UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Machine learning strategies for diagnostic imaging support on histopathology and optical coherence tomography

March, 2022

ELECTRONIC ENGINEERING DEPARTMENT

Author: José Gabriel García Pardo

Supervisors: Prof. Valery Naranjo Ornedo
Dr. Adrián Colomer Granero

Acknowledgements

José Ortega y Gasset acuñó en 1914 la frase "Yo soy yo y mi circunstancia" para referirse a que su vida en general estaba condicionada, en mayor o menor medida, por una serie de factores externos que no dependían directamente de él. Algo similar ocurre con esta tesis doctoral. El trabajo que aquí se recoge es fruto de la implicación de muchas personas que, por suerte, se han cruzado en mi camino para contribuir, de una forma u otra, a su desarrollo. Por eso, estas líneas van destinadas a reconocer y agradecer a todas esas personas que, en su paso, han dejado un poquito de ellas en esta tesis.

Me gustaría empezar dándole las gracias a Valery. Lo que viene después de esta página estaría en blanco de no ser por ella. Conocer a Valery fue, además de un privilegio, un punto de inflexión que decantó por completo mi vida profesional, haciendo de mi trabajo mi pasión. Para mí, siempre ha sido como una segunda madre, velando por mi bien y enseñándome cada día. Su inteligencia combinada con su inquietud, ilusión y personalidad es el cóctel perfecto del que espero continuar aprendiendo durante muchos años más. Muchas gracias por todo, Valery; este Gabri sería muy diferente sin ti.

También quiero agradecer a Adri, que siempre ha sido, además de mi codirector, un buen amigo y un espejo al que mirar. Él es quien ha seguido más de cerca cada avance, proponiendo siempre soluciones óptimas y aportando ese punto de calidad que hace de esta tesis un trabajo del que sentirse orgulloso. Recuerdo cuando, hace ya más de cuatro años, como cotutor de mi TFM, me dijo: "Juntos vamos a formar un buen equipo, ya lo verás". Bueno, no me corresponde a mí juzgar si lo hemos conseguido o no, pero sí puedo garantizar que esta tesis no habría sido la misma sin tu ayuda. Gracias, Adri.

Me gustaría también agradecer a todos mis compañeros del grupo CVBLab por su apoyo, por su amistad y por hacer que las rutinas siempre tengan un sabor diferente. Gracias, Sandra, Elena, Cristian, Laëtitia, Alejandro, María y, especialmente, Fer y Julio. Gracias, Fer, por ser esa persona en la que siempre puedo encontrar un amigo, con un punto de vista prudente, racional y

coherente. Y gracias, Julio, por aportar siempre ese toque de humor diferente, con ideas originales, a la par que brillantes, que tengo la suerte de escuchar cada día al otro lado de la mesa. Recordar también con especial cariño a los que emprendieron otro rumbo, no sin antes dejar un pedacito de ellos en La Sala. Gracias por los momentos compartidos, Félix, Reinier, Andrés y Fran.

También quiero darle las gracias a José Dolz, mi supervisor durante la estancia de investigación. Una estancia totalmente atípica, marcada por el coronavirus y otros infortunios que limitaron las oportunidades que, a priori, se presentaban. No obstante, si echo la vista atrás, solo puedo estar agradecido a José, por su apoyo y su implicación que, a pesar de los obstáculos, contribuyeron enormemente a darle un valor añadido a esta tesis. Gracias, José.

No puedo dejar de agradecerle a mi mejor amigo, por ese apoyo incondicional, por cuidarme y estar siempre pendiente de mí, por solo saber decirme que sí, por sumar siempre en todos los aspectos de mi vida. Gracias, Carlos Javier, ya sabes lo importante que eres para mí. También quiero agradecer al resto de mis amigos por enseñarme que siempre es buen momento para volver a Iniesta, que siempre hay gente en casa. Gracias, chicos.

"Yo soy yo y el sacrificio de mis padres". La vida es demasiado corta para agradecer lo suficiente a mis padres, quienes siempre han dado más de lo que tenían para que yo pueda estar donde estoy. Su educación, su cariño, sus siempre sabios consejos y su buen querer, han hecho de mí quien soy. Solo espero estar algún día a la altura de vuestras enseñanzas y demostraros que vuestro esfuerzo mereció la pena. Gracias por todo. Os quiero mucho.

Gracias a mi hermano Rafael, por enseñarme más de la vida de lo que ningún libro podrá jamás enseñarme. Gracias por cuidarme siempre, por ser mi mejor complemento y mi mayor apoyo. Gracias por ser ese perfecto equilibrio que ha hecho posible compaginar esta tesis con el resto de las cosas importantes en la vida. Te admiro casi tanto como te quiero. Gracias también al resto de mi familia: a mis tíos, a mis primas y a mis abuelos. Con vuestro apoyo incondicional, me habéis ayudado más de lo que os podéis imaginar.

Por último, me gustaría dedicar unas palabras de especial agradecimiento a mi Rho, a esa persona especial que me acompaña cada día para ayudarme a crecer, tanto personal como profesionalmente. Gracias, cariño, por hacer de mí una mejor versión, por contagiarme tu locura y alegría, a la vez que tu responsabilidad. Gracias por ser una fuente de inspiración y por estar siempre a mi lado, apoyándome en los momentos malos y celebrando conmigo en los buenos. Gracias por la vida que me das. Te quiero, mi ρ.

Abstract

Machine learning (ML) is one of the most important areas in the field of artificial intelligence (AI), which is increasingly present in our daily lives. From the beginning, ML algorithms have played an important role in the development of computer-aided diagnostic systems aimed at improving the efficiency and accuracy of experts. In this context, medical imaging has been of particular interest, as computer vision (CV) techniques can automatically perform pattern recognition tasks to associate certain biomedical structures with a specific disease.

Over time, different imaging modalities have been used to address a wide range of diseases under the umbrella of CV. In this thesis, we focus on two important research areas at the forefront of medical imaging: digital pathology and ophthalmology. Specifically, we use histological images to assist pathologists in the diagnosis of prostate and bladder cancer, and optical coherence tomography (OCT) data to help ophthalmologists in glaucoma decision making. We propose different cutting-edge solutions based on traditional and deep learning methods, as well as hybrid approaches to leverage the strengthens of each.

In order to exploit the potential of AI in diagnostic imaging, we address several learning paradigms to cover different supervision scenarios. For histological imaging, we propose fully supervised methods for the segmentation and classification of prostate-specific structures, as well as fully unsupervised techniques for bladder-specific histological pattern recognition. Moreover, for glaucoma assessment, we rely on recurrent learning to detect glaucoma from OCT volumes in the spectral domain (SD), and on few-shot learning to determine the severity level of glaucoma from circumpapillary B-scans.

In the prostate-based studies, we provide a comparison between hand-driven and deep learning methods for identifying the earliest stage of prostate cancer. The conventional ML approach shows better performance than deep learning in distinguishing between artefacts (false glands), benign and

pathological glands, as hand-crafted features allow the computation of spatial hierarchies and orientations that are essential in this multi-class scenario. The proposed end-to-end system contributes to the accurate localisation and classification of prostate histological structures, achieving an accuracy of 88.30% in discriminating normal and Gleason grade 3 glands. In contrast, the proposed deep unsupervised algorithm far outperforms other conventional clustering algorithms in the classification of muscle-invasive bladder cancer (MIBC). Here, we resort to high-resolution histological samples stained with immunohistochemistry techniques to self-recognise non-tumour, mild and infiltrative MIBC patterns. The proposed model achieves a multi-class accuracy of 90.31% without incurring prior annotation steps, which bridges the gap with respect to training the model on labelled data.

Regarding glaucoma detection from SD-OCT volumes, we propose the combination of convolutional neural networks (CNNs) with long short-term memory (LSTM) units to find glaucoma-specific spatial dependencies between adjacent 2D slides of the SD-OCT cube. Important contributions to glaucoma detection are included in the architectures of both the slide-level feature extractor and the volume-based predictive model. The proposed recurrent learning system improves on other state-of-the-art approaches based on 3D architectures, reaching an accuracy of 81.25% in discerning between healthy and glaucomatous SD-OCT volumes. Delving deeper into glaucoma assessment, we address a novel learning strategy to discriminate, for the first time, between different levels of glaucoma severity from circumpapillary OCT B-scans. We propose a new hybrid backbone to optimise the feature extraction process and embed it in a novel few-shot learning scenario based on dynamic prototypical neural networks (PNN). The convolutional coefficients of the backbone are refined during model training according to the prototypical latent feature assignment, leading to higher performance compared to dense layers activated by softmax. At test time, the proposed model achieves accuracies of 96.97% and 87.88% in detecting and grading glaucoma, respectively.

In short, the AI methods proposed in this thesis contributes to the diagnosis of prostate and bladder cancer from histological images, as well as to glaucoma detection from OCT samples, making use of ML algorithms under different supervision scenarios.

Resumen

El *machine learning* (ML) es una de las áreas más importantes en el campo de la inteligencia artificial (IA), la cual está cada vez más presente en nuestra vida cotidiana. Los algoritmos de ML siempre han desempeñado un papel importante en el desarrollo de sistemas de ayuda al diagnóstico destinados a mejorar la eficacia y la precisión de los expertos. En este contexto, la imagen médica ha cobrado especial interés, ya que las técnicas de *computer vision* (CV) pueden realizar automáticamente tareas de reconocimiento de patrones para asociar determinadas estructuras biomédicas a una enfermedad específica.

A lo largo del tiempo, diferentes modalidades de imagen se han utilizado para abordar una amplia gama de enfermedades bajo el paraguas del CV. En esta tesis, nos centramos en dos importantes áreas de investigación en el campo de la imagen médica: la patología digital y la oftalmología. Usamos imágenes histológicas para ayudar a los patólogos en el diagnóstico del cáncer de próstata y de vejiga, y datos de tomografía por coherencia óptica (OCT) para ayudar a los oftalmólogos en la toma de decisiones relacionadas con el glaucoma. Para ello, proponemos diferentes soluciones de vanguardia basadas en métodos tradicionales de ML y de *deep learning*, así como enfoques híbridos.

Además, abordamos varios paradigmas de aprendizaje para cubrir diferentes escenarios de supervisión. Con respecto a las imágenes histológicas, proponemos métodos supervisados para segmentar y clasificar estructuras específicas de la próstata, así como técnicas totalmente no supervisadas para el reconocimiento de patrones histológicos de la vejiga. Con respecto al glaucoma, aplicamos aprendizaje recurrente para detectar la enfermedad en volúmenes OCT en el dominio espectral (SD), así como métodos de *few-shot learning* para determinar su gravedad a partir de imágenes OCT circumpapilares.

En los estudios sobre la próstata, ofrecemos una comparación entre los métodos de *hand-driven* y *deep learning* para identificar la etapa más temprana del cáncer de próstata. El enfoque convencional muestra un mejor rendimiento

a la hora de distinguir entre artefactos (glándulas falsas), glándulas benignas y patológicas, ya que las características codificadas manualmente permiten tener en cuenta el cálculo de jerarquías y orientaciones espaciales, lo cual es esencial en este escenario multiclase. El sistema propuesto contribuye a la precisa localización y clasificación de las estructuras histológicas de la próstata, logrando una precisión del 88.30% en la discriminación entre glándulas normales y de grado 3 de Gleason. Por su parte, el algoritmo propuesto de *deep learning* no supervisado supera con creces a otros métodos de *clustering* convencional en la clasificación del cáncer de vejiga músculo-invasivo (MIBC). En este caso, utilizamos muestras histológicas de alta resolución teñidas con técnicas de inmunohistoquímica para reconocer patrones de MIBC no tumorales, leves e infiltrativos. El modelo propuesto alcanza una precisión multiclase del 90.31% sin incurrir en pasos previos de anotación, lo cual reduce la brecha con respecto a entrenar el modelo utilizando datos etiquetados.

En cuanto a la detección de glaucoma a partir de volúmenes SD-OCT, proponemos la combinación de redes neuronales convolucionales (CNN) con algoritmos de memoria a corto plazo (LSTM) para encontrar dependencias espaciales específicas de glaucoma entre los cortes 2D. Se incluyen contribuciones clave para la detección del glaucoma en las arquitecturas tanto del extractor de características a nivel de imagen como del modelo predictivo basado en el volumen. El sistema propuesto centrado en el aprendizaje recurrente mejora otros enfoques del estado del arte basados en arquitecturas 3D, alcanzando una precisión del 81.25% en la clasificación entre volúmenes SD-OCT sanos y glaucomatosos. Profundizando en la evaluación del glaucoma, llevamos a cabo una novedosa estrategia de aprendizaje para discernir, por primera vez, entre diferentes niveles de gravedad del glaucoma a partir de imágenes OCT circumpapilares. Proponemos una nueva arquitectura híbrida para optimizar el proceso de extracción de características y la embebemos en un novedoso escenario de *few-shot learning* basado en redes neuronales prototípicas (PNN) dinámicas. Los coeficientes convolucionales de la arquitectura se refinan durante el entrenamiento del modelo de acuerdo con la asignación prototípica de las características latentes, lo que conduce a un mayor rendimiento en comparación con las capas densas activadas por softmax. Durante la etapa de test, el modelo propuesto alcanza precisiones del 96.97% y 87.88% en la detección y gradación del glaucoma, respectivamente.

En definitiva, los métodos de IA propuestos en esta tesis contribuyen al diagnóstico del cáncer de próstata y vejiga a partir de imágenes histológicas, así como a la detección del glaucoma a partir de muestras OCT, utilizando algoritmos de ML bajo diferentes escenarios de supervisión.

Resum

El *machine learning* (ML) és una de les àrees més importants en el camp de la intel·ligència artificial (IA), que està cada vegada més present en la nostra vida quotidiana. Des del principi, els algorismes de ML han exercit un paper important en el desenvolupament de sistemes d'ajuda al diagnòstic destinats a millorar l'eficàcia i la precisió dels experts. En aquest context, la imatge mèdica ha cobrat especial interés, ja que les tècniques de *computer vision* (CV) poden realitzar automàticament tasques de reconeixement de patrons per a associar determinades estructures biomèdiques amb una malaltia específica.

Al llarg del temps, diferents modalitats d'imatge s'han utilitzat per a abordar una àmplia gamma de malalties sota el paraigua del CV. En aquesta tesi, ens centrem en dues àrees d'investigació importants en el camp de la imatge mèdica: la patologia digital i l'oftalmologia. Usem imatges histològiques per a ajudar als patòlegs en el diagnòstic del càncer de pròstata i de bufeta, i dades de tomografia de coherència òptica (OCT) per a ajudar als oftalmòlegs en la presa de decisions sobre el glaucoma. Per a això, proposem diferents solucions d'avantguarda basades en mètodes tradicionals de ML i de *deep learning*, així com enfocaments híbrids.

A més, abordem diversos paradigmes d'aprenentatge per a cobrir diferents escenaris de supervisió. Respecte a les imatges histològiques, proposem mètodes totalment supervisats per a la segmentació i classificació d'estructures específiques de la pròstata, així com tècniques no supervisades per al reconeixement de patrons histològics de la bufeta. Respecte al glaucoma, recorrem al *recurrent learning* per a detectar la malaltia en els volums SD-OCT, així com a mètodes de *few-shot learning* per a determinar la seua gravetat a partir d'imatges OCT circumpapilares.

En els estudis sobre la pròstata, oferim una comparació entre els mètodes de *hand-driven* i *deep learning* per a identificar l'etapa més primerenca del càncer de pròstata. L'enfocament convencional mostra un millor rendiment

que el *deep learning* a l'hora de distingir entre artefactes (glàndules falses), glàndules benignes i patològiques, ja que les característiques codificades manualment permeten tindre en compte el càlcul de jerarquies i orientacions espacials, la qual cosa és essencial en aquest escenari multiclasse. El sistema proposat contribueix a la precisa localització i classificació de les estructures histològiques de la pròstata, aconseguint una precisió de 88.30% en la discriminació entre glàndules normals i de grau 3 de Gleason. Per contra, l'algorisme proposat de *deep learning* no supervisat supera amb escreix a altres mètodes de *clustering* convencional en la classificació del càncer de bufeta múscul-invasiu (MIBC). En aquest cas, utilitzem mostres histològiques tenyides amb tècniques d'immunohistoquímica per a reconéixer patrons de MIBC no tumorals, lleus i infiltrativos. El model proposat aconsegueix una precisió multiclasse del 90.31% sense incórrer en passos previs d'anotació, la qual cosa redueix la bretxa respecte a entrenar models supervisats.

Quant a la detecció de glaucoma a partir de volums SD-OCT, proposem la combinació de xarxes neuronals convolucionals (CNN) amb algorismes de memòria a curt termini (LSTM) per a trobar dependències espacials específiques de glaucoma entre els talls 2D. S'inclouen contribucions clau per a la detecció del glaucoma en les arquitectures tant de l'extractor de característiques a nivell de diapositiva com del model predictiu basat en el volum. El sistema proposat centrat en *recurrent learning* millora altres enfocaments de l'estat de l'art basats en arquitectures 3D, aconseguint una precisió del 81.25% en la classificació entre volums SD-OCT sans i glaucomatosos. Aprofundint en l'avaluació del glaucoma, duem a terme una nova estratègia d'aprenentatge per a discernir, per primera vegada, entre diferents nivells de gravetat del glaucoma a partir d'imatges OCT circumpapilares. Proposem una nova arquitectura híbrida per a optimitzar el procés d'extracció de característiques i l'embevem en un nou escenari de *few-shot learning* basat en xarxes neuronals prototípiques (PNN) dinàmiques. Els coeficients convolucionals de l'arquitectura es refinen durant l'entrenament del model d'acord amb l'assignació prototípica de les característiques latents, la qual cosa condueix a un major rendiment en comparació amb les capes denses activades per softmax. Durant l'etapa de test, el model proposat aconsegueix precisions 96.97% i del 87.88% en la detecció i gradació del glaucoma, respectivament.

En definitiva, els mètodes de IA proposats en aquesta tesi contribueixen al diagnòstic del càncer de pròstata i bufeta a partir d'imatges histològiques, així com a la detecció del glaucoma a partir de mostres de OCT, utilitzant algorismes de ML baix diferents escenaris de supervisió.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

This chapter introduces the motivation and the objectives pursued in this thesis, as well as the main contributions. It also includes the thesis framework and the thesis outline.

Contents

1.1 Motivation

Artificial intelligence (AI) is a part of computer science that focuses on the intelligent analysis of data. At the turn of the century, Kononenko [1] claimed that there is no intelligence without learning, which makes machine learning (ML) one of the most important and rapidly developing subfields of AI research. According to [1], ML algorithms were, from the beginning, designed and used to analyse medical datasets. In [2], Xu et al. also declared that the application of AI technology in healthcare enhances human abilities and improves the accuracy of medical treatment. From the 1970s onwards, with the digital revolution and the AI development, accessible means of data collection and storage began to appear. Hospitals were provided with well-equipped systems to gather and share a large amount of information [1]. Since then, a wide range of medical domains such as histopathology [3, 4], ophthalmology [5, 6], radiology [7, 8], odontology [9, 10] and embryology [11, 12], among many others [13], have been the subject of study for the development of diagnostic-aid systems under the umbrella of ML algorithms. An automatic diagnosis via machine learning is simplistically translated as a classification task in which medical diagnostic knowledge can be automatically derived from the description of cases solved in the past. The proposed models can be then inferred either to assist experts in diagnosing new patients or to train non-specialists as a support tool [1].

In the context of computer-aided diagnosis, image-based systems are increasingly relevant because ML algorithms enable automatic pattern recognition that allows intrinsic biomedical structures to be associated with a specific disease [2]. Different imaging modalities have been explored in the literature to aid decision making in countless diseases, e.g. fundus image, for ocular disorders such as diabetic retinopathy [14], age-related macular degeneration [15], glaucoma [16], etc.; X-ray imaging, for scoliosis [17], bone fracture [18], pneumonia [19], etc.; magnetic resonance imaging (MRI), for Alzheimer [20], colorectal cancer [21], Parkinson [22], etc.; computed tomography (CT), for lung cancer [23], Covid-19 [24], intracranial haemorrhage [25], etc.

Originally, ML approaches based on medical imaging advocated sequential learning in which a feature engineering process must be carried out before the classification stage [26]. This type of hand-driven learning usually requires the outcome of a prior image segmentation task to delimit the regions of interest (ROIs) from which extracting the relevant information for a particular disease [27]. Back in 2011, with the advent of deep learning running on GPUs, a new range of promising possibilities opened up in the field of AI. Hand-

crafted features began to be replaced by automatically learned embedding when Krizhevsky et al. [28] demonstrated that the use of deep convolutional neural networks (CNNs) provided a significant improvement in the *ImageNet*[1] classification challenge. However, it was not until around 2015, two years before the start of this PhD thesis, that deep learning began to gain a foothold in the research area [29]. At that moment, according to de Bruijne et al. [30], one of the main challenges facing deep learning was the lack of labelled data, as for data to drive learning, a lot of data is needed. This long-standing problem is now accentuated in the healthcare community, where technical and specific expertise is required to label data.

Under ideal conditions, it is possible to train deep learning models using millions of annotated samples in a fully supervised setting, but it is far from the real-life scenario that medicine currently faces in daily practice. This is just one of the reasons why traditional computer vision approaches can sometimes lead to a better solution than deep learning, as conventional algorithms are not class-specific, but work the same for any image. In contrast, the features learned from a deep neural network are dependent on the training dataset which, if it does not account for variability, is likely to perform poorly in predicting new samples. In [31], a comparison is made between deep learning and traditional ML algorithms providing limitations and advantages of each strategy. Mahony et al. [31] also discussed mixing hand-crafted and deep learning approaches in a hybrid framework for better performance, as conventional algorithms provide transparency and power efficiency, whereas deep learning offers higher accuracy and versatility [32]. Moreover, the issue of limited healthcare-related information opens the door to many other deep learning paradigms intended to deal with the lack of labelled data, e.g. self-training, unsupervised, semi-supervised, few-shot learning, etc.

In this PhD thesis, we explore all these families of learning methods to exploit the potential of the AI for improving diagnostic accuracy and expert efficiency. We propose innovative ML-based solutions to cover two interesting research areas at the leading edge of medical imaging [33]: digital pathology and ophthalmology. Regarding the former, histopathological images digitised from biopsied tissues have recently come under the spotlight to characterise different types of cancer via ML algorithms. Most of the state-of-the-art studies focus on prostate cancer to assist pathologists in Gleason grading of tumours [3, 4]. However, histological imaging allows the analysis of other types of cancer such as melanoma [34], osteosarcoma [35] and colorectal cancer [36], among others [37, 38]. In this PhD thesis, we mainly focus on prostate cancer

[1] https://image-net.org/download-images

using whole-slide images (WSIs) stained by Hematoxylin and Eosin (H&E) (Chapter 2). To go further in histological image-based diagnosis, we also propose a novel framework for bladder cancer grading using WSIs stained by immunohistochemistry techniques (Chapter 3). Concerning the ophthalmology field, AI-based methods have been also explored in the literature to analyse several optical disorders, especially from fundus image and OCT-based data. Since our aim is the assessment of glaucoma, we advocate the use of the OCT technology, as it is the imaging modality par excellence for glaucomatous damage evaluation [39]. To carry out a robust exploration of ML algorithms aimed at glaucoma detection, we cover both 3D and 2D data using spectral-domain (SD)-OCT volumes (Chapter 4) and circumpapillary OCT B-scans (Chapter 5), respectively.

Big players

According to [40], the recent impact of digital and computational pathology is reflected in academia through the increase in publications. Specifically, from the studies published in the last 10 years related to the terms of computational pathology and AI, PubMed and Google Scholar respectively report that 77.48% and 83.09% have been registered in the last 5 years, which is evidence of the growing interest of the scientific community in this field. In addition, big players such as Google Health [41, 42] and Philips [43], among others, are now at the forefront of computational pathology aimed at assisting pathologists through artificial intelligence. Other important entities also show interest in this field. The Food and Drug Administration (FDA) approved the first American digital pathology system for diagnostic use in 2017 [44]; the UK Life Sciences Industrial Strategy reinforced actions for the use of digital pathology [45]; the Royal College of Pathologists highlighted the need for investment to support digital pathology infrastructure [46] and the Innovative Medicines Initiative (IMI) launched an H2020 call focused on supporting the collaborative development of artificial intelligence in pathology in 2019 [47].

Similarly, the ophthalmology field has shown significant progress with the breakthrough of AI in the assessment of different optical diseases through diagnostic imaging, especially using fundus images and OCT data. As before, this is evident in academia from the percentage of publications in recent years. Looking at OCT and AI-related studies published in the last 10 years, PubMed and Google Scholar record, respectively, that 75.96% and 85.55% are from the last 5 years. Big players in this research area include DeepMind [48, 49] and IBM research [50], among others. Specifically, *DeepMind* has recent publications in leading journals applying deep learning algorithms on OCT samples to address age-related macular degeneration [48] and other retinal-

derived diseases [49]. Furthermore, *IBM research*, in collaboration with the New York University School of Medicine, is involved in glaucoma assessment, as reported in [50], where Maetschke et al. proposed the use of 3D convolutional architectures to detect glaucoma from OCT volumes. Additionally, some of the world's leading universities have shown interest in this research field in recent years. For instance, Harvard University, with the review of AI on ophthalmology carried out in [51], Stanford University, with different studies for glaucoma assessment using OCT [52, 53] and Columbia University, with other interesting works to evaluate glaucoma progression from B-scans [54, 55].

1.2 Objectives

The main objective of this PhD thesis is the design, development and validation of innovative diagnostic-aid systems to understand how ML algorithms can assist experts to manage specific diseases in clinical practice. We aim to address a wide range of learning methods to cover different imaging domains and problems via artificial intelligence. In particular, we set our sights on hand-driven, deep learning and hybrid approaches for segmentation and classification tasks. This thesis intends to frame these approaches in different learning paradigms such as supervised, self-training, recurrent, few-shot and fully unsupervised learning to include distinct supervision scenarios where the amount of annotated data varies. In pursuit of this, we aim to develop new ML algorithms able to address problems proposed by doctors and characterise common diseases of each of the proposed imaging modalities.

In each chapter of this PhD thesis, different objectives are proposed depending on the type of data. For histological imaging, the main target for both prostate and bladder cancer scenarios is to recognise and characterise tumour patterns in order to predict patient prognosis. In the prostate-based studies, we aim to identify the first stage of cancer by segmenting and classifying histological glandular structures via hand-driven and deep learning algorithms. In contrast, in the case of bladder cancer, we opt for a self-recognition of high-resolution patterns to determine tumour aggressiveness using fully unsupervised learning methods. Concerning OCT data, the main goal is to detect and grade glaucoma by applying ML algorithms on SD-OCT volumes and circumpapillary B-scans. Specifically, we propose a novel glaucoma detection framework focusing on recurrent knowledge to deal with SD-OCT cubes composed of unlabelled slides. Furthermore, we aim to discern, for the first time, between different levels of glaucoma severity by making use of circumpapillary images centred on the optic nerve head (ONH) of the retina. To achieve these objectives, each proposed

approach must undergo a sequential protocol based on the CRISP-ML(Q) methodology [56], as detailed below:

- Data understanding. An in-depth exploration of the histological patterns of prostate and bladder cancer should be carried out. Similarly, an exhaustive analysis of OCT-specific glaucoma-related anomalies should be performed.
- Data preparation. The databases corresponding to each particular disease should be prepared for defining the learning framework and training the models. Processing algorithms based on feature engineering should be included in this step to select, clean, construct and standardise data.
- Modelling. This stage includes the design and development of innovative ML-based predictive algorithms to provide optimal disease-specific approaches. Different learning paradigms should be implemented to handle the peculiarities of each disease and imaging modality. Previously, state-of-the-art methods should be explored to propose cutting-edge solutions for automatic image-based diagnostics.
- Evaluation. In this step, the robustness of the proposed models should be determined through validation techniques. Quantitative and qualitative results should be reported to compare the models' performance with other state-of-the-art methods. The identification of limitations and future research lines should be also performed in this stage to decide if the model can be deployed.

The last steps of the CRISP-ML(Q) methodology, i.e. *Deployment* and *Monitoring and Maintenance* are not presented in this PhD thesis.

1.3 Main contributions

As mentioned above, we focus on two specific imaging modalities: i) histopathology, to assist pathologists in the diagnostic process of prostate and bladder cancer, and ii) optical coherence tomography (OCT), to help ophthalmologists in glaucoma decision-making. In the course of this PhD thesis, a transition from conventional ML to deep learning algorithms was experienced through different biomedical image-based applications. In [57, 58], segmentation and classification tasks were addressed using conventional and deep learning approaches, respectively, to detect and categorise prostate-specific structures in histological images. In [59, 60], we carried out a

comparison between hand-driven and data-driven algorithms to identify, in a novel way, the first stage of prostate cancer by segmenting and classifying histological glands locally. In [61], we faced, for the first time, a fully unsupervised scenario employing deep clustering algorithms to self-recognise the aggressiveness of bladder cancer from high-resolution histopathological images. In an attempt to extend our AI knowledge to other imaging modalities, we made use of OCT data to aid in the diagnosis of glaucoma. In [62], we conducted a supervised pipeline for glaucoma detection using fully convolutional neural networks (CNNs). Further on, in [63, 64], we combined hand-crafted and automatic-learned OCT-specific features, giving rise to innovative-hybrid approaches for optimal glaucoma detection. As a novelty, in [65], we applied recurrent learning on unlabelled 2D images to find glaucoma-specific spatial dependencies in SD-OCT volumes. In addition, we explored advanced deep learning techniques for limited supervision. In [66], we proposed a self-training framework for glaucoma grading to take advantage of unlabelled samples through pseudo-labelling methods. Moreover, in [64], we resorted to a new learning strategy based on supervised prototypical neural networks (PNNs) to grade glaucoma under a redefined few-shot paradigm.

1.3.1 Contributions to prostate and bladder cancer from histological images

Prostate cancer is the second leading cause of cancer death in men and one of the most common types of cancers in the entire population [67]. Currently, the diagnosis of this chronic disease is made by pathologists after extracting and staining a sample of prostate tissue. Experts assign a specific score according to the Gleason scale, from 1 to 5, in ascending order of aggressiveness [68]. Gleason grades 1 and 2 closely resemble the structure of normal tissues, where individual and well-differentiated glands are evident. In contrast, Gleason grades 3, 4 and 5 correspond to cancerous tissues (from less to more aggressive). In Chapter 2, we aim to identify prostate cancer at its earliest stage by discerning between normal tissues and those with a Gleason pattern of grade 3. One of the main contributions lies in the classification framework carried out at the gland level. Unlike the state-of-the-art studies [69, 70], which addressed prostate cancer detection attending to a slide-level label, we focused on local gland patterns to determine the benign or malignant character of each prostate gland present in the tissue sample. This is possible because Gleason patterns 1, 2 and 3 have individual glands, while Gleason grades 4 and 5 show a different histological arrangement. As far as we are aware, our study [60] detailed in Chapter 2, was the first in addressing gland detection and

classification algorithms in a unified framework. For the gland segmentation task, we innovatively applied the *Locally Constrained Watershed Transform* (LCWT) method using the gland nuclei as the constraints required by the algorithm [57]. Note that all important tissue components were extracted by conventional clustering techniques. Further on, we improved the results of the proposed LCWT algorithm by a novel deep segmentation network based on residual and multi-resolution U-Net [59]. As for the classification task, both hand-driven and deep learning approaches are detailed in Chapter 2. The main contribution in this line is reported by the hand-crafted feature extraction algorithm, which provides a fingerprint composed of four types of descriptors. In addition to morphological, texture and contextual features, the measurement of the Hurst exponent deserves a special mention, as [60] was the first study in which the fractal dimension was considered for prostate cancer detection. We also addressed the discrimination between normal and Gleason 3 pathological tissues through deep learning methods, namely by training CNNs from scratch and fine-tuning the most popular ones in the literature [58]. Another related contribution was the publication of a new locally annotated prostate cancer database with artefacts, benign and pathological glands [60].

Bladder cancer is the second most common urinary tract cancer and the fifth most prevalent among men in developed countries [71]. As with prostate, definitive diagnosis of bladder cancer requires a biopsy to examine the tumour growth. Tissue samples are usually stained with hematoxylin and eosin (H&E) to enhance the histological properties; however, as a novelty, we resorted to digitised images stained with immunohistochemical CK AE1/3 to highlight in brown the antigen-antibody binding. Bladder cancer can be non-muscle invasive (NMIBC) and muscle-invasive (MIBC). We focus on the latter, as it has the worst prognosis and favours tumour dissemination to adjacent organs. MIBC corresponds to a high-grade urothelial carcinoma, but the patient's prognosis depends on the aggressiveness of the histological patterns [72]. Specifically, we can find three types of patterns: nodular, trabecular and infiltrative. In the study [61] detailed in Chapter 3, we coined the nodular and trabecular patterns as class "mild" because they are associated with a better prognosis. Contrarily, the infiltrative pattern denotes tumour budding, which refers to the most aggressive scenario with a high mortality rate. In Chapter 3, we provide a novel self-learning framework to discern between non-tumour, mild and infiltrative patterns. The study details a fully unsupervised learning methodology to recognise the MIBC-specific structures without the need for prior annotation steps. It consists of a novel deep clustering architecture based on an attention module aimed at refining the feature space and improving the classification of high-resolution histological patches of MIBC.

1.3.2 Contributions to glaucoma from optical coherence tomography samples

Glaucoma is a degenerative optic neuropathy characterised by causing several visual field defects, functional damage and structural changes in the optic nerve [73]. Nowadays, this chronic disease is the leading cause of blindness worldwide and is expected to affect 111.8 million people in 2040 [74, 75]. Current diagnosis of glaucoma involves a large number of hardworking tests such as pachymetry, tonometry and visual field testing, among others. Imaging-based techniques have become of particular interest for glaucoma diagnosis, especially the OCT modality, as it is able to measure the deterioration of retinal cell layers, which is closely related to glaucoma disease. To contribute to OCT-based glaucoma assessment, we propose two different ML frameworks to deal with 3D data, via SD-OCT volumes, and 2D images, via circumpapillary B-scans. Regarding the first one, our work [65] detailed in Chapter 4 addresses a novel glaucoma detection scenario from ONH-centred SD-OCT slides labelled at the volume level. For the first time, we advocated the use of a hybrid setup composed of CNNs and Long Short-Term Memory (LSTM) networks to extract the feature space from each slide and handle the information relative to the spatial dependencies across the SD-OCT volume. We assumed each spatial slide as a temporary instance, rather than resorting to 3D deep learning architectures as state-of-the-art studies [50, 76]. As a novelty, we proposed a 2D-CNN feature extractor that combines fine-tuned networks with convolutional blocks trained from scratch through residual connections. We also included an identity shortcut to refine the latent space extracted from each slide using an attention module. Additionally, in [65], we detail a new way of codifying the LSTM outputs by proposing a sequential-weighting module (SWM), which allows all the outputs of each LSTM cell to be considered in a weighted manner. This lead to an optimal convergence by providing more stable learning during the model training. To compute and measure the performance of the learning curves, we proposed two new metrics: i) OVFT, to quantify overfitting and ii) QLTY, to quantify the quality of training by considering the stability (STBL) and the amount of learning (LRNG). Finally, we report the class activation maps (CAMs) to evidence not only the ROIs in each slide of the volume, but also the most relevant slides to detect glaucoma from SD-OCT volumes without losing the spatial information.

Concerning the 2D approach based on circumpapillary OCT B-scans, we conducted two works ([62] and [63]) that form the basis of the study [64] detailed in Chapter 5. Specifically, in [62], we proposed a fully-supervised strategy for glaucoma detection by fine-tuning the most popular state-of-the-

art architectures. Contrarily, in [63], we used the same dataset to validate the performance of new CNNs trained from scratch. To cover all the possible learning scenarios, we also performed fully hand-driven and hybrid-based strategies to combine the hand-crafted and automatic features in the same pipeline. Inspired by the results achieved in these previous works, in [64], we proposed a hybrid network as a benchmark of a novel glaucoma-related paradigm aimed at measuring, for the first time, the aggressiveness of the disease from circumpapillary B-scans. The hybrid network is focused on the late fusion of the features extracted from both fine-tuned architectures conducted in [62] and the new hand-crafted methods proposed in [63]. One of the main contributions addressed in [64] is the use of tailored PNNs to optimise the classification performance between healthy, early and advanced glaucomatous cases. We detail different prototype-based solutions in a few-shot paradigm, in which the k-shot methodology was adapted to optimise glaucoma grading via supervised learning. Further on, in [66], we carried out a self-training framework for glaucoma grading under domain adaptation techniques. We proposed a two-step learning methodology to transfer, via pseudo-labelling, domain knowledge from an unlabelled dataset (target) to a model previously trained on a labelled dataset (source).

1.4 Framework

This PhD thesis is part of three different research projects, as detailed below:

- *SICAP* − Sistema de interpretación de imágenes histopatológicas para la detección del cáncer de próstata. This is a national project whose objective was to develop a diagnostic aid system for prostate cancer by classifying histopathological images from biopsies into different grades according to the Gleason scale. *SICAP* project was funded by the *Ministerio de Economía, Industria y Competitividad* (DPI2016-77869-C2-1-R). Chapter 2 contributes to this project in the design and development of a classification system capable of identifying the first stage of prostate cancer from specific histological structures.

- *GALAHAD* − Glaucoma-advanced, label-free high resolution automated OCT diagnostics. This is a European project that aimed to improve the axial resolution of OCT equipment and explore polarisation sensitive applications of OCT to develop an early warning glaucoma screening test. *GALAHAD* project was funded by the European Commission through Horizon 2020 [H2020-ICT-2016-2017, 732613]. Chapter 4 and Chapter 5

contribute to this project on glaucoma detection from SD-OCT volumes and glaucoma grading from circumpapillary B-scans, respectively.

- *CLARIFY* – Cloud artificial intelligence for pathology. This is another European project that proposes the creation of a research infrastructure based on AI and cloud-oriented data algorithms to facilitate the interpretation and diagnosis of triple-negative breast cancer (TNBC), high-risk non-muscle-invasive bladder cancer (HR-NMIBC) and spitzoid melanocytic lesions (SML) from histopathological images. *CLARIFY* project was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska Curie grant agreement (No 860627). Chapter 3 contributes to this project giving rise to a self-learning framework for the assessment of bladder cancer by applying unsupervised techniques on histological images.

1.5 Outline

This thesis is divided into 6 chapters. The current chapter introduces the motivation behind the research involved in this thesis, the proposed objectives and the main contributions. Subsequently, this chapter also details the framework and the thesis outline.

Chapter 2 corresponds to the paper: "First-Stage Prostate Cancer Identification on Histopathological Images: Hand-Driven versus Automatic Learning" [60]. It was published in the journal *Entropy* belonging to the editorial *Multidisciplinary Digital Publishing Institute* (MDPI). *Entropy* had an impact factor of 2.494 when the article was published in 2019, an impact score of 3.01 and an h-index of 74. The best rank was in the category *physics, multidisciplinary* with a percentile of 61.76 (Q2).

Chapter 3 corresponds to the paper: "A Novel Self-Learning Framework for Bladder Cancer Grading Using Histopathological Images" [61]. It was published in the journal *Computers in Biology and Medicine* (CIBM) belonging to the editorial *ELSEVIER*. The paper was published in 2021, but the following publication details correspond to 2020, as the most recent journal indexes date from that year. CIBM journal had an impact factor of 4.589, an impact score of 5.59 and an h-index of 94 in 2020. The top ranking was in the category *mathematical & computational biology* with a percentile of 88.79 (Q1).

Chapter 4 corresponds to the paper "Glaucoma Detection from Raw SD-OCT Volumes: a Novel Approach Focused on Spatial Dependencies" [65], published

in the journal *Computer Methods and Programs in Biomedicine* (CMPB). The paper was also published in 2021, but the publication data is for 2020. CMPB journal had an impact factor of 5.428, an impact score of 6.67 and an h-index of 102 in 2020. The best rank was in the category *computer science, theory & methods* with a percentile of 88.64 (Q1).

Chapter 5 corresponds to the paper "Circumpapillary OCT-Focused Hybrid Learning for Glaucoma Grading Using Tailored Prototypical Neural Networks" [64], published in the jorunal *Artificial Intelligence in Medicine* (AIIM). As before, the paper was published in 2021, but the publication details date from 2020. AIIM journal had an impact factor of 5.326, an impact score of 6.69 and an h-index of 87 in 2020. The top ranking was in the category *medical informatics* with a percentile of 78.33 (Q1).

Note that Chapters 2, 3, 4 and 5 base on the same communication structure. First, they present an abstract followed by an introduction containing a review of the literature and the contribution of the proposed work. Next, the material section explains the datasets used to train and evaluate the developed ML algorithms, which are explained in the following methodology part. Then, the performance reached by the proposed methods is presented and discussed in the results and discussion sections, respectively. At the end, a brief conclusion recapitulating the main results and contributions of each work is included.

In Chapter 6, we relate the findings from each paper with the global aim of this PhD thesis. We also collect final remarks from a global perspective and suggest future research lines. Then, in Merits, we include journal publications, national and international conferences, as well as research awards derived from this thesis. Finally, we display the Bibliography.

Chapter 2

First-Stage Prostate Cancer Identification on Histopathological Images: Hand-Driven versus Automatic Learning

The content of this chapter corresponds to the author version of the following published paper: García, G., Colomer, A. & Naranjo, V. First-stage prostate cancer identification on histopathological images: Hand-driven versus automatic learning. Entropy, 21(4), 356 (2019).

Contents

First-Stage Prostate Cancer Identification on Histopathological Images: Hand-Driven versus Automatic Learning

Gabriel García[1], Adrián Colomer[1] and Valery Naranjo[1]

[1]Instituto de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, 46022, Valencia, Spain

Abstract

Analysis of histopathological image supposes the most reliable procedure to identify prostate cancer. Most studies try to develop computer aid-systems to face the Gleason grading problem. Contrary, we delve into the discrimination between healthy and cancerous tissues in its earliest stage, only focusing on the information contained in the automatically segmented gland candidates. We propose a hand-driven learning approach, in which we perform an exhaustive hand-crafted feature extraction stage combining in a novel way descriptors of morphology, texture, fractals and contextual information of the candidates under study. Then, we carry out an in-depth statistical analysis to select the most relevant features that constitute the inputs to the optimised machine learning classifiers. Additionally, we apply for the first time on prostate segmented glands, deep learning algorithms modifying the popular VGG19 neural network. We fine-tuning the last convolutional block of the architecture to provide the model specific knowledge about the gland images. Hand-driven learning approach, using a non-linear Support Vector Machine, reports a slight outperforming over the rest of experiments with a final multi-class accuracy of 0.876 ± 0.026 in the discrimination between false glands (artefacts), benign glands and Gleason grade 3 glands.

2.1 Introduction

Nowadays, prostate cancer is one of the most diagnosed types of cancer in the world, according to the American Cancer Society (ACS) [67]. This study reveals that prostate cancer supposes the second cause of death related to cancer in men and the first type of cancer concerning the estimated new cases. The ACS also situates this disease as one of the most common types of cancer in USA concerning the general population, although it only affects men. The Spanish Society of Medical Oncology (SEOM) [77] exposes prostate cancer as a chronic disease due to their very high incidence and 5-years prevalence ratios. For this reason, it becomes necessary to carry out a fast and accurate diagnosis facilitating an early treatment to improve the quality of life of the patients with this chronic disease.

At present, the diagnostic procedure to detect prostate cancer is a very time-consuming task that is manually accomplished by pathologists or urologists. First, they carry out a rectal examination to find anomalies in the size of the prostate gland. The next step is to perform some analysis based on detecting specific antigens in the blood, such as Prostate-Specific Antigen (PSA) and Prostate-Cancer Antigen (PCA3). If all non-invasive tests are positive, experts extract a sample of tissue and submit it to a preparation process composed of four phases: fixation, inclusion, cutting and staining. Once the samples are stained using the Hematoxylin and Eosin (H&E) pigment, specialists perform an in-depth examination under the microscope, or by computer if the samples are previously digitised, to determine the definitive diagnosis. Finally, pathologists base on the Gleason classification system [68] to assign specific scores (from 1 to 5) to each tissue depending on the cancer aggressiveness (see Figure 2.1).

Gleason grades 1 and 2 closely resemble the structure of normal tissues, since both have large and well-defined gland units. In addition, lumens contain large areas surrounded by cytoplasmic complexes and usually by an epithelial multi-layer of nuclei. More specifically, Gleason pattern 1 corresponds to a well-differentiated (low grade) carcinoma, whereas Gleason pattern 2 corresponds to a moderately differentiated carcinoma. Gleason grades 3, 4 and 5 are related to cancerous tissues (from less to more severe). In particular, a Gleason pattern 3 is characterised by presenting dimensions of lumens and glands smaller and more circular. Besides, the cell density of the epithelial nuclei layer is lower in the pathological tissue. Gleason pattern 3 also corresponds to a moderately differentiated carcinoma. Regarding the Gleason grade 4, there are no gland units, but glandular regions composed of the fusion of not well-defined glands.

Contrarily, the Gleason grade 5 is easily differentiated by the presence of a large number of scattered nuclei along the stroma. Gleason grades 4 and 5 correspond to a poorly differentiated carcinoma and an anaplastic carcinoma, respectively. (See Figure 2.1).

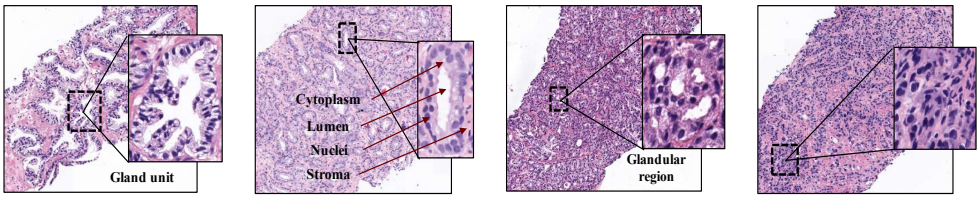


Figure 2.1: Histological prostate-specific patterns according to the Gleason scale. **(a)** Grade 2 (normal). **(b)** Grade 3. **(c)** Grade 4. **(d)** Grade 5.

2.1.1 Related work

Currently, the biopsy analysis entails a considerable subjectivity level between pathologists, besides a large workload. For this reason, there are a lot of studies in the state of the art whose goal is to provide automatic models capable of reporting an initial indicative diagnosis from histopathological images. In most of these studies, the authors tried to perform computer-aided prognosis to discern between all Gleason grades [78–81] or simply between cancer and non-cancer [69, 82, 83]. Nevertheless, we propose, for the first time, a gland classification system exclusively focused on patterns corresponding to moderately differentiated carcinomas. Our aim is to the pathologists to reduce the workload and the subjectivity level when they try to diagnose this type of heterogeneous structures.

In the literature, we can mainly find three different image-processing strategies for encoding the relevant information of the histological images: i) construction of patches, ii) detection of regions of interest (ROIs) and iii) segmentation of gland units. Note that the kind of descriptors to extract the key information from the image is highly dependent of the selected approach.

Image processing associated with the construction of patches lies in dividing the whole-slide image (WSI) into different sub-images from which to extract discriminant information. It is the simplest approach, as it evaluates the histological image without any previous identification of ROIs. In [84] the authors implemented an initial sliding window protocol of 512×512 pixels

with a 10% of overlap, from 12 histological scenes of 2295×4407 pixels. Later, using sub-regions of 100×100 pixels and applying textural and morphological descriptors, they achieved an accuracy of 0.79 in the distinction of stroma, normal and cancer tissues. In another study [85], the researchers analysed the fractal dimension of sub-bands derived from 1000 patches obtained with a magnifying factor of 40×. Support Vector Machine (SVM) classifiers were used to achieve an accuracy of 0.86 in the Gleason classification system. In [81], Farooq et al. reported an overall accuracy of 98.3% in the Gleason grading of 268 images of 2448×3264 pixels. They applied a k-nearest neighbour classifier (K-NN) using textural features based on Local Binary Patterns (LBPs) and Gabor Filters. In a more recent study [69], Esteban et al. used 45 WSIs from which they composed sub-regions of three different dimensions: 512×512, 1024×1024 and 2048×2048 pixels. The best accuracy (0.98) was achieved by applying a linear SVM classifier on the 1024×1024 images.

The strategy based on ROI identification enables the extraction of more specific features. Notwithstanding, it entails a clear limitation, as it requires either a manual identification, which supposes a tedious workload, or a previous automatic identification, which assumes an additional error during the process. This approximation is widely used in the state of the art because it allows for the training of a model exclusively based on relevant information. In [86], the authors tried to distinguish between stroma, benign epithelium, Gleason grade 3 and Gleason grade 4. They extracted morphological and textural features from 54 labelled tissue patches at 40× optical magnification, which resulted in an accuracy of 0.85. In [79], Tabesh et al. made use of 268 colour images from representative areas of hematoxylin and eosin to address the Gleason grading. They utilised colour, texture and morphometric features and achieved an accuracy of 0.81 by means of the K-NN and SVM classifiers. Arvaniti et al. [70] developed a Gleason grading system directly using a dataset of Gleason annotations that contained 5 tissue microarrays (TMAs) at 40× magnification. The authors divided each annotated TMA spot (3100×3100 pixels) in several patches (750×750 pixels), discarding those with multiple annotations. They trained a convolutional neural network (CNN) based on the MobileNet architecture. Then, the patches were reconstructed to the original spot size to evaluate them in terms of average recall. The reported results showed an accuracy of 0.7 concerning to the Gleason grading task.

The strategy related to gland unit segmentation is similar to the previous one, as it also lies in identifying specific parts of the tissue that contain information of interest. Whereas in ROI-based detection, several patterns associated to different classes must be found, e.g. a group of poorly defined

glands or an accumulation of nuclei, in the gland unit segmentation approach, a single type of structures (individual glands) should be segmented along the whole tissue. In [82], the authors applied a KNN classifier on an imbalanced dataset of 199 images and they reported an accuracy of 0.95 per image, only discerning between benign and malignant tissues. First, they used texture-based techniques to segment the prostate glands and they extracted several shape features to address the classification process. In [83], Kwak et al. performed a tissue segmentation using 189 features based on the intensity and texture of the pixels. From the segmented regions of lumens and nuclei structures, the authors applied a total of 670 morphological features to detect prostate cancer. They reported an Area Under the ROC Curve (AUC) of 0.98 using a multi-view boosting classifier. In another recent study [87], Leo et al. made use of the Nuclei-Lumen Association (NLA) algorithm, proposed in [88], to carry out the gland unit segmentation. Then, the authors extracted features based on graphs and texture, as well as shape and orientation disorder of the glands, to distinguish between cancerous and non-cancerous regions. Leo et al. finally reported an AUC of 0.98 by applying random forest classifiers on a database composed of 398 regions at 20× and 40× optical magnification. Similar state-of-the-art studies [78, 80, 88] advocate for a similar strategy to ours, as they used the lumen area as a starting point to address the gland unit segmentation. This approach also entails the segmentation of specific objects called "artefacts", which are characterised by presenting a very similar structure and colour to the lumen elements. However, artefacts are differentiated because they are not surrounded by cytoplasm and epithelial nuclei components, unlike the lumens. Naik et al. [78] implemented a Bayesian classifier to detect the lumens and perform the gland segmentation from them. Nevertheless, the results showed important limitations, as the gland boundaries were delimited by the cytoplasm structure, instead of by the nuclei components according to the medical literature [80]. From the previously segmented glands, Naik et al. applied morphological and textural features to a database composed of 44 images with an optical magnification of 40×. They reported an accuracy of 0.86 using a SVM classifier to distinct between benign and Gleason grade 3 tissues. In contrast, Nguyen et al. [88] developed the NLA algorithm to segment the gland units, which consisted in the linear union of the nuclei surrounding the lumen of a prostate gland. After the segmentation stage, the database was composed of 525 artefacts, 931 benign glands and 1375 cancer glands. The authors extracted a total of 22 features based on the context of each gland candidate and they reported an accuracy of 0.77 by applying a SVM classifier to distinguish between the three classes. Note that, unlike the classification processes of the aforementioned works focused on a gland-unit level, the validation of the proposed methodologies were performed in a

patch or image-wise way. The study performed in [88] is one of the few studies (along with ours) in which the predictive models were created and evaluated from the individual gland candidates, instead of from regions. This strategy entails lower overall accuracy, but more reproducible and reliable results. The study performed in [89] is another example where the authors also carried out a gland classification by discriminating between benign and pathological glands. However, in that case, the artefacts elements were not considered. The researchers extracted colour distributions, textural and structural features from the previously segmented 159 benign and 108 pathological glands. They finally reported an accuracy of 0.86 via SVM and boosted trees classifiers.

2.1.2 Contribution of this work

We present in this paper an extended version of our previous work [90], in which we made a comparison between traditional hand-driven and deep learning techniques for histological glands classification. The main difference with respect to all aforementioned studies is that we only focus on the identification and classification of individual glands what correspond to moderately differentiated carcinomas. This is because such individual glands, in spite of presenting a relatively homogeneous structure, contain the necessary information to discern between normal and pathological tissues with a low grade of cancer, according to the stipulated by the specialists in pathological anatomy. Note that the discrimination between normal and cancerous tissues in a first stage could be decisive in the treatment of the patient.

We address the classification task from the individual gland candidates automatically segmented by means of the *Locally Constrained Watershed Transform* (LCWT) [91], which was applied for the first time on histological images in our previous work [57]. The proposed LCWT algorithm reported better results, in terms of intersection over union (Jaccard index), than the popular NLA algorithm [88] which, in turn, outperformed the methods developed in [78, 92].

Regarding the approach related to the hand-driven learning, we incorporate a novel hand-crafted feature extraction stage based on four global kinds of descriptors. Besides morphological and textural-based descriptors applied on most of the studies [78, 82, 84, 86, 87], we also use descriptors related to the fractal dimension, like [79, 85, 93], and contextual features, similarly to [80, 88]. Until now, each kind of descriptor had been implemented separately, but in this paper, we build a hybrid fingerprint able to encode all the relevant information included in the gland units.

According to the state-of-the-art studies [70, 94, 95], which demonstrated the high viability of the deep learning techniques in the field of the histopathological prostate image, we also include in this paper a convolutional neural network (CNN) to detect the first stage of prostate cancer. Notably, to the best of the author's knowledge, this is the first study that purely use the gland candidates as an input of a CNN, instead of patches or ROIs.

Another contribution of this work is the presentation of a new database composed of prostate glands candidates, which is divided per categories into: artefacts, benign and pathological glands. Note that we provide two images per gland candidate: the bounding box enveloping the candidate and the same bounding box, but masked according to the outputs from the gland segmentation stage.

2.2 Material

The database used in this paper consists of 35 WSIs (17 corresponding to healthy tissues and 18 containing tumour prostate areas) like the one shown in Figure 2.2 (a), which was pixel-wise annotated by an expert pathologist of the *Hospital Clínico Universitario de València*. The samples belong to 8 healthy and 17 patients with prostate cancer in an initial stage. From each slide, we performed a pre-processing step to identify the bounding box containing tissue information (Figure 2.2 (a)). Once the useless information was removed, we implemented a sliding window protocol to work with sub-images of reduced size that allow for improved performance in terms of resolution and local information (Figure 2.2 (b)). In particular, we divided the detected bounding box from the WSI in patches of 1024×1024 pixels ($10\times$ magnification). Next, aimed at improving the computational cost without affecting the study, we applied another processing step to discard those patches with less than 5% of tissue pixels. This resulted in a database (prior to the segmentation process) composed of 854 benign and 614 Gleason grade 3 sub-images of 1024×1024 pixels (see Figure 2.2 (c)).

After the segmentation stage, we obtained the final database for the classification task, which contains 3.195 benign glands, 3.000 cancerous glands of Gleason grade 3 and 3.200 artefacts (false glands). Note that we randomly selected some artefacts from the total obtained (22045) in order to balance the number of samples per class. The resulting database, including original gland candidates and the corresponding segmented masks, can be downloaded from `https://cvblab.synology.me/PublicDatabases/ProstateGlandDB.zip`.

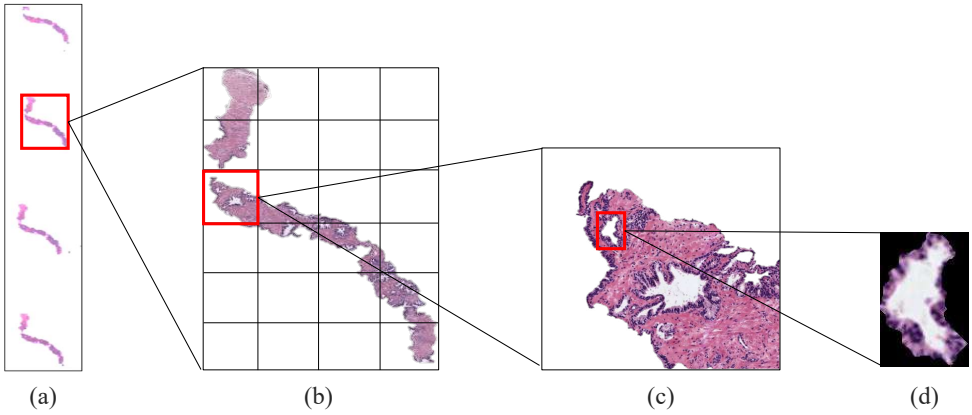


Figure 2.2: Material. **(a)** Example of a WSI. **(b)** Region of interest to perform the sliding window protocol. **(c)** Sub-image of 1024×1024 pixels to address the segmentation task. **(d)** Gland candidate achieved after applying the LCWT segmentation method.

2.3 Methods

The methodology implemented in this paper is represented in the diagram exposed in Figure 2.3, in which we show the two strategies carried out from the segmented gland candidates. Regarding the hand-driven learning approach, we initially applied a hand-crafted feature extraction stage based on four families of descriptors. Then, we performed a statistical analysis to select the best features in terms of correlation and discriminatory ability. Once the normalised matrix of key features was extracted, we carried out a data partitioning to divide the items into different sets in order to build reliable predictive models for the classification stage. We achieved results per gland from two different machine learning classifiers. Concerning the deep learning approach, we directly addressed the data partitioning from the previous segmented gland candidates. We created 5 sets of images corresponding to such candidates, which constitute the input for the implemented CNN. Specifically, the network includes both the feature extraction stage, defined by the combination of four convolutional blocks, and the classification stage composed of two fully-connected layers. Finally, the performance of the proposed hand-driven and deep learning approaches was compared and best model was selected to predict samples from new patients.

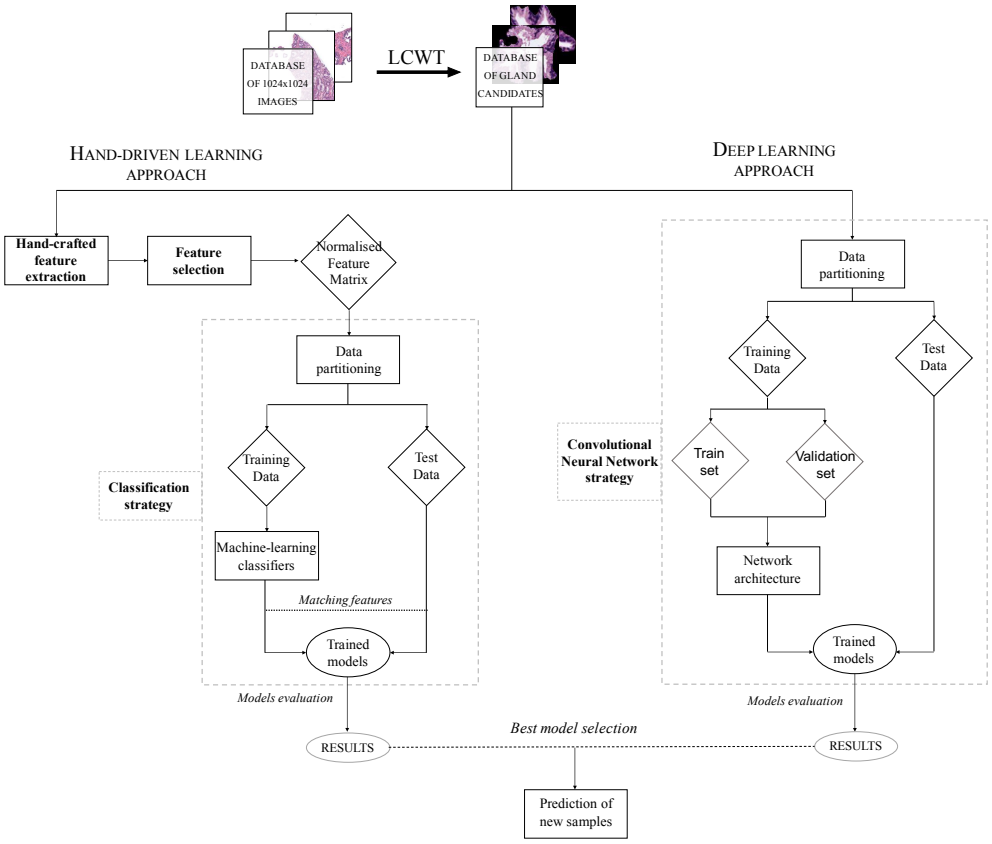


Figure 2.3: Flowchart of the two proposed learning approaches to perform the gland candidates classification.

Background. The first step was to separate the four main components that appear in the H&E prostate images, i.e. lumens, nuclei, cytoplasm and stroma, in order to compute different features from each component, as well as from the relation between them. Similarly to [87, 88], we applied clustering algorithms based on the k-means technique to carry out the identification of each tissue component. However, unlike the previous studies that only used the RGB image as an input for the clustering step, we used different colour spaces depending on the component mask that we wanted to extract. In particular, we made use of the saturation channel S_{HSV} from the HSV (Hue, Saturation and Value) colour space to detect the lumen objects; the cyan channel C_{CMYK} from the CMYK (Cyan, Magenta, Yellow and Key) colour space to identify

the cytoplasm and stroma components; and a reshaped RGB image V_{RGB} to achieve the nuclei elements, as depicted in Figure 2.4 (a). From an initial RGB image I_{RGB} of dimensions 1024×1024, we applied a reduction factor of 50% and performed the colour transformations to obtain each one of the four candidate maps. Once the different channels were computed, we selected the number of clusters k in which the pixels of I_{RGB} should be grouped. Specifically, we established $k = 3$ to obtain the map of lumen candidates L_m, as well as the maps of the cytoplasm C_m and stroma S_m candidates. In contrast, we set $k = 4$ to acquire the nuclei map N_m. We ran k-means algorithm on the S_{HSV}, C_{CMYK} and V_{RGB} images to group the pixels of each image into k different classes according to their intensity level. After the clustering stage, we obtained three labelled images LI (one from each colour space), whose pixels can take values from 1 to k, depending on the previous unsupervised pixel classification (see Figure 2.4 (b)). From here, we carried out a thresholding of each LI image by determining the value of the pixels associated with a specific tissue component, as shown in Figure 2.4 (c). For instance, to achieve the lumen structure L_m, we binarised the LI_{HSV} image according to Eq. 2.1, as the darkest pixels of the LI_{HSV} correspond to the lumen structures.

$$L_m = \begin{cases} 1, & \text{if } LI_{HSV} = min(LI_{HSV}) \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

Additionally, we performed a post-processing stage, from the outputs of the clustering phase, to reduce noise and provide the final binary maps: L_{map}, C_{map}, S_{map} and N_{map}, exposed in Figure 2.4 (d). We applied different morphological operations depending of the tissue map. To achieve the map of lumen candidates, we implemented a filtering operation called *area opening* (γ^a_λ) from the set of pixels $X \subseteq L_m \subset \mathbb{R}^2$. This morphological operation is defined by the union of all connected components of X whose area is greater than a specific number of pixels λ, according to Eq. 2.2:

$$\gamma^a_\lambda(X) = \bigcup \{X_i \mid i \in I,\ Area(X_i) \geq \lambda\} \tag{2.2}$$

After applying the area opening filter with a specific $\lambda_{lum} = 20$ pixels, we implemented a *dilation* operation described as: $\delta_B(X) = X \oplus B$, where B is a structuring element (SE) with radius $r = 1$ and disk shape. Regarding the maps of cytoplasm C_m and stroma S_m, we carried out a filtering phase based on an *opening* operation defined by $\gamma_B(X) = (X \ominus B) \oplus B$, followed by another *area opening* with $\lambda_{cyto} = \lambda_{str} = 20$ pixels to discard the non-consistent sets of

pixels. With respect to the nuclei elements, we first applied a *dilation* using the SE B to emphasise the importance of the nuclei around the gland. Then, we again implemented an *area opening* γ_λ^a with $\lambda_{nuc} = 20$ to discard non-epithelial nuclei scattered across the stroma region.

Note that both clustering and post-processing algorithms were carried out from the 1024×1024 RGB images. The resulted binary maps of each tissue component was employed to address the segmentation stage via LCWT algorithm [57]. Then, LCWT outcomes were used as a starting point of the proposed classification framework aimed at distinguishing between artefacts, benign glands and pathological glands.

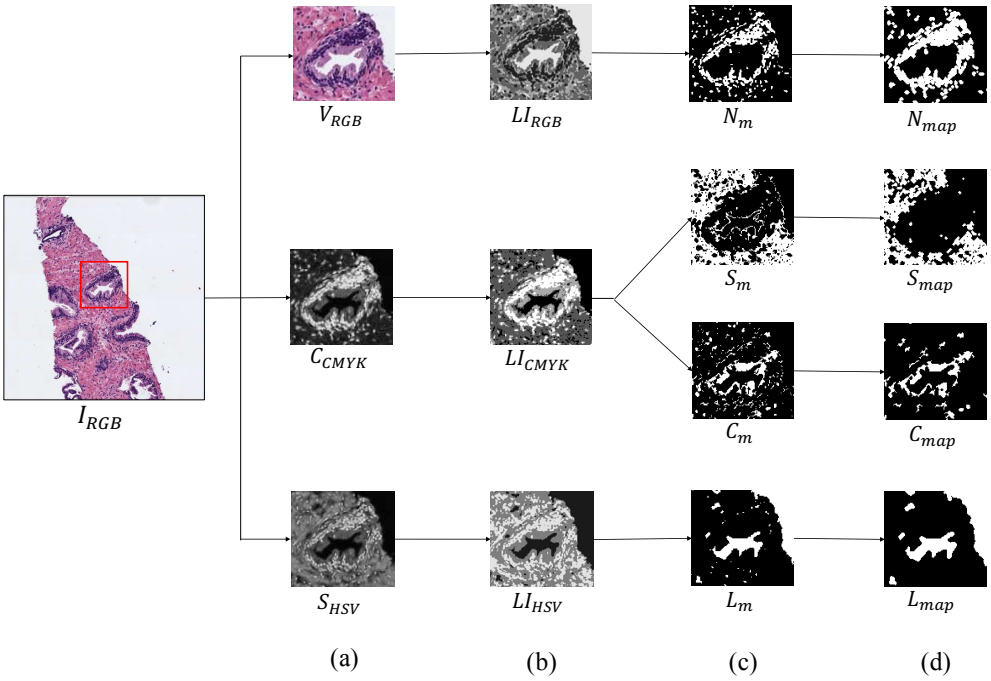


Figure 2.4: Process to obtain the binary map of each tissue component. **(a)** Outputs after the colour space transformations. **(b)** Labelled images achieved from the clustering stage. **(c)** Masks obtained after the binarisation. **(d)** Final maps of each tissue component.

2.3.1 Hand-driven learning approach

2.3.1.1 Feature extraction

This stage supposes one of the most remarkable novelties of this paper since, for the first time, we include information relative to the combination of 241 features extracted from four different families of descriptors.

Morphological descriptors. Several state-of-the-art studies demonstrated the viability of this kind of features in the characterisation of the histological prostate image [78, 82, 83]. Therefore, we used a total of 20 features related to the morphology of the glands and their respective lumens, taking into account the differences of the Gleason grades detailed in Section 2.1. From the gland candidates segmented by the proposed LCWT method (see Figure 2.5 (a)), we acquired gland and lumen masks to extract the same 10 shape and geometry-based features from each specific structure. In this way, let $Gland_{RGB}$ be a colour image of dimensions $M \times N$, in which each pixel $p(i,j)$ is denoted by the system (Eq. 2.3), we decomposed the image into its three colour components (R, G, B), and added them according to Eq. 2.4.

$$p(i,j) = \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} R_{i,j} \\ G_{i,j} \\ B_{i,j} \end{bmatrix}, \qquad where \quad [a,b,c] \in [0,255] \tag{2.3}$$

$$RGB_s = \sum_{i=1}^{M} \sum_{j=1}^{N} R_{i,j} + G_{i,j} + B_{i,j} \tag{2.4}$$

We applied the *Otsu* method [96] to identify an optimal threshold for gland mask extraction. Then, we carried out a filtering procedure based on *area opening* with 4-connectivity and $\lambda = 20$ pixels, followed by a flood-fill operation on the background pixels. Once the gland mask G_{mask} (represented by the white pixels in Figure 2.5 (b)) was extracted, we achieved the lumen mask L_{mask} (Figure 2.5 (c)) by identifying the lumen pixel candidates included inside the coordinates of the gland mask previously detected. Note that the name of each morphological feature is accompanied by G or L (gland or lumen) depending on the element under study. To explain the morphological features, we only define below those corresponding to the gland masks, but the same operations were computed for lumen masks.

(a) (b) (c)

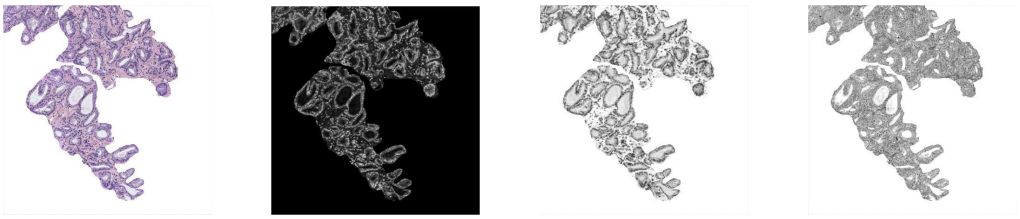
Figure 2.5: Gland processing. **(a)** Segmented Gland. **(b)** Gland mask. **(c)** Lumen mask.

- G_{area}. Number of pixels containing a specific gland candidate.
- $G_{convexArea}$. Number of pixels in the region known as *convex hull* that is defined by the smallest convex polygon around the gland.
- G_{eccent}. Ratio of the distance between the centre of E_G and its major axis length, where E_G is the ellipse adjusted to the gland area with the same second moments.
- $G_{equivDiam}$. Diameter of a circle with the same area as the gland, defined by: $G_{equivDiam} = \frac{4*G_{area}}{\pi}$.
- G_{extent}. Ratio of pixels between G_{area} and the area of the bounding box G_{bBox} that contains the gland. It is computed via $G_{extent} = \frac{G_{area}}{G_{bBox}}$.
- $G_{orientation}$. Angle between the x-axis and the major axis of E_G.
- $G_{perimeter}$. Number of pixels describing the edge of the gland.
- $G_{solidity}$. Proportion of the pixels in the convex hull also included inside the area of the gland. It is described as: $G_{solidity} = \frac{G_{area}}{G_{convexArea}}$.
- $G_{roundness}$. Scalar that measures the compact character of the gland by $G_{roundness} = \frac{R_G * G_{perimeter}}{G_{area}}$, where R_G is the radius of the gland.
- $G_{compactness}$. Scalar indicating how round the gland is via $G_{compactness} = \frac{G_{perimeter}}{\sqrt{G_{area}}}$.

Fractal analysis. Several state-of-the-art studies, such as [79, 83, 85, 93], used features based on different fractal dimensions to address the classification of histological prostate images. We applied, for the first time on this kind of

images, a fractal analysis based on the Hurst exponent H [97]. We made use of three different grey-scale images (cyan, hematoxylin and eosin) in order to take into account the contributions of different colour spaces (see Figure 2.6). Cyan was used because it is the channel in which the differences between the four tissue components can be better differentiated. In addition, we computed the hematoxylin and eosin colour images by means of the colour deconvolution method proposed in [98], which was also implemented in other studies of the state of the art related with this research field [69, 99]. The colour deconvolution method enables the separation of the contributions of each stain by calculating the Optical Density (OD) parameter (Eq. 2.5); where A is the absorbance and C_s is the concentration of a specific stain s.

$$OD_s = A * C_s \tag{2.5}$$



(a) (b) (c) (d)

Figure 2.6: Colour space transformation. **(a)** RGB image. **(b)** Cyan channel. **(c)** Hematoxylin stain. **(d)** Eosin contribution.

For each colour image, we extracted the Hurst exponent H, which is related to the fractal Brownian motion (fBm). This fBm is a Gaussian, self-similar and non-stationary process $B_H(t)$ on [0,T], whose co-variance function (Eq. 2.6) was introduced in [100].

$$\rho(s,t) = E[B_H(t)B_H(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}), \quad \forall H \in [0,1] \tag{2.6}$$

where $0 < s \leq t$ and H corresponds to the aforementioned Hurst exponent, which governs the stochastic representation of the fBm and allows for the measurement of the tortuosity of the images. H is related to the fractal dimension such that $H = E + 1 - FD$, where E is the Euclidean dimension and FD the fractal dimension that takes higher values when the signal is more complex.

Being X the grey-scale bounding box of dimensions $M \times N$ that contains a specific gland candidate (Figure 2.7 (b)), and taking into account that the fBm is a non-stationary process, it is more convenient to analyse the incremental process of the fBm (Figure 2.7 (c)), i.e. the fractional Gaussian noise (fGn) defined as follows:

$$G_t(a) = B_H(a+1) - B_H(a) \tag{2.7}$$

Once the fGn was computed, we calculated the discrete Fourier transform (Eq. 2.8), from the 1D signal $d[n]$ (Eq. 2.9), where m and n denotes the row and column indices, respectively. We expose in Figure 2.7 (d) an example of a 1D signal for $m = 1$.

$$D[a] = \sum_{n=0}^{N-1} d[n] e^{\frac{-j2\pi an}{N}} \tag{2.8}$$

$$d[n] = X[m, n+1] - X[m, n] \tag{2.9}$$

From the fGn function and being $D'[a, b]$ a 2D matrix composed of the M rows corresponding to all 1D signals previously achieved, we calculated the average Power Spectral Density (PSD) as follows:

$$PSD[a] = \frac{1}{M} \sum_{t=0}^{M-1} D'[a, b] \tag{2.10}$$

In a log-log scale, the PSD function of the fGn corresponds to a line of slope $1 - 2H$, which we obtained via linear regression, as shown in Figure 2.7 (e). The Hurst exponent H was finally calculated to determine if the pixels of the gland candidates followed purely random patterns or kept underlying trends. In particular, we considered L = 5 directions to calculate the Hurst exponent along each one of them: $H = \{H^0, H^{30}, H^{45}, H^{60}, H^{90}\}$. Our aim was to cover different patterns in the orientation of the glands. Note that H was extracted from the cyan, hematoxylin and eosin channels, so that 15 features related to the fractal dimension (H_{cyan}, H_{hmtx} and H_{eosn}) were finally computed.

Texture descriptors. In this work, we propose two kinds of descriptors for encoding the textural information related to the artefacts, benign and pathological glands. Specifically, we used Gray-Level Co-occurrence Matrix

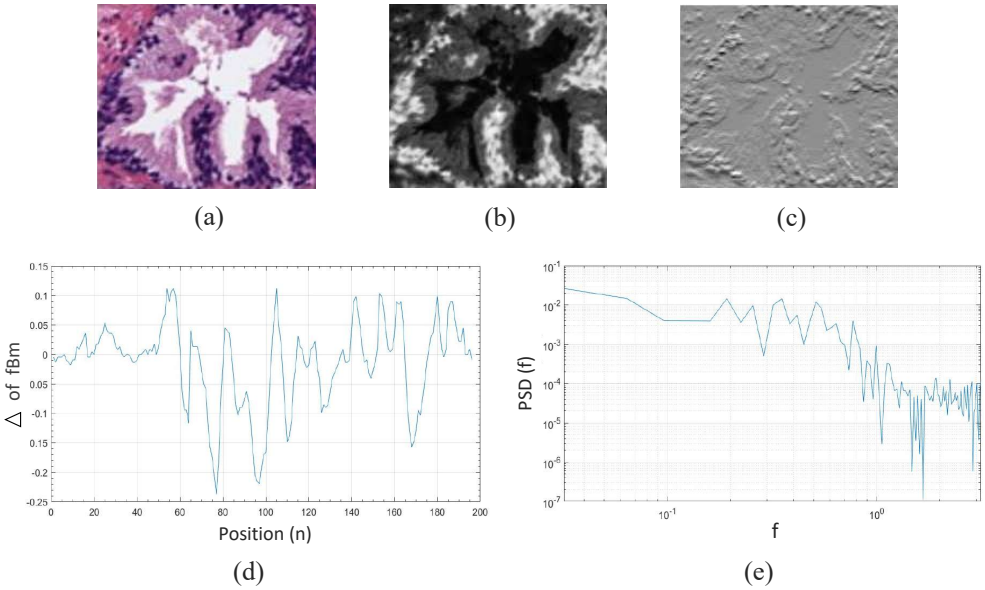


Figure 2.7: Fractal analysis. **(a)** Original bounding box corresponding to a RGB gland candidate image. **(b)** Cyan channel of the specific gland candidate. **(c)** Increments of the fBm corresponding to the fractional Gaussian noise fGn. **(d)** 1D signal calculated from the fGn for $m = 1$. **(e)** PSD of the increments of all rows from the gland candidate.

(GLCM), similarly to Leo et al. [87], who calculated a 18×18 co-occurrence matrix to obtain information about the glands orientation. In other studies such as [84, 86, 101], the authors also applied GLCM-based techniques on histological regions, instead of on gland units. Additionally, we used Local Binary Patterns (LBP) to extract local intensity changes of the gland candidates, unlike other works which used LBP to segment different tissue structures [83, 99] or to discriminate between cancerous and non-cancerous patches [69, 81]. Note that, in this case, we also made use of the cyan, hematoxylin and eosin channels to extract a total of 186 textural features.

Gray-Level Co-occurrence Matrix (GLCM) is a matrix of frequencies, in which it is represented (on the (i,j)-position) the number of times that a pixel with an intensity value i is adjacent to another pixel with intensity value j. During the GLCM creation, we specified the number of adjacent pixels D that must have the intensity value $j = i$, as well as the direction (angle) in which such pixels are considered adjacent. In Figure 2.8 (a), we show an example of a GLCM obtained from a specific image I with an angle of 0^{o} and $D = 1$ pixels.

Particularly, we applied a number of adjacent pixels $D = 2$ and established two different directions corresponding to angles of 0^{o} and 45^{o} to consider the orientation trend of the gland candidates. The two directions are represented by an offset of [0,2], relative to the angle of 0^{o}, and another offset of [-2,2], corresponding to an angle of 45^{o}, as detailed in Figure 2.8 (b). Note that the dimension of the GLCMs computed in this paper for each colour image was always 8×8 pixels.

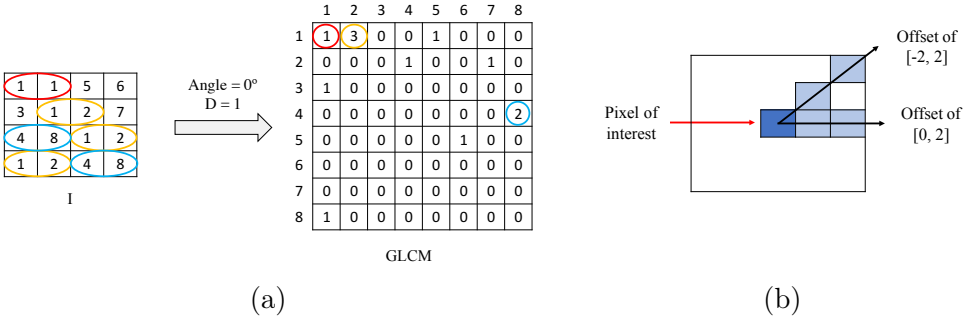


Figure 2.8: GLCM illustration. **(a)** Example of GLCM achieved from a certain image I using an offset of [0,1]. **(b)** Two offset implemented in this paper to obtain each GLCM.

Once the GLCM was obtained, we normalised it via Eq. 2.11, where $sGLCM = GLCM + GLCM^{T}$ corresponds to the symmetric GLCM [102].

$$nGLCM = \frac{sGLCM}{\sum sGLCM} \tag{2.11}$$

From the normalised $nGLCM_r^s$ corresponding to an offset s and a colour image r, we extracted 21 different features. The use of two offset values and three colour images allows for the extraction a total of 126 features related to the GLCM for each gland candidate. We detail below the 21 different variables, for an offset of [0,2], denoted by 0^{o}, and a colour image corresponding to the cyan channel C. Remember that these features are equally extracted from the rest of offsets and colour images.

- *Homogeneity.* It reaches higher values when the occurrence is focused along the normalised GLCM diagonal. Being $p(i, j)$ the probability of the grey-occurrence in each pixel, the homogeneity is calculated as follows:

$$Homo_{nGLCM_C^{0^o}} = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \tag{2.12}$$

- *Contrast*. It measures the local variation of a specific image. It is the opposite to the homogeneity and it is computed by:

$$Cont_{nGLCM_C^{0^o}} = \sum_{i,j} |i-j|^2 \; p(i,j) \tag{2.13}$$

- *Energy*. The energy, a.k.a. the angular second moment, takes smaller values the more similar the inputs are. It is denoted as follows:

$$Ener_{nGLCM_C^{0^o}} = \sum_{i,j} p(i,j)^2 \tag{2.14}$$

- *Correlation*. The correlation indicates how much similar information provides a pixel over the whole image regarding its neighbour, such that:

$$Corr_{nGLCM_C^{0^o}} = \sum_{i,j} \frac{(i-\mu i)(j-\mu j)p(i,j)}{\sigma_i \sigma_j} \tag{2.15}$$

- *Entropy*. It is a measure related to the uniformity of the image. The entropy takes small values when the inputs of $nGLCM$ are close to 0 or 1. It follows the next equation:

$$Entr_{nGLCM_C^{0^o}} = \sum_{i,j} -p(i,j) ln(p(i,j)) \tag{2.16}$$

- $Mean(\mu)$. This feature corresponds to an average by columns of the grey values of the 8×8 $nGLCM$. We obtained $N = 8$ values, according to:

$$\mu^j_{nGLCM_C^{0^o}} = \frac{1}{N} \sum_{i=1}^{N} p(i,j) \tag{2.17}$$

- *Standard deviation*(σ). We also computed 8 values relative to the standard deviation of the 8×8 $nGLCM$ calculated by columns:

$$\sigma^{j}_{nGLCM^{\varrho^{\varrho}}_{C}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}|p(i,j)-\mu|^2} \tag{2.18}$$

Local Binary Patterns (LBP) were also used to recognise local textures and specific shapes in the gland candidates images. The basic LBP operator proposed in [103] consists of computing the difference between the value of the pixel of interest and the value of its neighbours. The pixels under study are binarised to 0 or 1 depending on whether the resultant values are negative or positive, as follows:

$$LBP_{P,R}(i,j) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \quad s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{2.19}$$

where P is the number of neighbour pixels with a grey value g_p inside a circle of radius R, respecting to the central pixel $p(i,j)$ with a grey value g_c. In this way, from the created binary string, we calculated the new value of the pixel of interest by performing a conversion to a decimal value. However, we did not implement the basic LBP, but we used the $LBP^{riu2}_{P,R}$ operator (Eq. 2.20), proposed by Ojala et al. in [104]. This is characterised by being uniformly invariant to rotation transforms for grey-scale images. $LBP^{riu2}_{P,R}$ operator allows for the calculation of P+2 different output values [105]. Specifically, we used $P = 8$ and $R = 1$ to extract 10-bin $LBP^{riu2}_{P,R}$ histograms from each colour image corresponding to each gland candidate. As we made use of the cyan, hematoxylin and eosin colour channels, we extracted a total of 30-bin histograms uniformly invariant to rotation transforms.

$$LBP^{riu2}_{P,R}(i,j) = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) & \text{if } U(LBP_{P,R}) \leq 2 \\ P+1 & \text{otherwise} \end{cases} \tag{2.20}$$

where,

$$U(LBP_{P,R}) = |s(g_{p-1}-g_c)-s(g_0-g_c)| + \sum_{p=1}^{P-1}|s(g_p-g_c)-s(g_{P-1}-g_c)| \tag{2.21}$$

In addition, we introduced the operator *Rotational Invariant Local Variance* (VAR) (Eq. 2.22), which is commonly implemented in combination to the $LBP^{riu2}_{P,R}$, as it is invariant to contrast changes.

$$VAR_{P,R}(i,j) = \frac{1}{P}\sum_{p=0}^{P-1}(g_p - \mu)^2, \quad where \quad \mu = \frac{1}{P}\sum_{p=0}^{P-1} g_p \tag{2.22}$$

From the $LBP^{riu2}_{P=8,R=1}$ and $VAR_{P=8,R=1}$ images of dimensions $M \times N$, we extracted the LBP variance (LBPV) histogram (Eq. 2.23), which was proposed by Guo et al. [106]. It consists in the accumulation of the $VAR_{P,R}$ value for each $LBP^{riu2}_{P,R}$ label according to:

$$LBPV_{P,R}(k) = \sum_{i=1}^{M}\sum_{j=1}^{N} w(LBP_{P,R}(i,j),k), \quad k \in [0,K] \tag{2.23}$$

where,

$$w(LBP_{P,R}(i,j),k) = \begin{cases} VAR_{P,R}(i,j) & \text{if } LBP_{P,R}(i,j) = k \\ 0 & \text{otherwise} \end{cases} \tag{2.24}$$

In particular, we computed 10-bin $LBPV$ histograms for each colour image (cyan, hematoxylin and eosin). In Figure 2.9, we show an example of the $LBPV_{8,1}$ variable extracted from the cyan channel of the three types of gland candidates, i.e. an artefact, a benign gland and a Gleason grade 3 gland. Taking into account the 30-bin histograms corresponding to the $LBP^{riu2}_{8,1}$ and the 30-bin histograms relative to the $LBPV_{8,1}$, we achieved a total of 60 features related to the LBP descriptor.

Contextual features. Nguyen et al. [80, 88] used structural features to address a classification problem from two different approaches. Particularly, in [80], the authors extracted 15 structural variables from the previous detected glandular regions. However, such study was performed exclusively making use of 82 ROIs carefully selected by hand, whereas in our work we propose an automatic end-to-end approach. We extracted a total of 20 hand-crafted features related to the context of each gland candidate image. For explanation purposes, we group the features into three different sets of features.

The first set contains 10 features related to the nuclei components:

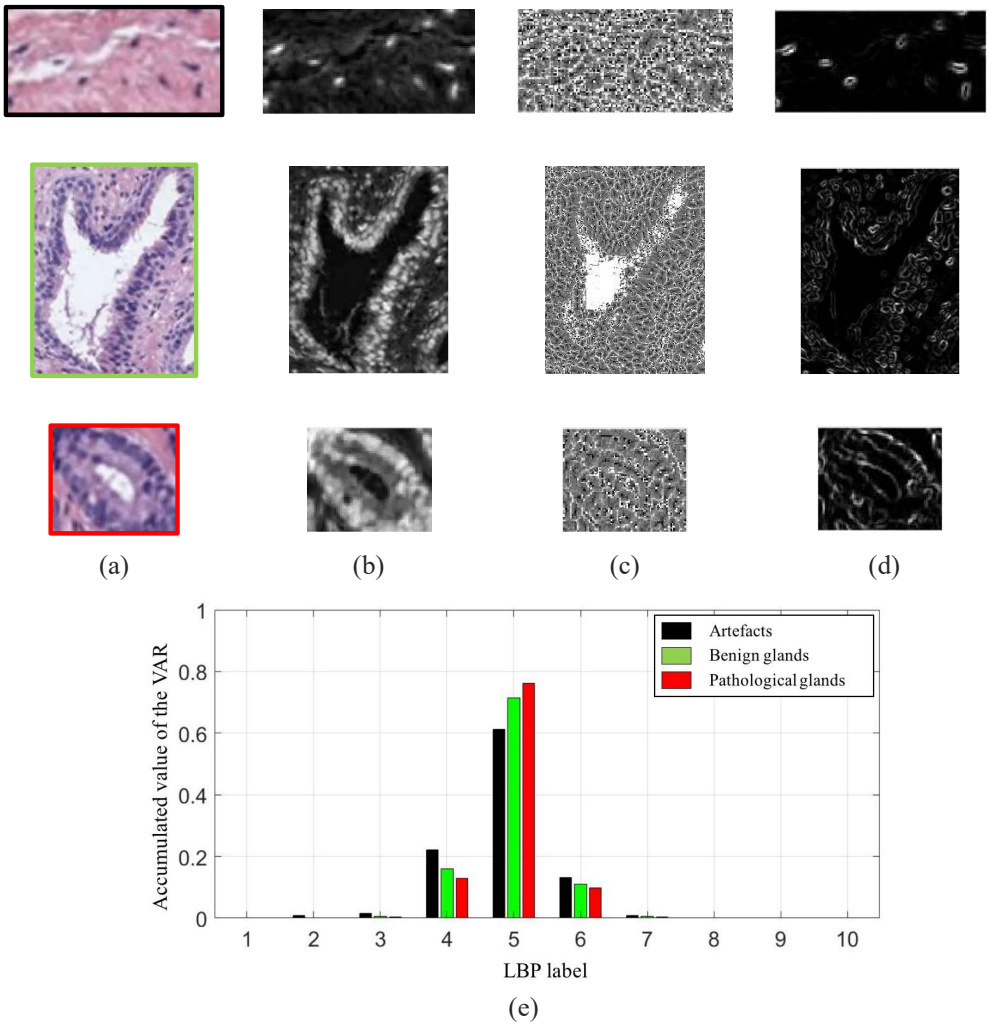


Figure 2.9: LBP-based feature extraction. **(a)** Gland candidates related to an artefact, a benign and a pathological gland highlighted in black, green and red, respectively. **(b)** Cyan channel of the the gland candidates. **(c)** $LBP^{riu2}_{8,1}$ image. **(d)** $VAR_{8,1}$ image. **(e)** 10-bin histograms of the $LBPV_{8,1}$ after combining the images (c-d).

- $nuclei^{num}_{BB}$ Eq. 2.25. It corresponds to the quantity of nuclei elements inside the area of the bounding box of the segmented gland G^{area}_{bBox}.

$$nuclei_{BB}^{num} = \sum NE, \qquad where \;\; NE \subseteq G_{bBox} \tag{2.25}$$

where $NE = \{p_1, p_2, ..., p_T\} \subseteq G_{bBox}^{area}$ is a group of all nuclei elements p_i with 8-connectivity inside the G_{bBox}^{area}

- $nucleiRatio_{BB}^{num}$ Eq. 2.26. It refers to the ratio between the number of nuclei elements and the G_{bBox}^{area}.

$$nucleiRatio_{BB}^{num} = nuclei_{BB}^{num} / G_{bBox}^{area} \tag{2.26}$$

- $nuclei_{Gland}^{num}$ Eq. 2.27. It corresponds to the number of nuclei objects inside the gland candidate area G_{area}, instead of inside the G_{bBox}^{area}.

$$nuclei_{Gland}^{num} = \sum NE, \qquad where \;\; NE \subseteq G_{area} \tag{2.27}$$

- $nucleiRatio_{Gland}^{num}$ Eq. 2.28. This refers to the ratio betewen the number of nuclei objects and the gland candidate area G_{area}.

$$nucleiRatio_{Gland}^{num} = nuclei_{Gland}^{num} / G_{area} \tag{2.28}$$

- $nucleiRatio_{Gland-BB}^{num}$ Eq. 2.29. It denotes the number of nuclei elements inside the G_{area} in relation to the number of nuclei elements inside the G_{bBox}^{area}.

$$nucleiRatio_{Gland-BB}^{num} = nuclei_{Gland}^{num} / nuclei_{BB}^{num} \tag{2.29}$$

Notably that the remaining 5 variables related to the nuclei components were calculated in the same way as before, but taking into account the number of pixels of the nuclei elements, instead of the number of nuclei elements themselves. Therefore, being $pix = \sum_{i=1}^{T} p_i$, where $p_i \subseteq NE$, we extracted the following features: $nuclei_{BB}^{pix}$, $nucleiRatio_{BB}^{pix}$, $nuclei_{Gland}^{pix}$, $nucleiRatio_{Gland}^{pix}$ and $nucleiRatio_{Gland-BB}^{pix}$.

Regarding the second set of variables, we computed 5 features relative to the cytoplasm structure. In a similar way as before, we extracted the variables from the pixels of the cytoplasm component to obtain: $Cyto_{features} = \{cyto_{BB}^{pix}, \; cytoRatio_{BB}^{pix}, \; cyto_{Gland}^{pix}, \; cytoRatio_{Gland}^{pix}, \; cytoRatio_{Gland-BB}^{pix}\}$.

Finally, the third set of variables corresponds to 5 contextual features associated with specific relations between the lumen, nuclei and cytoplasm components, as we detail below:

- $Ratio^{pix}_{L-G}$ Eq. 2.30. It corresponds to the proportion between the lumen and gland areas, in terms of number of pixels.

$$Ratio^{pix}_{L-G} = L_{area}/G_{area} \tag{2.30}$$

- μ_{L-Edge} Eq. 2.31. It refers to the euclidean average distance between the centroid of a lumen and the pixels of its edge.

$$\mu_{L-Edge} = \frac{\sum_{i=1}^{N} \sqrt{(c_x - x_i)^2 + (c_y - y_i)^2}}{N} \tag{2.31}$$

  where N is the total number of pixels of the lumen edge, (c_x, c_y) are the coordinates corresponding to the centroid of the lumen and (x_i, y_i) are the coordinates of the pixel i relative to the lumen edge.

- σ_{L-Edge} Eq. 2.32. It is the standard deviation of the euclidean distance between the centroid and the edge of each lumen.

$$\sigma_{L-Edge} = \frac{1}{N-1} \sqrt{\sum_{i=1}^{N} |v_i - \mu_{L-Edge}|^2} \tag{2.32}$$

  where,

$$v_i = \sqrt{(c_x - x_i)^2 + (c_y - y_i)^2} \tag{2.33}$$

- Tor_{C-N} Eq. 2.34. It corresponds to the number of pixels of the nuclei N_{mask} and cytoplasm C_{mask} contained in a toroid region $Toroid$ achieved by subtracting the gland G_{mask} and lumen L_{mask} masks.

$$Tor_{C-N} = \sum Toroid \cap (C_{mask} + N_{mask}) \tag{2.34}$$

- $ToroidRatio_{C-N}$ Eq. 2.35. It is the ratio between the number of pixels in C_{mask} and N_{mask} inside the $Toroid$ with respect to the area of such region Tor_{area}.

$$RatioTor_{C-N} = Tor_{C-N}/Tor_{area} \tag{2.35}$$

2.3.1.2 Feature selection

Once the 241 variables were computed and stored, we addressed an exhaustive statistical analysis based on parametric and non-parametric tests to select the most relevant features, in terms of independence between pairs of variables and dependence between each feature and the ground truth label. The first step was to normalise the variables, for assigning the same relevance to each of them, according to the *z-score* parameter described by: $z_i = \frac{x_i - \mu}{\sigma}$. z_i is the normalised number of the x_i value for a specific variable with mean μ and standard deviation σ. After normalising the variables, we performed the non-parametric *Kolmogorov-Smirnov test* to carry out a hypothesis contrast in which the null hypothesis H_0 maintains that the variables follow a normal distribution N(0,1). In Figure 2.10 (a), we represent the histogram concerning the first 20 computed variables over a Gaussian function, a.k.a. Gauss bell, which characterises the normal distribution N(0,1) to evidence the differences between both distributions. Depending on the previous statistical test, we performed another contrast hypothesis aimed at analysing the discriminatory ability of each variable with respect to the class. We carried out a comparison of means making use of the *ANOVA* test (ANalysis Of VAriance) if the variable follows a distribution N(0,1), or a comparison of medians using the *Kruskal-Wallis* test otherwise. Using a significance level $\alpha = 10^{-6}$, which denotes a confidence value of $99,9999\%$, we compared α with the *p-value* obtained from the *ANOVA* or *Kruskal-Wallis* test to determine the independence level between each variable and its class. Finally, if *p-value* $\leq \alpha$, we rejected the null hypothesis H_0 that holds the independence variable-class, as there is a significant evidence to ensure that the variable under study is dependent on the class. Therefore, this variable is selected due to its high discriminatory ability. Contrarily, we discarded the variables whose *p-value* was greater than α, as they were not class dependent. As depicted in Figure 2.10 (b), we show the box plot corresponding to the first 10 variables to qualitatively evidence the differences between the values obtained by each feature depending on the ground truth label.

In addition, we performed a third hypothesis contrast to analyse the independence level between pairs of variables. We discarded the features correlated with others to avoid redundant information. To do this, we calculated the Pearson's correlation coefficient r and the *p-value* from the correlation matrix to remove those variables that meet both *p-value* $\leq \alpha$ and $|r| \geq 0.95$. In Figure 2.10 (c), we show an example of the correlation matrix obtained during the feature selection stage. Finally, from the 241 variables, we achieved a total of 136 relevant features that we list below in Table 2.1.

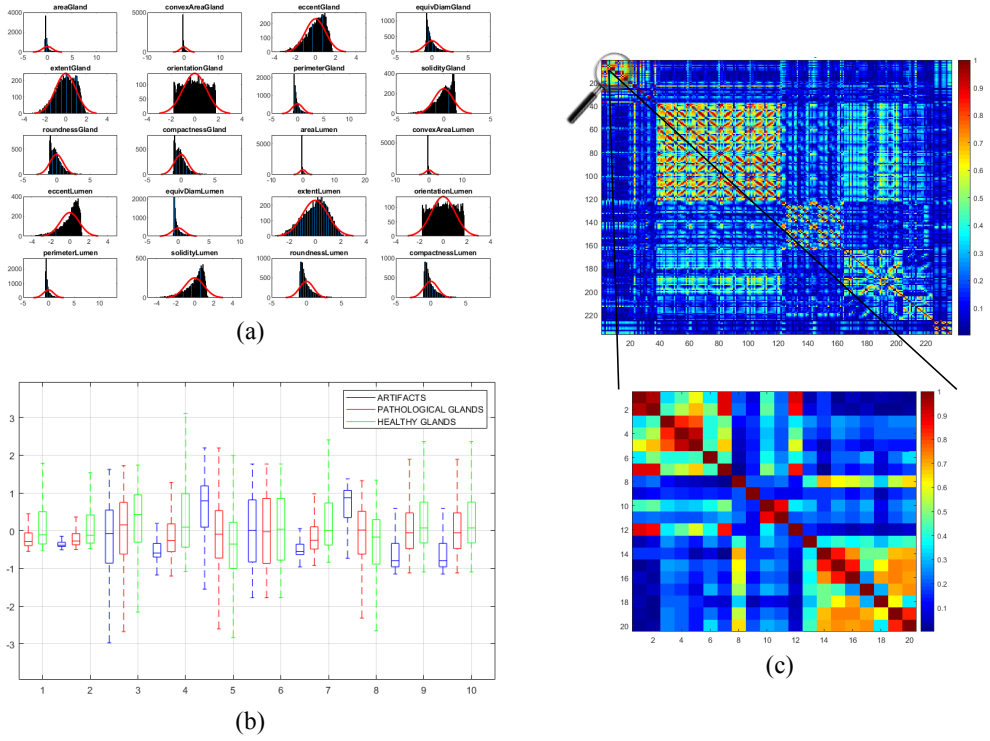


Figure 2.10: Qualitative outcomes from the feature selection stage.**(a)** Bar chart corresponding to the distribution of the first 20 variables overlapped on the Gaussian bell represented in red. **(b)** Box plot relative to the discriminatory ability of the first 10 variables. **(c)** Correlation matrix evidencing the independence level between pairs of variables.

Table 2.1: Features selected from the statistical analysis.

Morphological features (13)	*Gland (6)*	G_{area}, G_{eccent}, $G_{equivDiam}$, G_{extent}, $G_{solidity}$, $G_{roundness}$
	Lumen (7)	L_{area}, L_{eccent}, $L_{equivDiam}$, L_{extent}, $L_{perimeter}$, $G_{solidity}$, $G_{roundness}$
Fractal analysis (11)	*Cyan (3)*	H^{30}_{cyan}, H^{45}_{cyan}, H^{60}_{cyan}
	Hematoxylin (5)	H^{0}_{hmtx}, H^{30}_{hmtx}, H^{45}_{hmtx}, H^{60}_{hmtx}, H^{90}_{hmtx}
	Eosin (3)	H^{30}_{eosn}, H^{45}_{eosn}, H^{60}_{eosn}
Textural features (94)	*Cyan (32)*	$Homo_{nGLCM^{0}_{C}}$, $Cont_{nGLCM^{0}_{C}}$, $Ener_{nGLCM^{0}_{C}}$, $Corr_{nGLCM^{0}_{C}}$, $\mu^{j}_{nGLCM^{0}_{C}}(1-3;7-8)$, $Homo_{nGLCM^{45}_{C}}$, $Cont_{nGLCM^{45}_{C}}$, $Corr_{nGLCM^{45}_{C}}$, $LBP^{riu2}_{8,1}(1-10)$, $LBPV_{8,1}(1-10)$
	Hematoxylin (27)	$Homo_{nGLCM^{0}_{H}}$, $Ener_{nGLCM^{0}_{H}}$, $Corr_{nGLCM^{0}_{H}}$, $\mu^{j}_{nGLCM^{0}_{H}}(1-2;7)$, $Corr_{nGLCM^{45}_{H}}$, $LBP^{riu2}_{8,1}(1-10)$, $LBPV_{8,1}(1-10)$
	Eosin (35)	$Homo_{nGLCM^{0}_{E}}$, $Cont_{nGLCM^{0}_{E}}$, $Corr_{nGLCM^{0}_{E}}$, $\mu^{j}_{nGLCM^{0}_{E}}(1-8)$, $Homo_{nGLCM^{45}_{E}}$, $Cont_{nGLCM^{45}_{E}}$, $Corr_{nGLCM^{45}_{E}}$, $Ener_{nGLCM^{45}_{E}}$, $LBP^{riu2}_{8,1}(1-10)$, $LBPV_{8,1}(1-10)$
Contextual features (18)	*Nuclei (9)*	$nuclei^{num}_{BB}$, $nucleiRatio^{num}_{BB}$, $nuclei^{num}_{Gland}$, $nucleiRatio^{num}_{Gland}$, $nucleiRatio^{num}_{Gland-BB}$, $nuclei^{pix}_{BB}$, $nucleiRatio^{pix}_{BB}$, $nucleiRatio^{pix}_{Gland}$, $nucleiRatio^{pix}_{Gland-BB}$
	Cytoplasm (5)	$cyto^{pix}_{BB}$, $cytoRatio^{pix}_{BB}$, $cyto^{pix}_{Gland}$, $cytoRatio^{pix}_{Gland}$, $cytoRatio^{pix}_{Gland-BB}$
	Lumen-Nuclei-Cytoplasm (4)	$Ratio^{pix}_{L-G}$, μ_{L-Edge}, σ_{L-Edge}, $RatioTor_{C-N}$

2.3.1.3 Classification Strategy

Data partitioning. From the selected hand-crafted feature matrix, we divided the different items (rows) into 5 datasets taking into account diverse criteria. On the one hand, we included a similar number of items in each set (fold), attending to the class of such items to create balanced groups. On the other hand, performed a partitioning separating the data according to the medical history of the patients. Thus, all the gland candidates features corresponding to the same patient were stored in the same fold. Then, we carried out a nested cross-validation strategy to remove the randomness effect in the data partitioning. First, we performed an external *5-fold cross-validation* to train the models using four partitions and validate their performance with the remaining fold. This process was repeated 5 times to ensure that all the samples were used to train and test. Additionally, in each iteration, we also applied an internal *10-fold cross-validation* to optimise the parameters of the classifiers using the 10% of the training data as a validation set.

Machine learning classifiers. In order to address the classification stage through hand-driven learning methods, we made use of two different classifiers widely used in the literature [69, 80, 83, 87]. Specifically, we optimised a Support Vector Machine (SVM) classifier, which is able to find the optimal hyperplane h that separates two regions of the input space by maximising the distance between two *support vectors*, one of each class [107]. In particular, due to the complexity of the problem under study, we applied a quadratic kernel to perform the multi-class classification from a non-linear approach. Note that

kernels allow a D-dimensional input space to be projected to another $M>D$, according to $\phi = \mathbb{R}^D \rightarrow \mathbb{R}^M$. Te goal is to map the latent features in a transformation space in which data can be linearly separated, as shown in Figure 2.11.

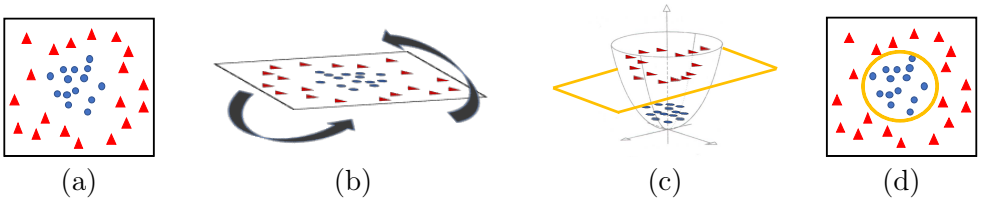
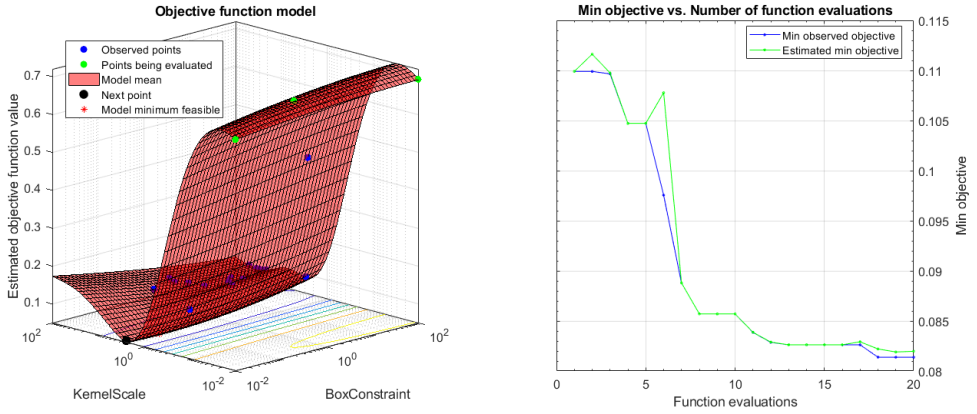


Figure 2.11: Operation of the non-linear SVM classifier. **(a)** Original input space. **(b)** Rotation of the plane of data. **(c)** 3D space transformation where the features are linearly separable. **(d)** Representation of the classification boundary on the 2D plane.

Taking into account that SVMs are non-parametric classifiers that were originally formulated to face binary problems, we divided the initial multi-class problem (artefacts, healthy glands and pathological glands) into two binary sub-problems (artefacts Vs glands and healthy glands Vs pathological glands), through applying a *OneVsOne* (OvO) strategy. Besides the polynomial order and the binary type of classification, SVM classifiers also enable the optimisation of other hyperparameters such as the *box constraint* C and the *kernel scale* γ. C helps to prevent overfitting by managing the maximum penalty imposed on margin-violating observations. The higher value of C the fewer support vectors are assigned, which leads to a smaller overfitting but longer training times. In contrast, γ defines the influence of a single training example. High values of γ decrease the computational cost, but it also implies that the weights of the observations closest to the decision boundary have greater importance than the rest. We specified the same range of values for both $C \in [10^{-2}, 10^2]$ and $\gamma \in [10^{-2}, 10^2]$ hyperparameters, in order to find a trade-off solution between computational cost and robustness of the models. In particular, we made use of a Bayesian optimisation algorithm that attempts to minimise a scalar objective function through evaluating the expected amount of improvements and modifying the behaviour when the classifier estimates that a local area is over-exploited. As shown in Figure 2.12, we found $C = 99.30$ and $\gamma = 28,88$ as the optimal values to train a new classification model composed of all 4-fold training data (see Subsection 2.3.1.3)

A second hand-driven learning classifier based on the Multi-Layer Perceptron (MLP) was implemented. According to [108], MLP is one of the best algorithms



(a) (b)

Figure 2.12: SVM training process. **(a)** 3D objective function that shows how the model find the optimal minimum by modifying the hyperparameters C and γ. **(b)** Diagram that shows how to reach the minimum objective as the number of time increases.

for pattern recognition tasks focused on hand-crafted features. This type of neural network consists of a set of logistic regressions in which the output of each is the input to the next one. MLP classifier is based on a forward-backward propagation algorithm that allows for the updating of the weights and bias during the models' training. We applied a MLP architecture with one hidden layer and fifteen neurons, as shown in Figure 2.13, taking into account that more hidden layers would provide considerable overfitting. The inputs $X = \{X_1, X_2, ..., X_i, ..., X_{136}\}$, correspond to the latent features extracted from each gland candidate G_i. Particularly, we performed the update of the weights ω by means of the stochastic gradient descent (SGD) algorithm with a momentum $\alpha = 0.9$ and an adaptive learning rate back propagation, according to Eq. 2.36.

$$\omega(t+1) = \alpha\omega(t) + \eta\alpha\frac{d}{d\omega(t)}\mathcal{L}(y, \hat{y}) \tag{2.36}$$

where $\omega(t)$ initially takes low random values and $\mathcal{L}(y, \hat{y})$ is the loss function corresponding to the categorical cross-entropy that measures the performance of the classifier according to Eq. 2.37.

$$\mathcal{L}(y, \hat{y}) = -\sum_{i} y_i \; log(\hat{y}_i), \tag{2.37}$$

where y_i corresponds to the output (i.e. the prediction) achieved from the model after each epoch, and $\hat{y}_i$ is the ground truth label corresponding to the target of the training data. Besides, $\eta = 10^{-3}$ is the initial learning rate that varies for each epoch following the next low: if the loss function error decreases up to the goal, the learning rate is increased by a factor $\mu_u = 1.5$, whereas if the performance increases by more than $m = 1.04$, then the learning rate is decreased by a factor $\mu_d = 0.5$.

Unlike in the SVM-based approach, in this case, we did not perform internal cross-validation techniques, as we initially defined all the hyperparameters. However, we included a validation subset for each training set to analyse the behaviour of the model and supervising if overfitting phenomenon appears along the epochs. Additionally, we set the maximum number of epochs $N = 1000$ and impose a stop criterion to end the training process if the performance of the neural network does not improve after 20 epochs.

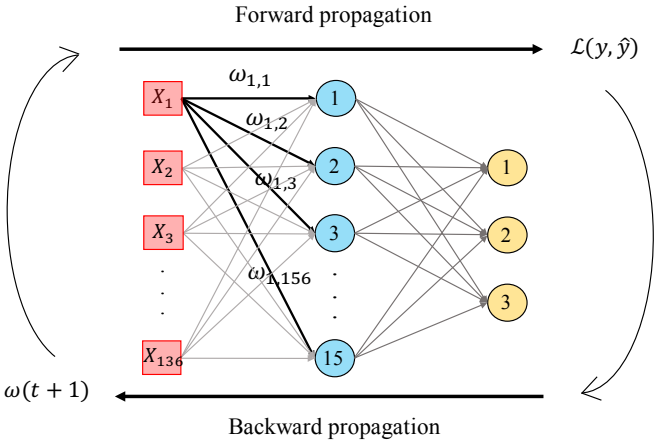


Figure 2.13: Operation of the forward-backward algorithm using a MLP architecture composed of one hidden layer with fifteen neurons and an output layer with three classes.

2.3.2 Deep learning approach

2.3.2.1 Convolutional Neural Network strategy

Data partitioning. In this case, we directly constructed the predictive models from the previous segmented gland candidates images, as the computed CNN are able to automatically extract the key features by means of convolutional operations performed in the base model. Similarly to the hand-driven learning strategy, we accomplished a data partitioning stage based on the medical history of each patient via 5-fold cross-validation. However, instead of carrying out the internal data-partitioning from the extracted features, we did it from the gland candidates images. The training process in this case lied in the optimisation of the convolutional coefficients during the forward-backward propagation process.

Network architecture. To compare the results achieved from the hand-driven learning approach, we made use of a very popular neural network called *Very Deep Convolutional Networks for Large-Scale Image Recognition (VGG19)*, which was proposed by Krizhevsky et al. [28] to face the challenge ILSVRC-2012[1]. It is important to remark that Simonyan and Zisserman [109] modified some hyperparameters of this architecture with the aim of reducing the classification error in the ImageNet dataset. They discovered that the performance of the neural network improved by moving each filter of the convolutional layer along the image by means of a sliding window with a small step size (stride). In addition, the authors achieved better results using smaller receptive fields, i.e. smaller sets of pixels inside the sliding window in each epoch. For these reasons, they designed the convolutional layers using receptive fields of size 3×3 and a stride $s = 1$, instead of 11×11 and $s = 4$, as was originally proposed in [28]. The researchers developed six different CNN architectures with the same basis and modified the depth in each of them. Finally, they proposed the VGG16 and VGG19 networks to face the ImageNet 2014 challenge and reached the first and the second places in the localisation and classification tracks. In this work, we based on the VGG19 architecture also using 3×3 receptive fields and a stride $s = 1$, as shown in Figure 2.14. However, we included some modifications in the classification stage (a.k.a top model) with respect to the original architecture. Specifically, we defined two fully-connected layers, but we applied dropout layers with coefficients of 0.5 and 0.25 after each dense layer. Note that these dropout layers aim to avoid the overfitting by randomly disconnecting the 50% and the 25% of the neurons, respectively. In addition, we defined the second hidden layer with 2048 neurons,

[1] www.image-net.org/challenges/LSVRC

instead of 4096 as the original network, and we modified the output layer using only three neurons to discern between artefacts, benign glands and pathological glands (see Figure 2.14).

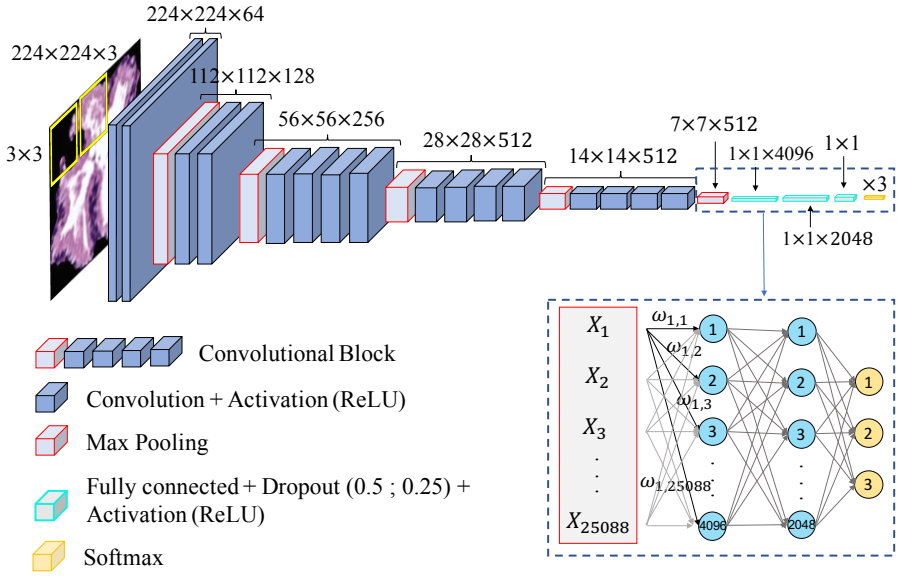


Figure 2.14: Architecture of the fine-tuned VGG19 used to develop deep learning models from gland candidates of histopathological prostate images.

Given the reduced number of gland candidate instances of our database to train from scratch an architecture with a such high depth, we made use of fine-tuning techniques [110] which are able to transfer the knowledge acquired by the VGG19 when it was trained on the *ImageNet* dataset to the problem under study. We used the pre-trained weights ω and applied a *deep fine-tuning* [111] strategy to freeze only the coefficients of the three first convolutional blocks. The filters of the two remaining blocks of the CNN were retrained using the specific knowledge of the gland candidate images. During the training phase, the update of the filter weights was performed in a similar way as in the MLP case by setting a batch size $b = 16$. The categorical cross-entropy (Eq. 2.37) was used as a loss function to calculate the error between the predictions and the ground truth. The SGD optimiser (Eq. 2.38) was applied to update the weights during the backward propagation step. We used a momentum of 0.98 to improve the convergence rate of the CNN with a learning rate $\eta = 10^{-5}$.

We trained the deep learning models during a maximum of 100 epochs with an early stopping of 15 iterations.

$$\omega(t+1) = \omega(t) + V(t+1), \tag{2.38}$$

where,

$$V(t+1) = \gamma V(t) - \eta \frac{d}{d\omega} \mathcal{L}(y, \hat{y}), \tag{2.39}$$

We applied *data augmentation* techniques [112] to increase the number of specific images. This technique allows for the creation of artificial samples similar to the original ones by performing different transformations. We defined the aggressive factor ratio as $t = 0.02$ to create the synthetic samples with the same label as the samples from which they were generated.

2.4 Results

In this section, we present a comparison of the results achieved from both hand-driven and deep learning approaches. We evaluate the 5-trained models over each 5-external cross-validation datasets. Table 2.2 reports the average results using different figures of merit such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F-score, area under the ROC curve (AUC) and accuracy. In addition, we also report the ROC curves (see Figure 2.15) aimed at showing the discriminatory ability of the proposed classifiers according to each problem. Note that, although the proposed approach correspond to a multi-class framework, we expose the results distinguishing, on the one hand, between artefacts and glands and, on the other hand, between benign and Gleason grade 3 glands.

Noticeably, this is one of the few studies that provides results per gland candidate, instead of per patch. For this reason, we expose in Table 2.3 the results achieved by our best predictive model to compare them with those reached by other authors who also dealt with this type of classification. It should be noted that we conduct an indirect comparison due to the private character of the code and databases used in [88] and [89].

To elucidate the statistical differences between the performance of the three classification methods, we carried out a statistical analysis similar to the

Table 2.2: Classification results obtained per gland candidate.

	Artefacts Vs Glands			*Benign Vs Pathological*		
	SVM	**MLP**	**VGG19**	**SVM**	**MLP**	**VGG19**
Sensitivity	**0.945** $\pm$ 0.016	0.930 $\pm$ 0.011	0.901 $\pm$ 0.020	**0.802** $\pm$ 0.058	0.802 $\pm$ 0.079	0.747 $\pm$ 0.109
Specificity	**0.952** $\pm$ 0.026	0.952 $\pm$ 0.028	0.939 $\pm$ 0.026	**0.873** $\pm$ 0.073	0.819 $\pm$ 0.094	0.780 $\pm$ 0.123
PPV	0.911 $\pm$ 0.058	**0.913** $\pm$ 0.053	0.884 $\pm$ 0.059	**0.845** $\pm$ 0.069	0.796 $\pm$ 0.084	0.779 $\pm$ 0.101
NPV	**0.964** $\pm$ 0.018	0.956 $\pm$ 0.019	0.945 $\pm$ 0.016	**0.843** $\pm$ 0.052	0.837 $\pm$ 0.059	0.769 $\pm$ 0.047
F-Score	**0.927** $\pm$ 0.029	0.921 $\pm$ 0.030	0.891 $\pm$ 0.032	**0.820** $\pm$ 0.030	0.793 $\pm$ 0.014	0.753 $\pm$ 0.059
AUC	0.984 $\pm$ 0.011	**0.987** $\pm$ 0.007	0.974 $\pm$ 0.010	**0.922** $\pm$ 0.045	0.912 $\pm$ 0.042	0.889 $\pm$ 0.036
Accuracy	**0.946** $\pm$ 0.017	0.943 $\pm$ 0.030	0.925 $\pm$ 0.018	**0.883** $\pm$ 0.026	0.853 $\pm$ 0.020	0.817 $\pm$ 0.031

Table 2.3: State-of-the-art comparison in terms of accuracy per gland.

	Xia et al. [89]	**Nguyen et al. [88]**	**Proposed Model**
Artefacts Vs. Glands	-	0.93 $\pm$ 0.04	**0.946** $\pm$ 0.017
Benign Vs. Pathological	0.86 $\pm$ 0.02	0.79 $\pm$ 0.08	**0.883** $\pm$ 0.026
Multi-class classification	-	0.77 $\pm$ 0.07	**0.876** $\pm$ 0.026

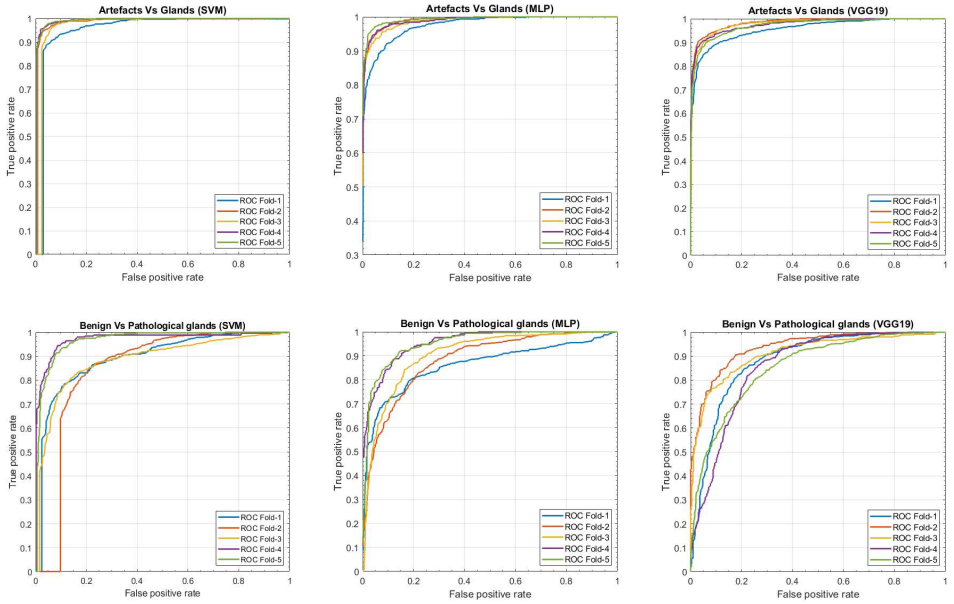


Figure 2.15: ROC curves from hand-driven and deep learning strategies after evaluating each binary sub-problem.

developed in Subsection 2.3.1.2. In this case, we calculated for each test dataset not only the prediction labels, but also the scores, i.e. the probabilities of each gland candidate of belonging to a specific class. The aim was to determine the discriminatory ability of the different classification models taking into account the scores and the ground-truth labels. Particularly, for each classification model (SVM, MLP and VGG19), we used the scores as a variable to study its independence level with respect to the targets. First, we performed a *Kolmogorov-Smirnov test* to determine if the variable under study followed a normal distribution N(0,1) or not. Depending on it, we made use of the *ANOVA* or *Kruskal-Wallis* test, comparing the mean or median of such variable with respect to the categorical classes, respectively. These tests reported a *p-value* for each possible class to measure the performance of each classifier when predicting the label of such class. We repeated this process 5 times (one per partition) and reported the averaged p-values to show the differences between the classifiers (see Table 2.4).

Table 2.4: Average of *p-values* achieved after calculating the independence level between the probability of each class and the targets, from each classifier.

	Artefact	**Benign gland**	**Pathological gland**
SVM	$4.2813 \cdot 10^{-223}$	$3.3720 \cdot 10^{-118}$	$2.1903 \cdot 10^{-210}$
MLP	$6.3328 \cdot 10^{-203}$	$2.8052 \cdot 10^{-143}$	$7.3978 \cdot 10^{-195}$
VGG19	$3.9919 \cdot 10^{-220}$	$1.2796 \cdot 10^{-137}$	$5.8481 \cdot 10^{-192}$

Prediction phase. At the test time, we made use of histological samples of new patients from which we only know if they are healthy or suffer for prostate cancer in an initial stage, but we do not have a ground truth per gland. Particularly, we selected the best classification models to perform a committee of evaluation responsible of predicting the label of the new samples. The final objective was to help the pathologists by means of an automatic system able to identify benign and pathological glands from histological prostate images. In Figure 2.16, we expose the prediction carried out for some 1024×1024 representative samples from different patients, in which we remark the automatically segmented glands according to the prediction carried out by the trained models. The boundary of the glands predicted as healthy are highlighted in green, whereas the contours of the glands predicted as pathological are highlighted in red. It is important to note that an expert pathologist annotated the samples reported in Figure 2.16 as: fully benign pattern (Figure 2.16 (a-c)), fully pathological pattern (Figure 2.16 (d-f)) and mixed pattern (Figure 2.16 (g-i)).

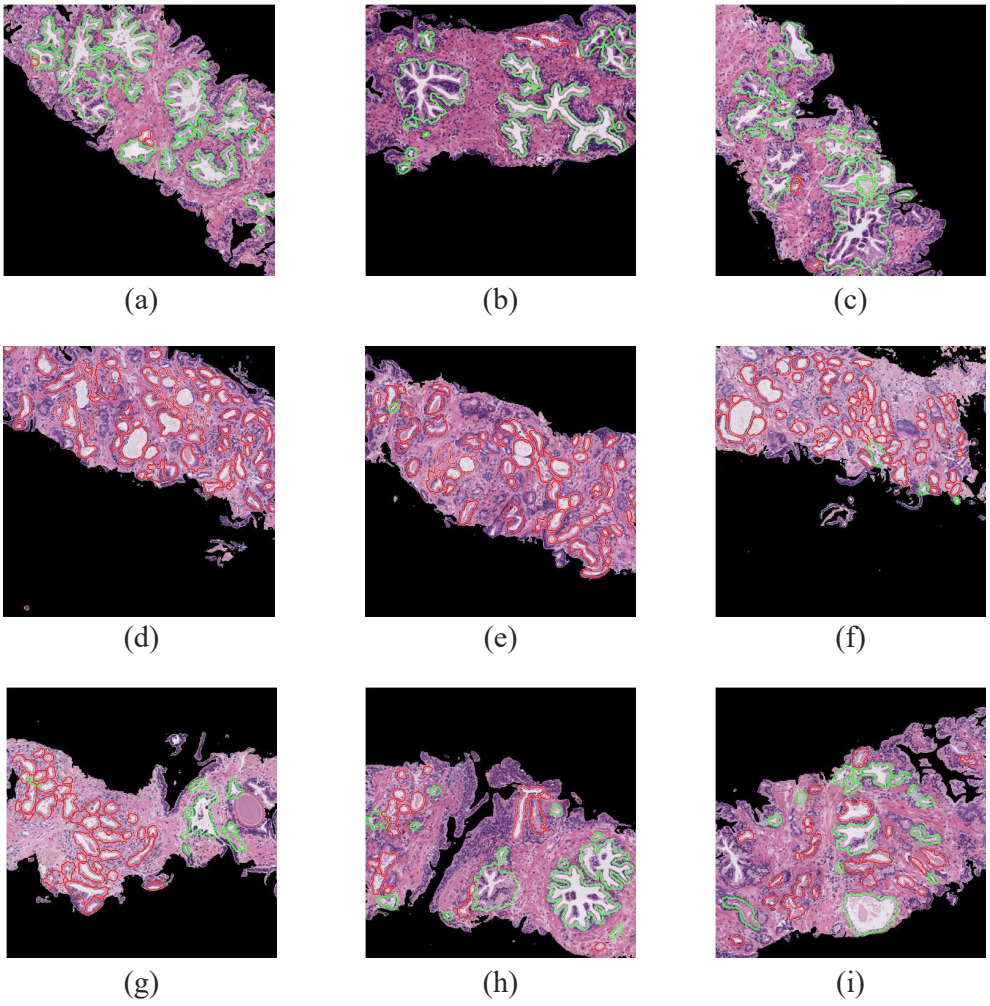


Figure 2.16: Qualitative prediction results about the gland classification showing in green and red the structures predicted as benign and pathological glands, respectively.

Computational cost. In Table 2.5, we list the average and standard deviation corresponding to the temporal intervals spent in each of the four processes that compose the hand-driven algorithm (clustering, segmentation, feature extraction and prediction) when analysing each patch.

Table 2.5: Computational cost of the proposed hand-driven learning algorithm. The results refer to the time (in seconds) required to classify the gland candidates of each 1024×1024 as artefact, benign or pathological glands.

	Healthy tissues (s)	**Cancerous tissues(s)**
Clustering stage	2.79 ± 0.58	2.39 ± 0.39
Gland segmentation	25.24 ± 29.63	9.69 ± 18.99
Feature extraction	9.75 ± 7.75	5.88 ± 7.48
Prediction	0.04 ± 0.01	0.04 ± 0.01
End-to-end algorithm	38.68 ± 34.44	18.87 ± 26.12

Table 2.5 shows that the proposed algorithm needs 38.68 ± 34.44 seconds to make the prediction of a benign 1024×1024 patch. Contrarily, the averaged computational cost to automatically examine a pathological patch is 18.87 ± 26.12 seconds. Note that the time analysis was performed on an Intel i7 @4.00 GHz of 16 GB of RAM with a Titan V GPU. The classification strategies related to the SVM and MLP models were executed in MATLAB 2018b, whereas the methodology based on the VGG19 architecture was implemented in Python 3.5, using Keras framework with Tensorflow as backend.

2.5 Discussion

From Table 2.2 and Table 2.4, we observe that the hand-driven learning approach, using a SVM classifier with a quadratic kernel, seems to provide a very slight superiority in the binary classification of both scenarios: artefacts Vs. glands and benign Vs. pathological glands. As depicted in Table 2.2, the results achieved from the VGG19 architecture decline a bit compared to the SVM and MLP classifiers, which may be due to the exhaustive feature extraction and selection methods carried out during the hand-driven learning.

Notably, the values obtained for all figures of merit in Table 2.2 are similar in the first binary classification (i.e. artefacts Vs. glands), but the differences are more prominent when discriminating between normal and cancerous samples. This is understandable taking into account that artefacts and glands present very different contextual circumstances, as artefacts are not usually surrounded by cytoplasm and nuclei components, unlike the lumen of the glands. Nevertheless, artefacts could be confused with lumen structures when an accumulation of glands envelop specific broken areas of tissue, which have the same colour as the lumen objects. These broken areas result in artefacts (false glands) because the model associates the nuclei of the surrounding glands

as the nuclei of such broken areas. In this case, the features extracted from artefacts and glands are more similar to each other, which can lead to an error in the prediction. It should be noted that artefact elements usually are involved by pixels corresponding to stroma components, as shown in Figure 2.9, what is decisive to the correct classification of the false glands. The visible differences between artefacts and glands enables a successful classification from both conventional and deep learning approaches. We can observe in Figure 2.15 (bottom row) and in Table 2.2 that the performance of the three learning strategies is closely similar attending to the ROC curves. In contrast, the distinction between benign and pathological glands is more complex, as both correspond to moderately differentiated carcinomas. In Figure 2.15, we can see how the performance of the classifiers fall due to the similarities of the benign and pathological glands.

It can be striking that the SVM-based conventional algorithm reports better results than the proposed deep learning approach. However, this makes sense taking into account that CNNs have specific limits and drawbacks related to the orientation and spatial relationships between objects. CNNs do not consider important spatial hierarchies between simple and complex objects. For this reason, the deep learning approach successfully discriminates between artefacts and glands, as it finds differences in the composition of the images. Nevertheless, its performance considerably decreases when facing the benign Vs. pathological scenario, as it can not manage the differences concerning to the spatial dimensions, orientations and sizes of the glands. In the problem under study, the size of the lumens and glands is really decisive to determine the class of a specific gland, as explained in Section 2.1. Thereby, since the hand-driven learning approach consider the dimensions of the gland candidates in the feature space, it is coherent that the SVM classifier provides better results than the proposed deep learning network.

The state-of-the-art comparison (Table 2.3) was carried out in an indirect way, as there are no public databases of gland units to contrast our results. Notwithstanding, the approach presented in this work outperforms the methods proposed by other state-of-the-art studies that also follow an approach at the gland unit level. In relation to the artefacts Vs. glands problem, Xia et al. [89] did not report results because they did not implement an approach based on an initial lumen identification, but they classified all segmented glands into two classes: benign and pathological. In contrast, Nguyen et al. [88] reported an accuracy of 0.93 in the discrimination between artefacts and glands. We improve their results surpassing the accuracy by 1.6%. However, the real challenge lies in discerning between healthy and

Gleason grade 3 glands. We surpass by 2.3% and 9.3% the accuracy with respect to [89] and [88], respectively, and improve the multi-class accuracy reached in [88] in a 10.6%.

Regarding the simulation of the clinical practice, we show in Figure 2.16 the ability of the proposed computer-aid system to identify and discriminate the healthy and pathological gland units, which are highlighted in green and red, respectively. Attending to Figure 2.16, we can determine that the proposed model performs successfully in detecting specific areas susceptible to cancer.

Concerning the computational cost, the automatic analysis of pathological patches requires half the time compared to the healthy images. This is mainly propitiated by the segmentation stage, which takes more time to analyse benign glands due to its larger size and fusiform appearance. Attending to the temporal intervals (Table 2.5), we can observe high standard deviation values, especially in the segmentation and feature extraction stages. This occurs because the computational cost is closely related to the number of glands in each patch, i.e. the more number of glands per image, the more time is required.

2.6 Conclusion

This work proposes two novel approaches to automatically identify the first stage of prostate cancer from images of gland candidates previously segmented. In the first approach, a combination of four kinds of descriptors based on morphology, texture, fractals and contextual information has been computed for an optimal hand-crafted feature extraction. Regarding the second approach, a CNN built upon VGG19 architecture has been used to automatically extract and classify the relevant features from artefacts, benign and pathological glands. The hand-driven learning approach, making use of the SVM with a quadratic kernel, provides the best classification performance and also outperforms the most relevant methods proposed in the state of the art. Promising results have been achieved in the prediction of samples from new patients, however, it would be necessary to carry out additional tests, as well as retraining the predictive models, using larger datasets of prostate WSIs. In future research lines, we propose to improve the deep learning results by training other popular state-of-the-art CNNs (e.g. ResNet, Xception, etc.). We also propose to go further into the framework of prostate cancer grading by considering all the patterns of the Gleason scale.

Chapter 3

A Novel Self-Learning Framework for Bladder Cancer Grading Using Histopathological Images

The content of this chapter corresponds to the author version of the following published paper: García, G., Esteve, A., Colomer, A., Ramos, D. & Naranjo, V. A novel self-learning framework for bladder cancer grading using histopathological images. Computers in Biology and Medicine, 104932 (2021).

Contents

A Novel Self-Learning Framework for Bladder Cancer Grading Using Histopathological Images

Gabriel García[1], Anna Esteve[1,2], Adrián Colomer[1], David Ramos[2] and Valery Naranjo[1]

[1]Instituto de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, 46022, Valencia, Spain

[2] Hospital Universitario y Politécnico La Fe, Avinguda de Fernando Abril Martorell, 106, 46026, Valencia, Spain.

Abstract

In recent times, bladder cancer has increased significantly in terms of incidence and mortality. Currently, two subtypes are known based on tumour growth: non-muscle invasive (NMIBC) and muscle-invasive bladder cancer (MIBC). In this work, we focus on the MIBC subtype because it has the worst prognosis and can spread to adjacent organs. We present a self-learning framework to grade bladder cancer from histological images stained by immunohistochemical techniques. Specifically, we propose a novel Deep Convolutional Embedded Attention Clustering (DCEAC) which allows for the classification of histological patches into different levels of disease severity, according to established patterns in the literature. The proposed DCEAC model follows a fully unsupervised two-step learning methodology to discern between non-tumour, mild and infiltrative patterns from high-resolution 512×512 pixel samples. Our system outperforms previous clustering-based methods by including a convolutional attention module, which enables the refinement of the features of the latent space prior to the classification stage. The proposed network surpasses state-of-the-art approaches by $2-3\%$ across different metrics, reaching an average accuracy of 0.9034 in a multi-class scenario. The reported heat maps evidence that our model is able to learn by itself the same patterns that clinicians consider relevant, without requiring previous annotation steps. This represents a breakthrough in MIBC grading that bridges the gap with respect to training the model on labelled data.

3.1 Introduction

Bladder cancer arises from uncontrolled proliferation of the urothelial bladder cells, which leads to tumour development. A significant increase in adult incidence and mortality has been observed during the last several years. Recent studies state that bladder cancer is the second most common urinary tract cancer and the fifth most prevalent among men in developed countries [71, 72].

Nowadays, the diagnostic procedure for bladder cancer involves several time-consuming tests. First, urine cytology is performed to determine the presence of cancer cells [113]. Subsequently, vesico-prostatic and renal ultrasound are employed to locate the tumour and assess the type of growth, which can be used to determine the grade and prognosis of the patient. If the tumour cannot be located at the previous stage, MRI urography is carried out to analyse possible local spread [114]. If there is evidence of bladder cancer, the urologist usually performs a cystoscopy based on the transurethral resection technique [115], which allows for the extraction of a sample of abnormal bladder tissue to determine the type of tumour growth. After the preparation process, the biopsied tissue is usually stained with hematoxylin and eosin (H&E) to enhance its histological properties. Finally, an additional staining process can be adopted to highlight special structures associated with the problem under study. The immunohistochemical CK AE1/3 technique was applied on the histological images used in this work to highlight the cancer cells by providing a brown hue when the antigen-antibody binding occurs.

The two kinds of bladder cancer, non-muscle invasive (NMIBC) and muscle-invasive (MIBC), are distinguished depending on the level of invasion of tumour growth within the bladder wall. Currently, 75% and 25% of bladder cancer cases correspond to NMIBC and MIBC, respectively [72]. In this study, we focus on the MIBC category as it has the worst prognosis and favours tumour dissemination to adjacent organs. According to [117], MIBC does not usually present low-grade malignancy, but rather high-grade urothelial carcinomas. Following the classification criteria proposed by the World Health Organisation (WHO) [118], these can be classified as grade 2 or 3. Jimenez et al. [116] described three different histological patterns which correlate with the patient outcome. Specifically, histopathological images stained with CK AE1/3 were annotated by a pathologist with more than 20 years of expertise considering nodular, trabecular and infiltrative patterns, as shown in Figure 3.1. The nodular pattern (yellow box) is defined by the presence of well-delineated, circular nests of tumour cells. The trabecular pattern is characterised by tumour cells arranged in interconnected bands. The infiltrative pattern, also

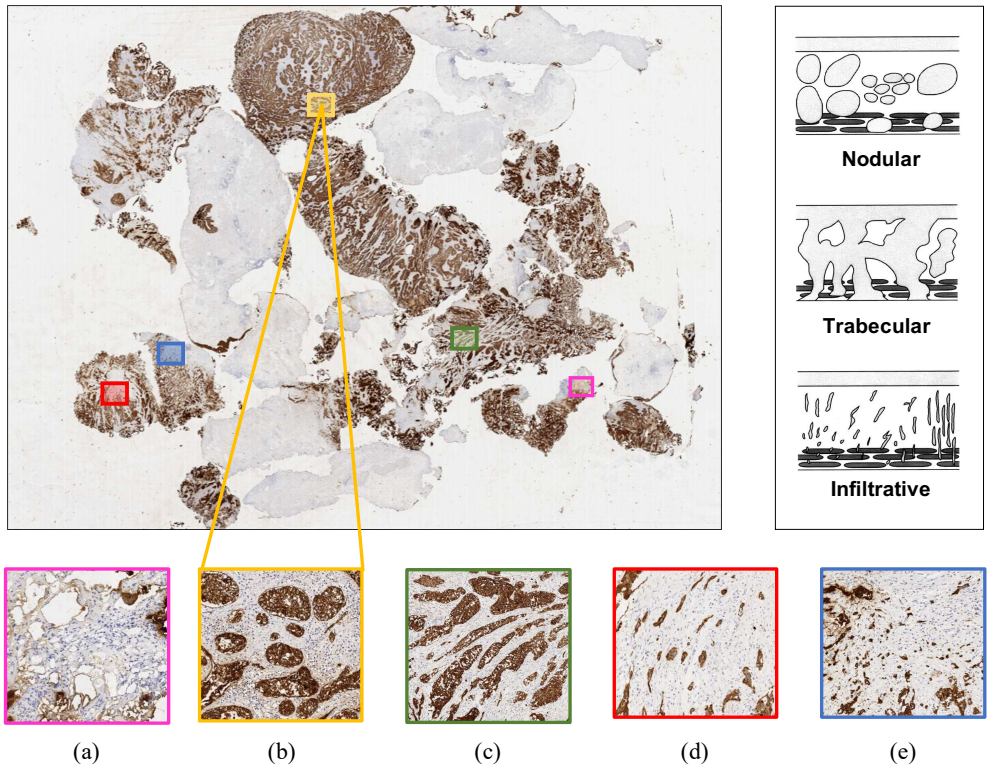


Figure 3.1: Histological bladder patterns. The larger image corresponds to a WSI of a patient suffering from muscle-invasive bladder cancer. The top right of the figure (modified from [116]) is a diagrammatic representation of the theoretical arrangement of the patterns. The patches marked with colours denote different growth patterns. (a) non-tumour pattern, (b) nodular arrangement (mild pattern), (c) trabecular arrangement (mild pattern), (d) tumour cell cords (infiltrative pattern) and (e) isolated tumour cells (infiltrative pattern).

known as tumour budding, is composed of cords of tumour cells (red box) or a small cluster of isolated cells called *buds* (blue box). The infiltrative pattern represents the most aggressive scenario and the worst prognosis for the patient [119–122]. Therefore, we combined nodular and trabecular structures into a single specific class (mild pattern) to grade the severity of MIBC according to the prognosis of the disease. We also considered a non-tumour pattern (pink box) to cover cases where the patient shows no signs of tumour. Thus, a multi-class scenario is conducted in this paper to classify bladder cancer into non-tumour (NT), mild (M) and infiltrative (I) patterns.

3.1.1 Related work

Accurate diagnosis of bladder cancer is a time-consuming task for expert pathologists and lacks reproducibility, leading to significant differences in histological interpretation [123, 124]. Many state-of-the-art studies have proposed artificial intelligence algorithms to assist pathologists in terms of cost-effectiveness and subjectivity ratio. Most of these approaches focused on machine learning techniques applied to H&E-stained histological images for segmentation [125–127] and classification [123, 124, 128–133].

Beginning with the segmentation-based studies, Lucas et al. [125] used the popular U-Net architecture to segment normal and malignant cases of bladder images. They then used the common VGG16 network [28] as a backbone to extract histological features from patches of 224×224 pixels. The resulting features were combined with other clinical data to carry out a classification step using bidirectional GRU networks [134]. The proposed algorithm reported an accuracy of 0.67 for 5-year survival prediction. In [126], the authors carried out an end-to-end approach to discern between MIBC and NMIBC categories from H&E images. First, they performed a segmentation process to distinguish tissue from image background. Patches of 700×700 pixels were used to perform both manual and automatic feature extraction. The hand-crafted learning was conducted via contextual features such as nuclear size distribution, crack edge, sample ratio, etc., whereas the data-driven learning was conducted using the VGG16 and VGG19 architectures. During the classification stage, different machine learning classifiers such as support vector machine (SVM), logistic regression (LR) and random forest (RF), among others, were used to determine the bladder tissue type. Manual approaches showed superior performance over deep learning models, reaching an accuracy of $91 - 96\%$.

Most of the classification-intended studies also focused on H&E-stained histological images, as in the segmentation frameworks. In [128], researchers proposed a multi-class scenario to detect the molecular subtype in MIBC cases. They applied the ResNet architecture to patches of 512×512 pixels, achieving results for the area under the ROC curve (AUC) of 0.89 and 0.87 in terms of micro- and macro-average, respectively. In [129], the authors made use of the Xception network as a feature extractor from H&E-stained patches of 256×256 pixels. An SVM classifier was then implemented to discern between high and low mutational burden, reaching values of 0.73 and 0.75 for accuracy and AUC, respectively. Harmon et al. [130] proposed a classification scenario to detect lymph node metastases from H&E patches of 100×100 pixels. A combination of the ResNet-101 architecture with AdaBoost classifiers reported an AUC of 0.678 at test time. Another study [123] carried out

a classification approach to categorise tissue type into six different classes: urothelium, stroma, damaged, muscle, blood and background. To this end, the authors combined supervised and unsupervised deep learning techniques on patches of 128×128 pixels stained with H&E. Specifically, they trained an autoencoder (AE) from the unlabelled images and used the encoder network to address the classification through features extracted from the labelled samples. They achieved multi-class scores of $0.936, 0.935$ and 0.934% for precision, recall and F1-score metrics, respectively. One of the most prominent state-of-the-art studies (focusing on H&E-stained histological images of bladder cancer) was conducted in [124]. In this study, Zhang et al. compiled a large database of Whole-Slide Images (WSIs) with the aim of discerning between low and high grades of disease. They used an autoencoder network to identify possible areas with cancer. They then fed 1024×1024 regions of interest (ROIs) into a Convolutional Neural Network (CNN) for classification into low and high classes. The proposed system obtained an average accuracy of 94%, compared to 84.3% achieved by pathologists. The findings from this study reveal that there exists a significant subjectivity among experts in diagnosing from histological images of bladder cancer, as discussed in [123].

In addition to histopathological samples, other imaging modalities are also considered in the literature for bladder cancer analysis, e.g. magnetic resonance imaging (MRI) [127], cystoscopy [132, 133] or computed tomography (CT) [131]. In particular, Dolz et al. [127] applied deep learning algorithms to detect bladder walls and tumour regions from MRI samples. In [132, 133], different deep learning architectures were implemented to distinguish between healthy and bladder cancer patients using cystoscopy samples. Yang et al. [131] outlined a classification between NMIBC and MIBC categories from CT images. Although immunohistochemical techniques are widely used in the literature to detect tumour budding, most state-of-the-art works applied them on colorectal cancer imaging [135–138]. However, there are relatively few immunohistochemistry-based studies for the diagnosis of bladder cancer in the literature. As far as we are aware, only the study conducted in [139] proposed the use of sample stained via immunofluorescence techniques to quantify tumour budding for the prognosis of MIBC via machine learning algorithms. Specifically, the authors aimed to establish a relationship between tumour budding and assessed survival in patients with MIBC. To do this, they carried out learning methods based on detecting nuclei and segmenting the tumour into stroma regions to count the isolated tumour budding cells. The authors proposed a survival decision function based on random forest, reporting a hazard ratio of 5.44.

3.1.2 Contribution of this work

To the best of our knowledge, no previous works have been conducted to analyse the severity of bladder cancer using histological images stained with cytokeratin AE1/AE3 immunohistochemistry. Moreover, all the state-of-the-art studies focused on supervised learning methods to find dependencies between the inputs and the predicted class [123, 124, 129, 130]. Some of them [123, 124] also considered using unsupervised techniques in the first methodological steps to find possible ROIs with cancer, but required labelled data to build the definitive predictive models. In addition, pattern recognition tasks aimed at grading bladder cancer have not been previously addressed.

To fill these gaps in the literature, we present in this paper a self-learning framework for bladder cancer growth patterns, which focuses on fully unsupervised learning strategies applied on CK AE3/1-stained WSIs. We propose a Deep Convolutional Embedded Attention Clustering (DCEAC) that boosts the performance of the classification model without incurring the cost of labelled data. In the literature, deep-clustering algorithms have demonstrated a high rate of performance for image classification [140–142], image segmentation [143], speech separation [144, 145] and data analysis [146], among other tasks. Inspired by [140], we propose a tailored algorithm capable of competing with the state-of-the-art results achieved by supervised approaches. As a novelty, we include a convolutional attention module to refine the features embedded in the latent space. Additionally, we are the first to focus on the arrangement of histological structures contained in the high-resolution patches to classify them into non-tumour (NT), mild (M) and infiltrative (I) patterns, according to the criteria proposed in [116]. We also computed a class activation map (CAM) algorithm [147] to evidence that the proposed network focuses on specific structures that match with the clinical patterns associated with bladder cancer aggressiveness.

The proposed end-to-end framework provides a reliable benchmark for making diagnostic suggestions without involving a pathologist, which adds significant value to the body of knowledge. In summary, the main contributions of this work are listed below:

- For the first time, we make use of CK AE3/1-stained images to enable the automatic diagnosis of bladder cancer using machine learning algorithms.
- We base on advanced unsupervised deep learning techniques to address bladder cancer classification without the need for prior annotation steps.

- We propose a new deep-clustering architecture capable of improving the representation space via convolutional attention modules, resulting in better unsupervised classification.
- We focus on high-resolution histological patches to learn specific-bladder cancer patterns and stratify different levels of disease severity according to the literature.
- We include heat maps highlighting decisive areas to incorporate an explainable component for network prediction. This provides an interpretability perspective that matches the clinicians' criteria.

3.2 Material

This study made use of a private database of 136 WSIs (one per patient) from the Hospital Universitario y Politécnico La Fe (Valencia, Spain). The WSIs were stained by immunohistochemistry, and were digitised using an intelligent scanner (LEICA BIOSYSTEMS – Aperio CS2) providing optical magnifications of 20× (0.5 μm/pixel) and 40× (0.25 μm/pixel) with a fast network interface of 1GB/second. Specifically, the 40× resolution was selected to take advantage of the inherent structure of the bladder patterns associated with each grade of disease, as high image resolution is necessary in order to achieve an accurate diagnosis of bladder cancer. This is because the class dependencies are only evident in the high frequency of the image, especially the details of the tumour budding.

In the first step of the database preparation (Figure 3.2), an expert from the Pathological Anatomy Department performed a manual segmentation to indicate possible areas of interest. At this point, it is important to highlight that the segmentation was carried out in a very rough manner, as observed in the green areas of Figure 3.2, in order to reduce the expert's annotation time as much as possible. The software used to perform the rough annotations was GIGAVISION: a system for labelling tumour regions in gigapixel histological images [148]. A patching algorithm was then applied to extract cropped images with an optimal block size in terms of computational efficiency and structural content. Specifically, patches of dimensions 512×512 pixels were extracted, according to some of the most recent studies focusing on histopathological images [4, 34, 128]. Next, useless regions (WSI background) were discarded by selecting only those patches that contained more than 75% annotated tissue. After this, a total of 2995 representative patches composed the unsupervised framework. For validation purposes, an expert pathologist with more than 20

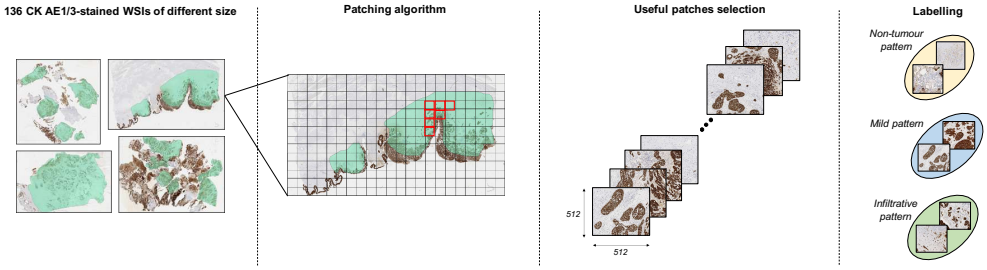


Figure 3.2: Database preparation process. The left-hand side corresponds to the rough segmentation (in green) carried out by the pathologist. Note that each WSI has a different size. Next, a patching algorithm was applied on the 136 WSIs stained with CK AE1/3 to extract sub-images of 512×512 pixels. The red rectangles correspond to some examples of useful patches. Finally, a labelling step was conducted for validation purposes. The resulting 2995 sub-images were classified by the expert as non-tumour (NT), mild (M) or infiltrative (I) pattern to give rise to a multi-class scenario for bladder cancer grading.

years of experience manually labelled each patch as non-tumour (NT), mild (M) or infiltrative (I) classes, according to the pattern criteria previously detailed in Section 3.1. The labelling process resulted in a dataset of 763 NT, 1470 M and 762 I cases, as reported in Figure 3.2. It is essential to remark that we did not have access to the labelled data during the training phase, as we propose a fully unsupervised strategy to achieve self-learning of the patterns. The labels were only considered at test time to evaluate the models' performance.

Concerning the software and hardware aspects, all models were developed using TensorFlow 2.3.1 on Python 3.6. The experiments were performed on an Intel(R) Core(TM) i7-9700 CPU @3.00GHz machine with 16 GB of RAM. For deep learning algorithms, a single NVIDIA DGX-A100 Tensor Core with cuDNN 7.6.5 and CUDA Toolkit 10.1 was used.

3.3 Methods

Recently, deep-clustering algorithms have risen to the forefront of image-based unsupervised techniques, as they are able to enhance feature learning while improving the clustering performance in a unified framework [140]. In this work, we address a fully unsupervised self-learning strategy to cluster a large collection of unlabelled images into $k = 3$ groups corresponding to three different severity levels of MIBC. Inspired by [140], we propose a novel Deep Convolutional Embedded Attention Clustering ($DCEAC$) in which the feature

space is updated in an end-to-end manner to learn stable representations for the clustering stage. Unlike conventional approaches [142], the proposed *DCEAC* algorithm optimises the latent space by preserving the local data structure, which helps stabilise the clustering-learning process without distorting the embedding properties [140] (see Subsection 3.3.2).

Self-learning methods aim at learning useful representations by leveraging the domain-specific knowledge from unlabelled data to accomplish downstream tasks. This training procedure is usually tackled by solving pretext tasks [149], relational reasoning [150] or contrastive learning [151] approaches. In our bladder cancer scenario, we advocate for a sequential strategy that uses image reconstruction as an unsupervised prior task. Specifically, we carry out a two-step learning methodology. First, a convolutional autoencoder (CAE) is trained to incorporate information about the properties of the histological domain (Subsection 3.3.1). Second (Subsection 3.3.2), a clustering branch is included at the output of the CAE bottleneck to provide the class information from the embedded features, which are updated by re-training the CAE on a combined network. In the following sections, we detail both learning steps.

3.3.1 CAE pre-training

Autoencoder (AE) is one of the most common techniques for data representation and aims to minimise the reconstruction error between X inputs and R outputs. AEs consist of two training stages: the encoder $f_\phi(\cdot)$ and the decoder $g_\theta(\cdot)$, where ϕ and θ are learnable parameters. The encoder network applies a non-linear mapping function to extract a feature space Z from the input samples X, such that $f : X \to Z$. The decoder structure is intended to reconstruct the input data from the embedded representations; $R = g_\theta(Z)$. The learning procedure is carried out by minimising a reconstruction loss function.

AE architectures are typically defined by fully connected layers aimed at reducing the dimensionality of the feature space [142, 146], or by convolutional layers acting to extract features from 2D or 3D input data [140]. Like [140], we adopted a CAE architecture to address the reconstruction of the histological patches as a pretext task. However, our CAE differs from the current literature in a specific aspect of the network: the bottleneck. Unlike Guo et al. [140], who combined flatten operations with fully connected layers at the central part of the CAE, we introduced a convolutional attention module through a residual connection to improve the latent space for the subsequent clustering task. As seen in Figure 3.3, the proposed CAE consists of three main structures: encoder, bottleneck and decoder. The encoder is composed of three stacked

convolutional layers with a 3×3 receptive field (blue boxes). At the bottleneck, we defined an attention block that allows the embedded features to be refined in the spatial dimension. Specifically, the proposed module combines 1×1 convolutions (green boxes) with a sigmoid function (purple layer) intended to re-calibrate the inputs. The inclusion of an identity shortcut forces the network to stabilise the feature space by propagating larger gradients to previous layers via skip connections. An additional 1×1 convolutional layer was included at the end of the bottleneck to extract the latent space (z_i) without affecting the dimensions of the feature maps. In the decoding stage, we applied regularisation operations between the transposed convolutional layers (yellow-contour boxes) throughout Batch Normalisation (BN) to avoid the internal covariate shift [152]. Notably, no pooling or up-sampling layers were used to adapt the dimensions of the feature maps after each convolutional step. Instead, we worked with a *stride* > 1 to provide a more transformable network by learning spatial sub-sampling [140].

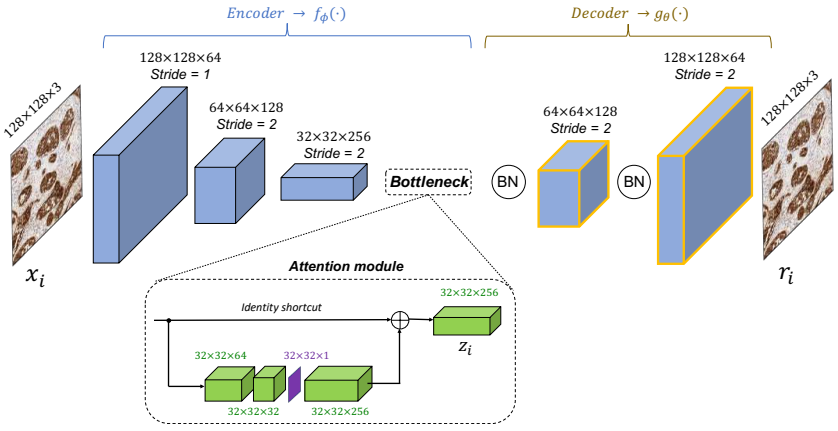


Figure 3.3: Architecture of the proposed CAE used for image reconstruction as a pretext task during the learning process.

As observed in Figure 3.3, given an input set of patches $X = \{x_1, ..., x_i, ..., x_N\}$, with N samples per batch, the encoder network maps each input $x_i \in \mathbb{R}^{M \times M \times 3}$ into an embedded feature space $z_i = f_\phi(x_i)$ resulting from the attention module. At the end of the autoencoder network, the decoder function was trained to provide a reconstruction map $r_i = g_\theta(z_i)$ trying to minimise the mean squared error (MSE) between the input x_i and the output r_i, according to Eq. 3.1. Note that the histological patches were resized from $M_0 = 512$ to $M = 128$ to alleviate GPU constraints during model training.

$$L_r = \frac{1}{N}\sum_{i=1}^{N} ||x_i - g_\theta(f_\phi(x_i))||^2 \tag{3.1}$$

Learning details for the CAE pre-training. Given a training set $\mathcal{X} = \{X_1, ..., X_b, ..., X_B\}$ composed of 2995 histological patches, the proposed CAE was trained during $\epsilon = 200$ epochs by applying a learning rate of 0.5 on $B = 94$ batches, with $X_b \subset \mathcal{X}$ being a single batch composed of $N = 32$ samples. The Adadelta optimiser [153] was used to update the reconstruction weights by minimising the MSE loss function L_r after each epoch e, as detailed in Algorithm 1. Loading the 2995 histological samples in memory takes 156.59 seconds at the beginning of the model's training. Then, each epoch takes 6.87 seconds to train the $B = 94$ batches.

Algorithm 1 CAE training

Data: Unlabelled training dataset $\mathcal{X} = \{X_1, ..., X_b, ..., X_B\}$
Results: Trained CAE parameters ϕ and θ.

$\phi, \theta \leftarrow$ random;
for $e \leftarrow 1$ **to** ϵ **do**
  for $b \leftarrow 1$ **to** B **do**
    $X \leftarrow X_b \subset \mathcal{X}$;
    for $i \leftarrow 1$ **to** N **do**
      $r_i \leftarrow g_\theta(f_\phi(x_i))$;
  $\mathcal{L}_r \leftarrow \frac{1}{N}\sum_{i=1}^{N} ||x_i - r_i||^2$;
  Update ϕ, θ using $\nabla_{\phi,\theta}\mathcal{L}_r$;

3.3.2 DCEAC training

In the pioneer deep-clustering work [142], the authors proposed a Deep Embedded Clustering (DEC) algorithm in which the decoder structure was discarded during the second stage of clustering training. However, Guo et al. [140] demonstrated that fine-tuning only the encoder network could distort the feature space and hurt the classification performance. Instead, they kept the decoder untouched, as AE architectures can avoid embedding distortion by preserving the local information of the data [154]. We also propose a simultaneous learning process for both reconstruction and clustering branches to avoid feature space corruption, similar to the approach taken in[140].

Once the CAE was pre-trained in the first stage (Algorithm 1), we incorporated a clustering branch at the output of the CAE bottleneck giving rise to the

proposed *DCEAC* model able to provide a soft label of class dependency. From the embedded representations $z_i = \{z_{i,1}, ..., z_{i,k}, ..., z_{i,C}\}$, with $C = 256$ the number of feature maps $z_{i,k} \in \mathbb{R}^{H \times W}$, we performed a spatial squeeze to obtain a feature vector $z'_i \in \mathbb{R}^C$ leading to a better label assignment. As depicted in Figure 3.4, a Global Average Pooling (GAP) layer (faded green) was used to reduce the feature maps $z_{i,k} \in \mathbb{R}^{H \times W}$, with $H = W = 32$, into the feature vector $z'_{i,k} \in \mathbb{R}^{1 \times 1}$ (see Eq. 3.2).

$$z'_{i,k} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} z_{i,k}(h, w) \tag{3.2}$$

After the GAP operation, a clustering layer (red box in Figure 3.4) was included to map each embedded representation z'_i onto a soft label $q_{i,j}$, which represents the probability of z'_i belonging to cluster j. In accordance with Eq. 3.3, $q_{i,j}$ was calculated via Student's T-distribution [155], keeping the cluster centres $\{\mu_j\}_1^K$ as trainable parameters.

$$q_{i,j} = \frac{(1 + ||z'_i - \mu_j||^2)^{-1}}{\sum_j (1 + ||z'_i - \mu_j||^2)^{-1}} \tag{3.3}$$

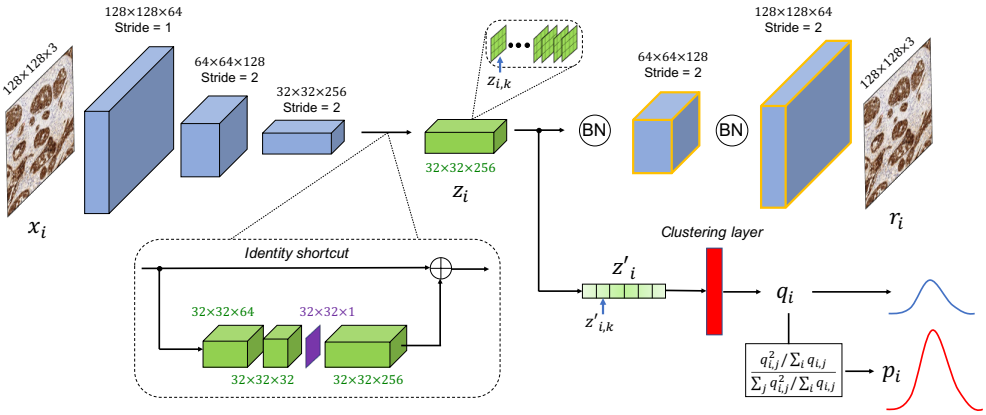


Figure 3.4: Architecture of the proposed DCEAC algorithm. The model is trained in an end-to-end manner by minimising both reconstruction and clustering loss functions. The reconstruction pretext task stabilises the feature space z_i avoiding the embedding distortion, while the clustering term predicts the soft-class assignments q_i for each input x_i.

Note that the cluster centres were initialised by running *kmeans* on the embedded features z'_i, as detailed in Algorithm 2. From here, a normal target distribution $p_{i,j}$ (defined in Eq. 3.4) was used as the ground truth during model training.

$$p_{i,j} = \frac{q_{i,j}^2 / \sum_i q_{i,j}}{\sum_j q_{i,j}^2 / \sum_i q_{i,j}} \tag{3.4}$$

The learning framework for the proposed *DCEAC* (Algorithm 2) was carried out by minimising a custom loss function (Eq. 3.5), where $\mathcal{L}_r$ and $\mathcal{L}_c$ are the reconstruction and clustering losses, respectively. $\gamma > 0$ is a temperature parameter used to prevent the distortion of the feature space, as $\gamma = 0$ would be equivalent to training just the CAE architecture.

$$\mathcal{L} = \mathcal{L}_r + \gamma \mathcal{L}_c \tag{3.5}$$

Specifically, the clustering loss was defined as Kullback-Leibler divergence ($KL = (P||Q)$) according to Eq. 3.6, whereas the MSE was used as a reconstruction loss function.

$$\mathcal{L}_c = \sum_i \sum_j p_{i_j} log \frac{p_{i,j}}{q_{i,j}} \tag{3.6}$$

As mentioned above, the autoencoders are responsible for preserving the local structure of the data, so the clustering term must provide only a slight contribution to the updating of the weights in order to avoid latent space corruption. Therefore, we empirically set $\gamma = 0.3$ for all experiments in the training process detailed in Algorithm 2.

Learning details for DCEAC training. As in the previous CAE pre-training, given an input batch X_b of $N = 32$ samples, we made use of the Adadelta optimiser with a learning rate of 0.5 to minimise the custom loss function $\mathcal{L}$ detailed in Algorithm 2. In this case, a single epoch takes 12.42 seconds to train each batch of $B = 94$ histological samples.

Algorithm 2 DCEAC training.

Data: Unlabelled training dataset $\mathcal{X} = \{X_1, ..., X_b, ..., X_B\}$
Results: Cluster assignment $\hat{y}_i$ for each histological sample x_i.

Step 1: Cluster centres initialisation
$\phi, \theta \leftarrow$ pre-trained CAE parameters;
$\mathcal{Z} \leftarrow f_\phi(\mathcal{X})$;
$\{\mu_j\}_{j=1}^{K} \leftarrow kmeans(\mathcal{Z})$;

Step 2: DCEAC training
for $e \leftarrow 1$ **to** ϵ **do**
  for $b \leftarrow 1$ **to** B **do**
    $X \leftarrow X_b \subset \mathcal{X}$;
    for $i \leftarrow 1$ **to** N **do**
      $z_i \leftarrow f_\phi(x_i)$;
      $r_i \leftarrow g_\theta(z_i)$;
      $z'_i \leftarrow GAP(z_i)$;
      $q_{i,j} \leftarrow \frac{(1+||z'_i - \mu_j||^2)^{-1}}{\sum_j (1+||z'_i - \mu_j||^2)^{-1}}$;
      $p_{i,j} \leftarrow \frac{q_{i,j}^2 / \sum_i q_{i,j}}{\sum_j q_{i,j}^2 / \sum_i q_{i,j}}$;
  $\mathcal{L}_r \leftarrow \frac{1}{N} \sum_{i=1}^{N} ||x_i - r_i||^2$;
  $\mathcal{L}_c \leftarrow \sum_i \sum_j p_{ij} log \frac{p_{i,j}}{q_{i,j}}$;
  $\mathcal{L} \leftarrow \mathcal{L}_r + \gamma \mathcal{L}_c$;
  Update ϕ, θ, μ_j using $\nabla_{\phi,\theta,\mu_j} \mathcal{L}$;

Step 3: Label prediction
for $b \leftarrow 1$ **to** B **do**
  $X \leftarrow X_b \subset \mathcal{X}$;
  for $i \leftarrow 1$ **to** N **do**
    $z'_i \leftarrow GAP(f_\phi(x_i))$;
    $q_{i,j} \leftarrow \frac{(1+||z'_i - \mu_j||^2)^{-1}}{\sum_j (1+||z'_i - \mu_j||^2)^{-1}}$;
    $\hat{y}_i \leftarrow argmax_j(q_{i,j})$;

3.4 Experimental results

3.4.1 Comparison with other state-of-the-art methods

In this section, we show a comparison between the proposed *DCEAC* model and the most relevant deep clustering-based works in the literature. In particular, we adapted the study carried out in [142], where the authors proposed a two-step learning strategy based on a Deep Embedded Clustering (DEC) model composed of fully connected layers. In the first step, they trained

the autoencoder network to extract domain knowledge from the unlabelled images. In the second step, after encoding the specific image information, Xie et al. [142] discarded the decoder structure to directly address the clustering phase from the learned feature space without considering the reconstruction error. However, later works such as [140] claimed that CAEs are more powerful than fully connected AEs for dealing with images. Thus, we adapted the previous DEC methodology by including convolution operations instead of fully connected layers. To do this, we followed the methodology proposed in [156], where stacked CAEs were originally proposed for hierarchical feature extraction. To perform a reliable state-of-the-art comparison, we fused both clustering [142] and CAE [156] to provide a refined DEC model (*rDEC*).

We also replicated the experiments conducted by Guo et al. [140], who proposed a hybrid learning for deep-clustering with convolutional autoencoders. The main difference with respect to the previous *rDEC* is that [140] kept the decoder term untouched during model training, resulting in a hybrid framework that combines reconstruction L_r and clustering L_c losses. The idea behind this is that the feature space embedded in *rDEC* could be distorted if only clustering-oriented loss is used. Therefore, they proposed leveraging the decoder structure to avoid latent space corruption by also considering the reconstruction error. Note that one of the main contributions of Guo et al. [140] lies in the proposed bottleneck, as they forced the dimension of the embedded features to be equal to the number of clusters along the fully connected layers. However, this is not scalable to other classification problems with higher-dimensionality input images or with a reduced number of clusters. Specifically, they applied the algorithms on the MNIST dataset composed of samples $x_i \in \mathcal{R}^{28\times28\times1}$ and provided an embedded space z_i with 10 features, depending on the $k = 10$ number of clusters. In our case, we deal with $128 \times 128 \times 3$ pixel images, where the high resolution is essential for the classification performance, unlike in the MNIST dataset. Furthermore, our goal is to classify the histological samples into $k = 3$ classes, so replicating the architecture of [140] is unfeasible as the decoder term would be unable to reconstruct the images from only three feature values. Therefore, to drive a convincing comparison with [140], we kept the same architectures and training details proposed in this work, but removed the convolutional attention module as it is one of our own main contributions. Henceforth, we will refer to this approach as *rDCEC*.

3.4.2 Quantitative results

In this section, we report the unsupervised classification performance achieved by the aforementioned *rDEC* [142] and *rDCEC* [140] algorithms in comparison to our proposed *DCEAC* model. Conventional methods based on running the clustering algorithms (*kmeans*, *spectral* and *agglomerative*) on the feature space were also considered to find out the performance difference between the proposed model and traditional techniques. These conventional approaches will be referred to as *AE+kmeans*, *AE+spectral* and *AE+agg*, respectively. In Table 3.1, we present the class performance obtained from the conventional clustering methods to show how well the three algorithms classify the 2995 histological patches with non-tumour (NT), mild (M) and infiltrative (I) patterns. Similarly, we also evaluate the per-class behaviour of the deep clustering-based algorithms (*rDEC*, *rDCEC* and *DCEAC*) in Table 3.2. In addition, the micro- and macro-average classification results are reported in Table 3.3. Both metrics provide information about the overall average performance of the classification models, but the micro-average takes into account the imbalance between classes, which allows for a truer picture of the models' behaviour than does the macro-average. Comparison among the different methods is handled by means of different figures of merit, such as sensitivity (SN), specificity (SP), F-score (FS), accuracy (ACC) and area under the ROC curve (AUC).

To improve the comparison between the six learning approaches, we represent in Figure 3.5 the latent space laid out by each clustering model with its respective confusion matrix. While the confusion matrix provides information about the classification ability of each model, the representation of the embedded features contributes to a more comprehensive clustering scenario for bladder cancer grading. Thus, while the latent space representation would fit better in the qualitative section, the confusion matrix provides a quantitative perspective that aids the interpretation of the embedded feature map.

Table 3.1: Unsupervised results per class from conventional methods.

	NON-TUMOUR			**MILD**			**INFILTRATIVE**		
	AE+kmeans	*AE+spectral*	*AE+agg*	*AE+kmeans*	*AE+spectral*	*AE+agg*	*AE+kmeans*	*AE+spectral*	*AE+agg*
SN	0.9345	0.9227	**0.9869**	0.5020	0.6327	**0.7415**	0.5105	**0.5827**	0.2533
SP	0.9870	0.9718	**0.9960**	0.7659	**0.8210**	0.6249	0.6556	0.7398	**0.8307**
FS	0.9475	0.9203	**0.9875**	0.5754	0.6958	**0.6960**	0.4052	**0.4969**	0.2896
ACC	0.9736	0.9593	**0.9937**	0.6364	**0.7285**	0.6821	0.6187	**0.6938**	0.6838

Table 3.2: Unsupervised results per class from deep-clustering methods.

	NON-TUMOUR			MILD			INFILTRATIVE		
	rDEC	*rDCEC*	*DCEAC*	*rDEC*	*rDCEC*	*DCEAC*	*rDEC*	*rDCEC*	*DCEAC*
SN	**1**	**1**	0.9987	0.8082	0.8952	**0.9041**	0.4659	0.5105	**0.6168**
SP	0.9319	0.9780	**0.9978**	**0.8262**	0.7862	0.8118	0.8782	0.9319	**0.9364**
FS	0.9094	0.9689	**0.9961**	0.8129	0.8458	**0.8613**	0.5112	0.5971	**0.6841**
ACC	0.9492	0.9836	**0.9980**	0.8174	0.8397	**0.8571**	0.7733	0.8247	**0.8551**

Table 3.3: Unsupervised results in terms of micro- and macro-average achieved from both conventional and deep clustering methods.

	MICRO-AVERAGE						MACRO-AVERAGE					
	AE+kmeans	*AE+spectral*	*AE+agg*	*rDEC*	*rDCEC*	*DCEAC*	*AE+kmeans*	*AE+spectral*	*AE+agg*	*rDEC*	*rDCEC*	*DCEAC*
SN	0.6144	0.6938	0.6798	0.7699	0.8240	**0.8551**	0.6490	0.7127	0.6606	0.7580	0.8019	**0.8399**
SP	0.8072	0.8469	0.8399	0.8850	0.9120	**0.9275**	0.8028	0.8442	0.8172	0.8788	0.8987	**0.9153**
FS	0.6144	0.6938	0.6798	0.7699	0.8240	**0.8551**	0.6427	0.7043	0.6577	0.7445	0.8039	**0.8472**
ACC	0.7429	0.7959	0.7865	0.8466	0.8827	**0.9034**	0.7429	0.7959	0.7865	0.8466	0.8827	**0.9034**
AUC	0.7259	0.7784	0.7389	0.8184	0.8503	**0.8776**	0.7259	0.7784	0.7389	0.8184	0.8503	**0.8776**

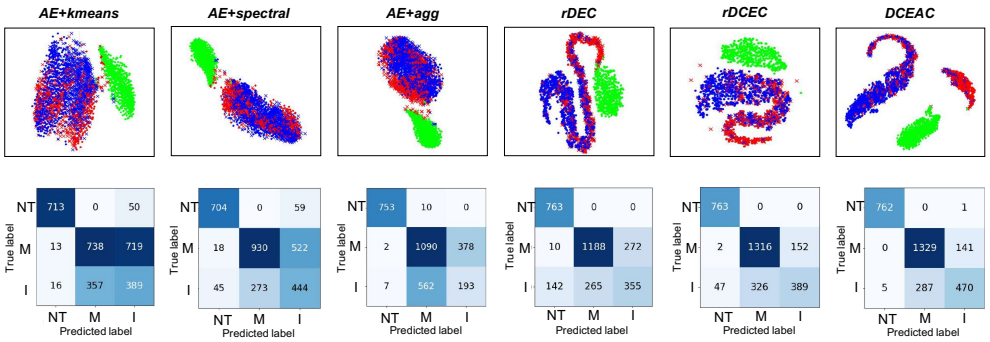


Figure 3.5: TSNE outcomes and confusion matrices. The top row with the scatter graphics corresponds to the latent space representation from the clustering classification achieved by each method. The T-distributed Stochastic Neighbour Embedding (TSNE) tool was employed to illustrate the feature space in a 2D map. Well- and missclassified embedded features are represented by spots and crosses, respectively. The green, blue and red colours refer to the non-tumour (NT), mild (M) and infiltrative (I) patterns. In the bottom row, a confusion matrix per method shows the ability of each one to discern the MIBC aggressiveness.

3.4.3 Qualitative results

In an attempt to incorporate an interpretative perspective for the reported quantitative results, we computed the class activation maps (CAMs), which highlight the regions to which the model pays attention in order to predict the class of each sample. This often helps to find hidden patterns associated with a specific class or to determine whether the label prediction is based on the

same patterns as the clinicians' findings. In this way, the reported heat maps lead to a better understanding of the embedded feature space by pinpointing areas of the histological patches that are decisive in cluster assignment.

As can be deduced from Figure 3.5, the major challenge in the classification of MIBC lies in distinguishing mild (M) and infiltrative (I) cancerous patterns, as expected. For this reason, in Figure 3.6 we present several examples of heat maps corresponding to missclassified samples to elucidate why the proposed model is flawed. We also show examples of well-predicted CAMs to evidence the relevant structures to which the network pays attention when predicting correctly. Specifically, we show five examples per case to make clear the criteria followed by the proposed model to determine the class. In the green frame of Figure 3.6, we illustrate well-classified mild (a-e) and infiltrative (f-j) histological patterns. Additionally, in the red frame, we show bladder cancer samples with a mild pattern missclassified as tumour budding (k-o), and vice versa (p-t). The findings from the CAMs will be discussed in Section 3.5.

3.5 Discussion

3.5.1 On quantitative results

From Table 3.1 and 3.2, we can observe that all models work well for detecting the non-tumour class. In the conventional approaches (Table 3.1), the *AE+agg* algorithm shows slightly better performance, but at the cost of greatly compromising detection of the rest of the classes. In contrast, the proposed *DCEAC* model (Table 3.2) achieves the highest performance for all metrics except for sensitivity, as the model missclassifies a non-tumour sample as an infiltrative case (see Figure 3.5). Concerning the mild (M) and infiltrative (I) patterns, the deep-clustering models notably improve the success of unsupervised classification compared to conventional approaches. Regarding the mild class, *AE+spectral* and *AE+agg* clustering methods show similar behaviour, but the proposed model provides the highest performance with an increase in accuracy of 12.85% with respect to the best conventional approach. Only *rDEC* surpasses it in any metric (by 1% in specificity), but in exchange for a 10% drop in sensitivity relative to the proposed *DCEAC*. During the evaluation of the infiltrative class, the *AE+agg* method drops strongly, which places *AE+kmeans* and *AE+spectral* as much more reliable conventional clustering algorithms. Comparing all algorithms side-by-side, the proposed *DCEAC* method shows the best results for all metrics, specially the F-score, where *DCEAC* outperforms the other approaches by more than 10%.

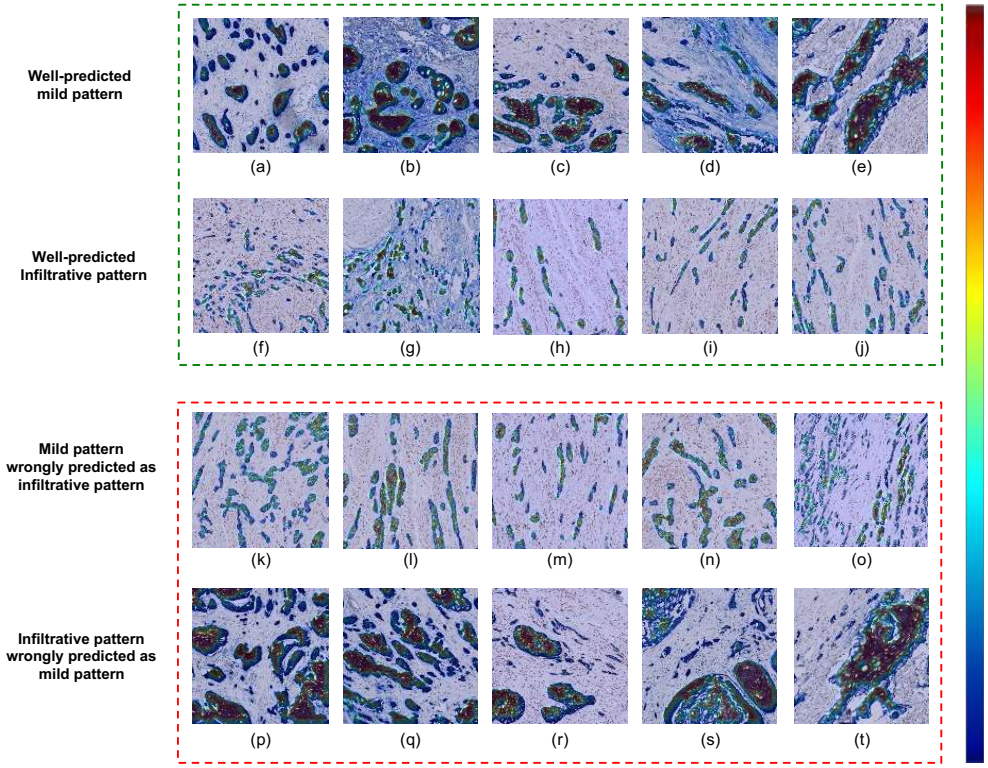


Figure 3.6: Class activation maps highlighting the regions that the proposed DCEAC model considers relevant for the class prediction. The green frame refers to well-predicted images with mild (M) and infiltrative (I) patterns, whereas the red frame corresponds to missclassified samples where disease aggressiveness has been confused. The more important the areas, the warmer the colour in which they are represented in the heat maps, so that the blue tones denote less important regions, and red tones refer to more important ones.

Table 3.3 reports the overall performance of the models, in terms of micro- and macro-average. As mentioned above, the micro-average results take into account the imbalance between classes, which is an important aspect in this study, as the samples with mild pattern are over-sampled. Nevertheless, the proposed *DCEAC* model consistently outperforms the other clustering methods by 2-3% across both micro- and macro-averaging, as can be seen in Table 3.3. As a final remark on the quantitative results, it is worth noting that the expert's decision coincides with the proposed artificial intelligence system in 90.34% of the cases, according to the average accuracy.

A reinforcement of the quantitative results is reported in Figure 3.5. In the confusion matrices, it is clear from the range of colours that all models tend to confuse mild cancerous and infiltrative patterns. Conventional algorithms demonstrate a very low ability to discern between cancerous samples as most of the images are predicted as a mild pattern due to oversampling of that class. This changes when deep-clustering algorithms are profiled. Specifically, the *rDEC* model improves the classification of carcinogenic images but greatly compromises that of the non-tumour class by missclassifying samples with infiltrative pattern. In contrast, the *rDCEC* model improves the results by significantly decreasing the instances of tumour budding samples erroneously predicted as non-tumour cases. In addition, *rDCEC* increases the number of true positives for tumour samples. However, this model presents a major shortcoming in predicting infiltrative patterns, as a large number of them are wrongly labelled as mild. Unequivocally, the proposed *DCEAC* model provides the best classification results. The number of samples with tumour budding missclassified as non-tumour cases decreases to a minimum, in contrast to the aforementioned methods. Moreover, the number of true positives increases for both mild and infiltrative patterns compared to the results of the other methods, while false positives and false negatives are reduced.

The representation of the embedded feature space offers a visual perspective of the quantitative results. We can observe that the conventional approaches are able to roughly discern between non-tumour and carcinogenic histological samples. However, the point clouds are too fuzzy to separate mild from infiltrative classes. Contrarily, the *rDEC* model shows a better distribution of the embedded data, although the features relative to each class are still close together in the latent space. This improves in the case of the *rDCEC* model, where independent clusters start to become apparent. The non-tumour features (shown in green) become unmarked in the representation space and the embedded tumour samples start to disperse into different classes of cluster. Indisputably, the *DCEAC* algorithm provides the best embedding representation as the features are distributed throughout the latent space, forming independent clusters according to a specific class. This further strengthens our confidence in the ability of the proposed model to discern between non-tumour, mild and infiltrative histological patterns.

From the above in-depth analysis of the quantitative and interpretative results, several conclusions can be drawn. The first is that the use of deep learning techniques improves classification performance compared to conventional clustering approaches. As expected, all deep clustering-based methods, i.e. *rDEC*, *DCEC* and *DCEAC*, outperform the baseline based

on the traditional *kmeans*, *spectral* and *agglomerative* algorithms. This is because the deep-clustering models allow for a more extensive learning stage in which the embedded features conform to a target distribution, unlike conventional algorithms, which modify the clusters iteratively without updating the feature learning. Additionally, we can observe that models with both the reconstruction and clustering branches integrated into a unified framework provide better results than the *rDEC* model, which carries out the learning process in two independent stages. The reason behind *rDCEC* and *DCEAC* outperforming *rDEC* lies in the preservation of the local structure of the embedded data. Since *rDCEC* and *DCEAC* models have a connected output between the clustering and reconstruction stages, the clustering term can transfer class information to the reconstruction term, which is responsible for updating the weights of the encoder network. In this way, the embedded features can be optimised by incorporating the class prediction without distorting the latent space, thanks to the decoder structure. Finally, the proposed *DCEAC* model shows substantial performance improvements over the rest of the approaches. This is due to the inclusion of the convolutional attention block, which allows for the refinement of the latent space to provide more suitable features for the clustering phase.

3.5.2 On qualitative results

As observed in Figure 3.6, the proposed *DCEAC* model focuses on tumour cell nests (Figure 3.6, a-c) and interconnected tumour bands (Figure 3.6, d-e) when predicting samples with a mild pattern. This implies that the proposed network has learnt by itself to associate nodular and trabecular structures with a mild pattern of disease. Furthermore, the *DCEAC* model recognises small clusters of isolated buds (Figure 3.6, f-g) or tumour cell cords (Figure 3.6, h-j) as structures characteristic of the infiltrative pattern. These findings are evidenced in the heat maps corresponding of the well-predicted samples.

In the case of the wrong predictions (red frame in Figure 3.6), we can observe that the proposed network maintains consistency in determining the class of each sample. The histological patches in Figure 3.6 (k–o), in which the network highlights small filaments reminiscent of tumour budding structures, are similar in appearance to infiltrative patterns. However, the true label assigned by the expert for these samples was a mild pattern, as trabecular structures are often difficult to distinguish from the cell cords of the infiltrative pattern. The qualitative results thus present an opportunity here for the model's suggestions to serve a role similar to that of a second opinion, and lead pathologists to reconsider their diagnoses.

In contrast, in the cases of the Figure 3.6 (p-t), samples with an infiltrative pattern are erroneously predicted by the model as mild cases. In these histological patches, the proposed network focuses on larger structures related to nodular or trabecular patterns, but ignores small, isolated tumour cells that lead to increased severity of bladder cancer. It follows that, although the final prediction could be wrong, the pattern recognition accomplished by the model maintains consistency. Note that the model is wrong because patterns that belong to a different class coexist in the same histological patch, so we will face this problem in future research lines.

In summary, the proposed *DCEAC* model demonstrates, through heat maps, a high-confidence prediction as it is able to focus on the same patterns as the clinicians, without having prior information from them. As mentioned above, the expert's opinion and the proposed model coincide in most cases (specifically, 90.34% of cases). Thus, the artificial intelligence system could help as a computer-aided system for process review, which would lead to an improvement in the quality of diagnosis without the need to involve other experts. In addition, the proposed system could be used to help inexperienced pathologists by suggesting annotations for specific areas of interest.

3.6 Conclusion

In this paper, we have proposed a novel self-learning framework based on deep-clustering techniques to grade the severity of bladder cancer through histological samples. Immunohistochemistry staining methods have been considered to enhance the NT, M and I patterns, according to the literature. The proposed DCEAC outperforms other conventional and deep clustering-based methods, achieving an average accuracy of 0.9034 for MIBC grading. The CAMs show that the proposed system is able to self-learn the same structures as clinicians to associate patterns with the correct aggressiveness level, without incurring prior annotation steps. Therefore, our fully unsupervised approach bridges the gap with respect to other supervised algorithms, as the proposed system does not require the involvement of experts for model training.

In future research lines, we will work on improving the accuracy of tumour sample classification when structures of different growth patterns appear in the same image. We will propose the use of convolutional variational autoencoders considering probabilistic and deterministic attention modules. Finally, we will pursue an end-to-end system in which no prior raw annotations are necessary.

Chapter 4

Glaucoma Detection from Raw SD-OCT Volumes: a Novel Approach Focused on Spatial Dependencies

The content of this chapter corresponds to the author version of the following published paper: García, G., Colomer, A. & Naranjo, V. Glaucoma detection from raw SD-OCT volumes: A novel approach focused on spatial dependencies. Computer Methods and Programs in Biomedicine, 200, 105855 (2021).

Contents

Glaucoma Detection from Raw SD-OCT Volumes: a Novel Approach Focused on Spatial Dependencies

Gabriel García[1], Adrián Colomer[1] and Valery Naranjo[1]

[1]Instituto de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, 46022, Valencia, Spain

Abstract

Glaucoma is the leading cause of blindness worldwide. Many studies based on OCT imaging have been developed in the literature to help ophthalmologists through AI techniques. Currently, 3D spectral-domain (SD)-OCT samples have become more important, as they could enclose promising information for glaucoma detection. To analyse the glaucoma-specific knowledge hidden in 3D scans, we propose for the first time a deep learning methodology based on leveraging the spatial dependencies of the features extracted from the slides of the volumes. The experiments were performed on a database composed of 176 healthy and 144 glaucomatous SD-OCT volumes centred on the optic nerve head. The proposed methodology consists of a two-step learning methodology to train: i) a slide-level feature extractor, and ii) a volume-based predictive model. The slide-level discriminator consists of a new architecture designed to extract the useful information from the slides of the SD-OCT cube. The volume-based predictive model bases on Long Short-Term Memory (LSTM) networks to combine the recurrent dependencies extracted from the slide-derived feature space. A sequential-weighting module (SWM) is proposed to map the latent embedding into a refined feature vector before the volume-level classification. The discriminator reports AUC values higher than 0.93 both in the primary and external test sets. The end-to-end system (discriminator + LSTMs + SWM) achieves an AUC of 0.8847 in the glaucoma prediction from SD-OCT volumes, which outperforms other state-of-the-art studies focused on 3D deep learning architectures. The class activation maps evidence that the proposed model pays attention to the same OCT-specific areas that experts focus on for glaucoma detection.

4.1 Introduction

Glaucoma is a group of progressive optic neuropathies that affects the optic nerve causing several visual field defects and structural changes [73]. This chronic disease is the leading cause of blindness worldwide [74], with a number of estimated cases of 111.8 million in 2040, according to [157]. Early diagnosis of glaucoma is essential for timely treatment in order to avoid the irreversible vision loss [74]. Currently, there is no single accurate test to certify glaucoma, so that the procedure includes a lot of hardworking tests such as pachymetry (to measure the thickness of the cornea), tonometry (to assess the intraocular pressure), visual field tests and a subjective examination and interpretation of optical features from different experts who often disagree [158]. In this context, techniques based on image analysis like fundus image and optical coherence tomography (OCT) have become very important for the diagnosis and management of this degenerative disease. OCT imaging modality [159] is a non-contact and non-invasive technique able to quantify several retinal structures through generating high-resolution 2D and 3D images of the retina. Ophthalmologists usually make use of these 2D-OCT images centred on the optic disc to analyse structural changes in the retinal nerve fibre layer (RNFL) and in the ganglion cell inner plexiform layer (GCIPL). Both structures are reported as useful biomarkers of glaucoma for the disease progression [160]. Otherwise, fundus image analysis is placed as a great cost-effectiveness technique which has reported promising results in the detection of several eye-focused diseases [5, 161, 162]. Fundus image-based studies are cheaper than OCT, but OCT is the quintessential imaging technique for glaucomatous damage evaluation [39]. This is because fundus photography is colour-dependent on the training dataset and its interpretation remains subjective [163, 164], whereas OCT modality can provide reproducible and objective measurements of optic nerve head (ONH) and RNFL thickness [165]. Besides, glaucoma disease is evident in the deterioration of the cell layer around the optic disc, which is very hard to distinguish in the 2D projection of the fundus images. Therefore, as OCT imaging modality focuses on the depth axis to identify structural retinal changes, glaucoma disease can be easier detected via OCT, instead of fundus image. Furthermore, OCT system can provide high-resolution three-dimensional images of the macula and ONH in the spectral domain (SD), which emerges as a powerful tool for detecting glaucoma [39]. However, due to around 30 million of OCT scans are acquired each year, experts rarely scroll through the entire cube because it supposes a workload difficult to face [166]. In this paper, we propose a promising volume-based predictive model to claim the added value that SD-OCT volumes can provide for glaucoma diagnosis.

4.1.1 Related work

4.1.1.1 2D-OCT approximation for glaucoma detection

Many OCT-based studies have been proposed in the literature to address the automatic detection of glaucoma aimed at reducing the workload and the rate of discordance between experts.

Hand-driven learning on 2D-OCT projection. Most of glaucoma-intended studies made use of 2D-OCT scans centred on the optic disc, a.k.a circumpapillary images, due to their known potential when diagnosing [167]. To the best of the authors' knowledge, all the circumpapillary-based works were performed via hand-driven learning methods [16, 168], which requires hand-crafted encoding phases before the classification stage for glaucoma detection.

Deep learning on 2D-OCT projection. The use of convolutional neural networks (CNNs) is another way to address the glaucoma detection from circumpapillary OCT images. This approach directly operates on the 2D-OCT scans without defining previous biomarkers, as we proposed in our previous study [62]. All the studies found in the literature under these conditions are based on fundus images [5, 169] or RNFL probability maps [54, 170] combining fundus images and OCT B-scans; however, no previous studies were conducted just from circumpapillary OCT images. This fact could be explained taking into account that the researchers focus their efforts on identifying useful patterns (e.g. RNFL and GCIPL) capable of providing a tangible interpretation for the clinicians. For that reason, many other studies were carried out for the sole purpose of segmenting retinal layers[171, 172].

4.1.1.2 3D-OCT approximation for glaucoma detection

Going deeper into the glaucoma detection, the real challenge today lies in the analysis of the unknown potential enclosed in the 3D-OCT scans, as the specialists claim that SD-OCT volumes hide a key knowledge that is not currently being traced due to their associated heavy workload. Therefore, we propose here a clinical decision support system based only on the analysis of ONH-centred cubes to claim the importance of the 3D cross-sectional information about the glaucoma diagnosis.

Hand-driven learning on 3D-OCT approach. Similarly to the 2D approximation, some studies in the literature applied hand-crafted algorithms on 3D scans to face the glaucoma discrimination [173–175]. Both [173] and

[174] manually extracted features related to the RNFL and the optic nerve throughout the cube. The authors proposed a similar methodology, but they tested their models on different databases. In [173], an AUC of 0.877 was achieved using a random forest classifier from a database composed of 46 healthy and 57 glaucomatous patients, whereas in [174], the same researchers provided an AUC of 0.818 by applying bagging methods on a database of 48 and 62 healthy and glaucomatous patients, respectively. Xu et al. [175] made use of a superpixel segmentation technique before addressing the feature extraction stage. They combined the features extracted from the superpixel maps with other RNFL measurements to feed an adaptive boosting classifier. An AUC of 0.855 was reported from a database of 44 healthy and 89 glaucomatous eyes.

Deep learning on 3D-OCT approach. The use of deep learning methods to address the glaucoma detection via SD-OCT volumes has been increased in recent times. Most of the recent studies claims the current interest of OCT volumes for glaucoma diagnosis [50, 53, 76]. A research group from Hong Kong deserves a special mention because most of the contributions in this field come from their work. They carried out two closely similar studies, [53] and [76] to detect glaucoma by means of 3D-CNNs. The main differences between them lied in the database and inclusion/exclusion criteria, as the authors concluded in [53]. Noury et al. [53] made use of a private database composed of 316 glaucomatous and 247 healthy eyes from people of different ethnicity. They developed an end-to-end classification model based on the network proposed in [49]. The researchers achieved an AUC of 0.8883 in the primary test set, but this value was lower when testing external datasets. In contrast, the authors in [76] applied similar techniques on a homogeneous database only composed of Chinese Asian people. Particularly, 2926 glaucomatous and 1961 healthy eyes. This work reported an AUC of 0.969, a sensitivity of 0.89, a specificity of 0.96 and an accuracy of 0.91 when testing the primary dataset. However, the results fell when assessing the network using an external database from Stanford, reaching values of 0.893, 0.78, 0.79 and 0.80, respectively. More recent works from the same authors [52, 75] performed a multi-output architecture by including other well-known glaucoma-specific indices such as Visual Field (VF), Mean Deviation (MD) and Pattern Standard Deviation (PSD). In their method, a neural branch of the network was responsible for the classification between normal and glaucomatous cases, whereas the other branch was intended to regression tasks for predicting VF, MD and PSD values. Therefore, the model was fed with information from VF, MD and PSD metrics during the backward propagation step to update the weights in each epoch taking into account interesting parameters associated with the glaucoma disease. However, these two last studies are not comparable

with our work because additional information was used besides the raw OCT volumes, unlike the works [53, 76] accomplished by the same research group. Another interesting study was carried out by the IBM research group in [50], where the authors made a comparison between hand-driven and data-learning approaches. They proposed a 3D-CNN architecture trained from scratch and they achieved an AUC of 0.94 in the prediction of the test set. However, it should be noted that, in this case, the experiments were performed on a significant unbalanced database, whose test set was composed of 17 healthy and 93 glaucomatous patients.

In this context, other works could be mentioned because they also applied deep learning techniques on SD-OCT volumes, but with other purposes. For example, in [176], Ran et al. proposed an algorithm for discriminating ungradable OCT optic disc scans. In [166], the Kurmann et al. implemented deep learning techniques to detect specific Age-Related Macular Degeneration (AMD) patterns in the B-scans of the 3D cubes. Also, De Fauw et al. in [177] applied artificial intelligence algorithms on OCT volumes to diagnosis several retinal injuries via tissue segmentation.

4.1.2 Contribution of this work

This paper documents several key contributions concerning the glaucoma detection from SD-OCT volumes. Unlike the previous studies that addressed the problem using 3D CNNs, we reveal a new approach based on extracting features from the B-scans by an innovative 2D-CNN, and preserving the embedding dependencies via LSTM networks [178]. The combination of CNNs and LSTM networks has been successfully performed in recent studies to identify pathological biomarkers associated to AMD and diabetic macular edema (DME) [166], as well as to predict the progression of other ophthalmic diseases using different slit-lamp images [179]. However, as far as we are aware, this is the first study that suggests the combination of CNNs and LSTMs to address the glaucoma detection, by assuming each spatial slide of the volume as a temporary instance. As a novelty, to attain the feature-extraction stage, we propose a new slide-level discriminator based on a pre-trained 2D-CNN model able to discern between healthy and glaucomatous cases just from raw circumpapillary OCT images. The proposed 2D-CNN feature extractor is composed of a novel combination of pre-trained convolutional blocks in parallel with residual modules trained from scratch. Additionally, an attention block was included via skip-connections to focus on local specific-glaucoma areas during the training phase. We also propose an innovative way of codifying the LSTM outputs by means of a sequential-weighting module (SWM), which

allows for the refinement of the feature space before the classification stage. The flowchart of the designed end-to-end system is exposed in Figure 4.1, where we represent how the pre-trained 2D-CNN extracts the features from the SD-OCT slides and how the 3D information is analysed making use of LSTM networks to finally predict the class of each specific ONH-centred cube.

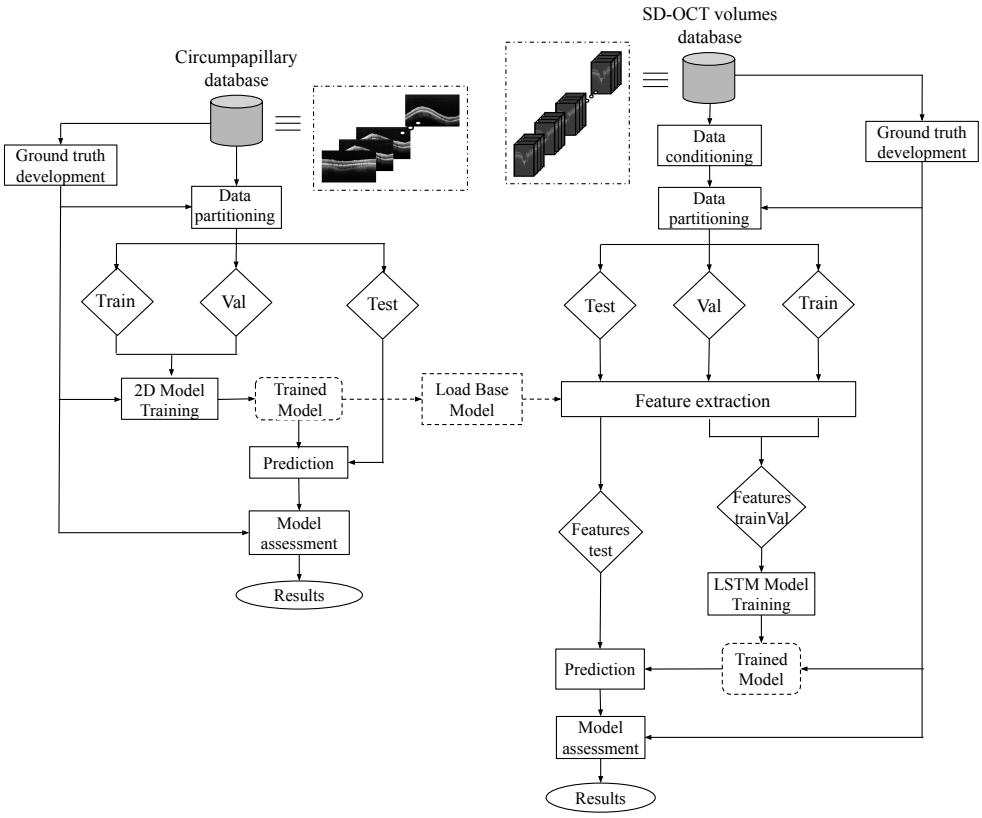


Figure 4.1: Flowchart of the proposed glaucoma detection framework based on combining CNN and LSTM networks to distinguish between healthy and glaucomatous eyes from raw SD-OCT volumes.

In [50], Maetschke et al. defended the use of 3D convolutions to be able to accomplish the 3D Class Activation Maps (CAMs), as otherwise, the resulting CAMs would be 2D and the depth information would be lost. Contrarily, our LSTM-based model is able to leverage the spatial dependencies extracted from the SD-OCT slides to compute the 2D-CAMs sequentially. Thereby, we

enable an interpretation of SD-OCT volumes based not only on identifying the regions of interest (ROIs) of each slide, but also the most relevant B-scans of the volume for glaucoma classification. It is important to note that we also replicated several architectures proposed in the literature to make a direct comparison between different methods. We tested in our database the state-of-the-art architectures intended to glaucoma detection just from raw SD-OCT volumes, i.e. the works performed in [76] and [50].

4.2 Material

Three different and independent databases were employed to accomplish this study, as indicated in Table 4.1. Two of them are related to circumpapillary OCT images and they were used to train and validate the proposed slide-level feature extractor. The third database is composed of the SD-OCT volumes from which we built the predictive models for glaucoma detection. Both the circumpapillary and SD-OCT volumes databases are centred around the ONH of the retina to extract the B-scans. Note that, although circumpapillary and volume slides are extracted following different acquisition options, the structure involved in both kind of scans is the same, i.e. the fibre layers of the retina. For that reason, we used the models trained on circumpapillary images as a feature extractor of the volume slides.

The first dataset (*circ-DB-1*), intended to train the slide-level discriminator, consists of 249 cross-sectional images around the ONH of the retina. Specifically, 156 healthy and 93 glaucomatous samples from 174 patients were labelled by an expert to create the ground truth for training and evaluating the models. The second circumpapillary database (*circ-DB-2*), which comes from another hospital, was used to perform an external validation of the proposed feature extractor. Particularly, *circ-DB-2* is composed of 336 OCT images (143 glaucomatous and 193 healthy cases from 199 patients) which were annotated by another ophthalmologist. It should be noted that a *Heidelberg Spectralis OCT*-called system was used to extract the B-scans (496×768 pixels) from both databases with an axial resolution of 4-5 μm. Patients with primary open-angle glaucoma (POAG) were included in the study, whereas subjects with other eye diseases, e.g. cataract, closed-angle glaucoma, and pseudo exfoliation syndrome were excluded. More information related to the age and gender of the patients is detailed in Table 4.2.

The third database (vol-DB-3) contains the spectral-domain OCT samples that we used to develop our volume-based predictive model. It consists of

320 OCT scans centred on the ONH which were captured on a *Topcon* 2000 OCT machine. This equipment enables measures up to 45^{o} and a depth of 2.3 mm with a resolution of less than 6 μm. Specifically, vol-DB-3 is composed of 176 healthy and 144 glaucomatous 3D-scans of $885 \times 512 \times 128$ voxels per volume, as detailed in Table 4.1. An expert ophthalmologist performed a volume-level annotation of the database containing cases of 200 patients with an age comprised between 18 and 80 years (see Table 4.2). Concerning the inclusion and exclusion criteria for diagnosis, the healthy group included samples with best-corrected visual acuity 20/40 or better, normal intra-ocular pressure (IOP) and normal-appearing optic nerves. Contrarily, scans with refractive error of >5D of the sphere, history retinal disease, intra-ocular pressure >21 mmHg or unusable OCT were excluded from the study. For the glaucoma group, the inclusion criteria lied in any glaucomatous visual field defect, whereas the exclusion rules comprised refractive error of >5D of the sphere, optic nerve-related diseases and unusable OCT samples.

Table 4.1: Breakdown of the databases used for glaucoma detection. circ-DB-1 refers to the database employed to train the feature extractor model. circ-DB-2 corresponds to the dataset used in the external validation of the feature extractor. circ-DB-3 refers to the database utilised to train and evaluate the volume-based predictive model.

Database	Label	Patients	Samples	Dimensions
circ-DB-1	Healthy Glaucoma	107 (61.49%) 67 (38.51%)	156 (62.65%) 93 (37.35%)	496×768
circ-DB-2	Healthy Glaucoma	99 (49.75%) 100 (50.25%)	193 (57.44%) 143 (42.56%)	496×768
vol-DB-3	Healthy Glaucoma	100 (50.00%) 100 (50.00%)	176 (55.00%) 144 (45.00%)	$885 \times 512 \times 128$

Table 4.2: Demographic data related to the age and gender of the patients.

	Age		Gender	
	Range	$\mu \pm \sigma$	**Male**	**Female**
circ-DB-1	[15-89]	55.95±18.77	74 (42.54%)	100 (57.47%)
circ-DB-2	[24-93]	60.82±12.32	86 (43.22%)	113 (56.78%)
vold-DB-3	[18-80]	50.13±15.54	89 (44.50%)	111 (55.50%)

The protocol used for glaucoma labelling was carried out following the European guideline for Glaucoma diagnosis. A thorough examination includes intra-ocular pressure analysis (using Goldmann applanation tonometry), study of the central corneal thickness, assessment of the anterior chamber angle (Gonioscopy), optic nerve head assessment (via slit lamp examination), Standard Automated Perimetry (using Octopus system) and measurement

of the thickness of retinal nerve fibre layer and ganglion cell layer (with OCT+HRT equipment). Based on these examinations, B-scans and SD-OCT volumes were labelled as healthy or glaucoma. Note that any of the following criteria, if repeatable, was considered sufficient evidence of glaucomatous visual field defect: glaucoma hemifield test outside normal limits, pattern standard deviation with p-value < 0.05 or a cluster of three points or more in the pattern deviation in a single hemifield (superior or inferior) with p-values < 0.05, one of which must have a p-value < 0.01.

4.3 Methodology

4.3.1 Slide-level feature extractor design

The objective here is to build a 2D-CNN architecture able to extract discriminatory features from the slides of the SD-OCT volumes. In our previous work [62], we carried out a validation of different architectures making use of the raw circumpapillary OCT samples. We considered some of the most common state-of-the-art architectures, as well as other CNNs trained from scratch. As detailed in [62], we proposed shallow networks to deal with the small amount of information and used data augmentation techniques to alleviate that problem. Additionally, we fine-tuned some of the most popular architectures of the literature, such as VGG16, VGG19, ResNet50, InceptionV3 and Xception [180], to take advantage of the wide knowledge acquired by these networks when they were trained on the *ImageNet* dataset. Thus, we loaded the weights ω pre-trained with around 14 million of natural images to initialise the coefficients of the networks. Then, we performed a *deep fine-tuning* strategy [111] to freeze the coefficients of the three first convolutional blocks and retrain the last ones making use of the specific samples. Note that, we replicated $\times 3$ the channels of the grey-scale images to adapt the input dimensionality to the fine-tuned CNNs. We also applied a $\times 0.5$ down-sampling to deal with the GPU memory constraints, which usually appear when data volumes are addressed.

In [62], the VGG family of networks reported the best glaucoma detection performance from raw circumpapillary OCT images. Therefore, to accomplish this study, we made use of these family of architectures as a starting point to develop the new feature extractor. We kept the same *deep fine-tuning* strategy previously conducted in [62]; however, in this paper, we propose two innovative modules to improve the models' performance through residual convolutions.

The first module (M_{res}) consists of a combination of the fine-tuned VGG16 architecture with a residual structure applied in parallel to the unfrozen blocks, followed by a 1×1 convolution layer, as depicted in Figure 4.2. We connected the fine-tuned structure with other convolutional blocks to propagate the information from initial to final layers, using residual connections in a novel way. This makes possible to mitigate the problem of vanishing gradients by allowing the shortcut to flow through the gradient of a deeper architecture. Unlike the traditional skip-connections defined in [181], where a specific input fed the network at two different points, the proposed system introduces a convolutional shortcut inspired by the basic structure of the ResNet-50 architecture. Such structure aims to optimise the dimensionality of the filters by alternating convolution layers of 1×1 and 3×3 kernel sizes, which are represented in Figure 4.2 in green and blue boxes, respectively. In addition, an initial batch normalisation layer (in brown) and a final max-pooling layer (in red) were implemented as a part of the residual block (see Figure 4.2). Note that kernel and stride sizes of 4×4 were specified for the max-pooling layer, ensuring the consistency of the filter dimensions to concatenate the residual features with the output from the Block_5 VGG16. Finally, the M_{res} module containing a 1×1 convolution layer generates a volume of features $G = \{g_1, g_2, ..., g_k, ..., g_C\}$, where $C = 512$ is the number of filters in the volume, and g_k the *k-th* feature map with dimensions $H \times W = 7 \times 12$.

The second module (M_{att}) of the proposed network includes an attention block characterised by a succession of 1×1 convolutional layers intended to refine the features in the spatial dimension. The proposed module is a kind of bottleneck architecture composed of a batch normalisation layer followed by two successive ReLu-activated convolutions, in which the size of the filters is decreased progressively, as observed in Figure 4.2. Also, a 1×1 convolution layer with a unique filter passing through a sigmoid function (purple) was used to recalibrate the inputs. At the end of the bottleneck, it includes another 1×1 convolution layer which increases the number of filters to make possible the concatenation between the inputs and the outputs of the attention block. In this way, a basic skip-connection was implemented to flow larger gradients to previous layers by learning an identity function as a shortcut, as depicted in Figure 4.2. Finally, the second module M_{att} is provided with another 1×1 convolution layer to obtain a feature volume map $F = \{f_1, f_2, ..., f_k, ..., f_C\}$. Regarding the top model, a spatial squeeze was performed by a Global Average Pooling (GAP) layer, which provides a vector $x \in \mathbb{R}^{1 \times 1 \times C}$ according to Eq. 4.1. Finally, we defined a softmax-activated dense layer with two neurons corresponding to the two classes (healthy and glaucoma) in which the OCT images had to be classified.

$$x_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_k(i,j). \tag{4.1}$$

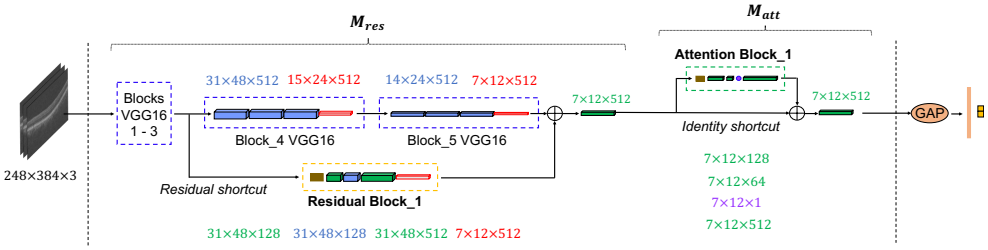


Figure 4.2: Architecture of the proposed slide-level discriminator used to distinguish between healthy and glaucomatous eyes from raw circumpapillary OCT images. Blue, red, brown and green colours denote 3×3 convolution, max-pooling, batch normalisation and 1×1 convolution layers, respectively.

As summarised in Figure 4.3, given an input image $I \in \mathbb{R}^{H' \times W' \times C'}$, being $H' \times W' \times C' = 248 \times 384 \times 3$ the dimensions of I, the first module M_{res} of the feature extractor generates a volume map $G \in \mathbb{R}^{H \times W \times C}$, $M_{res} : I \to G$. From here, G becomes the input to the second module M_{att}, which provides a refined output $F \in \mathbb{R}^{H \times W \times C}$, $M_{att} : G \to F$ that corresponds to the feature volume embedded in the latent space, which will be used to extract information from each B-scan of the SD-OCT volumes.

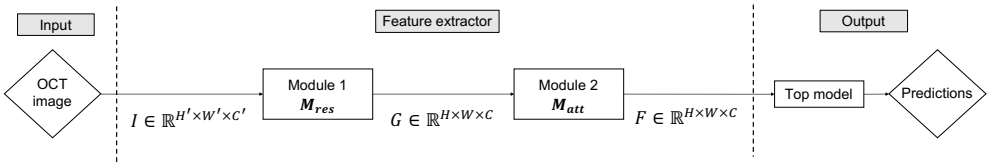


Figure 4.3: Flowchart of the proposed circumpapillary architecture highlighting the connections between the different modules of the feature extractor.

4.3.2 Volume-based predictive model development

4.3.2.1 Data-volume conditioning

As reported by Maetschke et al. [50], it was necessary to prepare the SD-OCT volumes database to face the constraints of the GPU memory caused by a large amount of data. In this paper, we propose a conditioning step for the volume slides based on extracting the useful information from each B-scan, instead of down-sampling the images [50, 53]. First, we discarded the 32 initial and the 32 final slides from the total of 128 because the glaucoma-specific information is located around the ONH, i.e. around the central slides of the cube [52].

Then, we developed a series of algorithms to remove useless pixels from each slide ensuring the same dimensions for all the slides. In this way, given $S = \{s_1, s_2, s_3, ..., s_P\}$ and $V = \{v_1, v_2, v_3, ..., v_Q\}$, where S and V are sets of slides and volumes composed of P and Q instances, respectively, the algorithm is able to reduce the dimensions $M \times N$ of each slide $s_i \in v_j$, with $i = \{1, 2, 3, ..., P\}$ and $j = \{1, 2, 3, ..., Q\}$, to dimensions $m \times N$ (Algorithm 3). To calculate m, it was first necessary to extract the dimensions of each specific bounding box $B_i \subset s_i$ corresponding to the region of the slide s_i that maximises the target retina area. Specifically, B_i was obtained by applying the *ROIret* function, which consists of a succession of morphological operations followed by the Otsu's binarisation method, as detailed in Algorithm 3.

Algorithm 3 *ROIret* function to extract the retina region B_i.

Data: Specific slide $s_i \in v_j$.
Functions:
Otsu, to find the optimal threshold.
Rectangle, to extract the region from a mask.
Result: Bounding box $B_i \subset s_i \in v_j$.

Initialisation:
$E_O \leftarrow 1$; *%Disk structuring element for opening*
$E_D \leftarrow [10, 20]$; *%Rectangular structuring element for dilation*
$E_C \leftarrow 10$; *%Disk structuring element for closing*

Bounding Box extraction:
$I_{open} \leftarrow (s_i \ominus E_O) \oplus E_O$;
$I_{dil} \leftarrow I_{open} \oplus E_D$;
$I_{close} \leftarrow (I_{dil} \oplus E_C) \ominus E_C$;
$th \leftarrow Otsu(I_{close})$;
$Mask \leftarrow I_{close} \geq th$;
$B_i \leftarrow Rectangle(Mask)$;

Once all $B_{i,j}$ were achieved, m was defined by the dimensions of the largest $B_{i,j}$, according to Algorithm 4. Then, we extracted a new set of slides $I = \{I_1, I_2, ..., I_i, ...I_P\}$, where $I_i \subset s_i \in v_j$ corresponds to the region $m \times N$ centred on the computed B_i, as detailed in Algorithm 4. After the conditioning phase, each OCT volume was defined by $P = 64$ B-scans of dimensions $m \times N = 550 \times 512$. However, to adapt the input dimensionality to the trained circumpapillary feature extractor, we resized each new slide $I_i \in v_i$ to 248×384 pixels, as illustrated in Figure 4.4.

Algorithm 4 Data-volume conditioning to remove useless pixels from the slides of the SD-OCT volumes.

Data: Slides $S \in V$ with dimensions $M \times N$.
Functions:
Centroid, to extract the centroid of the $B \subset s_i \in v_j$.
ROIret, to extract the bounding box B_i from each slide $s_i \in v_j$ (Algorithm 3).
Result: New slides $I \subset S \in V$ with dimensions $m \times N$.

Data conditioning:
for $j \leftarrow 1$ **to Q do**
  for $i \leftarrow 1$ **to P do**
    $B_{i,j} \leftarrow ROIret$ from $s_i \in v_j$;
    $x, y \leftarrow Centroid$ from $B \subset s_i \in v_j$;
    $c_{i,j} \leftarrow x$;
    $d_{i,j} \leftarrow |B(1,1) - B(end,1)|$;
$m \leftarrow MAX(d)$
for $j \leftarrow 1$ **to Q do**
  for $i \leftarrow 1$ **to P do**
    $I_{i,j} \leftarrow s_i(c_{i,j} - \frac{m}{2}$ to $c_{i,j} + \frac{m}{2}, 1$ to $N) \in v_j$;

4.3.2.2 LSTM network construction

In this stage, we propose the use of LSTM networks to feed the model with the spatial dependencies of the latent features extracted from each new slide $I_i \in v_j$. LSTM is a kind of Recurrent Neural Network (RNN) for sequence modelling, widely used in handwriting recognition [182], speech recognition [183] and video classification [184], among other tasks. Unlike traditional RNNs, LSTM networks contain a memory cell c_t able to accumulate the state information to avoid the long-term dependency problem. A common LSTM unit is composed of a series of *gates* that control the flow of information around the cell. An *input gate* i_t regulates the new information that enters the cell to be accumulated. The activation of a *forget gate* f_t determines whether the

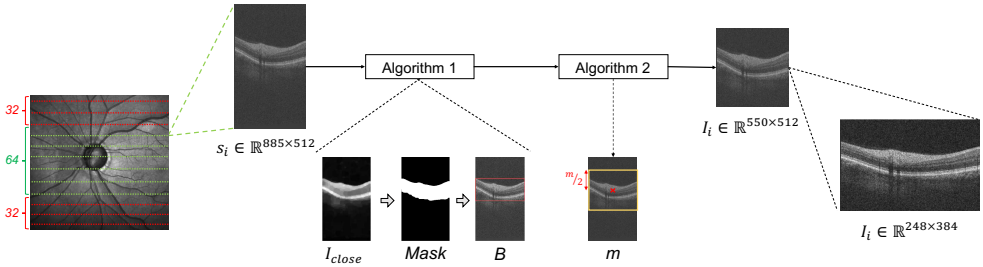


Figure 4.4: Data volume conditioning to adapt the B-scans of the volumes to the slide-level discriminator dimensions. This new set of volumes of dimensions $P \times m \times N$ will constitute the input to the feature extraction stage.

past cell status c_{t-1} is forgotten or not. Finally, an *output gate* o_t controls the propagation of the latest cell output c_t to the final state h_t, being t each temporary instance.

In this paper, we follow the CNN-LSTM strategy carried out in [184] to consider sequences of CNN activations. Unlike the aforementioned video-based study, the proposed approach is intended to OCT-volume classification, so that we consider each slide as a frame, i.e. each spatial dependency as a temporary instance. Once the slide-level discriminator and data-volume conditioning stages were performed in the previous sections, we used the pre-trained base model to extract the feature space from each conditioned volume slide. As illustrated in Figure 4.3, given an input image $I_i \in v_j$, an embedding f_i was obtained after the feature extraction phase. Then, 1D array a_i was generated from each feature volume f_i by flattening their dimensions $7 \times 12 \times 512$, which correspond to the output of the last 1×1 convolution layer of the slide-level discriminator (see Figure 4.5). According to this, the input to the LSTM network consists of an array $A = \{a_1, a_2, ..., a_i, ..., a_P\}$, being $P = 64$ the number of slides per volume. The output of each LSTM memory cell h_i corresponds to the concatenation of all the outputs h_{i_u} obtained from each LSTM unit u, so that $h_i = [h_{i_1}, h_{i_2}, ..., h_{i_u}, ..., h_{i_U}]$, where U is the number of specified LSTM units, as deduced from Figure 4.5. Note that each h_i constitutes the input (together with the a_{i+1}) to the next LSTM memory cell c_t, which is graphically represented by the discontinue lines in Figure 4.5. Traditional LSTM networks can return, per volume v_j, either a set of all spatial dependencies $H_j = \{h_1, h_2, ..., h_i, ..., h_P\}$ or just the last one h_P, which contains information from all the previous B-scans. As a novelty, we developed in this case a sequential-weighting module (SWM) to take into account all LSTM

outputs by weighting them, in a sequential way, to provide a holistic feature vector O_j before the top model (Algorithm 5). Then, a flatten operation was applied to concatenate the LSTM outputs $H_j = \{h_1, h_2, ..., h_i, ..., h_P\}$ into an array R_j of length $T = P * U$. At this point, a weighting layer $W = \{\frac{1}{T}, \frac{2}{T}, ..., \frac{k}{T}, ..., 1\}$, with $k = \{1, 2, 3, ..., T\}$, was included to generate a vector $L_j = R_j \circ W$, so that L_j became a weighting output from the initial LSTM output vector H_j. Additionally, as observed in Figure 4.5, a skip-connection module was defined to map F_j into a squeezed array Z_j via 3D global average-pooling layer (3DGAP). Finally, a holistic feature vector O_j per volume was obtained by concatenating Z_j and L_j outputs (see Figure 4.5). Regarding the top model structure, a classifier based on a Multi-Layer Perceptron (MLP) was implemented to achieve the probability p_j corresponding to the class predicted from each specific volume v_j (Algorithm 5). Aspects related to the hyper-parameters of the architecture and the top model are detailed in Subsection 4.4.2.2.

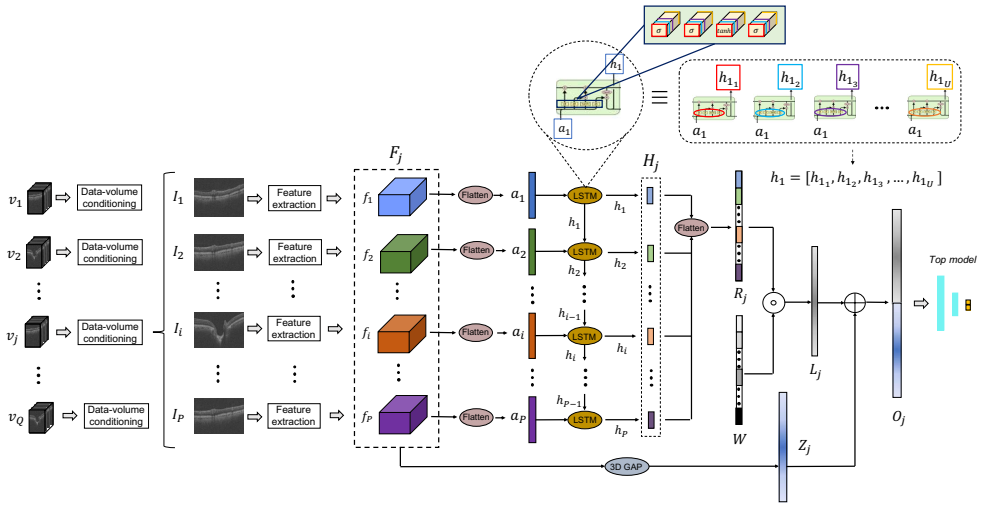


Figure 4.5: Architecture of the proposed volume-based predictive model to address the glaucoma detection just from raw SD-OCT volumes centred on the ONH.

Algorithm 5 Pipeline of the end-to-end volume-based methodology to predict glaucoma per volume.

Data: P slides per Q raw OCT volumes centred on ONH.
Functions:
$DC \equiv$ Data Conditioning stage.
$FE \equiv$ Feature Extraction phase
$MLP \equiv$ Multi-Layer Perceptron classifier
Result: Predictions p_j from each SD-OCT volume v_j.

End-to-end LSTM approach:
for $j \leftarrow 1$ **to Q do**
  for $i \leftarrow 1$ **to P do**
    $I_i \leftarrow DC(s_i)$, where $s_i \in v_j$;
    $f_i \leftarrow FE(I_i)$, where $I_i \subset s_i \in v_j$;
    $a_i \leftarrow Flatten(f_i)$;
    $h_i \leftarrow LSTM(a_i)$;
  Sequential-Weighting Module (SWM):
  $R_j \leftarrow Flatten(H_j)$;
  $L_j \leftarrow R_j \circ W$;
  $Z_j \leftarrow 3DGAP(F_j)$;
  $O_j \leftarrow Concatenate(L_j, Z_j)$;
  Top model:
  $p_j \leftarrow MLP(O_j)$;

4.4 Results

4.4.1 Slide-level feature extractor

4.4.1.1 Data partitioning

To provide robust and reliable results about the feature extractor, we performed a patient-level data partitioning of the *circ-DB-1* database. In particular, 52 circumpapillary images (20 glaucomatous and 32 healthy) were grouped into an independent set to test the model. With the rest of the data (training set), we conducted an internal 5-fold cross-validation (ICV) stage to optimise the hyper-parameters of the neural networks. Specifically, in each iteration, $\frac{4}{5}$ of the training set (58 glaucomatous and 99 healthy eyes) were employed to train a specific model and $\frac{1}{5}$ (15 with glaucoma and 25 normal images) as a validation subset to monitor and prevent overfitting. Subsequently, we used the entire training dataset (197 circumpapillary samples) to train the final model with the architecture and

parameters that reported the best performance during the ICV stage. The final model was validated using the test set and evaluated with the external *circ-DB-2* database, which is composed of 143 glaucomatous and 193 healthy circumpapillary OCT images.

4.4.1.2 Validation phase

In our previous work [62], we carried out a rigorous comparison between different neural networks for glaucoma detection just from raw circumpapillary OCT images. As VGG family of networks reported the best performance, in this paper, we perform a comparison between them and the proposed *Residual Attention Glaucoma Network* (RAGNet).

Ablation experiments. In the exploration of the optimal hyper-parameters for the feature extractor, we contemplated different training configurations. In Table 4.3, we show a summary of the main hyper-parameters considered to build the slide-level discriminator, as well as the range of values used during the exploration. Additionally, in Table 4.4, we expose the configuration selected to carry out each approach under study. The abbreviations set out in the Table 4.3 correspond as follows: SGD - Stochastic Gradient Descent; MSE - Mean Squared Error; WBCE - Weighted Binary Cross Entropy; MLP - Multi-Layer Perceptron; GMP - Global Max Pooling; GAP - Global Average Pooling.

Table 4.3: Possible training configurations for the slide-level discriminator.

Hyper-params	Range	Top model	Range	Fine-tuning	Range
Learning rate	$[5e^{-1}, 1e^{-5}]$	*Initial dropout*	[0, 0.5]	*Unfrozen VGG blocks*	1, 2, 3, 4, 5
Optimizer	SGD Adadelta Adam	*Structure*	Flatten + MLP GMP + Dense GAP + Dense	-	-
Loss function	MSE WBCE Hinge	*Final Activation*	Softmax Sigmoid	-	-
Batch size	8, 16, 32, 64	-	-	-	-
Number of epochs	[50, 500]	-	-	-	-

Based on the experiments accomplished in [62], we performed a deep fine-tuning strategy of the VGG16 and VGG19 architectures. For both CNNs, data augmentation techniques [112] were implemented to face the overfitting problem by increasing the number of images of the database with synthetic samples. A factor ratio of 0.2 was applied to perform random geometric and

Table 4.4: Selected configuration for the slide-level discriminator after the empirical exploration of each feature extractor approach.

		VGG16	VGG19	RAGnet (with VGG16)	RAGnet (with VGG19)
Model hyperparameters	*Learning rate*	0.001	0.001	0.0005	0.0005
	Optimizer	Adadelta	Adadelta	SGD	SGD
	Loss function	WBCE	WBCE	WBCE	WBCE
	Batch size	16	16	16	16
	Number of epochs	125	125	120	120
Top Model	*Initial dropout*	0.4	0.4	0.4	0.4
	Structure	GAP	GAP	GAP	GAP
	Final activation	Softmax	Softmax	Softmax	Softmax
Fine-tuning	*Unfrozen blocks*	3	3	3	3

dense elastic transformations from the original images, according to [62]. As observed in Table 4.4, we unfroze the two last convolutional blocks of the VGGs to retrain the weights with the specific information contained in the circumpapillary OCT images. Additionally, a weighted binary cross-entropy (WBCE) was used as a loss function by employing an optimal balanced factor $\alpha = [1.35, 0.79]$, following the Eq. 4.2 and 4.3, to alleviate the unbalanced problem between glaucoma and healthy classes, respectively (see Table 4.4). The models reached the best performance when they were trained during 125 epochs, trying to minimise the WBCE loss function, using Adadelta optimiser with a learning rate of 0.001 and a batch size of 16. It is noticeable that we added an initial dropout layer with a coefficient of 0.4 before the top model, as specified in Table 4.4.

$$WBCE = -\alpha_2 \; y_i \; log(\hat{y}_i) - \alpha_1(1 - y_i)log(1 - \hat{y}_i) \; , \quad being \tag{4.2}$$

$$\alpha_c = \frac{Ns}{Nc \sum y_i^c} \; , \qquad with \; c \in [1, 2] \tag{4.3}$$

where $\hat{y}$ and y represent the outputs and the ground truth, respectively. Ns denotes the total number of samples and $Nc = 2$ the number of classes, being $c = 1$ and $c = 2$ the glaucoma and healthy classes.

Regarding the proposed model (RAGNet), we performed the same deep fine-tuning strategy as before. Besides, the same data augmentation processing and weighted loss function were specified to make an objective comparison.

The major innovation of the proposed method lies in the inclusion of the residual M_{res} and attention M_{att} modules, whose filters and dimensions parameters were illustrated in Figure 4.2. The combination of the best hyper-parameters was carried out following the same empirical exploration as before (see Table 4.3). In this case, we trained the models (using VGG16 and VGG19 architectures as a baseline) during 120 epochs, instead of 125 like in the VGGs case. WBCE (Eq. 4.2) was selected as the loss function to be minimised, using SGD optimiser with a learning rate of 510^{-4} and a batch size of 16 (see Table 4.4). As before, we also included a dropout layer of 0.4 to address the classification stage. Concerning the top model, we made use of the same structure for all approaches, which is described in Table 4.4. Particularly, we included a Global Average Pooling (GAP) layer to obtain a spatial squeeze before the softmax-activated dense layer, which is composed of two neurons referring to healthy and glaucoma classes.

Quantitative results. Different figures of merit, such as sensitivity (SN), specificity (SP), F-score (FS) and area under the ROC curve (AUC) are reported in Table 4.5 to provide objective results. Note that AUC metric was calculated by means of a polynomial approximation using a gradient descent method, according to [185]. As a novelty, we also calculated quantitative metrics related to the learning curves achieved during the training of the models (see Table 4.6). In particular, we consider the validation accuracy (ACC) and loss (LOSS) values and propose two additional measures to quantify the overfitting (OVFT) and the quality (QLTY) of the validation learning curves. According to Eq. 4.4, OVFT indicator aims to provide measurable information related to the generalisation ability of the models, so that the closer OVFT is to 0 the better. OVFT < 0 indicates underfitting, whereas OVFT > 0 denotes overfitting, proportionally.

$$OVFT = \frac{1}{\epsilon - p} \sum_{e=p}^{\epsilon} (VL_e - TL_e), \quad OVFT \in (-\infty, \infty) \tag{4.4}$$

VL_e and TL_e correspond to the validation and training loss curves at the epoch e, respectively. ϵ denotes the total number of epochs and p the position (epoch) in which the validation loss curve VL reaches the global minimum.

QLTY metric provides information about how stable is the model and how much does it learn. Therefore, we measured the stability (STBL) of the model taking into account the variations between adjacent epochs of the validation loss curve. Specially, STBL was calculated according to Eq. 4.5, where $D = \{d_1, d_2, ..., d_e, ..., d_{\epsilon-1}\}$ is a vector containing the differences between

adjacent epochs. Each $d_e = VL_e - VL_{e+1}$ denotes the difference between the values of the validation loss curve achieved for the epochs e and $e+1$. The closer STBL is to 0, the more stable is the validation curve. Regarding the amount learned by the model (LRNG), it was calculated as the difference between the validation loss values at the initial and final epochs, according to Eq. 4.6. In this case, negative values correspond to a significant overfitting, whereas higher values denote greater learning. Finally, QLTY metric was achieved following the Eq. 4.7, where LRNG and STBL measures were combined to evaluate the quality of the model training. In this way, lower variations of validation loss between adjacent epochs and higher positive differences between initial and final epochs would entail a QLTY value closer to 1, which is associated with a more reliable model.

$$STBL = \sqrt{\frac{\sum_{e=1}^{\epsilon-1}(d_e - \bar{D})^2}{\epsilon - 1}}, \quad STBL \in [0, \infty) \tag{4.5}$$

$$LRNG = VL_1 - VL_{\epsilon-1}, \quad with \quad LRNG \in [-\infty, \infty) \tag{4.6}$$

$$QLTY = \begin{cases} \frac{1}{1+e^{-\frac{LRNG}{STBL}}} & \text{if } STBL > 0 \\ \frac{1}{1+e^{-\frac{LRNG}{10^{-\epsilon}}}} & \text{if } STBL = 0 \end{cases} \quad QLTY \in [0, 1] \tag{4.7}$$

Table 4.5: Cross-validation results reached from the *circ-DB-1* dataset. Metrics are reported in terms of mean ± standard deviation followed by a 95% confidence interval (CI).

	VGG16	VGG19	RAGNet-VGG16	RAGNet-VGG19
SN	0.81±0.11 (0.70-0.91)	0.77±0.17 (0.60-0.93)	**0.86±0.11 (0.75-0.97)**	0.73±0.12 (0.61-0.84)
SP	0.96±0.03 (0.93-0.99)	0.95±0.03 (0.92-0.98)	0.95±0.04 (0.91-0.99)	**0.97±0.02 (0.95-0.98)**
FS	0.86±0.08 (0.79-0.93)	0.82±0.11 (0.72-0.93)	**0.88±0.04 (0.85-0.92)**	0.81±0.08 (0.73-0.89)
AUC	0.88±0.05 (0.85-0.91)	0.86±0.08 (0.83-0.88)	**0.91±0.04 (0.87-0-94)**	0.85±0.06 (0.82-0.88)

Table 4.6: Learning curves behaviour from the *circ-DB-1* dataset. Metrics are reported in terms of average ± standard deviation followed by a 95% CI.

	VGG16	VGG19	RAGNet-VGG16	RAGNet-VGG19
ACC	0.90±0.05 (0.86-0.95)	0.88±0.06 (0.82-0.94)	**0.92±0.02 (0.90-0.94)**	0.89±0.05 (0.83-0.92)
LOSS	0.27±0.11 (0.16-0.37)	0.34±0.18 (0.17-0.5)	**0.25±0.09 (0.16-0.35)**	0.38±0.18 (0.2-0.54)
OVFT	**0.11±0.10 (0.02-0.20)**	0.16±0.14 (0.02-0.29)	0.12±0.08 (0.04-0.20)	0.17±0.10 (0.08-0.26)
STBL	**0.01±0.01 (0.01-0.02)**	0.02±0.01 (0.02-0.03)	0.07±0.03 (0.05-0.10)	0.08±0.01 (0.07-0.08)
QLTY	**1±3e^{-13} (0.99-1)**	0.99±0.01 (0.98-1)	1±2e^{-3} (0.99-1)	0.89±0.21 (0.69-1)

4.4.1.3 Prediction stage

In this section, we detail the results for the prediction of the primary and external test sets. Specifically, the results achieved when evaluating the primary test set from the *circ-DB-1* database are reported in Table 4.7. The results corresponding to the external validation are exposed in Table 4.8. The goal of this section is to demonstrate that the proposed slide-level discriminator could be valid to perform the feature extraction from the B-scans of the SD-OCT volumes. Therefore, we made use of the *circ-DB-2* database as an external test set to check how did the proposed feature extractor work with new OCT samples centred on the ONH of the retina.

Table 4.7: Test results achieved on *circ-DB-1* dataset.

	VGG16	VGG19	RAGNet-VGG16	RAGNet-VGG19
SN	0.8500	0.8500	0.8500	**0.9000**
SP	0.8750	0.8438	**0.9375**	0.8750
FS	0.8293	0.8095	**0.8718**	0.8571
AUC	0.8625	0.8469	**0.8938**	0.8875
ACC	0.8654	0.8462	**0.9038**	0.8846

Table 4.8: External test results achieved on *circ-DB-2* dataset

	VGG16	VGG19	RAGNet-VGG16	RAGNet-VGG19
SN	0.8741	**0.8951**	**0.8951**	0.8741
SP	0.8446	0.8238	**0.8756**	0.8653
FS	0.8389	0.8393	**0.8678**	0.8503
AUC	0.8593	0.8595	**0.8789**	0.8697
ACC	0.8571	0.8542	**0.8839**	0.8690

4.4.2 Volume-based predictive model

4.4.2.1 Data partitioning

As before, we also carried out a data partitioning stage of the *vol-DB-3* database to guarantee the rigour of the experiments performed with the SD-OCT volumes. Particularly, $\frac{1}{5}$ of the data (29 glaucomatous and 35 healthy) was used as a test set to evaluate the proposed model. The rest of the database was used to train the model through a 5-fold cross-validation technique. In each of the five iterations, $\frac{4}{5}$ of the training data (92 glaucomatous and 113 healthy eyes) were employed to develop a specific model, whereas $\frac{1}{5}$ (23 with glaucoma and 28 normal volumes) was used as a validation set to control overfitting

and optimise the hyper-parameters. The final model was built via training the best architecture (optimised during the ICV stage) with the samples from both training and validation sets.

4.4.2.2 Validation phase

In this section, we detail and compare the structure and the hyper-parameters that compose the proposed architecture in relation to the state of the art. Specifically, we show the differences between the developed method and other basic LSTM structures to evidence the added value that the proposed SWM-based structure introduces for glaucoma detection using SD-OCT volumes. Note that RAGNet architecture (via fine-tuning the VGG16 network) was selected as the slide-level discriminator to extract the features used as an input to the LSTM networks.

Ablation experiments. To perform a comparison as reliable as possible between different approaches, we established some fixed conditions by means of an initial random exploration of several hyper-parameters. Firstly, we fixed the input and output dimensions, as detailed in Table 4.9, to elucidate which method extracted the most discriminatory features. Regarding the model hyper-parameters, a nested loop was sweeping different loss functions, gradient-based learning algorithms and sizes of batches to select the most promising. In particular, we found WBCE loss function, Adadelta optimiser and batch size of 16 the best hyper-parameter combination to address the next phase. In this case, we calculated an optimal weighting factor $\alpha = [1.11, 0.91]$ to balance the glaucomatous and healthy samples, respectively (see Table 4.9). In addition, the number of training epochs and the learning rate were specified depending on the approach. Concerning the LSTM architecture, we fixed specific hyper-parameters, such as an input dropout of 0.3 to prevent overfitting, whereas the number of LSTM units was adapted to each approximation to provide a holistic feature map of size $T = 512$ before the classification stage. We also specified a constant top-model structure defined by an MLP algorithm composed of two fully-connected layers with 256 and 32 units, followed each one by a dropout layer with a coefficient of 0.25. Finally, a softmax layer with two neurons (healthy and glaucoma) was included to achieve the predictions per volume, as collected in Table 4.9.

Once the aforementioned hyper-parameters were established, we compared several useful LSTM options such as the shape of the final LSTM outputs or the use of Bidirectional layers, as proposed in [166], where the authors also performed a CNN-LSTM strategy to identify biomarkers associated with

Table 4.9: Configuration of the volume-based predictive model..

Data shape	
Input shape	$Q \times P \times 248 \times 384 \times 3$
Output shape	$Q \times 512$ (before the top model)
Model hyper-parameters	
Loss function	WBCE
Weighting factor	α=[1.11, 0.91]
Optimiser	Adadelta
Batch size	16
Number of epochs	*Variable* $\rightarrow$ [50, 500]
Learning rate	*Variable* $\rightarrow$ $[5e^{-1}, 1e^{-5}]$
Architecture hyper-parameters	
Feature extractor	RAGNet (with VGG16)
Input dropout	0.3
LSTM units	*Variable* $\rightarrow$ 4, 8, 256, 512
Top model	
Dense units	1×(256, 32)
Dropout coefficients	2×(0.25)
Final activation	Softmax (2 neurons)

the AMD disease. Regarding the output shape, basic LSTM networks can provide 3D or 2D arrays depending on whether all LSTM outputs H_j are considered or just the last one h_P, which contains information from all the previous slides. Both approaches are analysed in this study, under the name of *OS3D* and *OS2D*, respectively (see Table 4.10 and 4.11). The use of bidirectional layers (Bi) was also contemplated for both previous approaches to determine the performance of this kind of layers. As evidenced in Table 4.10 and 4.11, bidirectional layers provide a slight outperforming, so that we carried out another experiment, with the best-reported conditions, based on stacked LSTM layers to consider more complex and deeper architectures. Note that for all the aforementioned experiments, the models were trained during 60 epochs with a learning rate of 0.01, trying to optimise the compromise between accuracy and overfitting metrics. The number of LSTM units defined to adapt the size of the feature embedding space to $T = 512$ were 8, 512, 4, 256 and 1×(256, 512) for *OS3D*, *OS2D*, *Bi+OS3D*, *Bi+OS2D* and *stacked Bi+OS2D* approaches, respectively. Note that *OS3D*-based models require an extra layer to flattening the features extracted from all the slides, unlike the *OS2D*-based methods which only output the features that come directly from the last slide. For this reason, the LSTM units related to *OS3D*-based models need to be lower than the associated with *OS2D*-based approaches. In contrast, when bidirectional layers are included, LSTM units must be halved

to keep the output dimensionality, as bidirectional layers generate a reverse copy of the input sequence. An additional experiment using the conventional VGG16 network as a feature extractor (*VGG16+SWM*) was carried out to find out how much the prediction results may differ depending on the slide-level discriminator.

In the proposed strategy based on the combination of RAGNet-VGG16, LSTM and SWM structures, we kept most of the hyper-parameters constant, but thanks to the developed SWM, it was possible to decrease the learning rate to perform a more stable training during more epochs, without reporting overfitting. Specifically, we trained the models during 150 epochs with a learning rate of 0.005. We also contemplated the use of bidirectional layers combined with the designed SWM-based model (see Table 4.10 and 4.11). Following the same criteria as before, we defined 4 and 8 LSTM units for the proposed method with and without bidirectional layers, respectively.

Quantitative results. We show the results provided by the different approaches during the ICV stage of the volume-based predictive model. The training performance and the behaviour of the validation learning curves are reported in Table 4.10 and 4.11, respectively.

Table 4.10: Cross-validation results reached from the *vol-DB-3* database. Metrics are reported in terms of average ± standard deviation followed by a 95% CI.

	SN	SP	FS	AUC
OS3D	0.68±0.04 (0.64-0.72)	0.77±0.09 (0.68-0.86)	0.70±0.065 (0.65-0.74)	0.73±0.05 (0.68-0.79)
OS2D	0.73±0.12 (0.60-0.86)	0.79±0.03 (0.77-0.82)	0.73±0.08 (0.64-0.82)	0.76±0.07 (0.70-0.81)
Bi+OS3D	0.72±0.05 (0.67-0.77)	0.75±0.08 (0.68-0.82)	0.71±0.06 (0.66-0.77)	0.74±0.05 (0.68-0.80)
Bi+OS2D	0.66±0.14 (0.51-0.81)	**0.85±0.11 (0.74-0.97)**	0.71±0.07 (0.63-0.79)	0.76±0.04 (0.72-0.80)
Stacked Bi+OS2D	0.75±0.13 (0.63-0.86)	0.71±0.20 (0.52-0.90)	0.72±0.06 (0.66-0.77)	0.73±0.06 (0.68-0.76)
VGG16+SWM	0.70±0.11 (0.59-0.80)	0.69±0.05 (0.64-0.73)	0.67±0.07 (0.60-0.73)	0.69±0.04 (0.65-0.73)
Proposed	**0.76±0.11 (0.64-0.87)**	0.79±0.08 (0.70-0.87)	**0.75±0.05 (0.70-0.80)**	**0.78±0.04 (0.72-0.82)**
Proposed + Bi	0.75±0.09 (0.66-0.83)	0.77±0.15 (0.63-0.91)	0.74±0.07 (0.68-0.80)	0.76±0.05 (0.71-0.80)

4.4.2.3 Prediction stage

This section consists of two parts. In the first one, we compare the results achieved on the *vol-DB-3* test set from the different approaches (see Table 4.12). In the second part, we show the computed CAMs, which highlight the ROIs obtained from the SD-OCT volumes. CAMs allow for the identification of the areas in which the proposed deep learning model pays attention to classify each SD-OCT volume as healthy or glaucomatous. As the

Table 4.11: Learning curves behaviour from the *vol-DB-3* dataset. Metrics are reported in terms of average ± standard deviation followed by a 95% CI.

	ACC	LOSS	OVFT	STBL	QLTY
OS3D	0.71±0.06 (0.65-0.78)	0.57±0.09 (0.48-0.67)	0.28±0.08 (0.19-0.37)	0.04±0.02 (0.02-0.07)	0.82±0.21 (0.61-1)
OS2D	0.75±0.07 (0.68-0.82)	0.54±0.16 (0.38-0.72)	0.27±0.13 (0.14-0.40)	0.12±0.04 (0.08-0.16)	0.70±0.29 (0.42-0.98)
Bi+OS3D	0.74±0.05 (0.68-0.80)	0.53±0.08 (0.45-0.61)	0.21±0.08 (0.13-0.30)	0.05±0.01 (0.04-0.05)	0.91±0.16 (0.75-1)
Bi+OS2D	0.77±0.33 (0.73-0.80)	0.55±0.12 (0.42-0.67)	0.28±0.08 (0.19-0.37)	0.15±0.01 (0.14-0.16)	0.71±0.17 (0.55-0.88)
Stacked Bi+OS2D	0.73±0.07 (0.66-0.80)	0.59±0.18 (0.40-0.79)	0.20±0.10 (0.10-0.31)	0.10±0.02 (0.09-0.12)	0.66±0.35 (0.33-0.99)
VGG16+SWM	0.69±0.04 (0.65-0.73)	0.64±0.07 (0.57-0.71)	0.41±0.06 (0.35-0.47)	**0.01±0.01 (0.01-0.02)**	**0.99±0.01 (0.99-1)**
Proposed	**0.77±0.04 (0.73-0.81)**	**0.49±0.09 (0.39-0.59)**	**0.13±0.09 (0.04-0.23)**	0.02±0.01 (0.01-0.03)	0.99±0.03 (0.96-1)
Proposed+Bi	0.77±0.07 (0.69-0.84)	0.53±0.15 (0.37-0.69)	0.18±0.11 (0.11-0.26)	0.05±0.01 (0.04-0.07)	0.92±0.13 (0.77-1)

proposed SWM-based method outperforms the rest of approaches, we report the CAMs extracted from the test set making use of the developed volume-based predictive model (without bidirectional layers). The combination of CAMs and LSTM units enables the detection of the ROIs of each B-scan, as well as the key slides inside the volume. In Figure 4.6, we show several examples of the heat maps generated by the model from random SD-OCT volumes of the test set. In particular, we show a sweep of several heat maps of representative slides I_i corresponding to four randomly selected volumes from each class v_r^c, being $r \in [1, Q]$ a random integer and c the class. Note that the hot-coloured areas indicate greater discriminatory differences.

Table 4.12: Test results achieved on *vol-DB-3* dataset.

	SN	SP	FS	AUC	ACC
OS3D	0.7586	0.7982	0.7547	0.7793	0.7813
OS2D	0.6897	0.6857	0.6667	0.6877	0.6875
Bi+OS3D	0.7586	0.8286	0.7719	0.7936	0.7969
Bi+OS2D	0.7586	0.7714	0.7458	0.7650	0.7656
Stacked Bi+OS2D	0.72414	0.8286	0.7500	0.7764	0.7813
VGG16+SWM	**0.8276**	0.6857	0.7500	0.7567	0.7500
Proposed	0.7586	**0.8571**	**0.7857**	**0.8079**	**0.8125**
Proposed+Bi	0.7586	0.7714	0.7458	0.7650	0.7656

In Table 4.13, we compare the proposed method with other state-of-the-art studies. It is important to highlight that, to the best of the authors' knowledge, there are no public databases of ONH-centred SD-OCT volumes to make possible a direct comparison. For that reason, we have faithfully replicated the experiments carried out in the SD-OCT volume-based works of the literature intended to glaucoma detection. As we mentioned in Subsection 4.1.1, two

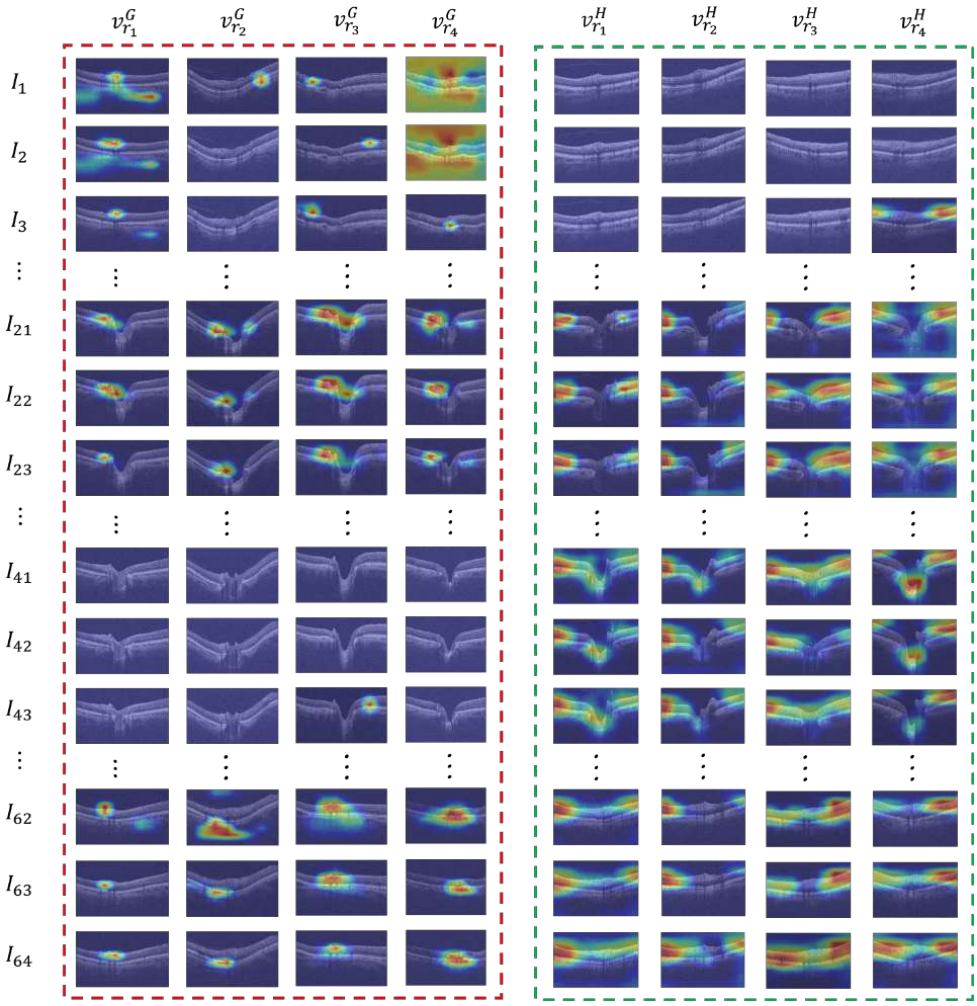


Figure 4.6: Heat maps from the proposed volume-based predictive model to highlight the most important regions of each B-scan when classifying specific random volumes as glaucomatous or healthy samples. Four first columns of volumes $v^G_{r_1}, ..., v^G_{r_4}$ bounded by the red rectangle correspond to glaucomatous eyes, whereas the four last columns of volumes $v^H_{r_1}, ..., v^H_{r_4}$ marked with the green rectangle correspond to healthy samples.

recent studies [50, 76] could be directly compared with ours, as they focused just on raw SD-OCT volumes to address the glaucoma detection, without including other variables extracted from the patient, such as visual field, intraocular pressure, mean deviation or fundus images, among others.

The state-of-the-art experiments were replicated making use of the same deep learning architecture and maintaining the original hyper-parameters during the training of the models. However, some specific conditions were adapted in each case to make possible the comparison between methods. In particular, the batch size and loss function were not reported in [50], so that we defined them according to our work, i.e. a batch size of 16 and the WBCE loss function. In contrast, to replicate the network performed in [76], we used 4 samples per batch aimed at facing the GPU memory problems associated with such a deep neural network. We conserved the original architectures to provide a comparison as objective as possible. In Table 4.13, we report the test results by comparing the proposed methodology with the different state-of-the-art approaches intended to glaucoma detection just from raw SD-OCT volumes. Note that all the experiments were performed on an Intel i7 @4.00 GHz of 16 GB of RAM with a Titan V GPU of 12 GB of RAM.

Table 4.13: State-of-the-art comparison on *vol-DB-3* test set

	SN	**SP**	**FS**	**AUC**	**ACC**
Maetschke et al.	0.6207	0.8000	0.6667	0.7103	0.7188
Ran et al.	0.6348	0.7286	0.5858	0.6817	0.6873
Proposed method	**0.7586**	**0.8571**	**0.7857**	**0.8079**	**0.8125**

4.5 Discussion

4.5.1 On the slide-level feature extractor

Contrarily to the state-of-the-art studies, which performed the glaucoma detection from SD-OCT volumes through 3D architectures, in this paper, we propose a new way of addressing this task by using the spatial dependencies between 2D images, instead of operating in the three-dimensional space. To do this, We have developed a new slide-level discriminator able to extract the features from the slides of the SD-OCT volumes. The proposed RAGNet model was compared with other validated architectures, which reported the best performance in our previous work [62] for glaucoma detection using circumpapillary OCT samples. In Table 4.5, different figures of merit extracted from the 5-fold cross-validated iterations are exposed to compare the different approaches. The proposed RAGNet method, characterised by the combination of residual connections and convolutional attention blocks, reported a significant outperforming with respect to the rest of networks. Specifically, the RAGNet model (with the fine-tuned VGG16) achieved the best

results for the most of metrics, except for SP measure, whose highest value was reached by the RAGNet model (with the fine-tuned VGG19). However, taking into account that this approach showed a significantly worse performance for the rest of metrics, and the RAGNet model (with the fine-tuned VGG16) provided a SP value closely similar to the best approach, the proposed network could be considered superior. In any case, the inclusion of the proposed residual and attention modules outperforms the popular pre-trained architectures of the state of the art.

As a novelty, besides the accuracy and loss values, we also introduced new metrics, such as OVFT, STBL and QLTY related to the learning of the models, as observed in Table 4.6. These additional measures provide information about the generalisation ability of the models to predict new samples. In this case, RAGNet (with VGG16) was consolidated as the best network, as it reported higher values for accuracy and loss metrics, besides the aforementioned. Additionally, for OVFT, STBL and QLTY measures the proposed model achieves closely similar results in relation with the reached by the best architecture (pre-trained VGG16). The small reported differences (0.01, 0.06 and 0.002 for OVFT, STBL and QLTY indicators, respectively) are negligible in the model's performance, as all of them represented a stable learning, which can be deduced from Table 4.7 and 4.8.

To verify how the models would work with new OCT samples, we carried out a prediction stage in which we evaluated the models on the primary test set of the *circ-DB-1* database (see Table 4.7). In line with the results obtained during the ICV stage, RAGNet-based approaches also surpassed the results for all figures of merit. In particular, RAGNet (with VGG16) worked as a more specific model, whereas RAGNet (with VGG19) provided a more sensible behaviour. For the rest of metrics, RAGNet (with VGG16) stood out for FS and ACC measures, but it reported lower values estimating AUC. Nevertheless, both RAGNet-based methods showed excellent performance, with results around 0.9 for all figures of merit. The results of the experiments make evident that the proposed RAGNet approach based on residual and skip-connections improved the performance of the traditional networks previously validated in [62].

To check the generalisation ability of the proposed slide-level discriminator, we performed an external validation using the *circ-DB-2* database. Note that the final success of the volume-based predictive model largely depends on the feature extractor performance. For that reason, an in-depth validation is necessary to ensure that the proposed slide-level discriminator can predict independent OCT samples centred in the ONH of the retina. Results corresponding to this stage are detailed in Table 4.8, where the outperforming

of the proposed RAGNet (with VGG16) model is evident. Additionally, the results reached in the prediction of the primary and external test sets are closely similar, which suggests that the proposed feature extractor is perfectly applicable to other OCT databases. Furthermore, it has demonstrated to be robust to the acquisition machine, as deduced from the volume-level results reported in Table 4.12. After the validation process, we found the proposed RAGNet model (with VGG16) as the best feature extractor, surpassing the performance reported by the previously trained models in [62]. For this reason, we made use of the proposed slide-level discriminator to extract the latent features from the B-scans of the SD-OCT volumes.

4.5.2 On the volume-based predictive model

According to the state-of-the-art works [50, 53], a pre-processing stage is necessary to face the GPU memory problems due to the large amount of data contained in the SD-OCT volumes. However, unlike the aforementioned studies where each OCT scan of the database was a cube of resolution of $200 \times 200 \times 1024$, the proposed approach was addressed from volumes of $885 \times 512 \times 128$. Specifically, [50] and [53] applied a down-sampling step to obtain volumes of dimensions $64 \times 64 \times 128$ and $100 \times 100 \times 128$ voxels, respectively. In contrast, the proposed method is able to take better advantage of the useful information from each slide by focusing on the retina regions around the ONH, as detailed in Figure 4.4.

To the best of the authors' knowledge, we are the first that propose the use of LSTM networks to address glaucoma detection just from raw SD-OCT volumes. We introduce some novelties with respect to the basic LSTM networks. In particular, we propose a SWM structure which to refine the LSTM outputs to control the overfitting and improve the learning of the models via skip-connections. SWM makes possible that each LSTM output h_i directly contributes to the volume classification, but in a weighted way. To demonstrate the outperforming of the proposed SWM-based approach, we compared it with other basic LSTM structures, as observed in Table 4.10. As LSTM networks can output 3D or 2D arrays depending on the specified output shape, we analysed both options (OS3D and OS2D, respectively) and we also included bidirectional layers (Bi), according to the architecture used in [166]. As appreciated in Table 4.10, the use of bidirectional layers surpasses in both cases the results achieved with the same models without including these layers. For that reason, we based on the best model from the four previous experiments to build deeper LSTM networks via stacking two LSTM memory cells. The results reached by the proposed methods with and without bidirectional layers

are reported in Table 4.10. We can observe in that table that the *Bi+OS2D* model reports better specificity, but in exchange for compromising the rest of the metrics. However, the proposed model provides higher values for the SN, FS and AUC metrics. Additionally, when the traditional VGG16 architecture is used as a feature extractor, the performance of the model notably decreases for all figures of merit in contrast to the proposed method.

In Table 4.11, the superiority of the proposed model is accentuated as it achieves the best results related to the learning stage for most of the metrics. Note that, in line with the results reported during the evaluation of the feature extractors, when the traditional VGG16 architecture is used, the stability (STBL) and the quality (QLTY) of the models are better. However, in this case, the end-to-end system with VGG16 as a feature extractor reports a lot of overfitting (OVFT), which explains the poor performance of the model affecting to the rest of figures of merit. The proposed method also stands out for STBL and QLTY metrics, besides the OVFT, being the differences in this case remarkably enough to affect the future behaviour of the model when predicting new SD-OCT volumes. Specially, OVFT metric shows a notorious better performance, which allowed for the convergence of the model during more epochs with a lower learning rate. This resulted in a higher quality (QLTY) of the training model, as greater and more stable learning (higher LRNG and lower STBL values) was reached. In Table 4.11, we can also see that the proposed model reaches the best results for validation accuracy and loss measures.

Finally, we carried out a prediction stage to evaluate the models' performance making use of the test set. The SWM-based models stand out for all the metrics, as expected. According to Table 4.12, the VGG16+SWM reports higher sensitivity values, but in exchange of greatly compromising other indices. In contrast, the proposed method (using RAGNet-VGG16 as a feature extractor) surpasses the rest of the metrics, without unduly affecting sensitivity. Therefore, the results reported during both training and testing phases place the proposed model as the best system for glaucoma detection using SD-OCT volumes, taking into account that they are not being currently traced due to the large involved workload.

From the results achieved by the end-to-end system, we can conclude that the proposed feature extractor is not camera-specific because, although it was developed using circumpapillary OCT images, the latent features extracted from each cross-sectional slide of the volume lead to a high performance after including the LSTM and SWM elements. This evidences that the proposed discriminator is also robust to different types of acquisition cut-offs

(circumpapillary or linear), as it founds the relevant information around the ONH of the retina independently of the acquisition mode.

Unlike the rest of the state-of-the-art works, which reported CAMs from single slides of each volume, we show different representative slides for several random volumes to determine which B-scans become more relevant for glaucoma diagnosis (see Figure 4.6). In particular, the first slides of the volumes $(I_1, I_2, I_3, ...)$ do not seem to matter much when predicting the healthy class, in contrast to the glaucoma label, as the heat maps highlight specific areas around the RNFL. Central slides $(..., I_{21}, I_{22}, I_{23}, ...)$ are more interesting because, in the case of healthy volumes, the LSTM model pays attention to the left and right bounds of the retina areas, whereas central regions corresponding to the optic disc cupping seem more discriminative for the glaucoma class. Additionally, a prominent activation usually appears highlighting the neuroretinal rims for glaucomatous volumes, especially on the left part. More advanced central slides, i.e. $..., I_{41}, I_{42}, I_{43}, ...$, reports a clear discriminatory ability to determine the healthy class, but no obvious signs of glaucoma are evidenced by the proposed model. Concerning the last slides $(..., I_{62}, I_{63}$ and $I_{64})$, the heat maps also manifest differences depending on the class, as the proposed model tends to highlight the external areas of the retina for healthy slides, and the central zones for glaucomatous samples. In summary, the findings achieved by the CAMs are directly in line with the reported in the literature [50, 76], as the heat maps focus on the edges of the retinal layers in the normal volumes, whereas retinal structures such as RNFL, neuroretinal rims and lamina cribrosa are evident in the glaucomatous cases.

As no public SD-OCT databases are available, in an attempt to compare our method with other state-of-the-art studies, we replicated the experiments performed in [50, 76] to objectively contrast the differences (Table 4.13). The results show a clear outperforming of the proposed model with respect to the rest of the state-of-the-art methods. In Table 4.13, we can observe the superiority of the proposed RAGNet+SWM model in detecting glaucoma from SD-OCT volumes, where the differences between exceeds more than 10% for all the figures of merit. Note that the results reported by the other networks could be underperformed, as we trained them on our database, but they were originally intended to be trained on larger datasets. Leaving aside the limitations during the state-of-the-art comparison, it can be concluded that the proposed method, based on the combination of LSTM and SWM networks using the RAGNet-VGG16 as a baseline, is the best system to detect glaucoma from SD-OCT volumes, surpassing other methods focused on 3D architectures.

4.6 Conclusion

In this paper, we have proposed a predictive model based on a novel two-step learning methodology to detect glaucoma just from raw SD-OCT volumes. The proposed slide-level feature extractor (RAGNet with VGG16) has improved the discrimination between healthy and glaucomatous circumpapillary OCT B-scans thanks to the attention module included via residual connections. Regarding the volume-based predictive model, the proposed recurrent framework, focused on transferring the knowledge of the spatial dependencies across the slides of the SD-OCT cube, has been shown to outperform other state-of-the-art architectures focused on 3D deep learning architectures. Additionally, the proposed sequential-weighting module (SWM) has led to a better performance by refining the feature space extracted from the volume slides. The reported CAMs could suppose a promising tool for an easier 3D scan analysis, as the ophthalmologists could scroll the heat maps that highlight the areas of interest to determine the class of each sample. Taking into account that we did not include VF, IOP or other external tests to develop the predictive models, we can conclude that SD-OCT volumes could provide a great added value for glaucoma diagnosis and help ophthalmologists to face the workload associated with the analysis of the cross-sectional OCT images.

As future research lines, better results for SD-OCT volumes could be reported by training a slide-level discriminator focused on the specific knowledge of the SD-OCT B-scans, instead of the circumpapillary images. The proposed volume-based predictive model could be considered as a good starting point to build a reliable computer-aided diagnosis system. This would require a significant increase in the number of samples to train the predictive model, as well as labelled external data sets to test it.

Chapter 5

Circumpapillary OCT-Focused Hybrid Learning for Glaucoma Grading Using Tailored Prototypical Neural Networks

The content of this chapter corresponds to the author version of the following published paper: García, G., Del Amor, R., Colomer, A., Verdú-Monedero, R., Morales-Sánchez, J. & Naranjo, V. Circumpapillary OCT-focused hybrid learning for glaucoma grading using tailored prototypical neural networks. Artificial Intelligence in Medicine, 118, 102132 (2021).

Contents

Circumpapillary OCT-Focused Hybrid Learning for Glaucoma Grading Using Tailored Prototypical Neural Networks

Gabriel García[1], Rocío del Amor[1], Adrián Colomer[1], Rafael Verdú-Monedero[2], Juan Morales-Sánchez[2] and Valery Naranjo[1]

[1]Instituto de Investigación e Innovación en Bioingeniería, Universitat Politècnica de València, 46022, Valencia, Spain

[2]Departamento de Tecnologías de la Información y las Comunicaciones, Universidad Politécnica de Cartagena, 30202, Cartagena, Spain

Abstract

Glaucoma is one of the leading causes of blindness worldwide and Optical Coherence Tomography (OCT) is the quintessential imaging technique for its detection. Unlike most of the state-of-the-art studies focused on glaucoma detection, in this paper, we propose, for the first time, a novel framework for glaucoma grading using raw circumpapillary B-scans. In particular, we set out a new OCT-based hybrid network which combines hand-driven and deep learning algorithms. An OCT-specific descriptor is proposed to extract hand-crafted features related to the retinal nerve fibre layer (RNFL). In parallel, an innovative CNN is developed using skip-connections to include tailored residual and attention modules to refine the automatic features of the latent space. The proposed architecture is used as a backbone to conduct a novel few-shot learning based on static and dynamic prototypical networks. The $k-$shot paradigm is redefined giving rise to a supervised end-to-end system which provides substantial improvements discriminating between healthy, early and advanced glaucoma samples. The training and evaluation processes of the dynamic prototypical network are addressed from two fused databases acquired via Heidelberg Spectralis system. Validation and testing results reach a categorical accuracy of 0.9459 and 0.8788 for glaucoma grading, respectively. Besides, the high performance reported by the proposed model for glaucoma detection deserves a special mention. The findings from the class activation maps are directly in line with the clinicians' opinion since the heatmaps pointed out the RNFL as the most relevant structure for glaucoma diagnosis.

5.1 Introduction

Glaucoma is a chronic and progressive disease that affects the optic nerve head (ONH) of the retina causing several structural changes and functional damage [73]. Nowadays, this optic neuropathy has become the leading cause of blindness worldwide, according to [74]. Recent studies suggest that the impact of this disease will continue to rise, affecting 111.8 million people in 2040 [157]. Therefore, early diagnosis of glaucoma could be essential for timely treatment in order to prevent irreversible vision loss [74].

Currently, there is no single accurate test to certify glaucoma, the diagnostic procedure includes several time-consuming tests such as pachymetry, tonometry and visual field tests, as well as the examination of different kinds of interpretable retinal images. Specifically, techniques based on image analysis like fundus photography and optical coherence tomography (OCT) have become very important in the context of glaucoma detection. Fundus image is a great cost-effectiveness technique which has reported promising results in the diagnosis of several eye-focused diseases, e.g. diabetic retinopathy [14, 186] and age-related macular degeneration [15, 187]. However, OCT imaging modality [159] is the quintessential technique for glaucomatous damage evaluation [39], as it enables the quantification of glaucoma-specific regions such as retinal nerve fibre layer (RNFL) and ganglion cell inner plexiform layer (GCIPL), which are useful biomarkers for the progression of this disease [160]. Additionally, glaucoma is evident in the deterioration of the cell layers around the optic disc, whose information could be exploited by the OCT imaging modality, as it focuses on the depth axis of the retina to identify structural changes, unlike the 2D projection of fundus image (see Figure 5.1).

Note that, although fundus image modality is cheaper than OCT, it is colour-dependent on the training dataset and its interpretation remains subjective [163, 164]. In contrast, OCT is a non-contact and non-invasive technique that provides objective information about the ONH and RNFL structures [165]. Glaucoma detection entails a subjective examination from different experts, whose mismatch ratio is usually high [158]. Consequently, many state-of-the-art studies developed different machine learning algorithms intended to detect glaucoma via fundus image and OCT samples.

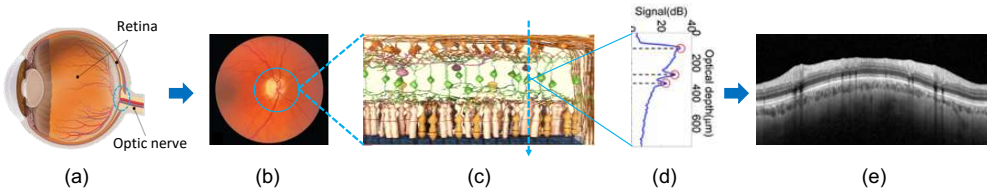


Figure 5.1: Illustration of the OCT images provenance. (a) Eyeball with interesting regions. (b) Fundus image focused on the optic nerve head (ONH). (c) Arrangement of the retinal fibre layers. (d) A-scan corresponding to the depth of retinal cells at a specific point. (e) B-scan representing the fibre layers of the retina with different grey-intensity levels.

5.1.1 Related work

According to [188], where a review of deep learning methods for glaucoma detection is described, spectral domain (SD) OCT has become the most widespread diagnostic tool for analysing glaucomatous pathologies. OCT system is routinely used in clinical practice for determining glaucoma severity, as it is able to emphasise the significance of the structural changes of the retina [189]. There are several clinical studies which claim the potential of the OCT imaging modality in the glaucoma detection paradigm. Medeiros et al. [190] enhanced the importance of the RNFL structure demonstrating that an early pathological degeneration of the retinal cells is associated with a thinning of the RNFL. Leung et al. [189] evidenced the usefulness of the RNFL thickness to determine pathological variations of the retina associated with different glaucomatous stages. Ojima et al. [191] also demonstrated that the RNFL thickness has a higher potential for glaucoma diagnosis than a complete macular volume. Moreover, in [192], the authors declared that RNFL features extracted from SD-OCT scans are a powerful indicator to detect glaucoma, which is directly in line with our previous works [63, 65].

Inspired by the aforementioned clinical studies, many researchers have developed predictive models to discern between normal and glaucomatous patients via OCT scans analysis. Most of them focused on the optic disc due to the known potential of the circumpapillary OCT images for glaucoma detection [167]. In [16, 168], authors extracted hand-crafted features from the B-scans and combined them with transformations of A-scans and visual field parameters to discriminate between healthy and glaucoma classes. They used different machine learning classifiers such as support vector machine (SVM), random forest (RF) and k-nearest neighbour (KNN). In [193], an automated framework to estimate the thickness of specific OCT layers was presented

in order to monitor retinal abnormalities, similarly to [194, 195], where the proposed methods were intended to quantify retinal layers thicknesses in search of abnormal retinal pathologies. Gao et al [196], made a comparison between the thickness measurements automatically reported by the Topcon OCT system and its end-to-end framework composed of a segmentation stage followed by a feature extraction based on the average of the RNFL thicknesses. This kind of hand-driven learning methodologies usually requires a previous manual or automatic segmentation of the retinal layers to delimit the regions from which extracting the discriminative features [188]. For that reason, several studies have been proposed in the literature for the sole purpose of addressing this task [171, 172, 197–202].

However, segmentation algorithms should entail errors which are transferred to the feature extraction stage [188]. To avoid this shortcoming, deep learning techniques arise to provide alternative ways of quantifying structural damage, as they can learn features from data automatically without reliance on previous segmentation stages or predefined features. In this context, ophthalmology has risen to a forefront in which application of deep learning (DL) algorithms can boost to a better artificial intelligence-based diagnosis in the medical fields. Specifically, glaucoma broadly meets the conditions to aid in the management of the vast amount of information coming from SD-OCT scans, according to the review outlined in [188]. Following the deep learning trend, Thompson et al. [203] demonstrated in a recent study that its segmentation-free DL algorithm (trained with raw SD-OCT B-scans) surpassed the conventional RNFL thickness parameters to directly discern between glaucomatous and healthy eyes. Maetschke et al. [50] also conducted a comparison between hand-driven and data-driven learning strategies applied on raw OCT volumes around the ONH of the retina for glaucoma detection. The main outcomes revealed that the deep learning approach outperformed conventional SD-OCT parameters to discriminate glaucomatous from normal samples. It should be noted that there are other studies which addressed the binary classification between healthy and glaucomatous cases by applying deep learning algorithms on SD-OCT volumes [52, 53, 75], including our previous work [65].

To the best of the authors' knowledge, deep learning is not very widespread on circumpapillary OCT images. A recent review of the literature found that most of the deep learning-focused studies using B-scans are conducted in combination with fundus image to detect glaucoma via RNFL probability maps [54, 170, 204]. To fill this gap in the literature, we previously proposed different hand-driven and automatic learning strategies for glaucoma diagnosis in [62, 63, 65], whose outcomes support the basis of this paper.

5.1.2 Contribution of this work

Inspired by the high performance reported from our previous circumpapillary-based studies, we have expanded the database of B-scans around the ONH to go deeper into the glaucoma paradigm. We propose in this paper an innovative framework based on prototypical networks for glaucoma grading using raw circumpapillary images. As far as we know, this work is the first OCT-focused study intended to grade the glaucoma severity by discerning between healthy, early and advanced classes, which adds significant value to the body of knowledge. Most of the previous studies centred on B-scans are designed to discriminate glaucoma from healthy samples [203, 205–209]. Other state-of-the-art studies also pursued the classification of healthy and glaucomatous cases, but using the OCT volumes as an input to their models [50, 52, 53, 65, 75]. Additional glaucoma-related studies were addressed from B-scans to accomplish different discrimination tasks such as pre-perimetric vs perimetric glaucoma [210, 211], progressing vs non-progressing glaucoma [212, 213] and close angle vs open-angle glaucoma [214, 215], among others. Furthermore, there are studies which used a different kind of input data, e.g. visual field tests [216, 217] to discern between healthy and glaucomatous patients. On a wider level, the sub-classification of early and advanced glaucoma has already been conducted in the literature, but throughout fundus image [218–220].

Noticeably, a very recent work [6] also proposes a kind of glaucoma grading set up making use of the Armed Forces Institute of Ophthalmology (AFIO) dataset [221], which contains OCT scans centred on the ONH. However, it pursues the discrimination between healthy, suspects and glaucomatous samples by computing the distance of different retinal layers of interest previously segmented. In particular, Raja et al. in [6] propose an encoder-decoder architecture to carry out the glaucoma classification. The encoder structure was used to provide a feature map capable of discerning between healthy and glaucoma classes via softmax function; whereas the decoder component was intended to segment the interesting layers of the retina to distinguish between suspect and glaucoma cases from the samples previously predicted as glaucoma. To accomplish this part, the mean of the segmented layers was used as a feature input of an SVM classifier. Unlike AFIO database, whose labels correspond to non-successive phases of the disease (healthy, suspect and glaucoma), our dataset includes different levels of glaucoma severity annotated according to the medical literature [222, 223]. Contrary to Raja et al. [6], we can conduct a learning framework that enables the analysis of glaucoma progression by differentiating between healthy, early and advanced glaucomatous samples. Another essential difference with respect

to [6] is that we develop the predictive models from raw gray-scale B-scans, whereas Raja et al. [6] made use of pre-processed RGB scans containing manual annotations and highlighted structures. Furthermore, the proposed work documents additional key contributions concerning the deep learning application in the glaucoma field. For the first time, we raise a glaucoma scenario based on prototypical neural networks (PNN) [224], which have demonstrated a high rate of performance in recent image analysis tasks, such as domain adaptation [225], noisy evaluation [226], text classification [227], etc. Note that PNN are usually formulated as a baseline within the few-shot paradigm [228–231], but in this paper, we exploit the prototypical concept in the k-shot methodology to optimise the learning process for glaucoma grading.

Tatham et al. [232] argued that circumpapillary RNFL (cpRNFL) thickness was the best structure to measure glaucoma progression and the most widely used parameter in clinical practice. According to this, and inspired by our previous works [62, 63], we outline in this paper a novel OCT-based hybrid backbone as a feature extractor of the prototypical framework. From [63], we observed that hand-crafted features could outperform the automatic features extracted by deep learning models trained from scratch. Nevertheless, from the study carried out in [62], we detected that fine-tuned models also improved the model's performance compared with algorithms trained from scratch. For that reason, in this approach, we propose a novel backbone composed of pre-trained deep learning networks (with additional attention modules and residual blocks) in a combination of hand-crafted RNFL-based features, similar to the hybrid methodology that reported the best performance in [63].

In summary, the main contributions of this work are:

- Raw circumpapillary OCT images are used for the first time to measure the glaucoma severity.
- Tailored prototype-based solutions are formulated in a novel framework for glaucoma grading.
- An adapted k-shot supervised learning, inspired by the few-shot paradigm, is conducted to exploit the specific-glaucoma knowledge.
- A new OCT-based hybrid backbone is proposed as a feature extractor to combine automatic and hand-crafted information from B-scans.

5.2 Methods

5.2.1 Backbone development

In this paper, we pay special attention to the base encoder development, as the performance of the prototypical framework largely depends on the representation vectors encoded in the latent space by the feature extractor. The original study of the PNNs [224], as well as others derived from it [225, 228], made use of a 4-layer CNN trained from scratch as an encoder of the feature representation. However, we observed from our previous glaucoma-based works [62, 63] that deep learning models trained from scratch reported the poorest performance in comparison to fine-tuning the models or even extracting hand-crafted features. For that reason, we built on our previous experience to propose in this work a new tailored backbone able to capture the OCT-specific cues for an optimal glaucoma grading. Specifically, we inspired on the OCT hybrid methodology followed in [63], but using pre-trained networks according to [62, 65], to provide a novel base encoder Ψ_ϕ with key novelties that lead to a better performance. The proposed backbone is composed of two learning branches giving rise to a multi-input feature model, as observed in Figure 5.2. Each independent module is detailed in the following subsections.

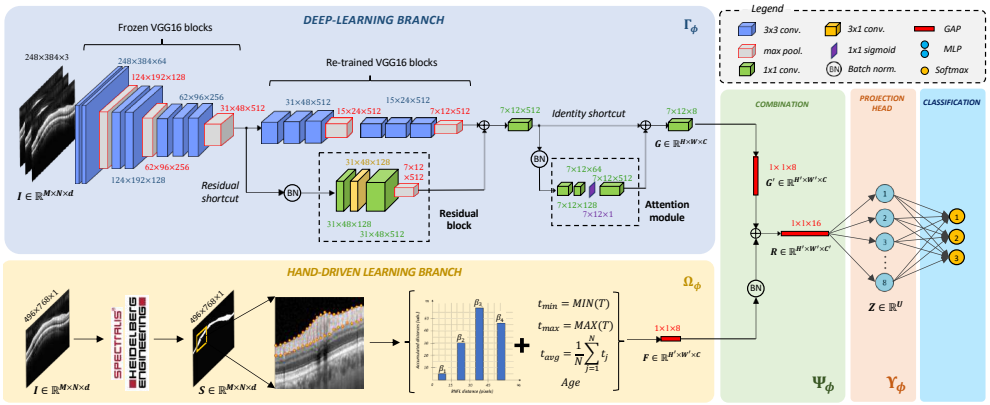


Figure 5.2: Illustration of the end-to-end backbone proposed as a benchmark to conduct the prototype-based learning strategies. Blue, yellow and green frames correspond to the base encoder network which consists of deep learning and hand-driven learning branches followed by a combination module, respectively. A projection head (red) maps the feature embedding in a lower-dimensional space to optimise the classification stage (cyan).

5.2.1.1 Deep learning branch

From the sweep of the pre-trained networks carried out in [62], we selected the VGG16 architecture as the baseline of our deep learning branch Γ_ϕ. In particular, we applied a deep fine-tuning strategy [111] to freeze the weights of the three first convolutional blocks, which were pre-trained with around 14 million of natural images corresponding to the *ImageNet* dataset. As a novelty, we propose a glaucoma-specific residual structure which allows the information from the initial VGG layers to be propagated to the last ones, via convolutional-skip connections. This shortcut flows through the gradient of a deeper network to mitigate the problem of vanishing gradients. The proposed residual block aims to optimise the dimensionality of the filters by combining 1×1 convolutions (green boxes) with a customised 3×1 convolutional layer (yellow box). This tailored kernel size is able to take advantage of the domain-specific knowledge of the OCT images to provide local cues for the glaucoma learning process. In this way, the network is encouraged to focus on the OCT vertical axis to learn the glaucoma-specific information underlying the contrast differences of the retinal layers. An additional 1×1 convolutional layer was included to reduce the filters' dimension after concatenating the feature maps from the VGG and residual structures. At this point, we also included an attention module via skip-connections to refine the features in the spatial dimension. The proposed module works as a kind of autoencoder composed of 1×1 convolutions in which the filters are decreased and increased, respectively. At the bottleneck, a single convolutional filter activated by a sigmoid function (purple layer) was included to recalibrate the inputs and forcing the network to learn useful properties from the input representations. The skip-connection propagates larger gradients to previous layers throughout an identity shortcut. At the end of the deep learning branch (see Figure 5.2), another 1×1 convolutional layer was defined to provide a volume map $G = \{g_1, g_2, ..., g_k, ..., g_C\}$, where $C = 8$ is the number of feature maps g_k, of size $H \times W = 7 \times 12$, which compose the set of the deep learning representations $\Gamma_\phi : I \rightarrow G$.

5.2.1.2 Hand-driven learning branch

Inspired by the clinical study [232], we conducted an additional hand-driven learning branch Ω_ϕ focused on hand-crafted features extracted from the RNFL. We made use of an innovative RNFL descriptor, which was proposed in our previous work [63] to include RNFL thickness-based information by computing a bespoke OCT-specific histogram. Unlike [6, 16, 168, 192, 233], among

others, where descriptors were based on the RNFL mean, the proposed method leverages the individual information provided by the RNFL thickness at each point of the B-scan by considering bags of similar thicknesses values. Thus, let I be a raw circumpapillary image of dimensions $M \times N \times d$, and S its corresponding RNFL mask segmented automatically by the Heidelberg Spectralis OCT system, a vector of thicknesses $T = \{t_1, t_2, ..., t_j, ..., t_N\}$ was computed, where t_j is the thickness value at the position j of the B-scan I, with $j = 1, 2, 3, ..., N$. The proposed histogram-based descriptor is able to quantify the RNFL information into $b = 4$ bags depending on the thickness values. In this way, each bag β_b collects the number of thicknesses t_j whose value is ranged between D_b and D_{b+1}, being $D = [0, 15, 30, 45, \infty]$ a vector of relevant distances optimised on the training images. Additionally, the minimum, maximum and average of the RNFL thickness, besides the age of the patients, were also considered according to the equations formulated in Figure 5.2. Finally, the hand-driven learning branch provides a feature vector F consisting of $C = 8$ RNFL-specific features, such that $\Omega_\phi : I \rightarrow F$.

5.2.1.3 Combination module

Once automatic and hand-crafted features were extracted from their respective branches, a simple combination module was proposed to join the embedded information in a holistic representation R composed of $C' = 16$ variables per learning instance, as observed in Figure 5.2. In particular, the feature volume G extracted from the deep learning branch was mapped to a vector G' by a Global Average Pooling (GAP) layer. This operation computes a spatial squeeze from $H \times W$ to $H' \times W'$ that enables the concatenation between the features G and F coming from the different branches. The proposed hybrid encoder Ψ_ϕ learns the embedding representations R from the input I as follows: $R = \Psi_\phi(I) = \Gamma_\phi(I) \oplus \Omega_\phi(I)$, where $\oplus$ denotes a concatenation operation.

5.2.1.4 Projection head module

We instantiates a projection head network Υ_ϕ that maps the representations R to an embedding vector Z where the classification stage is addressed in a lower-dimensional space. The projection head Υ_ϕ is comprised of a small multi-layer perceptron (MLP) with one hidden layer activated by a ReLU function (see Figure 5.2). The use of a projection head network is widely used in very recent state-of-the-art techniques, such as contrastive learning [151, 234, 235], to maximise the classification agreement. In this paper, we project the representations of the latent space via $Z = \Psi_\phi(R)$ to evidence that

the new backbone Ψ_ϕ is better than the previous feature extractors proposed in [62, 63, 65]. According to the aforementioned studies [151, 234, 235], the projection head network is then discarded during the prototypical learning stage to measure the distances from the representations $R = \Psi_\phi(I)$ to each prototype. For comparison purposes, a softmax function with three neurons was defined in the last dense layer to contrast a conventional classification approach with the proposed both static and dynamic prototypical frameworks.

As a summary of the detailed base encoder backbone, we show a pipeline in Figure 5.3 that collects the essential information. Given an input B-scan $I \in \mathbb{R}^{M \times N \times d}$, with $M \times N \times d = 248 \times 384 \times 3$ the dimensions of I, a feature embedded map $R \in \mathbb{R}^{H' \times W' \times C'}$ is provided by the encoder $\Psi_\phi : I \rightarrow R$. Then, the projection head module Υ_ϕ maps R to a metric vector $Z \in \mathbb{R}^U$, $\Upsilon_\phi : R \rightarrow Z$, where $U = 8 < C'$ denotes a lower dimensional space than the latent space R. At the end of the convolutional network, a softmax-activated dense layer is applied to address the classification stage.

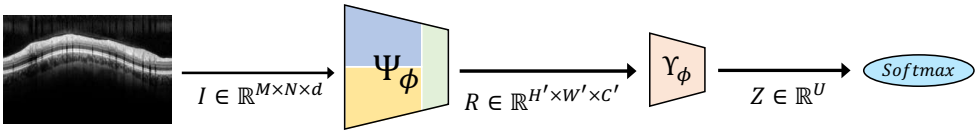


Figure 5.3: Pipeline showing the backbone architecture composed of the base encoder Ψ_ϕ, the projection head network Υ_ϕ and the softmax function.

5.2.2 Prototype-based learning strategies development

PNNs were born from the idea that there exists an embedding space in which the features from the same class cluster around a single latent group, a.k.a prototype [224]. In this paper, we conduct an experimental methodology to analyse the performance of two different prototype-based solutions with respect to conventional learning strategies for glaucoma grading. To this end, the traditional approach was defined according to the backbone architecture exposed in Figure 5.3. In contrast, for the prototype-based strategies, we propose two novel approaches -static and dynamic- that focus on the embedding space R to determine the glaucoma severity.

5.2.2.1 Static prototypes

In this paper, we introduce the concept "*static*" to reference the use of rigid prototypes extracted from an encoder network Ψ_ϕ whose weights ϕ were pre-trained in a previous stage through a conventional approach. Thereby, the frozen weights (denoted by Θ) are inferred to the base encoder model to extract the embedding representations R from both training τ and validation ν sets. Note that R^τ is used to obtain the prototypes and R^ν to find the nearest class prototype in the latent space.

A similar procedure based on measuring the similarity of the latent features from different input data has been applied on different state-of-the-art techniques, such as contrastive learning [151, 234, 235] or content-based image retrieval (CBIR) [236–238]. In contrastive learning tasks, an encoder network followed by a projection head is trained to differentiate between positive and negative samples, being positive samples the augmented version of a query (or samples with the same label in the case of supervised contrastive [234]), and negative samples the entire remainder of the batch. Contrarily, CBIR studies train convolutional autoencoders and extract the latent space from the encoder structure to find relevant images retrieved from a set of reference that shares similar embedding features with the query image. Previous systems present some similarities with the proposed static prototype approach, as all the methods train an encoder network that is then frozen to face a second classification stage. In the case of contrastive-learning studies, the feature map extracted by the encoder architecture is used to predict the class of the query sample via MLP, KNN or inferred prototypes, according to [234]. In contrast, CBIR studies are intended to find the reference set samples that most closely resemble the query image by measuring the similarity of the embedded representations extracted by the encoder structure. The proposed approach differs from the previous ones in a critical point: the encoder network is trained during the first stage for the same objective to be achieved in the second one, i.e. glaucoma grading. Oppositely, contrastive-learning and CBIR-based studies use backbones that were pre-trained for a different task during the first stage.

Below, we detail the training of the proposed approach based on static prototypes, which is composed of two (online and offline) stages, according to Figure 5.4. In the online stage, a conventional classification pipeline was conducted to optimise the weights of the proposed OCT-hybrid backbone by minimising the categorical cross-entropy (CEE) loss function $\mathcal{L}(y^\tau, \hat{y}^\tau)$ in each training epoch $e = 1, 2, 3, ..., \epsilon$, as detailed in Algorithm 6. In the offline stage, the weights of the pre-trained encoder Ψ_ϕ were frozen

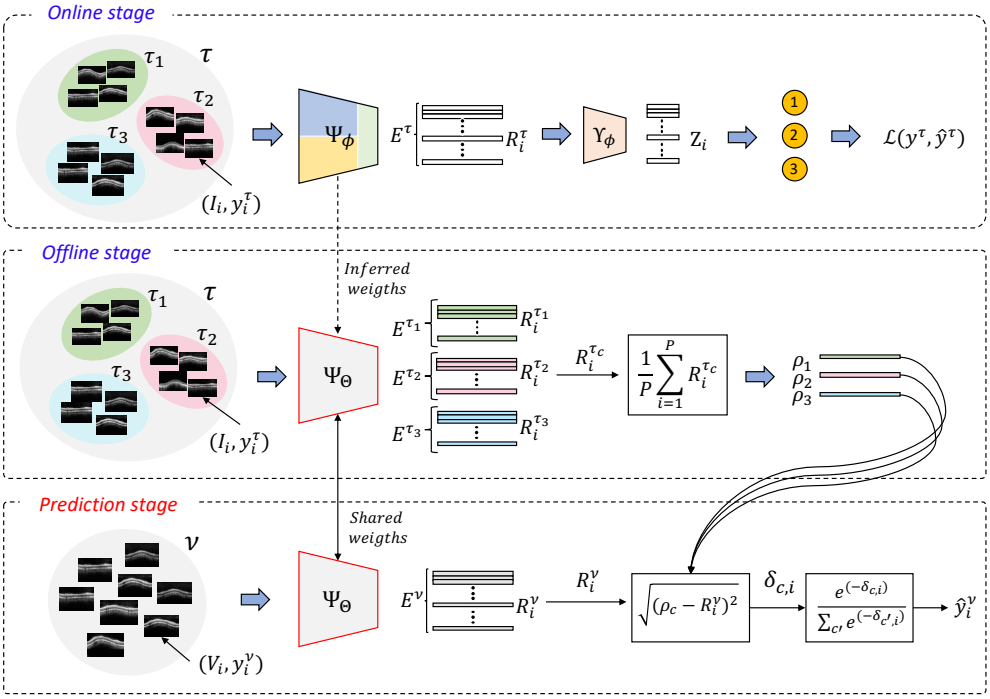


Figure 5.4: Proposed static prototype-based learning strategy. A conventional approach is conducted in the online stage to optimise the encoder. In the offline stage, the weights of the pre-trained backbone are inferred to extract a prototype ρ_c per class. In the prediction phase, the class of a validation B-scan V_i is determined by measuring the latent distance between each prototype ρ_c and the embedding representations $R_i^{\nu} = \Psi_{\phi}(V_i)$.

$(\phi \rightarrow \Theta)$ and the projection head Υ_{ϕ} and softmax modules were discarded to avoid the non-linearity of the top model. The embedding representations $E^{\tau} = \{R_1, R_2, ..., R_i, ..., R_{P^{\tau}}\}$, with P^{τ} the number of samples of the minority class in the training set τ, were used to infer the rigid prototypes. In particular, each ρ_c was calculated as the mean of the latent representations $R_i^{\tau_c} = \Psi_{\phi}(I_i)$, where $I_i \in \tau_c \subset \tau$ denotes the i-sample of the training set τ associated with the class c (see Eq. 5.1).

$$\rho_c = \frac{1}{P^{\tau_c}} \sum_{i=1}^{P^{\tau_c}} R_i^{\tau_c} \tag{5.1}$$

During the prediction phase, a matrix of distances $\delta_{c,i}$ was achieved by measuring the Euclidean distance (Eq. 5.2) between each prototype ρ_c and the embedding representation $R_i^\nu = \Psi_\phi(V_i)$, where $V_i \in \nu$ corresponds to the i-scan of the validation subset ν. A probability of belonging to each class was calculated (Eq. 5.3) to determine the predicted class $\hat{y}_i^\nu$ (Algorithm 6).

$$\delta_{c,i} = \sqrt{(\rho_c - R_i^\nu)^2} \tag{5.2}$$

$$p_{i,c} = \frac{\exp\left(-\delta_{c,i}\right)}{\sum_{c'} \exp\left(-\delta_{c',i}\right)} \tag{5.3}$$

Algorithm 6 Static prototype-based learning strategy.

Data: Training $\tau = \{(I_1, y_1^\tau), (I_{P^\tau}, y_{P^\tau}^\tau)\}$ and validation $\nu = \{(V_1, y_1^\nu), (V_{P^\nu}, y_{P^\nu}^\nu)\}$ sets.
Results:
Online stage ← Trained base encoder Ψ_ϕ;
Offline stage ← Inferred prototypes ρ_c;
Prediction stage ← Predicted labels $\hat{y}_i^\nu$;

Algorithm:
Online stage:
$\phi \leftarrow$ random;
for $e \leftarrow 1$ **to** ϵ **do**
  for $i \leftarrow 1$ **to** $\mathbf{P}^\tau$ **do**
    $R_i^\tau \leftarrow \Psi_\phi(I_i)$;
    $Z_i \leftarrow \Upsilon_\phi(R_i^\tau)$;
    $\hat{y}_i^\tau \leftarrow \text{softmax}(Z_i)$;
  $\mathcal{L}(y^\tau, \hat{y}^\tau) \leftarrow -\sum_i y_i^\tau \; log(\hat{y}_i^\tau)$;
  Update ϕ using $\nabla_\phi \mathcal{L}$;

Offline stage:
for $c \leftarrow 1$ **to 3 do**
  $\rho_c \leftarrow \frac{1}{P^\tau} \sum_{i=1}^{P^\tau} \Psi_\Theta(I_i)$;

Prediction phase:
for $i \leftarrow 1$ **to** $\mathbf{P}^\nu$ **do**
  $R_i^\nu \leftarrow \Psi_\Theta(V_i)$;
  for $c \leftarrow 1$ **to 3 do**
    $\delta_{i,c} \leftarrow \sqrt{(\rho_c - R_i^\nu)^2}$;
    $p_{i,c} \leftarrow \frac{\exp\left(-\delta_{i,c}\right)}{\sum_{c'} \exp\left(-\delta_{i,c'}\right)}$;
  $\hat{y}_i^\nu \leftarrow argmax(p_{i,c})$;

5.2.2.2 Dynamic prototypes

Inspired by [224], where PNNs were proposed for few-shot learning, we present in this paper a PNN-based framework for grading glaucoma by exploiting the k-shot methodology. The main difference with respect to the previous static approach lies in the online stage, as dynamic prototypes are trained in an end-to-end manner, such that prototypes are updated after each epoch e. In this way, the base encoder network can be optimised according to latent distances, instead of a conventional classification top model, as observed in Figure 5.5.

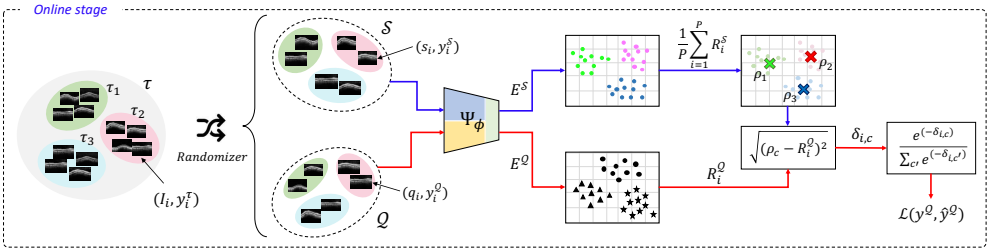


Figure 5.5: Proposed dynamic prototype-based learning strategy. In the online stage, a support set $\mathcal{S}$ and a query set $\mathcal{Q}$ are randomly selected from the training set τ to develop a supervised PNN. The hybrid OCT-based backbone Ψ_ϕ is used to extract the embedding representations from both $\mathcal{S}$, to determine the prototypes ρ_c, and $\mathcal{Q}$, to map the latent representations from the query samples. At the end of the training process, a softmax function is applied to predict the query label based on the latent distances $\delta_{i,c}$.

From here on, we will employ the terminology ($\mathcal{N}$-way, $\mathcal{K}$-shot) used in the literature for few-shot learning, where $\mathcal{K}$ is the number of labelled samples and $\mathcal{N}$ the number of classes in the training set [224]. In the state-of-the-art studies based on prototypical networks [224, 239, 240], a labelled support set $\mathcal{S} = \{(s_1, y_1), ..., (s_i, y_i), ..., (s_{\mathcal{K}}, y_{\mathcal{K}})\}$ and an unlabelled query set $\mathcal{Q} = \{q_1, q_2, ..., q_i, ..., q_{\mathcal{U}}\}$, are considered to train the PNN-based models in a few-shot scenario, being $\mathcal{U}$ the number of unlabelled samples selected from the training set τ. Specifically, $\mathcal{S}$ is used to extract each of $\mathcal{N}$ prototypes ρ_c, whereas $\mathcal{Q}$ is employed to find the nearest class prototype for an embedded query point $R_i^{\mathcal{Q}} = \Psi_\phi(q_i)$. A negative log-probability loss is updated according to the distance metrics from ρ_c and $R_i^{\mathcal{Q}}$.

The proposed dynamic prototype-based learning strategy differs from the state of the art in multiple ways. Unlike aforementioned studies [224, 239, 240], where authors selected a specific number of $\mathcal{U} = 5, 10$ or 15 unlabelled samples, depending on the database addressed, in this study we made use of $\mathcal{U} = P - \mathcal{K}$

labelled query samples to get the most out of the training set τ. Note that we propose, for the first time, the use of a labelled query set $\mathcal{Q}$, giving rise to a novel Supervised Prototypical Neural Network for Glaucoma Grading (SPNN-GG). This results in another difference concerning the literature, as a supervised approach allows the ground truth label $y^{\mathcal{Q}}$ to be included in the loss function, so that learning proceeds by minimising the CEE, instead of a log-probability function. As a novelty, we move away from the few-shot setting by proposing an optimal $k-$shot scenario in which $\mathcal{K} \in [1, P - \mathcal{U}]$ is optimised as an additional hyper-parameter, unlike in the previous studies where $\mathcal{K} \in [1, 5]$.

The online training of the proposed dynamic prototype-based approach is conducted in a supervised end-to-end way, as observed in Figure 5.5. The main difference between static and dynamic prototypes lies here, as the static approach infers rigid prototypes from a backbone trained by a projection head module, whereas the dynamic strategy updates the prototypes during the training phase by optimising a backbone based on latent distances. Note that the offline and prediction phases are addressed as in the static approach, i.e. using the entire training set τ to extract the prototypes ρ_c (Algorithm 7).

From the training set τ containing P_c^{τ} samples for each class c, a support set $\mathcal{S} = \{(s_1, y_1^{\mathcal{S}}), ..., (s_{\mathcal{K}}, y_{\mathcal{K}}^{\mathcal{S}})\}$ and a query set $\mathcal{Q} = \{(q_1, y_1^{\mathcal{Q}}), ..., (q_{\mathcal{U}}, y_{\mathcal{U}}^{\mathcal{Q}})\}$, with $\mathcal{U} = P - \mathcal{K}$, are obtained after a randomisation process at every epoch e. The hybrid OCT-based backbone Ψ_ϕ was used as the encoder network to extract the embedding representations from the support s_i and query q_i samples. Specifically, latent support set-coming features $E^{\mathcal{S}} = \Psi_\phi(\mathcal{S}) = \{R_1^{\mathcal{S}}, ..., R_i^{\mathcal{S}}, ...R_{\mathcal{K}}^{\mathcal{S}}\}$ are used to extract each class-prototype ρ_c as the mean of the embedded representations $R_i^{\mathcal{S}} = \Psi_\phi(s_i)$. In contrast, query representations $E^{\mathcal{Q}} = \Psi_\phi(\mathcal{Q}) = \{R_1^{\mathcal{Q}}, ..., R_i^{\mathcal{Q}}, ...R_{\mathcal{U}}^{\mathcal{Q}}\}$ are mapped in the latent space to find the closest prototype ρ_c, in terms of Euclidean distance. Then, a softmax function is applied to determine the probability of belonging to each class $p(\hat{y}_i^{\mathcal{Q}} = c|R_i^{\mathcal{Q}})$. During the backward-propagation step, the embedding representations are refined by updating the weights of the base encoder network at every epoch, according to the CEE loss function, denoted by $\mathcal{L}(y^{\mathcal{Q}}, \hat{y}^{\mathcal{Q}})$. In this way, the prototypes are optimised under the hypothesis that each class can be described by just one subspace. Therefore, the learning progresses by minimising the latent distances between $R_i^{\mathcal{Q}_c \subset \tau_c}$ and ρ_c, unlike in the static approach where the encoder was updated according to the embedding representations $Z = \Upsilon(\Psi(I_i))$ extracted from the projection network.

Algorithm 7 Dynamic prototype-based learning strategy.

Data: Training $\tau = \{(I_1, y_1^\tau), ..., (I_{P^\tau}, y_{P^\tau}^\tau)\}$ and validation $\nu = \{(V_1, y_1^\nu), ..., (V_{P^\nu}, y_{P^\nu}^\nu)\}$ sets.
Results:
Online stage $\leftarrow$ Trained base encoder Ψ_ϕ.
Offline stage $\leftarrow$ Refined prototypes ρ_c.
Prediction stage $\leftarrow$ Predicted labels $\hat{y}_i^\nu$.

Online stage:
$\phi \leftarrow$ random;
for $e \leftarrow 1$ **to** ϵ **do**
  for $i \leftarrow 1$ **to** $\mathcal{K}$ **do**
    $(s_i, y_i^{\mathcal{S}}) \leftarrow (I_{random(i)}, y_{random(i)}^\tau)$;
    $R_i^{\mathcal{S}} \leftarrow \Psi_\phi(s_i)$;
  for $i \leftarrow 1$ **to** $\mathcal{U}$ **do**
    $(q_i, y_i^{\mathcal{Q}}) \leftarrow (I_{random(i)}, y_{random(i)}^\tau) \notin \mathcal{S}$;
    $R_i^{\mathcal{Q}} \leftarrow \Psi_\phi(q_i)$;
    for $c \leftarrow 1$ **to** $\mathcal{N}$ **do**
      $\rho_c \leftarrow \frac{1}{\mathcal{K}} \sum_{i=1}^{\mathcal{K}} R_i^{\mathcal{S}}$;
      $\delta_{i,c} \leftarrow \sqrt{(\rho_c - R_i^{\mathcal{Q}})^2}$;
      $p_{i,c} \leftarrow \frac{\exp(-\delta_{i,c})}{\sum_{c'} \exp(-\delta_{i,c'})}$;
    $\hat{y}_i^{\mathcal{Q}} \leftarrow argmax(p_{i,c})$;
  $\mathcal{L}(y^{\mathcal{Q}}, \hat{y}^{\mathcal{Q}}) \leftarrow -\sum_u y_i^{\mathcal{Q}}\ log(\hat{y}_i^{\mathcal{Q}})$;
  Update ϕ using $\nabla_\phi \mathcal{L}$;

Offline stage:
for $c \leftarrow 1$ **to** N **do**
  $\rho_c \leftarrow \frac{1}{P^\tau} \sum_{i=1}^{P^\tau} \Psi_\Theta(I_i)$;

Prediction phase:
for $i \leftarrow 1$ **to** $\mathbf{P}^\nu$ **do**
  $R_i^\nu \leftarrow \Psi_\Theta(V_i)$;
  for $c \leftarrow 1$ **to 3 do**
    $\delta_{i,c} \leftarrow \sqrt{(\rho_c - R_i^\nu)^2}$;
    $p_{i,c} \leftarrow \frac{\exp(-\delta_{i,c})}{\sum_{c'} \exp(-\delta_{i,c'})}$;
  $\hat{y}_i^\nu \leftarrow argmax(p_{i,c})$;

5.3 Ablation experiments

5.3.1 Datasets

Two private databases coming from different sources were used to develop and evaluate the predictive models for glaucoma grading. Both datasets (*Database 1* and *Database 2*) contain high-resolution SD-OCT scans, which were acquired from healthy, early and advanced-glaucomatous patients with an axial resolution of 3-4 μm using the Heidelberg Spectralis OCT system. This equipment provides circumpapillary B-scans centred on the ONH of the retina throughout a super-luminescence diode with an infrared beam of an average wavelength of 870 nm and a bandwidth of 25 nm. The samples were extracted with a resolution of 496×768 pixels.

Subjects with open-angle glaucoma (POAG) were included in the study, whereas patients suffering other ocular disorders, such as closed-angle glaucoma or pseudoexfoliation syndrome were discarded from both databases. Also, patients with media opacity were excluded if the opacity disturbs the B-scan OCT imaging critically. Thus, B-scans with a poor-quality OCT image were discarded from the study. Two different senior ophthalmologists (with more than 25 years of professional experience in clinical ophthalmology) carried out the annotation of the databases. Specifically, each expert manually labelled just one dataset, following the European guideline for Glaucoma diagnosis. The examination included several tests such as Goldman applanation tonometry, gonioscopy, slit lamp examination, standard automated perimetry and thickness measurement of specific retinal layers of interest. According to the clinical literature [222, 223], the mean deviation (MD) score plays an essential role in the glaucoma grading scale, such that the severity of the glaucoma depends on the range in which the MD value is found. Conforming to [222], $MD >= -6dB$ is Early; $-6dB > MD \geq -12dB$ is Moderate; $-12dB > MD \geq -20dB$ is Advanced and $MD < -20dB$ is Severe. As our objective is contributing to the glaucoma grading just from OCT images, without performing additional time-consuming tests, we simplified the glaucoma staging scale by labelling as Advanced those glaucomatous samples with an $MD < -6dB$. Note that the proposed system is not intended to serve as a definitive glaucoma diagnosis, but as a diagnostic tool able to guide the expert's decision through approximate but reliable OCT-based results. Based on the above analyses, all B-scans were classified as healthy, early or advanced. In Tables Table 5.1 and 5.2, we show more information about the specifications of each dataset.

Table 5.1: Number of patients (pat.) and samples (samp.) in each database grouped by categories, according to the experts' annotation.

	Healthy *(pat./samp.)*	**Early** *(pat./samp.)*	**Advanced** *(pat./samp.)*	**TOTAL** *(pat./samp.)*
Database 1	32 / 41	28 / 35	25 / 31	85 / 107
Database 2	26 / 49	24 / 37	21 / 26	71 / 112
TOTAL	58 / 90	52 / 72	46 / 57	156 / 219

Table 5.2: Additional demographic information about the age and gender of the patients.

	Age		**Gender**	
	Range	$\mu \pm \sigma$	*Male*	*Female*
Database 1	[19-88]	60.45±16.54	46 (54.12%)	39 (45.88%)
Database 2	[30-90]	64.80±13.93	26 (36.62%)	45 (63.38%)
TOTAL	[19-90]	62.44±15.51	72 (46.15%)	84 (53.85%)

It is important to note that, as claimed in [6], there are no public glaucoma-labelled OCT databases that enable an objective comparison with our work. To the best of the authors' knowledge, AFIO dataset [221] is the only publicly available repository of ONH-centred SD-OCT scans of healthy and glaucomatous subjects. However, the differences between those B-scans and ours make the direct application of our algorithms impossible for multiple reasons: i) B-scans from AFIO dataset present manual annotations and highlighted structures of interest. ii) OCT images have been pre-processed showing an RGB colour mode in which cup-to-disk regions appears remarked. iii) AFIO database was acquired using Topcon 3D OCT-1000 machines. iv) The experts' annotations include healthy, glaucoma and suspect labels. Differently, the databases used here contain the grey-scale OCT samples extracted in raw by the Heidelberg Spectralis OCT equipment, without any manual annotation or pre-processing. In addition, our database is explicitly labelled into healthy, early and advanced glaucoma classes, according to the visual field-based criteria of the medical literature [222, 223]. Another OCT database (OCTID) (with raw ONH B-scans similar to ours) is publicly available in [241]. However, OCTID dataset includes normal, macular hole, age-related macular degeneration, central serous retinopathy and diabetic retinopathy classes, but it does not contemplate the glaucoma class.

Data partitioning. In this paper, we fuse *Database 1* and *Database 2* to increase the number of samples from which to develop our machine learning algorithms and evidence the reliability of the predictive models using two datasets coming from different sources. Making use of the entire fused

database, we conducted a patient-level data partitioning procedure to separate training and testing sets. Specifically, $\frac{1}{6}$ of the data was used to test the models, whereas the remainder of the database was employed to train the algorithms. From the training set, we randomly split the data again into training and validation subsets, according to Table 5.3, to optimise the models' hyper-parameters and monitor the overfitting.

Table 5.3: Data partitioning to train and evaluate the predictive models.

	Healthy	**Early**	**Advanced**
Training	60	48	41
Validation	15	12	10
Test	15	12	6

5.3.2 Backbone selection

Unlike most of the state-of-the-art studies which used as a feature extractor either well-known architectures such as ResNet or VGG [234, 240] or simpler CNNs trained from scratch [224, 225], we pretend to exploit the feature extraction stage to get the most out from the circumpapillary OCT scans. For this reason, we propose a novel OCT-hybrid backbone inspired by ophthalmic clinical studies [232] and our previous glaucoma detection-based experience applying hand-driven [63] and deep learning algorithms [62, 65]. To address an objective comparison with other state-of-the-art studies, we contrast in Table 5.4 and 5.5 the validation results achieved by different architectures trained in a multi-class scenario. The comparison was handled by means of different figures of merit, such as sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F-score (FS) and accuracy (ACC). Notably, the backbone reporting the best performance during the validation stage was selected as the base encoder network to address the next prototypical-based learning strategy. Particularly, we compared four approaches to select the best backbone:

1. *RNFL features.* A simple MLP was trained using the output of the hand-driven learning branch $\Omega_\phi(I)$ as an input data, similarly to [63].

2. *Fine-tuned VGG16.* This very popular architecture was used (freezing the three first convolutional blocks) as a feature extractor followed by the same MLP classifier as before, according to [62].

3. *Fine-tuned RAGNet.* An expanded version of the previous approach was conducted by including residual and attention modules in the feature extraction architecture. This approach, corresponding to the deep learning branch $\Gamma_\phi(I)$, was introduced in our previous work [65].

4. *OCT-hybrid network.* This approach corresponds to the end-to-end backbone $\Psi_\phi(I)$ exposed in Figure 5.2, which proposes a combination of hand-crafted and automatic features before the top model.

Table 5.4: Backbone selection: Validation results per class achieved from different architectures in a multi-class scenario.

	HEALTHY				EARLY GLAUCOMA				ADVANCED GLAUCOMA			
	RNFL features [63]	*Fine-tuned VGG16 [62]*	*Fine-tuned RAGNet [65]*	*OCT-hybrid network*	*RNFL features [63]*	*Fine-tuned VGG16 [62]*	*Fine-tuned RAGNet [65]*	*OCT-hybrid network*	*RNFL features [63]*	*Fine-tuned VGG16 [62]*	*Fine-tuned RAGNet [65]*	*OCT-hybrid network*
SN	**1**	**1**	0.9333	**1**	0.6667	0.5833	**0.7500**	**0.7500**	0.8000	**0.9000**	**0.9000**	**0.9000**
SP	0.8636	0.8636	**0.9545**	0.9091	0.9200	**0.9600**	0.9200	**0.9600**	**0.9630**	0.9259	0.9259	**0.9630**
PPV	0.8333	0.8333	**0.9333**	0.8824	0.8000	0.8750	0.8182	**0.9000**	0.8889	0.8182	0.8182	**0.9000**
NPV	**1**	**1**	0.9545	**1**	0.8519	0.8276	0.8846	**0.8889**	0.9286	0.9615	0.9615	**0.9630**
FS	0.9091	0.9091	0.9333	**0.9375**	0.7273	0.7000	0.7826	**0.8182**	0.8421	0.8571	0.8571	**0.9000**
ACC	0.9189	0.9189	**0.9459**	**0.9459**	0.8378	0.8378	0.8649	**0.8919**	0.9189	0.9189	0.9189	**0.9459**

Table 5.5: Backbone selection: Average validation results achieved from different approaches in a multi-class scenario.

	Micro-Average				Macro-Average			
	RNFL features [63]	*Fine-tuned VGG16 [62]*	*Fine-tuned RAGNet [65]*	*OCT-hybrid network*	*RNFL features [63]*	*Fine-tuned VGG16 [62]*	*Fine-tuned RAGNet [65]*	*OCT-hybrid network*
SN	0.8378	0.8378	0.8649	**0.8919**	0.8222	0.8278	0.8611	**0.8833**
SP	0.9189	0.9189	0.9324	**0.9459**	0.9155	0.9165	0.9335	**0.9440**
PPV	0.8378	0.8378	0.8649	**0.8919**	0.8407	0.8422	0.8566	**0.8941**
NPV	0.9189	0.9189	0.9324	**0.9459**	0.9268	0.9297	0.9336	**0.9506**
FS	0.8378	0.8378	0.8649	**0.8919**	0.8262	0.8221	0.8577	**0.8852**
ACC	0.8919	0.8919	0.9099	**0.9279**	0.8919	0.8919	0.9099	**0.9279**

Training details. All the approaches were implemented using Tensorflow 2.3.1 on Python 3.6. Experiments were conducted on a machine with Intel(R) Core(TM) i7-9700 CPU @3.00GHz processor and 16GB RAM. A single NVIDIA GeForce RTX 2080 having cuDNN 7.5 and a CUDA Toolkit 10.1 was used to develop the deep learning algorithms. All the models were trained during 200 epochs using a learning rate of 0.0005 with a batch size of 16. Stochastic gradient descent optimiser was applied trying to minimise the CEE loss function at every epoch. The rest of the architecture hyper-parameters and input dimensions are shown in Figure 5.2.

5.3.3 Prototype-based learning strategies

In this section, we report the validation performance of the static and dynamic prototype-based methods in comparison to the conventional multi-class approach. The comparison was conducted using the proposed OCT-hybrid backbone as a feature extractor for all the scenarios. Following the organisation of the previous section, Table 5.6 and 5.7 show the comparison between the three involved learning strategies during the validation phase.

Table 5.6: Learning strategy: Validation results per class using the proposed OCT-hybrid backbone for glaucoma grading.

	HEALTHY			EARLY GLAUCOMA			ADVANCED GLAUCOMA		
	Conventional multi-class	*Static prototypes*	*Dynamic prototypes*	*Conventional multi-class*	*Static prototypes*	*Dynamic prototypes*	*Conventional multi-class*	*Static prototypes*	*Dynamic prototypes*
SN	**1**	**1**	**1**	0.7500	0.7500	**0.8333**	**0.9000**	0.8000	**0.9000**
SP	0.9091	0.9091	**1**	**0.9600**	0.9200	**0.9600**	**0.9630**	0.9630	0.9259
PPV	0.8824	0.8824	**1**	0.9000	0.8182	**0.9091**	**0.9000**	0.8889	0.8182
NPV	**1**	**1**	**1**	0.8889	0.8846	**0.9231**	**0.9630**	0.9286	0.9615
FS	0.9375	0.9375	**1**	0.8182	0.7826	**0.8696**	**0.9000**	0.8421	0.8571
ACC	0.9459	0.9459	**1**	0.8919	0.8649	**0.9189**	**0.9459**	0.9189	0.9189

Table 5.7: Learning strategy: Average validation results using the proposed OCT-hybrid backbone as a feature extractor.

	Micro-Average			Macro-Average		
	Conventional multi-class	*Static prototypes*	*Dynamic prototypes*	*Conventional multi-class*	*Static prototypes*	*Dynamic prototypes*
SN	0.8919	0.8649	**0.9189**	0.8833	0.8500	**0.9111**
SP	0.9459	0.9324	**0.9595**	0.9440	0.9307	**0.9620**
PPV	0.8919	0.8649	**0.9189**	0.8941	0.8631	**0.9091**
NPV	0.9459	0.9324	**0.9595**	0.9506	0.9377	**0.9615**
FS	0.8919	0.8649	**0.9189**	0.8852	0.8541	**0.9089**
ACC	0.9279	0.9099	**0.9459**	0.9279	0.9099	**0.9459**

Training details. The same hardware and software systems as before were used to accomplish this section. However, some differences in the dynamic prototypical approach are worth noting. Dimensions of the input image must be downsized to $124 \times 192 \times 3$ to face the GPU memory constraints. Additionally, a decreased learning rate of 0.001 allowed the convergence of the model in 50 training epochs. A batch size of 16 samples was defined to minimise the CEE loss function using the SGD optimiser. Regarding the specific parameters of the dynamic prototype-based strategy, the number of $\mathcal{K}$ shots and the number of $\mathcal{U}$ query samples were determined after an optimisation process, according to Figure 5.6. Specifically, $\mathcal{K} = 20$ support samples

were selected to extract the prototypes ρ_c as the mean of the embedding representations $E^{\mathcal{S}}$, whereas $\mathcal{U} = 21$ query samples were used to measure the Euclidean distance between the latent features E^{Ω} and each ρ_c. Note that $P = 41$ denotes the number of training samples of the minority class (see Table 5.1). In addition, other statistics and distance metrics were considered during the optimisation of the models, as observed in Table 5.8. The rest of the hyper-parameters related to the dynamic prototypes is detailed in Subsection 5.2.2.2.

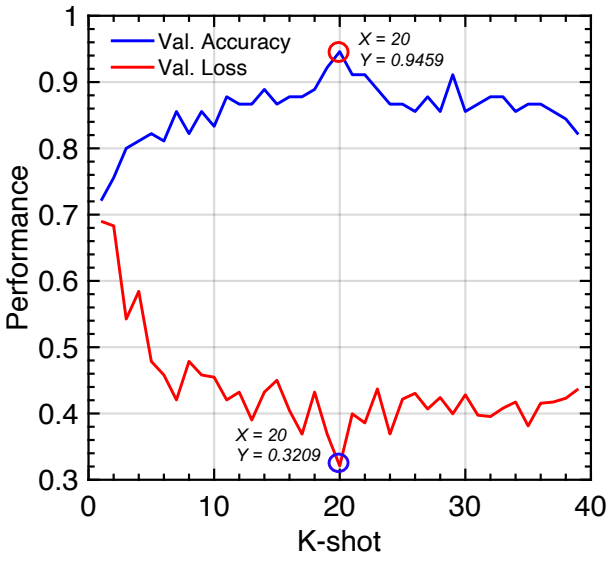


Figure 5.6: Performance of the proposed model depending on $\mathcal{K}$. The graphic shows validation accuracy and loss results (in blue and red, respectively) using $\mathcal{K}$ samples from each class c to define the prototypes ρ_c, and $\mathcal{U} = P - \mathcal{K}$ query samples to measure the embedding distances to minimise the loss function.

Table 5.8: Validation accuracy reached by the dynamic prototypical approach using different statistics and distance metrics.

		Distance			
		Euclidean	*Cosine*	*Manhattan*	*Canberra*
Statistic	*mean*	**0.9459**	0.9099	0.9279	0.8739
	median	0.9279	0.9279	0.9099	0.8559

5.4 Prediction results

Quantitative results

In this section, we show the quantitative results achieved by the three learning strategies conducted during the prediction of the test set. All the approaches contrasted here were addressed using the proposed OCT-based hybrid backbone as a feature extractor, as it reported the best results in the validation phase. As before, we evaluate the models' performance both per class (Table 5.9) and in terms of micro- and macro-average (Table 5.10).

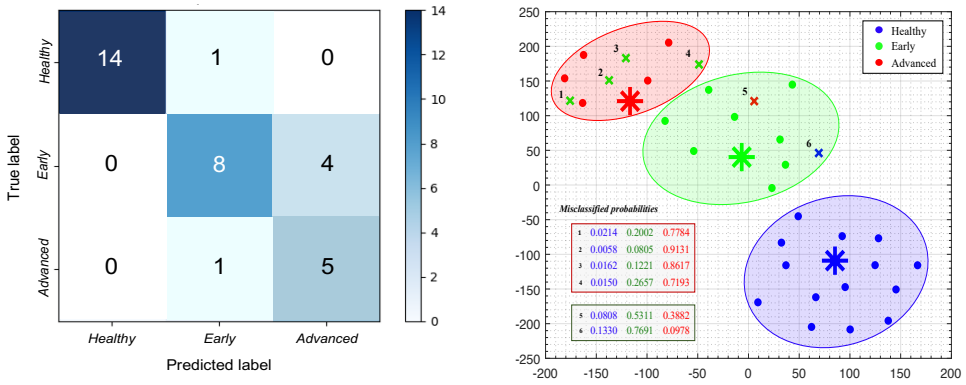
Table 5.9: Prediction stage: Test results per class reached by the different proposed learning strategies for glaucoma grading.

	HEALTHY			EARLY GLAUCOMA			ADVANCED GLAUCOMA		
	Conventional multi-class	*Static prototypes*	*Dynamic prototypes*	*Conventional multi-class*	*Static prototypes*	*Dynamic prototypes*	*Conventional multi-class*	*Static prototypes*	*Dynamic prototypes*
SN	**0.9333**	**0.9333**	**0.9333**	0.5000	0.4167	**0.6667**	**1**	0.8333	0.8333
SP	**1**	**1**	**1**	**0.9524**	0.9048	0.9048	0.7778	0.7407	**0.8519**
PPV	**1**	**1**	**1**	**0.8571**	0.7143	0.8000	0.5000	0.4167	**0.5556**
NPV	**0.9474**	**0.9474**	**0.9474**	0.7692	0.7308	**0.8210**	**1**	0.9524	0.9583
FS	**0.9655**	**0.9655**	**0.9655**	0.6316	0.5263	**0.7273**	**0.6667**	0.5556	**0.6667**
ACC	**0.9697**	**0.9697**	**0.9697**	0.7879	0.7273	**0.8182**	0.8182	0.7576	**0.8485**

Table 5.10: Prediction stage: Average test results achieved by the different proposed learning strategies at the test time.

	Micro-Average			Macro-Average		
	Conventional multi-class	*Static prototypes*	*Dynamic prototypes*	*Conventional multi-class*	*Static prototypes*	*Dynamic prototypes*
SN	0.7879	0.7273	**0.8182**	0.8111	0.7278	**0.8112**
SP	0.8939	0.8636	**0.9091**	0.9101	0.8818	**0.9189**
PPV	0.7879	0.7273	**0.8182**	0.7852	0.7103	**0.7857**
NPV	0.8939	0.8636	**0.9091**	0.9055	0.8768	**0.9106**
FS	0.7879	0.7273	**0.8182**	0.7446	0.6825	**0.7865**
ACC	0.8586	0.8182	**0.8788**	0.8586	0.8182	**0.8788**

In Figure 5.7 (a), we show the confusion matrix obtained by the best approach, i.e. the dynamic prototype-based model, to evidence the overall behaviour of the proposed method when predicting new samples. To provide a more comprehensive interpretation of the glaucoma grading scenario, we illustrate in Figure 5.7 (b) a 2D map corresponding to the latent space arranged by the dynamic learning. The prototypes (denoted by asterisks) were calculated from the training and validation sets, whereas spots and crosses make reference to the embedding representations of the well and missclassified test data, respectively. In addition, the Euclidean distance-based probabilities from the missclassified samples are detailed in Figure 5.7 (b) to manifest the confidence of the dynamic model when it is wrong in the prediction.



(a) (b)

Figure 5.7: Confusion matrix and TSNE. (a) Confusion matrix obtained by the dynamic prototype-based approach during the prediction phase. (b) Latent space showing the prototypes and the embedding test features on a 2D map via t-distributed Stochastic Neighbour Embedding (TSNE) tool. Blue, green and red colours denote the ground truth labels of the healthy-, early- and advanced-derived representations, respectively.

Qualitative results

Class activation maps (CAMs) [147] were computed to remark the regions in which the proposed dynamic prototypical network paid attention to predict the class of the test samples. The reported heat maps allow for a better understanding of the CNN-extracted features by highlighting the most relevant information of the B-scan for the predictions. These activated maps may provide an additional interpretation of the results for glaucoma grading depending on the patterns highlighted for each class. However, it should be mentioned that CAMs usually do not present a high precision at pixel-level, as the heat maps are created from the last layers of the proposed model, which output a down-sampled map from the input image.

In Figure 5.8, we show several examples of CAMs corresponding to correctly and wrongly predicted samples to elucidate the relevant patterns found by the dynamic prototypical network to address the B-scans prediction. Specifically, we expose three well-classified heat maps for each class to demonstrate the criteria followed by the proposed model to determine the predicted labels. Also, we report in the red frame of Figure 5.8 some examples of missclassified B-scans to visually evidence the reason why the model gets wrong in the prediction. Rows and columns represent the predictions and labels, respectively.

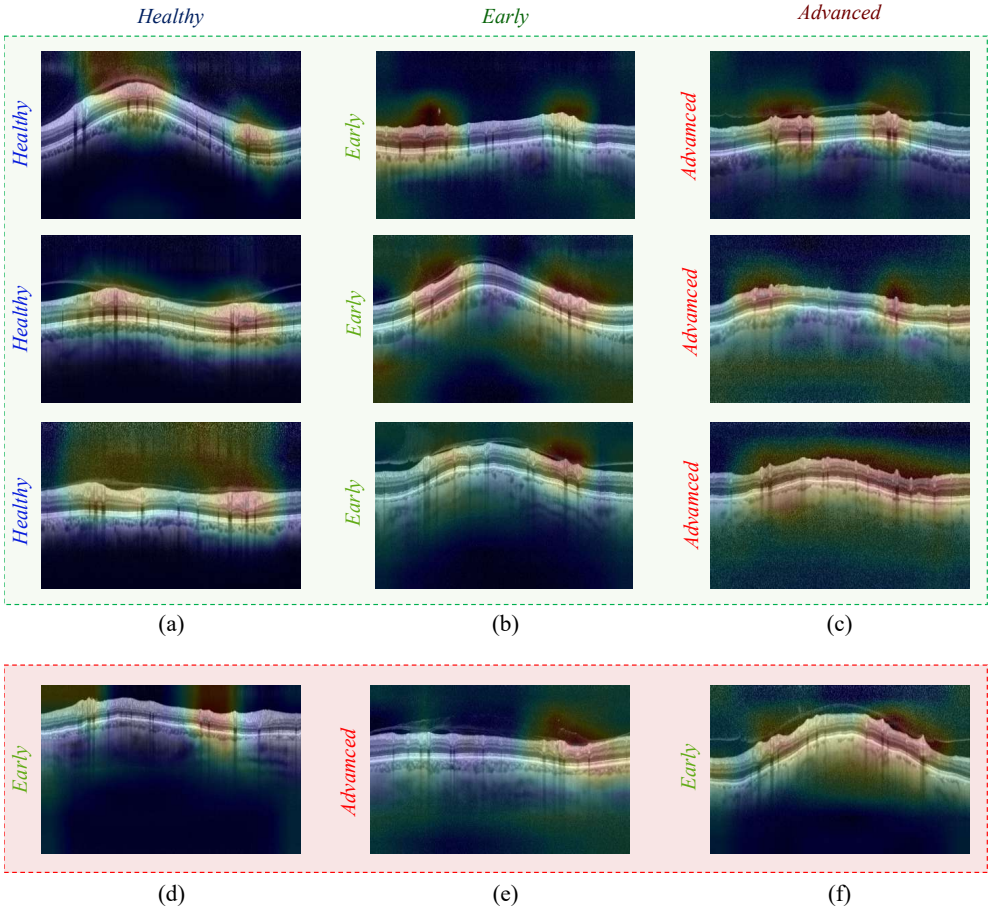


Figure 5.8: Class activation maps showing the regions of interest in which the proposed prototypical-dynamic model pays attention to predict each class. Heat maps over the green background (a-c) correspond to well-classified images, whereas the red frame (d-f) represents the missclassified samples.

5.5 Discussion

OCT B-scans contain information too limited to provide a complex diagnosis based on glaucoma grading. For this reason, additional tests are usually carried out to accomplish a more exhaustive and reliable glaucoma grading diagnosis considering all the stages of the disease. Currently, the OCT systems can provide a result indicating if the B-scan is within normal limits, borderline or

outside normal limits, which is equivalent to discern between healthy, suspect and glaucomatous cases. At this point, our system has demonstrated to provide a very high performance achieving an accuracy of 100% and 96, 97% during the validation and testing phases, respectively. However, with the proposed system, we pretend to go beyond the binary classification of healthy vs glaucoma by means of a prototypical framework able to discern early from advanced glaucoma only using OCT samples, which adds significant value to the body of knowledge. Below, we discuss the different CNN configurations and learning strategies carried out in this work for glaucoma grading.

5.5.1 On ablation experiments

Backbone selection. One of the main novelties addressed in this paper lies in the proposed OCT-based hybrid network, which combines hand-crafted and automatic features extracted from the input B-scans. In order to elucidate the superiority of the proposed model with respect to similar approaches, we raise a multi-class glaucoma-grading scenario comparing different methodologies. Note that a direct comparison with other state-of-the-art studies was not possible because of the lack of public repositories with annotations of different glaucoma severity levels, as claimed in [6]. For that reason, in this paper, we contrasted the proposed model with other own glaucoma-detection methods recently published in [62, 63, 65]. It should be mentioned that all the approaches were adapted to the multi-class environment to provide a reliable comparison during the validation stage, as detailed in Table 5.4 and 5.5. Particularly, from Table 5.4 we can observe that the fine-tuned RAGNet model [65] introduces slight improvements compared to the traditional VGG16 architecture and the hand-crafted RNFL features. The use of tailored residual blocks and attention modules lead to a better results for most of the metrics, by providing more distinctive feature maps. However, the proposed OCT-based hybrid network outperforms the rest of the models in the discrimination of all the classes. Focusing on the healthy column, RNFL features, VGG16 architecture and the proposed hybrid backbone report a more sensible behaviour highlighting for the SN and NPV metrics, whereas RAGNet model showcases more specific results by outstanding for the SP and PPV figures of merit. More global metrics, such as FS and ACC show higher values for RAGNet and OCT-hybrid approaches. Contrarily, the models differ to a greater extent in the discrimination of the glaucoma severity grades. Specifically, the proposed OCT-hybrid network report the best performance distinguishing between early and advanced glaucomatous samples for all the measures. The results detailed in Table 5.5 further strengthens

our confidence in the proposed OCT-hybrid network, as it reports the higher values for all the figures of merit in terms of micro- and macro-average. Table 5.5 is especially interesting to compare the model's performance, as it gives an idea of their overall precision. In such a table, we can observe that hand-driven and basic deep learning approaches present a similar behaviour, whereas more sophisticated CNNs provide substantial performance improvements. Nevertheless, the combination of hand-crafted RNFL-based features and refined CNN architectures yields the best model for glaucoma grading, achieving an average accuracy of 0.9279.

Prototype-based learning strategy. The best approach reported during the validation phase was used as a feature extractor to address the next stage corresponding to the prototype-based learning strategy. Therefore, according to the previous statements, the proposed OCT-hybrid network Ψ_ϕ was selected as the backbone of the prototypical architectures. At this point, several contributions are proposed in a novel framework for glaucoma grading. The method based on static prototypes was conducted for the first time by inferring the weights ($\phi \rightarrow \Theta$) of class-based prediction networks, instead of auto-encoders, as in the case of CBIR-based studies [236–238], or architectures intended to discern between positive and negative classes, as in the contrastive learning works [151, 234, 235]. Notwithstanding, the main novelties related to the prototypical environment were introduced in the dynamic approximation. As detailed in Subsection 5.2.2.2, state-of-the-art studies usually made use of $\mathcal{K} \in [1, 5]$ labelled images to extract the prototypes and $\mathcal{U} = 5$, 10 or 15 unlabelled query samples to measure the latent distance to the prototypes during the training of the models. Unlike them, in this paper, the few-shot paradigm was redefined by dealing the $\mathcal{K}$ and $\mathcal{U} = P - \mathcal{K}$ variables as any other hyper-parameter to be optimised in the validation phase. As another novelty, we made use of labelled samples to build the query set $\mathcal{Q}$ in order to make the most of all available information about the B-scans for glaucoma grading. Our hypothesis claimed that the use of more than $\mathcal{K} = 5$ support images to extract the prototypes would provide better results. However, it would be necessary to find the optimal balance between the number of support and query samples, because the data used to extract the prototypes ρ_c is just as important as the data employed to know how well the prototypes work. For that reason, we show in Figure 5.6 the categorical accuracy and the loss value achieved by the dynamic prototypical model using $\mathcal{K} \in [1, P]$ samples, being $P = 41$ the total number of the training samples of the minority class (see Table 5.3). As expected, the model's performance using a few $\mathcal{K}$ samples remains low and it improves as the number of support samples increases. However, the model reaches the best in the middle of the plotting, and then, the curve

stabilises or even worsens. This fact consolidates our hypothesis which holds the importance of a similar splitting of the support and query samples to train the models. In particular, the highest performance is reached using $\mathcal{K} = 20$ shots and $\mathcal{U} = 21$ query scans, which reports a validation accuracy of 0.9459 and a loss value of 0.3209, according to Figure 5.6.

In Table 5.6 and 5.7, both static and dynamic prototype-based approaches are compared with a conventional multi-class system based on the OCT-hybrid network trained for glaucoma grading. The classification of healthy samples in Table 5.6 deserves special attention, as the fully supervised dynamic PNN perfectly works for detecting the non-glaucomatous data, achieving the 100% of performance for all the figures of merit. The discrimination of the middle class, i.e. early class, makes the outperforming of the dynamic prototypical network evident. However, the conventional multi-class strategy surpasses the results for advanced glaucoma cases in all the metrics. In addition, Table 5.7 provides a comparison of the models' behaviour in terms of micro- and macro-average. Specifically, the dynamic prototype reaches the best results with values higher than 0.90 for all the indices. It should be highlighted that the baseline approach outperforms the strategy based on static prototypes, which reveals that end-to-end methodologies provide more robust models, as expected.

Furthermore, we carried out an empirical exploration of different distance metrics and statistical parameters to maximise the classification agreement in the latent space between the prototypes and the embedding query representations. Particularly, statistics based on mean and median operations were considered to extract the prototypes from the support samples, whereas Euclidean, Cosine, Manhattan and Canberra distances were subject to study to determine the class of the query representations according to the closest prototype in the latent space. As detailed in Table 5.8, the highest validation accuracy was reached using the Euclidean distance and the mean statistic, which reported a validation accuracy of 0.9459.

5.5.2 On prediction results

Quantitative results. In this section, we compare the performance of the different proposed learning strategies during the prediction of the test set. As in the validation phase, results per class and in terms of micro and macro-average are reported in Table 5.9 and 5.10, respectively. According to the previous section, the discrimination between healthy and glaucoma classes is successfully accomplished by achieving results higher than 0.93 in the prediction of healthy samples for all the figures of merit (see Table 5.9). It is remarkable that

the three contrasted models provide the same effectiveness in distinguishing the healthy class. However, the glaucoma grading results make clear the differences between the models. In particular, the dynamic PNN shows better SN and NPV results, whereas the baseline approach out stands for SP and PPV metrics in the early glaucoma detection. The opposite happens in the advanced glaucoma case, as the conventional approach reveals a more sensitive behaviour (higher recall and NPV), whereas the dynamic prototypical model provides more specific results (better specificity and precision). Note that F-score (FS) and Accuracy (ACC) metrics reach higher values using the dynamic PNN-based strategy. Additionally, the results in Table 5.10 are directly in line with the reported in the validation phase, as the dynamic prototypical method is consolidated as the best model, followed by the conventional multi-class approach and leaving the static prototypical network in the last position. From here, we can conclude that end-to-end trained systems are more compelling and allow providing more reliable and robust models.

From the confusion matrix in Figure 5.7 (a), promising results are evidenced for glaucoma detection, as the dynamic prototype strategy almost perfectly distinguishes glaucomatous from healthy samples. Only a specific healthy B-scan was wrongly predicted as a sample with early glaucoma. However, the proposed model makes more mistakes predicting the early class since it sometimes confuses early and advanced patterns of the disease. In the case of the advanced glaucoma class, the prototypical network only disagrees with the experts once, in which a severe glaucomatous scan was classified as early glaucoma. To visualise the results of the confusion matrix in a more interpretable scenario, we illustrated the latent space environment arranged by the proposed dynamic SPNN-GG in Figure 5.7 (b). The TSNE technique shows the distribution of the embedding test representations on a 2D map, as well as the prototypes extracted from the training samples. Three well-differentiated sub-spaces arise on the 2D map locating the healthy and advanced representations in the opposite corners and the prototypes of early and advanced classes very close, as expected. Additionally, early glaucoma-derived features are mapped in the middle of the plotting, which matches with the order of the glaucoma severity scale. The prediction probabilities for the missclassified samples are detailed in Figure 5.7 to demonstrate the coherence of the proposed model using the embedding features for glaucoma grading. For example, the nicknamed representations $1-4$ correspond to early glaucomatous samples which were missclassified as advanced glaucoma items with a level of confidence less than 0.8 in some cases. Regarding the tagged representation 5, the dynamic prototypical network wrongly predicted an early B-scan as an advanced sample, but with serious difficulties, as the prediction

probabilities are 0.53 and 0.39 for early and advanced classes, respectively. Something similar occurs for the nicknamed point 6, which was missclassified as an early glaucoma sample when it was actually healthy. In spite of this, the representation 6 is located near the decision boundary between early and healthy classes, which manifests the robustness of the proposed model.

Qualitative results. The findings from the CAMs reported in Figure 5.8 keep consistency with the clinical interpretation provided by the ophthalmologists, who claim that a thinning of the RNFL structure is usually associated with glaucomatous patterns, whereas a thickening of the RNFL denotes cues of healthy samples [189, 190]. As appreciated in Figure 5.8, the proposed model pays special attention to specific regions of the RNFL depending on the reported prediction, according to the outcomes found in our previous work [62]. In the green frame of Figure 5.8, three random examples are illustrated to demonstrate that there is a repeating pattern for each of the classes. Well-classified samples corresponding to the healthy class (Figure 5.8. (a)) show heat maps with highlighted pixels in regions characterised by a thickening of the RNFL. Oppositely, well-classified early and advanced glaucomatous samples (Figure 5.8. (b), (c)) provide apparent cues of RNFL deterioration, which is visible in the heat maps highlighting regions where a thinning of the RNFL is evident. As observed, the degradation level of the RNFL thickness is accentuated the greater the severity of the disease.

The wrongly predicted CAMs can be also appreciated in the red frame of Figure 5.8. Particularly, the example showed in Figure 5.8 (d) shows a healthy B-scan missclassified as an early glaucoma sample. Nevertheless, the model's decision shows to be coherent with the established patterns, as the B-scan presents several areas in which the RNFL appears slightly thinned. Something similar happens with the remainder of cases of the red frame. Specifically, the example in Figure 5.8 (e) is striking because the RNFL seems to be widely deteriorated throughout the entire B-scan, so that the model associates the image with an advanced glaucomatous sample. The opposite occurs in Figure 5.8 (f), where an advanced glaucoma OCT image is predicted with the early label. In this case, the B-scan shows several regions in which the RNFL thickness agrees more with early than advanced glaucoma patterns. Therefore, from the red frame of Figure 5.8, it is possible to elucidate the complexity in the diagnosis of the different grades of glaucoma, even for expert ophthalmologists, who often disagree. In future research lines, it would be interesting to include additional data to provide the model with the necessary information that clinicians take into account to determine the glaucoma severity.

5.6 Conclusion

In this paper, we have proposed several artificial intelligence solutions for glaucoma grading using raw circumpapillary OCT images. A new base encoder network has been developed improving the multi-class methodologies addressed to date for glaucoma severity detection. The proposed model introduces a residual convolutional block with tailored kernel sizes to get the most out from the B-scans, and an attention module able to refine the features of the latent space to maximise the classification agreement. The encoder network uses, for the first time, a combination of hand-crafted and automatic features giving rise to an OCT-based hybrid model which adds substantial improvements in the final prediction. As a novelty, this architecture was employed as the backbone of a novel framework based on prototypical learning, in which the limits of the few-shot paradigm have been redefined for an optimal glaucoma grading procedure. This innovative approach carried out in an end-to-end manner has surpassed the multi-class baseline and it has reported promising results for both glaucoma detection and glaucoma grading scenarios, achieving testing accuracy of 0.9697 and 0.8788, respectively. In future research lines, the efforts should be focused on improving the discrimination of the different grades of glaucoma severity by including additional information outside the SD-OCT images.

Chapter 6

Final conclusions

This chapter relates the findings from each paper to the final aim of the PhD thesis. It summarises concluding remarks and suggests future research lines for each proposed learning framework.

Contents

6.1 Global remarks

In this thesis, we have designed, developed and validated different ML algorithms for imaging-based diagnostic support in order to improve expert efficiency and accuracy. We have proposed several cutting-edge AI-based solutions to aid decision-making of the two most important research areas in the field of medical imaging: histopathology and ophthalmology. For this purpose, different modelling strategies from different learning paradigms have been considered. Specifically, this PhD thesis compiles from conventional approaches focusing on hand-driven learning to data-driven algorithms based on deep learning techniques, as well as hybrid frameworks by combining both manual and automatic learning perspectives. Regarding the conventional approaches, novel descriptors have been proposed to perform robust hand-crafted feature extraction. Moreover, in-depth statistical analyses have been performed for feature selection, and classical ML algorithms such as SVM and MLP have been implemented to address the classification stage. As for deep learning-based approaches, innovative methods from different supervision scenarios have been proposed to deal with the lack of labelled data. In addition to the supervised setting performed under ideal conditions, this thesis also includes other methods based on fully unsupervised learning via deep clustering, recurrent learning from 2D slides with volume-level labels, few-shot learning through prototypical neural networks, and self-training under the presence of domain shift.

6.2 Specific remarks

In Chapter 2, we have presented a comparison between hand-driven and deep learning methods for identifying the first stage of prostate cancer from previously segmented histological gland candidates. Both learning strategies have reported similar behaviour in distinguishing between artefacts and glands. However, the conventional computer vision approach focusing on morphological, textural, fractal-related and contextual features has outperformed the deep learning methodology, as hand-crafted features can take into account the size of specific elements, as well as spatial hierarchies and orientations, which play an essential role in discriminating between healthy and pathological glands. Additionally, the SVM classifier using a second-order non-linear kernel has reported better results than other ML classifiers based on artificial neural networks. The proposed system has also demonstrated higher performance than other state-of-the-art methods based on gland classification, surpassing multi-class accuracy by more than 10.4%. This study contributes

to the accurate localisation of histological structures and their classification into normal or pathological glands with an accuracy of 88.30%.

In Chapter 3, we have compared conventional clustering algorithms with unsupervised deep learning methods for bladder cancer assessment, also using histological images. In this case, an additional immunohistochemistry staining was employed to enhance the cancer cells with the aim of self-recognising different high-resolution bladder patterns, where the aggressiveness of each pattern leads to a different clinical prognosis. In contrast to the previous chapter, where the classification framework was approached from individual structures, in this study we focused on a self-learning strategy carried out from histological patches. Conventional and deep clustering methods have been able to discern between tumour and non-tumour patterns with a very high success rate, which is not a major finding as the differences are obvious. The real challenge lies in distinguishing between mild and infiltrative patterns, where the tumour structures are similar, but the patient's prognosis is completely different. In this context, the proposed deep learning method (DCEAC), using an attention module to refine the features at the bottleneck, has reported significant outperformance, especially regarding conventional approaches, as they were unable to learn the differences between the cancer patterns on their own. The proposed DCEAC model has shown an increase in accuracy of 12.85% compared to the classical clustering approaches and $2-3\%$ concerning other deep learning state-of-the-art methods. In addition, the feature representation provided by the proposed DCEAC showcases the high ability of the model to discern between non-tumour, mild and infiltrative patterns, as the features derived from each of them appear as independent clusters in the latent space. The heat maps extracted also reveal that the proposed system can learn by itself the same histological structures as clinicians to determine the aggressiveness of bladder cancer. This contributes significantly to the body of knowledge, as a multi-class accuracy of 90.31% has been reached without incurring prior annotation steps. Therefore, the proposed AI-based solution is valuable in addressing human eye fatigue and assisting inexperienced pathologists by suggesting inadvertent areas of interest to avoid biased diagnosis.

In the previous chapters, different learning models, performing under different setups, have been applied on histopathological images to demonstrate the potential of artificial intelligence for cancer diagnosis, in particular prostate and bladder cancer. Otherwise, Chapter 4 and Chapter 5 move away from digital pathology to put AI knowledge at the service of glaucoma screening through OCT-based data. In Chapter 4, we have proposed a supervised framework to

detect glaucoma from annotated SD-OCT volumes composed of unlabelled 2D slides. The combination of a novel 2D-CNN, acting as a slide-level feature extractor, with LSTM networks, responsible for leveraging the spatial dependencies between adjacent slides, results in a model with better predictive capability than other state-of-the-art architectures based on 3D-CNNs. For the first time, to build the slide-level discriminator, we combined fine-tuned architectures with residual and attention modules, which have been shown to lead to better convergence by optimising the dimensionality of the filters during the model training. For the volume-level prediction, the proposed recurrent framework using a novel sequential-weighting module (SWM) has provided higher performance regarding other LSTM-based configurations, as all the slide-derived features contribute to the final prediction but in a sequential way. Taking into account that SD-OCT volumes are not currently being traced by ophthalmologists due to the large workload associated, the reported heat maps highlighting the regions of interest at the slide level could lead to a promising tool for the screening of the SD-OCT cubes. Therefore, this study has revealed that SD-OCT volumes enclose valuable knowledge for glaucoma assessment, as the proposed model has achieved an accuracy of 81.25%.

Contrary to the previous chapter, where glaucoma detection is addressed from SD-OCT volumes, in Chapter 5, we have oriented glaucoma assessment using 2D circumpapillary images, rather than three-dimensional data. In this case, we based on our previous glaucoma-related experience to propose a new hybrid backbone to optimise the feature extraction process. The inclusion of hand-crafted features, intended to enhance essential glaucoma-specific structures associated with the deterioration of the retinal cells, has resulted in substantial prediction improvements regarding other approaches only focused on deep learning architectures. Nevertheless, results have proven to be even higher when the proposed hybrid backbone was used as a feature extractor in a novel few-shot learning framework based on prototypical neural networks (PNNs). The learning process is shown to maximise classification agreement using dynamic PNNs, instead of the non-linear operations associated with softmax activation of the multi-layer perceptron. The performance improvement is because the proposed dynamic PNN model allows the convolutional coefficients to be refined during the back-propagation step as a function of the success of latent embedding assignment. By providing the prototypical learning for supervision and optimising the few-shot paradigm, the proposed system has reported very promising results on both binary and multi-class classification tasks. At the time of testing, an accuracy of 96.97% has been achieved in discerning between healthy and glaucomatous samples. However, the most

remarkable feat is the 87.88% accuracy achieved for OCT-based glaucoma grading, which was addressed for the first time in this work.

In summary, the AI-based methods detailed in this thesis have proven to be a valuable tool to assist in diagnostic imaging, whether for prostate and bladder cancer diagnosis making use of histological samples or for glaucoma assessment using OCT-based data. Nevertheless, this thesis has covered just a very small part of the full potential of AI for medical imaging.

6.3 Future work

Many research lines remain open for future works in the fields of digital pathology and ophthalmology, as well as for other imaging modalities. Currently, histopathological imaging is being widely studied in the literature using ML algorithms. Contributions in this line should be directed towards the development of computer-aid systems capable of processing complete digitised biopsies without resorting to patches or local structures to aid decision making. This would represent a huge leap in quality and lead to more reliable models for application in clinical practice. A powerful hardware setting would be needed to cope with the large size of images (in the order of GB) and the large number of operations involved in this learning process. The CVBLab research group has recently acquired a hardware machine capable of supporting that computing power: the NVIDIA DGX A100 system, which integrates eight tensor cores with up to 320GB of GPU memory. The exclusive use of such a system would allow the training of the models using as input data the whole-slide images (WSIs). However, this solution would take up a significant part of the CVBLab's hardware, leaving the rest of the team's components without important resources. The centre's hardware infrastructure is expected to be upgraded to more computationally powerful individual experiments. Additionally, interesting contributions could also be made in developing unsupervised ML algorithms, similar to those carried out in Chapter 3, that can compete against the fully supervised approaches. It would suppose a breakthrough in cancer detection from histological images, as no prior annotation steps would be necessary, thus relieving some of the experts' workload. To address this research line, algorithms based on weakly-supervised segmentation, unsupervised anomaly detection or domain adaptation, among others, would be of interest.

Regarding glaucoma assessment, there is still much work to be done before an automated AI-based diagnosis system using only OCT data can be

implemented in ophthalmology departments. The proposed frameworks provide a reliable starting point to assist experts in decision-making. However, it would be necessary to significantly increase the number of samples in the training dataset, as well as to test the reliability of the proposed models with external databases. This is a long-standing problem, as there are no public OCT-specific datasets with glaucoma annotations. We have already started to work in this direction by compiling annotated databases for glaucoma detection. Currently, we are collaborating with the Polytechnic University of Cartagena to bring together the OCT databases with glaucoma annotations and make them public to contribute to the scientific community. Furthermore, one of the main challenges facing OCT-based glaucoma diagnosis is the discrimination of different stages of the disease. This is an arduous task because OCT only provides simple data with sparse and very localised information, making it difficult for CNNs to discover hidden patterns associated with different patterns of aggressiveness. In our works [64, 66], we propose some initial solutions to discern between different levels of glaucoma severity just from OCT-based data. In future works, it would be interesting to combine the OCT images with another kind of data, such as fundus images or indicator parameters like the visual field or intraocular pressure. It would provide a more reliable decision as it would be supported by the combination of different diagnostic tests with richer information when finding distinct levels of aggressiveness. Another major challenge lies in providing high-confidence glaucoma predictions from SD-OCT volumes. Our work [65] suggests a novel way to address this problem, but additional datasets containing other ethnic groups would be necessary to strengthen the reliability of the proposed model. This would add significant value to glaucoma diagnosis, as experts would draw on additional information that they are not using today due to the associated high workload. Finally, a relevant contribution could be made in the field of glaucoma diagnosis if the evolution of a specific patient would be traced. To this end, different OCT samples should be collected in several temporary instances for a particular patient. Then, by applying recurrent learning, e.g. LSTMs or transformers-based algorithms, it would be possible to determine a patient's prognosis taking into account its temporal evolution.

Merits

Journal papers

García, G., Colomer, A., & Naranjo, V. First-stage prostate cancer identification on histopathological images: Hand-driven versus automatic learning. Entropy, 21, 356 (2019).

García, G., Colomer, A., & Naranjo, V. Glaucoma detection from raw SD-OCT volumes: A novel approach focused on spatial dependencies. Computer Methods and Programs in Biomedicine, 200, 105855 (2021).

García, G., Del Amor, R., Colomer, A., Verdú-Monedero, R., Morales-Sánchez, J., & Naranjo, V. Circumpapillary OCT-focused hybrid learning for glaucoma grading using tailored prototypical neural networks. Artificial Intelligence in Medicine, 118, 102132 (2021).

García, G., Esteve, A., Colomer, A., Ramos, D., & Naranjo, V. A novel self-learning framework for bladder cancer grading using histopathological images. Computers in Biology and Medicine, 104932 (2021).

Berenguer-Vidal, R., Verdú-Monedero, R., Morales-Sánchez, J., Sellés-Navarro, I., Del Amor, R., **García, G.**, & Naranjo, V. Automatic segmentation of the retinal nerve fiber layer by means of mathematical morphology and deformable models in 2D optical coherence tomography imaging. Sensors, 23, 8027 (2021).

International conferences

García, G., Colomer, A., Naranjo, V., Peñaranda, F., & Sales, M. Á. *Identification of individual glandular regions using LCWT and machine learning techniques* in *International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)* (2018), 642-650.

García, G., Colomer, A., López-Mir, F., Mossi, J. M., & Naranjo, V. *Computer aid-system to identify the first stage of prostate cancer through deep-learning techniques* in *2019 27th European Signal Processing Conference (EUSIPCO)* (2019) 1-5.

García, G., del Amor, R., Colomer, A. & Naranjo, V. *Glaucoma Detection From Raw Circumpapillary OCT Images Using Fully Convolutional Neural Networks* in *2020 IEEE International Conference on Image Processing (ICIP)* (2020), 2526–2530.

García, G., Colomer, A. & Naranjo, V. *Analysis of Hand-Crafted and Automatic-Learned Features for Glaucoma Detection Through Raw Circumpapillary OCT Images* in *International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)* (2020), 156–164.

Silva-Rodríguez, J., Payá-Bosch, E., **García, G.**, Colomer, A., & Naranjo, V. *Prostate Gland Segmentation in Histology Images via Residual and Multi-resolution U-NET* in *International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)* (2020), 1-8.

García, G., Colomer, A., Verdú-Monedero, R., Dolz, J., & Naranjo, V. *A self-training framework for glaucoma grading in OCT B-scans* in *2021 29th European Signal Processing Conference (EUSIPCO)* (2021), 1281-1285.

National conferences

García, G., Colomer, A., Naranjo, V., Sales-Maicas, M.A. & García-Morata, F.J. *Comparación de estrategias de machine learning clásico y de deep learning para la clasificación automática de estructuras glandulares en imágenes histológicas de próstata* in *Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)* (2018), 357-360.

García, G., Colomer, A., Naranjo, V., Sales-Maicas, M.A. & García-Morata, F.J. *Desarrollo de un sistema automático para la segmentación de glándulas en imágenes histológicas de próstata* in *Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)* (2018), 345-348.

Research awards

García, G. (2019). Desarrollo de un sistema para la detección automática de glándulas cancerosas en imágenes histológicas de próstata. *XIV Premi d'Investigació Cientificotècnic "Ciutat d'Algemesí" per a joves investigadors.*

García, G. (2020). Multi-class deep learning approach to identify Covid-19 pneumonia from X-ray images. *Second prize of the "Best Oral Presentation Award 2020"* at the VI International Conference IMFAHE.

Silva-Rodríguez, J., Payá-Bosch, E., **García, G.**, Colomer, A., & Naranjo, V. (2020). Prostate Gland Segmentation in Histology Images via Residual and Multi-resolution U-NET. *Best Paper on Image Processing Award.*

Bibliography

1. Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine* **23,** 89–109 (2001).
2. Xu, J., Xue, K. & Zhang, K. Current status and future trends of clinical diagnoses via image-based deep learning. *Theranostics* **9,** 7556 (2019).
3. Silva-Rodríguez, J., Colomer, A., Sales, M. A., Molina, R. & Naranjo, V. Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine* **195,** 105637 (2020).
4. Silva-Rodriguez, J., Colomer, A., Dolz, J. & Naranjo, V. Self-learning for weakly supervised Gleason grading of local patterns. *IEEE Journal of Biomedical and Health Informatics* (2021).
5. Diaz-Pinto, A. *et al.* Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Transactions on Medical Imaging* **38,** 2211–2218 (2019).
6. Raja, H., Hassan, T., Akram, M. U. & Werghi, N. Clinically Verified Hybrid Deep Learning System for Retinal Ganglion Cells Aware Grading of Glaucomatous Progression. *IEEE Transactions on Biomedical Engineering* (2020).
7. Wang, S. & Summers, R. M. Machine learning and radiology. *Medical Image Analysis* **16,** 933–951 (2012).
8. McBee, M. P. *et al.* Deep learning in radiology. *Academic radiology* **25,** 1472–1480 (2018).
9. Montenegro, R. D., Oliveira, A. L., Cabral, G. G., Katz, C. R. & Rosenblatt, A. *A comparative study of machine learning techniques for caries prediction* in *2008 20th IEEE International Conference on Tools with Artificial Intelligence* **2** (2008), 477–481.

10. Khanagar, S. B. *et al.* Application and performance of artificial intelligence technology in forensic odontology-a systematic review. *Legal Medicine,* 101826 (2020).
11. Bori, L. *et al.* Novel and conventional embryo parameters as input data for artificial neural networks: an artificial intelligence model applied for prediction of the implantation potential. *Fertility and Sterility* **114,** 1232–1241 (2020).
12. Zaninovic, N. & Rosenwaks, Z. Artificial intelligence in human in vitro fertilization and embryology. *Fertility and Sterility* **114,** 914–920 (2020).
13. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
14. García-Floriano, A., Ferreira-Santiago, Á., Camacho-Nieto, O. & Yáñez-Márquez, C. A machine learning approach to medical image classification: Detecting age-related macular degeneration in fundus images. *Computers & Electrical Engineering* **75,** 218–229 (2019).
15. Grassmann, F. *et al.* A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology* **125,** 1410–1420 (2018).
16. Kim, S. J., Cho, K. J. & Oh, S. Development of machine learning models for diagnosis of glaucoma. *PloS one* **12,** e0177726 (2017).
17. Wu, H., Bailey, C., Rasoulinejad, P. & Li, S. Automated comprehensive adolescent idiopathic scoliosis assessment using MVC-Net. *Medical Image analysis* **48,** 1–11 (2018).
18. Lum, V. L. F., Leow, W. K., Chen, Y., Howe, T. S. & Png, M. A. *Combining classifiers for bone fracture detection in X-ray images* in *IEEE International Conference on Image Processing 2005* **1** (2005), 1–1149.
19. Rajpurkar, P. *et al.* Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv: 1711.05225* (2017).
20. Coupé, P., Manjón, J. V., Lanuza, E. & Catheline, G. Lifespan changes of the human brain in Alzheimer's disease. *Scientific reports* **9,** 1–12 (2019).
21. Ferrari, R. *et al.* MR-based artificial intelligence model to assess response to therapy in locally advanced rectal cancer. *European journal of radiology* **118,** 1–9 (2019).

22. Manjón, J. V. *et al.* pBrain: A novel pipeline for Parkinson related brain structure segmentation. *NeuroImage: Clinical* **25,** 102184 (2020).
23. Kadir, T. & Gleeson, F. Lung cancer prediction using machine learning and advanced imaging techniques. *Translational lung cancer research* **7,** 304 (2018).
24. Li, W., Cui, H., Li, K., Fang, Y. & Li, S. Chest computed tomography in children with COVID-19 respiratory infection. *Pediatric radiology,* 1 (2020).
25. Arbabshirani, M. R. *et al.* Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ digital medicine* **1,** 1–7 (2018).
26. Chauhan, N. K. & Singh, K. *A review on conventional machine learning vs deep learning* in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)* (2018), 347–352.
27. Arunkumar, N. *et al.* Fully automatic model-based segmentation and classification approach for MRI brain tumor using artificial neural networks. *Concurrency and Computation: Practice and Experience* **32,** e4962 (2020).
28. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25,** 1097–1105 (2012).
29. Chen, D. *et al.* Deep learning and alternative learning strategies for retrospective real-world clinical data. *NPJ digital medicine* **2,** 1–5 (2019).
30. De Bruijne, M. Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis* **33,** 94–97 (2016).
31. O'Mahony, N. *et al. Deep learning vs. traditional computer vision* in *Science and Information Conference* (2019), 128–144.
32. Arrieta, A. B. *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58,** 82–115 (2020).
33. Razzak, M. I., Naz, S. & Zaib, A. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps,* 323–350 (2018).
34. Del Amor, R. *et al.* An Attention-based Weakly Supervised framework for Spitzoid Melanocytic Lesion Diagnosis in Whole Slide Images. *Artificial Intelligence in Medicine,* 102197 (2021).

35. Anisuzzaman, D., Barzekar, H., Tong, L., Luo, J. & Yu, Z. A deep learning study on osteosarcoma detection from histological images. *Biomedical Signal Processing and Control* **69,** 102931 (2021).

36. Ohata, E. F. *et al.* A novel transfer learning approach for the classification of histological images of colorectal cancer. *The Journal of Supercomputing,* 1–26 (2021).

37. Ortega, S. *et al. Hyperspectral imaging and deep learning for the detection of breast cancer cells in digitized histological images* in *Medical Imaging 2020: Digital Pathology* **11320** (2020), 113200V.

38. Ker, J., Bai, Y., Lee, H. Y., Rao, J. & Wang, L. Automated brain histology classification using machine learning. *Journal of Clinical Neuroscience* **66,** 239–245 (2019).

39. Bussel, I. I., Wollstein, G. & Schuman, J. S. OCT for glaucoma diagnosis, screening and detection of glaucoma progression. *British Journal of Ophthalmology* **98,** ii15–ii19 (2014).

40. Morales, S., Engan, K. & Naranjo, V. Artificial intelligence in computational pathology–challenges and future directions. *Digital Signal Processing* **119,** 103196 (2021).

41. Nagpal, K. *et al.* Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA oncology* **6,** 1372–1380 (2020).

42. Steiner, D. F. *et al.* Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. *JAMA network open* **3,** e2023267–e2023267 (2020).

43. *What is computational pathology?* `https://www.philips.com.au/healthcare/sites/pathology/about/computational-pathology`. Accessed: 2021-11-24.

44. *Food and Drug Administration (FDA). Philips intellisite pathology solution* `https://www.accessdata.fda.gov/cdrh_docs/pdf16/DEN160056.pdf`. Accessed: 2021-11-24.

45. *Life Sciences Industrial Strategy. A report to the Government from the life sciences sector* `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650447/LifeSciencesIndustrialStrategy_acc2.pdf`. Accessed: 2021-11-24.

46. *The Royal College of Pathologists. Digital pathology* `https://www.rcpath.org/profession/digital-pathology.html`. Accessed: 2021-11-24.

47. *Innovative Medicines Initiative (IMI). Central repository of digital pathology slides to support the development of artificial intelligence tools H2020-JTI-IMI2-2019-18-01* `https://cordis.europa.eu/programme/id/H2020_IMI2-2019-18-01`. Accessed: 2021-11-24.

48. Yim, J. *et al.* Predicting conversion to wet age-related macular degeneration using deep learning. *Nature Medicine* **26,** 892–899 (2020).

49. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* **24,** 1342 (2018).

50. Maetschke, S. *et al.* A feature agnostic approach for glaucoma detection in OCT volumes. *PloS one* **14** (2019).

51. Dahrouj, M. & Miller, J. B. *Artificial Intelligence (AI) and Retinal Optical Coherence Tomography (OCT)* in *Seminars in Ophthalmology* **36** (2021), 341–345.

52. Wang, X. *et al.* Towards multi-center glaucoma OCT image screening with semi-supervised joint structure and function multi-task learning. *Medical Image Analysis* **63,** 101695 (2020).

53. Noury, E. *et al.* Detecting Glaucoma Using 3D Convolutional Neural Network of Raw SD-OCT Optic Nerve Scans. *arXiv preprint arXiv:1910.06302* (2019).

54. Thakoor, K. A., Li, X., Tsamis, E., Sajda, P. & Hood, D. C. *Enhancing the Accuracy of Glaucoma Detection from OCT Probability Maps using Convolutional Neural Networks* in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2019), 2036–2040.

55. Grossman, J. L. *et al.* Identifying patterns of glaucomatous progression in the circumpapillary retinal nerve fiber layer using OCT circle scans. *Investigative Ophthalmology & Visual Science* **62,** 1840–1840 (2021).

56. Studer, S. *et al.* Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction* **3,** 392–413 (2021).

57. García, G., Colomer, A., Naranjo, V., Peñaranda, F. & Sales, M. Á. *Identification of individual glandular regions using lcwt and machine learning techniques* in *International Conference on Intelligent Data Engineering and Automated Learning* (2018), 642–650.

58. García, G., Colomer, A., López-Mir, F., Mossi, J. M. & Naranjo, V. *Computer aid-system to identify the first stage of prostate cancer through deep-learning techniques* in *2019 27th European Signal Processing Conference (EUSIPCO)* (2019), 1–5.

59. Silva-Rodríguez, J., Payá-Bosch, E., García, G., Colomer, A. & Naranjo, V. *Prostate Gland Segmentation in Histology Images via Residual and Multi-resolution U-NET* in *International Conference on Intelligent Data Engineering and Automated Learning* (2020), 1–8.

60. García, G., Colomer, A. & Naranjo, V. First-stage prostate cancer identification on histopathological images: Hand-driven versus automatic learning. *Entropy* **21,** 356 (2019).

61. García, G., Esteve, A., Colomer, A., Ramos, D. & Naranjo, V. A novel self-learning framework for bladder cancer grading using histopathological images. *Computers in Biology and Medicine,* 104932 (2021).

62. García, G., del Amor, R., Colomer, A. & Naranjo, V. *Glaucoma Detection From Raw Circumpapillary OCT Images Using Fully Convolutional Neural Networks* in *2020 IEEE International Conference on Image Processing (ICIP)* (2020), 2526–2530.

63. García, G., Colomer, A. & Naranjo, V. *Analysis of Hand-Crafted and Automatic-Learned Features for Glaucoma Detection Through Raw Circumpapillary OCT Images* in *International Conference on Intelligent Data Engineering and Automated Learning* (2020), 156–164.

64. García, G. *et al.* Circumpapillary OCT-focused hybrid learning for glaucoma grading using tailored prototypical neural networks. *Artificial Intelligence in Medicine* **118,** 102132 (2021).

65. García, G., Colomer, A. & Naranjo, V. Glaucoma Detection from Raw SD-OCT Volumes: A Novel Approach Focused on Spatial Dependencies. *Computer Methods and Programs in Biomedicine* **200** (2021).

66. García, G., Colomer, A., Verdú-Monedero, R., Dolz, J. & Naranjo, V. A self-training framework for glaucoma grading in OCT B-scans. *arXiv preprint arXiv: 2111.11771* (2021).

67. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA: a cancer journal for clinicians* (2019).

68. Gleason, D. Histological grading and clinical staging of prostatic carcinoma. *Urologic pathology. The prostate* **171** (1977).

69. Esteban, Á., Colomer, A., Naranjo, V. & Sales, M. *Granulometry-Based Descriptor for Pathological Tissue Discrimination in Histopathological Images* in *2018 25th IEEE International Conference on Image Processing (ICIP)* (2018), 1413–1417.

70. Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N. J., Hermanns, T., *et al.* Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports,* 280024 (2018).

71. Antoni, S. *et al.* Bladder cancer incidence and mortality: a global overview and recent trends. *European urology* **71,** 96–108 (2017).

72. Lorenzo, L. *Valor pronóstico de la presencia de un componente tumoral indiferenciado ("tumor budding") en pacientes con carcinoma vesical músculo-invasivo* PhD thesis (2018).

73. Weinreb, R. N. & Khaw, P. T. Primary open-angle glaucoma. *The Lancet* **363,** 1711–1720 (2004).

74. Jonas, J. B. *et al.* Glaucoma–Authors' reply. *The Lancet* **391,** 740 (2018).

75. Wang, X. *et al. Unifying Structure Analysis and Surrogate-driven Function Regression for Glaucoma OCT Image Screening* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), 39–47.

76. Ran, A. R. *et al.* Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis. *The Lancet Digital Health* **1,** e172–e182 (2019).

77. SEOM. *Las cifras del cáncer en España* `https://seom.org/seomcms/images/stories/recursos/Las_Cifras_del_cancer_en_Espana2018.pdf`. Accessed: 25-10-2018. 2018.

78. Naik, S., Doyle, S., Feldman, M., Tomaszewski, J. & Madabhushi, A. *Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information* in *MIAAB workshop* (2007), 1–8.

79. Tabesh, A. *et al.* Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Transactions on Medical Imaging* **26,** 1366–1378 (2007).

80. Nguyen, K., Sabata, B. & Jain, A. K. Prostate cancer grading: Gland segmentation and structural features. *Pattern Recognition Letters* **33,** 951–961 (2012).

81. Farooq, M. T., Shaukat, A., Akram, U., Waqas, O., *et al. Automatic gleason grading of prostate cancer using Gabor filter and local binary patterns* in *Telecommunications and Signal Processing (TSP), 2017 40th International Conference on* (2017), 642–645.

82. Farjam, R., Soltanian-Zadeh, H., Jafari-Khouzani, K. & Zoroofi, R. A. An image analysis approach for automatic malignancy determination of prostate pathological images. *Cytometry Part B: Clinical Cytometry: The Journal of the International Society for Analytical Cytology* **72,** 227–240 (2007).

83. Kwak, J. T. & Hewitt, S. M. Multiview boosting digital pathology analysis of prostate cancer. *Computer Methods and Programs in Biomedicine* **142,** 91–99 (2017).

84. Diamond, J., Anderson, N. H., Bartels, P. H., Montironi, R. & Hamilton, P. W. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human pathology* **35,** 1121–1131 (2004).

85. Tai, S.-K., Li, C.-Y., Wu, Y.-C., Jan, Y.-J. & Lin, S. C. *Classification of prostatic biopsy* in *Digital Content, Multimedia Technology and its Applications (IDC), 2010 6th International Conference on* (2010), 354–358.

86. Doyle, S. *et al. Automated grading of prostate cancer using architectural and textural image features* in *4th IEEE international symposium on biomedical imaging: from nano to macro (ISBI)* (2007), 1284–1287.

87. Leo, P. *et al.* Stable and discriminating features are predictive of cancer presence and Gleason grade in radical prostatectomy specimens: a multi-site study. *Scientific reports* **8,** 14918 (2018).

88. Nguyen, K., Sarkar, A. & Jain, A. K. *Structure and context in prostatic gland segmentation and classification* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2012), 115–123.

89. Xia, T., Yu, Y. & Hua, J. *Automatic detection of malignant prostatic gland units in cross-sectional microscopic images* in *2010 17th IEEE International Conference on Image Processing (ICIP)* (2010), 1057–1060.

90. García, G., Colomer, A., Naranjo, V., Sales, M. & García-Morata, F. *Comparación de estrategias de machine learning cásico y de deep learning para la clasificación automática de estructuras glandulares en imágenes histológicas de próstata* in *XXXVI Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)* (2018), 357–360.

91. Beare, R. A locally constrained watershed transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28,** 1063–1074 (2006).

92. Monaco, J. P., Tomaszewski, J. E., Feldman, M. D., Hagemann, I., Moradi, M., *et al.* High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models. *Medical Image Analysis* **14,** 617–629 (2010).

93. Huang, P.-W. & Lee, C.-H. Automatic classification for pathological prostate images based on fractal analysis. *IEEE Transactions on Medical Imaging* **28,** 1037–1050 (2009).

94. Zhou, N., Fedorov, A., Fennessy, F., Kikinis, R. & Gao, Y. Large scale digital prostate pathology image analysis combining feature extraction and deep neural network. *arXiv preprint arXiv:1705.02678* (2017).

95. Del Toro, O. J. *et al. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score* in *Medical Imaging 2017: Digital Pathology* **10140** (2017), 101400O.

96. Otsu, N. A threshold selection method from gray-level histograms. *IEEE ransactions on Systems, Man, and Cybernetics* **9,** 62–66 (1979).

97. Hurst, H. E. Long term storage. *An experimental study* (1965).

98. Ruifrok, A. C., Johnston, D. A., *et al.* Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology* **23,** 291–299 (2001).

99. Gertych, A., Ing, N., Ma, Z., Fuchs, T. J., Salman, S., *et al.* Machine learning approaches to analyze histological images of tissues from radical prostatectomies. *Computerized Medical Imaging and Graphics* **46,** 197–208 (2015).

100. Mandelbrot, B. B. & Van Ness, J. W. Fractional Brownian motions, fractional noises and applications. *SIAM review* **10,** 422–437 (1968).

101. DiFranco, M., O'Hurley, G., Kay, E., Watson, W. & Cunningham, P. *Automated gleason scoring of prostatic histopathology slides using multi-channel co-occurrence texture features* in *Proceedings of international workshop on microscopic image analysis with applications in biology (MIAAB'08)* (2008).

102. Presutti, M. La matriz de co-ocurrencia en la clasificación multiespectral: tutorial para la enseñanza de medidas texturales en cursos de grado universitario. *4ª Jornada de Educação em Sensoriamento Remoto no Âmbito do Mercosul* (2004).

103. Ojala, T., Pietikäinen, M. & Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition* **29,** 51–59 (1996).

104. Ojala, T., Pietikainen, M. & Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24,** 971–987 (2002).

105. Pietikäinen, M., Ojala, T. & Xu, Z. Rotation-invariant texture classification using feature distributions. *Pattern Recognition* **33,** 43–52 (2000).

106. Guo, Z., Zhang, L. & Zhang, D. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing* **19,** 1657–1663 (2010).

107. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2,** 27 (2011).

108. Bishop, C. M. Pattern recognition. *Machine learning* **128** (2006).

109. Simonyan, K. & Zisserman, A. *Two-stream convolutional networks for action recognition in videos* in *Advances in neural information processing systems* (2014), 568–576.

110. Hoo-Chang, S. *et al.* Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging* **35,** 1285 (2016).

111. Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging* **35,** 1299–1312 (2016).

112. Wong, S. C., Gatt, A., Stamatescu, V. & McDonnell, M. D. *Understanding data augmentation for classification: when to warp?* in *2016 international conference on digital image computing: techniques and applications (DICTA)* (2016), 1–6.

113. Feil, G. & Stenzl, A. Pruebas de marcadores tumorales en el cáncer de vejiga. *Actas Urológicas Españolas* **30,** 38–45 (2006).

114. Sharma, S., Ksheersagar, P. & Sharma, P. Diagnosis and treatment of bladder cancer. *American family physician* **80,** 717–723 (2009).

115. Richards, K. A., Smith, N. D. & Steinberg, G. D. The importance of transurethral resection of bladder tumor in the management of nonmuscle invasive bladder cancer: a systematic review of novel technologies. *The Journal of urology* **191,** 1655–1664 (2014).

116. Jimenez, R. E. *et al.* Grading the invasive component of urothelial carcinoma of the bladder and its relationship with progression-free survival. *The American journal of surgical pathology* **24,** 980–987 (2000).

117. Stenzl, A. *et al.* Guía clínica sobre el cáncer de vejiga con invasión muscular y metastásico. *European Association of Urology* (2010).

118. Busch, C. & Algaba, F. The WHO/ISUP 1998 and WHO 1999 systems for malignancy grading of bladder cancer. Scientific foundation and translation to one another and previous systems. *Virchows Archiv* **441,** 105–108 (2002).

119. Almangush, A., Karhunen, M., Hautaniemi, S., Salo, T. & Leivo, I. Prognostic value of tumour budding in oesophageal cancer: a meta-analysis. *Histopathology* **68,** 173–182 (2016).

120. Karamitopoulou, E. *et al.* Tumour budding is a strong and independent prognostic factor in pancreatic cancer. *European journal of cancer* **49,** 1032–1039 (2013).

121. Masuda, R. *et al.* Tumor budding is a significant indicator of a poor prognosis in lung squamous cell carcinoma patients. *Molecular medicine reports* **6,** 937–943 (2012).

122. Fukumoto, K. *et al.* Tumor budding, a novel prognostic indicator for predicting stage progression in T1 bladder cancers. *Cancer science* **107,** 1338–1344 (2016).

123. Wetteland, R., Engan, K., Eftestøl, T., Kvikstad, V. & Janssen, E. A. Multiclass Tissue Classification of Whole-Slide Histological Images using Convolutional Neural Networks. *ICPRAM* **1,** 320–327 (2019).

124. Zhang, Z. *et al.* Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence* **1,** 236–245 (2019).

125. Lucas, M. *et al.* Deep Learning-based Recurrence Prediction in Patients with Non-muscle-invasive Bladder Cancer. *European Urology Focus* (2020).

126. Yin, P.-N. *et al.* Histopathological distinction of non-invasive and invasive bladder cancers using machine learning approaches. *BMC medical informatics and decision making* **20,** 1–11 (2020).

127. Dolz, J. *et al.* Multiregion segmentation of bladder cancer structures in MRI with progressive dilated convolutional networks. *Medical physics* **45,** 5482–5493 (2018).

128. Woerl, A.-C. *et al.* Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *European urology* **78,** 256–264 (2020).

129. Xu, H., Park, S., Lee, S. H. & Hwang, T. H. Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. *bioRxiv,* 554527 (2019).

130. Harmon, S. A. *et al.* Multiresolution application of artificial intelligence in digital pathology for prediction of positive lymph nodes from primary tumors in bladder cancer. *JCO clinical cancer informatics* **4,** 367–382 (2020).

131. Yang, Y., Zou, X., Wang, Y. & Ma, X. Application of Deep Learning as a Noninvasive Tool to Differentiate Muscle-invasive Bladder Cancer and Non-muscle-invasive Bladder Cancer with CT. *European Journal of Radiology,* 109666 (2021).

132. Ikeda, A. *et al.* Support system of cystoscopic diagnosis for bladder cancer based on artificial intelligence. *Journal of endourology* **34,** 352–358 (2020).

133. Yang, R. *et al.* Automatic recognition of bladder tumours using deep learning technology and its clinical application. *The international journal of medical robotics and computer assisted surgery,* e2194 (2020).

134. Cho, K., Van Merriënboer, B., Bahdanau, D. & Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

135. Prall, F., Nizze, H. & Barten, M. Tumour budding as prognostic factor in stage I/II colorectal carcinoma. *Histopathology* **47,** 17–24 (2005).

136. Lugli, A. *et al.* CD8+ lymphocytes/tumour-budding index: an independent prognostic factor representing a 'pro-/anti-tumour' approach to tumour host interaction in colorectal cancer. *British journal of cancer* **101,** 1382–1392 (2009).

137. Ogawa, T. *et al.* Tumor budding is predictive of lymphatic involvement and lymph node metastases in submucosal invasive colorectal adenocarcinomas and in non-polypoid compared with polypoid growths. *Scandinavian journal of gastroenterology* **44,** 605–614 (2009).

138. Zlobec, I., Bihl, M. P., Foerster, A., Rufle, A. & Lugli, A. The impact of CpG island methylator phenotype and microsatellite instability on tumour budding in colorectal cancer. *Histopathology* **61,** 777–787 (2012).

139. Brieu, N. *et al.* Automated tumour budding quantification by machine learning augments TNM staging in muscle-invasive bladder cancer prognosis. *Scientific reports* **9,** 1–11 (2019).

140. Guo, X., Liu, X., Zhu, E. & Yin, J. *Deep clustering with convolutional autoencoders* in *International conference on neural information processing* (2017), 373–382.

141. Guo, X., Zhu, E., Liu, X. & Yin, J. *Deep embedded clustering with data augmentation* in *Asian Conference on Machine Learning* (2018), 550–565.

142. Xie, J., Girshick, R. & Farhadi, A. *Unsupervised deep embedding for clustering analysis* in *International Conference on Machine Learning* (2016), 478–487.

143. Enguehard, J., O'Halloran, P. & Gholipour, A. Semi-supervised learning with deep embedded clustering for image classification and segmentation. *IEEE Access* **7,** 11093–11104 (2019).

144. Hershey, J. R., Chen, Z., Le Roux, J. & Watanabe, S. *Deep clustering: Discriminative embeddings for segmentation and separation* in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), 31–35.

145. Prasetio, B. H., Tamura, H. & Tanno, K. *A Deep Time-delay Embedded Algorithm for Unsupervised Stress Speech Clustering* in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (2019), 1193–1198.

146. Del Amor, R., Colomer, A., Monteagudo, C. & Naranjo, V. A Deep Embedded Refined Clustering Approach for Breast Cancer Distinction based on DNA Methylation. *arXiv preprint arXiv:2102.09563* (2021).

147. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. *Learning deep features for discriminative localization* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 2921–2929.

148. Colomer, A., Naranjo, V. & Fuentes, F. *GIGAVISION: Sistema para el marcado de regiones tumorales en imágnes histológicas gigapixel.*

149. Gidaris, S., Singh, P. & Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* (2018).

150. Patacchiola, M. & Storkey, A. Self-supervised relational reasoning for representation learning. *arXiv preprint arXiv:2006.05849* (2020).

151. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. *A simple framework for contrastive learning of visual representations* in *International Conference on Machine Learning* (2020), 1597–1607.

152. Ioffe, S. & Szegedy, C. *Batch normalization: Accelerating deep network training by reducing internal covariate shift* in *International Conference on Machine Learning* (2015), 448–456.

153. Zeiler, M. D. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).

154. Peng, X., Xiao, S., Feng, J., Yau, W.-Y. & Yi, Z. *Deep Subspace Clustering with Sparsity Prior* in *IJCAI* (2016), 1925–1931.

155. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9** (2008).

156. Masci, J., Meier, U., Cireşan, D. & Schmidhuber, J. *Stacked convolutional auto-encoders for hierarchical feature extraction* in *International conference on artificial neural networks* (2011), 52–59.

157. Tham, Y.-C. *et al.* Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology* **121,** 2081–2090 (2014).

158. National, G. A. U. Glaucoma: diagnosis and management (2017).

159. Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., *et al.* Optical coherence tomography. *Science* **254,** 1178–1181 (1991).

160. Medeiros, F. A. *et al.* Detection of glaucoma progression with stratus OCT retinal nerve fiber layer, optic nerve head, and macular thickness measurements. *Investigative Ophthalmology & Visual Science* **50,** 5741–5748 (2009).

161. Sinthanayothin, C. *et al.* Automated detection of diabetic retinopathy on digital fundus images. *Diabetic medicine* **19,** 105–112 (2002).

162. Walter, T. *et al.* Automatic detection of microaneurysms in color fundus images. *Medical Image Analysis* **11,** 555–566 (2007).

163. Lichter, P. R. Variability of expert observers in evaluating the optic disc. *Transactions of the American Ophthalmological Society* **74,** 532 (1976).

164. Varma, R., Steinmann, W. C. & Scott, I. U. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology* **99,** 215–221 (1992).

165. Jaffe, G. J. & Caprioli, J. Optical coherence tomography to detect and manage retinal disease and glaucoma. *American Journal of Ophthalmology* **137,** 156–169 (2004).

166. Kurmann, T. *et al. Fused Detection of Retinal Biomarkers in OCT Volumes* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), 255–263.

167. Hood, D. C. & Raza, A. S. On improving the use of OCT imaging for detecting glaucomatous damage. *British Journal of Ophthalmology* **98,** ii1–ii9 (2014).

168. Bizios, D., Heijl, A., Hougaard, J. L. & Bengtsson, B. Machine learning classifiers for glaucoma diagnosis based on classification of retinal nerve fibre layer thickness parameters measured by Stratus OCT. *Acta ophthalmologica* **88,** 44–52 (2010).

169. Medeiros, F. A., Jammal, A. A. & Thompson, A. C. From Machine to Machine: An OCT-Trained Deep Learning Algorithm for Objective Quantification of Glaucomatous Damage in Fundus Photographs. *Ophthalmology* **126,** 513–521 (2019).

170. An, G. *et al.* Glaucoma Diagnosis with Machine Learning Based on Optical Coherence Tomography and Color Fundus Images. *Journal of Healthcare Engineering* **2019** (2019).

171. Fang, L. *et al.* Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomedical optics express* **8,** 2732–2744 (2017).

172. Pekala, M. *et al.* Deep learning based retinal OCT segmentation. *Computers in Biology and Medicine* **114,** 103445 (2019).

173. Barella, K. A. *et al.* Glaucoma diagnostic accuracy of machine learning classifiers using retinal nerve fiber layer and optic nerve data from SD-OCT. *Journal of Ophthalmology* **2013** (2013).

174. Vidotti, V. G. *et al.* Sensitivity and specificity of machine learning classifiers and spectral domain OCT for the diagnosis of glaucoma. *European Journal of Ophthalmology* **23,** 61–69 (2013).

175. Xu, J. *et al.* Three-dimensional spectral-domain optical coherence tomography data analysis for glaucoma detection. *PloS one* **8,** e55476 (2013).

176. Ran, A. R. *et al.* Artificial intelligence deep learning algorithm for discriminating ungradable optical coherence tomography three-dimensional volumetric optic disc scans. *Neurophotonics* **6,** 041110 (2019).

177. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* **24,** 1342–1350 (2018).

178. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Computation* **9,** 1735–1780 (1997).

179. Jiang, J. *et al.* Predicting the progression of ophthalmic disease based on slit-lamp images using a deep temporal sequence network. *PloS one* **13** (2018).

180. Khan, A., Sohail, A., Zahoora, U. & Qureshi, A. S. A survey of the recent architectures of deep convolutional neural networks. *arXiv preprint arXiv:1901.06032* (2019).

181. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 770–778.

182. Graves, A. *et al.* A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31,** 855–868 (2008).

183. Sak, H., Senior, A. W. & Beaufays, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling (2014).

184. Yue-Hei Ng, J. *et al. Beyond short snippets: Deep networks for video classification* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), 4694–4702.

185. Calders, T. & Jaroszewicz, S. *Efficient AUC optimization for classification* in *European Conference on Principles of Data Mining and Knowledge Discovery* (2007), 42–53.

186. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA Ophthalmology* **316,** 2402–2410 (2016).

187. Burlina, P. M., Joshi, N., Pekala, M., Pacheco, K. D. & et al. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmology* **135,** 1170–1176 (2017).

188. Thompson, A. C., Jammal, A. A. & Medeiros, F. A. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Translational Vision Science & Technology* **9,** 42–42 (2020).

189. Leung, C. K.-S. *et al.* Retinal nerve fiber layer imaging with spectral-domain optical coherence tomography: patterns of retinal nerve fiber layer progression. *Ophthalmology* **119,** 1858–1866 (2012).

190. Medeiros, F. A., Zangwill, L. M., Bowd, C., Mansouri, K. & Weinreb, R. N. The structure and function relationship in glaucoma: implications for detection of progression and measurement of rates of change. *Investigative Ophthalmology & Visual Science* **53,** 6939–6946 (2012).

191. Ojima, T. *et al.* Measurement of retinal nerve fiber layer thickness and macular volume for glaucoma detection using optical coherence tomography. *Japanese Journal of Ophthalmology* **51,** 197–203 (2007).

192. Abd El-Naby, A. E., Abouelkheir, H. Y., Al-Sharkawy, H. T., Mokbel, T. H., *et al.* Correlation of retinal nerve fiber layer thickness and perimetric changes in primary open-angle glaucoma. *Journal of the Egyptian Ophthalmological Society* **111** (2018).

193. Ometto, G. *et al.* ReLayer: A free, online tool for extracting retinal thickness from cross-platform OCT images. *Translational Vision Science & Technology* **8,** 25–25 (2019).

194. Kamal Abdellatif, M. & Abdelmaguid Mohamed Elzankalony, Y. e. a. Outer retinal layers' thickness changes in relation to age and choroidal thickness in normal eyes. *Journal of Ophthalmology* **2019** (2019).

195. Hassan, T., Akram, M. U., Masood, M. F. & Yasin, U. Deep structure tensor graph search framework for automated extraction and characterization of retinal layers and fluid pathology in retinal SD-OCT scans. *Computers in Biology and Medicine* **105,** 112–124 (2019).

196. Gao, E. *et al.* Comparison of retinal thickness measurements between the topcon algorithm and a graph-based algorithm in normal and glaucoma eyes. *PloS one* **10,** 1–13 (2015).

197. Niu, S. *et al.* Automated retinal layers segmentation in SD-OCT images using dual-gradient and spatial correlation smoothness constraint. *Computers in Biology and Medicine* **54,** 116–128 (2014).

198. Kromer, R., Rahman, S., Filev, F. & Klemm, M. An Approach for Automated Segmentation of Retinal Layers In Peripapillary Spectralis SD-OCT Images Using Curve Regularisation. *Insights in Ophthalmology* **1,** 1–6 (2017).

199. Duan, W. *et al.* A generative model for OCT retinal layer segmentation by groupwise curve alignment. *IEEE Access* **6,** 25130–25141 (2018).

200. Devalla, S. K. *et al.* DRUNET: a dilated-residual U-Net deep learning network to segment optic nerve head tissues in optical coherence tomography images. *Biomedical optics express* **9,** 3244–3265 (2018).

201. Zang, P., Wang, J. & Hormel, T. T. e. a. Automated segmentation of peripapillary retinal boundaries in oct combining a convolutional neural network and a multi-weights graph search. *Biomedical optics express* **10,** 4340–4352 (2019).

202. Mariottoni, E. B. *et al.* Quantification of Retinal Nerve Fibre Layer Thickness on Optical Coherence Tomography with a Deep Learning Segmentation-Free Approach. *Scientific reports* **10,** 1–9 (2020).

203. Thompson, A. C., Jammal, A. A., Berchuck, S. I., Mariottoni, E. B. & Medeiros, F. A. Assessment of a segmentation-free deep learning algorithm for diagnosing glaucoma from optical coherence tomography scans. *JAMA Ophthalmology* **138,** 333–339 (2020).

204. Shehryar, T. *et al.* Improved automated detection of glaucoma by correlating fundus and SD-OCT image analysis. *International Journal of Imaging Systems and Technology* (2020).

205. Asaoka, R. *et al.* Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *American Journal of Ophthalmology* **198,** 136–145 (2019).

206. Leung, C. K. *et al.* Impact of age-related change of retinal nerve fiber layer and macular thicknesses on evaluation of glaucoma progression. *Ophthalmology* **120,** 2485–2492 (2013).

207. Iverson, S. M., Feuer, W. J., Shi, W., Greenfield, D. S., for Glaucoma Study Group, A. I., *et al.* Frequency of abnormal retinal nerve fibre layer and ganglion cell layer SDOCT scans in healthy eyes and glaucoma suspects in a prospective longitudinal study. *British Journal of Ophthalmology* **98,** 920–925 (2014).

208. Naghizadeh, F., Garas, A., Vargha, P. & Holló, G. Detection of early glaucomatous progression with different parameters of the RTVue optical coherence tomograph. *Journal of Glaucoma* **23,** 195–198 (2014).

209. Hammel, N. *et al.* Comparing the rates of retinal nerve fiber layer and ganglion cell–inner plexiform layer loss in healthy eyes and in glaucoma eyes. *American Journal of Ophthalmology* **178,** 38–50 (2017).

210. Na, J. H. *et al.* Detection of glaucoma progression by assessment of segmented macular thickness data obtained using spectral domain optical coherence tomography. *Investigative Ophthalmology & visual science* **53,** 3817–3826 (2012).

211. Na, J. H. *et al.* Rates and patterns of macular and circumpapillary retinal nerve fiber layer thinning in preperimetric and perimetric glaucomatous eyes. *Journal of Glaucoma* **24,** 278–285 (2015).

212. Na, J. H. *et al.* Detection of glaucomatous progression by spectral-domain optical coherence tomography. *Ophthalmology* **120,** 1388–1395 (2013).

213. Wessel, J. M. *et al.* Longitudinal analysis of progression in glaucoma using spectral-domain optical coherence tomography. *Investigative Ophthalmology & Visual Science* **54,** 3613–3620 (2013).

214. Xu, B. Y. *et al.* Deep learning classifiers for automated detection of gonioscopic angle closure based on anterior segment OCT images. *American Journal of Ophthalmology* **208,** 273–280 (2019).

215. Fu, H. *et al.* A deep learning system for automated angle-closure detection in anterior segment optical coherence tomography images. *American Journal of Ophthalmology* **203,** 37–45 (2019).

216. Li, F. *et al.* Automatic differentiation of Glaucoma visual field from non-glaucoma visual filed using deep convolutional neural network. *BMC Medical Imaging* **18,** 35 (2018).

217. Kucur, Ş. S., Holló, G. & Sznitman, R. A deep learning approach to automatic detection of early glaucoma from visual fields. *PloS one* **13,** e0206081 (2018).

218. Ahn, J. M. *et al.* A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PloS one* **13,** e0207982 (2018).

219. Zhen, Y., Wang, L., Liu, H., Zhang, J. & Pu, J. Performance assessment of the deep learning technologies in grading glaucoma severity. *arXiv preprint arXiv:1810.13376* (2018).

220. Serener, A. & Serte, S. *Transfer Learning for Early and Advanced Glaucoma Detection with Convolutional Neural Networks* in *2019 Medical Technologies Congress (TIPTEKNO)* (2019), 1–4.

221. Raja, H. *et al.* Data on OCT and fundus images for the detection of glaucoma. *Data in brief,* 105342 (2020).

222. Mills, R. P. *et al.* Categorizing the stage of glaucoma from pre-diagnosis to end-stage disease. *American Journal of Ophthalmology* **141,** 24–30 (2006).

223. Susanna Jr, R. & Vessani, R. M. Staging glaucoma patient: why and how? *The open ophthalmology journal* **3,** 59 (2009).

224. Snell, J., Swersky, K. & Zemel, R. *Prototypical networks for few-shot learning* in *Advances in neural information processing systems* (2017), 4077–4087.

225. Pan, Y. *et al. Transferrable prototypical networks for unsupervised domain adaptation* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), 2239–2247.

226. Gao, T., Han, X., Liu, Z. & Sun, M. *Hybrid attention-based prototypical networks for noisy few-shot relation classification* in *Proceedings of the AAAI Conference on Artificial Intelligence* **33** (2019), 6407–6414.

227. Sun, S., Sun, Q., Zhou, K. & Lv, T. *Hierarchical attention prototypical networks for few-shot text classification* in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), 476–485.

228. Fort, S. Gaussian prototypical networks for few-shot learning on omniglot. *arXiv preprint arXiv:1708.02735* (2017).

229. Boney, R. & Ilin, A. Semi-supervised few-shot learning with prototypical networks. *CoRR abs/1711.10856* (2017).

230. Lu, J., Cao, Z., Wu, K., Zhang, G. & Zhang, C. *Boosting few-shot image recognition via domain alignment prototypical networks* in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)* (2018), 260–264.

231. Wang, J. & Zhai, Y. *Prototypical siamese networks for few-shot learning* in *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)* (2020), 178–181.

232. Tatham, A. J. & Medeiros, F. A. Detecting structural progression in glaucoma with optical coherence tomography. *Ophthalmology* **124,** S57–S65 (2017).

233. Mayer, M. A., Hornegger, J., Mardin, C. Y. & Tornow, R. P. Retinal nerve fiber layer segmentation on FD-OCT scans of normal subjects and glaucoma patients. *Biomedical optics express* **1,** 1358–1383 (2010).

234. Khosla, P. *et al.* Supervised contrastive learning. *arXiv preprint arXiv:2004.11362* (2020).

235. Le-Khac, P. H., Healy, G. & Smeaton, A. F. Contrastive Representation Learning: A Framework and Review. *IEEE Access* (2020).

236. Petscharnig, S., Lux, M. & Chatzichristofis, S. *Dimensionality reduction for image features using deep learning and autoencoders* in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing* (2017), 1–6.

237. Song, J. Binary generative adversarial networks for image retrieval. *arXiv preprint arXiv:1708.04150* (2017).

238. Daoud, M. I., Saleh, A., Hababeh, I. & Alazrai, R. *Content-based image retrieval for breast ultrasound images using convolutional autoencoders: A feasibility study* in *2019 3rd International Conference on Bioengineering for Smart Technologies (BioSMART)* (2019), 1–4.

239. Simon, C., Koniusz, P. & Harandi, M. Projective subspace networks for few-shot learning (2018).

240. Simon, C., Koniusz, P., Nock, R. & Harandi, M. *Adaptive Subspaces for Few-Shot Learning* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 4136–4145.

241. Gholami, P., Roy, P., Parthasarathy, M. K. & Lakshminarayanan, V. OCTID: Optical coherence tomography image database. *Computers & Electrical Engineering* **81,** 106532 (2020).