea (2020)	Artificial intelligence to detect papilledema from ocular fundus photographs	1. Papilloedema versus none 2. Abnormal versus normal optic disc	Colour fundus photographs	14341	Papilloedema (n=2148) Other disc abnormalities (n=3037) Normal (n=9156)	1. AUC 0.96, sensitivity 96.4%, specificity 84.7% 2. AUC 0.98, sensitivity 86.6%, specificity 95.3%

3.1.2.1 Datasets

All datasets were made up of colour fundus photographs except for one record [152] that utilised optical coherence tomography images. One dataset included fluorescein angiography images as well as colour fundus photographs [153]. The datasets ranged in size from 331 total images to 14341 total images. The mean dataset size was 2980.14 total images.

Five records specifically included papilloedema labels (Avramidis et al, Biousse et al., Lin et al, Liu et al, Milea et al) [3,154–157]. Of those, three met the same inclusion criteria as defined within this labelling protocol for papilloedema that proven evidence of intracranial pressure had to be defined either by positive neuroimaging or raised opening pressure on lumbar puncture (Avramidis, Biousse, Milea) [3,154,157]. Lin et al. reported that the label of papilloedema was determined by a local neuro-ophthalmologist with full access to the patient's clinical records including any brain imaging and lumbar puncture results, however, it is not specified that evidence of raised intracranial pressure was required in order to apply the diagnosis of papilloedema [155]. Papilloedema was included as one of many pathologies under the label abnormal Liu et al and the method in which papilloedema was confirmed was not provided [156]. The number of papilloedema images per dataset ranged from 50 to 2148 total images. The mean total number of papilloedema images was 536. The percentage distribution of papilloedema labels per dataset ranged from 3.11 - 45% with a mean percentage of 22.84% which can be visualised in Table 6.

Table 6. Percentage distribution of the 'papilloedema' label per dataset

Papilloedema versus no papilloedema	**Papilloedema images (percentage of total dataset)**
Avramidis et al. (2022)	149 (45%)
Biousse et al. (2024)	50 (3.11%)
Lin et al. (2024)	254 (28.29%)
Milea et al. (2020)	2148 (14.98%)

Three records specifically included pseudopapilloedema (Avramidis, Diener, Nowomiejska) in their datasets [152–154]. Avramidis used the label pseudopapilloedema while Diener and Katarzyna included Optic Disc Drusen which as outlined in the Introduction is a form of pseudopapilloedema. Avramidis reported 182 labels of pseudopapilloedema, Diener reported to have 100 Optic Disc Drusen Fluorescein Angiography image labels and 100 Optic Disc Drusen colour fundus photograph labels in their dataset which were divided between the training and validation, and external validation processes. Katarzyna reported 1023 Optic Disc Drusen optical coherence tomography images within their dataset.

Biousse, Lin, Liu and Milea included the label abnormal optic disc or other disc abnormality. This included papilloedema images for Liu among other pathologies [156]. This was distinct from papilloedema for Biousse, Lin and Milea [3,155,157]. The number of abnormal optic disc labels within a dataset varied between records from 86 to 3037 total images with a mean number of 704.57 images.

3.1.2.2 Deep learning approach

All records used bespoke deep learning requiring data scientist input. Three publications utilised the Modified Version of the Brain and Optic Nerve Study Artificial Intelligence (BONSAI) deep learning system (Biousse, Lin and Milea). This is a deep learning system that was developed through collaboration across a number of international neuro-ophthalmology sites in 12 centres in 8 countries to train the model on a dataset of 6443 fundus images including 3563 abnormal disc images [158]. The other models used a variety of different deep learning models which require very specialised expertise and are beyond the scope of this research.

3.1.2.3 Model performance

Papilloedema versus no papilloedema

Avramidis et al reported an AUC of 78.3% and best accuracy of 78.1% [154]. Biousse et al. reported an AUC of 0.97, sensitivity of 84.0% and specificity of 98.9% [157]. Lin et al. reported an AUC of 0.99, sensitivity of 98.0% and specificity of 94.1% [155]. Milea et al. reported an AUC of 0.99. External testing demonstrated an AUC of 0.96, sensitivity of 96.4% and specificity of 84.7% [3].

Normal versus abnormal optic discs

Biousse et al. reported an AUC of 0.92, sensitivity of 75.6% and specificity of 89.6% [157]. Diener et al. reported a sensitivity of 85%, specificity 100% and accuracy of 92.5% for colour fundus photograph images and 100% accuracy, sensitivity and specificity for fluorescein angiography images [153]. Nowomiejska et al. reported a mean accuracy of 85.45% for the detection of healthy discs and 87.54% for the detection of optic disc drusen with a precision of 89.94% for healthy discs and 97.16% for optic disc drusen and a recall of 97.06% for healthy discs and 90.49% for optic disc drusen [152]. Lin et al. reported a performance AUC of 0.99, sensitivity of 87.3% and specificity of 98.5% [155]. Liu et al reported an AUC-ROC of 0.99, sensitivity of 94% and specificity of 96% on initial training with external validation performance reported as AUC-ROC 0.87, sensitivity 90% and specificity 69% [156]. Milea et al. reported an AUC of 0.98 on external testing with sensitivity 86.6%, specificity 95.3% and accuracy 91.8% [3].

3.2 Dataset and participants

As described in the Methods, there were two datasets used in this research:

1. Dataset I: A retrospectively curated dataset from the Electronic Healthcare Record of Moorfields Eye Hospital, NHS Foundation Trust
2. Dataset II: A publicly available dataset downloaded from a freely available online repository

Herein, each dataset will be described in detail individually.

3.2.1 Dataset I: Moorfields Eye Hospital

3.2.1.1 Dataset curation

An initial record search was performed on all electronic patient letters, OpenEyes (ToukanLabs), using Magic xpa from the 2nd April 2008 until the 4th May 2022 using the keywords ‘papilledema’ and ‘papilloedema’ yielding a result of 9691 letters of 7148 patients. This search was expanded on the 8th August 2022 to include the keyword ‘OCT’ in combination with ‘papilloedema’ OR ‘papilledema.’ The results were exported as a spreadsheet which contained 12101 distinct patient images with the corresponding letter diagnosis. There were 9193 optical coherence tomography images including 4598 right eyes and 4595 left eyes. Clinic letter diagnoses included 1238 normal optic discs, 2299 papilloedema, 1552 previous papilloedema, 201 suspected papilloedema, 3021 pseudopapilloedema, 729 diagnoses of other disc abnormalities and 153 with no record. There were 2908 colour fundus photographs corresponding to 1454 left eyes and 1454 right eyes. The clinic letter diagnoses for the colour fundus photographs included 1687 normal optic discs, 3250 papilloedema, 2084 previous papilloedema, 313 suspected papilloedema, 3630 pseudopapilloedema, 917 other disc abnormality and 220 no record.

Any imaging protocol that was not disc centred ART or rectangular volume was excluded which resulted in 8239 images: 5437 optical coherence tomography images and 2802 colour fundus photographs. All diagnoses of suspected papilloedema and previous papilloedema were excluded which resulted in 6339 images corresponding to 4229 optical coherence tomography

images and 2110 colour fundus photographs. This resulted in the following labels for manual review and application of the inclusion criteria as described in the Methods:

- Normal: 778 optical coherence tomography images, 442 colour fundus photographs
- Papilloedema: 1440 optical coherence tomography images, 919 colour fundus photographs
- Pseudopapilloedema: 1611 optical coherence tomography images, 570 colour fundus photographs
- Other disc abnormality: 400 optical coherence tomography images, 179 colour fundus photographs

With data scientist support, a Structured Query Language search was performed of all patients attending the Accident and Emergency Department who underwent a disc-centred optical coherence tomography scan and the results were exported in a spreadsheet which corresponded to 17000 images. The diagnosis from the electronic patient letter at time of imaging was also exported. These results were cross-referenced against the spreadsheet created from the OpenEyes electronic letter search to ensure all cases of papilloedema had been identified. All disc-centred optical coherence tomography images that at time of diagnosis had been labelled as normal fundus examination and normal optical coherence tomography image were labelled as “A&E normal.”

Finally, only the first visit from each patient was included to avoid data leakage. For the purposes of training a deep learning model, the dataset is divided into separate training, validation and test sets. The model uses the training set to develop its parameters while the validation and test sets are used for fine tuning. Data leakage occurs when there is overlap of data between the training and test sets and can result in a falsely high performance as the model simply recognises the previous image and returns the same prediction [159]. For this reason, only one image per eye was used and only the first visit from each patient was included in the curation of this dataset.

Following exclusion of any duplicate attendances and manual review to ensure each label met the inclusion criteria, the final dataset consisted of 6980 optical coherence tomography images and 6982 colour fundus photographs. The breakdown per label can be visualised in Table 7.

Table 7. Number of images per label(s) in Dataset I			
Label I	**Label II**	**Optical coherence tomography**	**Colour fundus photograph**
A&E Normal	-1	6556	6558
Normal	0	120	120
Papilloedema	1	138	138
Pseudopapilloedema	2	143	143
Other abnormalities	3	23	23

3.2.1.2 Dataset processing

As described within the Methods, the curated dataset was subsequently shared with a third party organisation, Softwire, who performed deidentification of the dataset verified by Moorfields Eye Hospital NHS Foundation Trust. Following deidentification, the dataset was uploaded to the Moorfields Google Cloud Platform within a secure storage bucket on a shared account with restricted permissions for access. In order to comply with data governance regulations, the dataset could only be accessed using a Virtual Machine within the Moorfields Eye Hospital network. A specific Project for this research was created within the Google Cloud Platform and the dataset was copied into a new storage bucket within the same Moorfields Eye Hospital Google Cloud. Following this, all tasks were performed within the designated Project.

Along with the de-identified dataset, a comma separated values file was provided with the new deidentified file names and corresponding label as provided from the original dataset. The labels were as follows: A&E normal = -1, Normal = 0, Papilloedema = 1, Pseudopapilloedema = 2 and Other abnormalities = 3 (Table 7). In accordance with the clinical use cases to train code-free deep learning models to classify papilloedema versus no papilloedema; and to classify normal optic discs versus abnormal optic discs, labels -1 and 0 were merged to create a singular ‘Normal’ label and labels 2 and 3 were merged to create a singular ‘Other optic disc abnormality’

label. The updated labels with the corresponding image file names were saved in a new master spreadsheet. The dataset was subsequently processed from DICOM format to PNG format in a variety of forms with data scientist support, as described in detail in the Methods.

3.2.2 Dataset II: Publicly-available dataset

The publicly-available dataset ‘Identification of Pseudopapilloedema’ containing 1369 disc-centred colour fundus photographs was downloaded to the local computer. This included 779 images labelled ‘Normal,’ 295 images labelled ‘Papilloedema’ and 295 images labelled ‘Pseudopapilloedema.’ As described within the Methods, this dataset was subsequently uploaded to the specific Project within the Google Cloud Platform onto a storage bucket with formatting of the corresponding comma separated values files.

3.3 Experiments and results: Model development

Sixteen binary cloud-hosted code-free deep learning models were trained in total using the high accuracy feature within AutoML Vertex AI. Eight models were trained to classify between the presence of papilloedema versus no papilloedema. Eight models were trained to distinguish between normal versus abnormal optic discs. Eight models were trained using optical coherence tomography images and eight models were trained on disc-centred colour fundus photographs. The performance of each model is provided at a default confidence threshold of 0.5. This was manually adjusted for each model to 0.25 to examine model performance at this threshold also because a lower threshold is relevant to the clinical use case of optic disc abnormality detection and need for high sensitivity. Sensitivity and specificity were manually calculated for each model based on the metrics provided in a contingency table using the formula.

The process of model development will be described in detail for two models (model one and three) using different imaging modalities (colour fundus photographs and optical coherence tomography). The key results for the remaining models will be summarised thereafter in Tables 8-11.

3.3.1 Model training

3.3.1.1 Papilloedema versus no papilloedema

Model one: High resolution colour fundus photographs for the classification of papilloedema versus no papilloedema

Model one was trained using Google Cloud AutoML training as an image classification (Single-label) model using high resolution colour fundus photographs. This model took 3 hours 23 minutes to train and used 24 node hours. A total of 6,976 images were uploaded to the model. Automatic random data split was performed by Google Cloud AutoML Vertex AI assigning 80% of the images to the training set, 10% to the validation set and 10% to the test set corresponding to 5580 images for training, 698 images for validation and 698 images as test items.

Model one performance: all labels

At a confidence threshold of 0.5, precision was 97.99%, recall was 97.99%, PR AUC was 0.991 and log loss was 0.105. Following adjustment of the confidence threshold to 0.25, precision was 97.3, recall was 98.3 and the average precision was 0.991.

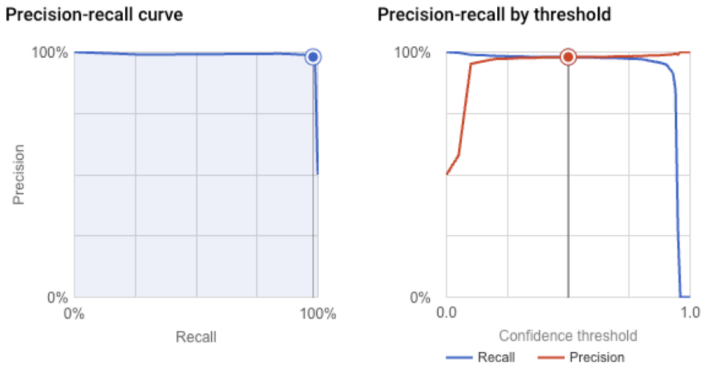

Figure 8. Precision-recall curve and precision-recall at confidence threshold 0.5 for model one for all labels (Adapted from Google Cloud AutoML Vertex AI)

Model one performance: no papilloedema label

At a confidence threshold of 0.5, average precision was 0.993, precision was 98.3% and recall was 99.7%. On adjustment of the confidence threshold to 0.25, average precision was 0.993, precision was 98% and recall was 99.9%.

Model one performance: papilloedema label

At a confidence threshold of 0.5, average precision was 0.218, precision was 50% and recall was 14.3%. At a confidence threshold of 0.25, average precision was 0.218, precision was 37.5% and recall was 21.4%.

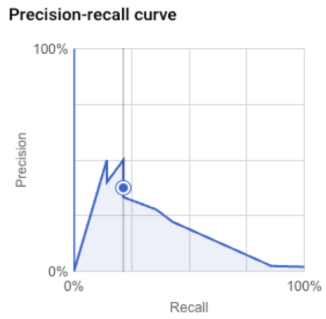

Figure 9. Model one precision-recall curve for papilloedema label at 0.5 confidence threshold

Model one performance: sensitivity and specificity

The sensitivity and specificity of model one for the identification of papilloedema was manually calculated from the confusion matrix provided by Google Cloud AutoML demonstrating how the model classified each label in the test dataset (n=698) at the 0.5 confidence threshold. Model one sensitivity for papilloedema was 14.29% and specificity was 99.7%.

Models one, two, nine and fifteen: colour fundus photographs for the classification of papilloedema versus no papilloedema

The results for each model trained to differentiate between the presence of papilloedema or no papilloedema is summarised below in Table 8. Each of these models was trained on colour fundus photographs using either Dataset I or Dataset II.

Table 8. Models trained using colour fundus photographs (CFP) for the classification of papilloedema versus no papilloedema

Image Type	**Classification**	**Model**	**Data**	**Evaluation Metrices**									
			Total number of images (80% training, 10% each for testing and validation)	Sensitivity	Specificity	TP	FP	FN	TN	Precision	Recall	F1	Specificity
CFP	Papilloedema versus no papilloedema	Model 1: High resolution images	6977	14.29%	99.70%	2	2	12	682	50%	14%	22 %	100%
CFP	Papilloedema versus no papilloedema	Model 2: Square re-sized 512x512mm	6976	21.42%	98.54%	3	20	11	674	23%	21%	22 %	99%
CFP	Papilloedema versus no papilloedema	Model 9: High resolution images following random undersampling	413	85.71%	74.07%	12	7	2	20	63%	86%	73 %	74%
CFP	Papilloedema versus no papilloedema	Model 15: Publicly-available data (Dataset II)	1368	86.21%	100.00%	25	0	4	107	100%	86%	93 %	100%

Colour fundus photograph (CFP); True positive (TP); False positive (FP); False negative (FN); True negative (TN)

Model three: Optical coherence tomography enface 512x512 images for the classification of papilloedema versus no papilloedema

Model three was trained using Google Cloud AutoML training as an image classification (Single-label) model using enface optical coherence tomography images. This model took 3 hours 28 minutes to train and used 24 node hours. A total of 6,973 images were uploaded to the model. Automatic random data split was performed by Google Cloud AutoML Vertex AI assigning 80% of the images to the training set, 10% to the validation set and 10% to the test set corresponding to 5578 images for training, 698 images for validation and 697 images as test items.

Model three performance: All labels

At the 0.5 confidence threshold, precision was 97.99%, recall was 97.99% PR AUC was 0.997 and log loss was 0.088. Following adjustment of the confidence threshold to 0.25, model performance metrics changed to an average precision of 0.997, precision of 96.4% and recall of 98.9%.

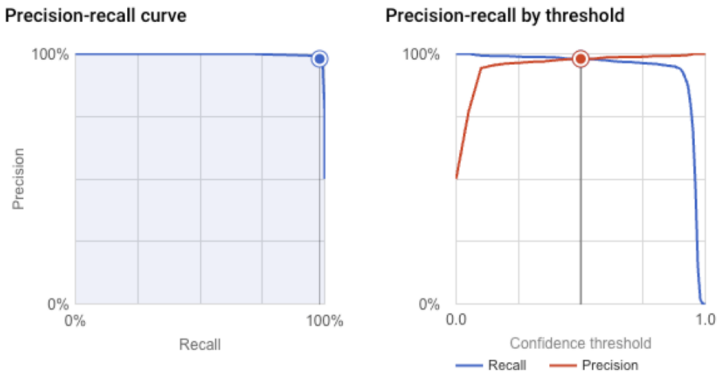

Figure 10. Precision-recall curve and precision-recall at confidence threshold 0.5 for model three for all labels (Adapted from Google Cloud AutoML Vertex AI)

Model three performance: No papilloedema label

At a confidence threshold of 0.5, average precision was 0.999, precision was 98.1% and recall was 99.9%. At a confidence threshold of 0.25, average precision was 0.999, precision was 98% and recall was 100%.

Model three performance: Papilloedema label

At a confidence threshold of 0.5, average precision was 0.331, precision was 50% and recall was 7.1%. At the confidence threshold of 0.25, average precision was 0.331, precision was 33.3% and recall was 42.9%.

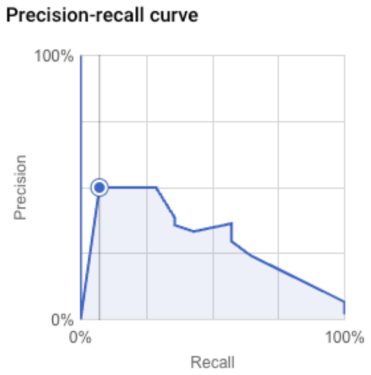

Figure 11. Model three precision-recall curve for papilloedema label at 0.5 confidence threshold (Adapted from Google Cloud AutoML Vertex AI)

Model three performance: sensitivity and specificity

The sensitivity and specificity of model one for the identification of papilloedema was manually calculated from the confusion matrix provided by Google Cloud AutoML demonstrating how the model classified each label in the test dataset (n=697) at the 0.5 confidence threshold. Model three sensitivity for papilloedema was 7.14% and specificity was 99.85%.

Models three, four, ten and eleven: optical coherence tomography images for the classification of papilloedema versus no papilloedema

Table 9. Models trained using optical coherence tomography (OCT) images for the classification of papilloedema versus no papilloedema

Image Type	Classification	Model	Data	Evaluation Metrices									
			Total number of images (80% training, 10% each for testing and validation)	Sensitivity	Specificity	TP	FP	FN	TN	Precision	Recall	F1	Specificity
OCT	Papilloedema versus no papilloedema	Model 3: Enface 512x512 images	6973	7.14%	99.85%	1	1	13	682	50%	7%	13%	100%
OCT	Papilloedema versus no papilloedema	Model 4: Middle b scan square 512x512 images	6974	0.00%	100.00%	0	0	14	683	0%	0%	0%	100%
OCT	Papilloedema versus no papilloedema	Model 10: Enface 512x512 images following random undersampling	413	100.00%	81.48%	14	5	0	22	74%	100%	85%	81%
OCT	Papilloedema versus no papilloedema	Model 11: Middle b scan square images following random undersampling	413	92.86%	74.08%	13	7	1	20	65%	93%	76%	74%

Optical coherence tomography (OCT)); True positive (TP); False positive (FP); False negative (FN); True negative (TN)

3.3.1.2 Normal versus abnormal

Models five, six, twelve and sixteen: colour fundus photographs for the classification of normal versus abnormal optic discs

Table 10. Models trained using colour fundus photographs (CFP) for the classification of normal versus abnormal optic discs

Image Type	Classification	Model	Data	Evaluation Metrices									
			Total number of images (80% training, 10% each for testing and validation)	Sensitivity	Specificity	TP	FP	FN	TN	Precision	Recall	F1	Specificity
CFP	Normal versus abnormal	Model 5: High resolution images	6976	40%	98.35%	12	11	18	656	52%	40%	45 %	98%
CFP	Normal versus abnormal	Model 6: Square resized 512x512	6976	23.30%	98.80%	7	8	23	659	47%	23%	31 %	99%
CFP	Normal versus abnormal	Model 12: High resolution images following random under-sampling	911	73.33%	86.89%	22	8	8	53	73%	73%	73 %	87%
CFP	Normal versus abnormal	Model 16: Publicly-available dataset (Dataset II)	1368	98.31%	100.00%	58	0	1	78	100%	98%	99 %	100%

Colour fundus photograph (CFP); True positive (TP); False positive (FP); False negative (FN); True negative (TN)

Models seven, eight, thirteen and fourteen: optical coherence tomography images for the classification of normal versus abnormal optic discs

Table 11. Models trained using optical coherence tomography (OCT) images for the classification of normal versus abnormal optic discs

Image Type	Classification	Model	Data	Evaluation Metrices									
			Total number of images (80% training, 10% each for testing and validation)	Sensitivity	Specificity	TP	FP	FN	TN	Precision	Recall	F1	Specificity
OCT	Normal versus abnormal	Model 7: Enface 512x512 images	6973	36.67%	98.20%	11	12	19	655	48%	37%	42%	98%
OCT	Normal versus abnormal	Model 8: Middle b scan square 512x512 images	6974	50.00%	97.45%	15	17	15	650	47%	50%	48%	97%
OCT	Normal versus abnormal	Model 13: Enface 512x512 images following random undersampling	909	90.00%	91.67%	27	5	3	55	84%	90%	87%	92%
OCT	Normal versus abnormal	Model 14: Middle b scan square images following random undersampling	910	76.67%	91.80%	23	5	7	56	82%	77%	79%	92%

Optical coherence tomography (OCT)); True positive (TP); False positive (FP); False negative (FN); True negative (TN)

3.4 External Validation

External validation was carried out using synchronous online predictions via the Google Cloud AutoML Vertex AI graphical user interface. This did not require any coding expertise. External validation was carried out on all code-free deep learning models developed within this thesis that achieved performances comparable or superior to the clinical literature: model nine, model twelve, model fifteen and model sixteen. As models one to eight, ten, eleven, thirteen and fourteen performed poorly in comparison to the reported bespoke deep learning models, they did not proceed to external validation.

3.4.1 Dataset I

The dataset I models were externally validated using fifteen images from Dataset II. These fifteen images were copied into a new folder on the local computer and included 5 normal colour fundus photographs, 5 pseudopapilloedema colour fundus photographs and five papilloedema colour fundus photographs. The selected images can be visualised in Figure 12.

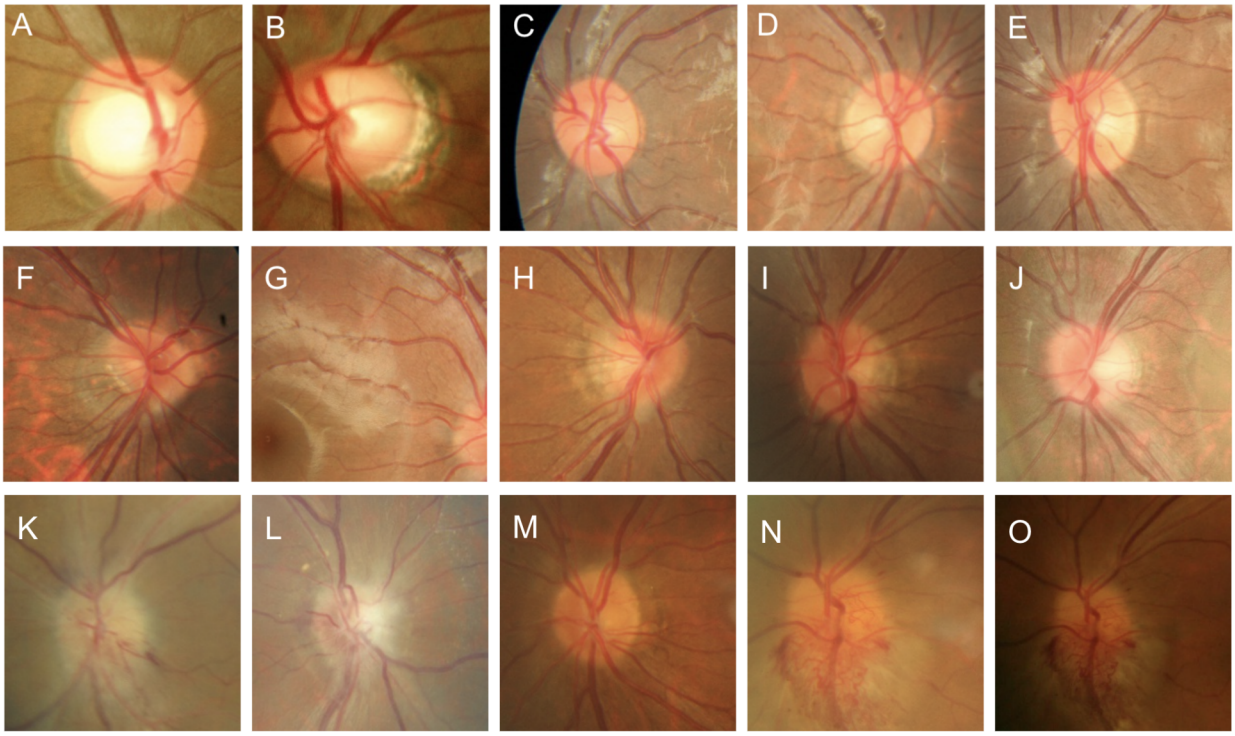

Figure 12. External validation dataset 1. **A-E.** show the Normal labelled images, **F-J.** show the Pseudopapilloedema labelled images, **K-O.** show the Papilloedema labelled images.

Model nine: High resolution colour fundus photographs for the classification of papilloedema versus no papilloedema following random undersampling

Model nine was located within the Model Registry on Google Cloud AutoML Vertex AI and ‘Deploy and Test’ was selected via the graphical user interface. Notification was provided once the model had been actively deployed. Images could then be uploaded to the model with predictions returned in real-time. The label predictions were supplied with a confidence score that indicates how confident the model is that the predicted label is the true label for each image. When tested on normal colour fundus photographs, model nine achieved a confidence score of >0.5 (0.547-0.705) for all images with the prediction that there was no papilloedema. It performed poorly when presented with pseudopapilloedema images and reported a confidence score of >0.69 (0.699-0.953) for the predicted label of papilloedema. Performance when provided with papilloedema colour fundus photographs was variable with confidence scores for the prediction of papilloedema ranging from 0.311-0.963. Sensitivity and specificity were manually calculated. Model nine achieved a sensitivity of 80% and specificity of 100% in the prediction of papilloedema.

Table 12. Model nine external validation performance by image			
	MODEL NINE		
	True label	**Prediction: No papilloedema**	**Prediction: Papilloedema**
1	Normal	0.547	0.453
2	Normal	0.572	0.428
3	Normal	0.633	0.367
4	Normal	0.705	0.295
5	Normal	0.605	0.395
6	Pseudopapilloedema	0.219	0.781
7	Pseudopapilloedema	0.100	0.900
8	Pseudopapilloedema	0.047	0.953
9	Pseudopapilloedema	0.301	0.699
10	Pseudopapilloedema	0.201	0.799
11	Papilloedema	0.073	0.927
12	Papilloedema	0.689	0.311
13	Papilloedema	0.448	0.552
14	Papilloedema	0.037	0.963
15	Papilloedema	0.174	0.826

Model twelve: High resolution colour fundus photographs for the classification of normal optic discs versus abnormal optic discs following random undersampling

Model twelve was located within the Model Registry on Google Cloud AutoML Vertex AI and 'Deploy and Test' was selected via the graphical user interface. Notification was provided once the model had been actively deployed. Images could then be uploaded to the model with predictions returned in real-time. The label predictions were supplied with a confidence score that indicates how confident the model is that the predicted label is the true label for each image. When tested on normal colour fundus photographs, model twelve incorrectly predicted the abnormal label for three of the five images it was tested on (confidence score range from 0.708-0.942). The first two normal colour fundus photographs it was tested on were correctly predicted as normal with a confidence score of 0.824 and 0.517, respectively. The model achieved a greater performance in the classification of four pseudopapilloedema images as abnormal with a confidence score of 0.920-0.45. One image was incorrectly predicted as normal with a confidence score of 0.852. Model twelve demonstrated very poor performance in the identification of papilloedema colour fundus photographs as abnormal. It predicted the normal label for all five images it was tested on with confidence predictions of >0.650 (0.650-0.934). Sensitivity and specificity were manually calculated. Model twelve achieved a sensitivity of 40% and specificity of 40% for the prediction of abnormal optic discs.

Table 13. Model twelve external validation performance by image			
	MODEL TWELVE		
	True label	**Prediction: Normal**	**Prediction: Abnormal**
1	Normal	0.824	0.176
2	Normal	0.517	0.483
3	Normal	0.058	0.942
4	Normal	0.167	0.833
5	Normal	0.292	0.708
6	Pseudopapilloedema	0.070	0.930
7	Pseudopapilloedema	0.080	0.920
8	Pseudopapilloedema	0.080	0.920
9	Pseudopapilloedema	0.852	0.148
10	Pseudopapilloedema	0.055	0.945
11	Papilloedema	0.650	0.350
12	Papilloedema	0.748	0.252
13	Papilloedema	0.929	0.071
14	Papilloedema	0.934	0.066
15	Papilloedema	0.870	0.130

3.4.2 Dataset II

Dataset II was externally validated using a curated dataset of colour fundus photographs identified through a Google Image search with the image location and label confirmed on the host website. This dataset was made up of three colour fundus photographs of normal optic discs, three colour fundus photographs of optic discs with papilloedema and four colour fundus photographs of other disc abnormalities including optic disc drusen and tilted disc. The selected images for this dataset can be visualised in Figure 16.

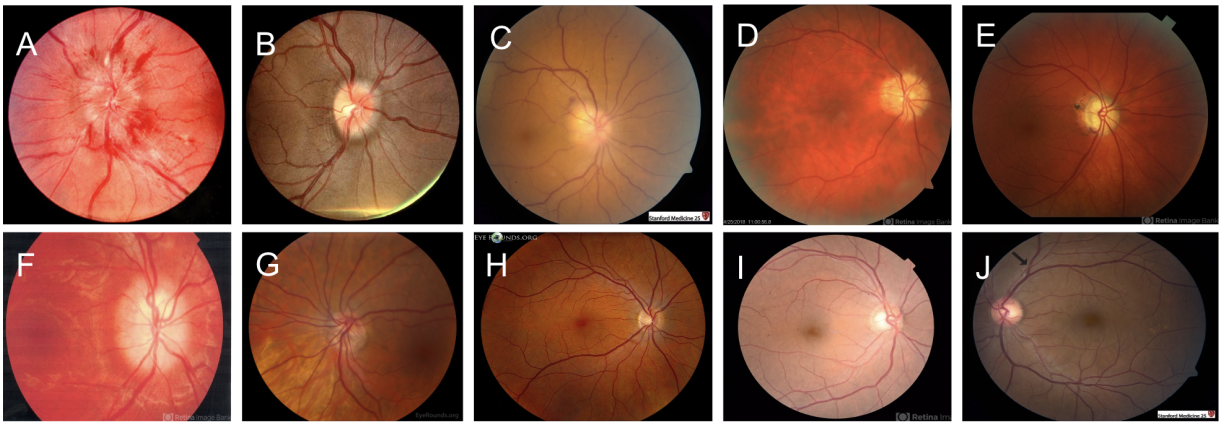

Figure 16. External dataset 2. **A-C.** show the Papilloedema labelled images, **D-G.** show Pseudopapilloedema labelled images (D-F. optic disc drusen; G. tilted disc), **H-J.** show the Normal labelled images.

Model fifteen: Publicly-available colour fundus photographs for the classification of papilloedema versus no papilloedema

Model fifteen was located within the Model Registry on Google Cloud AutoML Vertex AI and 'Deploy and Test' was selected via the graphical user interface. Notification was provided once the model had been actively deployed. Images could then be uploaded to the model with predictions returned in real-time. The label predictions were supplied with a confidence score that indicates how confident the model is that the predicted label is the true label for each image. Model fifteen showed good performance with a confidence score of >0.648 (0.648-0.923) for the prediction of no papilloedema when tested on normal colour fundus photographs.

Similarly, it correctly predicted all pseudopapilloedema images as no papilloedema with a confidence score ranging from 0.639 to 0.868. Model fifteen performed poorly in the correct prediction of papilloedema from colour fundus photographs. Two of the three papilloedema colour fundus photographs the model was tested on were incorrectly predicted as no papilloedema with confidence scores of 0.972 and 0.669, respectively. It correctly predicted one of the papilloedema images with a confidence score of 0.885. Sensitivity and specificity were manually calculated. Model fifteen achieved a sensitivity of 33.3% and specificity of 100% for the prediction of papilloedema.

Table 14. Model fifteen external validation performance by image

	MODEL FIFTEEN		
	True label	**Prediction: No papilloedema**	**Prediction: Papilloedema**
1	Normal	0.923	0.077
2	Normal	0.648	0.352
3	Normal	0.662	0.338
4	Pseudopapilloedema	0.824	0.176
5	Pseudopapilloedema	0.650	0.350
6	Pseudopapilloedema	0.868	0.132
7	Pseudopapilloedema	0.639	0.361
8	Papilloedema	0.115	0.885
9	Papilloedema	0.972	0.028
10	Papilloedema	0.669	0.331

Model sixteen: Publicly-available colour fundus photographs for the classification of normal optic discs versus abnormal optic discs

Model sixteen was located within the Model Registry on Google Cloud AutoML Vertex AI and ‘Deploy and Test’ was selected via the graphical user interface. Notification was provided once the model had been actively deployed. Images could then be uploaded to the model with predictions returned in real-time. The label predictions were supplied with a confidence score that indicates how confident the model is that the predicted label is the true label for each image. Model sixteen incorrectly predicted all normal images as abnormal with a confidence score ranging from 0.619 to 0.930. Performance was better when tested on colour fundus photographs of pseudopapilloedema correctly identifying three of the four as abnormal (confidence score 0.915-0.938) and one incorrectly as normal (confidence score 0.720). Papilloedema predictions were poor. Two of the three papilloedema colour fundus photographs were predicted as normal with confidence scores of 0.885 and 0.896, respectively. One colour fundus photograph of papilloedema was correctly predicted as abnormal with a confidence score of 0.877. Sensitivity and specificity were manually calculated. Model sixteen achieved a sensitivity of 57.14% and specificity of 0% for the prediction of abnormal optic discs.

Table 15. Model sixteen external validation performance by image			
	MODEL SIXTEEN		
	True label	**Prediction: Normal**	**Prediction: Abnormal**
1	Normal	0.070	0.930
2	Normal	0.154	0.846
3	Normal	0.381	0.619
4	Pseudopapilloedema	0.085	0.915
5	Pseudopapilloedema	0.064	0.936
6	Pseudopapilloedema	0.720	0.280
7	Pseudopapilloedema	0.062	0.938
8	Papilloedema	0.885	0.115
9	Papilloedema	0.896	0.104
10	Papilloedema	0.123	0.877

Chapter 4: Discussion

4.1 Overview

Optic disc abnormalities, particularly papilloedema, are associated with significant diagnostic uncertainty. The potentially fatal implications of misdiagnosing papilloedema leads to a huge burden of referrals to tertiary ophthalmic centres for specialised assessment [73]. The aim of this research was to investigate the performance of code-free deep learning in the binary classification of optic disc abnormalities using colour fundus photographs and optical coherence tomography images. Although the application of deep learning in this task has been previously explored within the clinical literature, the use of code-free deep learning has not been reported on. The research presented in this thesis describes the curation and identification of appropriate datasets, and the development of sixteen code-free deep learning models for the task of either classifying optic discs as normal or abnormal, or the prediction of papilloedema versus no papilloedema. The development of the code-free deep learning models was carried out by a clinician with no coding expertise.

A systematic review of the clinical literature has established that deep learning can be applied to the task of papilloedema and/or optic disc abnormality classification with varying degrees of success. All identified records utilised datasets of colour fundus photographs except for Nowomiejska et al. who curated a dataset of optical coherence tomography images [152]. Dataset size ranged from 331 images [154] to 14341 images [3] with a mean number of 2980.14 images per dataset and median of 944 images. Both datasets used within this thesis for model development contain a greater number of images than the median reported within the clinical literature. Dataset I contains 6981 colour fundus photograph images and 6980 optical coherence tomography images while Dataset II contains 1369 colour fundus photographs.

Model performance varied significantly both within the clinical literature and among the models described in this thesis. Sensitivity and specificity were manually calculated for each of the code-free deep learning models (models one to sixteen) described in this thesis to allow for comparison with those reported in the clinical literature. The eight code-free deep learning models trained initially on Dataset I demonstrated poor sensitivity in all tasks. Sensitivity ranged

from 0-21.42% for the classification of papilloedema versus no papilloedema while specificity for this task ranged from 98.54-100% (models one, two, three and four). Sensitivity for the prediction of normal versus abnormal optic discs ranged from 23.3-50% while specificity range was 97.45-98.8% (models five, six, seven and eight). This sensitivity is significantly worse than any of the bespoke deep learning models reported in the clinical literature which ranges from 84-100% (where provided). Following random undersampling, code-free deep learning performance improved considerably. Sensitivity was 85.71-100% for the binary classification of papilloedema versus no papilloedema (models nine, ten, eleven) and sensitivity was 73.33-90% for the binary classification of abnormal versus normal optic discs (models twelve, thirteen and fourteen). This implies that class imbalance played a negative role in the performance of code-free deep learning models one to eight further supported by the robust performance of both code-free deep learning models trained using Dataset II. Models fifteen and sixteen were trained using Dataset II and demonstrated excellent performance. Model fifteen achieved a sensitivity of 86.21% in the classification of papilloedema and model sixteen achieved a sensitivity of 98.31% in the classification of abnormal optic discs.

Despite the strong performance metrics of models nine, twelve, fifteen and sixteen in validation testing, external validation exposed inherent flaws within each model. Model nine showed strong confidence at correctly predicting colour fundus photographs with a normal optic disc but performed poorly at the classification of pseudopapilloedema and papilloedema. Model twelve successfully predicted the majority of pseudopapilloedema images as abnormal but incorrectly predicted nearly all of the colour fundus photographs containing normal optic discs as abnormal and papilloedema colour fundus photographs as normal. Model fifteen showed good performance in the classification of normal colour fundus photographs and pseudopapilloedema colour fundus photographs. However, when presented with papilloedema images, it incorrectly predicted no papilloedema for the majority of images on which it was tested.

Herein, the findings from this thesis will be examined in greater detail and compared against the clinical literature identified in the systematic review.

4.2 Summary and implications of key findings

The systematic review identified seven publications meeting the inclusion criteria. Of these seven publications, none utilised code-free deep learning for the classification of optic disc abnormalities. All models were trained using bespoke deep learning, i.e., a deep learning algorithm designed and developed by a data scientist with deep learning expertise. All included papers from the systematic review used colour fundus photographs except Nowomiejska et al. who used optical coherence tomography images [152]. All records used bespoke datasets for the purpose of their deep learning model development. Liu et al., also incorporated the publicly-available datasets DRIONS and ACRIMA for inclusion of normal optic disc colour fundus photographs [156] and Nowomiejska et al. used the online repository, Kaggle, for the normal optical coherence tomography images within their dataset.

4.2.1 Application of deep learning for the diagnosis of optic disc abnormalities in adults

4.2.1.1 Deep learning for the prediction of papilloedema versus no papilloedema

Five records identified in the systematic review specifically examined the performance of deep learning at the classification of papilloedema. Avramidis et al. and Lin et al. both used paediatric colour fundus photograph datasets which will be later discussed in greater detail [154,155]. Biousse et al. and Milea et al. examined the application of deep learning to the differentiation of papilloedema versus no papilloedema using colour fundus photographs [3,157]. Liu et al. examined the binary classification of abnormal optic discs versus normal optic discs, including papilloedema within their abnormal image label [156]

Milea et al. trained a bespoke deep learning model using 14,341 mydriatic colour fundus photographs retrospectively collected via international collaboration from 19 different sites. This corresponded to 9156 normal optic disc photographs, 2148 colour fundus photographs with papilloedema (with confirmed intracranial hypertension) and 3037 images of other disc

abnormalities. The authors reported a sensitivity of 96.4% and specificity of 84.7% for the validation test set in the classification of papilloedema versus no papilloedema [3]. Biousse et al. retrospectively analysed the colour fundus photographs from the FOTO-ED study, a three phase cross-sectional study specifically examining direct ophthalmoscopy and nonmydriatic ocular fundus photographs in the emergency department, which is referenced in the Introduction [84,86,160]. The authors included nonmydriatic ocular fundus photographs from the study containing the optic disc with sufficient clinical information within the medical record to assign a definitive diagnosis in line with the inclusion criteria. As described, an image could only be labelled as papilloedema if there was confirmed evidence of raised intracranial pressure. Biousse et al. reported a sensitivity of 84% and specificity of 98.9% for the classification of papilloedema versus no papilloedema [157].

Within this thesis, models one, two nine, fifteen and sixteen were all trained on colour fundus photographs using code-free deep learning to classify between papilloedema versus optic discs without papilloedema. Models one and two demonstrated poor performance in the classification of papilloedema. Model one was trained using high resolution colour fundus photographs and achieved a sensitivity of 14.29% and specificity of 99.7%. At a 0.5 confidence threshold, precision for the prediction of papilloedema was 50% and recall was 14.3%. Model two was trained using square resized colour fundus photographs and achieved a sensitivity of 21.42% and specificity of 98.54%. At a confidence threshold of 0.5, precision for the prediction of papilloedema was 23.1% and recall was 21.4%. Conversely, at the same confidence threshold, for both models one and two in the prediction of no papilloedema, the precision was 98.3% and 98.4% respectively, and the recall was 99.7% and 98.5%, respectively. For the binary classification of papilloedema versus no papilloedema, Model fifteen, which was trained on Dataset II, demonstrated poorer sensitivity (86.21%) but higher specificity (100%) in primary validation compared to Milea et al. who reported primary validation sensitivity of (93.2%) and specificity (95.1%) [3]. Model Fifteen demonstrated superior performance to the model described by Biousse et al. who reported a sensitivity of 84.0% for the classification of papilloedema versus no papilloedema with specificity of 98.9%.

The significance of these findings are not yet clear but it must be observed that the bespoke deep learning models reported by Biousse et al., Lin et al. and Milea et al. were trained using the BONSAI-DLS [3,155,157]. Brain and Optic Nerve Study with Artificial Intelligence - Deep Learning System (BONSAI-DLS). The BONSAI consortium was a collaborative project which

resulted in the collection of 14341 colour fundus photographs from 6779 patients across 19 sites. This contained 9156 normal optic discs, 2148 papilloedema discs and 3037 colour fundus photographs of other disc abnormalities. Comparable to the research reported in this thesis, papilloedema could only be diagnosed with proven evidence of raised intracranial pressure [158]. Milea et al. trained their deep learning model using this dataset [3]. Biousse et al. developed an updated version of the BONSAI-DLS, training a bespoke deep learning model on 17,368 colour fundus photographs including 9210 normal optic discs, 3146 papilloedema colour fundus photographs and 5012 other disc abnormalities, applying the same stringent inclusion criteria for papilloedema images [157]. Lin et al. also used an updated version of the BONSAI-DLS to train their bespoke deep learning models [155]. Using a relatively small dataset, code-free deep learning models could be trained by a clinician with no coding expertise with similar performance to those described above.

4.2.1.2 Deep learning for the prediction of normal optic discs versus abnormal optic discs

Two records examined the application of deep learning specifically for the detection of pseudopapilloedema. Nowomiejska et al. used a dataset of optical coherence tomography images to develop a deep learning model to distinguish between optic disc drusen and healthy optic discs. The authors curated a dataset of 116 optical coherence tomography optic disc drusen images from three different ophthalmic institutions (Poland, Italy, Lithuania). A technique called data augmentation, which is commonly used in deep learning to enlarge and diversify datasets, was subsequently applied. Data augmentation typically involves applying changes to the original dataset to create new artificial data including image cropping, resizing, scaling and flipping horizontally or vertically [161]. Following data augmentation, Nowomiejska et al. reported a dataset of 1023 optic disc drusen optical coherence tomography images. The authors then combined this with a publicly available dataset of normal optic disc optical coherence tomography images obtained from the same freely available online repository, Kaggle, whereby Dataset II was identified for the research in this thesis. The authors reported an accuracy of 87.54% for the detection of optic disc drusen with a precision of 97.06% and recall of 90.49% [152]. The performance of this bespoke deep learning is superior to any of the code-free deep learning models trained within this thesis. Models seven and eight demonstrated considerably

poorer performance with precision at 47.8% and 46.9%, respectively, and recall at 36.7% and 50%, respectively, measured at a confidence threshold of 0.5. Adjusting the confidence threshold did not improve precision or recall to a performance greater than the bespoke deep learning model developed by Nowomiejska et al. Random undersampling of the optical coherence tomography datasets did improve performance for both enface optical coherence tomography images and middle b scan optical coherence tomography images as demonstrated by model thirteen (precision 84.4%, recall 90% at 0.5 confidence threshold) and model fourteen (precision 82.1%, recall 76.7% at confidence threshold of 0.5). However, code-free deep learning did not outperform the bespoke deep learning models for the task of abnormal versus normal optic disc classification when trained using optical coherence tomography images. It could be argued that a direct comparison of the bespoke deep learning model trained by Nowomiejska et al. should not be carried out against models seven, eight, thirteen and fourteen. A large proportion of the images Nowomiejska et al. trained their deep learning model with were artificially created via the technique of data augmentation whereas models seven, eight, thirteen and fourteen were trained on real-world data collected from Moorfields Eye Hospital within Dataset I with a higher likelihood of artefacts or variable quality. It would be interesting to apply code-free deep learning to the dataset curated by Nowomiejska et al. to more directly compare performances at the same task.

Diener et al. curated two datasets from the Muenster University Hospital database (Germany) for patients attending between January 2015 and January 2020. The colour fundus photograph dataset contained 80 optic disc drusen images and 80 healthy optic disc photographs. The fluorescein angiography dataset contained 80 optic disc drusen images and 80 healthy disc images. As the application of deep learning to fluorescein angiography images is beyond the scope of this research, the results of that section will not be examined. The deep learning models trained using the colour fundus photograph images showed a sensitivity of 85%, specificity of 100% and accuracy of 92.5%. The authors acknowledge that the small dataset is a limitation and the fact that all images came from the same imaging system may negatively impact the model performance when presented with a dataset acquired from a different imaging system [153]. In comparison to the models trained within this thesis, the results reported by Diener et al. are relatively good. Model five trained on Dataset I showed a very poor sensitivity of 40% with specificity of 98.35% when trained with high resolution colour fundus photographs. Model six similarly showed an even worse sensitivity of 23.3% with specificity at 98.8% when trained using square resized colour fundus photographs. Random undersampling increased sensitivity

performance to 73.33% with specificity measured at 86.89% for model twelve, however, the performance of this code-free deep learning model is still worse than that reported by Diener et al. for their bespoke deep learning algorithm. Only model sixteen outperformed that bespoke deep learning model which as described had a sensitivity of 98.31% and specificity of 100% when trained using Dataset II.

Biousse et al., Liu et al., and Milea et al., all also examined deep learning performance at classifying optic discs as normal versus abnormal in non-paediatric patients. Within their study, Biousse et al. labelled any non-papilloedema disc abnormalities as 'other optic disc abnormalities' including both pseudopapilloedema and other disc pathologies such as anterior ischemic or compressive optic neuropathies. The authors reported a sensitivity of 75.6% and specificity of 89.6% for the classification of normal versus abnormal optic discs [157]. Similarly, Milea et al. defined other optic disc abnormalities to include ischaemic optic neuropathies, pseudopapilloedema and optic disc atrophy. The deep learning model trained to differentiate normal from abnormal optic discs described by the authors had a sensitivity of 93.5% and specificity of 96.2% in the primary validation set [3]. Liu et al. curated a training dataset of 944 colour fundus photographs containing 364 abnormal images and 580 normal optic disc photographs. The authors collected all abnormal photographs from the same neuro-ophthalmological practice with a variety of pathologies represented including papilloedema (22.8%), hypoplasia (16.5%), hereditary optic neuropathy (11.5%) and others. Normal optic disc colour fundus photographs were downloaded from publicly available databases DRIONS [162] and ACRIMA (https://figshare.com/s/c2d31f850af14c5b5232). The authors also identified a third database containing both normal and abnormal colour fundus photographs of optic discs [163]. In primary validation testing, Liu et al. reported a sensitivity of 94% and specificity of 96% for classification between normal and abnormal optic disc [156].

Performance metrics for both models four and five were all worse than for the bespoke deep learning models described above in the classification of normal versus abnormal optic discs. As discussed, random undersampling improved performance and model twelve ultimately achieved a sensitivity (73.33%) closer to that reported by Biousse et al. (75.6%) but still worse than all reported bespoke deep learning models. The primary validation sensitivity (98.31%) and specificity (100%) from model sixteen trained on Dataset II was superior to the primary validation sensitivity and specificity reported by Biousse et al., Liu et al. and Milea et al. for the classification of normal versus abnormal optic discs using colour fundus photographs [3,156,157].

4.2.2 Application of deep learning for the diagnosis of optic disc abnormalities in children

Two of the records identified in this systematic review examined the use of deep learning for optic disc abnormality detection within the paediatric population.

Avramidis et al. used the smallest dataset (n=331) in comparison to the other records included in this review and the two datasets used for the research within this thesis. The authors collected paediatric colour fundus photographs from 105 patients attending Children's Hospital Los Angeles and University of California, Los Angeles between 2011 and 2021. As discussed, the authors applied the same rigorous inclusion criteria for the diagnosis of papilloedema as described in this research. The authors reported an AUC of 78.3% and 78.1% accuracy. The authors propose that their deep learning model performance is comparable to clinical standards and that it could serve as a useful ancillary screening tool. However, it is questionable whether the reported accuracy within this paper is truly acceptable to be utilised in clinical practice for the detection of a potentially fatal condition, particularly in children. The authors do acknowledge that the small size of their dataset is a limitation and that the model may experience challenges when faced with a different patient population [154]. Undoubtedly, if the accuracy were to decrease any further, it would not be an appropriate screening assistant for the detection of papilloedema.

Lin et al. also examined the application of deep learning for the diagnosis of paediatric papilloedema using colour fundus photographs. The authors in this study used a larger dataset of 898 fundus photographs containing 558 normal optic discs, 254 optic discs with papilloedema and 86 colour fundus photographs depicting other optic disc abnormalities. As discussed, the authors did not explicitly define that evidence of raised intracranial pressure was required for the diagnosis of papilloedema, however, it is outlined that the labels were determined by neuro-ophthalmologists with full access to the patient record including any neuroimaging and/or lumbar puncture results. The deep learning system, BONSAI-DLS, was used and models were trained to differentiate between normal versus abnormal optic discs as well as papilloedema versus no papilloedema. The authors reported a sensitivity of 89% and specificity of 94.1% in the classification of papilloedema [155].

The application of deep learning to diagnose papilloedema or other disc abnormalities in paediatric patients is a potentially contentious topic. As outlined within the Introduction, referring

a child to a tertiary ophthalmic institution to investigate for papilloedema is associated with significant stress and anxiety both for the patient and the parents. In addition, very young children are often subject to a greater burden of investigations as general anaesthetic may be required for tests such as lumbar puncture or magnetic resonance imaging of the brain. Nonetheless, the implications of a missed diagnosis of raised intracranial pressure can be fatal. Children can present differently to adults. Papilloedema typically does not present in infancy prior to closure of the fontanelles [164]. This would make a deep learning screening tool for papilloedema largely inaccurate for those less than two years of age [165]. A systematic review and meta-analysis on the presentation of central nervous system tumours in childhood found papilloedema to be present in only 13% of patients with intracranial tumours whereas headache, nausea and vomiting were features in greater than 30%. Papilloedema was more common (34%) in posterior fossa tumours, however, symptoms of nausea and vomiting occurred in 75%, headaches in 67% and abnormal gait and coordination in 60% [166]. Therefore, specifically for a paediatric population, it is difficult to envisage a situation whereby a deep learning screening tool that specifically looks at the binary detection of papilloedema versus no papilloedema could reliably replace the clinical assessment whereby optic disc appearance is taken into consideration in conjunction with symptoms and other signs.

4.2.3 External validation

External validation should always be performed on all deep learning models in order to assess the accuracy and generalisability of a model when presented with new data. However, as one objective of this thesis was to directly compare the accuracy of code-free deep learning against bespoke deep learning, the models that demonstrated markedly poorer performance to the clinical literature on initial validation testing were not externally validated. Models one to eight demonstrated very poor sensitivity in the classification of papilloedema or abnormal optic discs. Models ten, eleven, thirteen and fourteen achieved considerably higher performances when trained on Dataset I following random undersampling, however, they still did not meet the performance metrics described by Nowomiejska et al. within the clinical literature [152].

An essential element of external validation is to provide the model with previously unseen data. For the external validation of models nine and twelve trained on Dataset I, the external validation dataset was curated from colour fundus photographs taken from Dataset II. Five

images from each class (normal optic disc, pseudopapilloedema, papilloedema) were copied into a new folder to allow upload to each model following deployment. As models fifteen and sixteen trained on Dataset II demonstrated a very strong performance on initial validation testing, it can be proposed that the images and labels within this dataset are of sufficient quality. The curation of an external validation dataset for models trained on Dataset II was more challenging as data governance approvals did not allow for Dataset I to be downloaded onto the local computer. A variety of ten colour fundus photographs were identified from a Google Image search including three normal optic disc, four pseudopapilloedema and three papilloedema colour fundus photographs. Although the source site and labels were examined, format and image quality was variable within the dataset.

Despite achieving comparable or superior metrics to the bespoke deep learning models within the clinical literature, code-free deep learning models nine, twelve, fifteen and sixteen demonstrated significantly poorer performances on external validation. Models nine and fifteen were developed as code-free deep learning models for the binary classification of papilloedema versus no papilloedema. Model nine correctly predicted papilloedema for four of the five colour fundus photographs of papilloedema achieving a sensitivity of 80%, however, a 20% failure rate is clearly not acceptable for a condition with such significant implications if missed. Furthermore, it predicted papilloedema for all of the five pseudopapilloedema colour fundus photographs on which it was tested. As described in the Introduction, the incorrect diagnosis of pseudopapilloedema as papilloedema can lead to a series of investigations with an associated burden of anxiety. While it is preferable that a model can accurately recognise an abnormal optic disc, model nine would not serve any purpose within the clinical care pathway to reduce the current burden of papilloedema referrals to tertiary centres. Model fifteen showed poor performance in the prediction of papilloedema, incorrectly labelling two of the three papilloedema colour fundus photographs it was tested on as having no papilloedema. Conversely to model nine, it correctly predicted all pseudopapilloedema colour fundus photographs as no papilloedema. In terms of patient outcomes, a missed diagnosis of papilloedema would clearly have a far greater significance than a missed diagnosis of pseudopapilloedema. Model fifteen also correctly predicted all normal optic disc colour fundus photographs as no papilloedema. This corresponded to a sensitivity of 33.3%. The possibility that model fifteen has merely learnt that the probability of predicting an image as no papilloedema is higher than papilloedema must be considered and the model can only be truly evaluated with a larger dataset of unseen images. This theory would be supported by the

manually calculated specificity of 100%. Models twelve and sixteen were developed to classify optic discs as normal or abnormal. Model twelve showed variable performance when tested with normal optic discs and pseudopapilloedema optic discs. Concerningly, model twelve predicted all papilloedema colour fundus photographs as normal. Model sixteen showed similarly poor performance in the classification of papilloedema predicting normal optic disc for two of the three images it was tested on. As described, the dataset that model sixteen was tested on contained images of variable quality and from different imaging systems and sources. However, the performances of both models twelve and sixteen were unacceptably poor for the use case of abnormal optic disc classification.

Within the clinical literature, bespoke deep learning model performance generally demonstrated a better performance than that visualised in the code-free deep learning models within this thesis. Diener et al. reported a sensitivity of 85%, specificity of 100% and accuracy of 92.5% for their bespoke deep learning model in the classification of colour fundus photographs with optic disc drusen. Although these figures are similar to the metrics achieved by model nine in external validation, Diener et al. tested their model with a larger external dataset of 40 colour fundus photographs [153]. Milea et al. reported a sensitivity of 96.4% and specificity of 84.7% on external testing of their papilloedema versus no papilloedema bespoke deep learning model and a sensitivity of 86.6% and specificity of 95.3% for the classification of normal versus abnormal optic discs. These models were externally validated using a separate dataset of 1505 photographs collected from five centres distinct from the initial dataset [3].

Therefore, all four code-free deep learning models demonstrated performances significantly worse than any of the bespoke deep learning models described within the clinical literature. Despite achieving comparable and/or superior performances to Milea et al. in initial validation testing, models fifteen and sixteen significantly underperformed on external validation.

4.3 Strengths and Limitations

4.3.1 Strengths

There are several strengths to this research. Firstly, this is the first known application of code-free deep learning to the binary classification of papilloedema and other optic disc abnormalities. It examines and describes the strengths and weaknesses code-free deep learning offers for this specific task and also closely examines the performance of bespoke deep learning models within the clinical literature trained for the same use cases. Secondly, a very large bespoke dataset (Dataset I) was curated from a tertiary ophthalmic institution containing over six thousand labelled images of colour fundus photographs and optical coherence tomography images. This is an invaluable resource that is representative of a large cohort of ethnically diverse patients and could be utilised in a variety of different deep learning projects provided data governance regulations are approved. It could also be further enlarged and enhanced in the future to improve quality and label balance.

4.3.2 Limitations

The limitations to this research can be divided into 1) limitations pertaining specifically to the research carried out within this thesis, and 2) limitations relating to code-free deep learning as a tool within healthcare.

4.3.2.1 Limitations to the research within this thesis

There are some limitations to this work that must be outlined. The limitations predominantly relate to Dataset I: The Moorfields Eye Hospital dataset, however, this research was also limited by Dataset II and the need for non-clinical input.

4.3.2.1.1 Limitations of Dataset I

This dataset confers strength to the research within this thesis as outlined above, however, it also plays a role in many of the limitations encountered. This is supported by the fact that the code-free deep learning models trained using Dataset II demonstrated a stronger performance

in optic disc classification using colour fundus photographs. The main limitations associated with Dataset I included challenges identifying images that would meet the strict inclusion criteria specified in the labelling protocol, class imbalance of the dataset, and a lack of metadata.

Inclusion criteria

As outlined within Methods, the inclusion criteria described in the Labelling Protocol was designed to ensure an accurate and high quality dataset. Any image labelled as papilloedema was obliged to have clear documented evidence of raised intracranial pressure either in the form of positive neuroimaging or raised opening pressure on lumbar puncture examination. This resulted in a robust dataset of papilloedema images. However, it also significantly limited the number of available images for inclusion. Moorfields Eye Hospital is a tertiary ophthalmic institution without the facilities to provide general medical acute investigations. Therefore, any patient presenting to the Eye Casualty with concerning features of papilloedema is sent to an alternate general medical institution for further work up. As there is no common shared Electronic Healthcare Record between NHS trusts, results of neuroimaging or lumbar puncture were unavailable for many of these patients unless they had been clearly documented within their Electronic Healthcare Record. This means that a large proportion of images from patients with clinically suspicious features of papilloedema could not be included resulting in a relatively small number of papilloedema images. Similarly, patients could only be labelled as pseudopapilloedema if the patient had been followed up over multiple visits to ensure no evolution of symptoms or clinical features. As discussed, patients with pseudopapilloedema are often sent to the Eye Casualty to exclude papilloedema. If a patient attended with obvious features of pseudopapilloedema and was reviewed by an experienced clinician, they may have been discharged after their first visit and therefore, not meet the inclusion criteria of multiple visits to ensure no evolution of symptoms.

Class imbalance

As discussed, the lack of available clinical metadata during the data collection process resulted in a relatively small cohort of papilloedema and pseudopapilloedema images. However, as Moorfields Eye Hospital is a major ophthalmic tertiary centre, there was no shortage of ‘Normal’ ocular imaging. The decision was made to keep the large dataset of ‘Normal’ images as larger datasets are generally believed to be more valuable for model training [167,168]. Nonetheless, the

class imbalance between labels likely played a negative role in the performance of each code-free deep learning model trained on Dataset I.

Class imbalance occurs when the number of observations in one class considerably outweighs the number of observations in the other class. In binary classification, they can be referred to as the majority class and the minority class. Class imbalance is a common issue in clinical artificial intelligence datasets where data for the target, or 'disease', class is less available than the control, 'normal', class. The reason class imbalance is a problem when the data is highly skewed is that the model will preferentially learn to predict the majority class rather than actually basing its classification on a feature within the image as the majority of the time this will be the correct prediction when there is a very small minority class and very large majority class. Class imbalance, therefore, results in a significantly negative effect on model performance [145,169].

The implications of class imbalance within this research can be visualised with the improved performance of all code-free deep learning models trained on Dataset I following random undersampling. Random undersampling was performed to create a more balanced dataset and the majority: minority class ratio was set at 2:1. Performances for all code-free deep learning models improved indicating that class imbalance within the initial dataset was negatively affecting accuracy.

Lack of metadata

Limited clinical metadata was collected at time of dataset curation due to logistical limitations, governance regulations and lack of available data as outlined within the Electronic Healthcare Record. Metadata refers to additional information relating to the patient including ethnicity and gender. Furthermore, the majority of the metadata that was collected was not processed and uploaded to the Cloud following deidentification of the dataset in order to comply with governance regulations. The images included in Dataset I were obtained from two different imaging systems: 3DOCT-2000 and Triton. It is possible that the deep learning models learnt to recognise different patterns between both imaging systems and made predictions based on which imaging system was used rather than specific disease features. For example, if the Triton imaging system is predominantly used in the Emergency Department or Neuroophthalmology clinics, there is a higher possibility that images from this system are papilloedema rather than the 3DOCT-2000 in an alternative department such as Medical Retina. The model may have

learnt that more images from the Triton imaging system contained papilloedema images, therefore, it would have a greater likelihood of accurate predictions when classifying Triton images with the papilloedema label. This is merely an example to demonstrate how mixed imaging systems within a dataset could be problematic. As the information pertaining to which imaging system the deidentified image came from was not available following upload to the cloud, it was not possible to examine the possibility of this. It is also possible that if there is varying quality between the imaging systems and the majority of high quality images were in the training set, the model may have poorer performance when tested with poorer quality images in the test set. Furthermore, the number of images included that had bilateral involvement was not specified, however, only one image from each patient was utilised within the dataset.

4.3.2.1.2 Limitations of Dataset II

Despite limitations to Dataset I, the labels were robustly confirmed to be accurate. A major limitation to Dataset II was the lack of available information on the labelling process. It is unknown whether these labels were obliged to have evidence of raised intracranial pressure. There was no metadata on what imaging system was used nor clinical metadata on ethnicity, diversity and pathology for pseudopapilloedema within the dataset.

4.3.2.1.3 External validation

There were limitations to the external validation carried out within this thesis. As described, external validation is an essential step in the assessment of a deep learning model. In order to truly assess model accuracy, external validation should be performed by batch prediction. This allows for large datasets to be presented to the model with corresponding predictions returned asynchronously. Efforts to carry out batch prediction without coding expertise were unsuccessful and significant additional data scientist input would have been required. This was considered beyond the scope of the research aims of this thesis. Therefore, the models were externally validated using the graphical user interface within Google Cloud Platform. Though this method of external validation is not as robust as batch prediction, it still provides insight into how the model performs when presented with new data. In addition, Dataset II was obtained from a Google Image search. Therefore, there was unavoidable artefact and mixed quality between images which may have affected model performance on external validation.

4.3.2.1.4 Additional input required

Code-free deep learning is promoted as a tool that is accessible to anyone without data scientist input. This is true for the specific model training aspect. However, in order to carry out the stages surrounding model development, specialised expert input was required. Data scientist input was required to process the images from DICOM to JPEG format and for support uploading the dataset to the Cloud. External support was also required in order to fulfil data governance regulations for de-identification and validation of dataset.

4.3.2.2 Limitations of code-free deep learning

Code-free deep learning faces the same barriers as bespoke deep learning in terms of dataset curation requirements, ethical and data governance regulations, stakeholder acceptability and challenges related to explainability. Despite the fact that accessibility to those without coding expertise is cited as an advantage of code-free deep learning, issues can arise when deep learning models are created by those without in-depth knowledge of these limitations and without the skills needed to interpret model performance accurately [109,170].

4.3.2.1 Explainability

Explainability is a term that describes how a model generates its prediction. Explainability remains a critical obstacle in the deployment of deep learning systems into clinical care. The issue of explainability is further increased by code-free deep learning where the tuning of hyperparameters, network architecture selection and model development is entirely automated with no human input. Abbas et al. when applying code-free deep learning to tabular data in the prediction of visual acuity in age-related macular degeneration patients acknowledge the challenge of explainability and the potential risk of racial biases within deep learning models. Using an open-source artificial intelligence interpretability tool, the What-if Tool, the authors examined predictions their model made across different ethnic groups. The authors reported a higher false negative rate in Asian patients which could be addressed by adjusting the threshold using the What-if Tool. The authors observe that the majority of patients in the major anti-VEGF trials were of White background which may have reinforced racial biases had the authors not made the appropriate adjustments [135].

4.3.2.2 Platform specific limitations

Despite promising results in the clinical literature, the specific platforms each have associated limitations. By nature, the user cannot select network architecture or manually tune the hyperparameters when training a code-free deep learning model. This is an appeal to those without the ability to develop their own models but it further limits explainability. On a general level, the platforms are not cohesive in how the model prediction results are reported. This makes it difficult to effectively compare model performance across platforms even if the same dataset is used. Some authors have overcome this difficulty by manually translating the results into F1 scores [110], however, for those with limited training in data science, this may not be an option [109].

For optimal results, many of the platforms still require some coding expertise, particularly when working with larger datasets for both data organisation, upload and external validation. System crashes have also been reported with attempts to upload larger datasets [111]. External validation is a critical component of the model development pipeline, however, this facility is not offered by all platforms. It has previously been reported that only two commercially available code-free deep learning platforms feature external testing (Google and MedicMind) [110,111,119]. Santomartino et al. report, however, that this feature remains limited for both with both systems crashing and coding expertise required for Google. In that respect, the authors conclude that “although code-free deep learning platforms present a compelling use case, our results indicate that they are not currently suitable for clinical use or easily accessible to clinicians without coding expertise.” The authors acknowledge greater accessibility associated with image classification models but report nonetheless that coding was still required [111].

Finally, though less expensive than bespoke deep learning algorithms, these platforms are commercial. There are typically free tiers or credits which progress into paid tiers once the credits have been used up. Although the costs reported within the clinical literature have been relatively low, costing approximately $100 per model [111], this can still represent a financial challenge to resource-limited research groups and hospitals. Even if the initial training of the model is moderate in cost, long-term usage, particularly cloud-hosting, can lead to significant expenses for the user [137]. In addition, the initial task of dataset curation can require significant resources as can data scientist input if required for dataset assistance or model interpretation [109].

4.3.2.3 Stakeholder acceptance - patient and clinician

For any form of deep learning to be successfully implemented into the clinical care pathway, stakeholder acceptance is essential. Without appropriate education of clinicians, it is difficult to expect support and mistrust is likely to prevail. Asan et al. list the following as key considerations in a stakeholder's trust in artificial intelligence: user education, past experiences, user biases, properties of artificial intelligence systems, transparency and associated risks. The authors emphasise that reliability will play a key role in securing or losing trust of clinicians and the general public. Ejaz et al. also make this observation that the issue of racial bias within deep learning algorithms has impacted stakeholder trust in clinical artificial intelligence which may particularly be an issue in lower income countries less-represented in artificial intelligence datasets. Conversely, it is not preferable that clinicians completely trust these algorithms. It is important that the models are continually critically appraised and that human judgement continues to play a role in patient care [171,172]. Therefore, it is believable that by maximising clinician education in clinical AI, healthcare workers will have the skills needed to confidently assess and appraise a deep learning system and autonomously decide whether or not to utilise it within their patient care pathway.

4.3.2.3 Limited education

At present, medical students and clinicians receive little to no training in clinical artificial intelligence [173,174]. Therefore, despite the fact that clinicians are 'use-case experts' in the development of useful and appropriate deep learning models for patients, their input remains limited and dependent on those with the necessary skills. In addition to this, there is a concern that limited knowledge may lead to the development of poorly designed deep learning models or that the results and performances of models trained using code-free deep learning may be incorrectly interpreted with incorrect conclusions. On one end of the scale, this could lead to poor use of resources but on the other end if implemented into direct patient care, they could have very significant consequences. It has already been discussed in detail the critical importance in curating a robust dataset reflective of the target population and the limited publicly-available datasets accessible to researchers. Users of code-free deep learning must be cognisant of their local governance guidelines and ensure that sensitive patient information is stored securely and utilised in line with regulations [105]. Clinicians also must be made aware of class imbalance problems both when curating datasets and training models. As outlined by Antaki et al. class imbalance problems may particularly arise with low-incidence diseases and

clinicians must be aware of when code-free deep learning is appropriate and when data scientist input is more appropriate [134]. Clinicians must be aware of issues such as discriminatory bias whereby the model will preferentially represent the majority class leading to poorer performance with under-represented populations. This can lead to overfitting when a non-representative dataset is used for a different population. Clinicians must be conscious that robust external validation is a critical part of the model development pipeline in order to truly assess the model performance [109].Clinical artificial intelligence must be covered in the curricula for medical schools and postgraduate training in order to both equip clinicians with skills and knowledge needed to navigate clinical artificial intelligence in general and to be able to efficiently devise and sufficiently interpret code-free deep learning models they independently develop.

4.4 Future research

This research demonstrates that code-free deep learning models can be developed by those without coding expertise to classify optic disc abnormalities. There are various directions this research could take in the future both specifically building on the work published in this thesis but also for code-free deep learning in general.

4.4.1 Future directions of this research

This research will be expanded upon to:

1. Compare the results of the code-free deep learning models against a bespoke deep learning model trained on the exact same datasets used within this thesis and
2. To utilise Dataset I to trial a new deep learning model designed to carry out image segmentation.

1. Dataset I: code-free deep learning versus bespoke deep learning

There are numerous potential applications for this dataset which was developed and curated as part of this thesis. As outlined, it is now a valuable resource containing thousands of labelled images from real-world patient data in a tertiary ophthalmic referral centre. In the first instance, further work is planned to compare how the code-free deep learning models described in this thesis perform against bespoke deep learning models trained on the same dataset to perform the same task. It has been outlined that the bespoke deep learning model published by Milea et al. showed superior performance to those described within this research, however, a direct comparison would be more valuable using the same training dataset for both. This may also further identify any additional areas of weakness within the dataset that could potentially be addressed. It would also be interesting to examine deep learning performance on a cohort of patients diagnosed and managed specifically within a neuro-ophthalmology clinic from the dataset including clinical meta-data such as age and gender. This would all require considerable data scientist input, therefore, external validation via batch prediction of the code-free deep learning models could also be performed for a comprehensive assessment of the model performance.

2. Segment Anything Model

Dataset I will also be utilised in research separate to code-free deep learning. Recently, deep learning has been applied to the task of image segmentation in medical imaging. Segmentation describes the identification and delineation of specific objects or features within an image. This could be anything in healthcare such as the identification and delineation of a naevus on a skin image or a pacemaker within a chest x-ray. Previously, this was performed manually which, similar to dataset labelling, is a time consuming task requiring expertise in the area. The automation of this process through the use of deep learning removes the burden of this task from experts allowing them to focus their efforts elsewhere [175]. The Segment Anything Model is the first deep learning algorithm of its type that aims to perform the task of image segmentation [176]. Future research will involve the application of the Segment Anything Model to Dataset I in order to perform optic disc segmentation. This may be useful in a number of different clinical scenarios within ophthalmology but particularly for glaucoma detection and monitoring.

4.4.2 Future directions of code-free deep learning

Code-free deep learning may take a number of different directions in the future. Herein, its potential applications will be discussed.

4.4.2.1 Medical education

As discussed, limited education in the field of clinical artificial intelligence poses a significant challenge to clinicians who will play a role of key stakeholder in any models implemented into future practice. This has gained recognition in recent years with many publications within the clinical literature detailing the need for artificial intelligence integration into medical school curricula in order to improve patient care [177–179].

The density of medical school curricula would not have space to include in-depth data science modules, however, code-free deep learning may serve as a useful tool through which this education can be facilitated. In 2022, Ejaz et al. published the results of a study of medical students' perspectives of artificial intelligence and medical education. The authors advised that curricula should prioritise clinical applications of artificial intelligence while also including students in the process of model development [171]. Code-free deep learning as a concept tool can teach users valuable lessons on the entire pipeline of model development. Students can be provided with a publicly available dataset, asked to appraise the quality and robustness of this dataset and then train a model and accurately interpret its performance. Exercises such as this would provide invaluable practical experience in the steps of algorithm development. This has also been observed by Thirunavukarasu et al. who outline that code-free deep learning "permits more learners to explore ideas practically rather than merely discussing them in theory. Learners can actively produce and modify models to demonstrate the importance of data set quality, validation, and explainability for themselves." [104]. In addition, users can independently learn to recognise problems associated with deep learning models such as inherent biases, limitations in poorly represented groups and the issues of explainability, as discussed.

Code-free deep learning could also be used directly to progress medical training. As discussed by Lazarus et al. artificial intelligence education could provide students with attention not feasible to human educators in terms of speed, scale, personalised approaches and consistency [180]. Educators could devise code-free deep learning models to monitor surgical progress, performance and complication rates [181] as outlined by Touma et al. who suggest that

the development of models to segment surgical videos into phases could provide a valuable training resource whereby ophthalmologists could review videos of a specific performance area they wish to improve [133]. Medical students could independently develop algorithms to assist when learning to identify various anatomical or pathological entities [182]. Artificial intelligence has also been applied to assist with post-graduate teaching [183] and, in time, clinicians could independently develop code-free deep learning tools to target areas of personal weakness within their training programmes.

4.4.2.2 Clinical research

Code-free deep learning may prove to be particularly useful in the field of clinical research. It allows researchers to develop a proof-of-concept tool prior to the development of a bespoke deep learning system. This could be useful in order to assess dataset robustness, or applicability of deep learning to a use case. Demonstration of a useful code-free deep learning model could assist researchers in applying for investment for the greater financial resources required for a bespoke system as well as equipment and personnel. Nonetheless, all research must still adhere to the same rigorous benchmarks expected of bespoke deep learning systems which are both transparent and adherent to governance regulations such as Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence (DECIDE-AI) [184]. Code-free deep learning may also expose weaknesses in a dataset prior to investing significant resources in the development of bespoke deep learning. This could allow more targeted investment in improving the robustness of the dataset or increasing balance of a certain class if necessary. Finally, deep learning models have the ability to detect unique patterns within images unrecognisable to humans [122]. This represents a huge potential in the advancement of disease recognition whereby a model can detect imperceptible changes within an image that could indicate the development or progression of a specific pathology [121]. By providing clinicians with the tools to develop their own code-free deep learning tools, the scope of application to a huge host of diseases is potentially vast.

4.4.2.3 Direct patient care

Due to the discussed limitations, it is unlikely that code-free deep learning will play a role in direct patient care in the near future. Nonetheless, the clinical literature has demonstrated a number of code-free deep learning models with comparable performance to bespoke algorithms trained on the same dataset [119,121] which is why that will form part of the future research arising from this thesis. Although explainability remains a critical issue in the deployment of these models, considerable work is being done in this area of artificial intelligence for bespoke deep learning [116,185,186], which, in time, may also prove useful in interpreting code-free deep learning models. As discussed, for safe and effective deployment into direct patient care, code-free deep learning models must receive approval for their intended use which ensures robustness and regulatory compliance [104]. These models must also be integrated into the clinical workflow in a manner that enhances efficiency rather than burdening the user with additional administrative tasks.

It is difficult to envision a situation whereby highly specialised data science expertise and computing resources could be outrivalled by code-free deep learning, however, as observed by Wagner et al. code-free deep learning could play a role in areas where resources are scarce. The authors outline how in diseases such as retinopathy of prematurity, disease natural history varies between ethnicities. Locally-trained code-free deep learning models might allow clinicians to develop models suited directly for their target population. These models, in turn, could be deployed on edge devices, further reducing the associated costs [125]. Another area they could be used in the clinical setting is in tracking of resources. Unadkat et al. discuss the application of code-free deep learning for surgical instrument tracking for robotic systems [137]. Banerjee et al. outline the importance in involving both clinicians and patients in the development of deep learning algorithms in order to harness trust in these systems. By providing clinicians with the ability to develop personalised models for their patients, this can involve all stakeholders in the algorithm development process and may, in time, improve trust and acceptability where appropriate [187].

4.5 Conclusion

The findings of this research indicate that bespoke deep learning may be useful in the screening for papilloedema and/or optic disc abnormalities but that code-free deep learning is not an appropriate alternative for this task at present. This thesis closely examines the use of both bespoke deep learning and code-free deep learning in the two separate tasks of 1) classification of papilloedema versus no papilloedema and, 2) classification of normal versus abnormal optic discs, using two different datasets and two different imaging modalities.

It is evident that bespoke deep learning can predict the presence or absence of papilloedema and/or the presence or absence of optic disc abnormalities. This is supported by the high sensitivity, specificity and accuracy reported in all bespoke deep learning records identified in the systematic review. The datasets differed in size and label distribution. The deep learning approach also varied between records, with those using the BONSAI deep learning system [3,155,157] achieving higher performance metrics. The record with the largest dataset exhibited the best performance [3]. There were no records identified within the clinical literature that examined the use of code-free deep learning for the defined tasks within this thesis. Therefore, the results from this research could not be directly compared against the performance of other code-free deep learning models previously described. As outlined in the Introduction, code-free deep learning has demonstrated impressive results when trained on other tasks [111,112,114,119,121,131,132].

The code-free deep learning models trained in this thesis demonstrated promising performances on initial validation testing. However, the results from external validation established that code-free deep learning performed poorly in comparison to bespoke deep learning. This applies to both the task of differentiating between normal and abnormal optic discs and the presence or absence of papilloedema. It must be acknowledged that the datasets used within this thesis are not particularly comparable with those described within the clinical literature. The optical coherence tomography dataset described by Nowomiejska et al. contained a large proportion of artificially curated images [152]. Conversely, code-free deep learning models three, four, seven, eight, ten, eleven, thirteen and fourteen were trained on Dataset I which was acquired from real-world patient data. This is more likely to contain artefacts or variation in the quality of imaging. Furthermore, the images from Dataset I were acquired from a variety of different imaging systems which may also impact the performance of the code-free deep learning model.

The performance of code-free deep learning in the classification of colour fundus photographs varied greatly between datasets. As evidenced, the worst performance came from models one, two, five and six which were trained on Dataset I. Performance improved following random undersampling to achieve sensitivity closer to those described within the clinical literature of the bespoke deep learning models. Only code-free deep learning models fifteen and sixteen achieved performance equal to or greater than the bespoke deep learning models for both papilloedema versus no papilloedema, and normal versus abnormal classification. These models were trained on Dataset II which came from a publicly-available dataset which was well-balanced in terms of label distribution. Nonetheless, models fifteen and sixteen similarly displayed poor predictive accuracy when tested on external validation meaning the dataset cannot be solely implicated in the performance of both models.

It must be concluded that code-free deep learning should not be considered an equal alternative to bespoke deep learning at present. This is particularly true for models that are potentially to be implemented into the direct patient pathway. Nonetheless, code-free deep learning offers significant promise within the field of artificial intelligence in healthcare for both clinical research and medical education. As demonstrated within this thesis, it allows clinicians without coding expertise to design and develop deep learning models for a clinical use case of their choosing. It also provides educators with a low-resource tool with which to enlighten future clinicians on the various steps in the path towards deep learning model development. Deep learning is likely to play a major role in the advancements of healthcare. However, advancement is not truly effective without input from specialty experts. Equipping clinicians with a core understanding of the deep learning development process will allow for meaningful collaboration with data scientists to further expedite and enrich the future of healthcare.

References

1. Dunn HP, Kang CJ, Marks S, Dunn SM, Healey PR, White AJ. Optimising fundoscopy practices across the medical spectrum: A focus group study. *PLoS One*. 2023;18(1):e0280937.

2. Hospital Outpatient Activity, 2022-2023: All attendances. NHS England. September 2023. Accessed January 2024. https://files.digital.nhs.uk/C9/49D3E3/hosp-epis-stat-outp-all-atte-2022-23-tab.xlsx

3. Milea D, Najjar RP, Zhubo J, et al. Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs. *N Engl J Med*. 2020;382(18):1687-1695.

4. Jones CH, Dolsten M. Healthcare on the brink: navigating the challenges of an aging society in the United States. *NPJ Aging*. 2024;10(1):22.

5. Vital statistics yearly summary 2022. CSO. Accessed June 21, 2024. https://www.cso.ie/en/releasesandpublications/ep/p-vsys/vitalstatisticsyearlysummary2022/

6. Births in the UK 2021. Statista. Accessed June 21, 2024. https://www.statista.com/statistics/281981/live-births-in-the-united-kingdom-uk/

7. Cristea M, Noja GG, Stefea P, Sala AL. The Impact of Population Aging and Public Health Support on EU Labor Markets. *Int J Environ Res Public Health*. 2020;17(4). doi:10.3390/ijerph17041439

8. Fleckenstein M, Keenan TDL, Guymer RH, et al. Age-related macular degeneration. *Nat Rev Dis Primers*. 2021;7(1):31.

9. Thomas CJ, Mirza RG, Gill MK. Age-Related Macular Degeneration. *Med Clin North Am*. 2021;105(3):473-491.

10. Li JQ, Welchowski T, Schmid M, Mauschitz MM, Holz FG, Finger RP. Prevalence and incidence of age-related macular degeneration in Europe: a systematic review and meta-analysis. *Br J Ophthalmol*. 2020;104(8):1077-1084.

11. National treatment purchase fund (NTPF). Accessed July 24, 2024. https://www.ntpf.ie/home/inpatient.htm

12. NHS England publishes guidance on reducing “did not attends” in outpatient services. The Royal College of Ophthalmologists. February 22, 2023. Accessed June 21, 2024. https://www.rcophth.ac.uk/news-views/nhs-england-publishes-guidance-on-reducing-did-not-attends-in-outpatient-services/

13. Marengoni A, Angleman S, Melis R, et al. Aging with multimorbidity: A systematic review of the literature. *Ageing Res Rev*. 2011;10(4):430-439.

14. Holman HR. The Relation of the Chronic Disease Epidemic to the Health Care Crisis. *ACR Open Rheumatol*. 2020;2(3):167-173.

15. Magliano DJ, Boyko EJ, IDF Diabetes Atlas 10th edition scientific committee. *IDF*

DIABETES ATLAS. International Diabetes Federation; 2021.

16. Hajat C, Stein E. The global burden of multiple chronic conditions: A narrative review. *Prev Med Rep*. 2018;12:284-293.

17. Chang D, Chang X, He Y, Tan KJK. The determinants of COVID-19 morbidity and mortality across countries. *Sci Rep*. 2022;12(1):5888.

18. The true death toll of COVID-19 estimating global excess mortality. Accessed June 19, 2024. https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality

19. Lopez-Leon S, Wegman-Ostrosky T, Perelman C, et al. More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. *Sci Rep*. 2021;11(1):16144.

20. Olmastroni E, Galimberti F, Tragni E, Catapano AL, Casula M. Impact of COVID-19 Pandemic on Adherence to Chronic Therapies: A Systematic Review. *Int J Environ Res Public Health*. 2023;20(5). doi:10.3390/ijerph20053825

21. Arsenault C, Gage A, Kim MK, et al. COVID-19 and resilience of healthcare systems in ten countries. *Nat Med*. 2022;28(6):1314-1324.

22. Transcript. IMF. January 28, 2021. Accessed June 19, 2024. https://www.imf.org/en/News/Articles/2021/01/28/tr012621-transcript-of-the-world-economic-outlook-update-press-briefing

23. Maslach C, Jackson SE, Leiter MP. Maslach Burnout Inventory: Third edition. In: Zalaquett CP, ed. *Evaluating Stress: A Book of Resources (pp*. Vol 474. Scarecrow Education, xvii; 1997:191-218.

24. Shanafelt TD, West CP, Dyrbye LN, et al. Changes in Burnout and Satisfaction With Work-Life Integration in Physicians During the First 2 Years of the COVID-19 Pandemic. *Mayo Clin Proc*. 2022;97(12):2248-2258.

25. Ortega MV, Hidrue MK, Lehrhoff SR, et al. Patterns in Physician Burnout in a Stable-Linked Cohort. *JAMA Netw Open*. 2023;6(10):e2336745.

26. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. *J Intern Med*. 2018;283(6):516-529.

27. National Academies of Sciences, Engineering, and Medicine; National Academy of Medicine; Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being. *Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being*. National Academies Press (US); 2019.

28. Welles CC, Tong A, Brereton E, et al. Sources of Clinician Burnout in Providing Care for Underserved Patients in a Safety-Net Healthcare System. *J Gen Intern Med*. 2023;38(6):1468-1475.

29. Nash L, Walton M, Daly M, et al. GPs' concerns about medicolegal issues - How it affects their practice. *Aust Fam Physician*. 2009;38(1-2):66-70.

30. Rigi M, Almarzouqi SJ, Morgan ML, Lee AG. Papilledema: epidemiology, etiology, and

clinical management. *Eye Brain*. 2015;7:47-57.

31. Pinto VL, Tadi P, Adeyinka A. Increased Intracranial Pressure. In: *StatPearls*. StatPearls Publishing; 2022.

32. Donaldson L, Margolin E. Absence of papilledema in large intracranial tumours. *J Neurol Sci*. 2021;428:117604.

33. Killer HE, Jaggi GP, Miller NR. Papilledema revisited: is its pathophysiology really understood? *Clin Experiment Ophthalmol*. 2009;37(5):444-447.

34. Benson JC, Madhavan AA, Cutsforth-Gregory JK, Johnson DR, Carr CM. The Monro-Kellie Doctrine: A Review and Call for Revision. *AJNR Am J Neuroradiol*. 2023;44(1):2-6.

35. Xie JS, Donaldson L, Margolin E. Papilledema: A review of etiology, pathophysiology, diagnosis, and management. *Surv Ophthalmol*. 2022;67(4):1135-1159.

36. Telano LN, Baker S. Physiology, Cerebral Spinal Fluid. In: *StatPearls*. StatPearls Publishing; 2023.

37. Wilson MH. Monro-Kellie 2.0: The dynamic vascular and venous pathophysiological components of intracranial pressure. *J Cereb Blood Flow Metab*. 2016;36(8):1338-1350.

38. Passi N, Degnan AJ, Levy LM. MR imaging of papilledema and visual pathways: effects of increased intracranial pressure and pathophysiologic mechanisms. *AJNR Am J Neuroradiol*. 2013;34(5):919-924.

39. Hayreh SS. Posterior ciliary artery circulation in health and disease: the Weisenfeld lecture. *Invest Ophthalmol Vis Sci*. 2004;45(3):749-757; 748.

40. Schirmer CM, Hedges TR 3rd. Mechanisms of visual loss in papilledema. *Neurosurg Focus*. 2007;23(5):E5.

41. Wang MTM, Bhatti MT, Danesh-Meyer HV. Idiopathic intracranial hypertension: Pathophysiology, diagnosis and management. *J Clin Neurosci*. 2022;95:172-179.

42. Crum OM, Kilgore KP, Sharma R, et al. Etiology of Papilledema in Patients in the Eye Clinic Setting. *JAMA Netw Open*. 2020;3(6):e206625.

43. Foley J. BENIGN FORMS OF INTRACRANIAL HYPERTENSION—"TOXIC" AND "OTITIC" HYDROCEPHALUS. *Brain*. 1955;78(1):1-41.

44. Wall M. Idiopathic intracranial hypertension. *Neurol Clin*. 2010;28(3):593-617.

45. Manfield JH, Yu KKH, Efthimiou E, Darzi A, Athanasiou T, Ashrafian H. Bariatric Surgery or Non-surgical Weight Loss for Idiopathic Intracranial Hypertension? A Systematic Review and Comparison of Meta-analyses. *Obes Surg*. 2017;27(2):513-521.

46. Abdelghaffar M, Hussein M, Abdelkareem SA, Elshebawy H. Sex hormones, CSF and serum leptin in patients with idiopathic intracranial hypertension. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*. 2022;58(1):39.

47. Boyter E. Idiopathic intracranial hypertension. *JAAPA*. 2019;32(5):30-35.

48. Wall M, Kupersmith MJ, Kieburtz KD, et al. The idiopathic intracranial hypertension treatment trial: clinical profile at baseline. *JAMA Neurol*. 2014;71(6):693-701.

49. Mollan SP, Markey KA, Benzimra JD, et al. A practical approach to, diagnosis, assessment and management of idiopathic intracranial hypertension. *Pract Neurol*. 2014;14(6):380-390.

50. Friedman DI, Jacobson DM. Diagnostic criteria for idiopathic intracranial hypertension. *Neurology*. 2002;59(10):1492-1495.

51. Prokop K, Opęchowska A, Sieśkiewicz A, Lisowski Ł, Mariak Z, Łysoń T. Effectiveness of optic nerve sheath fenestration in preserving vision in idiopathic intracranial hypertension: an updated meta-analysis and systematic review. *Acta Neurochir (Wien)*. 2024;166(1):476.

52. Pineles SL, Volpe NJ. Long-Term Results of Optic Nerve Sheath Fenestration for Idiopathic Intracranial Hypertension: Earlier Intervention Favours Improved Outcomes. *Neuroophthalmology*. 2013;37(1):12-19.

53. Almasoud F, Alabduljabbar A, Alotaibi A, et al. Efficacy and Safety of Optic Nerve Sheath Fenestration for Idiopathic Intracranial Hypertension. A Subgroup-Focused Systematic Review and Meta-Analysis. *ASIDE Intern Med*. 2025;1(3):7-14.

54. Liu KC, Bhatti MT, Chen JJ, et al. Presentation and Progression of Papilledema in Cerebral Venous Sinus Thrombosis. *Am J Ophthalmol*. 2020;213:1-8.

55. Capecchi M, Abbattista M, Martinelli I. Cerebral venous sinus thrombosis. *J Thromb Haemost*. 2018;16(10):1918-1931.

56. Park MG, Roh J, Ahn SH, Park KP, Baik SK. Papilledema and venous stasis in patients with cerebral venous and sinus thrombosis. *BMC Neurol*. 2023;23(1):175.

57. Ferro JM, Canhão P, Stam J, Bousser MG, Barinagarrementeria F, ISCVT Investigators. Prognosis of cerebral vein and dural sinus thrombosis: results of the International Study on Cerebral Vein and Dural Sinus Thrombosis (ISCVT). *Stroke*. 2004;35(3):664-670.

58. Ferro JM, Canhão P, Stam J, et al. Delay in the diagnosis of cerebral vein and dural sinus thrombosis: influence on outcome. *Stroke*. 2009;40(9):3133-3138.

59. Filippidis A, Kapsalaki E, Patramani G, Fountas KN. Cerebral venous sinus thrombosis: review of the demographics, pathophysiology, current diagnosis, and treatment. *Neurosurg Focus*. 2009;27(5):E3.

60. Lamonte M, Silberstein SD, Marcelis JF. Headache associated with aseptic meningitis. *Headache*. 1995;35(9):520-526.

61. Hanna LS, Girgis NI, Yassin MW, et al. The incidence of optic atrophy in meningitis. *Bull Ophthalmol Soc Egypt*. 1978;71(75):113-118.

62. Naranjo M, Chauhan S, Paul M. Malignant hypertension. Published online 2018. https://europepmc.org/article/nbk/nbk507701

63. Tajunisah I, Patel DK. Malignant hypertension with papilledema. *J Emerg Med*. 2013;44(1):164-165.

64. Mollan SP, Markey KA, Benzimra JD, et al. A practical approach to, diagnosis, assessment and management of idiopathic intracranial hypertension. *Pract Neurol*. 2014;14(6):380-390.

65. Gili P, Flores-Rodríguez P, Yangüela J, Herreros Fernández ML. Using autofluorescence to detect optic nerve head drusen in children. *J AAPOS*. 2013;17(6):568-571.

66. Lee KM, Woo SJ, Hwang JM. Differentiation of optic nerve head drusen and optic disc edema with spectral-domain optical coherence tomography. *Ophthalmology*. 2011;118(5):971-977.

67. Sarac O, Tasci YY, Gurdal C, Can I. Differentiation of optic disc edema from optic nerve head drusen with spectral-domain optical coherence tomography. *J Neuroophthalmol*. 2012;32(3):207-211.

68. Pineles SL, Arnold AC. Fluorescein angiographic identification of optic disc drusen with and without optic disc edema. *J Neuroophthalmol*. 2012;32(1):17-22.

69. Atta HR. Imaging of the optic nerve with standardised echography. *Eye* . 1988;2 (Pt 4):358-366.

70. Kovarik JJ, Doshi PN, Collinge JE, Plager DA. Outcome of pediatric patients referred for papilledema. *J AAPOS*. 2015;19(4):344-348.

71. McCafferty B, McClelland CM, Lee MS. The diagnostic challenge of evaluating papilledema in the pediatric patient. *Taiwan J Ophthalmol*. 2017;7(1):15-21.

72. Dahlmann-Noor AH, Adams GW, Daniel MC, et al. Detecting optic nerve head swelling on ultrasound and optical coherence tomography in children and young people: an observational study. *Br J Ophthalmol*. 2018;102(3):318-322.

73. Blanch RJ, Horsburgh J, Creavin A, DOPS Study Group, Burdon MA, Williams C. Detection of Papilloedema Study (DOPS): rates of false positive papilloedema in the community. *Eye* . 2019;33(7):1073-1080.

74. Sibony PA, Kupersmith MJ, Kardon RH. Optical Coherence Tomography Neuro-Toolbox for the Diagnosis and Management of Papilledema, Optic Disc Edema, and Pseudopapilledema. *J Neuroophthalmol*. 2021;41(1):77-92.

75. Chang MY, Velez FG, Demer JL, et al. Accuracy of Diagnostic Imaging Modalities for Classifying Pediatric Eyes as Papilledema Versus Pseudopapilledema. *Ophthalmology*. 2017;124(12):1839-1848.

76. Mackay DD, Garza PS, Bruce BB, Newman NJ, Biousse V. The demise of direct ophthalmoscopy: A modern clinical challenge. *Neurol Clin Pract*. 2015;5(2):150-157.

77. Kelly LP, Garza PS, Bruce BB, Graubart EB, Newman NJ, Biousse V. Teaching ophthalmoscopy to medical students (the TOTeMS study). *Am J Ophthalmol*. 2013;156(5):1056-1061.e10.

78. Mackay DD, Garza PS, Bruce BB, et al. Teaching ophthalmoscopy to medical students (TOTeMS) II: A one-year retention study. *Am J Ophthalmol*. 2014;157(3):747-748.

79. Fisayo A, Bruce BB, Newman NJ, Biousse V. Overdiagnosis of idiopathic intracranial

hypertension. *Neurology*. 2016;86(4):341-350.

80. Ang GS, Dhillon B. Do junior house officers routinely test visual acuity and perform ophthalmoscopy? *Scott Med J*. 2002;47(3):60-63.

81. Cordeiro MF, Jolly BC, Dacre JE. The effect of formal instruction in ophthalmoscopy on medical student performance. *Med Teach*. 1993;15(4):321-325.

82. Shuttleworth GN, Marsh GW. How effective is undergraduate and postgraduate teaching in ophthalmology? *Eye* . 1997;11 (Pt 5):744-750.

83. Roberts E, Morgan R, King D, Clerkin L. Funduscopy: a forgotten art? *Postgrad Med J*. 1999;75(883):282-284.

84. Bruce BB, Lamirel C, Biousse V, et al. Feasibility of nonmydriatic ocular fundus photography in the emergency department: Phase I of the FOTO-ED study. *Acad Emerg Med*. 2011;18(9):928-933.

85. Lamirel C, Bruce BB, Wright DW, Delaney KP, Newman NJ, Biousse V. Quality of nonmydriatic digital fundus photography obtained by nurse practitioners in the emergency department: the FOTO-ED study. *Ophthalmology*. 2012;119(3):617-624.

86. Bruce BB, Thulasi P, Fraser CL, et al. Diagnostic accuracy and use of nonmydriatic ocular fundus photography by emergency physicians: phase II of the FOTO-ED study. *Ann Emerg Med*. 2013;62(1):28-33.e1.

87. Poostchi A, Awad M, Wilde C, Dineen RA, Gruener AM. Spike in neuroimaging requests following the conviction of the optometrist Honey Rose. *Eye* . 2018;32(3):489-490.

88. Xu Y, Liu X, Cao X, et al. Artificial intelligence: A powerful paradigm for scientific research. *Innovation (Camb)*. 2021;2(4):100179.

89. Turing AM. Computing Machinery and Intelligence. In: Epstein R, Roberts G, Beber G, eds. *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Netherlands; 2009:23-65.

90. Moor J. The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AIMag*. 2006;27(4):87-87.

91. Jones LD, Golan D, Hanna SA, Ramachandran M. Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern? *Bone Joint Res*. 2018;7(3):223-225.

92. Fei-Fei L, Deng J, Li K. ImageNet: Constructing a large-scale image database. *J Vis*. 2009;9(8):1037-1037.

93. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis*. 2015;115(3):211-252.

94. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*. 2012;25:1097-1105.

95. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural*

Comput. 2006;18(7):1527-1554.

96. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.

97. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.

98. Venkatesh KP, Raza MM, Nickel G, Wang S, Kvedar JC. Deep learning models across the range of skin disease. *NPJ Digit Med*. 2024;7(1):32.

99. Litjens G, Ciompi F, Wolterink JM, et al. State-of-the-Art Deep Learning in Cardiovascular Image Analysis. *JACC Cardiovasc Imaging*. 2019;12(8 Pt 1):1549-1565.

100. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv [csCV]*. Published online November 14, 2017. http://arxiv.org/abs/1711.05225

101. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342-1350.

102. Placido D, Yuan B, Hjaltelin JX, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat Med*. 2023;29(5):1113-1122.

103. Volinsky-Fremond S, Horeweg N, Andani S, et al. Author Correction: Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nat Med*. 2024;30(7):2092.

104. Thirunavukarasu AJ, Elangovan K, Gutierrez L, et al. Democratizing Artificial Intelligence Imaging Analysis With Automated Machine Learning: Tutorial. *J Med Internet Res*. 2023;25:e49949.

105. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*. 2021;3(1):e51-e66.

106. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. *Patterns (N Y)*. 2021;2(10):100347.

107. Escalante HJ. Automated Machine Learning—A Brief Review at the End of the Early Years. In: Pillay N, Qu R, eds. *Automated Design of Machine Learning and Search Algorithms*. Springer International Publishing; 2021:11-28.

108. Shen Z, Zhang Y, Wei L, Zhao H, Yao Q. Automated Machine Learning: From Principles to Practices. *arXiv e-prints*. Published online October 1, 2018:arXiv:1810.13306.

109. O'Byrne C, Abbas A, Korot E, Keane PA. Automated deep learning in ophthalmology: AI that can build AI. *Curr Opin Ophthalmol*. Published online July 6, 2021. doi:10.1097/ICU.0000000000000779

110. Korot E, Guan Z, Ferraz D, et al. Code-free Deep Learning for Multi-Modality Medical Image Classification. *Nature Machine Intelligence*.

111. Santomartino SM, Hafezi-Nejad N, Parekh VS, Yi PH. Performance and Usability of Code-

Free Deep Learning for Chest Radiograph Classification, Object Detection, and Segmentation. *Radiol Artif Intell*. 2023;5(2):e220062.

112. Huemer J, Kronschläger M, Ruiss M, et al. Diagnostic accuracy of code-free deep learning for detection and evaluation of posterior capsule opacification. *BMJ Open Ophthalmol*. 2022;7(1). doi:10.1136/bmjophth-2022-000992

113. Jacoba CMP, Doan D, Salongcay RP, et al. Performance of Automated Machine Learning for Diabetic Retinopathy Image Classification from Multi-field Handheld Retinal Images. *Ophthalmol Retina*. 2023;7(8):703-712.

114. Touma S, Hammou BA, Antaki F, Boucher MC, Duval R. Comparing code-free deep learning models to expert-designed models for detecting retinal diseases from optical coherence tomography. *Int J Retina Vitreous*. 2024;10(1):37.

115. Zoph B, Le QV. Neural Architecture Search with Reinforcement Learning. *arXiv [csLG]*. Published online November 5, 2016. http://arxiv.org/abs/1611.01578

116. Siontis KC, Suárez AB, Sehrawat O, et al. Saliency maps provide insights into artificial intelligence-based electrocardiography models for detecting hypertrophic cardiomyopathy. *J Electrocardiol*. 2023;81:286-291.

117. Arun N, Gaw N, Singh P, et al. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiol Artif Intell*. 2021;3(6):e200267.

118. Merenda M, Porcaro C, Iero D. Edge Machine Learning for AI-Enabled IoT Devices: A Review. *Sensors* . 2020;20(9). doi:10.3390/s20092533

119. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digit Health*. 2019;1(5):e232-e242.

120. Kim IK, Lee K, Park JH, Baek J, Lee WK. Classification of pachychoroid disease on ultrawide-field indocyanine green angiography using auto-machine learning platform. *Br J Ophthalmol*. Published online July 3, 2020. doi:10.1136/bjophthalmol-2020-316108

121. Korot E, Pontikos N, Liu X, et al. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep*. 2021;11(1):10286.

122. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158-164.

123. Vasavada AR, Raj SM, Shah GD, Nanavaty MA. Posterior capsule opacification after lens implantation: incidence, risk factors and management. *Expert Rev Ophthalmol*. 2013;8(2):141-149.

124. Dammann O, Hartnett ME, Stahl A. Retinopathy of prematurity. *Dev Med Child Neurol*. 2023;65(5):625-631.

125. Wagner SK, Liefers B, Radia M, et al. Development and international validation of custom-engineered and code-free deep-learning models for detection of plus disease in retinopathy of prematurity: a retrospective study. *Lancet Digit Health*. 2023;5(6):e340-e349.

126. Lim JI, Regillo CD, Sadda SR, et al. Artificial Intelligence Detection of Diabetic Retinopathy: Subgroup Comparison of the EyeArt System with Ophthalmologists' Dilated Examinations. *Ophthalmol Sci*. 2023;3(1):100228.

127. Borkowski AA, Viswanadhan NA, Thomas LB, Guzman RD, Deland LA, Mastorides SM. Using Artificial Intelligence for COVID-19 Chest X-ray Diagnosis. *Fed Pract*. 2020;37(9):398-404.

128. Borkowski AA, Wilson CP, Borkowski SA, et al. Google Auto ML versus Apple Create ML for Histopathologic Cancer Diagnosis; Which Algorithms Are Better? *arXiv [q-bioQM]*. Published online March 19, 2019. http://arxiv.org/abs/1903.08057

129. Zeng Y, Zhang J. A machine learning model for detecting invasive ductal carcinoma with Google Cloud AutoML Vision. *Comput Biol Med*. 2020;122:103861.

130. Koga S, Ghayal NB, Dickson DW. Deep Learning-Based Image Classification in Differentiating Tufted Astrocytes, Astrocytic Plaques, and Neuritic Plaques. *J Neuropathol Exp Neurol*. 2021;80(4):306-312.

131. Jungo P, Hewer E. Code-free machine learning for classification of central nervous system histopathology images. *J Neuropathol Exp Neurol*. 2023;82(3):221-230.

132. Ashraf AR, Somogyi-Végh A, Merczel S, Gyimesi N, Fittler A. Leveraging code-free deep learning for pill recognition in clinical settings: A multicenter, real-world study of performance across multiple platforms. *Artif Intell Med*. 2024;150:102844.

133. Touma S, Antaki F, Duval R. Development of a code-free machine learning model for the classification of cataract surgery phases. *Sci Rep*. 2022;12(1):2398.

134. Antaki F, Kahwati G, Sebag J, et al. Predictive modeling of proliferative vitreoretinopathy using automated machine learning by ophthalmologists without coding experience. *Sci Rep*. 2020;10(1):19528.

135. Abbas A, O'Byrne C, Fu DJ, et al. Evaluating an automated machine learning model that predicts visual acuity outcomes in patients with neovascular age-related macular degeneration. *Graefes Arch Clin Exp Ophthalmol*. 2022;260(8):2461-2473.

136. Wang S, Xu J, Tahmasebi A, et al. Incorporation of a Machine Learning Algorithm With Object Detection Within the Thyroid Imaging Reporting and Data System Improves the Diagnosis of Genetic Risk. *Front Oncol*. 2020;10:591846.

137. Unadkat V, Pangal DJ, Kugener G, et al. Code-free machine learning for object detection in surgical video: a benchmarking, feasibility, and cost study. *Neurosurg Focus*. 2022;52(4):E11.

138. Scott CJ, Kardon RH, Lee AG, Frisén L, Wall M. Diagnosis and grading of papilledema in patients with raised intracranial pressure using optical coherence tomography vs clinical expert assessment using a clinical staging scale. *Arch Ophthalmol*. 2010;128(6):705-711.

139. Cloud computing services. Google Cloud. Accessed August 27, 2024. https://cloud.google.com/

140. Kim U. Machine learning for Pseudopapilledema. Published online August 1, 2018.

doi:10.17605/OSF.IO/2W5CE

141. Install the gcloud CLI. Google Cloud. Accessed July 30, 2024. https://cloud.google.com/sdk/docs/install

142. cp - Copy files and objects. Google Cloud. Accessed August 27, 2024. https://cloud.google.com/storage/docs/gsutil/commands/cp

143. Vertex AI with Gemini 1.5 pro and Gemini 1.5 flash. Google Cloud. Accessed August 27, 2024. https://cloud.google.com/vertex-ai?hl=en

144. Montesinos López OA, Montesinos López A, Crossa J. Overfitting, Model Tuning, and Evaluation of Prediction Performance. In: Montesinos López OA, Montesinos López A, Crossa J, eds. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing; 2022:109-139.

145. Thölke P, Mantilla-Ramos YJ, Abdelhedi H, et al. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *Neuroimage*. 2023;277:120253.

146. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *SIGKDD Explor Newsl*. 2009;11(1):10-18.

147. Weka. Accessed August 27, 2024. https://sourceforge.net/projects/weka/files/weka-3-8/3.8.6/weka-3-8-6-azul-zulu-arm-osx.dmg/download?use_mirror=deac-fra

148. Shams R. Weka Tutorial 33: Random Undersampling (Class Imbalance Problem). December 12, 2013. Accessed August 27, 2024. https://www.youtube.com/watch?v=ocOlm73HeNs

149. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14(1):49-58.

150. Get batch predictions and explanations. Google Cloud. Accessed August 25, 2024. https://cloud.google.com/vertex-ai/docs/tabular-data/classification-regression/get-batch-predictions

151. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.

152. Nowomiejska K, Powroźnik P, Skublewska-Paszkowska M, et al. Residual Attention Network for distinction between visible optic disc drusen and healthy optic discs. *Opt Lasers Eng*. 2024;176:108056.

153. Diener R, Lauermann JL, Eter N, Treder M. Discriminating Healthy Optic Discs and Visible Optic Disc Drusen on Fundus Autofluorescence and Color Fundus Photography Using Deep Learning—A Pilot Study. *J Clin Med Res*. 2023;12(5):1951.

154. Avramidis K, Rostami M, Chang M, Narayanan S. Automating Detection of Papilledema in Pediatric Fundus Images with Explainable Machine Learning. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE; 2022:3973-3977.

155. Lin MY, Najjar RP, Tang Z, et al. The BONSAI (Brain and Optic Nerve Study with Artificial Intelligence) deep learning system can accurately identify pediatric papilledema on standard ocular fundus photographs. *J AAPOS*. 2024;28(1):103803.

156. Liu TYA, Wei J, Zhu H, et al. Detection of Optic Disc Abnormalities in Color Fundus Photographs Using Deep Learning. *J Neuroophthalmol*. 2021;41(3):368-374.

157. Biousse V, Najjar RP, Tang Z, et al. Application of a Deep Learning System to Detect Papilledema on Nonmydriatic Ocular Fundus Photographs in an Emergency Department. *Am J Ophthalmol*. 2024;261:199-207.

158. The Brain and Optic Nerve Study with Artificial Intelligence (BONSAI): A Deep Learning System to Detect Papilledema on Ocular Fundus Photography (2743). Neurology. Accessed August 21, 2024. https://www.neurology.org/doi/abs/10.1212/WNL.94.15_supplement.2743

159. Brookshire G, Kasper J, Blauch NM, et al. Data leakage in deep learning studies of translational EEG. *Front Neurosci*. 2024;18:1373515.

160. Bruce BB, Bidot S, Hage R, et al. Fundus Photography vs. Ophthalmoscopy Outcomes in the Emergency Department (FOTO-ED) Phase III: Web-based, In-service Training of Emergency Providers. *Neuroophthalmology*. 2018;42(5):269-274.

161. Garcea F, Serra A, Lamberti F, Morra L. Data augmentation for medical imaging: A systematic literature review. *Comput Biol Med*. 2023;152:106391.

162. Carmona EJ, Rincón M, García-Feijoó J, Martínez-de-la-Casa JM. Identification of the optic nerve head with genetic algorithms. *Artif Intell Med*. 2008;43(3):243-259.

163. Ludwig CA, Murthy SI, Pappuru RR, Jais A, Myung DJ, Chang RT. A novel smartphone ophthalmic imaging adapter: User feasibility studies in Hyderabad, India. *Indian J Ophthalmol*. 2016;64(3):191-200.

164. Krahulik D, Hrabalek L, Blazek F, et al. Sensitivity of Papilledema as a Sign of Increased Intracranial Pressure. *Children*. 2023;10(4). doi:10.3390/children10040723

165. Boran P, Oğuz F, Furman A, Sakarya S. Evaluation of fontanel size variation and closure time in children followed up from birth to 24 months. *J Neurosurg Pediatr*. 2018;22(3):323-329.

166. Wilne S, Collier J, Kennedy C, Koller K, Grundy R, Walker D. Presentation of childhood CNS tumours: a systematic review and meta-analysis. *Lancet Oncol*. 2007;8(8):685-695.

167. Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data. *IEEE Intell Syst*. March-April 2009;24(2):8-12.

168. Sanaat A, Shiri I, Ferdowsi S, Arabi H, Zaidi H. Robust-Deep: A Method for Increasing Brain Imaging Datasets to Improve Deep Learning Models' Performance and Robustness. *J Digit Imaging*. 2022;35(3):469-481.

169. Gao L, Zhang L, Liu C, Wu S. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. *Artif Intell Med*. 2020;108:101935.

170. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.

171. Ejaz H, McGrath H, Wong BL, Guise A, Vercauteren T, Shapey J. Artificial intelligence and medical education: A global mixed-methods study of medical students' perspectives. *Digit Health*. 2022;8:20552076221089099.

172. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res*. 2020;22(6):e15154.

173. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. *BMC Med Educ*. 2022;22(1):772.

174. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing Artificial Intelligence Training in Medical Education. *JMIR Med Educ*. 2019;5(2):e16048.

175. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. 2024;15(1):654.

176. Huang Y, Yang X, Liu L, et al. Segment anything model for medical images? *Med Image Anal*. 2024;92:103061.

177. Keane PA, Topol EJ. AI-facilitated health care requires education of clinicians. *Lancet*. 2021;397(10281):1254.

178. Topol E, Others. The topol review. *Preparing the healthcare workforce to deliver the digital future*. Published online 2019:1-48.

179. Weidener L, Fischer M. Artificial Intelligence Teaching as Part of Medical Education: Qualitative Analysis of Expert Interviews. *JMIR Med Educ*. 2023;9:e46428.

180. Lazarus MD, Truong M, Douglas P, Selwyn N. Artificial intelligence and clinical anatomical education: Promises and perils. *Anat Sci Educ*. 2024;17(2):249-262.

181. Satapathy P, Hermis AH, Rustagi S, Pradhan KB, Padhi BK, Sah R. Artificial intelligence in surgical education and training: opportunities, challenges, and ethical considerations - correspondence. *Int J Surg*. 2023;109(5):1543-1544.

182. Abdellatif H, Al Mushaiqri M, Albalushi H, Al-Zaabi AA, Roychoudhury S, Das S. Teaching, Learning and Assessing Anatomy with Artificial Intelligence: The Road to a Better Future. *Int J Environ Res Public Health*. 2022;19(21). doi:10.3390/ijerph192114209

183. Duong MT, Rauschecker AM, Rudie JD, et al. Artificial intelligence for precision education in radiology. *Br J Radiol*. 2019;92(1103):20190389.

184. Vasey B, Nagendran M, Campbell B, et al. Publisher Correction: Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28(10):2218.

185. Beqiri S, Abbas A, Korot E, Wagner S, Struyven R, Kelly M. Investigating the impact of saliency maps on clinician's confidence in model predictions.

186. Zihni E, Madai VI, Livne M, et al. Opening the black box of artificial intelligence for clinical

decision support: A study predicting stroke outcome. *PLoS One*. 2020;15(4):e0231166.

187. Banerjee S, Alsop P, Jones L, Cardinal RN. Patient and public involvement to build trust in artificial intelligence: A framework, tools, and case studies. *Patterns (N Y)*. 2022;3(6):100506.