IFCA Instituto de Física de Cantabria

CSIC Consejo Superior de Investigaciones Científicas

UC Universidad de Cantabria

Robust and explainable deep learning for computed tomography-based diagnostic systems

Aprendizaje profundo explicable y robusto para sistemas de diagnóstico basados en tomografía computarizada

Ph.D. thesis submitted for the degree of Doctor of Science and Technology at the University of Cantabria

Memoria presentada para acceder al título de Doctor en Ciencia y Tecnología por la Universidad de Cantabria

Author: Miriam Cobo Cano

Supervisors: Lara Lloret Iglesias
Wilson José dos Santos Silva

July, 2025

Abstract

Artificial intelligence is transforming healthcare, particularly the field of medical imaging, providing powerful tools to process very high dimensional data, discover novel patterns, and assist clinicians in their work. The wide adoption of these tools is currently hindered by challenges related to the data (standardization, interoperability, privacy), the algorithms (robustness, transparency, fairness, reproducibility), and the lack of rigorous clinical validation (limited large-scale, prospective, and multi-center evaluations). Addressing these barriers is essential to ensure that artificial intelligence systems are trustworthy and broadly applicable across diverse clinical settings.

This thesis is focused on enhancing computed tomography-based diagnostic systems with deep learning techniques. The first part presents the foundations of deep learning for computer vision applications, the fundamentals of physics in medical imaging, together with current challenges and limitations to the clinical adoption of artificial intelligence. Guidelines are proposed to promote standardization, reproducibility and fairness, in addition to increasing awareness among developers, researchers and healthcare professionals. The second part explores specific applications of deep learning, first, in intracranial hemorrhage prognosis and, subsequently, in lung cancer early diagnosis.

The work on intracranial hemorrhage prognosis leverages prior knowledge on clinical and demographic variables highly correlated with prognosis to enhance the robustness of image models through a multitask learning approach. The project on early lung cancer diagnosis presents a novel multimodal dataset integrating annotated screening low-dose computed tomography scans and plasma proteomics data generated by proximity extension assay. This dataset is a highly valuable research tool to develop or validate individualized risk prediction models that could significantly advance early lung cancer detection and intervention strategies. In addition to its planned public release to support open research, the curated dataset is used to assess the generalizability of deep learning models to predict the risk of malignancy in pulmonary nodules. Furthermore, a multimodal joint fusion approach integrating deep learning features extracted from the scans and protein biomarkers is proposed to enhance early lung cancer diagnosis using this novel screening dataset.

The contributions of this thesis highlight the inherent interdisciplinary nature of artificial intelligence in healthcare, and the potential of prior knowledge to enhance the robustness of deep learning algorithms. While artificial intelligence systems can process vast amounts of data, it is the clinical context, provided by healthcare professionals, that gives this information meaning and relevance. Close collaboration between clinicians, physicists, biologists and developers is essential to identify real clinical needs, harness the potential of artificial intelligence for personalized medicine, in addition to ensuring safe and reliable clinical translation.

Resumen

La inteligencia artificial está transformando la atención sanitaria, en particular el área de imagen médica, al proporcionar potentes herramientas para procesar datos multidimensionales, descubrir nuevos patrones y ayudar a los profesionales clínicos en su trabajo. En la actualidad, la adopción generalizada de estas herramientas se ve limitada por problemas relacionados con los datos (estandarización, interoperabilidad, privacidad), los algoritmos (robustez, transparencia, equidad, reproducibilidad) y la falta de validación clínica rigurosa (escasas evaluaciones a gran escala, prospectivas y multicéntricas). Abordar estos obstáculos es esencial para garantizar que los sistemas de inteligencia artificial sean fiables y ampliamente aplicables en diversos entornos clínicos.

Esta tesis se centra en la mejora de los sistemas de diagnóstico basados en tomografía computarizada con técnicas de aprendizaje profundo. En la primera parte se presentan los fundamentos del aprendizaje profundo para aplicaciones de visión artificial, los principios físicos de las imágenes médicas, junto con los retos y las limitaciones actuales para la adopción clínica de la inteligencia artificial. Se proponen guías para promover la estandarización, reproducibilidad y equidad, además de aumentar la concienciación entre desarrolladores, investigadores y profesionales sanitarios. La segunda parte investiga aplicaciones concretas del aprendizaje profundo, primero, en el pronóstico de hemorragias intracraneales y, posteriormente, en el diagnóstico precoz del cáncer de pulmón.

El trabajo en pronóstico de hemorragia intracraneal aprovecha la información complementaria de variables clínicas y demográficas altamente correlacionadas con el pronóstico para mejorar la robustez de los modelos de imagen mediante un enfoque de aprendizaje multitarea. El proyecto en diagnóstico precoz de cáncer de pulmón presenta un novedoso dataset multimodal que integra tomografía computarizada de baja dosis y datos proteómicos medidos en plasma generados por un ensayo de extensión de proximidad. Este conjunto de datos es una herramienta de investigación muy valiosa para desarrollar o validar modelos individualizados de predicción del riesgo, que podrían suponer un avance significativo en las estrategias de detección precoz e intervención del cáncer de pulmón. Además de su próxima publicación para apoyar la investigación en abierto, el conjunto de datos se utiliza para evaluar la capacidad de generalización de modelos de aprendizaje profundo en la predicción del riesgo de malignidad en nódulos pulmonares. Asimismo, se presenta una estrategia de fusión multimodal que integra características de aprendizaje profundo extraídas de las imágenes y biomarcadores proteicos para mejorar el diagnóstico precoz del cáncer de pulmón utilizando este novedoso dataset.

Entre las contribuciones de esta tesis destacan la naturaleza interdisciplinar inherente a la inteligencia artificial en la atención sanitaria, y el potencial del conocimiento a priori para mejorar la robustez de los algoritmos de aprendizaje profundo. Aunque los sistemas de inteligencia artificial pueden procesar grandes cantidades de datos, es el contexto clínico, proporcionado por los profesionales sanitarios, el que da sentido y relevancia a esta información. La estrecha colaboración entre clínicos, físicos, biólogos y desarrolladores es esencial para identificar necesidades clínicas reales, aprovechar el potencial de la inteligencia artificial para la medicina personalizada, además de asegurar una traslación clínica segura y fiable.

Acknowledgements

Me gustaría comenzar dando las gracias a mi directora de tesis Lara, por su apoyo constante, dedicación y supervisión. Ha sido una suerte trabajar contigo estos años y aprender no solo de una gran investigadora, sino también de una bellísima persona. I would like to thank Wilson, my co-supervisor, who also supervised me during my research stays at the Netherlands Cancer Institute and Utrecht University, and who kindly welcomed me as part of his group. Under his guidance and support, I was able to engage in very valuable discussions with the group, which motivated me to explore new perspectives. Quiero agradecer a todas las personas que me han acompañado y ayudado durante este periodo, a los profesores que me inspiraron para seguir por este camino y, en especial, a Lalo y a Marcos.

Quiero dar las gracias a mis amigos, que han estado apoyándome durante estos años de doctorado. Ellos saben quiénes son, aunque he decidido no nombraros aquí uno a uno y decíroslo en persona, sin que ello reste importancia a cada uno de vosotros, pero evitando la presión de encontrar una frase a la altura o escribir la típica frase genérica. Gracias a vosotros, tengo la suerte de contar con personas increíbles a mi lado. Aun así, tengo que mencionar a una artista, Nuria, muchas gracias por compartir tus obras conmigo, y a una gran diseñadora, Andreea. I would like to thank all the people I met in Amsterdam and Utrecht, where I also had the opportunity to get to know amazing people. This thesis is dedicated to a very special person, Gab, you are always in my heart, forever.

Quiero dar las gracias de corazón a mi familia, a mis padres y mi hermano, sois mi apoyo incondicional. Esta tesis os la dedico a vosotros, os quiero mucho.

Contents

Glossary

The following glossary defines terms and acronyms used throughout this thesis:

Adam	Adaptive Moment Estimation (optimizer)
AdamW	Adaptive Moment Estimation decoupling Weight decay regularization (optimizer)
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
BINN	Biologically Informed Neural Network
CAD	Computer-aided diagnosis
CapsNet	Capsule network
CIMA	Center of Applied Medical Research, Navarra, Spain
CNN	Convolutional Neural Network
CSF	Cerebrospinal Fluid
CT	Computed Tomography
CUN	Clínica Universitaria de Navarra, Navarra, Spain
CV	Cross-validation
DECT	Dual-Energy Computed Tomography
DenseNet	Densely Connected Convolutional Network
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DMP	Data Management Plan
DNN	Dense Neural Network
DSC	Dice Similarity Coefficient
DTC	Decision Tree Classifier
EIBALL	European Imaging Biomarker ALLiance
ELISA	Enzyme-Linked ImmunoSorbent Assay (immunological test)
EM	Electromagnetic
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable (principles for data sharing)
FBP	Filtered Back Projection
FC	Fully Connected
FDA	U.S. Food and Drug Administration
FM	Foundational Models
FP	False Positive
FPR	False Positive Rate
FUTURE-AI	Fairness, Universality, Traceability, Usability, Robustness, and Explainability

	(structured framework for trustworthy and ethical AI in healthcare)
GAN	Generative Adversarial Network
GCS	Glasgow Coma Scale (clinical variable that describes the extent of impaired consciousness in all types of acute medical and trauma patients)
GDPR	General Data Protection Regulation
gLMs	genome Language Models
GNN	Graph Neural Network
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
HU	Hounsfield Unit
HUCA	Hospital Universitario Central de Asturias, Asturias, Spain
HUMV	Hospital Universitario Marqués de Valdecilla, Cantabria, Spain
IBSI	Image Biomarker Standardization Initiative
ICH	Intracranial Hemorrhage
ID	Identifier
INTEGRAL	Integrative Analysis of Cancer Risk and Etiology (research consortium program)
IPN	Indeterminate risk Pulmonary Nodule
IQR	Interquartile Range
IR	Interventional Radiology
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	Least Absolute Shrinkage and Selection Operator
LC	Lung Cancer
LDCT	Low-Dose Computed Tomography
LEON	Complejo Asistencial Universitario de León, León, Spain
LLM	Large Language Model
LLP	Liverpool Lung cancer Project
LR	Logistic Regression
Lung-RADS	Lung imaging Reporting and Data System
MAE	Mean Absolute Error
Mayo-PET	Mayo risk model incorporating Positron Emission Tomography
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MT	Multitask (referred to a multitask predictive model)
NIfTI	Neuroimaging Informatics Technology Initiative (file format)
NLP	Natural Language Processing
NLST	National Lung Screening Trial
NOS	Not Otherwise Specified
NPV	Negative Predictive Value
NPX	Normalized Protein eXpression
NRRD	Nearly Raw Raster Data (file format)
NRs	Neuroradiologists
OCT	Optical Coherence Tomography
ODE	Ordinary Differential Equation
PCCT	Photon-Counting Computed Tomography
PCR	Polymerase Chain Reaction
PDE	Partial Differential Equation
PEA	Proximity Extension Assay

P-ELCAP	Pamplona-Early Lung Cancer Action Program screening cohort
PET	Positron Emission Tomography
PIML	Physics-Informed Machine Learning
PINN	Physics-Informed Neural Network
P5 medicine	Predictive, preventive, personalized, participatory, psycho-cognitive medicine
PM	Personalized/Precision medicine
PNG	Portable Network Graphics (file format)
PPV	Positive Predictive Value
pre-LC	Lung cancer participants in previous years to lung cancer diagnosis
QA	Quality Assurance
QC	Quality Control
QIBA	Quantitative Imaging Biomarkers Alliance
QIN	Quantitative Imaging Network of the National Institute of Health, U.S.
qMRI	quantitative MRI
QWK	Quadratic Weighted Cohen Kappa score
ReLU	Rectified Linear Unit
ResNet	Residual Network
RF	Random Forest
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
RSF	Random Survival Forest
SD	Standard Deviation
SHAP	SHapley Additive exPlanations
SNR	Signal-to-Noise Ratio
SOTA	State-of-the-art
SPECT	Single-photon Emission Computed Tomography
T	Tesla (unit used to measure the strength of a magnetic field)
TE	Echo Time (in magnetic resonance imaging)
TNM	Tumor (T), Node (N), and Metastasis (M) (staging system for cancer)
TPR	True Positive Rate
TR	Repetition Time (in magnetic resonance imaging)
t-SNE	T-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
UOC	Uniform Ordinal Classification index
US	Ultrasound
ViT	Vision Transformer
WSI	Whole-Slide Imaging
XAI	Explainable artificial intelligence
YI	Youden's index

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Context and motivation

The development of artificial intelligence (AI) applications in healthcare has experienced unprecedented expansion in the last few decades. Computer-aided diagnostic (CAD) systems have demonstrated significant potential in automating and supporting the workflow of medical professionals, particularly in disease diagnosis and staging. Their growing impact is evident in digital imaging specialties such as radiology, where CAD systems enhance the interpretation of medical images, while reducing radiologists' workload and inter-reader variability. This thesis focuses on enhancing deep learning (DL) techniques for computed tomography (CT), which plays a crucial role in clinical decision-making. In particular, this research has explored the fields of oncology, pulmonology and neurology.

A key principle of translational medicine is the integration of diverse data sources, such as molecular biology, genetics, and medical imaging, to enhance disease diagnosis and treatment strategies. Personalized medicine (PM) leverages this multimodal approach to tailor treatments based on individual biomarkers. However, the complexity of patient data demands advanced techniques capable of processing different sources of data, i.e., machine learning (ML) and DL algorithms, to increase diagnostic accuracy by identifying meaningful data relationships. This multimodal strategy is essential for improving early detection, optimizing treatments, and driving better patient outcomes. In particular, this thesis focuses on lung cancer, which is one of the most commonly diagnosed cancers, and a leading cause of cancer-related mortality worldwide. The high prevalence on both men and women underscores the urgent need for improved diagnostic strategies to enhance patient survival. The integration of conventional screening low-dose computed tomography (LDCT) scans and detailed profiling of the blood proteome in novel multimodal AI algorithms can provide deeper insights into disease mechanisms in lung cancer screening population, facilitating early detection and individual risk assessment. Although lung cancer is the primary focus, this thesis also explores and contributes to a broader understanding of the challenges that exist for the full development of trustworthy AI algorithms in healthcare. Additionally, the work on this thesis describes a method to enhance the robustness and interpretability of image-based DL techniques in intracranial hemorrhage prognosis.

1.2 Objectives and document structure

The goal of this thesis is to study and develop robust machine learning and computer vision techniques to enhance radiological clinical decision making, and complement current image-based methods with explainable AI techniques, medical prior knowledge and multimodal approaches.

This document presents the main contributions related to the previous goals, and is divided into the following chapters:

- Chapter 2: provides an overview of deep learning focused on image analysis, with a particular emphasis on convolutional neural networks, transformers, and capsule networks, which are discussed and compared as state-of-the-art approaches in computer vision.
- Chapter 3: presents the fundamentals of physics in medical imaging for the main modalities in the clinics, and includes an overview of physics-informed machine learning methods.
- Chapter 4: delves into deep learning in medical imaging, first introducing conventional hand-crafted radiomics. Subsequently, the main limitations of both conventional radiomics and deep learning systems are explained. Next, guidelines are proposed to standardize medical imaging workflows. Finally, the chapter concludes with the concepts of explainability and trustworthy AI.
- Chapter 5: focuses on the specific challenges regarding clinical informatics data preparation and preprocessing, together with their impact on reproducibility and data leakage. To address current limitations, recommendations and guidelines are given to standardize reporting on data preparation stages in clinical informatics.
- Chapter 6: presents a multi-task learning approach to enhance intracranial hemorrhage prognosis. The method leverages prior knowledge on clinical and demographic variables highly correlated with prognosis to introduce them in a multi-task CT image-based deep learning model. Interpretability saliency maps are clinically assessed by a neuroradiologist.
- Chapter 7: describes a multimodal dataset for personalized LDCT-based lung cancer screening research collected and curated as part of this thesis. This dataset will be released in Zenodo to support research on lung cancer screening cohorts and advance the field of personalized medicine through AI-driven multimodal models in the context of LDCT-based lung cancer screening.
- Chapter 8: explores the generalizability of state-of-the-art 2D and 3D deep learning models within a medical informed framework for lung nodule malignancy risk assessment.
- Chapter 9: presents a novel multimodal approach to enhance early lung cancer diagnosis combining image-based lung nodule risk assessment with plasma proteomics biomarkers. Performance is compared with existing lung cancer risk prediction tools.
- Conclusions: summarizes the overall contributions, evaluates the extent to which the proposed objectives have been achieved, and outlines directions for future research.

1.3 Research outcomes

1.3.1 Research papers

The work of this PhD thesis has culminated in several papers published in high impact journals and proceedings of international conferences, as detailed in Table 1.1.

Table 1.1: Details of published papers.

Article	**Impact Factor**	**Cites**
Cobo, Miriam, *et al.* "Multi-task Learning Approach for Intracranial Hemorrhage Prognosis." *International Workshop on Machine Learning in Medical Imaging.* Cham: Springer Nature Switzerland, 2024. p. 12-21.	-	-
Fernández-Miranda, P. M., Fraguela, E. M., de Linera-Alperi, M. Á., **Cobo, Miriam**, *et al.* "A retrospective study of deep learning generalization across two centers and multiple models of X-ray devices using COVID-19 chest-X rays." *Sci Rep* **14**, 14657 (2024).	81.9	4
Cobo, Miriam, *et al.* "Enhancing radiomics and Deep Learning systems through the standardization of medical imaging workflows." *Sci Data* **10**, 732 (2023).	88.4	43
Cobo, Miriam, *et al.* "Novel deep learning method for coronary artery tortuosity detection through coronary angiography." *Sci Rep* **13**, 11137 (2023).	81.7	1

There are also a few papers under revision or planned to be submitted to high impact journals and proceedings of international conferences, as detailed in Table 1.2.

In addition, one dissemination book was published: Miriam Cobo Cano and Lara Lloret Iglesias. *Inteligencia artificial y medicina.* Editorial CSIC, 2023 (ISBN-10: 8413527252).

1.3.2 Conferences and workshops

The author of this thesis has co-organized the tutorial *Think your deep learning model works? Think again!* with Pietro Vischia (University of Oviedo) and Lara Lloret Iglesias, accepted at the European Conference on Artificial Intelligence (ECAI 2025). The author was also invited as a speaker in the Workshop *Artificial Intelligence in Biology* celebrated in the National Centre for Biotechnology (CNB-CSIC) in Madrid on February 14th, 2025. The talk was entitled *Trustworthy Artificial Intelligence for Multimodal Medical Data.* Additionally, the poster from the author *Reliable machine learning algorithms for medical imaging* presented at "X Jornadas Doctorales del G-9 y V Jornadas de Divulgación" in June 2023 organized by "Grupo 9 de Universidades (G-9)" received the Jury's Prize for the best poster in the Science area.

Table 1.2: Details of current papers under review or planned to be submitted.

Article	Status
Cobo, Miriam, *et al.* “Physical foundations for trustworthy medical imaging: a survey for artificial intelligence researchers.”	Under review in *Artificial Intelligence in Medicine.*
Cobo, Miriam, *et al.* “Applying the FAIR principles in clinical informatics data preprocessing for artificial intelligence algorithms.”	Planned to be submitted to *Sci Data.*
Cobo, Miriam, *et al.* “A novel open access multimodal dataset of nodule imaging and circulating proteome from a lung cancer screening cohort.”	Planned to be submitted to *Radiology: Artificial Intelligence.*
Cobo, Miriam, *et al.* “Validating a medical informed deep learning ordinal approach for lung nodule assessment.”	Planned to be submitted to *IEEE International Symposium on Biomedical Imaging 2026.*

1.4 Collaborations

Research in AI for medical applications is highly interdisciplinary, thus, collaborations are crucial to have access to high quality annotated datasets, understand the characteristics of the data from different modalities, and consequently design AI algorithms that adapt to their unique features. In addition, clinical validation and feedback are essential to build trustworthy AI systems. The work presented in this thesis is the result of several research collaborations with researchers and clinicians from other institutions.

1.4.1 Clínica Universitaria de Navarra and Center of Applied Medical Research

The main work of this thesis has been done in collaboration with Clínica Universitaria de Navarra (CUN) and Center of Applied Medical Research (CIMA), which provided the dataset P-ELCAP aimed at enhancing early lung cancer diagnosis in a screening population from a multimodal perspective. Among many others, the following researchers led or were particularly involved in the project:

- Prof. Luis Montuenga Badía, who was awarded the first Lung Ambition Alliance Prize for Research on Lung Cancer Early Detection (01/11/2021) as Principal Investigator. He has supervised the work since the beginning, from the dataset collection at CUN to contacting external institutions for external validation datasets, and supporting the research at all stages. He has provided translational research, clinical background supervision and feedback for all results, articles and applications we developed.
- Prof. Gorka Bastarrika Alemañ, who has supervised the LDCT collection and provided clinical feedback from the radiological point of view.
- Dr. Diego Serrano Tejero, who has collaborated particularly in the proteomics analysis, and provided clinical feedback for all results, articles and applications we developed.

1.4.2 Utrecht University and Netherlands Cancer Institute

The author of this thesis performed research stays at Utrecht University and the Netherlands Cancer Institute. The following researchers contributed to the progress made during this doctoral thesis:

- Prof. Wilson Silva, who co-supervised this thesis, providing technical supervision and feedback for all results, articles and applications we developed. The biweekly group meetings with external speakers fostered valuable discussions and the exchange of novel ideas among the group members, who actively participated and engaged in the meetings.
- Prof. Sanne Abeln, who leads the AI Technology for Life group at Utrecht University. The weekly group meetings provided very insightful discussions that enhanced the knowledge of AI in life sciences and multi-omics, while also encouraging innovative ideas. This was possible thanks to the active participation of fellow researchers, creating a welcoming, stimulating and supportive research environment.

1.4.3 Liverpool University

As part of this thesis, a validation dataset for lung cancer early diagnosis was obtained through a collaboration with Liverpool University. The following researchers were instrumental in enabling this data sharing agreement:

- Prof. John Field, who leads the Liverpool Lung cancer Project (LLP) and authorized the use of LLP as an external validation cohort.
- Dr. Michael Davies, who has supported the preparation of the LLP cohort and the establishment of the institutional agreement.

The LLP cohort is a lung cancer screening population that supports external validation under blinded conditions.

1.4.4 University hospitals and other institutions

Several radiologists contributed to this thesis by collecting datasets, annotating medical images, providing extensive clinical feedback, and participating in the research articles: Dr. Amaia Pérez del Barrio (Hospital Reina Sofía, Tudela, Spain), Dr. Pablo Menéndez Fernández Miranda (Hospital Universitario Rey Juan Carlos, Móstoles, Spain), Dr. Pablo Sanz Bellón (Hospital Universitario Marqués de Valdecilla, Santander, Spain), and M.D. David Corral Fontecha (Complejo Asistencial Universitario de León).

Additionally, a project proposal was submitted to the United States National Cancer Institute to obtain access to the National Lung Screening Trial (NLST) dataset, which was subsequently approved for use in the lung cancer project.

1.5 Funding

This work was funded by the Ministry of Education of Spain through the FPU (Training programme for Academic Staff) grant with reference FPU21-04458, and the FPU grant for research stays EST24/00533, as well as the project AI4EOSC "Artificial Intelligence for the European Open Science Cloud", that has received funding from the European Union's Horizon

Europe research and innovation programme under grant agreement number 101058593. Funding for the collaboration with CIMA-CUN has also been allocated from the Lung Ambition Alliance Prize for Research on Lung Cancer Early Detection (01/11/2021-01/11/2024).

Chapter 2

Overview of deep learning in image analysis

2.1 Introduction

The idea of using machines to replicate intelligent behavior and reasoning was first proposed by Alan Turing in 1950. In his seminal work "Computing Machinery and Intelligence", Turing introduced what would later be called the *Turing Test*, a method to measure the ability of a machine to exhibit intelligent behavior indistinguishable from that of a human [7]. In this test, a human evaluates natural language conversations with a real human and a machine, without knowing which is which, and has to determine which participant is the machine based solely on its responses. If the machine can convincingly mimic human replies, it is considered to have passed the test.

The term *artificial intelligence* (AI) was introduced during the Dartmouth Summer Workshop in 1956, where it was broadly referred to as *thinking machines* [8]. AI can be defined as the ability of a machine or computational model to recognize or *learn* patterns and relationships from representative data (training data), and apply this knowledge to make informed decisions on new, previously unseen inputs (test data) [8, 9]. Since the 1950s, AI has undergone significant growth, evolving over the past decades into increasingly complex algorithms, and creating new opportunities in many domains. This thesis focuses on one of the most promising areas of application: healthcare, with a particular emphasis on medical imaging and multimodal AI. From a broader perspective, AI can be divided into two main approaches: symbolic AI and machine learning (ML) [10], which will be defined throughout the next section.

In this chapter, we provide an overview on the main developments of AI since the beginnings until today, focusing on deep learning (DL), the branch within ML that has enabled many of the recent advances in computer vision and image analysis applied to medical imaging.

2.2 Learning paradigms

In the early beginnings, *symbolic* or *rule-based* AI was the dominant paradigm, encoding expert knowledge into explicit if–then rules and decomposing complex tasks into manually crafted subroutines. This approach consists in incorporating human knowledge and rules of behavior, acquired through experience, into computer programs, so-called *expert systems* [9]. The objective task is divided into smaller, simpler tasks and these are programmed manually. This paradigm was very successful in solving problems where the rules were very clear. In medical diagnosis, MYCIN was one of the earliest and most influential systems [11]. Developed in the 1970s, MYCIN is a computer-based expert system which leveraged a knowledge base of bacterial-infection rules to assist physicians in antibiotic treatment. In the field of medical imaging, pioneering computer-aided detection tools applied rule-based logic to mammograms to support radiologists, culminating in the 1998 FDA clearance of the first commercial CAD system, *ImageChecker M1000* [12]. This system detects microcalcifications by analyzing their brightness and size, and identifies larger masses by examining patterns of dense regions. The aforementioned achievements highlighted the strengths of explicit knowledge representation, but also revealed a critical limitation: rigid rule sets struggled to adapt to unforeseen scenarios. Such inflexibility paved the way for statistical ML and, subsequently, DL approaches, which offered more scalable, data-driven solutions for complex pattern recognition in medical imaging. Nevertheless, rule-based systems are still used today in certain medical imaging applications, particularly in the field known as radiomics, which will be explained in Section 4.2.

The revolution came with digitalization and the increase in computational power, in particular high-performance graphical processing units (GPUs) and distributed computing, which facilitated the processing of complex algorithms and large-scale data [13]. The MP-neuron, introduced by McCulloch and Pitts in 1943, represented one of the earliest models of artificial neurons and was used to simulate logic gates in early computational systems [14]. Their work served as a starting point for subsequent developments in neural network research in the following years. Backpropagation was first introduced in the 1970s, although it did not become widely known until a few years later, in the work published by Rumelhart, Hinton, and Williams in 1986 [15, 16]. This work laid the foundations of ML, especially regarding neural networks and supervised learning. In the ML paradigm, a computer system or model *learns* or improves its predictive performance by detecting patterns and correlations within the data and iteratively refining its internal representations based on feedback from previous iterations [9]. ML techniques can be categorized into three main groups: supervised learning (models learn from labeled data to predict outcomes, e.g., classification, detection or segmentation tasks), unsupervised learning (models create pseudo-labels from data to learn representations, e.g., generative models), and reinforcement learning (agents learn optimal actions by interacting with an environment, e.g., AlphaGo).

Within ML, the most powerful techniques and algorithms have emerged from DL, characterized by the use of neural networks with multiple layers, which allow for the hierarchical extraction of features from data. Its development has been driven by the growth in computational capacity, particularly the use of GPUs, the availability of large annotated datasets, along with continuous architectural and algorithmic improvements. Figure 2.1 shows the most relevant milestones since the 1950s until the current state of the art in DL, with particular attention to techniques applied to image analysis. Many of these models and architectures are explored in the next chapters of this thesis on medical imaging applications. The following section provides a brief overview of the main architectures and key breakthroughs in DL for computer vision and image analysis.

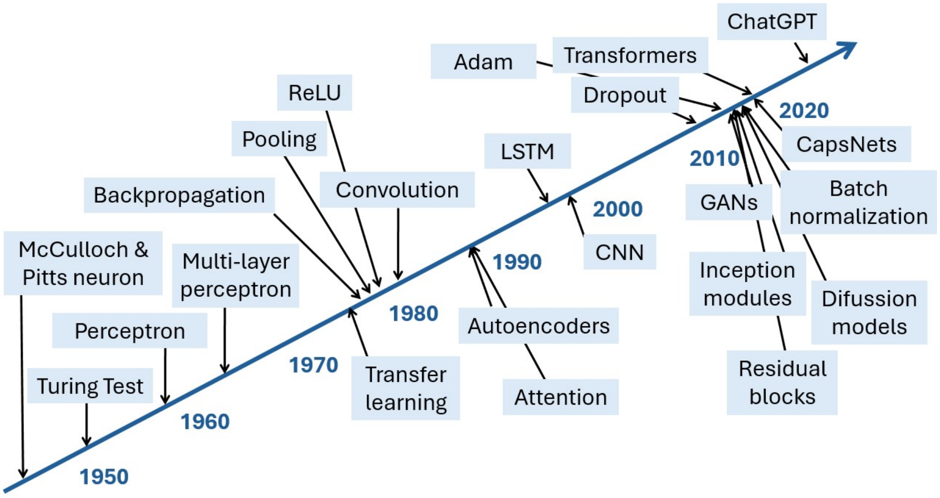


Figure 2.1: Milestones in the history of artificial intelligence, deep learning and neural networks. Adapted from Lones [4].

2.3 Deep learning in image applications

2.3.1 Convolutional neural networks

Convolutional Neural Networks (CNNs) are a hierarchical and cascade model that consists of convolutional layers, pooling layers and fully connected layers. Within convolutional layers, local feature maps are extracted by the convolutional operation of filters, which are then passed to the next layers to extract higher-level features [17]. Pooling layers reduce the dimension of the feature maps with a down-sampling operation (max, min or average pooling), enabling a considerable reduction in the number of computations needed [14]. The fully connected layers come after the final convolutional layer and transform the feature map into a one-dimensional vector to give the final output. As in any neural network the activation function introduces the non-linearity [18].

Weight sharing is an essential property of convolutional layers, allowing the same kernels to scan across the entire image. This approach not only significantly reduces the number of trainable parameters compared to fully connected layers, but also allows translation equivariance. Consequently, a spatial shift in the input of a convolutional layer results in a corresponding shift in the output feature maps, preserving the spatial structure of the input, as the convolutional operation applies the same filters uniformly across all spatial locations [6]. Subsampling techniques in CNNs, such as max-pooling, are typically used after the convolutional layer to reduce the spatial resolution of feature maps while preserving the most relevant activations. As a result, the output of a pooling unit remains unchanged regardless of the exact position of a feature within its pooling window, leading to local invariance to input translations in the activations in the next layer [6]. While pooling operations contribute to better generalization by discarding redundant information and reducing spatial complexity, they also lose fine details in the image that may be critical for certain tasks [19]. The combination of convolutional operations and pooling enables CNNs to learn hierarchical representations, as successive layers extract increasingly abstract representations of the input, from low-level textures and edges in the first layers to high-level semantic structures in the last ones.

CNNs are widely used in medical imaging for several tasks [17], such as classification [5, 20], segmentation [21], or detection. Transfer learning is a typically employed technique to optimize the learning task. This approach leverages a pretrained base model and, depending on the specific application, either retrains the entire network or freezes the initial layers while fine-tuning the final ones to adapt to the target task. Figure 2.2 shows an example of transfer learning in coronary artery tortuosity detection from ImageNet [22], the largest natural image database available.

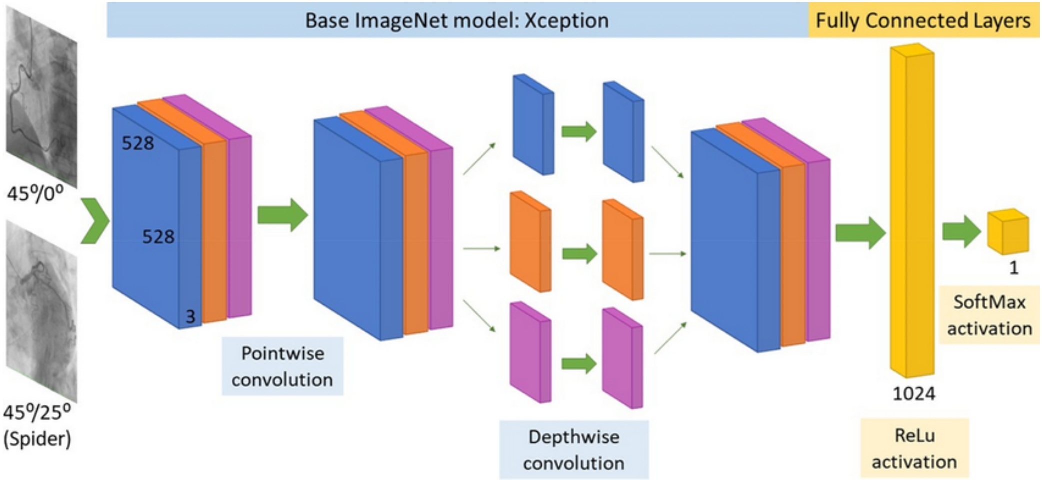


Figure 2.2: Transfer learning in a CNN Xception architecture for coronary artery tortuosity detection. Reproduced from Cobo *et al.* [5].

Numerous CNN architectures have been proposed in the literature. The primary ones used in this thesis are presented below:

- Residual network (ResNet) introduces shortcut connections, also known as skip connections, which allow the input of a block of layers to bypass one or more intermediate layers and be added directly to the output of a deeper layer [23]. This architecture addresses the *vanishing/exploding gradient* problems, enabling to train more effectively deep neural networks [23]. Popular architectures (e.g., ResNet-18, ResNet-34) with different network depths and number of parameters, pretrained on ImageNet [22] or Med3D [24] (a medical imaging database), are available in open source frameworks like Pytorch [25] or Monai [26], and can be selected according to specific computational or performance needs.

- Densely connected convolutional network (DenseNet) connects each layer to all subsequent layers in a feed-forward structure, allowing each layer to receive feature maps from all previous layers and pass its own outputs to all the following ones. This design improves gradient flow, enhances feature reuse, and significantly reduces the number of parameters [27]. Similar to ResNet, popular configurations like DenseNet-121, 169, and 201, with different network depths, are available in open source frameworks such as Pytorch [25] or Monai [26], and can be selected depending on specific computational or performance constraints.

- EfficientNet introduces a scaling method that uniformly scales and balances network depth, width and resolution to achieve better performance and efficiency [28]. The EfficientNet family includes models from EfficientNet-B0 to EfficientNet-B7, which combine efficient scaling of depth, width, and resolution with regularization techniques.

In the field of medical imaging, CNNs achieving expert-level performance on both skin cancer classification [29] and diabetic retinopathy classification [30] represented two significant milestones, demonstrating their value as diagnostic support tools in clinical practice. Additionally, CNNs are widely used for segmentation tasks, enabling the automated delineation of anatomical structures and pathological regions, which in medical imaging are crucial for treatment planning, disease monitoring, etc. Although segmentation is not the primary focus of this thesis, an important milestone in this field was the introduction of the U-Net architecture by Ronneberger *et al.* [21] in 2015, specifically designed for biomedical image segmentation.

2.3.2 Transformers

Transformers were originally proposed for natural language processing (NLP) in 2017 by Vaswani *et al.* [31]. This architecture relies entirely on attention mechanisms, in contrast with previous architectures that combined attention with recurrent and convolutional layers [31]. The attention mechanism maps a query vector and a set of key-value vector pairs to a single output vector. The output is generated by taking a weighted sum of the value vectors, with each weight determined by a compatibility score between the query and its corresponding key. This process enables the model to adjust the contribution of each input element based on its alignment with the query (what we would like to pay attention to).

The transformer architecture employs self-attention and multi-head attention [31]. Given a sequence of elements, self-attention computes the relevance of each element to all others: based on the query, it measures its similarity against the keys of all elements in the sequence and returns a weighted average of the value vector for each element. This allows transformers to capture and process relevant information in the input sequence, attending to the important elements. The most commonly used attention functions for computing similarity are additive attention [32] and scaled dot-product attention [31]. In the latter, attention is computed in parallel over a set of queries, keys, and values represented as matrices Q, K, and V, respectively. The output is given by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.1}$$

where d_k is the dimension of the queries and keys, while d_v is the dimension of the values. Multi-head attention extends the self-attention mechanism by using multiple heads, i.e., sets of different query-key-value projections on the same features, allowing the model to attend and capture diverse aspects of the input sequence simultaneously. Positional encodings are added to input embeddings to provide information about the position of the tokens (words or parts of words) in the sequence [31]. Transformers considerably improved computational efficiency and scalability compared to previous architectures in NLP. The transformer architecture laid the foundation for the development of large language models (LLMs) [33], now widely used in NLP tasks such as text generation, translation, summarization, and dialogue systems. LLMs are DL systems designed to process natural language at a large scale, enabling advanced capabilities in natural language understanding and generation [33]. In genomics, the success of LLMs has motivated the development of genome language models (gLMs), inspired by the analogy between natural language and the genome's biological code [34]. These gLMs are trained on DNA sequences utilizing transformer-based architectures.

Pure vision transformers (ViTs) for images were introduced by Dosovitskiy *et al.* in 2021 [35]. Their architecture splits an image into patches, and a sequence of their linear embeddings is used as input to a transformer, where each patch is treated the same way as tokens in NLP. Cordonnier *et al.* [36] showed that multi-head self-attention is a more general operation than convolution; given a sufficient number of heads, it can approximate any convolutional layer. This enables fully attentional models to learn to combine local context (similar to convolution) and global attention based on the input [36] for better representation learning. However, when trained on insufficient amounts of data, transformers struggle to generalize due to the absence of inductive biases inherent to CNNs [35], introduced by weight sharing and pooling operations, such as translation equivariance and locality. Despite this limitation, the minimal inductive bias in transformers' design achieves competitive performance on a wide range of tasks when combined with large-scale pretraining for NLP and vision applications.

In medical imaging, transformers are very popular in a wide range of tasks, including segmentation, detection, classification, reconstruction, synthesis, or registration, demonstrating similar or even superior performance compared to CNNs due to their ability to capture global context, as surveyed by Shamshad *et al.* [37]. A major advantage of transformers is that they inherently support multimodal applications by integrating embeddings from different modalities and enabling cross-modal attention [37]. There are highly promising applications integrating whole-slide images and bulk transcriptomics through multimodal transformers [38], a field that has been more extensively studied due to the availability of open-access datasets.

2.3.3 Capsule networks

Capsule networks (CapsNets) were implemented by Sabour, Frosst, and Hinton in 2017 [39] to address an important limitation of CNNs, which do not explicitly preserve the spatial relationships and relative positions between features. Pooling operations in CNNs lose valuable information about the precise position of an object within the region [39, 40]. This shortcoming requires larger amounts of training data or the replication of feature detectors for each possible variation of a transformation, which in CNNs translates into an exponential increase in the number of feature maps, thereby raising computational demands and the risk of overfitting [39]. An alternative but inefficient way to address this issue is through data augmentation; however, this approach is often impractical in real-world scenarios due to limited availability of diverse viewpoint data [6]. CapsNets were developed to improve the modeling of part-whole hierarchical relationships, and to learn more robust, pose-aware, and interpretable object-centric representations [6]. As an example, a CNN may classify an image as a face based solely on the presence of key features like eyes, a nose, and a mouth, regardless of whether these features

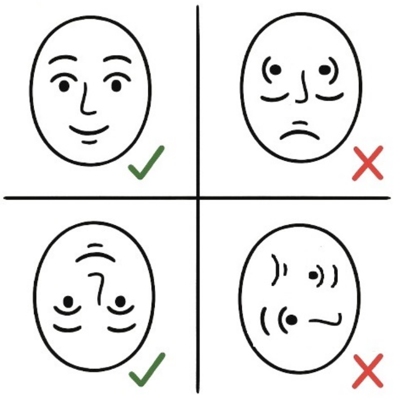


Figure 2.3: Classification of a face depends not only on detecting individual features but also on preserving their spatial configuration. Inspired by Figure 6 from De Sousa *et al.* [6].

are arranged correctly (e.g., eyes above the mouth) or randomly organized [14, 6]. In contrast, capsule networks ideally consider the spatial relationships between features, correctly identifying an image with properly arranged facial features as a face, while rejecting a disorganized configuration. A visual example is shown in Figure 2.3.

CNNs' limited ability to preserve spatial relationships is associated with their difficulty in generalizing to novel viewpoints not seen during training [6]. Convolution operations are only equivariant to translations, therefore, CNNs struggle to handle changes in viewpoint involving other types of transformations. This task is relatively trivial to humans and, unlike CNNs, CapsNets attempt to capture viewpoint invariance within the network's weights to produce viewpoint-invariant predictions [6], as will be explained in the next paragraphs. In contrast to CNNs, CapsNets replace scalar-output feature detectors with vector-output capsules and use routing-by-agreement instead of max-pooling [39]. Geirhos *et al.* showed that CNNs tend to focus more on textures and other local features when interpreting images [41], rather than object shapes, more in line with humans, which makes CNNs more vulnerable against adversarial examples, that are visually indistinguishable to humans [42].

The main components of CapsNets are capsules, capsule layers, the dynamic routing mechanism and the capsule vector activation:

- **Capsule.** A capsule is a group of artificial neurons that represent different properties of the same type of visual entity such as an object or an object part [39]. Their architecture is inspired by evidence from neuroscience, suggesting that groups of spatially proximate neurons (forming what is known as a *hyper-column*) are tightly interconnected and may somehow function as higher-level vector-valued units, capable of transmitting not just a single scalar value, but an entire set of coordinated values [43].

  Formally, a capsule receives a set of N input signal vectors $\mathbf{x}_i$, each of dimension D, and a corresponding set of transformation weight matrices $\mathbf{W}_i$, with dimension $P \times D$. The capsule processes these inputs through a *dynamic routing* process, which employs a set of *routing coefficients* $\gamma = \{\gamma_i\}_{i=1}^N$, with $\gamma_i \in [0, 1]$ to adjust the contribution of each transformed input [6]. The capsule can be described as a parameterized function $c(\mathcal{X}; \mathcal{W}) : \mathbb{R}^{N\times D} \to \mathbb{R}^P$, where $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^D$, and $\mathcal{W} = \{\mathbf{W}_i\}_{i=1}^N$ with $\mathbf{W}_i \in \mathbb{R}^{P\times D}$. Subsequently, a non-linear activation function $\varphi(\cdot)$ is applied to produce an output capsule vector $\mathbf{y} \in \mathbb{R}^P$ given by [6]

  $$\mathbf{y} = \varphi\left(\sum_{i=1}^{N} \gamma_i \mathbf{W}_\mathbf{i} \cdot \mathbf{x}_i\right). \tag{2.2}$$

  This non-linear activation function is the *squashing* function in Sabour *et al.* [39]. Unlike the artificial neuron that produces a scalar output $y \in \mathbb{R}$, a capsule generates a vector of neural activities $\mathbf{y} \in \mathbb{R}^P$.

- **Capsule vector activation.** The capsule vector activation proposed by Sabour, Frosst, and Hinton [39] is the non-linear *squashing* function. The goal is that the length of the output vector $\mathbf{y}_j$ of a capsule represents the probability that the entity is present or not in the input, while the length remains in the range $[0, 1]$:

  $$\mathbf{y}_j = \texttt{squash}(\mathbf{s}_j) = \frac{\|\mathbf{s}_j\|^2}{1 + \|\mathbf{s}_j\|^2} \frac{\mathbf{s}_j}{\|\mathbf{s}_j\|}. \tag{2.3}$$

where $\mathbf{s}_j = \sum_{i=1}^{N} \gamma_i \mathbf{W}_i \cdot \mathbf{x}_i$. This function scales short vectors (low confidence) down and long vectors (high confidence) towards unit length, preserving direction while encoding the presence probability in the length.

- **Capsule layer.** Each layer in a CapsNet contains many capsules where the outputs of one layer become the inputs to the next, enabling hierarchical modeling of lower-level *part* capsules and higher-level *object* capsules [6]. The first capsule layer, known as the *primary capsule*, comes generally from the application of convolutional operations directly to the raw input data (e.g., images, as will be discussed in Chapter 8). The outputs of these capsules serve as part representations, which are then passed hierarchically to the next layers. In each higher layer, the capsules act as parts for increasingly abstract entities, continuing this hierarchical composition until the final capsule layer is reached.

- **Dynamic routing mechanism.** The dynamic routing is the non-linear, clustering-like process that defines how the information flows from the lower layers to the higher layers of capsules [6], similar to transformers that employ self-attention to decide how to attend to different parts of the input. Unlike pooling operations in CNNs, which select the strongest activations, capsule routing operates in a dynamic, clustering-like manner. The goal is to assign lower-level *part* capsules to higher-level *object* capsules by optimizing a set of routing coefficients $\gamma \in \mathbb{R}^{N \times M}$, where $0 \leq \gamma_{ij} \leq 1$. These coefficients represent the affinity between each input signal $\mathbf{x}_i$ and the corresponding output capsule [6]. Capsule routing operates dynamically and aims to establish part-whole relationships through an iterative agreement mechanism, defined as *routing-by-agreement* by Sabour *et al.* [39].

  Initially, each part capsule i makes predictions (votes) about the pose of each object capsule j. If multiple part capsules produce votes that agree (i.e., point in similar directions), the corresponding routing coefficients γ_{ij} are reinforced, while others are suppressed. This enables the network to detect entire objects based on evidence from their parts, evaluating whether those parts form a consistent spatial configuration, rather than isolated feature activations like in pooling operations, as explained in Figure 2.3.

  Formally, during dynamic routing, the routing logits b_{ij} are iteratively updated according to the level of agreement between the output $\mathbf{y}_j$ of each higher-level capsule j and the prediction (vote) $\mathbf{v}_{j|i}$ made by each lower-level capsule i [6]. This agreement is computed as the scalar product $\alpha_{ij} = \mathbf{v}_{j|i} \cdot \mathbf{y}_j$ [39]. Intuitively, if both vectors $\mathbf{y}_j$ and $\mathbf{v}_{j|i}$ are unit vectors (i.e., $\|\mathbf{y}_j\| = 1$ and $\|\mathbf{v}_{j|i}\| = 1$), then α_{ij} corresponds to the cosine of the angle between them. A high value of α_{ij} indicates strong agreement, meaning that the vote vector and the output vector point in a similar direction. This agreement is accumulated into the routing logit b_{ij}, which is subsequently normalized using a softmax function to produce the routing coefficients γ_{ij}:

$$\gamma_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}. \tag{2.4}$$

  Through this iterative process, described in Algorithm 1, capsules learn to assign more weight to those higher-level capsules whose outputs are most consistent with their predictions. The number of routing iterations r suggested by Sabour *et al.* was 3 [39]. In Chapter 8, the number of routing iterations was reduced to 2 without compromising performance.

There are similarities between capsule routing and the self-attention mechanism in Transformers

Algorithm 1 Dynamic routing-by-agreement algorithm

```
function ROUTING($\mathbf{x}, \mathbf{W}, r$)
    for all capsule $i$ in layer $\ell$ and capsule $j$ in layer $\ell + 1$: $b_{ij} \leftarrow 0$
    for all $i, j$: $\mathbf{v}_{j|i} \leftarrow \mathbf{W}_{ij} \cdot \mathbf{x}_i$                ▷ voting
    for $r$ iterations do
        $\forall\, i \in \ell :\ \gamma_i \leftarrow \texttt{softmax}(b_i)$        ▷ routing weights in Equation 2.4
        $\forall\, j \in \ell + 1 :\ \mathbf{s}_j \leftarrow \sum_i \gamma_{ij} \mathbf{v}_{j|i}$
        $\forall\, j \in \ell + 1 :\ \mathbf{y}_j \leftarrow \texttt{squash}(\mathbf{s}_j)$        ▷ Equation 2.3
        $\forall\, i, j :\ \alpha_{ij} \leftarrow \mathbf{v}_{j|i} \cdot \mathbf{y}_j$        ▷ agreement
        $\forall\, i, j :\ b_{ij} \leftarrow b_{ij} + \alpha_{ij}$        ▷ update
    end for
    return $\mathbf{y}_j$
end function
```

discussed in the previous section. Inspired by neuroscience [43], both architectures use vector-valued units that represent entities rather than single scalar values. In CapsNets, capsules are vector-valued units, while in Transformers self-attention mechanisms are represented by key, value, and query vectors [6, 43].

Despite their conceptual potential, CapsNets are still in the early phases of development, as their intended advantages remain difficult to realize in practice. The additional complexity introduced by vector-valued neural activities, together with the computational overhead imposed by routing algorithms result in models that are often inefficient and challenging to train [6]. Moreover, the equivariance property that CapsNets aim to achieve is only approximate in practice, limiting their robustness under complex transformations [6]. In contrast, Transformers, introduced around the same time, have rapidly gained popularity and become the dominant architecture in various domains due to their scalability and superior performance. The effectiveness of DL architectures depends not only on their underlying mechanisms, but also on careful network engineering and design choices, which introduce meaningful inductive biases that play a crucial role in achieving robust performance in real-world applications [44].

In medical imaging, CapsNets have shown promising results in various applications, including pulmonary nodule evaluation in lung cancer imaging, as will be discussed in Chapter 8. Their ability to preserve hierarchical pose relationships and geometrical information between object parts has been particularly explored in medical image segmentation, demonstrating superior performance over CNNs on small datasets [45]. For instance, in lung segmentation, Lalonde *et al.* [46] introduced a novel encoder-decoder CapsNets architecture with an improved routing mechanism, achieving higher performance than the popular U-Net for biomedical image segmentation, while utilizing less than 5% of its parameters. In image classification tasks, Long *et al.* [47] proposed a CapsNet architecture for blood cell classification that outperformed CNN-based methods in low resolution and small datasets. Despite their potential, the high computational cost of CapsNets remains a limitation, suggesting the development of hybrid designs that combine them with conventional neural networks as a promising future research direction [45].

2.4 Discussion and conclusions

Artificial intelligence has made substantial progress since Turing's 1950 proposal of a test to evaluate machine intelligence through language. The *Turing Test* has gained renewed relevance in the era of LLMs, which have demonstrated an impressive ability to produce grammatically coherent and meaningful sentences. However, these systems continue to struggle with many basic aspects of arithmetic, logic, causal reasoning and physical common-sense [48]. Their tendency to produce confident but incorrect outputs (commonly referred to as *hallucinations*) raises concerns about their reliability, particularly in high-stakes domains like healthcare. To address the limitations of purely linguistic evaluation, the *embodied Turing test* has been proposed, shifting the focus to how AI systems interpret and interact with the physical world [48].

This broader view of intelligence reflects the historical development of AI methods. Early symbolic systems, which relied on manually encoded rules, were successful in structured, well-defined domains. In medicine, expert systems like MYCIN demonstrated the potential of rule-based reasoning, but were limited in flexibility and scalability. The emergence of ML represented a turning point towards data-driven models capable of recognizing patterns without explicit instructions. Yet, conventional ML methods still required hand-crafted features and struggled to model the complexity of high-dimensional image data.

Deep learning, particularly through CNNs, addressed many of these limitations by enabling models to learn hierarchical features directly from raw inputs. CNNs became the standard for image-based tasks, including classification, segmentation, and detection in medical imaging [17]. However, their architecture favors local context and may overlook global relationships or fine spatial details, which can be critical for accurate clinical interpretation. To mitigate these weaknesses, new architectures have emerged. On the one hand, transformers, through their attention-based mechanisms, are capable of modeling both local and global dependencies. This architecture is especially effective in multimodal scenarios that integrate imaging with other data sources. CapsNets, on the other hand, aim to preserve spatial relationships between features, providing robustness to variations in viewpoint, and potentially enabling more interpretable object-centric representations. Each of these models addresses different limitations of traditional CNNs, and their combined use is likely to define the next generation of AI systems for medical applications.

When considered as a whole, these architectural paradigms offer complementary strengths. CNNs are efficient and well-validated for local pattern detection. Transformers enable context-aware and multimodal processing. CapsNets preserve spatial relationships and learn pose-aware object representations. It is expected that future progress will not rely on one single model, but rather on hybrid approaches that combine the best properties of each architecture. For instance, recent studies explore the integration of convolutional encoders with transformer-based attention modules for medical image segmentation [49]; or the combination of capsule layers, multi-head attention and convolutional layers to improve generalization capabilities of CNNs for COVID-19 chest X-ray classification [50].

It is also important to emphasize that symbolic AI is not obsolete. While it may no longer dominate AI research, its principles continue to inform the development of more transparent and controllable systems. In medical imaging, symbolic knowledge can be integrated into DL pipelines in several ways: through the design of loss functions, the inclusion of prior knowledge in model architectures, or post-processing steps that align outputs with domain constraints. These hybrid strategies can enhance performance, especially in data-scarce scenarios or when dealing with rare conditions where purely data-driven learning may be insufficient to capture the clinical variability.

The goal of AI in medical imaging is not to replace human experts, but to support and extend their capabilities. Successful AI systems must be reliable, transparent, and aligned with clinical objectives. They must also be adaptable to the complex conditions of healthcare environments, which often involve imperfect data, variability across institutions, and the need for clear explanations of model outputs. As this chapter has shown, the foundations set by decades of AI research are now converging in powerful new tools. However, the journey is ongoing, and many open questions still remain regarding interpretability, fairness, generalization, and integration into clinical workflows.

The next chapter introduces the fundamental concepts of physics in medical imaging and examines how these can be incorporated into learning algorithms. In particular, physics-informed machine learning represents a promising direction, combining the strengths of domain knowledge with the flexibility of purely data-driven approaches. This reflects a growing evolution in AI towards models that are not only accurate, but also interpretable, uncertainty-aware and, overall, trustworthy, as will be described in Chapter 4.

Chapter 3

Fundamentals of physics in medical imaging

Main publication associated with this chapter: **Cobo, Miriam**, *et al.* "Physical foundations for trustworthy medical imaging: a survey for artificial intelligence researchers". *Under review.*

3.1 Introduction

Applications of AI in medical imaging have experienced unprecedented growth in the last decades, driven by rapid advancements in DL, and the increasing availability of high-performance computing. Beyond diagnostic purposes, medical imaging plays a crucial role in treatment planning, disease progression monitoring, and real-time guided interventions, together with research and educational purposes, such as functional imaging, and construction of population atlases.

Crucial aspects of medical images include finding a compromise between patient safety and image quality in the acquisition power and energy levels, together with the amount of time required to generate the images, since patient motion compromises the acquisition of higher resolution medical images. These elements are particularly important in those modalities that employ ionizing radiation. Yet, in some modalities such as nuclear medicine, the acquisition time is directly limited by physics [51]. The aforementioned factors are taken into account to find an optimal balance in the clinics, and they are essential in medical image reconstruction. Hence, medical images are optimized to perform the required task at the necessary image quality that ensures accurate diagnosis [51].

There exists a gap between research in AI for medical imaging applications and clinical translation, which is hindered by challenges in generalization, standardization, interoperability and privacy limitations, as will be discussed in the next chapters [52]. Moreover, discrepancies in how neural networks learn from different domains, in particular the adaptation between medical imaging and natural images domain are often overlooked when developing AI algorithms [53].

The existing literature tends to focus on domain-specific challenges in medical imaging. More general reviews published recently address either AI-driven innovations in healthcare [54], or practical and regulatory challenges in implementing AI in medical imaging [55, 56]. Varoquaux and Cheplygina [55] provide recommendations to avoid systematic challenges and to improve the clinical impact of ML in medical imaging. Saw *et al.* [56] provide an overview of the pitfalls of current state-of-the-art AI systems in medical imaging. However, these reviews overlook the gap between pure ML data-driven methods and the physical foundations of medical imaging, which hold significant potential to mitigate some of the existing limitations. This chapter addresses that gap by examining how physics knowledge can be integrated into AI models to improve trustworthiness and robustness in real-world data-limited medical settings.

The 2024 Nobel Prize in Physics, awarded to John J. Hopfield and Geoffrey E. Hinton for their foundational work in machine learning (ML) with artificial neural networks, underscores the fundamental role of physics in driving technological innovations. Physics-based methods integrated in AI algorithms provide enhanced reliability and system safety by accurately representing the underlying physical relationships in medical images. In this chapter, the physical properties of the main imaging modalities in the clinics are summarized, with a particular focus on CT images, as they represent the primary medical imaging modality studied in this thesis. The limitations of image quality and the potential of DL and generative AI to mitigate these challenges are further discussed. Moreover, the integration of physics knowledge into physics-inspired ML models is explored. These algorithms leverage physics-based constraints to enhance the learning of medical imaging features.

This chapter is structured as follows: Section 3.2 introduces the physics behind each medical imaging modality existing in the clinics (Section 3.2.1-3.2.6), as well as image quality challenges (Section 3.2.7). Section 3.3, presents physics-informed machine learning algorithms, detailing existing approaches (Section 3.3.1-3.3.3), and challenges (Section 3.3.4). Finally, Section 3.4 discusses the potential of physics-informed models, future trends and conclusions.

3.2 Physics behind medical imaging by modality

Radiation is the propagation of energy through space or matter. Electromagnetic radiation, subatomic particle radiation (nuclear imaging), and acoustic radiation (ultrasound imaging) are the main types of radiation for medical imaging. Figure 3.1 depicts the electromagnetic radiation spectrum for the different medical imaging modalities available (note that radiation ranges in ionizing imaging techniques are approximate and depend on the patient's size). Figure 3.2 shows ultrasound imaging in the acoustic spectrum. In this section we introduce all the clinical medical imaging modalities shown in Figures 3.1 and 3.2.

Given the variety of imaging techniques and the importance of consistency in medical diagnostics, it becomes crucial to have a standard for managing and sharing medical images. This is where DICOM (Digital Imaging and Communications in Medicine) comes into play. DICOM is the universal standard that defines and controls the formats for sending, distributing, and storing medical images across different machines, manufacturers, and imaging modalities [57]. It plays a central role in ensuring that images from various sources can be accessed, interpreted, and analyzed consistently, regardless of the technology used. DICOM is implemented in nearly all radiology, cardiology, and radiotherapy devices, and its use is expanding to other medical fields such as ophthalmology and dentistry [58].

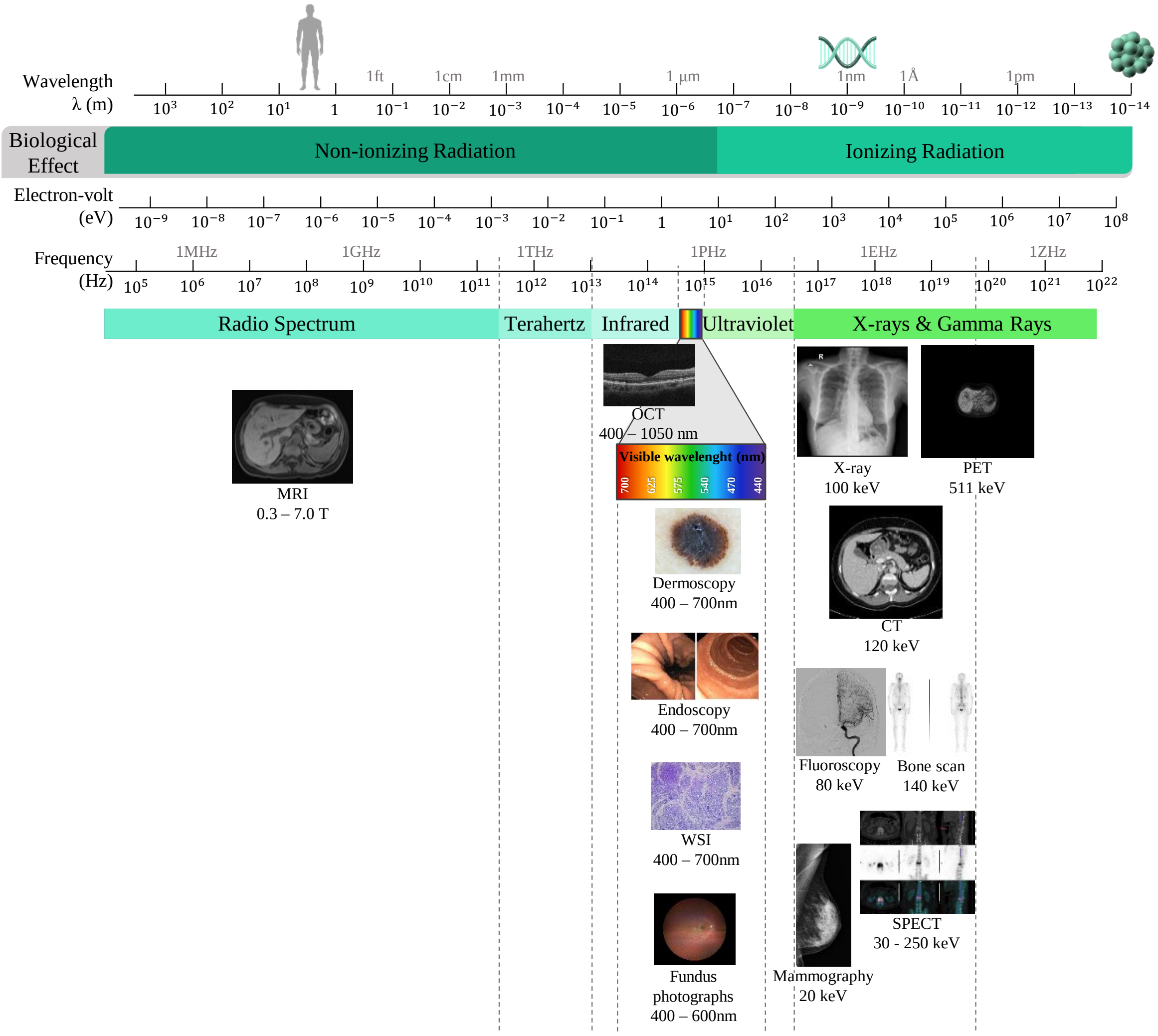


Figure 3.1: Main medical imaging modalities in the electromagnetic radiation spectrum. Abbreviations: CT = Computed Tomography, MRI = Magnetic Resonance Imaging, OCT = Optical Coherence Tomography, PET = Positron Emission Tomography, SPECT = Single-photon Emission Computed Tomography, WSI = Whole-Slide Imaging (in pathology, histology).

3.2.1 Visible spectrum images

Visible light is utilized to produce 2D images or videos in fields such as dermatology, gastroenterology, histology and ophthalmology. In dermatology, the most widely used technique to capture skin images is dermoscopy (a dermatoscope is simply a magnifying lens), followed by total-body digital photography [59]. Endoscopy uses visible light to illuminate different parts of the gastrointestinal tract, capturing images or videos of the structures of interest. Examples of applications of endoscopy are colonoscopy or laparoscopy, yet its utility spans a broad spectrum of medical procedures. Current trends in AI applications in endoscopy can be found in Chahal *et al.* [60]. Histology frequently relies on visible light microscopy to examine tissue and cells stained with specific dyes at different magnifications. The introduction of whole slide imaging (WSI) and digitalization in 1999 revolutionized pathology by enabling high-resolution digital slides [61]. Bahadir *et al.* [62] review the latest trends of AI

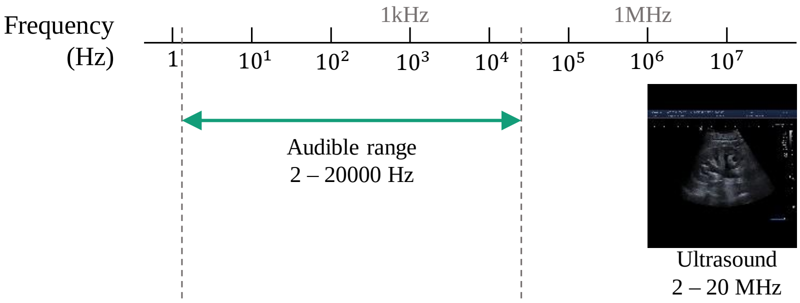


Figure 3.2: Ultrasound imaging in the acoustic spectrum.

in histopathology. Although light microscopy is the main diagnostic tool in histology, Transmission Electron Microscope is performed routinely on renal biopsies. In this context, recent work by Zhang *et al.* [63] applied DL techniques on electron microscopy images of renal biopsy. In ophthalmology, color fundus photography employs a fundus camera to record color images of the retina (fundus). Diverse applications of AI are reviewed by Grzybowski *et al.* [64]. Moreover, the latest studies showed that fundus photographs can be used to monitor the progression of neuro-degenerative disorders [64].

Optical Coherence Tomography

Optical coherence tomography (OCT) was developed in the 1990s for non-invasive cross-sectional imaging in biological systems [65]. Since then, it is evolving from near-infrared illumination to visible light optical coherence tomography [66], demonstrating its effectiveness in preclinical and ophthalmic imaging. OCT holds great promise for AI applications [67], with some recent advances in the field of neurological diseases in relation to OCT [68].

3.2.2 X-ray imaging

X-rays are most likely the best known medical imaging modality. The underlying physical principle is simple, and yet effective: X-rays are a type of electromagnetic radiation produced by high energy electrons. These electrons are produced due to the ionization of nitrogen and oxygen atoms, which attract positive ions to the cathode, and therefore inject electrons that are accelerated to the anode. The resulting X-rays can be directed to a patient, and then collected in a detector. X-rays are absorbed and scattered to different extents by various types of tissues, which allows to capture their interactions with the patient's anatomy in a radiographic image. There are four major types of interactions of photons with matter, but only three of them play a role in diagnostic radiology and nuclear medicine (Rayleigh scattering, Compton scattering and photoelectric effect) [51], while the last one (pair production, can only occur when the energy of photons exceeds 1.02 MeV) has only been simulated at a theoretical level for monitoring of radiotherapy dosing [69].

Radiography

Radiography was the first medical imaging technology. This technique captures the attenuation (absorption and scattering) of an homogeneous distribution of X-rays entering a patient, which is then modified by the interactions with the different tissues, resulting in an heterogeneous distribution emerging from the patient that is recorded in a 2D radiograph [51]. The kilovoltage, X-ray exposure time, and beam size are adjusted according to the anatomical area under study. As an example, in the field of dentistry, individual dental radiographs and orthopantomograms are performed. Orthopantomograms are panoramic dental radiographs produced by rotating the X-ray tube around the patient's head, generating a comprehensive 2D image of the dental and maxillofacial structures. [70]. Advances in AI have further enhanced the interpretation of these images, as reviewed by Costa *et al.* [71].

Computed Tomography

Computed tomography (CT) images are 3D images generated by producing multiple X-ray projection images across a broad angular range, typically 360°, rotating the X-ray tube and detector around the patient [51]. Simultaneous rotation of the X-ray source and translatory movement of the patient allow to achieve continuous data acquisition throughout the volume of interest [72]. This geometry corresponds to an helical CT scanner, which represents the vast majority of CT scanners in use today.

Previous to helical CT scanners, there were sequential CT scanners, which are still in use for cranial imaging [73]. In sequential CT scanners, the patient remains stationary during each full rotation of the X-ray tube, advancing incrementally to acquire axial slices. This technique enhances image resolution but results in increased radiation exposure and scanning time compared to helical CT.

The generation of the 3D image relies on advanced reconstruction algorithms, with the resulting voxel values represented in grayscale, ranging from -1000 to 1000. The grayscale in CT is named Hounsfield Unit (HU), after one of the main developers of this technology [51]. X-ray CT scanners typically operate at 120 kV, however, the voltage can be optimized based on the specific application and the patient's size [51]. Deep learning (DL) has shown potential to personalize optimization of imaging protocols to minimize radiation exposure while maintaining clinical image resolution [74].

Spatial resolution depends on several factors including physics related aspects such as X-ray focal spot size, number of projection views per rotation of the X-ray tube, detector cell size; together with reconstruction algorithms [75]. Until 2009, the lack of computational power prevented the clinical introduction of iterative algorithms for image reconstruction in CT, which rapidly replaced filtered back projection (FBP). In the latter, CTs were reconstructed from projections (sinograms) by applying a high-pass filter followed by a backward projection step [76]. The main drawback of FBP is the significant reduction in image quality when radiation dose is decreased, due to the increase in image noise. Recently, AI has emerged as a new promising technique to improve CT image reconstruction, showing potential to reduce CT radiation doses while speeding up reconstruction times [51, 76, 77, 78]. However, DL-based reconstruction algorithms require large training datasets, and are prompt to biases in different subpopulations if the training data significantly differs from the target population.

There are particularities of CT protocols depending on the clinical application. For lung cancer screening [79], low-dose computed tomography (LDCT) utilizes a lower dose of radiation to scan the patient, which is achieved lowering the X-ray flux. However, lowering the dose of radiation in LDCT increases image noise and, thus, reduces signal-to-noise ratio (SNR) and image quality. Physics-/model-based data-driven methods for LDCT are surveyed by Xia *et al.* [80].

The incorporation of contrast agents, which facilitate the evaluation of patient hemodynamics and the characteristic vascularization of tissues, has enabled the development of various advanced acquisition modalities. These include perfusion imaging, which assesses blood flow dynamics, aiding in the diagnosis of stroke and myocardial perfusion; virtual CT colonoscopy, which generates 3D images of the colon providing a non-invasive alternative for detecting polyps and tumors; and prospectively gated cardiac CT, which reduces motion artifacts by synchronizing image acquisition with the cardiac cycle, enhancing coronary artery evaluation [51, 81].

Next-generation modern CT systems have integrated dual-energy technology (DECT), wherein X-ray spectra are captured at both low and high energy levels. This approach enables the independent assessment of the contributions from photoelectric effect and Compton scattering. DECT facilitates material decomposition by leveraging differences in attenuation coefficients at varying energy levels, thereby distinguishing materials with similar HU but differing atomic compositions. This technology can produce various image types, such as virtual monoenergetic images, material-specific images (e.g., iodine maps), and virtual non-contrast images, significantly enhancing tissue characterization and lesion detection capabilities [82]. Recent advancements in generative AI have further improved these capabilities. Jeong *et al.* [83] explored the application of generative AI techniques to enhance DECT imaging, focusing on improving image quality and diagnostic accuracy through ML methods.

Additionally, as an emerging technology with the potential to change clinical CT, photon-counting CT (PCCT) integrates new energy-resolving X-ray detectors to count the number of incoming photons and measure their energy, resulting in higher contrast-to-noise ratio, improved spatial resolution, and optimized spectral imaging at a lower radiation exposure [84]. In conventional CT, finer detector space leads to larger datasets that allow to achieve higher resolution, increasing reconstruction time, which is not required for many clinical tasks [51]. PCCT energy-resolving detectors eliminate electronic noise, in comparison with the traditional energy-integrating detectors in CT. Greffier *et al.* [81] comprehensively review PCCT and compare their technical innovations against conventional CT. Nevertheless, as any innovative technology, PCCT also encounters challenges, such as detector charge-sharing effects or Compton scattering, which may lead to errors during the reconstruction process that degrade image quality. Yu *et al.* [85] proposed a novel physics-guided material decomposition model for PCCT, that leverages DL and incorporates critical physical parameters, which underlines the role of physics in building reliable DL systems in medical imaging.

Mammography

Mammography is an optimized radiography examination specifically designed for detecting breast cancer at an early stage. Screening mammography attempts to detect breast cancer in the asymptomatic population, while diagnostic mammography aims to assess and delineate lesions identified by the former [51]. AI holds significant potential for improving breast cancer screening, current state of the art and challenges are reviewed by Díaz *et al.* [86].

Modern mammography systems incorporate rotating X-ray tubes that facilitate tomosynthesis, a technique grounded in the same physical principles as sequential CT. This advancement enables the acquisition of more detailed and comprehensive imaging, offering enhanced diagnostic information in appropriately selected cases. Additionally, the introduction of contrast-enhanced mammography, combined with DL algorithms, has shown diagnostic performance comparable to magnetic resonance imaging (MRI), especially for evaluating dense breast tissue [87]. Furthermore, in breast imaging, ultrasound (US) is used as a supplemental screening alternative for women with dense breast tissue [51], and recently MRI screening has been recommended for women with extremely dense breasts [88].

Fluoroscopy

Fluoroscopy shows real-time X-ray imaging of internal anatomic structures for the placement of medical devices, such as catheters and stents, or the observation of temporal physiological phenomena in patients providing a dynamic display of the structure of interest [51] . Fluoroscopy systems can operate in two modes: (1) fluoroscopy, real-time imaging for positioning, which is usually not recorded and involves relatively low radiation exposure, and (2) fluorography, which records clinically relevant sequences using a pulsed radiographic mode, giving higher radiation levels, similar to radiographic imaging [51]. The use of contrast media is critical in fluoroscopy for enhancing the visibility of internal structures and improving diagnostic accuracy. Contrast agents, such as iodine-based compounds or barium sulfate, allow for the differentiation of soft tissues, blood vessels, and other anatomical features that would otherwise be indistinguishable in conventional X-ray imaging. In procedures such as angiography, the injection of iodinated contrast highlights the vascular system, enabling clinicians to visualize blood flow, detect blockages, and guide interventions with precision. Similarly, in gastrointestinal studies, barium-based contrast delineates the digestive tract, facilitating the assessment of structural and functional abnormalities [89] AI applications are more complex for interventional radiology (IR) than for diagnostic radiology. IR encompasses preprocedural diagnostic imaging, procedural imaging guidance, prosprocedural imaging evaluation, and therapeutic tools, mostly relying on unstructured data, which challenges AI applications. The potential of AI in the field of IR is reviewed by Glielmo *et al.* [90].

3.2.3 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) studies the magnetic properties of the nucleus of the atom through radio-frequency (RF) waves. The atomic nucleus is composed of protons and neutrons, which exhibit a magnetic field associated with their nuclear spin and charge distribution. In contrast with X-ray imaging, MRI does not employ ionizing radiation, however, it is a more expensive imaging modality compared to the former. The key components are the magnet, magnetic field gradient coil, and RF coils. In MRI, a strong external magnetic field generated by the magnets causes the individual nuclei to selectively absorb, and then release, energy unique to those nuclei and their surrounding environment. This energy coupling is known as resonance [51]. An RF coil is basically a resonant circuit, which is tuned to the resonance frequency of proton spins for a given magnet field (similar to a radio tuned to the frequency of a radio station) [91]. The typical magnetic field strengths for MRI systems range from 0.3 to 7.0 T, which require the electromagnet core wires to be superconductive [51]. Magnetic field gradients are essential to localize signals generated during MRI process, as these fields interact with the main (and much stronger) magnetic field [51].

MRI sequences leverage the resonance to generate images with varying contrasts based on tissue properties, in which each voxel (3D pixel representing a small unit of volume in an image) depends on the number of protons (proton density) and magnetic properties of the tissue in that voxel. MRI can produce high contrast images due to the distinct local magnetic field properties of different types of tissue (fat, white and gray matter in the brain, tumor, etc.) [51]. The two primary pulse sequences in MRI, spin echo and gradient echo, serve as the basis for generating different types of image contrast. Among these, T1-weighted and T2-weighted imaging provide distinct diagnostic information. T1-weighted imaging focuses on the longitudinal relaxation of protons, using short repetition time (TR) and short echo time (TE), where tissues with short T1 times appear bright and fluids appear dark, offering excellent anatomical detail. In contrast, T2-weighted imaging emphasizes transverse relaxation, employing long TR and long TE, where tissues with long T2 times and fluids appear bright, making it particularly effective for identifying pathology, inflammation, and edema. Additional sequences include Proton Density, which highlights proton concentration differences using long TR and short TE; FLAIR (Fluid-Attenuated Inversion Recovery), which suppresses cerebrospinal fluid (CSF) signals to enhance lesion detection near CSF spaces; and Diffusion-Weighted Imaging, which assesses water molecule diffusion, and is critical for early stroke diagnosis. [92]. MRI can also monitor blood flow in arteries (MR angiography), and blood flow in brain (functional MR) [51].

MRI data are initially stored in the raw spatial frequency domain, the k-space matrix. The k-space encodes spatial frequency values in a four quadrant 2D matrix of complex values, where the origin at the center represents frequency zero, the central region contains lower spatial frequencies, and the higher spacial frequencies are in the periphery [51]. Each point in the k-space corresponds to a specific spatial frequency component of the final image. In a conventional acquisition, the k-space matrix is filled one row at a time by systematically collecting data points through the application of gradient magnetic fields, which encode spatial information by varying the frequency and phase of the detected signals [51]. Once rows in the k-space matrix are fully populated, image reconstruction is performed using the inverse fast Fourier transform. This mathematical operation decodes the frequency-domain data in the k-space matrix to produce the spatial domain representation, revealing the anatomical structures of the scanned area. The final image is processed to represent photon density, T1, T2 and flow characteristics of the tissues using a grayscale range, with each pixel corresponding to a voxel [51]. The organization and density of data sampling in the k-space matrix directly influence image quality, resolution, and the presence of artifacts.

Image reconstruction in MRI is influenced by the physical effects that are included in the signal model [93], and presents challenges due to long acquisition times. Several research efforts have focused on accelerating MRI, i.e., developing methods to reconstruct images from under-sampled data [94]. Traditional reconstruction methods from under-sampled k-space data include parallel imaging and compressed sensing, widely used in the clinics [95], although they both encounter practical limitations [94]. DL methods have enabled transforming under-sampled or noisy data into high quality images, mitigating artifacts and accelerating the imaging process [94]. DL algorithms have also been used to reduce slice spacing in MRI and reconstruct higher-resolution volumes [96]. Generative adversarial neural networks (GANs) have been explored in MRI reconstruction to estimate missing k-space samples, and correct artifacts in the image space [97]. In a recent proof of concept, Okoli *et al.* [98] proposed a score-based diffusion model for accelerating MRI reconstruction, although they underscored the need for further research and clinical assessment. Furthermore, model-based methods considering physical effects and integrating neural networks have shown potential to improve image quality [93, 99]. Another recent study by Peng *et al.* [100] proposed a k-space acquisition optimization strategy conditioned on MRI physics for accelerated MRI reconstruction using a neural ordinary differential equation (ODE) combined with DL-based reconstruction. These works pave the way for further research in physics-based AI algorithms.

Patient motion represents a significant challenge in MRI, where acquisition times range from 20 to 60 minutes depending on the different sequences added. Quantitative MRI (qMRI) derives physical tissue properties from a set of qualitative images captured with different imaging settings, facilitating consistent measurement of biomarkers, in spite of longer acquisition times. In contrast to conventional MRI, which relies on relative signal intensities for visual interpretation, qMRI quantifies parameters such as T1 and T2 relaxation times, proton density, and diffusion coefficients in standardized units. This quantitative evaluation enables reproducible comparisons across subjects, scanners, and time points, while minimizing hardware-related variability. By employing multiple acquisitions with differing parameters, qMRI enhances measurement precision and specificity, allowing for the detection of subtle changes in tissue integrity and composition, such as myelin content and iron concentration [101, 102]. Recent work by Eichhorn *et al.* [103] proposed physics-informed motion correction (through a physics-informed loss) to leverage information from the MRI signal evolution to detect and exclude motion-corrupted k-space lines from a data consistent reconstruction.

MRI also employs contrast agents to enhance the visualization of internal structures during imaging examinations. These agents are typically administered intravenously, with gadolinium-based contrast agents (GBCAs) being the most commonly used. GBCAs primarily shorten the T1 relaxation time of tissues, leading to increased signal intensity on T1-weighted images. Additionally, iron oxide-based agents, including super-paramagnetic iron oxide (SPIO) and ultra-small super-paramagnetic iron oxide (USPIO), are utilized to reduce T2 signal intensity, while manganese-based agents enhance T1 signal intensity. The selection of a specific contrast agent depends on the clinical application, with some agents designed for targeted organ imaging, such as liver-specific contrast agents [104].

3.2.4 Nuclear images

Nuclear medicine is a specialized field of medical imaging that uses radioactive isotopes, known as radiotracers, to visualize physiological and metabolic processes within the body. These radiotracers may be bound or unbound to other molecules and interact at the cellular level, emitting ionizing radiation during nuclear decay. The emitted radiation consists of subatomic particles, such as alpha particles (helium nuclei), or beta particles (electrons or positrons), or gamma rays (photons). This radiation is detected to generate images that reflect metabolic activity [51]. While these images offer valuable functional information, they typically present lower anatomical resolution compared to other imaging modalities. Advances in nuclear medicine have led to the development of two primary imaging techniques: Single-Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET), each enabling unique diagnostic capabilities. Both SPECT and PET rely on the principle that ionizing radiation emitted by radiotracers traverses anatomical structures with varying attenuation based on tissue density [105]. These techniques provide critical insights into physiological and metabolic processes, playing a pivotal role in the diagnosis and management of numerous conditions, including cardiovascular diseases, neurological disorders, and cancers [106]. However, nuclear medicine techniques have several limitations, including lower spatial resolution compared to CT or MRI, which can hinder the detection of small structures [107]. The use of ionizing radiation requires careful dose management to minimize patient risk. Additionally, the short half-lives of radiotracers pose logistical challenges related to their production, distribution, and timely administration, while the high cost and limited availability of radiotracers can restrict access to imaging services [106]. Despite these challenges, nuclear medicine remains a crucial tool in modern diagnostics due to its ability to provide unique functional and metabolic information.

Single Photon Emission Computed Tomography

Single-Photon Emission Computed Tomography (SPECT) is a nuclear medicine imaging technique that provides 3D functional information about biological processes within the body. SPECT imaging relies on gamma-emitting radiotracers, such as technetium-99m (Tc-99m), which is the most commonly used isotope due to its favorable energy characteristics and high half-life [108]. Tc-99m decays to a more stable form, Tc-99, by emitting gamma rays (140 keV [109]) which pass through the body interacting with tissues according to their density and the radiation's penetration power. High-density structures, such as bones, attenuate more radiation than lower-density tissues, such as fat, a difference that is captured by detectors to produce images. Gamma cameras detect these photons using collimators that allow only photons traveling at specific angles to pass through, ensuring precise image formation [51]. By acquiring multiple 2D projections from different angles, tomographic reconstruction algorithms create detailed 3D images that enhance lesion localization and characterization. SPECT's versatility is demonstrated by its wide range of clinical applications as myocardial perfusion imaging, cerebral blood flow, pulmonary ventilation/perfusion, tumor detection or bone scintigraphy. Notably, SPECT bone scintigraphy plays a critical role in evaluating bone metabolism, facilitating the identification of fractures, infections, and metastases [110, 111].

Positron Emission Tomography

Positron Emission Tomography (PET) is an advanced molecular imaging technique that visualizes biological functions and metabolic processes with high sensitivity. PET relies on radiotracers labeled with positron-emitting radionuclides. The most commonly used radiotracer is fluorine-18-labeled fluorodeoxyglucose (18-FDG), which mimics glucose metabolism. During its decay to Oxygen-18, Fluorine-18 emits a positron that annihilates with an electron, resulting in the emission of two 511 keV gamma photons in opposite directions. In PET scanners, rings of detectors identify these photon pairs through a process known as coincidence detection, allowing for accurate localization of the positron-electron annihilation event [51]. Subsequently, advanced image reconstruction algorithms generate 3D images of the tracer distribution within the body, mapping metabolic activity at a molecular level.

PET imaging is highly versatile, utilizing a range of radiotracers tailored to specific clinical applications. 18-FDG is primarily used in oncology for cancer detection, staging, monitoring treatment response, and recurrence assessment [112]. In neurology, PET imaging evaluates brain function in conditions such as dementia, epilepsy, and neurodegenerative diseases [113]. In cardiology, PET is used to assess myocardial perfusion and viability, supporting the planning of coronary artery bypass graft procedures. Additionally, radiotracers such as Carbon-11, Nitrogen-13, and Oxygen-15 are utilized to study specific metabolic processes, while specialized tracers are designed for particular cancers or neurological disorders [113].

Hagos *et al.* [114] underscored the potential of generative AI and LLMs to advance nuclear medicine practices, opening new avenues for improving diagnostic accuracy and workflow efficiency.

3.2.5 Ultrasound

Ultrasound (US) waves are mechanical high frequency sound waves that require an elastic medium to spread over, unlike the previously described medical imaging modalities that leverage electromagnetic (EM) waves. Sound waves are longitudinal waves, oscillating in the direction of travel. In contrast, EM waves are transverse waves, oscillating perpendicular to the direction of travel [115]. In the process, an ultrasound transducer, that is in direct physical contact with the patient, generates a short-duration pulse of sound, which travels into the tissue and is reflected by internal structures and organs in the body, such that the transducer receives and records the amplitude and time delay of the echoes for a given direction of the pulse [51]. The repeated application of this process over a wider region within the anatomical area of interest enables the creation of an US image.

In modern US transducers, multiple elements can transmit and receive the pulses resulting in a brightness mode (B-mode) ultrasound image of a planar section of tissues [51]. Apart from the B-mode, there are also A-mode and M-mode. A-mode (amplitude) is the processed echo amplitude as a function of time, generated as output by the receiver. It is the simplest mode of US generation by a single transducer [51, 115]. This was the earliest application of US in medicine, and is now almost obsolete. However, A-mode is sometimes combined with M-mode imaging, and it is also used in ophthalmology for precise measurements of the eye, as well as in therapeutic US applications [51, 115]. M-mode (motion), also known as T-M mode (time-motion), is commonly used in echo-cardiography. It employs B-mode information from a stationary US beam to track velocities of echoes generated from moving structures acting as reflectors throughout the cardiac cycle [51, 115]. Additionally, Doppler modes are essential for evaluating blood flow. Color Doppler displays blood flow direction and velocity, while pulse wave and continuous wave Doppler measure the speed of blood flow or tissue movement, leveraging the Doppler effect to detect changes in frequency due to motion [51].

Optimizing US image quality involves selecting appropriate settings for the specific anatomical area under examination. To facilitate this process, manufacturers provide preset configurations tailored to different target areas, simplifying image acquisition for operators who may not have extensive knowledge of the underlying physics. Once the appropriate preset is selected, several key parameters can be adjusted to fine-tune image quality. Brightness is influenced by gain settings, including overall gain and time gain compensation, which compensates for signal attenuation at different depths. Lateral gain compensation adjusts brightness horizontally. Proper gain settings ensure that low echogenicity structures, such as fluids, appear black, while highly echogenic structures, like bones, appear white. Other critical parameters include depth, dynamic range, focal zone, and frequency. Higher frequencies provide better resolution but lower penetration, while dynamic range settings influence contrast and visibility of details [116]. To enhance image quality, tissue harmonic imaging (THI) is employed. This technique relies on the non-linear propagation of US waves through tissue, resulting in the generation of harmonic frequencies (multiples of the fundamental frequency). The second harmonic frequency is typically used for image formation, as higher harmonics are attenuated. THI improves image resolution, reduces artifacts, and enhances the visualization of deeper structures [117].

Building on the principles of harmonic imaging, ultrasound contrast agents (UCAs), administered intravenously, further enhance diagnostic capabilities by improving the visualization of blood flow and tissue perfusion. UCAs are gas-filled microbubbles, typically 1-8μm in diameter, stabilized by a phospholipid or protein shell. When exposed to US waves, these microbubbles oscillate and resonate, producing strong echoes due to their nonlinear behavior. This resonance generates harmonic frequencies that can be selectively detected, significantly increasing the contrast-to-tissue ratio. Techniques like harmonic imaging and pulse inversion exploit these properties to differentiate microbubble signals from background tissue. UCAs are employed in cardiology to define endocardial borders and assess myocardial perfusion, as well as in radiology to characterize liver lesions [118].

In quantitative ultrasound imaging, the goal is to quantify interactions between US and biological tissues, enhancing diagnostic capabilities [119]. Ultrasound elastography extends these applications by evaluating tissue stiffness, aiding in differential diagnoses and identification of biopsy targets. This combination of modes, techniques, and image optimization parameters makes US a versatile and indispensable tool in modern medical imaging, complementing other diagnostic modalities [120].

Recent work in US explored improving the quality of AI generated US images by introducing a physics-based diffusion model specifically designed for this modality [121]. The proposed customized noise scheduler simulated the attenuation of echoes returning to an US receiver. This work opened the door to further refinements based on the physics of US, such as reflection and scattering.

3.2.6 Combined imaging modalities

Combined imaging modalities integrate multiple imaging techniques to provide comprehensive anatomical and functional information, thereby enhancing diagnostic accuracy and clinical decision-making. Leveraging advanced computational methods and DL-based algorithms, combined imaging modalities improve image quality, diagnostic specificity, and workflow efficiency. These techniques also address the inherent limitations of single-modality imaging by enabling clinicians to obtain a more holistic view of patient anatomy and physiology, ultimately contributing to improved outcomes across various medical specialties [122, 123, 124]. These approaches include both hardware-based hybrid systems and software-based image fusion techniques, each offering distinct advantages and applications in medical diagnostics.

Hardware-based combined imaging modalities

Hardware-based hybrid systems combine two imaging modalities into a single device, enabling simultaneous or near-simultaneous image acquisition. Notable examples include PET/CT, SPECT/CT, and PET/MRI. PET/CT merges the metabolic insights of PET with the high-resolution anatomical detail of CT, playing a critical role in cancer detection, staging, and treatment monitoring. SPECT/CT integrates the functional information of SPECT with the anatomical detail of CT, enhancing attenuation correction and localization accuracy. It is extensively used in cardiology, oncology, and neurology [107]. PET/MRI integrates PET's molecular imaging capabilities with MRI's superior soft tissue contrast, providing reduced radiation exposure and enhanced diagnostic capabilities for neuroimaging and oncology [125]. These hybrid systems minimize patient movement artifacts, improve image co-registration, and streamline clinical workflows, resulting in more accurate and efficient diagnoses.

Recent work by Sudarshan *et al.* [126] introduced a deep neural network uncertainty-framework to predict standard-dose PET images from a combination of low-dose PET images and multi-contrast MRI (acquired during simultaneous PET-MRI). The proposed transform-domain loss was inspired by the physics of the image acquisition process, that modeled the underlying sinogram-based physics of the PET imaging system. Their approach enhanced the robustness to unseen out-of-distribution acquisitions that differed from the training set distribution, which could arise from variations in radiotracers, anatomy, pathology, photon counts, hardware, and reconstruction protocol [126].

Software-based combined imaging modalities

Software-based fusion techniques involve the computational combination of images obtained separately from different modalities. Examples include dual-energy CT (DECT) fusion, where images captured at different energy levels produce virtual monoenergetic images, material decomposition maps, and virtual non-contrast images [122]. AI-based algorithms, such as convolutional neural networks (CNNs), further enhance DECT fusion by improving material differentiation and image quality [127]. Zhao *et al.* [128] developed a DL approach for DECT imaging that predicts high-energy images from low-energy data, facilitating accurate virtual non-contrast imaging and iodine quantification. Similarly, Li *et al.* [129] proposed an iterative neural network incorporating CNNs for high-quality image-domain material decomposition, demonstrating superior performance over traditional methods. These advancements underscore the potential of AI to enhance DECT imaging, enabling more detailed and reliable diagnostic information. MRI fusion techniques combine sequences like T1-T2 fusion for enhanced tissue contrast and diffusion-perfusion fusion for stroke evaluation [127]. Multi-modality fusion methods, such as CT-MRI, PET-CT, and SPECT-CT, integrate functional and anatomical data, allowing more accurate diagnosis and improved treatment planning. Additionally, ultrasound-MRI fusion supports real-time image guidance for procedures like biopsies [130].

3.2.7 Image quality challenges: artifacts and technical limitations

Medical image quality is critical for the precise and reliable development of DL algorithms in diagnostic and analytical processes. Several inherent challenges in medical imaging modalities, such as artifacts, technical limitations, and data heterogeneity, have a substantial impact on diagnostic outcomes, as will be explained in the next paragraphs.

Artifacts

Artifacts in medical imaging are unintended distortions or errors which can compromise image quality and hinder accurate interpretation. They may result from various factors, including patient movement, physical limitations of the imaging modality, hardware or software anomalies, and image processing techniques [51]. Patient movement, such as involuntary motion during scans, is a common source of motion artifacts, leading to blurring or ghosting effects that degrade the quality of MRI and CT images [131]. Imaging physics constraints, such as the differential absorption of X-ray photons by tissues of varying density, result in beam-hardening artifacts in CT, which manifest as streaking patterns that obscure diagnostic details [131, 51]. Hardware limitations, including imperfections in imaging systems or detectors and software errors during data reconstruction, can also introduce image distortions [51, 131]. Additionally, image processing techniques, including under-sampling or compression, can lead to aliasing artifacts in MRI, where high-frequency signals are misrepresented as lower frequencies, complicating accurate image interpretation [51]. These diverse sources highlight the multifactorial nature of artifacts and their significant impact on diagnostic accuracy.

The latest advances in DL and generative AI have shown potential in reducing artifacts and enhancing image quality. CNNs, GANs and diffusion models have been successfully employed for artifact removal, de-noising and synthetic data generation [132]. Recent novel approaches, including StylEx [133] and Dual-Domain Optimization [134], further address challenges such as model interpretability and edge artifacts, improving the overall reliability of diagnostic imaging.

Technical limitations

Technical limitations of imaging modalities and physical protocols are another factor that can degrade image quality. Resolution constraints, such as limited spatial resolution, hinder the detection of small lesions or fine anatomical structures, particularly in modalities like US or LDCT imaging [51]. Similarly, insufficient voxel intensity contrast between tissues complicates the differentiation of structures with shared attributes, a common challenge in soft tissue imaging [135]. Discrepancies among imaging protocols, including the use of intravenous contrast agents, or sequence adjustments like acquisition time after administering contrast agents, significantly impact image characteristics. In addition, longitudinal inconsistencies caused by technological advancements (e.g., transitioning from 1.5T to 3T MRI systems) further exacerbate these issues [136].

Medical image data also presents challenges related to its dimensionality and associated metadata. Variability in voxel spacing, along with heterogeneity across imaging modalities and protocols, leads to notable differences in resolution and comparability between different studies [137]. Anisotropic voxels, which arise from discrepancies in slice thickness and inter-slice gaps in CT and MRI, increase the complexity of data analysis [138]. The presence of anisotropic voxels can distort the representation of anatomical structures, which is especially problematic when attempting to detect subtle changes, like those observed in brain tumor studies or vascular pathologies. Furthermore, inconsistencies in file formats, such as the original DICOM and its conversion into alternative simpler formats, hinder data sharing and interoperability, potentially resulting in critical misinterpretations of contextual information stored in metadata headers [52, 139]. Medical data interoperability is crucial not only for accurate diagnosis but also for continuity of care over time. Inconsistencies in formats can create gaps in critical information stored in metadata, which could lead to incorrect diagnoses or delays in patient care.

DL and generative AI algorithms have shown potential to address the aforementioned issues. Super-resolution techniques enable the reconstruction of high-resolution images from low-resolution inputs, facilitating the detection of small lesions and intricate anatomical details, particularly in LDCT and MRI [140]. Image enhancement and denoising methods, including recent algorithms that leverage autoencoders and GANs [141], improve tissue differentiation and overall image quality by mitigating contrast and noise limitations [142]. Moreover, multimodal fusion techniques integrate data from multiple imaging modalities, providing a more comprehensive representation of anatomical and pathological features [122]. Hybrid methods that incorporate physics-based constraints into DL frameworks, known as physics-informed machine learning (PIML), further enhance image reconstruction by combining data-driven and theoretical approaches for improved performance [142], as will be discussed in the next section. Furthermore, CNNs have demonstrated strong potential for precise segmentation of 3D images, enabling more accurate identification of complex structures like blood vessels and tumors. These techniques are beginning to revolutionize the fields of functional MRI and CT imaging, improving surgical planning and early disease detection.

3.3 Physics-informed machine learning

Throughout the previous section several examples of physics inspired AI algorithms in different medical imaging modalities have been presented. These algorithms leverage fundamental laws of physics underlying medical images to enhance AI applications, bridging the gap between natural image and medical image computer vision. The study of physics informed machine learning (PIML) has the potential to enhance explainability, consistency, physical plausibility, robustness and generalizability. Prior knowledge from medical images can act as a regularization mechanism to limit the range of acceptable solutions [143]. For example, physics can be leveraged in generative models in the form of constraints to avoid creating non-realistic images, such as modeling echo attenuation [121], reflection and scattering in US imaging, or simulating T1 and T2 relaxation phenomena in MRI [144], among many others. Hence, research into PIML algorithms has experienced an exponential growth in the last years, and it is expected to continue evolving. In this section, we will focus on the main characteristics of these hybrid methods where laws of physics are combined with AI to build more reliable algorithms.

Physics-informed learning can be defined as the process of leveraging prior knowledge derived from observational, empirical, physical, or mathematical understanding of the world to enhance the performance of a learning algorithm [145]. Existing approaches are grouped by their use of physical principles to modify input data (observation bias), training losses (learning bias), and network architectures (inductive bias), as explained in recent general reviews [146, 145], and a targeted review focused on medical image analysis tasks [147]. These approaches can be combined to enhance PIML systems, resulting in more sophisticated hybrid methods.

In addition to physical modeling, certain approaches incorporate domain-specific priors derived from empirical observations. For clarity, we distinguish between priors grounded in physical principles and those informed by heuristic or data-driven assumptions. The following paragraphs present a general overview of existing approaches to leverage prior knowledge in ML algorithms, concluding with a discussion of the key challenges in the context of medical imaging. Table 3.1 summarizes the types of biases and the origin of the prior knowledge, including examples of representative methods, particularly in medical imaging.

3.3.1 Observational biases

Observational data can implicitly encode domain knowledge, making it one of the simplest ways to introduce biases in AI algorithms [145]. While this approach does not impose physical laws explicitly, the training data may capture them to some extent. Deep neural networks, when exposed to diverse and representative data, can learn to approximate the underlying physical processes through pattern recognition [146]. The main challenge lies in acquiring sufficiently large and high-quality datasets to reinforce these biases and produce robust predictions. However, the success of this method relies on the availability of large, high-quality datasets. In medical imaging, acquiring large, diverse datasets is often challenging, which has

Table 3.1: Taxonomy of bias types relevant to machine learning algorithms, including their associated priors, representative methods, and examples in medical imaging applications.

Type of bias	**Nature of prior**	**Representative methods**	**Examples in medical imaging**
Observational bias	Empirical	Data augmentation; synthetic image generation (GANs, VAEs)	Synthesis of additional weighted images in MRI [148]. Synthetic dual-energy CT images generated from single-energy CT [83].
Learning bias	Physical or empirical	PINNs, regularization terms in the loss function	Physics-guided material decomposition model for PCCT [85]. Physics-informed loss for motion-corrected quantitative brain MRI reconstruction [103].
Inductive bias	Physical or heuristic	CNNs, CapsNets, Transformers; GNNs, neural ODEs, Hamiltonian and Lagrangian NNs	Physics-based diffusion model for US image generation [121]. Neural ODE for accelerated MRI reconstruction [100].

led to the growing use of AI-generated synthetic data to augment training sets and reduce dependence on real-world samples. While this empirical strategy can enhance generalization, producing realistic and clinically meaningful synthetic images remains a challenge [149].

3.3.2 Learning biases

Learning biases are prior assumptions that implicitly embed prior knowledge in the learning algorithm through soft penalty constraints (regularization) added in the loss function [146]. Learning biases can be expressed as integral, differential or even fractional equations to promote convergence towards physical plausible solutions. However, the underlying laws of physics can only be approximately satisfied [145]. Physics-informed neural networks (PINNs) incorporate the knowledge of the physics of the process in the form of partial differential equations (PDEs) that are embedded into the loss function using automatic differentiation to calculate differential operators [146]. More recently, sef-adaptive PINNs (SA-PINNs) allow adaptive training of neural networks by applying trainable weights to each training point, which enable to focus on challenging regions of the solution space [150].

Furthermore, learning biases can also be heuristic, when based on empirical observations or domain expert assumptions not explicitly derived from first principles. For example, medical domain knowledge can be integrated in the form of learning biases in ML models. This medical informed ML promotes adherence to clinical guidelines, and benefits ML models enhancing model performance, interpretability, data efficiency and generalization, particularly in scenarios with limited data or expert availability, medical uncertainty or poor data quality [151, 152]. Prior expert knowledge can be represented as equations, simulation results, spatial invariance, logic rules, knowledge graphs, probabilistic relations and human feedback [153]. The integration of medical knowledge in ML models is further explored in Chapters 6 and 8.

3.3.3 Inductive biases

Inductive biases can be seen as a hard generalization of learning biases, such that prior assumptions are directly forced into the architecture of the model through specific design interventions. These are typically hard-coded and, when rooted in physical principles, ensure that predictions inherently comply with a defined set of physical laws. Examples include Hamiltonian and Lagrangian neural networks, Neural ODEs, and more general PINNs, utilizing kernels directly derived from the fundamental physical principles of the problem, which encode conservation laws or energy-based formulations [100, 146, 145]. Alternatively, some inductive biases are more general-purpose and data-driven in nature. For instance, CNNs enforce translational invariance [6], while capsule networks (CapsNets) promote translational equivariance [154], and transformers impose permutation equivariance [155], as discussed in Chapter 2. Graph neural networks (GNNs) and kernel methods such as Gaussian processes may also incorporate domain-specific structural knowledge, though not necessarily grounded in physics.

In other fields such as biology, the architecture of biologically informed neural networks (BINNs) is explicitly constrained by biological pathway ontologies, which are designed using an underlying graph that encodes known pathway hierarchies from databases like Reactome, Gene Ontology or Kyoto Encyclopedia of Genes and Genomes (KEGG) [156, 157]. In the network, each node corresponds to a real-world biological entity, for instance, a gene, pathway, or biological process, while the edges represent established relationships between these entities [156]. The application of BINNs to circulating protein markers in a lung cancer screening cohort is detailed in Chapter 7.

3.3.4 Challenges

Introducing physics constraints into AI models to build more trustworthy medical image applications requires finding an optimal balance between the complexity of physics-based constraints and data-driven approaches to better capture real world dynamics and enhance generalization [146]. Domain expertise is necessary for selecting the most suitable physics prior to be modeled by the algorithm, and the best approach to introduce it in the model should be carefully considered. Incorporating excessive constraints during training can lead to over-fitting and over-regularization, therefore, ablation studies and quantitative assessments are necessary to evaluate the influence of such constraints on the model's performance [147]. Additionally, it is essential to consider the scalability of these systems in real-world medical environments, as incorporating too many physical constraints could limit the model's adaptability across different clinical settings. Although PIML represents a promising direction in medical imaging research, limitations remain in the explainability of these algorithms, particularly in understanding how physical constraints interact with learned features [146]; managing uncertainty to avoid overconfident models, and dealing with incompleteness of knowledge due to the inherent difficulties of modeling all possible phenomena.

3.4 Conclusions

This chapter has provided an overview of the underlying physical properties in medical imaging for AI applications. This knowledge enhances explainability, and enables more reliable DL architectural designs, particularly for applications in image generation and reconstruction. PIML represents a promising strategy for embedding prior knowledge into AI algorithms, contributing to the development of more trustworthy systems while bridging the gap between natural and medical image analysis.

This chapter has particularly focused on generative AI, which is transforming the landscape of medical imaging. Both image reconstruction and synthetic image generation require an understanding of the acquisition process specific to each modality. While the field of generative AI is evolving rapidly, a strong foundation in physics remains essential for reliable interpretation of medical images. At the same time, careful consideration is required when introducing generative models into clinical practice, as poor generalization or unrealistic outputs could mislead clinicians and compromise image integrity. Foundation models, that incorporate data from multiple modalities, have emerged to enhance medical image analysis and can further benefit from prior knowledge. These models combine different modalities (multimodal) with various data types (text, image, video) and scales (cell, tissue, organ, patient, population). By integrating prior domain knowledge in the form of physics, biology, or medical expertise, foundation models can improve diagnostic accuracy and contribute to more personalized and reliable clinical decision-making in the context of precision medicine.

The clinical translation of AI systems in healthcare relies on building trustworthy algorithms that capture the complexity of real world data. Medical expert knowledge, physics in medical imaging, and biology in life sciences, play a fundamental role in embedding prior knowledge into AI algorithms, providing complementary information to current data-driven methods with the aim of enhancing the learning process. The properties of trustworthy AI will be presented in the next chapter, while Chapters 6 and 8 will illustrate how the incorporation of learning biases and inductive biases enhances the robustness and explainability of DL models in medical imaging.

Chapter 4

Challenges of deep learning and radiomics in medical imaging

Main publication associated with this chapter: **Cobo, Miriam**, *et al.* "Enhancing radiomics and Deep Learning systems through the standardization of medical imaging workflows." *Sci Data* **10**, 732 (2023). DOI: `https://doi.org/10.1038/s41597-023-02641-x`.

4.1 Introduction

Computer aided diagnosis (CAD) systems are transforming diagnostics and therapeutics in healthcare with autonomous systems that aim to assist clinicians in their work, improve patient care, and develop novel ways to discover new treatments and diagnostics in the laboratories [158]. However, there are several factors that currently hinder the generalized adoption of these systems in clinical practice, expanding from challenges in data collection to the effective implementation of AI algorithms in medical workflows.

Significant efforts have been dedicated to advancing computer vision technologies for radiology, an inherent digital image-based specialty, with increasing interest driven by the rising demand for clinical imaging and the global shortage of radiologists [159]. For example, in oncology, medical imaging is the reference to evaluate most cancers, in particular for lesion detection and staging, which proves the need for general standards and guidelines in radiology to advance research in CAD systems for digital diagnosis. Medical images play a key role not only in diagnosis, but also in monitoring the progression and development of tumors, in addition to supervising the response to therapy and risk of relapse [160, 161]. This role expands beyond oncology to other areas of medicine, such as neurology, cardiology and pulmonology. In this chapter, we will describe some of the existing limitations, in particular related to the lack of standardization and interoperability, and we will introduce trustworthy AI [162, 163], which is a multi-faceted concept grounded in several key principles, such as transparency, robustness, fairness, and accountability. These principles are essential to build trust and ensure acceptance in medical applications, both of which are necessary to achieve real-world clinical adoption.

4.2 A deep look into radiomics

Radiomics is the quantitative evaluation of medical images, which enables the extraction and analysis of predefined hand-crafted semi-quantitative (e.g., attenuation, shape, size, and location) and/or quantitative features (e.g., wavelet decomposition, histogram, and gray-level intensity) with the goal of developing predictive or prognostic models [164, 165]. Quantitative image descriptors in medical imaging have emerged as noninvasive prognosis phenotypes and predictive biomarkers [166, 167]. Particularly in oncology, these noninvasive techniques reach the whole tumor volume [168], in contrast with pathological examinations, which require biopsies or invasive surgeries to analyze only a limited sample of tumor tissue that may not be representative of the whole lesion due to its heterogeneity [166]. Radiomics and radiogenomics have shown potential to complement pathological diagnosis [169, 167, 170], yet the successful integration of these workflows into clinical practice requires addressing several standardization and interoperability challenges.

Conventional radiomics extracts pre-designed features from a segmented region of interest (ROI) corresponding to the tumor [165]. This approach heavily relies on the segmentation contour, and the image characteristics (type of scanner, acquisition protocol, etc.), which in many cases impede generalization to new settings, as will be discussed in the next section. Additionally, manual annotations increase the radiologists' workload, and are subjected to inter-observer variability [165]. Interestingly, radiomics is not exclusive to oncology, and can be applied to a wide range of medical imaging modalities, from MRI, CT, US, PET and SPECT [171, 172, 173]. The extracted radiomics features can be analyzed with statistical methods and ML models.

In contrast with conventional radiomics, deep neural networks do not necessarily need the segmented ROI, and the extracted features can be analyzed internally within the same model or go through a different analyzer [165]. Thus, features can be automatically extracted in an end-to-end process where no prior knowledge is necessary [165].

Throughout the next section, an overview of current methods in preprocessing and harmonization of radiomics features will be presented, alongside limitations of both radiomics and DL based CAD systems, emphasizing the need for standardized workflows in medical imaging.

4.3 Limitations of radiomics and deep learning

The translation of computer vision advances into clinical practice is currently being delayed due to the lack of standardization and harmonization of radiology clinical protocols and workflows [174]. The potential of AI to revolutionize the state of the art in medical imaging requires a paradigm shift from individual to collective standards, particularly in data collection and preprocessing. This shift will also enable the transition of research from retrospective studies to clinical trials and generalized adoption.

Several reviews of publications discussed by Hadjiiski *et al.* [175] reveal that most current ML models are far from being ready for real-world clinical deployment. These models lack sufficient reproducibility, rigorous validation, generalizability to external datasets, and robustness to translate into clinical practice [163].

4.3.1 Data acquisition and preprocessing

Regarding data collection and preprocessing, there is a wide variability between manufacturers that implement distinct reconstruction algorithms, and institutions that utilize different reconstruction parameters, which may also be customized for each patient [176]. Orlhac *et al.* [177] showed in CT that scanner parameters such as reconstruction kernel or slide thickness influence radiomics texture features. Moreover, Son *et al.* [178] revealed that similar CT protocols and same slice gaps in data from different hospitals led to an improved performance of ML algorithms. Rizzo *et al.* [176] proposed identifying and excluding radiomics features highly influenced by the acquisition and reconstruction parameters, however, this solution may limit the power of radiomics analyses. Image quality is another factor that impacts the performance of radiomics systems, particularly if the equipment has become obsolete compared to modern devices [175]. In case the images come from different sources (manufacturers, hospitals) a similar distribution of “positive” and “negative” cases needs to be ensured to train an AI algorithm [175]. Our research into the factors influencing internal validation and generalization of DL neural networks in chest X-rays revealed that model generalization is significantly impacted across devices with different types of response functions, followed by variations in image processing and inter-institutional differences [20].

Furthermore, preprocessing steps like filtering, resampling and morphological image processing also have an impact on radiomics features, as illustrated in Figure 4.1. In this context, Soleymani *et al.* [179] conducted a phantom study in CT to assess the reproducibility of radiomics features across varying ROI sizes, image resolutions, and Hounsfield unit (HU) values. The authors concluded that standardizing radiomics features is essential to ensure consistency across different imaging conditions. Finally, for AI systems, data augmentation and synthetic data generation should not alter the images in a way that the underlying biological or tissue properties become implausible [149, 175].

There have been some attempts in the literature to provide guidelines to preprocess medical images for conventional radiomics. Van Timmeren *et al.* [171] enumerate some of the necessary steps before radiomic feature extraction, such as interpolation, normalization and discretization. However, the authors highlight that many questions regarding these steps remain open. Aerts *et al.* [166] performed radiomics analysis from the raw imaging data (before the images are reconstructed), without any pre-processing or normalization, yet a strong dependence of their radiomic signature on tumor volume was later revealed by Vallieres *et al.* [180].

To standardize radiomics features, ComBat harmonization is a batch-effect correction [181] that aims to suppress batch effects by standardizing the means (location) and variances (scale) of each feature across batches to reduce the batch effect error [182, 183]. This algorithm is based on an empirical Bayes approach, originally developed for genomics data [184], later applied to reducing radiomics variability in PET [185], and CT [177, 186]. There are other variations of the algorithm, such as longComBat [182], developed for longitudinal data. Overall, ComBat is intended to harmonize radiomics features, thereby minimizing the impact of different acquisition protocols on radiomics feature extraction, which is particularly useful for retrospective studies, where it would be impractical -or even impossible- to re-image patients to a controlled imaging protocol [186]. Ligero *et al.* [183] applied ComBat considering different sources of variance as batches: manufacturer-dependent convolution kernel, slice thickness, and the combination of both. Their results showed that ComBat correction minimized radiomics data variability regardless of differences in CT acquisition protocols [183].

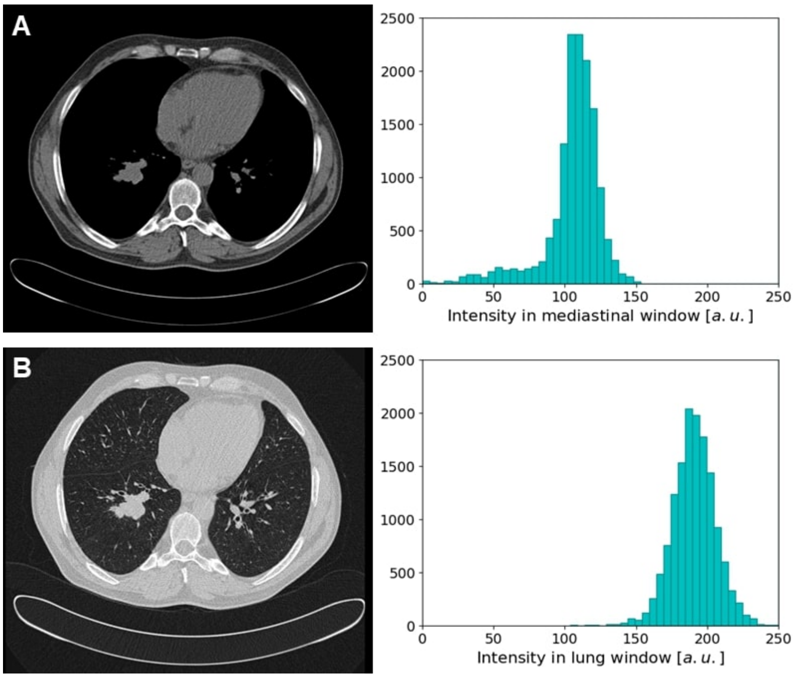

A
2500
2000
1500
1000
500
0
0
50
100
150
200
250
Intensity in mediastinal window [*a. u.*]
B
2500
2000
1500
1000
500
0
0
50
100
150
200
250
Intensity in lung window [*a. u.*]

Figure 4.1: Effect of different preprocessing steps on the same nodule and the corresponding histograms calculated for the nodule mask: (A) mediastinal window, (B) lung window (a.u. refers to arbitrary units).

In the study by Mahon *et al.* [186], ComBat harmonization proved to be effective by harmonizing radiomic features extracted from different imaging protocols, although emphasized that its effect on imaging feature–based predictive models requires further investigation. In fact, research is underway to analyze the power of ComBat harmonization in multicenter studies in various imaging modalities, for example, Leithner *et al.* [187] studied ComBat harmonization on PET/MRI and PET/CT for radiomics-based tissue classification. Furthermore, ComBat is generalizable to other imaging modalities as it makes no assumptions about the origin of the site effects [181].

The previous examples illustrate the need for general guidelines for medical image preprocessing in computer vision tasks, and the relevance of adopting standard scanning protocols across institutions to achieve consistency in the acquisition parameters.

4.3.2 Reproducibility and radiomics standardization initiatives

The clinical utility of an algorithm highly relies on the quality of the reference standard used in its training and evaluation [175]. Reference standards based on radiologists' opinion are subjective, especially if assessed only by a single expert, and should therefore be replaced whenever possible by objective reference standards, such as diagnostic tests and pathologic evaluation of biopsies or excised lesions, patient survival or time-to-progression for shorter-term reference standards [163, 175].

There are several standardization initiatives and imaging protocols investigating homogenization of image biomarkers and radiomics features, such as the Image Biomarker Standardization Initiative (IBSI) [188], the Quantitative Imaging Network of the National Institute of Health

(QIN) [189], the Quantitative Imaging Biomarkers Alliance (QIBA) [190], and the European Imaging Biomarker ALLiance (EIBALL) [191], among others. Harmonization of the extraction and validation of robust radiomics features is essential to achieve results that are reliable and reproducible [192, 177, 193, 194], although it does not address the systematic variations between patient subpopulations [175]. The range of different standardization initiatives shows the need to reach consensus among the radiomics research community on joint standards.

Radiomics signatures are intrinsically data driven, which poses several challenges as the high number of features is susceptible to overfitting and overinterpretation of the derived models [170]. The development of radiomics signatures is significantly affected by underlying dependencies between radiomics features, redundancies and multicollinearity, as outlined by Welch *et al.* [192]. ML algorithms can be effective to identify unexpected effects, such as volume-confounding features [193, 194]. Recently, the lack of biological meaning of current high-throughput agnostic radiomics analyses has raised concerns. Tomaszewski *et al.* [170] emphasized the need of supporting radiomics with biological validations to gain insights into the casual relationships of the features with the outcomes.

Most published radiomics studies lack independent validations of their signatures beyond a single external test set [170], which is insufficient for their deployment in clinical practice. Independent validations of radiomics signatures on different cohorts and multiple institutions are hindered by the lack of standardization in medical imaging, although Shi *et al.* [195] have already proposed an approach for distributed radiomics. Therefore, to achieve generalization and robustness of radiomics signatures, further efforts are required to homogenize image acquisition and preprocessing [177], in addition to controlling the effect of potential confounders [194].

The previous paragraphs highlight the factors that affect the reproducibility of conventional radiomics features. The next section focuses on the interpretability and generalizability challenges in DL algorithms.

4.3.3 Black-box nature and lack of generalizability

A crucial aspect that compromises the translation of radiomics and AI tools into clinical practice is the *black-box* nature of most current DL systems [196]. For instance, in the European Union (EU), the General Data Protection Regulation (GDPR) establishes that individuals have the right to receive a clear and understandable explanation of how AI is being used to make decisions that directly affect them [197]. Explainable AI (XAI) is essential to gain the trust of physicians and understand the reasons behind a prediction or decision [175]. Besides, interpretability can detect biases and problems such as unbalanced data, and explainable models are more robust against adversarial attacks [198]. Post-hoc explanations like saliency maps are insufficient to provide a full explanation of why and how the features are connected and weighted to identify the target lesion. Provided explanations should align with medical knowledge or be supported by clinical evidence [175]. XAI will be further explained in Section 4.5.

The shortage of large enough datasets to train and externally validate radiomics signatures in prospective multi-center studies also happens for medical AI devices [199]. Several of the devices approved by FDA for diagnostic use were trained on small datasets from a single center or from only two centers [200]. These algorithms are prone to biases and lack of generalizability outside the site where they were trained [20]. In general, AI tools deployed in new clinical settings should be evaluated for their local clinical validity, and re-calibrated if necessary [163].

Public databases provide free validation datasets to the medical imaging community, however, as argued by Hadjiiski *et al.* [175], the quality assurance (QA) process for data in a public database is often overlooked. For example, the well-known lung cancer LIDC-IDRI dataset [201] includes the manufacturer in DICOM metadata, but not demographic information such as patient age or gender [202], which can lead to unexpected biases when developing ML models.

As outlined by Hadjiiski *et al.* [175], even if a hospital could use a vendor-trained CAD-AI tool with multi-institutional data and approved for clinical use, its performance in the local population could not be the same as in the vendor's specifications. Hence, the hospital would have to evaluate the performance of this tool on their patients in an adjustment phase, achieving a deeper understanding of the system's performance in the local setting, while reducing unrealistic expectations and improper use of the CAD-AI tool [163, 175].

To ensure data availability, accessibility and reusability, radiomics signatures demand stability and reproducibility across different hospitals, scanners and acquisition protocols, that is, the adoption of FAIR principles (findable, accessible, interoperable, reusable), as described by Wilkinson *et al.* [203], in a manner that preserves patient privacy [174]. Data collection must also conform to the ethical considerations and legal framework of the country in which the data were obtained [175]. Standardization extends to validation and evaluation criteria, providing guidelines and contrasted metrics to reduce bias and overly optimistic results hiding the lack of generalization of certain models subjected to highly restrictive data conditions and insufficient reporting [171].

The promise of CAD systems lies in their potential for noninvasive automated evaluation of medical images. The price will be standardizing the different workflows in image acquisition, preprocessing, annotation, anonymization, metadata, and storage processes.

4.4 Guidelines to achieve standardization

There are several public databases available with medical images, such as The Cancer Imaging Archive [204] or Neurovault [205]. However, the lack of standardization in database formats, i.e., limited interoperability, hinders the simultaneous integration of multiple data sources within the same ML algorithm. [174]. Thus, the change of paradigm from visual assessment of medical images to computer-aided evaluation demands for methodological standardization of the workflows in medical imaging as proposed in Figure 4.2. This standardization should implement the FAIR principles to the extent that the requirements due to the nature of medical images (de-identification, security) allow.

Data collection is a crucial step to create computer vision models and involves different agents within the hospital: radiologists, technicians, nurses, general practitioners, etc. Data interoperability is vital to facilitate research and multicenter studies, therefore, all the involved agents in data collection should become aware of methodological standards when these are adopted. We believe radiologists will play a key role in ensuring the correct application of standards and the effective adoption of protocols. There are two levels at which standardization of the workflows in image analysis should be implemented: software (consistency of technical implementation among scanners and manufacturers) and human interaction (coherence between different observers and practitioners) [173].

At human interaction, we identify two levels at which radiological studies should be labeled: study level (e.g. brain MRI FLAIR sequence, chest radiography AP, etc.) and pathology level

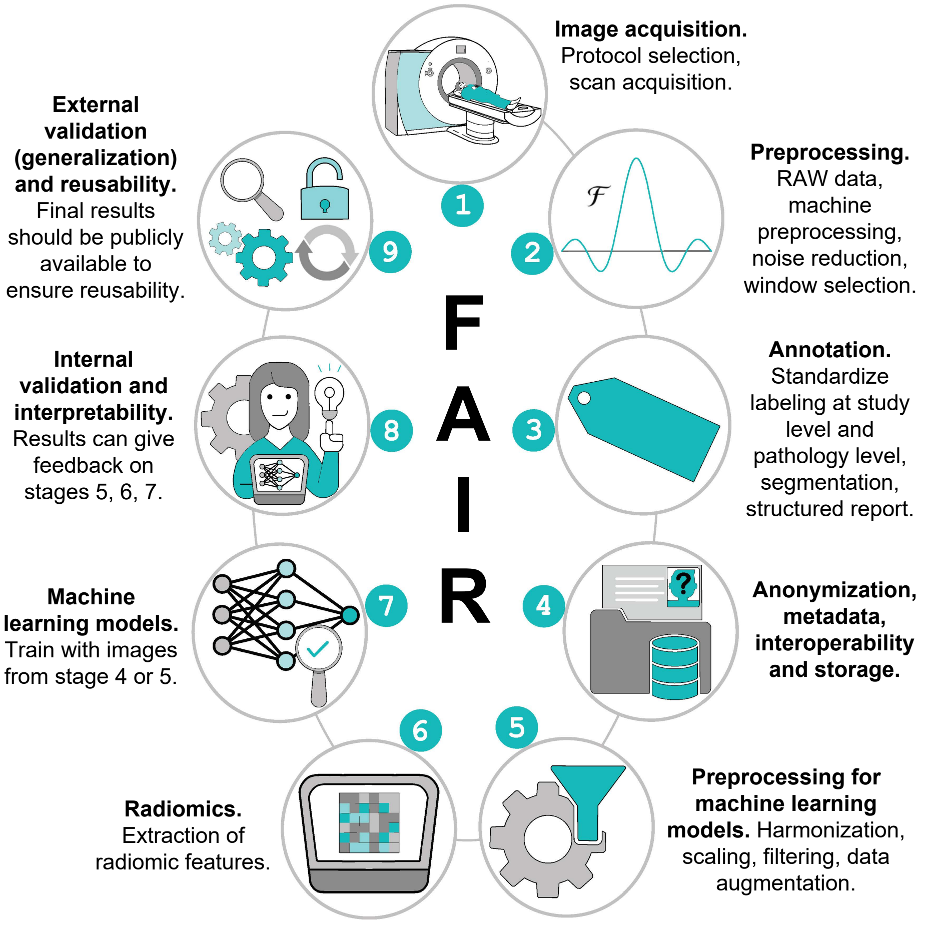


Image acquisition.
Protocol selection, scan acquisition.
Preprocessing.
RAW data, machine preprocessing, noise reduction, window selection.
Annotation.
Standardize labeling at study level and pathology level, segmentation, structured report.
Anonymization, metadata, interoperability and storage.
Preprocessing for machine learning models. Harmonization, scaling, filtering, data augmentation.
Radiomics.
Extraction of radiomic features.
Machine learning models.
Train with images from stage 4 or 5.
Internal validation and interpretability.
Results can give feedback on stages 5, 6, 7.
External validation (generalization) and reusability.
Final results should be publicly available to ensure reusability.
1
2
3
4
5
6
7
8
9
F
A
I
R

Figure 4.2: The nine stages of reaching standardization and making medical imaging data as FAIR as possible.

(e.g. tumor, benign nodule, etc.). The study level labeling relies on the work of technicians and nurses, who are responsible for the correct categorization of the data according to the type of study modality they have performed. Hence, in the study level labeling, the Series Description parameter in DICOM should correctly include the type of study modality that was carried out. Ultimately, the labeling at study level should be incorporated in the DICOM Study Description and Series Description fields, according to the RadLex lexicon [206] standard. Therefore, it is essential that this field is homogenized for each DICOM across all hospitals and scanners. In addition, the pathology level labeling should be incorporated into the structured report [207].

Regarding software, we believe that manufacturers' involvement in the process of standardization is essential, as they are in charge of bringing the latest technology to the clinic. To ensure their engagement, we propose that all leading radiological societies join forces to request the implementation of the necessary technology from the manufacturers. In particular, we acknowledge that standardization of MRI protocols for MRI-based radiomics is a challenge [208], due to the inherent versatility of this imaging modality. The experience of Sharma *et al.* [209] first reported a systematic inventory of MRI technology and personnel. They proposed the creation of a committee of stakeholders (radiologists, MRI physicists, technologists and scientists) committed to establishing and maintaining a standardized imaging strategy, with annual protocol reviews. In their conclusions, Sharma *et al.* [209] emphasized the need for better remote connectivity to MRI systems and increased automation in exam acquisition, including protocol selection, configuration, and parameter modification. In other medical imaging modalities, such as radiography or CT [210, 211, 212], the same process as in MRI could be followed, automating exam acquisition and parameter selection based on the patient's characteristics.

We propose the following guidelines to improve generalization of CAD radiomics and DL systems in radiology:

- Medical imaging datasets should always incorporate metadata information about the manufacturer and the acquisition protocol.
- Datasets' anonymization process should retain demographic information (e.g., age, gender, comorbidity, ethnicity) to avoid biases, as long as the patient cohort is sufficient to ensure patient de-identification.
- Datasets that include segmentations should provide metadata describing if the segmentation was manually performed, otherwise information describing the automatic or semiautomatic method that was used should be provided, including values of internal parameters in case of fine-tuning of the algorithm.
- Reference standards should be objective as far as possible, otherwise, independent evaluations should be secured from several experts with an assessment of the inter-reader variability.
- Hospitals should appoint a stakeholder committee within their staff to guide and monitor the standardization strategy, through a QA/QC process.
- All hospitals should adopt the same standards and guidelines to ensure interoperability.
- Radiomics and AI systems should include interpretable explanations in human understandable terms, similar to medical standards, on how and why they perform predictions or decisions to assist physicians.
- Datasets along with their metadata, and code if exists, should be made publicly available to allow reusability and reproducibility.

Standardization of computational statistics for radiomics-based systems should consider data balancing, sufficient patient population in size and diversity to prevent potential biases, interpretability, biological validation (relation of radiomic signature to cell morphology, density, distribution pattern, etc. [173]), generalization and suitability of performance metrics to the case of use, among other aspects. Ultimately, it is critical to continuously monitor the performance of radiomics and DL systems to ensure their efficiency does not degrade over time, the so-called data drift [163, 213], as clinical practices, protocols and patient demographics may change, with a corresponding impact on performance.

The constraints of conventional hand-crafted radiomics CAD systems, detailed in this chapter, and originally presented in the first research paper of this thesis [52], motivated a transition in the thesis towards DL-based approaches, with a particular emphasis on explainable AI algorithms. The following Section 4.5 defines the terms transparent, explainable and interpretable, while the final Section 4.6 introduces the multidimensional concept of trustworthy AI.

4.5 Explainable artificial intelligence

Deep neural networks are very powerful mathematical algorithms with thousands or millions of parameters. Their internal reasoning is difficult to interpret by a human, since they learn from patterns and correlations in the data through effective feature representations. Thus, DL algorithms have been criticized by their *black-box* nature [196], which hinders their widespread adoption in high-stakes decision-making processes, such as healthcare, as described in Section 4.3.3. The results of these models are often unexplainable, unjustifiable and unaccountable [214]. In contrast, XAI has emerged to open the *black-box* and provide understandable explanations to the decisions of DL models.

Within the literature, the terms 'transparency', 'explainability' and 'interpretability' are often used interchangeably. In this thesis, they will be defined following the same criteria as Angelov *et al.* [215]:

- Transparency: a model is considered transparent if, by itself, it has the potential to be understandable (as opposed to a *black-box* nature).
- Interpretability: the ability to provide interpretations in terms that are comprehensible by a human.
- Explainability: human-understandable interface between humans and the system [216]. It comprises AI systems that are accurate and comprehensible to humans, which encompasses summarizing the reasons for their behavior, and offering insights into the causes of their decisions [217].

XAI models are inherently interpretable, but not all interpretable models are necessarily explainable [217]. Both explainability and interpretability constitute concepts by themselves, rather than binary properties, and, therefore, they are considered multidimensional concepts [218]. In this regard, Nauta *et al.* identified 12 conceptual properties, a high level decomposition of explanation quality, such as completeness, correctness, and compactness, for a comprehensive evaluation of the quality of an explanation [218]. Given the overlapping characteristics of explainability and interpretability, these terms will be used interchangeably throughout this thesis [219].

Several techniques have been proposed to enhance explainability and interpretability in ML models. These methods can be divided into two main categories:

- *Post-hoc* explainability: an explanation method is applied in an attempt to gain insights into the learning process of an already trained DL model. This approach tries to provide faithful explanations to what the original model computes. However, these explanations cannot fully replicate the original model, since, if they could, the *black-box* model would be unnecessary [196]. Post-hoc methods are integrated after creating and training the model. Examples include interpretability saliency maps for images [220], or importance-based methods and SHAP values for computing feature importance in tabular data [221]. The aforementioned methods attempt to provide insights into which features are driving the predictions of the model [4].
- *In-model* explainability: the explainability is implemented by design into the DL model, building an inherently explainable model [218]. A common misbelief is that there exists a trade-off between accuracy and interpretability [196], however, this is merely a consequence of most research efforts being focused on non-explainable DL models.

> Explainable by-design models introduce interpretability constraints (e.g., sparsity, learning biases, as discussed in Chapter 3), attention mechanisms, example-based explanations, or can be directly *white-box* models such as linear regression [214, 218].

In both categories explanations can be local (referred to a single instance) and/or global (for the whole model). There is a third less common category which are supervised explanation methods, that train the model providing a ground-truth explanation [218].

Another useful technique for analyzing complex models is the ablation study, which involves successively removing components of the model to assess their contribution to the final decision. This approach can also facilitate the development of simpler models when the removal of certain parts does not result in a significant drop in performance.

XAI is necessary in high-stakes decisions, such as healthcare applications, to allow end users to interpret the model and its outputs, assess its strengths and limitations, and make informed decisions on its use, including whether to rely on it or not depending on the situation [163]. Incorporating end-user considerations into the design and development process of XAI models is essential [214]. The FUTURE-AI framework [163], that provides guidelines for trustworthy and ethical AI in healthcare, defines two recommendations for explainability: defining explainability needs and evaluating explainability.

4.6 The road to trustworthy artificial intelligence

Trustworthy AI is a fundamental concept to ensure the safe use of AI algorithms in the clinics, which involves a wide range of factors including robustness, security, transparency, explainability, fairness, and safety [222, 162]. In 2019, the European Commission published a guideline composed of seven key requirements that should be met for trustworthy AI [223], which complements the GDPR on individual's *right to explanations* for AI decisions [197]. The seven key requirements for trustworthy AI are:

- Human agency and oversight.
- Technical robustness and safety.
- Privacy and data governance.
- Transparency.
- Diversity, non-discrimination and fairness.
- Societal and environmental well-being.
- Accountability.

The FUTURE-AI framework structured the guidelines for trustworthy healthcare AI around six guiding principles: fairness, universality, traceability, usability, robustness, and explainability [163]. The next chapters of this thesis focus on enhancing robustness and transparency of DL models, which are related to two of the seven requirements for trustworthy AI, and two of the six guiding principles of the FUTURE-AI framework.

Chapter 5

FAIR principles in clinical informatics data preprocessing for artificial intelligence algorithms

Main publication associated with this chapter: **Cobo, Miriam**, *et al.* “Applying the FAIR principles in clinical informatics data preprocessing for artificial intelligence algorithms”. *To be submitted.*

5.1 Introduction

Reproducibility in medical imaging and omics data for deep learning algorithms strongly relies on preprocessing, yet this process lacks well-defined guidelines and standardization, in contrast with the well-known FAIR principles for research data, presented in Section 4.3.3. As datasets are shared to enable scientific discovery, preprocessing steps before training machine learning algorithms are crucial to ensure adaptability and reproducibility. For this purpose, it is necessary to guarantee that researchers report preprocessing pipelines in clinical informatics following FAIR principles. To encourage consistent and transparent FAIR reporting of data preparation in clinical informatics, this chapter presents a set of best practices designed to establish the minimum principles, increase awareness in this topic and foster further discussion.

5.2 Data preprocessing as the first step in machine learning pipelines

In the context of data preparation, **preprocessing** refers to the pipeline that encompasses the selection, preparation and curation of data for subsequent use in machine learning (ML) models. This process includes different steps in clinical informatics depending on the specific data modality, e.g., images, electronic health records, omics data, etc. The purpose is to transform the initial data, and the corresponding labels, into a suitable format and structure for input into the ML model [224]. Preprocessing plays a crucial role in the development of ML algorithms, since the quality of the input data directly influences the learning process and the output predictions. Therefore, regardless of the modality, preprocessing demands standardization in clinical informatics to ensure high quality data management, facilitate reusability and reproducibility [225].

The recent FUTURE-AI framework [163], presented in Section 4.6, emphasizes the need to systematically report preprocessing and annotation workflows to address technical and human biases. The widely known FAIR principles were designed to enhance the value and impact of scientific digital objects [203]. These principles have been widely adopted by researchers, organizations, and regulatory entities, as they provide a standardized framework for improving data collection, curation, organization, and storage [163]. The application of these principles naturally extends to preprocessing and data preparation pipelines before training and validating ML models. This chapter gathers a detailed list of best practices that take inspiration from the FAIR principles, extending the general guidelines in the FUTURE-AI framework to standardize the reporting of data preprocessing in clinical informatics.

Preprocessing is often an overlooked step in the literature, yet it plays a crucial role in the model's performance and generalizability, preventing data leakage and ensuring reproducibility, two interconnected challenges in ML research [4, 224]. Using the entire dataset for any pre-processing steps or ignoring temporal dependencies in time-series data results in data leakage [4, 226]. For instance, steps such as normalization, imputation of missing values, oversampling, data augmentation, dimension reduction or feature selection before data splitting cause an imperfect separation between training and test sets [4, 224, 226], ultimately leading to overoptimistic results and lack of generalizability and reproducibility. Data-informed splitting requires domain-specific knowledge and additional bioinformatics tools depending on the application [224]. Hence, a clear and standardized reporting of preprocessing steps is necessary to ensure high quality evaluation and auditing of ML models. However, in practice, this level of reporting is often absent or incompletely documented in the literature. This problem is particularly prevalent in conference proceedings with space limitations, where there is usually no dedicated section clearly detailing data preparation and preprocessing steps, which hinders reproducibility [3, 227, 228]. Even if the code is released, this remains suboptimal since it can be open to interpretations and, therefore, prone to errors, particularly when reproducing an existing model on new datasets that did not follow the exact same collection criteria as the original data. This scenario is common in clinical informatics, where procedures continuously evolve and adapt to enhance healthcare delivery. Thus, releasing the code is encouraged but does not guarantee reproducibility of the results [229], highlighting the importance of documenting preprocessing in both the main text and supplementary materials. In Chapter 4, guidelines were presented towards standardization of medical imaging workflows, and the impact of different preprocessing steps on radiomics features was examined. It was emphasized that medical images vary significantly depending on the selected Hounsfield Unit (HU) window (Figure 4.1), which affects the visualization, analysis and interpretation of anatomical structures. In this chapter, these concerns are generalized to preprocessing in clinical informatics data, spanning from medical images to omics datasets, providing general principles applicable to diverse data types. To facilitate understanding, visual examples are presented primarily from medical images, as they offer a more intuitive and straight-forward interpretation than other modalities, such as omics data. Figure 5.1 illustrates the relevance of preprocessing in the life-cycle of an ML model.

Foundational models (FM) have attracted significant attention in healthcare in the past few years, offering new opportunities for scalable, data-driven insights across a wide range of clinical tasks. These models have shown potential to mitigate generalization issues in applications

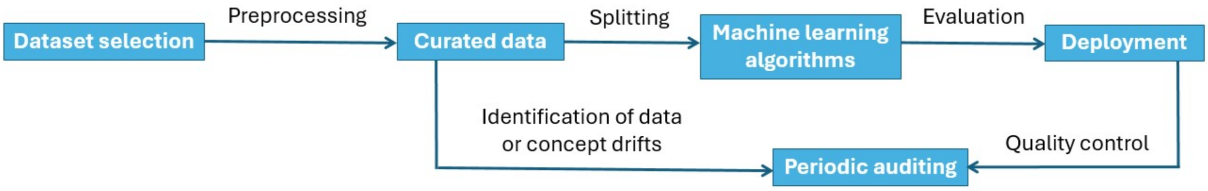


Figure 5.1: The role of preprocessing in the training, evaluation and auditing of machine learning algorithms.

involving small or imbalanced datasets [230]. Pretraining FM for downstream tasks in healthcare faces challenges associated with diverse data types, quality issues, inter-subject variability and integrating heterogeneously informative modalities [231]. Addressing these limitations requires comprehensive data preparation workflows, that will also benefit from standardized domain-specific preprocessing.

This chapter aims to raise awareness within the scientific community on the vital role of preprocessing in clinical informatics applications. Simple, yet effective, guidelines and best practices are proposed to promote a FAIR reporting of all the key preprocessing steps to prepare clinical informatics data for developing and validating ML algorithms.

5.3 An example of pitfalls in medical imaging preprocessing

Medical imaging preprocessing is a crucial step to ensure correct sizing and management of medical images before passing them to deep learning (DL) algorithms. Effective preprocessing enhances model performance by reducing variability, optimizing resolution and uniformity, while preserving clinically relevant information [140]. Seoni *et al.* [232] review image harmonization techniques in multi-center/device studies, emphasizing their impact on improving model performance, generalizability, and mitigating biases. For instance, histopathology images, characterized by their high resolution and large file sizes, require systematic preprocessing, including resizing, normalization, and color correction to ensure consistency, while preserving tissue morphology. Additionally, stain normalization techniques are commonly applied to standardize variations in staining protocols across different laboratories, improving the ability of DL models to learn meaningful patterns rather than artifacts. Similarly, in radiology, preprocessing can involve intensity normalization, HU windowing in computed tomography (CT) scans, or bias field correction in magnetic resonance imaging (MRI) to standardize contrasts and improve feature extraction, as described by Masoudi *et al.* [233]. Preprocessing strategies should always be assessed with clinical knowledge, and adapted depending on the task and type of data, since they can have an impact on the model's interpretation of the images [234]. Image registration, noise reduction, and segmentation are also critical steps in preparing medical images for ML analysis. Figure 5.2 shows an example of different preprocessing techniques applied to a lung nodule in a CT scan to illustrate the effect on how the algorithm "sees" the data. These examples are inspired by methods described in the literature, even though preprocessing steps were not consistently reported [3, 235]. Transparent reporting of each step performed is essential to avoid misinterpretations, promoting reuse of existing DL models.

In medical imaging, DICOM (Digital Imaging and Communications in Medicine) is the

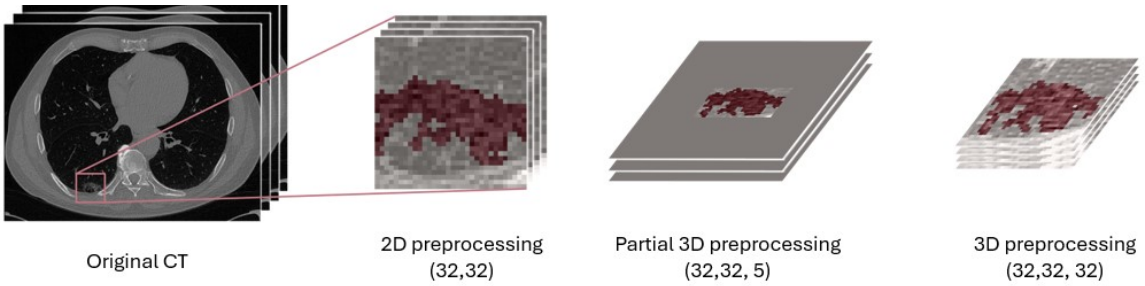


Figure 5.2: Different preprocessing techniques on the same nodule in a CT scan.

universal standard that defines and controls the formats for sending, distributing, and storing medical images across different machines, manufacturers, and imaging modalities [57]. The transformation of original DICOM files into other medical image formats should preserve metadata to document the steps performed during the conversion, as well as keep information regarding the original device, acquisition parameters and, if possible due to the nature of the data, demographic information [236, 52]. For instance, the selection of HU windowing should always be reported when converting DICOM to PNG to ensure reproducibility and proper interpretation. However, this standard practice is often overlooked in the literature [237, 238, 239], potentially causing inconsistencies in subsequent data analysis and limiting reusability. It should also be noted that some preprocessing steps can be irreversible if the dataset is released in the derived format, e.g., when performing intensity normalization the original intensity values cannot be recovered post-normalization [240].

This section contains examples focused on medical images, but the same process is generalizable to omics data [241], electronic health records [242] and, in general, any clinical informatics data modality. Overall, robust preprocessing requires the integration of domain knowledge with a comprehensive understanding of the data and existing standards.

5.4 Towards FAIRness in data preprocessing

Data preprocessing steps are performed to optimize the ultimate objective of the ML algorithm [233]: classification, detection, segmentation, and representation learning, including self-supervised approaches. These preparation and curation pipelines are expensive in terms of time, effort and resources. The **data management plan** (DMP) is a dynamic structured document plan that describes how data will be handled throughout a project, from collection and organization to quality control, documentation, and usage [243]. Additionally, it outlines strategies for data preservation and sharing, ensuring compliance with established policies and facilitating future accessibility. **Data provenance** refers to the origin, processing, movement and storage of data [244]. This concept is key to account for the distinct versions of the original data that can be derived from different subsequent preprocessing pipelines. Both the DMP and the data provenance are resources to document and standardize reporting of preprocessing pipelines. Fraga-González *et al.* recently provided a series of affordable approaches to achieve a reasonable degree of *FAIRness* (not to be confused with fairness in the context of algorithms or ethics, which refers to the absence of bias or discrimination in decision-making processes [163]), enabling data reusability after publication [245]. These recommendations are applicable to any scientific field, and we particularly emphasize in the next lines the relevance of FAIR reporting in clinical informatics data preprocessing.

Findability. The dataset preprocessing and metadata should be easy to find, including explicitly the identifier of the original raw data. We encourage researchers in clinical informatics to integrate in their pipelines DMPs, and we suggest international conferences and journals to enforce sharing the DMP's link, together with code and dataset's repository if applicable, detailing preprocessing steps and data provenance with the public code in persistent identifiers. This way, space limitations will not prevent authors from achieving high-quality reporting.

Accesibility. The preprocessing and associated metadata describing all the methods, tools and conversions should be preserved and archived in a long-term registry or repository. Whenever possible, metadata associated with preprocessing steps should be reported in a standardized manner. For example, in medical images, DICOM supports versatile and standardized big data management [236], and we recommend keeping this original format and leveraging existing software tools for reading and writing DICOM fields.

Interoperability. Preprocessing should be performed with standard tools that allow correct versioning reporting and automation, including explicit references to these tools and their internal parameters in the metadata.

Reusability. This principle ensures that data preprocessing pipelines are both usable (i.e., reproducible) and reusable, allowing them to be efficiently applied to new datasets or adapted to different research contexts. This can be achieved with provenance recording in DMPs.

Transparent and FAIR reporting of preprocessing pipelines facilitates the identification of potential flaws in the subsequent development of ML models. We illustrate this need with a recent example from the literature. Vandewiele *et al.* [246] reported a critical metodological flaw in electrohysterography recordings to estimate the risk on preterm birth: applying over-sampling before partitioning the data into mutually exclusive training and testing sets. This methodological flaw led to a large number of studies reporting near-perfect performance, resulting in biased results that, when corrected, were in many cases not significantly better than random guessing [246]. Such flaws could be easily detected by reviewers if authors follow FAIR preprocessing guidelines.

Recently, international renowned conferences, such as NeurIPS, are encouraging authors to generate a *Croissant* machine-readable metadata file to document their datasets [247]. This metadata format creates a shared representation across ML tools, frameworks and platforms. We highlight the potential of this tool and advocate for further efforts to extend its application to preprocessing and the field of clinical informatics.

5.5 Best practices for reporting FAIR preprocessing pipelines

Data preparation and preprocessing pipelines involve multiple components and stages (including the data itself, the associated labels, and the labeling process) that influence the input to ML models; all these parts of the pipeline should be reported following the FAIR principles. We propose the following best practices to ensure FAIR reporting of preprocessing steps in clinical informatics:

1. **Data splitting strategy.** Separating an independent subset of the data for testing the model at the beginning of the project [4]. This should be guided by domain knowledge, taking into account temporal dependencies when present (e.g., time-series data,

longitudinal records), specific experimental conditions under which the data were generated (e.g., laboratory settings), and, in some cases, data characteristics that require stratification (e.g., avoiding a test set with a disproportionate number of samples from one gender, or predominantly easier cases) to prevent data leakage and shortcut learning. Division of the data into train, test, and validation splits, should describe if any cross-validation is performed, and include the samples identifiers to allow full reproducibility.

2. **Label adaptation.** Curation of the original labels should also be explained if any transformation is performed, e.g., when there are multiple readers for the same medical image how is the ground truth derived to train the algorithm. Uncertainty and noise in ground truth labels, as well as labeling of edge cases should be considered and discussed, especially when the gold label is not available (e.g., radiological visual evaluation of malignancy in a lung nodule instead of pathological examination).

3. **Format conversion.** Original formats and subsequent derivations should be justified, i.e., conversion of DICOM to NRRD or PNG formats should report the HU window, as well as changes in data types (e.g., integer to float, log-transformed values), for example, in omics data.

4. **Annotation, feature selection.** Data annotation and feature selection process should be documented. For example, in medical images, steps such as skull-stripping or segmentation (either manual, semi-automatic or fully automatic) of the relevant parts should be described including the software and hyperparameters. A rationale should be given for the tools and libraries employed, e.g., if Monai [26] or other typical tools were used, an explanation of the choice should be included to facilitate understanding and promote reproducibility. Additional relevant information on the centers where the data was acquired, different machines or methodology should be documented if necessary.

5. **Feature normalization.** Manipulation of the data to train an ML model should always be clearly explained step by step, i.e., normalization/cropping/downsampling/ upsampling/padding medical images, handling missing data and outliers in omics data, converting raw sequencing reads to gene expression matrices or normalizing proteomics intensity values. This includes reasoning behind the choice of hyperparameters in normalization (type of normalization used, ranges chosen, etc.). Specific choices should be grounded on domain-specific knowledge.

6. **Data augmentation.** Data augmentation and resampling strategies should be documented and, if possible, including examples in the supplementary materials (e.g., for images).

7. **Metadata traceability.** Recording metadata changes and documenting data provenance after pre-processing to ensure traceability. These should also be documented in the DMP.

8. **Auditing.** In the clinical deployment of the algorithm, describing the data for auditing, quality control tests that will be performed, together with measures to identify data and concept drifts [248].

Authors are encouraged to briefly summarize all these steps in the main manuscript and to present full details in supplementary or external documentation. To promote full engagement, conferences and journals should consider enforcing these best practices through their official submission guidelines. Additionally, the afore-mentioned best practices can be used in combination with existing checklists for conducting and reporting ML-based science, such as REFORMS [249].

As illustrated in Figure 5.3, the proposed best practices are grounded on several principles from the FUTURE-AI [163] and FAIR frameworks [203].

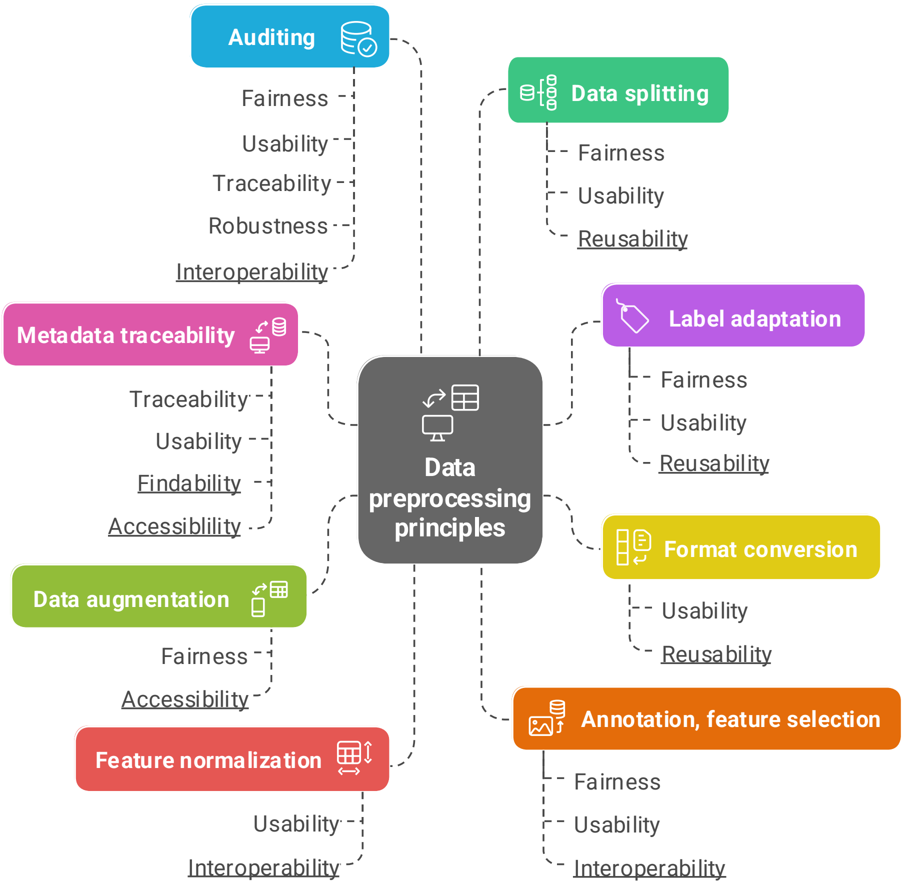


Figure 5.3: Best practices in data preprocessing associated with the corresponding key principles from FUTURE-AI and FAIR frameworks.

5.6 Conclusion

Data preparation and preprocessing in clinical informatics workflows are key steps before the development of ML algorithms. Transparent preprocessing pipelines foster the advancement of scientific research, promoting reproducibility, accessibility and validity of ML results. Current state of the art in clinical informatics tends to overlook the impact of data preprocessing and data quality on the robustness of ML algorithms, hindering clinical translation. In this chapter, examples are provided to emphasize the relevance of leveraging domain-knowledge in clinical informatics to perform informed data preprocessing as a crucial step to ensure consistent results across different research institutions.

The proposed best practices intend to support standardized, reusable, and reproducible application of data preparation steps. This is a fundamental requirement for developing trustworthy AI systems in healthcare, since systematic reporting following FAIR principles is essential to detect and prevent common reproducibility issues that threaten ML research, such as data leakage. Promoting and enforcing these guidelines in journals and conferences will require time, but we believe current efforts are already in this direction, and educating the ML and bioinformatics community is essential to drive further progresses.

Chapter 6

Multitask learning to enhance image-based intracranial hemorrhage prognosis

Main publication associated with this chapter: **Cobo, Miriam**, *et al.* "Multi-task Learning Approach for Intracranial Hemorrhage Prognosis." *International Workshop on Machine Learning in Medical Imaging.* Cham: Springer Nature Switzerland, 2024. p. 12-21. DOI: https://doi.org/10.1007/978-3-031-73290-4_2.

6.1 Introduction

Intracranial hemorrhage (ICH) is a leading cause of death and disability worldwide, characterized by high mortality rates and significant long-term neurological impairments [250]. This condition poses a substantial burden on healthcare systems, and presents complex challenges in terms of timely diagnosis, effective treatment, and rehabilitation strategies [251]. Computed tomography (CT) is the standard imaging modality for evaluating ICH, in which acute hematomas appear as high-density regions [252].

The incidence of ICH is projected to rise due to aging populations and increasing prevalence of risk factors such as hypertension, coagulopathy, and cerebral amyloid angiopathy [253]. Yet, prognostic predictors of ICH are significantly under-explored, hindering patient stratification by severity and the evaluation of the efficacy of emerging therapeutic interventions [254]. This prognostic uncertainty sometimes leads to a lack of consensus among clinicians on treatment [255]. Thus, there is an urgent need to improve the understanding of the relationship between clinical and demographic variables with ICH imaging features, in order to gain insights into the underlying factors in prognosis through imaging.

Deep convolutional neural networks (CNNs) are able to extract meaningful feature representations from medical images for classification tasks [256]. Recently, multimodal fusion models are gaining importance to exploit information across different modalities, and to build more precise and robust models [257]. Identifying medical knowledge relevant to image analysis tasks might be complex. While some of this knowledge can be directly learned from the training data, other aspects are not easily captured by the DL model, making it necessary to promote their learning [258]. Transforming medical knowledge into valuable representations to enhance the performance of DL image models also requires a careful understanding of the data [258]. In this regard, ordinal learning approaches are gaining more attention in the medical imaging field [259, 260], since they leverage the ordinal relationships among different stages of disease severity.

This chapter focuses on enhancing feature representation from CT scans by incorporating clinical and demographic variables strongly associated with ICH prognosis. These variables are introduced in the image model as learning biases (see Section 3.3.2) through soft penalty constraints added in the loss function in a binary or ordinal approach. In this chapter, learning biases will be based on clinical evidence, in contrast with PIML methods, which rely on physical constraints, discussed in Chapter 3. The joint learning of a shared embedding that incorporates prognosis and highly correlated relevant variables enhances the interpretability and robustness of the DL image model.

6.2 Statement of the problem

The potential of clinical context for ICH prognosis has already been studied in previous works [255, 261]. Perez *et al.* [255] proposed a hybrid model following a joint fusion approach to classify patients into good and poor prognosis, using both CT images and clinical variables. Shan *et al.* [261] simplified the task introducing a multimodal DL algorithm integrating both brain CT image data and Glasgow Coma Scale (GCS) score to improve prognosis. GCS is a clinical variable that describes the extent of impaired consciousness in all types of acute medical and trauma patients, ranging from 3 (worst) to 15 (highest) [262]. In a related study, Ma *et al.* [263] proposed a generative prognostic model for predicting ICH treatment outcomes utilizing imaging and tabular data. They used a variational autoencoder model to generate a low-dimensional prognostic score, and combined the multi-modality distributions into a joint distribution. All these methods require both tabular and imaging modalities as model inputs.

Current ICH fusion models fail to explore the entanglement between clinical information and medical images, overlooking the extent to which images alone can contribute to prognosis. In routine practice, neuroradiologists (NRs) integrate tabular clinical variables with imaging data to make prognostic assessments and stratify patients [251]. Hence, the exploration of how clinical information can be inferred directly from images to enhance prognostic accuracy represents a significant area of interest, particularly in scenarios where certain medical variables cannot be measured due to the patient's condition, such as in intubated ICH patients. Zhou *et al.* [264] showcased the value of incorporating domain knowledge in dermatology proposing a multi-task model to mimic dermatologists' diagnostic procedures, achieving state-of-the-art recognition performance.

In this chapter, we propose learning clinical information in the form of discrete binary and ordinal variables to improve feature representation of CT scans in an end-to-end multi-task ICH prognosis model. Multitask learning is a regularization approach that increases the robustness of DL algorithms by simultaneously learning the main task and auxiliary tasks that provide complementary information to the main task in a joint feature space. Our contributions can be summarized as (1) evaluating the clinical and demographic variables with the highest impact on ICH prognosis through machine learning (ML) tabular models, and their best encoding for the multi-task models; (2) introducing the two primary tabular variables driving the prognosis (GCS and age) in two multi-task prognostic image models (binary and ordinal); (3) performing ablations to show the predictive power of the proposed multi-task models; (4) assessing interpretability saliency maps [220] and their alignment with neuroradiologist's knowledge, ultimately comparing the prognostic capabilities of the models with four board-certified NRs.

6.3 Materials and methods

6.3.1 Data

We used a publicly available single center dataset [1] of 261 brain CT scans in NIfTI format, demographic and clinical variables (in the form of tabular data) from the patients' medical history for ICH prognosis, as described in Perez *et al.* [255]. These variables are collected in Table 6.1. The dataset was collected at Hospital Universitario Marqués de Valdecilla (HUMV), located in Santander, Spain. Ground truth classification labels are based on hospital survival: 99 patients with good prognosis (label 0), and 162 patients with poor prognosis (label 1).

Table 6.1: Categorical and numerical variables available in the dataset by Pérez *et al.* [1]. [a]Glasgow Glasgow Coma Scale is an ordinal categorical variable.

Categorical variables	**Numerical variables**
Sex	Age (years)
Smoker	Systolic arterial pressure
Alcohol	Diastolic arterial pressure
Head trauma	Oxygen saturation
Hypertension	Temperature
Diabetes Mellitus	Heart rate
Dyslipidemia	Respiratory frequency
Medical history of intracranial hemorrhage	Fibrinogen
Medical history of cardiovascular diseases	Glucose
Medical history of neurologic diseases	Creatinine
Medical history of dementia	Sodium
Medical history of cancer	Potassium
Medical history of hematologic diseases	White blood cells
Medical history of other major diseases	Hemoglobin
Anticoagulant drugs	Platelets
Antiaggregant drugs	Mean corpuscular volume
Antihypertensive drugs	Red blood cell distribution width
Calcium antagonist drugs	International normalized ratio
Alpha-blockers drugs	Mean platelet volume
Physical exploration with neurological signs and symptoms	Mean corpuscular hemoglobin concentration
Glasgow Coma Scale[a]	Urea

6.3.2 Method

Previous work conducted on this dataset [255] proposed a fusion model concatenating image features extracted by a custom 3D CNN with tabular data extracted by a Dense neural network (DNN). However, significant limitations were identified: training curves were close to random guess in the image model, and oversampling was performed in training, validation and test sets. These shortcomings undermined our confidence in the reported results. Thus, we repeated the baseline experiments in Perez *et al.* [255] performing 10 fold cross-validation (CV), and limited oversampling to the training set for class balance in prognosis. The 27 CT scans from one test fold were further labelled by four board-certified NRs with an average experience of 7 years (5, 5, 8 and 10 years) for benchmarking.

The proposed method aims to enhance the image model feature representation by learning a shared loss regularization across the main decision-driving variables in the ICH prognosis tabular models. To this end, the prognostic capability of the tabular variables available was first evaluated. Subsequently, a 3D DenseNet121 model [26] was used as feature extractor, and we designed two multi-task image models that aggregated the loss in the prognosis task with the loss of one clinical and one demographic variable, which was back-propagated through the image model. The method is presented in Figure 6.1, and explained below.

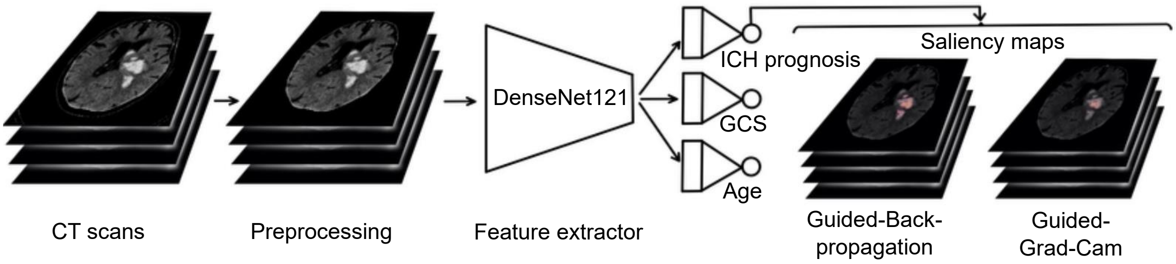


Figure 6.1: Proposed multi-task image model integrating Glasgow Coma Scale (GCS) and age as outputs to regularize the learning and enhance the prognosis task. In the saliency maps, brighter colors mean higher importance.

Preprocessing

Following Perez *et al.* [255], numerical variables were normalized using min-max normalization. In the CT scans, skull was stripped [265], and preprocessing [26] reproduced the steps in Perez *et al.* [255] for the best windowing selection. CT scans were downsampled to $301 \times 301 \times 40$ before feeding them to the models.

Tabular models

First, Perez *et al.* [255] DNN tabular model was reproduced using the variables in Table 6.1, then each variable's importance was evaluated using Shapley additive explanations (SHAP values) [266]. Subsequently, ML tabular models, including Random Forest (RF), XGBoost (XGB), Logistic Regression (LR), and Decision Tree Classifier (DTC), were trained on the most relevant variables to identify the prognostic decision boundaries, which were integrated into the multi-task image models. The process is described in Figure 6.2.

Image models

For evaluation purposes, the baseline image model was replicated as described by Perez *et al.* [255]. Yet, we also introduced a new baseline model based on a 3D DenseNet121 backbone feature extractor [26], which mitigates the vanishing-gradient problem and promotes feature

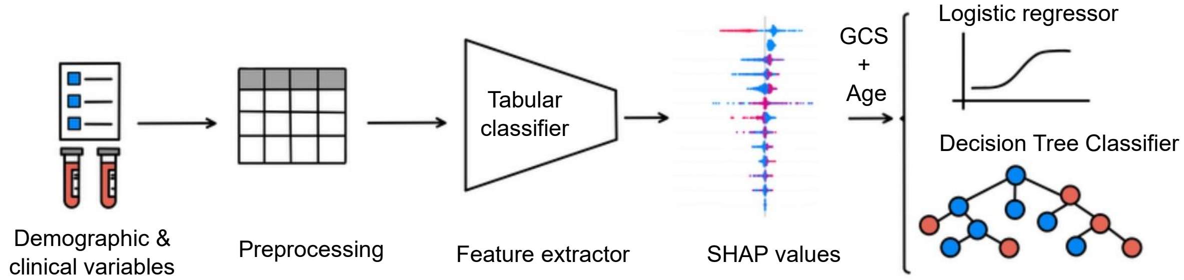


Figure 6.2: Tabular models to identify the most relevant variables driving ICH prognosis predictions, reproduced with logistic regressor and decision tree classifier models trained on the two main variables guiding the decisions: Glasgow Coma Scale (GCS) and age.

reuse [267]. The two key prognostic variables in the tabular models, GCS (clinical) and age (demographic), were incorporated as outputs into the baseline model, transforming it into a multi-task model (Figure 6.1). The aim was to regularize the model's learning process, exploring the shared knowledge between imaging-based prognosis, GCS, and age.

The learning of GCS and age was promoted by encoding them as discrete binary and ordinal variables since, as demonstrated by the DTC, these variables have clear decision boundaries that enable their discretization, which at the same time prevents unnecessary complexity in the multi-task model. For the binary classification scenario, we followed the DTC's decision boundaries, i.e., GCS from 3 to 8 was set to 1, while GCS from 9 to 15 was set to 0. For the three ordinal classes scenario, we followed Jain and Iverso [262] GCS division for a common classification of acute traumatic brain injury. Thus, severe GCS from 3 to 8 was set to 2; moderate GCS from 9 to 12 was set to 1; and mild GCS from 13 to 15 was set to 0. Then, we encoded GCS to preserve ordinality ($0 \rightarrow [0, 0]$, $1 \rightarrow [1, 0]$, and $2 \rightarrow [1, 1]$). Age was binarized according to the decision boundaries in the DTC: age below 80 was established as 0, otherwise it was set to 1.

The first multi-task model predicted prognosis, binary GCS and binary age, hereafter referred to as MT (bin GCS, bin age). The second multi-task model integrated prognosis, three class ordinal GCS and binary age, hereafter referred to as MT (ord GCS, bin age). Both models used a DenseNet121 backbone for feature extraction, and the loss was combined following Equation 6.1 to enhance the feature representation for each task:

$$\mathcal{L} = \lambda_{prog} \cdot \mathcal{L}_{prog} + \lambda_{GCS} \cdot \mathcal{L}_{GCS} + \lambda_{age} \cdot \mathcal{L}_{age} \tag{6.1}$$

where $\mathcal{L}_{prog}$ is the loss in prognosis, $\mathcal{L}_{GCS}$ is the loss in GCS, and $\mathcal{L}_{age}$ is the loss in age. They are all binary cross-entropy losses, since we previously encoded three class ordinal GCS. The hyperparameters in Equation 6.1 were empirically optimized through several experiments: $\lambda_{prog} = 0.4$, $\lambda_{GCS} = 0.3$, and $\lambda_{age} = 0.3$.

All image models used early stopping with a patience of 20 epochs, evaluated on Balanced Accuracy in validation, and a dropout of 0.2 in the DenseNet121 feature extractor to prevent overfitting. Small data augmentations applied to the training set included rotations (up to 5°), zoom (up to 10%), and Gaussian noise (mean: 0.0, standard deviation: 0.01), with a probability of 0.5, using MONAI [26]. Models were coded in Pytorch (version 1.13.1) [268] and trained on NVIDIA Tesla T4$^{\text{TM}}$ 16GB GPU, utilizing a batch size of 8, three steps of gradient accumulation, and AdamW optimizer [269] (learning rate of 0.001, weight decay of 0.0001).

6.3.3 Evaluation

The performance of the models was assessed on the CV test sets, for a conservative threshold of 0.5, utilizing the following classification metrics: Area Under the Curve (AUC), Accuracy (Acc.), Balanced Accuracy (B. Acc.), Specificity (Spec.), Negative Predictive Value (NPV), Precision (Prec.), Recall, F1-score (F1-sc.). In the case of the three class ordinal GCS, we used: Acc., B. Acc., Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Uniform Ordinal Classification index (A_{UOC}) [270], and quadratic weighted Cohen's Kappa score [271, 272]. An ablation analysis was performed to evaluate the contribution of the loss terms in Equation 6.1.

The impact of the multi-task regularization was further assessed computing interpretability saliency maps, specifically Guided-Backpropagation and Guided-Grad-Cam [273]. To retrieve the most relevant content, saliency maps were normalized and thresholded. One of the NRs examined the saliency maps for the baseline and multi-task image models in the same test fold labelled by four NRs.

Neuroradiologist criteria to evaluate the saliency maps

In relation to the prognosis of intracranial hemorrhage (ICH), Glasgow Coma Scale (GCS), age, hematoma volume, and the presence of intraventricular and/or infratentorial hemorrhage are factors that have been associated with poor prognosis in various studies [274].

Moreover, in recent years, several potential imaging-specific prognostic factors have been identified. These include the irregularity of hematoma margins and internal hypodensities [275], the spot sign (a hyperdense focus within the hematoma on contrast-enhanced scans that correlates with an increased risk of hematoma expansion) [276], and heterogeneous density in subdural hematomas [277]. These factors increasingly underscore the prognostic value of imaging.

6.4 Results

6.4.1 Tabular models

Table 6.2 presents the performance of tabular models with mean and standard deviation (SD). XGB was omitted since performance was slightly inferior to RF. SHAP values indicated that GCS and age were the main predictors. LR trained only on GCS and age features reproduced the DNN model by Perez *et al.* [255]. Hence, GCS and age were integrated in the proposed multi-task image models.

6.4.2 Image models

A comparison of the image models' performance with the four NRs in the specific test fold they evaluated is shown in Table 6.3. This test fold was randomly selected. Ablation analysis on the 10-fold CV test sets is shown in Figure 6.3. Separate baseline image models were trained on each outcome variable: prognosis, GCS (binary and ordinal), and age. Results are detailed in Tables 6.4 and 6.5.

We evaluated the impact of GCS and age on the regularization and explainability of the MT image models computing Guided-Backpropagation and Guided-Grad-Cam saliency maps. Four examples of clinically relevant saliency maps selected by the neuroradiologist (who reviewed all

Table 6.2: Tabular models 10-fold cross-validation performance for intracranial hemorrhage prognosis (mean and standard deviation). Best results are highlighted in **bold**. Abbreviations: DNN: Dense Neural Network, LR: Logistic Regressor, DTC: Decision Tree Classifier, RF: Random Forest.

Model	**DNN**	**RF**	**LR**	**LR**	**DTC**	**RF**
Features	All	All	GCS	GCS & age	GCS & age	GCS & age
AUC	0.79 ± 0.06	0.78 ± 0.07	0.70 ± 0.09	**0.80 ± 0.07**	0.74 ± 0.06	0.79 ± 0.05
Acc.	0.72 ± 0.05	0.72 ± 0.09	0.62 ± 0.13	**0.73 ± 0.05**	0.71 ± 0.06	**0.73 ± 0.06**
B. Acc.	0.70 ± 0.04	0.71 ± 0.09	0.65 ± 0.10	**0.74 ± 0.04**	0.69 ± 0.06	0.72 ± 0.07
Spec.	0.66 ± 0.12	0.64 ± 0.12	0.77 ± 0.15	**0.79 ± 0.10**	0.60 ± 0.11	0.71 ± 0.12
NPV	0.65 ± 0.12	**0.66 ± 0.13**	0.53 ± 0.13	0.63 ± 0.08	0.63 ± 0.12	0.63 ± 0.10
Prec.	0.79 ± 0.05	0.78 ± 0.07	0.79 ± 0.12	**0.85 ± 0.06**	0.76 ± 0.04	0.81 ± 0.06
Recall	0.75 ± 0.14	**0.78 ± 0.13**	0.5 ± 0.2	0.70 ± 0.11	0.77 ± 0.11	0.74 ± 0.09
F1-sc.	0.76 ± 0.06	**0.77 ± 0.09**	0.6 ± 0.2	0.76 ± 0.06	0.76 ± 0.06	**0.77 ± 0.06**

Table 6.3: Comparison between models' performance and neuroradiologists (NRs). For the models, 95% confidence intervals estimated by boostrapping are shown in brackets, calculated generating 1000 bootstrap samples from the original test set. For the NRs, mean and standard deviation are given. Best results are highlighted in **bold**. Abbreviations: MT: multitask, bin: binary, ord: ordinal, GCS: Glasgow Coma Scale.

Fold 1	**Baseline [255]**	**Baseline DenseNet121**	**MT (bin GCS, bin age)**	**MT (ord GCS, bin age)**	**NRs**
Acc.	0.48(0.38-0.57)	0.63(0.53-0.73)	**0.70**(0.61-0.79)	0.66(0.57-0.75)	0.60 ± 0.06
B. Acc.	0.57(0.50-0.63)	0.52(0.47-0.58)	0.60(0.54-0.67)	**0.67**(0.58-0.76)	0.64 ± 0.05
Spec.	**0.90**(0.80-0.98)	0.10(0.02-0.21)	0.20(0.08-0.33)	0.70(0.53-0.84)	0.80 ± 0.08
NPV	0.41(0.30-0.52)	0.50(0.11-0.88)	**1.00**(1.00-1.00)	0.54(0.40-0.67)	0.48 ± 0.05
Prec.	0.79(0.60-0.95)	0.64(0.54-0.74)	0.68(0.58-0.77)	0.79(0.67-0.89)	**0.81** ± 0.06
Recall	0.23(0.13-0.34)	0.94(0.88-0.99)	**1.00**(1.00-1.00)	0.64(0.53-0.75)	0.49 ± 0.12
F1-sc.	0.36(0.22-0.49)	0.76(0.68-0.83)	**0.81**(0.73-0.87)	0.71(0.62-0.79)	0.60 ± 0.10

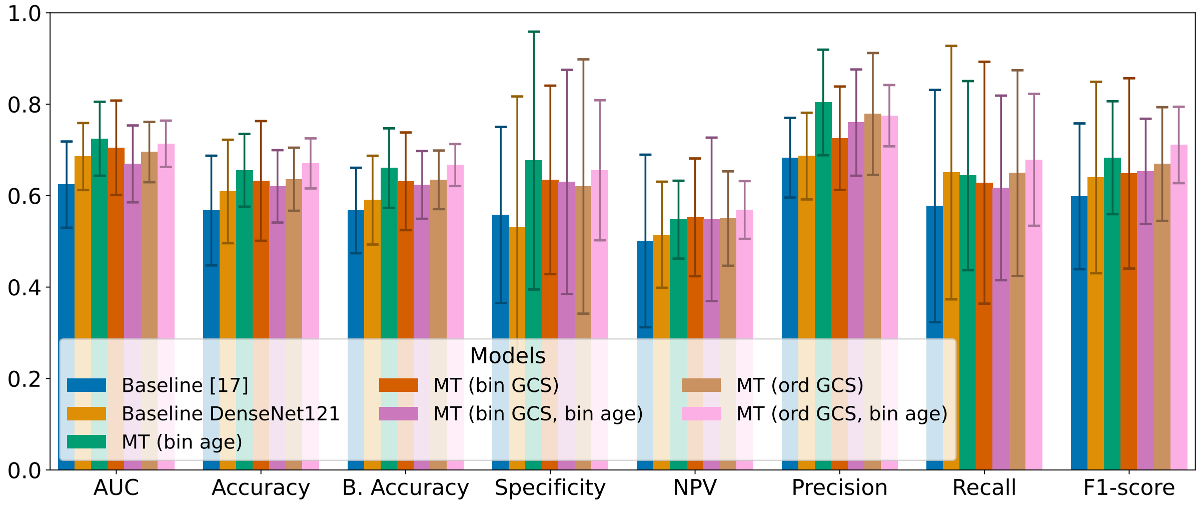


Figure 6.3: Ablation analysis of each component of the loss in our method.

Table 6.4: Performance on test for the baseline image models trained on each of the outcome variables separately for 10 fold cross-validation (mean and standard deviation).

Baseline Task	**Prognosis [255]**	**Prognosis DenseNet121**	**Binary GCS**	**Binary age**
AUC	0.62 ± 0.09	0.69 ± 0.07	0.71 ± 0.15	0.70 ± 0.13
Acc.	0.57 ± 0.12	0.61 ± 0.11	0.68 ± 0.08	0.61 ± 0.15
B. Acc.	0.57 ± 0.09	0.59 ± 0.10	0.66 ± 0.13	0.64 ± 0.11
Spec.	0.56 ± 0.19	0.5 ± 0.3	0.71 ± 0.10	0.6 ± 0.2
NPV	0.50 ± 0.19	0.51 ± 0.12	0.86 ± 0.09	0.86 ± 0.09
Prec.	0.68 ± 0.09	0.69 ± 0.09	0.38 ± 0.15	0.42 ± 0.14
Recall	0.6 ± 0.3	0.7 ± 0.3	0.6 ± 0.2	0.7 ± 0.3
F1-sc.	0.60 ± 0.16	0.6 ± 0.2	0.45 ± 0.16	0.50 ± 0.17

Table 6.5: Performance on test for the baseline image model trained on three class ordinal GCS model for 10 fold cross-validation (mean and standard deviation).

Baseline Task	**Ordinal GCS**
Accuracy	0.49 ± 0.11
Balanced accuracy	0.44 ± 0.10
MAE	0.65 ± 0.13
RMSE	0.98 ± 0.06
A_{UOC} **index**	0.69 ± 0.08
Cohen Kappa score (quadratic weighted)	0.19 ± 0.18

the slices per patient) are depicted in Figure 6.4. These saliency maps were generated using Guided Backpropagation, which provided the most informative activations. Although Figure 6.4 A is a challenging case to label focusing only on ICH imaging features (some expansivity is shown), the neuroradiologist believes that the MT models' ability to incorporate GCS (15) and age (78 years) was key to recognizing it as a good prognosis. Interestingly, in Figure 6.4 B, the MT models highlight the presence of posterior fossa and intraventricular hemorrhage component compared to the baseline model, that does not detect it. Related to Figure 6.4 C and D, MT (ord GCS, bin age) model additionally detects the lateral component of the subdural hematoma, and shows activations in the adjacent grooves to the subdural hematoma, that is, in the expansivity component, while MT (bin GCS, bin age) and baseline models only detect the medial component of the subdural hematoma. Overall, the neuroradiologist concluded that the MT (ord GCS, bin age) model exhibited fewer arbitrary activations, and probably detected better intraventricular hemorrhage and expansivity signs. The corresponding Guided-Grad-Cam saliency maps are additionally presented in Figure 6.5.

Neuroradiologist assessment of relevant saliency maps

The quantity of clinically relevant Guided-Backpropagation saliency maps per patient identified by the neuroradiologist for each model is shown in Table 6.6. Occasionally, no saliency maps were clinically relevant, because there were also activations in the background.

Comments from the neuroradiologist on the saliency maps. The neuroradiologist reviewed all the Guided-Backpropagation saliency maps of the three models: baseline

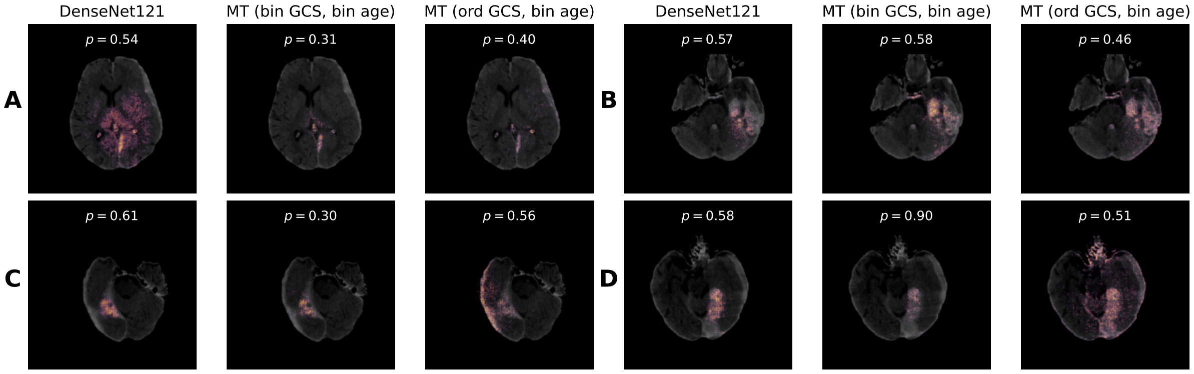


Figure 6.4: Guided-Backpropagation saliency maps (p is the output probability, $p < 0.5$ corresponds to good prognosis). **A: Good prognosis.** Correctly labelled by 4/4 neuroradiologists. **B: Good prognosis.** Incorrectly labelled by 4/4 neuroradiologists. **C: Poor prognosis.** Incorrectly labelled by 4/4 neuroradiologists. **D: Poor prognosis.** Correctly labelled by 3/4 neuroradiologists.

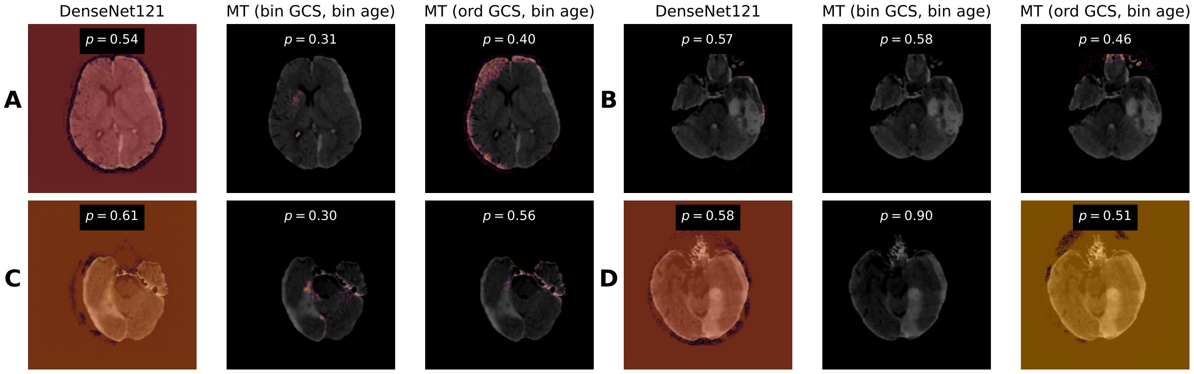


Figure 6.5: Guided-Grad-Cam saliency maps (p indicates the output probability of each model). Patient ID is indicated in brackets. **A: Good prognosis (34).** Correctly labelled by 4/4 neuroradiologists. **B: Good prognosis (93).** Incorrectly labelled by 4/4 neuroradiologists. **C: Poor prognosis (59).** Incorrectly labelled by 4/4 neuroradiologists. **D: Poor prognosis (140).** Correctly labelled by 3/4 neuroradiologists.

DenseNet121, MT (bin GCS, bin age), MT (ord GCS, bin age), for the 27 patients in the test fold that was further labelled by the four neuroradiologists. Moreover, the neuroradiologist selected an extra set of relevant saliency maps and commented on them. The most representative slice of these saliency maps is depicted in Figure 6.6. For completeness, we include the neuroradiologist's comments on Figure 6.6:

- A. MT (ord GCS, bin age) does not show as many activations as baseline model, which is compatible with a good prognosis.
- B. All models detect the intraventricular component.
- C. Baseline shows more activations, possibly arbitrary, compared to the MT models.
- D. Similar activations in all the models.

Table 6.6: Number of clinically relevant Guided-Backpropagation saliency maps slices per patient and model according to the evaluation criteria from the neuroradiologist. Patient ID corresponds to the same ID as in the public dataset [1].

Patient ID	**Baseline**	**MT (bin GCS, bin age)**	**MT (ord GCS, bin age)**
12	9	8	6
20	5	8	12
34	12	14	17
51	13	15	16
56	0	6	6
59	14	7	17
71	16	14	7
90	0	16	9
92	0	11	12
93	14	17	14
106	17	18	17
116	19	19	19
128	18	17	19
137	15	19	12
140	18	16	16
152	7	6	5
170	21	15	21
187	13	14	14
197	0	10	8
203	6	8	6
214	0	6	19
245	15	15	11
251	14	12	12
258	12	0	8
265	12	13	12
268	8	0	6
292	11	13	11
Total	289	317	332

- E. MT (ord GCS, bin age) detects more the midline intraventricular component compared to the rest.
- F. MT (ord GCS, bin age) highlights the highest density component of the bihemispheric subdural hematomas present in the patient. Baseline and MT (bin GCS, bin age) show less useful saliency maps.
- G. MT (ord GCS, bin age) highlights hematoma with internal hypodensities. Baseline shows too many activations.
- H. MT (ord GCS, bin age) detects all the intraventricular hemorrhage component. Baseline and MT (bin GCS, bin age) do not clearly detect all the intraventricular hemorrhage component or the intraparenchymal hematoma.

- I. In addition to the hemorrhage, MT (ord GCS, bin age) also highlighted the expansiveness over the sulci. Baseline did not detect hematoma correctly.
- J. In this case the activations in baseline are more meaningful than in the multi-task models.
- K. MT (ord GCS, bin age) highlights the expansive effect of a second hematoma located in the posterior fossa (not shown in this slice, but the neuroradiologist spotted it during the revision of all the slices). Too many activations in baseline.
- L. MT models highlight more than the baseline model the intraventricular hemorrhage on occipital right horn of the lateral ventricle, but MT (bin GCS, bin age) shows too many activations.

6.5 Discussion and conclusion

To the best of our knowledge, this is the first time that the most relevant variables to ICH prognosis (GCS and age) are integrated into an end-to-end multi-task prognostic model to regularize the shared feature representation from CT scans, outperforming baseline models [255], and four board-certified NRs. The decision boundaries of GCS and age, established independently by a DTC and supported by the literature [251], focus the learning of the clinical context by allowing the discretization of these variables. The ablation analysis in Figure 6.3 highlights that the MT (ord GCS, bin age) model exhibits smaller variances compared to all other models, particularly in specificity, recall and F1-score, showing that our approach increased the robustness of the feature representation. The metrics were set for a conservative threshold of 0.5, but could be optimized to maximize recall at the expense of precision, as typically done in clinical application settings [278].

The proposed multi-task models demonstrate comparable performance to tabular models, which exhibit strong prognostic capabilities in ICH [279]. However, for reliable prognosis, imaging is essential to evaluate parameters such as the hematoma's volume, expansivity, and possible presence of tumors or other pathologies, all of which contribute more significantly to prognosis and treatment decisions than some other clinical data. Quantifying the information displayed in the image would be more time-consuming (e.g., requiring segmentation of the hematoma for volume quantification), subjective due to the difficulty in precisely characterizing these parameters, and further constrained by the lack of standardization in ICH imaging. Direct inclusion of the data in the image itself offers more exhaustive and informative insights. Thus, our approach mimics clinical decision-making requiring only CT scans as input, and enhances the interpretability of the image model incorporating GCS and age predictions.

The comprehensive evaluation in one of the test folds by four NRs provides additional validation for the multi-task image models, alleviating concerns about the study's single-center data and limited patient sample [280]. The neuroradiologist concluded that the saliency maps across all models are generally similar. Yet, the MT (ord GCS, bin age) model was more selective or specific, and presented fewer random activations than the MT (bin GCS, bin age) model, and even more so compared to the baseline model. In summary, the baseline model showed more activations, whereas the proposed multi-task models focused more exclusively on the hematoma (or hematomas) and its expansive effect. These findings remain qualitative. Thus, future work could focus on quantifying the extent to which the proposed GCS and age variables contribute to

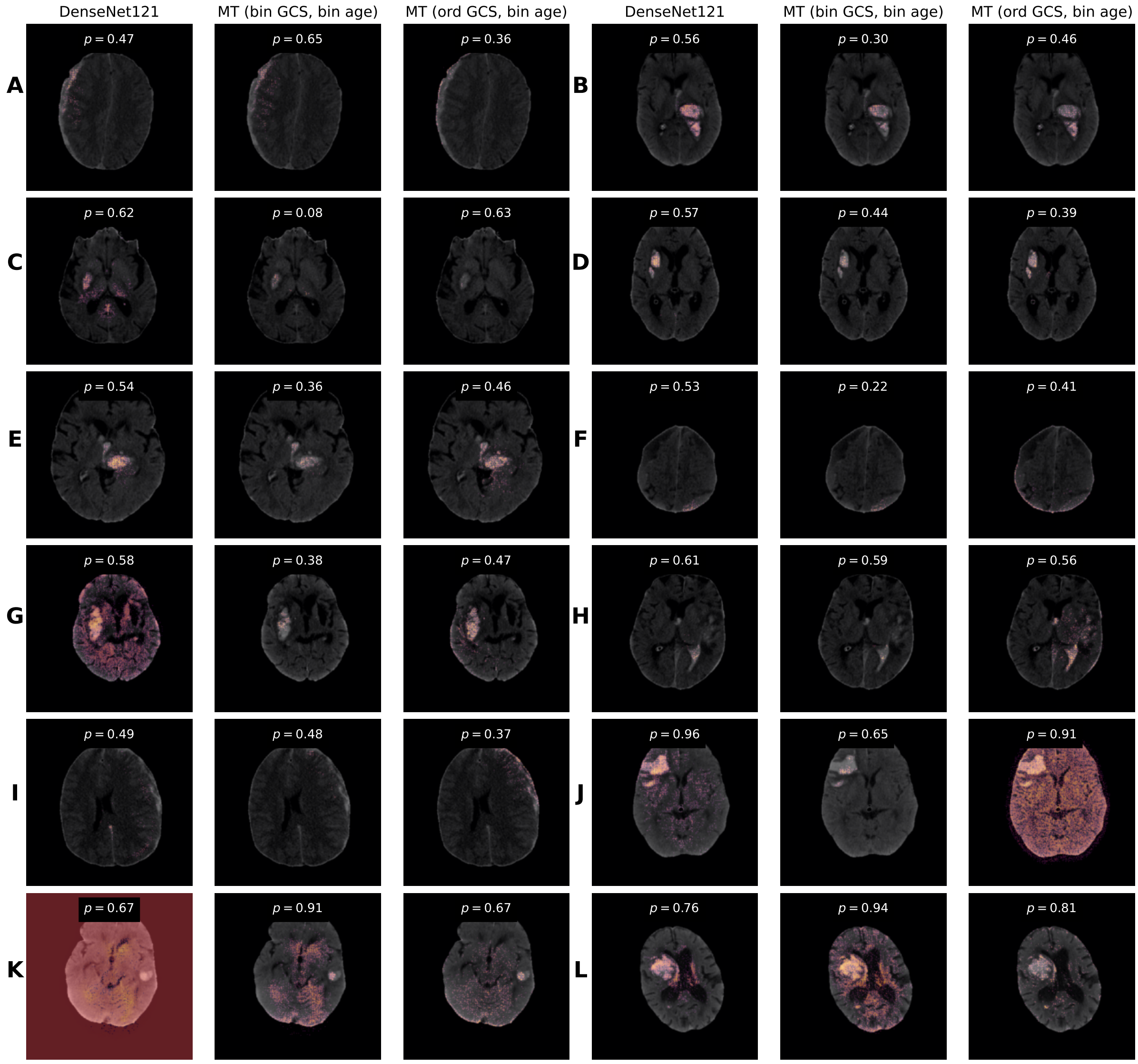


Figure 6.6: Guided-Backpropagation saliency maps (p is the output probability, $p < 0.5$ corresponds to good prognosis). Patient ID is indicated in brackets. **A: Good prognosis (51).** Correctly labelled by 4/4 neuroradiologists. **B: Good prognosis (71).** Correctly labelled by 3/4 neuroradiologists. **C: Good prognosis (152).** Correctly labelled by 4/4 neuroradiologists. **D: Good prognosis (203).** Correctly labelled by 4/4 neuroradiologists. **E: Poor prognosis (12).** Correctly labelled by 2/4 neuroradiologists. **F: Poor prognosis (20).** Correctly labelled by 1/4 neuroradiologists. **G: Poor prognosis (56).** Incorrectly labelled by 4/4 neuroradiologists. **H: Poor prognosis (106).** Correctly labelled by 4/4 neuroradiologists. **I: Poor prognosis (128).** Correctly labelled by 4/4 neuroradiologists. **J: Poor prognosis (170).** Incorrectly labelled by 4/4 neuroradiologists. **K: Poor prognosis (197).** Correctly labelled by 2/4 neuroradiologists. **L: Poor prognosis (258).** Correctly labelled by 3/4 neuroradiologists.

image-based prognosis using ground truth segmentations, and assessing model generalizability to new datasets.

In conclusion, this chapter introduces a novel multi-task method for ICH prognosis leveraging GCS and age, the two main variables driving the decisions in the tabular ICH prognosis models. The proposed multi-task image models regularize the loss using the clinical information embedded in GCS and age outputs, and learn more robust feature representations than state-of-the-art approaches.

6.6 Future work

External validation of the MT (ordinal GCS, binary age) model for ICH prognosis is currently underway using data from two additional hospitals: Complejo Asistencial Universitario de León (LEON), located in León, Spain, and Hospital Universitario Central de Asturias (HUCA), located in Oviedo, Spain. This will enable a more comprehensive evaluation of the model's generalizability across heterogeneous clinical settings, including variations in imaging protocols and patient populations.

The data available from the new hospitals is shown in Table 6.7, including the number of patients with good and poor prognosis, following the same criteria as in the HUMV training dataset.

Table 6.7: Data description from the new hospitals.

Hospital	**# Good prognosis**	**# Poor prognosis**	**Total patients**
HUCA	59	55	114
LEON	211	140	351

The preliminary results on evaluating the generalizability of the MT (ord GCS, bin age) model to the external hospitals (HUCA and LEON), in comparison with the performance on the internal test subset (HUMV) are shown in Table 6.8.

Table 6.8: Comparison between MT (ord GCS, bin age) model's performance on in-distribution (HUMV) and out-distribution (HUCA and LEON) datasets. For the models, 95% confidence intervals estimated by boostrapping are shown in brackets, calculated generating 1000 bootstrap samples from the original test set.

Fold 1	**HUMV**	**HUCA**	**LEON**
Validation set	Internal	External	External
Acc.	0.66(0.57-0.75)	0.71(0.63-0.80)	0.67(0.58-0.76)
B. Acc.	0.67(0.58-0.76)	0.71(0.63-0.80)	0.68(0.58-0.77)
Spec.	0.70(0.53-0.84)	0.75(0.61-0.86)	0.64(0.51-0.76)
NPV	0.54(0.40-0.67)	0.69(0.56-0.80)	0.78(0.66-0.89)
Prec.	0.79(0.67-0.89)	0.74(0.61-0.86)	0.57(0.44-0.69)
Recall	0.64(0.53-0.75)	0.68(0.56-0.80)	0.72(0.58-0.86)
F1-sc.	0.71(0.62-0.79)	0.71(0.60-0.80)	0.63(0.51-0.74)

Results on the external datasets are consistent with the metrics obtained on the internal dataset, indicating the MT (ord GCS, bin age) model is robust across variations in patient populations. Future work will involve additional experiments to further quantify the model's generalizability and to explore the similarities and differences among the three populations. As part of this extension, uncertainty quantification techniques will be incorporated to assess the reliability and robustness of the model's predictions in external data.

Chapter 7

Design and curation of a database from a screening cohort for personalized lung cancer diagnosis from a multimodal perspective

Main publication associated with this chapter: **Cobo, Miriam**, *et al.* “A novel open access multimodal dataset of nodule imaging and circulating proteome from a lung cancer screening cohort”. *To be submitted.*

7.1 Introduction

Translational research in medicine leverages insights from different fields, such as molecular biology, genetics, and medical images, to advance disease diagnosis, prognosis, risk assessment, prevention, and treatment in highly interdisciplinary environments. These collaborative approaches have led to personalized medicine (PM), also referred to as precision medicine [281], or P5 medicine (predictive, preventive, personalized, participatory, psycho-cognitive) [282]. PM seeks to tailor medical treatments to the individual characteristics of each patient, by utilizing biological information and biomarkers, including molecular pathways, genetic variants, protein expression, and metabolic profiles [283, 281]. The growing complexity of patient data, however, requires the use of advanced analytical tools. Research in both traditional and modern AI and ML methods is increasingly contributing to the success of translational medical research in drug discovery, imaging, and genomics [158].

This chapter focuses on PM to improve early lung cancer diagnosis, through the design, collection, curation and future release of a new dataset from a clinical screening cohort. This well curated dataset is a valuable research resource that will be made publicly available to advance scientific discovery and support the development of personalized diagnostic tools. The dataset, curated by the author of this thesis, aims to contribute specifically to the development of personalized multimodal AI models to enhance early detection and intervention strategies in lung cancer screening. The curation process included defining and refining inclusion and exclusion criteria, annotating imaging data, selecting relevant clinical variables, as well as identifying and understanding the clinical questions and needs.

Low-dose computed tomography (LDCT) lung cancer screening has significantly enhanced early detection and patient survival rates in the population at risk. Current screening methods, that primarily rely on LDCT imaging, will very likely benefit from molecular biomarkers to achieve a more comprehensive, accurate, personalized and non-invasive risk assessment leveraging multimodal tools. In this chapter, we present a novel multimodal (imaging, proteomics, clinical and demographic) dataset designed to provide an available research resource on LDCT-based early lung cancer detection. The dataset consists of annotated screening LDCT scans and plasma proteomics data generated by proximity extension assay (PEA) with most of the Olink Target 96 platforms (1078 individual proteins across 12 panels focused on a specific area of disease or biology) for a total of 211 screening participants. There are 67 lung cancer patients, 68 controls with benign pulmonary nodules, 71 controls without nodules, matched by sex, age and time between image and sample collection, and 5 surgically excised false positive lesions. Both radiological and molecular signatures were collected within a six month window, providing detailed insights into disease progression. Nodules were considered as lung cancer cases if biopsy-confirmed lung cancer was diagnosed within 5 years after imaging, enabling the study of longitudinal biomarker evolution and its correlation with imaging findings. To complement the dataset, clinical and demographic data will also be available in open access, providing a detailed overview of patient characteristics. The informed consent signed by the participants allows for unrestricted open access for requests directly or indirectly related to lung cancer research. Experiments were conducted to assess the technical quality and demonstrate the dataset's usability as a proof of concept, showing alignment with findings from previously published studies. This comprehensive dataset aims to facilitate research towards the development of personalized multimodal AI models, and support the investigation of the relationship between imaging and molecular data, paving the way for more accurate understanding of early lung cancer biology. Finally, the dataset presented in this chapter may help to develop or validate individualized risk prediction models that could significantly advance early lung cancer detection and intervention strategies.

7.2 Statement of the problem

Lung cancer is a leading cause of cancer death worldwide [284, 285], and the second most commonly diagnosed cancer in both men and women in the United States in 2023 [286]. LDCT screening has improved early diagnosis of lung cancer, decreasing lung cancer mortality among population at high risk [79, 287, 288]. The population at risk is defined by different risk models, all of which include age, smoking exposure and other clinical or demographic traits. The United States Centers for Medicare & Medicaid Services have defined their eligibility criteria for LDCT-based lung cancer screening for beneficiaries who are 50 to 77 years old, asymptomatic, current smokers or those who have quit within the past 15 years, and have a smoking history of at least 20 pack-years [289]. Other models include additional risk components, or different ranges of age and smoking to further refine inclusion criteria [290]. One of the unmet clinical needs in lung cancer screening is assessing and managing indeterminate risk pulmonary abnormalities (nodules) found in the CT images. These indeterminate risk (neither clearly benign, nor highly suspicious of malignancy) pulmonary nodules (IPNs) pose a challenge, since they are found in approximately 20%−40% of screening participants [291]. A similar clinical situation may happen outside screening, when incidental lung nodules are found as a result of CT-based diagnostic studies performed for other conditions [292].

The Lung CT Screening Reporting and Data System (Lung-RADS) was developed to standardize the reporting and management of screen-detected pulmonary nodules, and improve risk assessment [293]. LungRADS classifies lung nodules on a scale from 1 to 4X, where 1-2 indicate benign nodules with minimal risk, 3 represents indeterminate nodules requiring follow-up, and 4A-4X suggest a high suspicion of malignancy, warranting further diagnostic evaluation. Highly suspicious lung nodules typically require a biopsy for definitive diagnosis, usually through an invasive procedure. Thus, precise evaluation of risk in pulmonary nodules found in the screening process (and also in the routine clinical detection of incidental nodules) is crucial to improve patient survival and avoid unnecessary biopsy procedures. Comprehensive evaluation after an IPN identification requires a close follow-up at different time points, in order to identify differential characteristics and progression patterns that enable monitoring risk, and eventually early detection of a malignant lesion.

Radiologist visual assessment of LDCT screening scans remains the gold standard approach for evaluating lung nodules in the clinics. This evaluation is labor-intensive, time-consuming, and subject to certain degree of subjectivity; variability in expertise and subtle pixel-level differences in malignant nodules make accurate assessment a challenging task, ultimately compromising the reliability of this approach. Several initiatives have aimed to standardize the quantitative CT measurements of lung nodules [294, 295]. Computer-aided diagnosis (CAD) systems, particularly those leveraging deep learning (DL) techniques, have demonstrated strong potential for evaluating pulmonary nodule malignancy, providing valuable support to radiologists and improving the management of IPNs [296, 297, 298].

The majority of DL-based computer aided diagnosis (CAD) models for lung nodule risk assessment in the literature rely on the use of the publicly available LIDC-IDRI dataset [201], which was collected by the US National Cancer Institute in 2011, as training tool. This database includes nodule malignancy annotations (ranging from label 1, indicating low risk, to label 5, corresponding to high risk) provided by four independent radiologists. However, this dataset lacks gold standard labels from pathological examination for the majority of cases, particularly for nodules rated as highly suspicious of malignancy (labels 4 and 5). There are few published publicly available datasets that contain gold standard labels for lung cancer nodules. The most comprehensive in terms of data volume is the National Lung Screening Trial (NLST) [299], a randomized controlled clinical trial of screening tests for lung cancer, where a subset of 28 000 LDCT images can be granted access through the Cancer Data Access System, including 623 participants with screen-detected cancer. However, nodule annotations are not available, although some studies have obtained them for subsets of the dataset [300, 301]. LUNGx challenge dataset [302] released a total of 83 scans in 2014, but all except 13 were contrast enhanced, which is not compatible with the usual lung cancer screening protocol. LIDP dataset [303] was claimed to be released after an embargo period. Yet, to the best of our knowledge, it is still not publicly available. Recently, two other CT datasets were published: one specifically annotated with histopathology-based information [304], and the other one was a cross spatio-temporal lung nodule dataset with pathological information for nodule identification [305]. These two latter datasets incorporating pathological gold standard labels constitute a step forward to improving CAD systems for lung cancer detection, but they were not obtained in the context of LDCT lung cancer screening. Both datasets were collected in Chinese cohorts, where existing guidelines for nodule management developed on US [79] and European [287] cohorts have shown suboptimal performance due to variations in the incidence of lung cancer, specifically in what relates to non- or low dose-smoking sensitivity, among Asian populations, particularly within the Chinese population [296, 306]. Hence, to ensure broad generalizability to other populations [307], further evaluation across diverse cohorts is essential, highlighting the need to expand publicly available datasets to support the scientific community in this field. This need is especially urgent in the context of LDCT-based lung cancer screening, where the development of AI-driven image analysis algorithms is accelerating, and relies on broad access to robust training datasets.

Pathological examination is the gold standard to confirm the presence of tumor in highly suspicious nodules. In contrast, benign nodules are not routinely biopsied, as the procedure is invasive, carries a risk of complications, and is generally reserved for cases where malignancy is suspected. As a result, there is no pathological confirmation ensuring the absence of malignancy in nodules classified as benign. Instead, their classification is based on long-term (at least two years) follow-up and nodule image stability over time. This follow-up process provides a temporary evaluation of nodules, allowing for the identification of those that remain stable and can be classified as benign, while highly suspicious cases are ultimately referred for biopsy. Thus, in radiological clinical practice, nodules can be categorized as highly suspicious, indeterminate (some of which may be considered suspicious in subsequent follow-ups), and low-risk (those that remain stable over time). In a few cases, biopsy excludes lung cancer in nodules initially deemed suspicious, revealing them as imaging false positives. Limiting dataset annotations to only those confirmed by histopathology, as in the dataset of Jian *et al.* [304, 305], reduces the effectiveness of CAD systems, since they are trained only on nodules that were already flagged as highly suspicious and referred to biopsy by the radiologist. This approach hinders the clinical usefulness, since in real-world scenarios there is a high proportion of IPNs, and they should be also used to train CAD systems. The absence of gold standard for highly suspicious nodules, as seen in LIDC-IDRI [201], means there is no definitive pathological confirmation, therefore, there exists the possibility of misdiagnosis [303]. This lack of gold standard introduces noise and uncertainty into CAD systems, leading to potential increases in false positives and false negatives.

Biomarkers have been proposed as potential adjuncts for risk characterization, contributing additional information to existing risk models to refine inclusion criteria, and as clinical tools to support the assessment of the potential malignancy of IPNs [308, 309]. Gene expression–based strategies have revolutionized breast cancer care by guiding treatment decisions with greater precision [310, 311]. Similarly, current lung cancer screening research is exploring circulating biomarkers in blood as non-invasive tools for diagnosis and risk estimation. These biomarker approaches, and specifically the analysis of circulating proteomics, aim to bring a higher degree of personalized care to lung cancer in line with genomics strategies in breast cancer. The Integrative Analysis of Cancer Risk and Etiology (INTEGRAL) research consortium program [312] comprises three projects, among them the Risk Biomarker project aims to investigate novel prediagnostic circulating protein biomarkers to enhance lung cancer risk estimation [313]. The objectives of the second INTEGRAL initiative, known as the Nodule Malignancy Project, are, among others, to identify circulating proteins that can differentiate benign versus malignant pulmonary nodules following an initial LDCT scan. The identification and validation of protein biomarkers provides additional information on lung cancer risk, and is a promising direction towards improving current image-based risk prediction models [312]. The present study describes the data included in the LDCT Pamplona-ELCAP screening cohort (P-ELCAP), collected at Clínica Universitaria de

Navarra (CUN), as one of the four screening cohorts included in the Nodule Malignancy project in the INTEGRAL program. The circulating proteome in prediagnostic plasma samples collected during LDCT screening was quantified with Olink proteomics platform using the proximity extension assay (PEA), as described in previous works [2, 312].

In this chapter, the goal is to develop and publicly release an open-access dataset, including images and proteomics, of the novel P-ELCAP lung cancer screening cohort for which multimodal data have been collected. The dataset incorporates the annotated LDCT image, the relative concentration of more than 1000 blood plasma circulating proteins markers, as well as several clinical and demographic variables. The dataset comprises data from 211 P-ELCAP LDCT-based screening participants, including 67 lung cancer cases, 68 matched controls with benign pulmonary nodules confirmed in subsequent follow-ups, and 71 matched controls without lung nodules. Moreover, there are 5 imaging false positives. LDCT imaging and blood collection took place within a similar time range of ± 6 months, always before surgery in participants diagnosed with cancer. In the case of these biopsy confirmed tumors, the image and blood were collected within 5 years prior to diagnosis. Individuals selected for both control groups (with benign nodules and without nodules) were matched with cancer cases for age, sex and time of sample/image collection. In total, there are 138 nodules segmented in the dataset, 65 malignant (a few were not visible at the LDCT corresponding to the blood collection), 68 benign and 5 corresponding to imaging false positives. The dataset has been fully anonymized to prevent exposure of sensitive patient information. As aforementioned, this study is part of the P-ELCAP cohort collected at CUN (Spain) since 2000. To the best of our knowledge, this is the first pulmonary nodule dataset to incorporate annotated LDCT images together with molecular protein biomarkers to leverage multimodal early lung cancer diagnosis methods. We further demonstrate, through a proof-of-concept analysis, the potential of this dataset for personalized lung cancer risk assessment, supported by technical validation using machine learning (ML) models applied to LDCT images and Olink circulating protein markers.

The contributions and innovative aspects of this chapter can be summarized as follows:

- We have curated and will provide open access to a novel multimodal early lung cancer screening dataset. The dataset incorporates annotated screening LDCT scans and plasma proteomics data measured with the Olink Target 96 platform (>1000 proteins) for a total of 211 screening participants.

- For lung cancer cases, pathological gold standard label confirmed the presence and subtype of tumor. There are 67 lung cancer patients (38 adenocarcinoma, 12 squamous cell carcinoma, 9 large cell carcinoma, 4 small cell carcinoma, 2 mixed adeno-squamous cell carcinoma and 2 other/NOS), 68 controls with benign pulmonary nodules, 71 controls without nodules and 5 surgically excised false positive lesions.

- Several experiments were conducted using ML models on the developed dataset. The results highlight inherent challenges associated with the limited dataset size. Nevertheless, due to its high quality and careful curation, the dataset holds significant value for external validation in other cohorts. Consequently, this dataset is highly valuable and represents an important step towards advancing the field of personalized medicine through AI-based systems in the context of LDCT-based lung cancer screening.

7.3 Materials and methods

The development and open access release of P-ELCAP dataset consisted of four steps: lung cancer screening sub-cohort selection of three matched groups of participants plus false positives; data acquisition and anonymization; data selection, curation and protein analysis; data annotation, preprocessing and dataset submission to the repository, where it will be publicly released upon acceptance of the associated paper. Furthermore, two additional steps involve developing a multimodal nodule risk malignancy tool: implementation of baseline ML models, and final development of a fusion module ready for validation. The fusion model strategy integrating radiological and molecular signatures is a highly promising objective, which will be described in Chapter 9. The six steps of the multimodal nodule risk scoring tool development process are depicted in Figure 7.1.

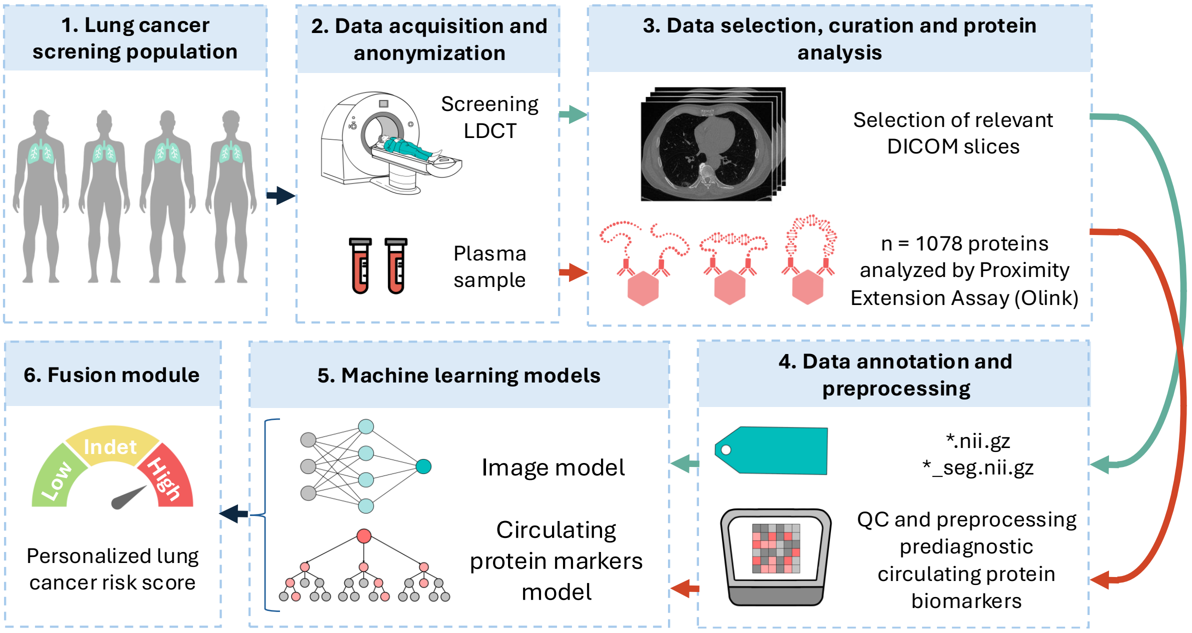

1. Lung cancer screning population
2. Data acquisition and anonymization
Screening LDCT
Plasma sample
3. Data selection, curation and protein analysis
Selection of relevant DICOM slices
n = 1078 proteins analyzed by Proximity Extension Assay (Olink)
4. Data annotation and preprocessing
*.nii.gz
*_seg.nii.gz
QC and preprocessing prediagnostic circulating protein biomarkers
5. Machine learning models
Image model
Circulating protein markers model
6. Fusion module
Low
Indet
High
Personalized lung cancer risk score

Figure 7.1: Workflow diagram. This diagram provides an overview of the dataset preparation and validation stages, including Lung cancer screening population, Data acquisition and anonymization, Data selection, curation and protein analysis, Data annotation, and preprocessing, Machine learning models and Fusion module.

7.3.1 Collection principles

To ensure precise subject selection and discard the impact of irrelevant factors, we adhered to the following principles in our case selection process:

Inclusion criteria. Participants were selected from a lung cancer screening population including men and women aged 40 years or older who were current or former smokers with a minimum smoking history of 10 pack-years, and who exhibited no symptoms of lung cancer at the time of enrollment. The specific inclusion criteria for P-ELCAP are further described in previous works by Sanchez-Salcedo *et al.* [314] and Mesa-Guzmán *et al.* [315]. Individuals with benign nodules were included as controls if the nodules remained stable in size during a follow-up period of at least two years.

Exclusion criteria. Participants with a prior history of lung cancer were excluded, as well as those diagnosed with any form of cancer within four years before or after the time of LDCT imaging and blood collection, with the exception of non-melanoma skin cancer.

Pre-treatment imaging. LDCT scans were obtained before the administration of any relevant treatments and surgery, ensuring that the nodule's imaging features were not altered by prior medical interventions.

Quality control. A comprehensive evaluation of LDCT images was conducted, discarding those with missing or incorrect layering to preserve the integrity and completeness of the lung nodule dataset.

7.3.2 Data collection and annotation

This cohort is part of the International Early Lung Cancer Action Program (P-ELCAP) conducted at Clínica Universidad de Navarra, Spain. This specific P-ELCAP subcohort was retrospectively collected, and was also included in the Integrative Analysis of Lung Cancer Risk and Etiology (INTEGRAL) research program. Ethical approval was obtained from the University of Navarra Research Ethics Board (ref 2020.251, January 25th 2021). Informed consent was received from each individual participant. The dataset includes a total of 67 lung cancer cases collected within 5 years prior to diagnosis, 68 matched controls with benign pulmonary nodules, and 71 matched controls without lung nodules. For each case, sex, age and plasma sample collection time with respect to LDCT imaging were used to match controls with benign nodules and controls without nodules within 5 years. Lung cancer cases were considered if biopsy-confirmed lung cancer was diagnosed within 5 years. Additionally, there are 5 imaging false positive participants, corresponding to patients who underwent surgery to remove a suspicious nodule and, following resection, lung cancer was conclusively excluded by the pathologist as a diagnosis. The information available for each case was the result of proteomics analysis performed by 12 Olink Target 96 panels (more than 1000 proteins, further details are provided in the subsequent section), as well as LDCT imaging data, together with clinical and demographic tabular data.

7.3.3 Data records and code

We will release the fully anonymized dataset, containing the aforementioned information for each case, in the EU-accredited repository Zenodo [316]. The dataset is deposited under a Creative Commons CC-BY-NC-SA license. This license enables users to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and provided proper attribution is given to the creator both for the material or its modifications. The latest version of the informed consent and patient information datasheet was approved by the University of Navarra Research Ethics Committee on its June 2nd 2022 session. In the informed consent, the screening participants provided permission for their data to be used exclusively in lung cancer research. As we do not foresee any applications beyond lung cancer, it is our view that all prospective users fall within the intended scope. Hence, a limitation in the Zenodo web access for lung cancer research is implemented solely through a requester intention confirmation box, without requiring a formal Declaration of Use. Thus, users will be granted access to the dataset in Zenodo upon stating through the request process that the dataset will be used exclusively for lung cancer research. The dataset's file structure is shown in Fig 7.2, the different data modalities can be matched using the participant's anonymized ID in the .csv and .xlsx files.

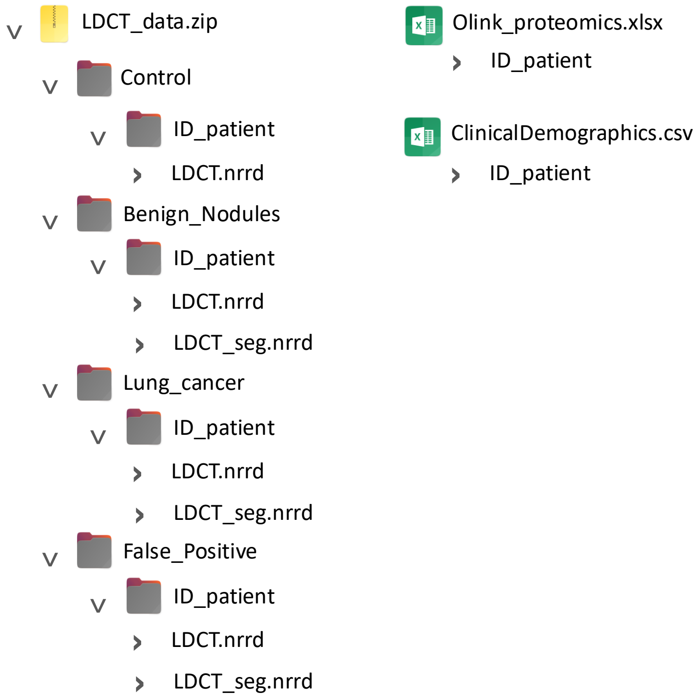


Figure 7.2: Structure and hierarchy of dataset files.

To ensure easy access to the code to read and process this dataset, we have made it available in a GitHub repository `https://github.com/MiriamCobo/P-ELCAP_Dataset.git`. The repository provides Python sample code demonstrating how to access and utilize the data, facilitating understanding and reuse.

7.4 Results

7.4.1 Dataset properties

Most participants were men (171 men compared to 40 women), and the median age at blood collection was 61 years (interquartile range [IQR], 13 years). The median time between pre-diagnostic blood collection and diagnosis was 1.4 years (IQR 3 years, range: 0-5 years, by design). Of the participants diagnosed with lung cancer, 31 cases occurred at baseline or within the first year (using the date of blood and image collection as the reference baseline point). The remaining diagnoses occurred over subsequent years: 8 after one year, 9 after two years, 10 after three years, 7 after four years, and 2 after five years. Details on the characteristics of the P-ELCAP cohort population are described in Table 7.1. The different data modalities included in the dataset are comprehensively documented in the following paragraphs.

Olink proteomics. Relative protein concentrations of up to 1078 individual proteins across 12 panels focused on a specific area of disease or biology (Cardiometabolic, Cardiovascular II, Cardiovascular III, Development, Immune Response, Inflammation, Metabolism, Neurology,

Characteristic	**Cancer**	**Benign nodules**	**Controls**	**False positives**	**Total**
Sex n (%)					
Female	12 (18.5%)	13 (19.1%)	13 (18.3%)	2 (40.0%)	40 (19.1%)
Male	55 (82.1%)	55 (80.9%)	58 (81.7%)	3 (60.0%)	171 (81.0%)
Age (years)					
Median (IQR)	61.0 (14.5)	61.5 (14.0)	60.0 (12.0)	68.0 (17.0)	61.0 (13.0)
Follow-up time (years)					
Median (IQR)	8.0 (7.0)	7.5 (6.0)	7.0 (7.5)	13.0 (7.0)	7.0 (6.5)
Follow-up survival time (years)					
Median (IQR)	2.0 (4.0)	-	-	-	2.0 (4.0)
N/A[a] n (%)	46 (68.7%)	68 (100%)	71 (100%)	5 (100%)	190 (90.0%)
Smoking status n (%)					
Current	34 (50.7%)	31 (45.6%)	42 (59.2%)	2 (40.0%)	109 (51.7%)
Former	33 (49.3%)	37 (54.4%)	29 (40.8%)	3 (60.0%)	102 (48.3%)
Smoking duration (years)					
Median (IQR)	41.3 (14.4)	40.3 (15.2)	38.6 (15.7)	38.0 (10.0)	40.4 (15.0)
Smoking pack years					
Median (IQR)	41.0 (36.1)	31.8 (28.1)	26.8 (22.4)	42.0 (14.4)	35.4 (28.0)
Smoking quit years (former)					
Median (IQR)	8.0 (11.8)	8.9 (16.3)	15.2 (19.1)	12.7 (2.1)	9.7 (15.0)
Emphysema n (%)					
Yes	30 (44.8%)	15 (22.1%)	15 (21.1%)	0 (0.0%)	60 (28.4%)
No	37 (55.2%)	53 (77.9%)	56 (78.9%)	5 (100.0%)	151 (71.6%)
COPD n (%)					
Yes	41 (61.2%)	23 (35.4%)	23 (37.7%)	2 (40.0%)	89 (42.2%)
No	26 (38.8%)	42 (64.6%)	38 (62.3%)	3 (60.0%)	109 (51.7%)
Unknown[b]	-	3 (4.4%)	10 (14.1%)	-	13 (6.2%)
Body Mass Index					
Median (IQR)	26.2 (6.4)	27.7 (5.1)	27.3 (4.5)	23.5 (4.0)	26.9 (5.1)
Family history of lung cancer n (%)					
Yes	17 (25.4%)	9 (13.2%)	10 (14.1%)	0 (0.0%)	36 (17.1%)
No	50 (74.6%)	59 (86.8%)	61 (85.9%)	5 (100.0%)	175 (82.9%)
Personal history of cancer n (%)					
Yes	4 (6.0%)	8 (12.1%)	5 (7.0%)	3 (60.0%)	20 (9.5%)
No	63 (94.0%)	58 (87.9%)	66 (93.0%)	2 (40.0%)	189 (89.6%)
Unknown[b]	-	2 (2.9%)	-	-	2 (0.9%)
Stage n (%)					
Early stage (TNM I/II)	53 (79.1%)				
Advanced stage (TNM III/IV)	14 (20.9%)	-	-	-	-
Histology n (%)					
Adenocarcinoma	38 (56.7%)	-	-	-	-
Squamous cell carcinoma	12 (17.9%)	-	-	-	-
Large cell carcinoma	9 (13.4%)	-	-	-	-
Small cell carcinoma	4 (6.0%)	-	-	-	-
Mixed adenocarcinoma and squamous cell carcinoma	2 (3.0%)	-	-	-	-
Other/NOS	2 (3.0%)	-	-	-	-

Table 7.1: Key characteristics of 67 lung cancer cases, 68 matched controls with benign pulmonary nodules, 71 matched controls without lung nodules and 5 false positives from P-ELCAP cohort. [a] N/A indicates that the patient is still alive; [b] Unknown indicates that there was no data available. Age was measured at the time of blood plasma collection.

Oncology II, Oncology III, Organ Damage, NeuroExploratory) in prediagnostic plasma samples were measured with Olink Target 96 platform. Quantification measurement was carried out by Olink in their central laboratories in Uppsala, Sweden [312]. This high-throughput technology is based on the highly sensitive PEA technique, as previously described by the INTEGRAL consortium [312, 2]. Relative protein concentrations are reported as normalized protein expression (NPX) values on a $\log_2$ scale. These values are derived from quantitative polymerase chain reaction (PCR) cycle threshold measurements and were standardized for subsequent analysis [312]. A detailed description of the Olink methodology is available in the white paper published on Olink's official website [317].

LDCT scans. For each LDCT, we selected the series with the thinnest CT image slices, defined as having a slice thickness of $\leq$ 3 mm, for inclusion in the dataset. LDCT scans were collected within a 6-month window (184 days), either before or after the date of blood plasma collection. For 5 lung cancer cases the nodule was not visible in the LDCT scan at the time of the plasma sample. Lung nodules were annotated by an experienced radiology technician, and subsequently supervised, refined, and approved by a certified experienced thoracic radiologist, who is a member of the P-ELCAP LDCT multidisciplinary team. Figure 7.3 shows horizontal bar plots of nodule dimensions (width, height, and depth), and the anatomical location of the nodules. We additionally show one example of LDCT scans for a participant belonging to each of the groups in Figure 7.4.

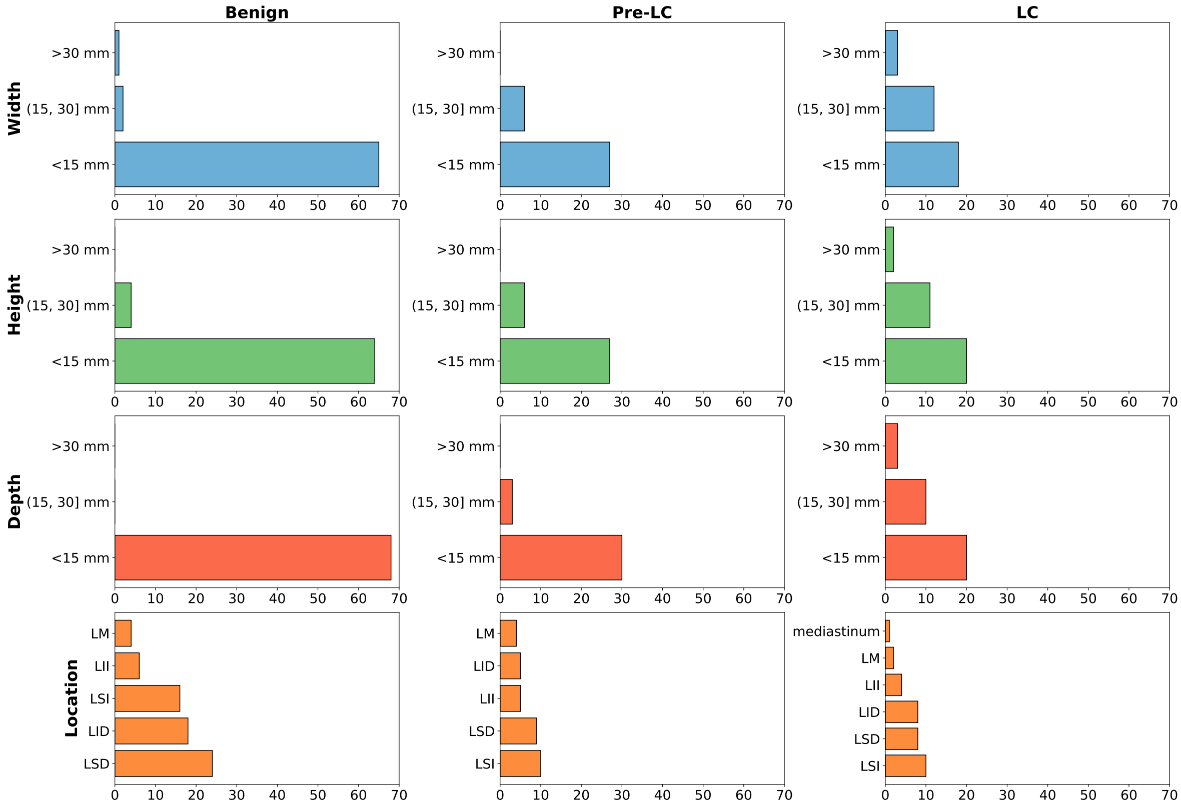


Figure 7.3: Distribution of lung nodule sizes and locations across diagnostic groups. The x-axis denotes the nodule count. These groups include controls with benign nodules ("Benign"), lung cancer from 1-5 years before diagnosis ("Pre-LC"), and lung cancer less than 1 year before diagnosis ("LC").

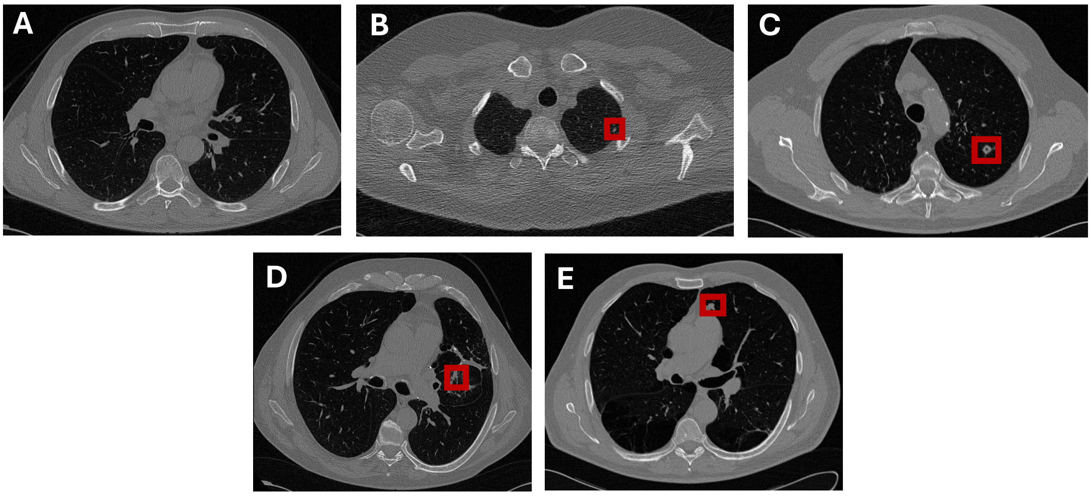


Figure 7.4: Example of a LDCT image. **A:** Control without nodules. **B:** Control with benign nodule. **C:** Imaging false positive. **D:** Lung cancer from 1-5 years before diagnosis. **E:** Lung cancer less than 1 year before diagnosis.

Clinical and demographic data. The clinicians involved in the P-ELCAP multidisciplinary team (mainly neumologists and specialized nurses) collected the demographic/clinical information of the study participants (age, sex, smoking questionnaire, comorbidities, exposures, lifestyle, etc.). Basic demographic information (age, sex, smoking status) is provided in the publicly available data for each individual in the Zenodo dataset. All cases enrolled in the study are of Caucasian ethnicity. The participant age in the dataset corresponds to the date of blood collection. We also include in the open access dataset the difference in years and in months between the date of blood collection and the date of diagnosis, rounding to the nearest month unit. For control participants (with or without nodules), this time was set to the maximum follow-up time observed among lung cancer participants (5 years or 60 months) to enable the use of time-to-event models such as random survival forest. The following variables are incorporated: Group, Age, Sex, Stage, Smoking (in pack years), and the aforementioned time in years and months.

7.4.2 Technical evaluation

To validate the dataset proposed in this study, a proof-of-concept and a technical validation were conducted using the data available in the open-access resource. The aim is to perform two baseline studies for technical validation: image-based deep learning (DL) models from the available LDCT images; and ML protein models from Olink proteomics data. Data was divided into 3 fold cross-validation (CV) train and test splits, and in the image models a validation subset representing 10% of training data was further separated. Data was stratified according to Figure 7.5. The same 3 folds were used to train and evaluate all the methods, filtering by the groups of participants depending on the case. All the analyses were done in Python. Performance was evaluated using binary classification metrics: area under the curve (AUC), accuracy, balanced accuracy, precision (i.e., positive predictive value, PPV), recall (i.e., sensitivity), F1-score, specificity and negative predictive value (NPV). Unless otherwise specified, a threshold of 0.5 was used to compute the binary values. In the following sub-sections we describe in more detail the two proof-of-concept studies.

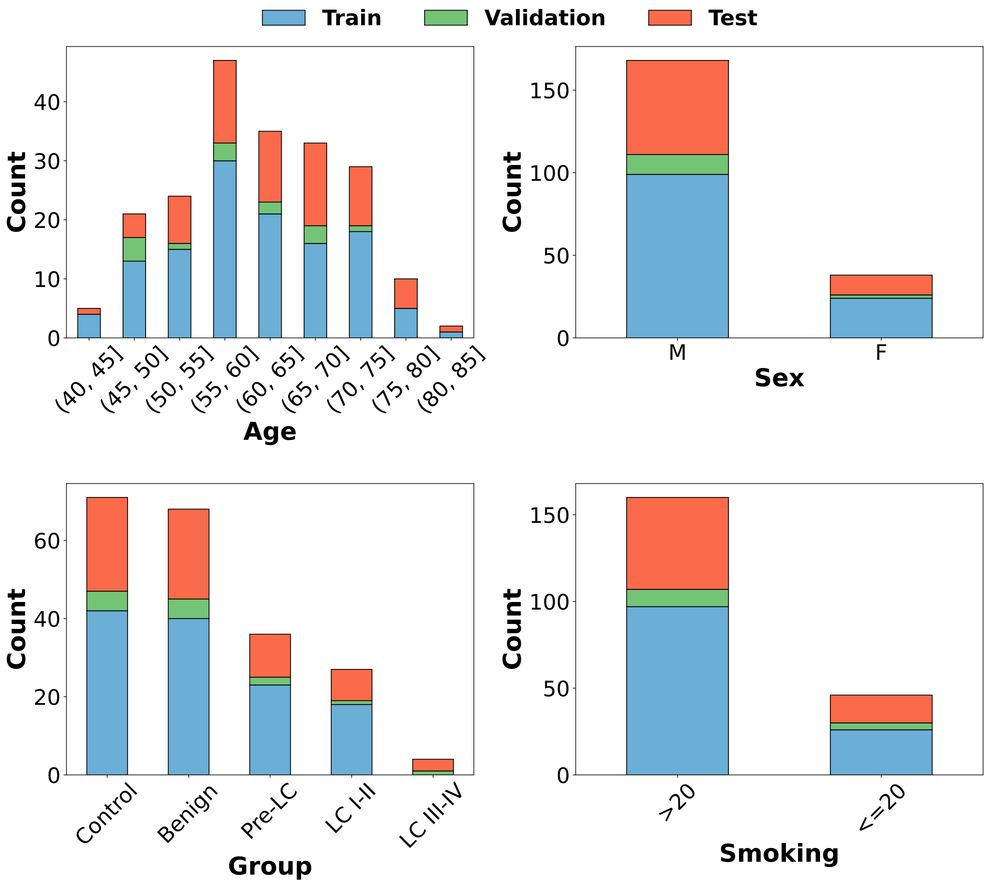


Figure 7.5: Stratification of data splits (train, validation, test) in the first fold according to age, sex, group and smoking (in pack years) variables. In the panel labeled “Group”, “Control” corresponds to controls without nodules, “Benign” to controls with nodules, “Pre-LC” to lung cancer cases from 1 to 5 years prior diagnosis, “LC I-II” to lung cancer cases at stages I-II, and “LC III-IV” to lung cancer cases at stages III-IV.

Image models

LDCT images from control participants with benign nodules and the corresponding lung cancer participants diagnosed in the same year as the blood collection were included in this analysis. We explored the effectiveness of 3D DL architectures for lung nodule classification. A crop around the nodule’s centroid, determined with the segmentation, with 3D size $[32, 32, 32]$ in pixels was performed to include surrounding information. Nodules larger than the specified size in any dimension were down-sampled in that dimension to ensure compatibility with the network’s architecture. We developed the models (DenseNet121, EfficientNet, ResNet34) using MONAI [26] and Pytorch, incorporating pretrained weights from Med3D when available to enhance performance [24]. A weighting strategy was used in the loss function to improve learning from imbalanced class distributions. We included small data augmentations in the train subset with a probability of 50% (random rotations in a range of 30° and isotropic scaling by a factor uniformly sampled between 0.9 and 1.1.), as well as dropout to enhance the learning of the model [4] and improve generalization. Table 7.2 presents the performance metrics. The results are promising in some of the comparative indicators (AUC, accuracy, PPV, NPV) especially in the EfficientNet analysis, but at the same time reveal the limitations of training deep architectures with a small dataset size [4]. What is clear is the potential of this open dataset if used in combination with other datasets, which eventually may become available in the future by other groups conducting research on lung cancer screening cohorts.

Table 7.2: Results of the deep learning models trained on low-dose computed tomography data for predicting benign participants and patients diagnosed with lung cancer in the same year as the plasma collection, evaluated using 3-fold cross-validation test sets. All models were implemented in their 3D architecture.

Metric	**DenseNet121**	**EfficientNet**	**ResNet34**
Pretraining	No	No	Yes
AUC	0.50 ± 0.07	0.86 ± 0.04	0.66 ± 0.05
Accuracy	0.43 ± 0.06	0.80 ± 0.04	0.60 ± 0.08
Balanced Accuracy	0.51 ± 0.10	0.76 ± 0.09	0.61 ± 0.09
Precision/PPV	0.33 ± 0.07	0.74 ± 0.05	0.3 ± 0.3
Recall/Sensitivity	0.76 ± 0.19	0.6 ± 0.2	0.6 ± 0.5
F1-score	0.46 ± 0.09	0.66 ± 0.14	0.4 ± 0.3
Specificity	0.26 ± 0.01	0.88 ± 0.07	0.6 ± 0.4
NPV	0.72 ± 0.16	0.84 ± 0.08	0.84 ± 0.16

Protein models

Almost 1000 circulating proteins were used in this second part of the proof-of-concept study. Markers with missing values or that failed to reach Olink's quality control were excluded from the technical validation. However, in the open database available at Zenodo, the raw values of these proteins are also provided, as they may be useful to support some other type of analyses, allowing for missing values. This filtering resulted in a total of 970 proteins. The same 3 fold CV previously described with training and test subsets was used. The NPX data (refer to Olink's white paper for further details [317]) were processed using standard scaler trained on the training subset (comprising both the training and validation sets), and then subsequently applied to the test subset. The best model for each fold was found through another grid search 3 fold CV on the training subset, and then it was evaluated on the separated test subset of that fold. We trained three ML models: random forest (RF), xgboost (XGB) and penalized regression (LASSO) to distinguish between benign and lung cancer participants diagnosed in the same year as the plasma collection. We additionally trained a novel architecture on biologically informed neural networks (BINNs) [157], that leverages the pathways from Reactome pathway database [318] to create a sparse neural network. Results are shown in Table 7.3. Furthermore, a random survival forest (RSF) was trained on all lung cancer and benign participants, oversampling lung cancer participants in the training set, that were repeated. Results for participants in years 0, 1, and 2 previous to lung cancer diagnosis are reported in Table 7.4. The most relevant proteins identified by the models are listed in Table 7.5. The selection involved first identifying the 100 most relevant proteins in each fold, after which those consistently present across all three folds were included in the final list.

Table 7.3: Results of the machine learning models trained on protein data for predicting benign participants and patients diagnosed with lung cancer in the same year as the plasma collection, evaluated using 3-fold cross-validation test sets.

Metric	**RF**	**XGB**	**LASSO**	**BINN**
AUC	0.65 ± 0.08	0.65 ± 0.03	0.71 ± 0.07	0.59 ± 0.19
Accuracy	0.72 ± 0.08	0.71 ± 0.06	0.71 ± 0.05	0.57 ± 0.13
Balanced Accuracy	0.61 ± 0.05	0.63 ± 0.01	0.62 ± 0.08	0.57 ± 0.16
Precision/PPV	0.7 ± 0.3	0.58 ± 0.14	0.52 ± 0.11	0.38 ± 0.16
Recall/Sensitivity	0.32 ± 0.05	0.41 ± 0.07	0.39 ± 0.19	0.57 ± 0.20
F1-score	0.42 ± 0.07	0.47 ± 0.01	0.44 ± 0.17	0.45 ± 0.18
Specificity	0.90 ± 0.09	0.84 ± 0.10	0.85 ± 0.02	0.57 ± 0.10
NPV	0.74 ± 0.04	0.76 ± 0.02	0.76 ± 0.06	0.75 ± 0.11

Table 7.4: Results of the random survival forest trained on protein data for predicting benign participants and patients diagnosed with lung cancer at different time points, evaluated using 3-fold cross-validation test sets. The classification threshold was empirically set at 0.6, as this value provided a better trade-off between sensitivity and specificity across the evaluated folds.

Metric	**Year 0**	**Year 1**	**Year 2**
AUC	0.65 ± 0.10	0.68 ± 0.09	0.62 ± 0.06
Accuracy	0.75 ± 0.09	0.68 ± 0.10	0.55 ± 0.06
Balanced Accuracy	0.61 ± 0.06	0.63 ± 0.06	0.57 ± 0.05
Precision/PPV	0.78 ± 0.05	0.75 ± 0.05	0.67 ± 0.06
Recall/Sensitivity	0.92 ± 0.09	0.78 ± 0.17	0.47 ± 0.11
F1-score	0.84 ± 0.06	0.76 ± 0.10	0.55 ± 0.08
Specificity	0.29 ± 0.04	0.49 ± 0.06	0.67 ± 0.07
NPV	0.7 ± 0.3	0.6 ± 0.2	0.47 ± 0.06

Table 7.5: Proteins ranked among the top 100 in predicting benign participants and patients diagnosed with lung cancer in the same year as the plasma collection across all three folds based on feature importance analysis. In **bold**, proteins that are repeated across different methods. In *italic*, proteins that were in the list of the 10 protein markers most relevant for imminent lung cancer in a previous work within the INTEGRAL consortium [2].

Model	**Most relevant proteins**
RF	**AP-N**, CCL3.1, CD93, CDH2, FR-alpha, IGFBP3, IL15, MCP-1, **PVRL4**, **TCL1A**, TF
XGB	**AP-N**, BNP, F11, FCRL6, SEMA7A
LASSO	*CEACAM5*, **EGFR**, ENPP7, *FASLG*, NCR1
RSF	AZU1, IL18.1, KIRREL2, **PVRL4**, **TCL1A**
BINN	**EGFR**, EIF4G1, HB-EGF, TR

7.5 Discussion and conclusions

P-ELCAP dataset is a valuable lung cancer screening cohort to support open access research. This dataset comprises a small-medium size cohort including 67 lung cancer patients, 68 controls with benign pulmonary nodules, 71 controls without nodules and 5 imaging false positives. The sample size of cancer cases in the cohort reflects the characteristics of lung cancer screening populations, where only approximately 1-3% of individuals are diagnosed with the disease, most often at early stages. Specifically, the CUN has screened a total of 5647 individuals over the past 25 years, diagnosing lung cancer in 158 of them. This low prevalence poses limitations for the development and validation of predictive models, particularly for early-stage malignancy detection. To the best of our knowledge, this is the first multimodal open access database including long-term follow-up benign and biopsy confirmed malignant pulmonary nodules, as well as controls without nodules. Furthermore, this is the first publicly available lung cancer screening dataset incorporating over 1000 circulating protein markers generated by PEA. The applications of this dataset for lung cancer research extend beyond the scope of this work, which focuses on assessing data quality, and leaves further discoveries to future investigations.

The technical validation results underscore both the potential of the dataset and the inherent challenges associated with its limited size. The LDCT image-based models can be further refined leveraging other publicly available datasets to enhance current results, that as a proof-of-concept were only trained on P-ELCAP dataset. The proposed protein models identified a set of proteins that consistently appeared across different methods, demonstrating robustness and alignment with findings from previous studies [2]. Moreover, the availability of a set of approximately 1000 circulating proteins in individuals with malignant in contrast to benign nodules may open a wealth of research hypothesis or help as a validation tool for discovery projects performed in similar or even distant clinical settings. The field of characterization of incidental nodules may highly benefit from the data that will be made available in the present database.

P-ELCAP dataset holds significant value for external validation in other screening cohorts, and represents an important step towards advancing the field of personalized medicine through AI-driven multimodal models in the context of LDCT-based lung cancer screening. This dataset will also help in the development and validation of novel multimodal strategies in early lung cancer screening. Future work may enhance the proposed image models through pretraining in larger publicly available LDCT imaging datasets, with evaluation on the P-ELCAP cohort to investigate the generalizability of existing DL approaches for multimodal lung nodule detection and risk characterization. The following Chapters, 8 and 9, present, respectively, the validation of a deep learning ordinal approach for lung nodule risk assessment and a multimodal approach for early lung cancer diagnosis. Although the P-ELCAP dataset represents an important resource for research in this field, further studies and external validation using independent cohorts are necessary to achieve real-world clinical utility of multimodal AI tools in LDCT-based lung cancer screening.

Chapter 8

Validation of a medical informed deep learning ordinal approach for lung nodule assessment

Main publication associated with this chapter: **Cobo, Miriam**, *et al.* "Validating a medical informed deep learning ordinal approach for lung nodule assessment". *To be submitted.*

8.1 Introduction and motivation

In early lung cancer LDCT-based imaging, AI models are often trained and evaluated on a single dataset, typically LIDC-IDRI [201], focusing primarily on optimizing specific performance metrics rather than meeting broader clinical needs, such as generalizability and real-world applicability. Both a deep understanding of clinical workflows and robust external validation are critical for the effective translation of AI models into clinical settings. This chapter investigates the generalizability of state-of-the-art 2D and 3D deep learning models within a medical informed framework for lung nodule malignancy risk assessment. An ordinal learning approach is adopted, reflecting the inherently ordinal nature of screening pulmonary nodules, addressing clinical needs, and resembling radiological decision-making. The algorithms are externally validated under identical conditions on the P-ELCAP dataset, a well-curated lung cancer screening cohort described in Chapter 7, which complies with standard clinical guidelines for lung nodule management. Additionally, a 3D capsule network architecture is explored, achieving the highest metrics in the external validation dataset.

In this chapter, learning biases are informed by clinical prior knowledge, similarly to the shared embedding incorporating prognosis and the highly correlated relevant variables proposed in Chapter 6. The ordinal regularization of the feature space promotes coherence with established clinical protocols in lung nodule management, enhancing the robustness, clinical applicability and generalizability of DL models.

Before investigating on the proposed methodology, an existing imaging model from the literature was externally evaluated on P-ELCAP dataset. This assessment revealed several limitations that hindered clinical applicability and translation, which in turn motivated the development of the medical informed ordinal approach described in the remainder of this chapter. The following paragraphs summarize the shortcomings of the existing image model.

8.1.1 Limitations of state-of-the-art approaches

Deep learning (DL) image algorithms for early lung cancer diagnosis in LDCT screening can be categorized according to the level at which predictions are performed:

- Local pulmonary nodule malignancy assessment (nodule level) [3, 228, 235, 296]. This level corresponds to the majority of existing DL approaches in the literature, since indeterminate risk pulmonary nodules (IPNs) are typically smaller than 30mm, and thus represent a minor part of the overall LDCT volume. This strategy is explored in the following sections of this chapter.
- Global lung cancer risk prediction from a single volumetric LDCT scan acquired years prior to diagnosis (scan level). Two relevant recent approaches were further studied within this group: Ardila *et al.* [319], who developed a DL algorithm that analyzes pulmonary nodules to predict lung cancer within 1 and 2 years, and Mikhael *et al.* [301], who extended their work to predict lung cancer occurring 1-6 years after the screen. Both algorithms were trained on LDCT images from the National Lung Screening Trial (NLST) [299].

Additionally, in the literature there are other approaches that investigate combinations of the aforementioned categories, such as Zhang *et al.* [320], who compared four ResNet image models at different levels (whole lung, image patch where the nodule was located, nodule box and follow-up nodule box) to predict malignancy of sub-centimeter pulmonary nodules using CT images. Their findings concluded that the performance of the models at the nodule box level was significantly better than the image patch or the whole lung models [320].

Methodology

In the preliminary analysis of image models, Sybil [301] was selected for evaluation on P-ELCAP dataset, as it offered a broader prediction range of lung cancer from 1 to 6 years before diagnosis, and represented a natural extension of the work by Ardila *et al.* [319]. Additionally, the code and the weights of the model were available, facilitating usability and reproducibility. However, certain limitations were identified in the reported results, as the evaluation of performance metrics reported only concordance index (c-index), receiver operating characteristic (ROC) curve and AUC score [301].

The c-index measures the discriminatory ability of a risk prediction model in survival analysis, evaluating how well predicted risk scores align with observed survival outcomes while accounting for censoring [321]. A pair of subjects is considered comparable if one experienced an event before the other or if one was censored after the other had an event. Among comparable pairs, a pair is said to be *concordant* if the subject with the shorter observed survival time has a higher predicted risk score, *discordant* if the opposite is true, and *tied* if both have the same predicted score (within a numerical tolerance), contributing half a point to the total. Let N_c be the number of concordant pairs, N_d the discordant pairs, and N_t the tied pairs. The total number of comparable pairs is $N_{\text{comp}} = N_c + N_d + N_t$. The c-index is then defined as [321]:

$$\texttt{c-index} = \frac{N_c + \frac{1}{2}N_t}{N_{\text{comp}}} \tag{8.1}$$

A value of `c-index` $=$ 1.0 indicates perfect concordance, `c-index` $=$ 0.5 corresponds to no predictive discrimination power, and `c-index` $<$ 0.5 suggests worse than random performance [321].

Regarding the aforementioned limitations, firstly, c-index is a measure of rank correlation that considers the order of event times, but does not measure the event status itself, hence, it is often not clinically meaningful in assessing model performance [322]. Secondly, ROC curves are not the most suitable metric for imbalanced datasets [323], in particular if reported alone. The standard form of ROC curves without threshold information hinders direct comparison of model performance at specific operating points, thereby limiting the interpretability [324]. Thus, the performance of risk prediction models for decision making, which is the essence of clinical translation for DL-based systems, has to be conditional on a risk threshold to fix classification costs [324]. Sybil was trained on NLST, a screening cohort that is, by nature, highly imbalanced, with approximately 5% positive LDCTs corresponding to lung cancers diagnosed within the following 6 years [301]. Given that Sybil [301] does not include risk thresholds, in its current form, the clinical applicability is very limited.

To mitigate the previous shortcomings, Sybil was evaluated on P-ELCAP including the following standard metrics for binary classification: accuracy (Acc.), balanced accuracy (B. Acc.), specificity (Spec.), negative predictive value (NPV), precision (Prec.), recall and F1-score.

Data

All LDCT scans from P-ELCAP cohort were included in the analysis with Sybil (67 lung cancer, 68 controls with benign pulmonary nodules, 71 controls without nodules and 5 false positives, as described in Chapter 7). The goal was to assess whether Sybil was capable of predicting lung cancer in screening participants from P-ELCAP cohort. Patients diagnosed in the same year as the blood collection were considered positive in all the years of Sybil's prediction.

Results and discussion

Table 8.1 shows the performance of Sybil in P-ELCAP cohort for a standard threshold of 0.5. Results indicate that this threshold is not suitable for evaluating Sybil. Therefore, to find the most suitable threshold, Youden's index (YI) was calculated to identify the point on the ROC curve that maximizes the difference between the true positive rate (TPR) and the false positive rate (FPR) , thereby optimizing the balance between sensitivity and specificity:

$$\mathrm{YI} = \mathrm{TPR} - \mathrm{FPR} \tag{8.2}$$

Table 8.2 shows the performance of Sybil on P-ELCAP evaluating binary metrics in Youden's index best threshold. The results presented in this table indicate a precision of 0.5 predicting lung cancer occurring one year after the screen, meaning that 50% of the positive predictions are false positives, which is insufficient for lung cancer screening. This limitation is critical, as accurate short-term prediction, especially within the year prior to diagnosis, is essential for timely clinical decision-making and improved patient outcomes.

In an attempt to improve the previous results, Sybil was re-calibrated on P-ELCAP; however, this approach did not yield significant improvements. As a final step, Sybil was fine-tuned on

Table 8.1: Sybil's performance on LDCT images from P-ELCAP dataset (threshold = 0.5). [a]Average is not applicable to c-index.

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Average
c-index							0.710[a]
AUC	0.839	0.814	0.814	0.811	0.807	0.808	0.815
Acc.	0.820	0.787	0.739	0.706	0.697	0.720	0.745
B. Acc.	0.513	0.531	0.537	0.532	0.534	0.576	0.537
Spec.	1.000	1.000	0.987	0.986	0.979	0.972	0.987
NPV	0.819	0.784	0.740	0.706	0.698	0.718	0.744
Prec.	1.000	1.000	0.714	0.714	0.667	0.750	0.808
Recall	0.026	0.062	0.086	0.077	0.090	0.179	0.087
F1-score	0.050	0.118	0.154	0.139	0.158	0.289	0.151

Table 8.2: Sybil's performance on P-ELCAP dataset (threshold is Youden's index). [a]Average is not applicable to c-index.

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Average
c-index							0.789[a]
AUC	0.839	0.814	0.814	0.811	0.807	0.808	0.815
Threshold	0.012	0.024	0.052	0.035	0.044	0.065	0.039
Acc.	0.815	0.801	0.806	0.749	0.749	0.749	0.778
B. Acc.	0.817	0.790	0.780	0.754	0.752	0.752	0.774
Spec.	0.814	0.810	0.837	0.740	0.743	0.743	0.781
NPV	0.952	0.923	0.889	0.878	0.870	0.870	0.897
Prec.	0.500	0.544	0.627	0.568	0.580	0.580	0.566
Recall	0.821	0.771	0.724	0.769	0.761	0.761	0.768
F1-score	0.621	0.638	0.672	0.654	0.658	0.658	0.650

P-ELCAP. Data was separated into 139 patients for training, 25 for validation and 47 for testing (including the 5 false positives), following the stratification procedure described in Section 7.4.2. Only the final layers were retrained, i.e., the fully connected and hazard layers [301]. This approach incorporated patients diagnosed in the same year as the LDCT scan acquisition (i.e., year 0). Table 8.3 shows Sybil's performance on the separated test subset of P-ELCAP, after retraining the final layers on the rest of the dataset.

The results presented in Table 8.3 were not satisfactory for clinical applicability, as the model exhibited poor precision and recall, particularly in the years immediately preceding lung cancer diagnosis. This is especially problematic in lung cancer screening, where low precision can result in a high rate of false positives, leading to unnecessary follow-ups or biopsies, with potential complications, increased patient anxiety, and inefficient use of medical resources.

The aforementioned limitations motivated the development of a new medical informed ordinal approach for lung cancer assessment at the nodule level, aligning with the majority of existing DL methods in the literature. This approach is further supported by prior evidence demonstrating improved performance at the local nodule level compared to the global LDCT scan level [320]. The proposed methodology is explained in the remainder of this chapter.

Table 8.3: Performance retraining the last layers of Sybil on LungAmbition dataset (best threshold) on 47 test patients. ^{a}Average is not applicable to c-index.

	Year 0	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Average**
c-index							0.676^a
AUC	0.704	0.693	0.667	0.677	0.677	0.677	0.683
Threshold	0.752	0.461	0.461	0.414	0.414	0.414	0.486
Acc.	0.872	0.745	0.723	0.702	0.702	0.702	0.741
B. Acc.	0.748	0.715	0.677	0.665	0.665	0.665	0.689
Spec.	0.925	0.763	0.771	0.758	0.758	0.758	0.789
NPV	0.925	0.906	0.844	0.806	0.806	0.806	0.849
Prec.	0.571	0.400	0.467	0.500	0.500	0.500	0.490
Recall	0.571	0.667	0.583	0.571	0.571	0.571	0.589
F1-score	0.571	0.500	0.519	0.533	0.533	0.533	0.532

8.2 Statement of the problem

LDCT screening has enhanced early diagnosis of lung cancer, reducing lung cancer mortality among population at high risk [79, 288, 287]. Precise evaluation of risk in pulmonary nodules found in the screening process is crucial to improve patient survival and avoid unnecessary biopsy procedures. One of the challenges of LDCT screening is the managing of indeterminate risk (neither clearly benign, nor highly suspicious of malignancy) pulmonary nodules (IPNs). The ultimate goal is to differentiate malignant IPNs from a vast majority (at least 95%) of benign nodules, and identify malignant IPNs at an early stage. Precise evaluation of risk in pulmonary nodules is crucial to improve patient survival and avoid unnecessary biopsy procedures. In clinical practice, benign nodules are typically considered non-suspicious if they remain stable in size in subsequent follow-ups. Deep learning (DL) has shown potential to evaluate pulmonary nodule malignancy in LDCT imaging at the nodule level [3, 228, 296]. However, most of existing DL methods are trained on a single dataset and lack validation on external datasets, hindering their clinical translation [163, 4], and expanding the existing gap between AI research in lung cancer imaging and real clinical implementation.

Many state-of-the-art (SOTA) DL approaches in lung nodule assessment [3, 228] are trained on the publicly available LIDC-IDRI dataset [201], which includes nodule malignancy annotations (ranging from label 1, indicating low risk, to label 5, corresponding to high risk) provided by four independent radiologists. Since its collection in 2011, the criteria used in LIDC-IDRI have been updated by the Lung CT Screening Reporting and Data System (Lung-RADS), which was developed to standardize the reporting and management of screen-detected pulmonary nodules, and improve risk assessment [293]. In clinical practice (Figure 8.1), following Lung-RADS guidelines, nodules are broadly categorized into three stages: benign or indeterminate (requiring risk-dependent follow-up), and malignant (necessitating immediate intervention). Approaching this classification from an ordinal learning perspective reflects how radiologists assess risk and enables the integration of domain-specific knowledge through ordinal regularization [259, 151]. Differing from prior work [3, 228, 235, 296, 320], that ignored the ordinal progression of malignancy risk and relied on binary or multiclass classification, our approach formulates lung cancer risk as an ordinal problem, more in line with clinical reasoning. Notably, earlier methods excluded the indeterminate label (3) in LIDC-IDRI [3, 228, 235, 259], overlooking its significance as a transitional risk category. This approach was already proposed by Wu *et al.* [325], who leveraged *unsure data* in LIDC-IDRI. The problem was formulated as a three-class ordinal regression task, showing the benefits in performance with respect to binary classification. However, their architecture, based on a 3D DenseNet, does not incorporate privileged information, lacks comparison with other baselines, and is not externally validated. To contextualize the problem, Figure 8.1 summarizes the motivation behind this research.

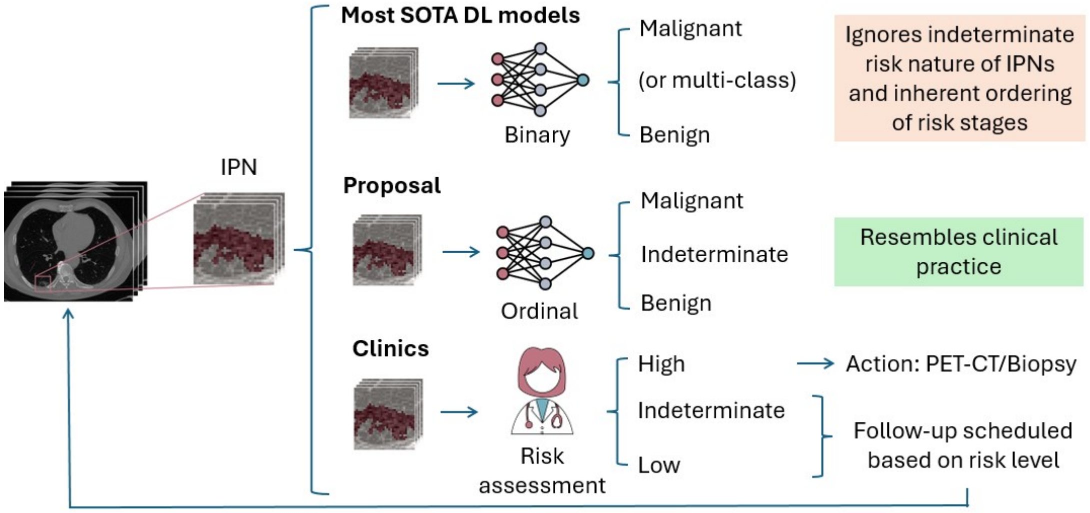


Figure 8.1: Illustration of how the proposed approach to manage indeterminate pulmonary nodules (IPNs) aligns with clinical practice, unlike current state-of-the-art methods.

DL architectures in medical imaging are typically based on convolutional neural networks (CNNs), e.g. DenseNet [27] and EfficientNet [28], or on attention mechanisms, as Vision Transformers (ViT) [35], discussed in Chapter 2. As an alternative, capsule networks (CapsNets, also described in Chapter 2) were introduced to overcome CNNs' limitations in modeling spatial hierarchies and part–whole relationships [6], and have shown promising results in various medical imaging applications, including lung cancer imaging. In LIDC-IDRI, recent approaches for binary lung nodule classification proposed 2D CapsNets that learned privileged information on nodule attributes [3, 228], annotated by the radiologists who labeled LIDC-IDRI dataset [201]. Afshar *et al.* [235] employed a 3D multi-scale CapsNet using patches centered on the nodule, comprising the central slice and four adjacent slices, without considering nodule attributes.

Developing and evaluating several models on the same dataset increases the likelihood that the best-performing model has simply overfitted to the test data (also known as *sequential overfitting*), rather than genuinely exhibiting better generalization [4]. Thus, conclusions drawn from benchmark results should be interpreted with caution, as minor performance improvements may not indicate truly superior models, and require confirmation through external validation. Shao *et al.* [303] evaluated the generalizability of SOTA models trained on LIDC-IDRI on a new lung CT image dataset with patological information (LIDP). They concluded that generalizability was very poor, due to significant differences in data distributions of both datasets, and lack of pathological confirmation in LIDC-IDRI compared to LIDP. However, several critical limitations were identified in this study: (1) the authors stated that the data originated from different distributions, but they did not thoroughly investigate how to better align radiological labels in LIDC-IDRI with pathological information in LIDP; (2) benign nodules confirmed with biopsy are biased to difficult samples and do not faithfully represent clinical practice; (3) label 3 in LIDC-IDRI was excluded, which represents the majority of IPNs and remains the primary unmet clinical need; (4) generalizability is constrained by the use of a single external dataset.

In this chapter, we validate medical informed DL ordinal approaches for lung nodule risk assessment on a novel well-curated screening cohort (P-ELCAP, described in Chapter 7). Beyond LIDC-IDRI [201], publicly available high-quality datasets that follow standard clinical guidelines for lung nodule management are scarce, and the availability of corresponding annotations is even more limited. Although LIDP [303] would be released after an embargo period, to the best of our knowledge, it is still not publicly accessible. Building on prior work [3, 228], we also propose a fully 3D CapsNet architecture that incorporates privileged information on nodule attributes. Our contributions can be summarized as (1) exploring how to leverage LIDC-IDRI radiological labels in an ordinal approach that aligns with clinical practice and targets unmet needs in lung nodule characterization; (2) benchmarking popular CNN architectures against ViTs and CapsNets to evaluate generalizability; (3) proposing a novel 3D CapsNet that achieves the highest generalizability.

8.3 Materials and methods

8.3.1 Data

LIDC-IDRI dataset

All models were trained on the LIDC-IDRI [201] dataset. To minimize label uncertainty, only nodules annotated by 3 to 4 radiologists with a standard deviation below 1 were considered. In the training set, we repeated the same nodule leveraging the different annotations available (with slightly different centroids) as a form of data augmentation, while in test and validation we took only the first annotation available. We considered the same labeling criteria as Wu *et al.* [325]: nodules with an average score above 3.5 were labeled as malignant (label 2), those below 2.5 were labeled as benign (label 0), and those within $[2.5, 3.5]$ were labeled as indeterminate (label 1), covering the full range of IPN classifications. Data was split into training (72%), validation (8%), and test (20%) using stratification based on the ordinal malignancy label, resulting in 3016 training, 91 validation and 223 test samples.

P-ELCAP dataset

All models were externally validated on P-ELCAP dataset, described in Chapter 7. We considered LDCT images from controls with benign nodules (68 cases), as well as lung cancer participants diagnosed in the same year as the LDCT acquisition (31 cases). Benign nodules classification is based on long-term follow-up and nodule image stability over time, while lung cancer is confirmed by pathological examination, following standard clinical practice. P-ELCAP is biased towards challenging-to-classify IPNs, reflecting the difficulty of distinguishing malignant IPNs from a majority of benign nodules in routine screening practice.

8.3.2 Method

Each model is trained with a total loss $\mathcal{L}_{\text{total}}$, consisting of various contributions, on the LIDC-IDRI dataset. A common prediction loss $\mathcal{L}_{\text{pred}}$ of nodule malignancy risk, representing the ordinal beta cross entropy (β_{CE}) loss from Vargas *et al.* [326], is optimized in all models. CapsNets include an additional attribute loss $\mathcal{L}_{\text{attr}}$. For the 2D CapsNet specifically, a reconstruction loss $\mathcal{L}_{\text{recon}}$ is incorporated, as in the original work [3]. The total loss is formally defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{attr}} \cdot \mathcal{L}_{\text{attr}} + \lambda_{\text{recon}} \cdot \mathcal{L}_{\text{recon}} \tag{8.3}$$

The nodule malignancy prediction loss $\mathcal{L}_{\text{pred}}$ is defined using the β_{CE} loss [326]:

$$\begin{aligned}\mathcal{L}_{\text{pred}} &= \beta_{\text{CE}}(y, \hat{y};\ \eta = 0.5,\ J = 3) \\ &= \sum_{i=1}^{J} q'(i) \cdot [-\log p(y = C_i \mid x)]\end{aligned} \tag{8.4}$$

where $J = 3$ is the number of ordinal classes, and $p(y = C_i \mid x)$ denotes the predicted probability for class C_i. To address class imbalance, a weighting factor based on the inverse frequency of each class was incorporated into the loss. The soft target distribution $q'(i)$ is defined as:

$$q'(i) = (1 - \eta)\delta_{i,y} + \eta f(x, a, b) \tag{8.5}$$

where $f(x, a, b)$ is the probability value sampled from a beta distribution centered in $x = \frac{2J-1}{2J}$, the parameters a and b are calculated as described by Vargas *et al.* [326], and $\delta_{i,y}$ is the Dirac delta function, which equals 1 if $i = y$ and 0 otherwise, and corresponds to the one-hot encoding of the ground-truth label. The parameter $\eta \in [0, 1]$ controls the linear combination between the one-hot ground-truth label and a unimodal prior derived from the beta distribution used to smooth the original one-hot label into a soft label [326]. In the experiments, we set $\eta = 0.5$ empirically. The attribute loss $\mathcal{L}_{\text{attr}}$ is only used in the 2D and 3D CapsNet models, with $\lambda_{\text{attr}} = 3$. This loss is defined as [3]:

$$\mathcal{L}_{\text{attr}} = \sum_{k=1}^{K} \text{MSE}\left(a_k^{\text{pred}}[i], a_k^{\text{gt}}[i]\right), \quad \forall i \in \mathcal{I}_{\text{attr}} \tag{8.6}$$

where $K = 8$ is the number of attributes, $a_k^{\text{pred}}[i]$ and $a_k^{\text{gt}}[i]$ denote the predicted and ground-truth values (mean attribute score by the radiologists), respectively, for the k-th attribute of sample i, and $\mathcal{I}_{\text{attr}}$ is the set of indices corresponding to the samples for which attribute annotations are available. The mean squared error (MSE) is computed independently for each attribute across the annotated samples. The reconstruction loss $\mathcal{L}_{\text{recon}}$ is only used in the 2D CapsNet to predict the segmentation mask of the nodule, with $\lambda_{\text{recon}} = 1$. This loss is defined as [3]:

$$\mathcal{L}_{\text{recon}} = \text{MSE}(x^{\text{recon}}, x) \tag{8.7}$$

where x is the segmentation mask and x^{recon} is the output produced by the decoder branch of the 2D CapsNet.

Models were trained for a maximum of 100 epochs, with early stopping based on validation accuracy and a patience of 20 epochs. Subsequently, independent external validation was performed on P-ELCAP dataset. All models were coded in Pytorch 2.5.1 [268] and trained on NVIDIA A30 GPU (24 GB VRAM), with a batch size of 256, and AdamW optimizer [269].

Baselines

Baseline models included two CNN-based architectures, DenseNet121 [27] and EfficientNet-B0 [28], as well as a Transformer architecture, ViT [35], in both 2D and 3D. We considered the CapsNet-2D architecture proposed by Gallée *et al.* [3], using the variant without prototypes, as their ablation study showed no significant performance difference, and this choice reduced complexity. Additional minor adjustments are described in the following section. Gallée *et al.* [3], which builds upon LaLonde *et al.* [228], attempts to predict the distribution of radiologists' scores, excluding nodules with a mean score of 3, and considering predictions within ±1 of the reference label (within-1 accuracy) as correct. However, this approach may lack sufficient discriminative power, as most nodules in LIDC-IDRI have a score between 2-4, and the model predominantly predicts the peak of the distribution in label 3.

Proposed method

We designed a 3D capsule network, as shown in Figure 8.2. We built upon the 2D implementation of Gallée *et al.* [3]. Experimental results justified reducing the number of feature maps in the convolutional layers from 256 to 32, and the routing iterations from 3 to 2, without compromising performance. The effect of different kernel sizes on performance was experimentally studied, and the best trade-off between performance and the number of capsules in the dense capsule layer was achieved with a kernel size of 9.

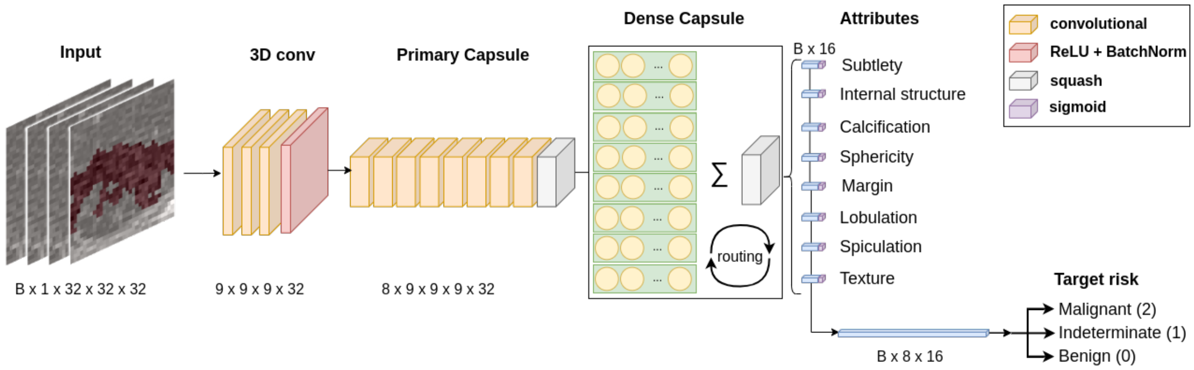


Figure 8.2: Capsule network architecture in 3D.

Preprocessing

A crop of size [32, 32, 32] pixels centered on the nodule's centroid was used to capture contextual information in 3D. The minority of nodules larger than this size in one or more dimensions were down-sampled in the corresponding axes to ensure compatibility with the model's input size. In 2D, only slices within this cube that contained the annotated nodule were retained. In contrast, Afshar *et al.* [235] used only one central and four neighboring slices, and Gallée *et al.* [3] performed tight 2D crops based on segmentation masks that were resized to the input size, limiting both contextual information.

Evaluation metrics

Previous work on LIDC-IDRI reported a single metric within-1-accuracy [3, 228]. However, this metric alone fails to capture overall model performance and is insufficient for a multi-class classification problem with an inherent ordinal nature [327]. Instead, we used three-class accuracy (Acc.), balanced accuracy (B. Acc.), mean absolute error (MAE), root mean squared error (RMSE), Uniform Ordinal Classification index (A_{UOC}) [270], and quadratic weighted Cohen's Kappa score (QWK) [271, 272] to evaluate on LIDC-IDRI test subset. In the external validation on P-ELCAP dataset, malignant nodules diagnosed in the same year as the LDCT acquisition were assigned the maximum risk label (2), while benign nodules were assigned the minimum risk label (0). We report the number of nodules predicted as indeterminate (label 1), along with how many of these correspond to true benign and true malignant cases. Binary classification metrics including Acc., B. Acc., sensitivity (Sens.), precision (Prec.), specificity (Spec.), negative predictive value (NPV) and F1-score (F1-sc.) were calculated excluding indeterminate predictions. For all the models, 95% confidence intervals estimated by bootstrapping were calculated generating 1000 bootstrap samples.

8.4 Results

8.4.1 Internal validation on LIDC-IDRI

The models were trained on LIDC-IDRI including 3016 samples in the training subset, 91 in validation, and 223 in test. Table 8.4 shows the performance of the models in LIDC-IDRI test subset.

Table 8.4: Performance on internal LIDC-IDRI test subset with ordinal approach. Total true indeterminates are 53.4% (119 cases). The aggregation method (Agg.) used in 2D models to compute the global score is indicated (Avg. = average, Max. = maximum). Subscripts indicate the maximum half-width of the 95% confidence interval centered at the mean. The best and second-best performances are denoted by **bold** and underlined, respectively. Abbreviations: DN = DenseNet121, EN = EfficientNet-B0, CN = CapsNet. To ensure correct implementation of CN-2D, the model was first reproduced as in the original work by Gallée *et al.* [3], achieving compatible within-1-accuracy in 2D (0.945).

Model	**DN-2D**	**EN-2D**	**ViT-2D**	**CN-2D**	**DN-3D**	**EN-3D**	**ViT-3D**	**CN-3D**
Agg.	Avg.	Avg.	Avg.	Max.	–	–	–	–
Acc.	$\mathbf{0.66}_{0.06}$	$0.58_{0.07}$	$\underline{0.64}_{0.07}$	$0.52_{0.07}$	$0.50_{0.07}$	$0.53_{0.07}$	$0.58_{0.06}$	$0.47_{0.07}$
B. Acc.	$\mathbf{0.68}_{0.07}$	$0.62_{0.07}$	$\underline{0.65}_{0.07}$	$\underline{0.65}_{0.05}$	$0.55_{0.07}$	$0.56_{0.06}$	$0.59_{0.07}$	$0.63_{0.05}$
MAE	$\mathbf{0.38}_{0.07}$	$\underline{0.43}_{0.07}$	$\mathbf{0.38}_{0.07}$	$0.53_{0.08}$	$0.55_{0.08}$	$0.50_{0.07}$	$0.44_{0.07}$	$0.54_{0.07}$
RMSE	$\underline{0.68}_{0.09}$	$\underline{0.68}_{0.07}$	$\mathbf{0.64}_{0.07}$	$0.79_{0.08}$	$0.81_{0.08}$	$0.75_{0.07}$	$0.69_{0.06}$	$0.76_{0.06}$
$\boldsymbol{A_{UOC}}$	$0.46_{0.07}$	$0.50_{0.06}$	$0.48_{0.08}$	$0.48_{0.06}$	$\mathbf{0.58}_{0.06}$	$\underline{0.55}_{0.06}$	$0.52_{0.06}$	$0.47_{0.06}$
QWK	$0.52_{0.12}$	$0.52_{0.11}$	$\mathbf{0.55}_{0.11}$	$0.50_{0.10}$	$0.39_{0.11}$	$0.42_{0.10}$	$0.48_{0.10}$	$\underline{0.54}_{0.09}$
% Pred indet.	51.1	48.4	51.1	19.3	39.5	52.9	53.4	12.6

Table 8.5 shows the performance of the models in LIDC-IDRI test subset excluding true indeterminates (ground truth label 1).

8.4.2 External validation on P-ELCAP

Table 8.6 shows generalizability to P-ELCAP dataset considering only benign and lung cancer nodules diagnosed in the same year as the LDCT acquisition.

Table 8.5: Performance on internal LIDC-IDRI test subset with ordinal approach, restricted to benign and lung cancer ground truth labels. Subscripts indicate the maximum half-width of the 95% confidence interval centered at the mean estimated using 1000 bootstrap samples. The best and second-best performances are denoted by **bold** and underlined, respectively. Abbreviations: DN = DenseNet121, EN = EfficientNet-B0, CN = CapsNet.

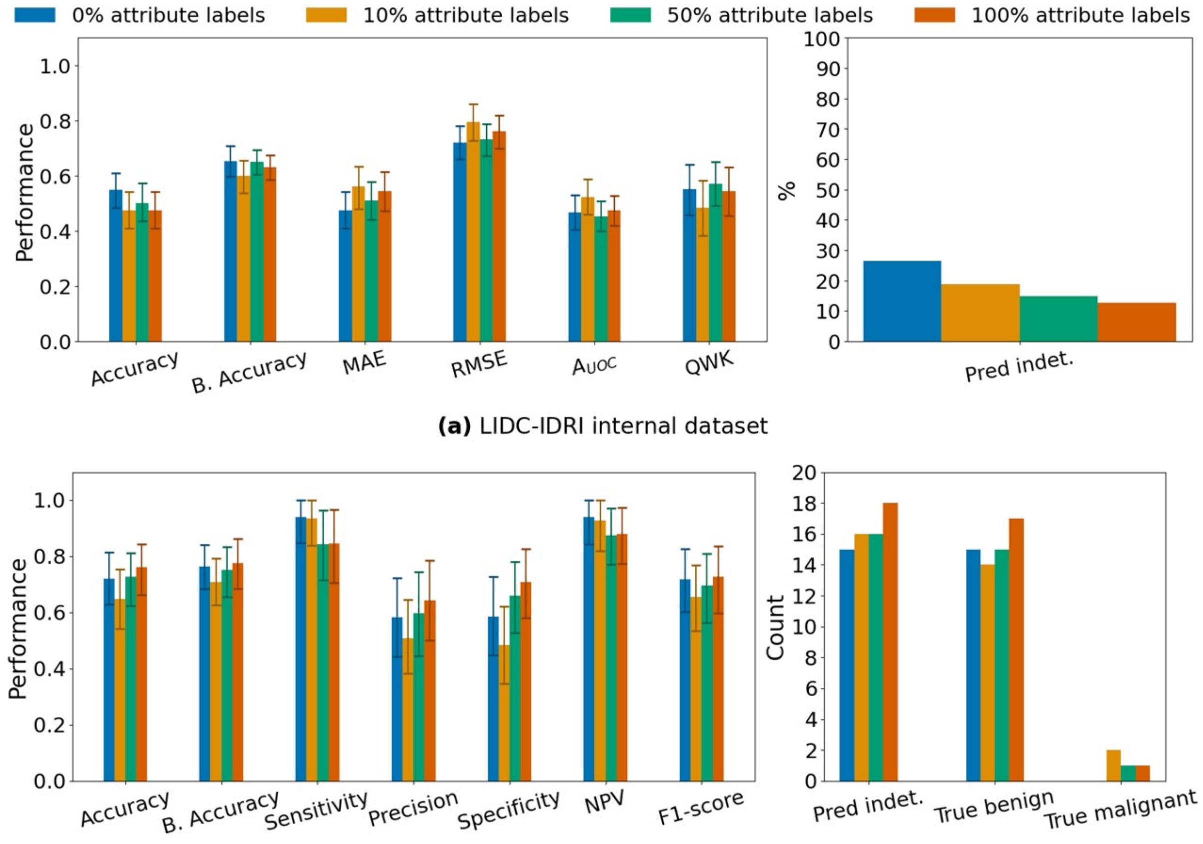
Model	**DN-2D**	**EN-2D**	**ViT-2D**	**CN-2D**	**DN-3D**	**EN-3D**	**ViT-3D**	**CN-3D**
Acc.	$0.88_{0.08}$	$\underline{0.94}_{0.08}$	$\underline{0.94}_{0.06}$	$0.88_{0.07}$	$0.84_{0.09}$	$0.88_{0.09}$	$0.93_{0.07}$	$\mathbf{0.96}_{0.04}$
B. Acc.	$0.88_{0.07}$	$\underline{0.94}_{0.06}$	$\underline{0.94}_{0.06}$	$0.90_{0.05}$	$0.84_{0.09}$	$0.84_{0.10}$	$0.92_{0.08}$	$\mathbf{0.96}_{0.05}$
Sens.	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$0.94_{0.09}$	$\mathbf{1.00}_{0.00}$	$0.89_{0.11}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\underline{0.97}_{0.06}$
Prec.	$0.80_{0.13}$	$0.90_{0.12}$	$\mathbf{0.94}_{0.09}$	$0.78_{0.12}$	$0.83_{0.13}$	$0.83_{0.12}$	$0.89_{0.11}$	$\underline{0.93}_{0.10}$
Spec.	$0.76_{0.14}$	$0.87_{0.13}$	$\mathbf{0.95}_{0.08}$	$0.80_{0.11}$	$0.78_{0.16}$	$0.69_{0.19}$	$0.83_{0.16}$	$\underline{0.94}_{0.07}$
NPV	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$0.95_{0.08}$	$\mathbf{1.00}_{0.00}$	$0.86_{0.13}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\underline{0.98}_{0.04}$
F1-sc.	$0.88_{0.08}$	$\underline{0.94}_{0.07}$	$\underline{0.94}_{0.07}$	$0.87_{0.08}$	$0.86_{0.10}$	$0.91_{0.07}$	$\underline{0.94}_{0.06}$	$\mathbf{0.95}_{0.06}$
# Pred indet.	31	39	36	10	35	48	49	11
# True benign	26	33	28	9	32	42	41	10
# True malig.	5	6	8	**1**	$\underline{3}$	6	8	**1**

Table 8.6: Performance on external P-ELCAP validation, restricted to benign and lung cancer ground truth labels. Subscripts indicate the maximum half-width of the 95% confidence interval centered at the mean estimated using 1000 bootstrap samples. The best and second-best performances are denoted by **bold** and underlined, respectively. Abbreviations: DN = DenseNet121, EN = EfficientNet-B0, CN = CapsNet. [a]NPV could not be computed due to absence of predicted negatives in some resamples.

Model	**DN-2D**	**EN-2D**	**ViT-2D**	**CN-2D**	**DN-3D**	**EN-3D**	**ViT-3D**	**CN-3D**
Acc.	$0.54_{0.14}$	$0.57_{0.14}$	$\mathbf{0.84}_{0.14}$	$0.59_{0.12}$	$0.65_{0.12}$	$0.67_{0.12}$	$0.57_{0.13}$	$\underline{0.76}_{0.10}$
B. Acc.	$0.59_{0.07}$	$0.63_{0.08}$	$\mathbf{0.85}_{0.11}$	$0.68_{0.07}$	$0.65_{0.09}$	$0.68_{0.10}$	$0.55_{0.10}$	$\underline{0.78}_{0.10}$
Sens.	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\underline{0.97}_{0.08}$	$0.96_{0.08}$	$0.92_{0.11}$	$0.85_{0.14}$
Prec.	$0.50_{0.15}$	$0.50_{0.16}$	$\mathbf{0.75}_{0.19}$	$0.47_{0.13}$	$0.59_{0.14}$	$0.61_{0.14}$	$0.56_{0.14}$	$\underline{0.64}_{0.15}$
Spec.	$0.17_{0.15}$	$0.25_{0.17}$	$\underline{0.69}_{0.21}$	$0.36_{0.14}$	$0.33_{0.16}$	$0.39_{0.18}$	$0.17_{0.18}$	$\mathbf{0.71}_{0.13}$
NPV	–[a]	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.00}_{0.00}$	$\underline{0.91}_{0.20}$	$\underline{0.91}_{0.20}$	–[a]	$0.88_{0.11}$
F1-sc.	$0.66_{0.14}$	$0.66_{0.15}$	$\mathbf{0.85}_{0.13}$	$0.64_{0.13}$	$0.73_{0.12}$	$\underline{0.74}_{0.11}$	$0.70_{0.13}$	$0.73_{0.13}$
# Pred indet.	46	50	62	23	39	45	51	18
# True benign	39	40	49	20	37	40	45	17
# True malig.	7	10	13	3	$\underline{2}$	5	6	**1**

8.4.3 Ablation study

An ablation analysis is shown in Figure 8.3 to evaluate the contribution of the attribute information in the proposed 3D CapsNet model.



(a) LIDC-IDRI internal dataset

(b) P-ELCAP external dataset

Figure 8.3: Ablation study of the CN-3D model when attribute information was only available as fractions of the training dataset.

8.4.4 Dimensionality reduction, attributes and interpretability analysis

Uniform Manifold Approximation and Projection (UMAP) was applied to visualize the concatenated feature representation from the output capsules used in the target risk computation both in LIDC-IDRI and P-ELCAP datasets. Figure 8.4 a) shows the UMAP embedding computed on the LIDC-IDRI training data and applied to the test data. The number of UMAP components was set to three, corresponding to the three expected diagnostic groups. However, the resulting projection reveals that indeterminate and benign cases are intertwined, while malignant cases appear more distinctly separated. In addition, violin plots are presented in Figure 8.4 b) for the two clusters identified by the k-means algorithm [328] in the test data, based on the UMAP projection. The same UMAP embedding, applied to P-ELCAP data, is shown in 8.5. Participants diagnosed with lung cancer are included from years 1, 2, and 3 prior to diagnosis (pre-LC), along with the corresponding Lung-RADS categories assigned to all detected nodules. Lung-RADS annotations were initially performed by an experienced radiology technician, and subsequently reviewed and approved by a board-certified thoracic radiologist, who is a member of the P-ELCAP LDCT multidisciplinary team.

The attributes in Figure 8.4 b) show that spiculation and lobulation are the most discriminative attributes for the identification of malignant nodules. Both radiological signs are indicators of

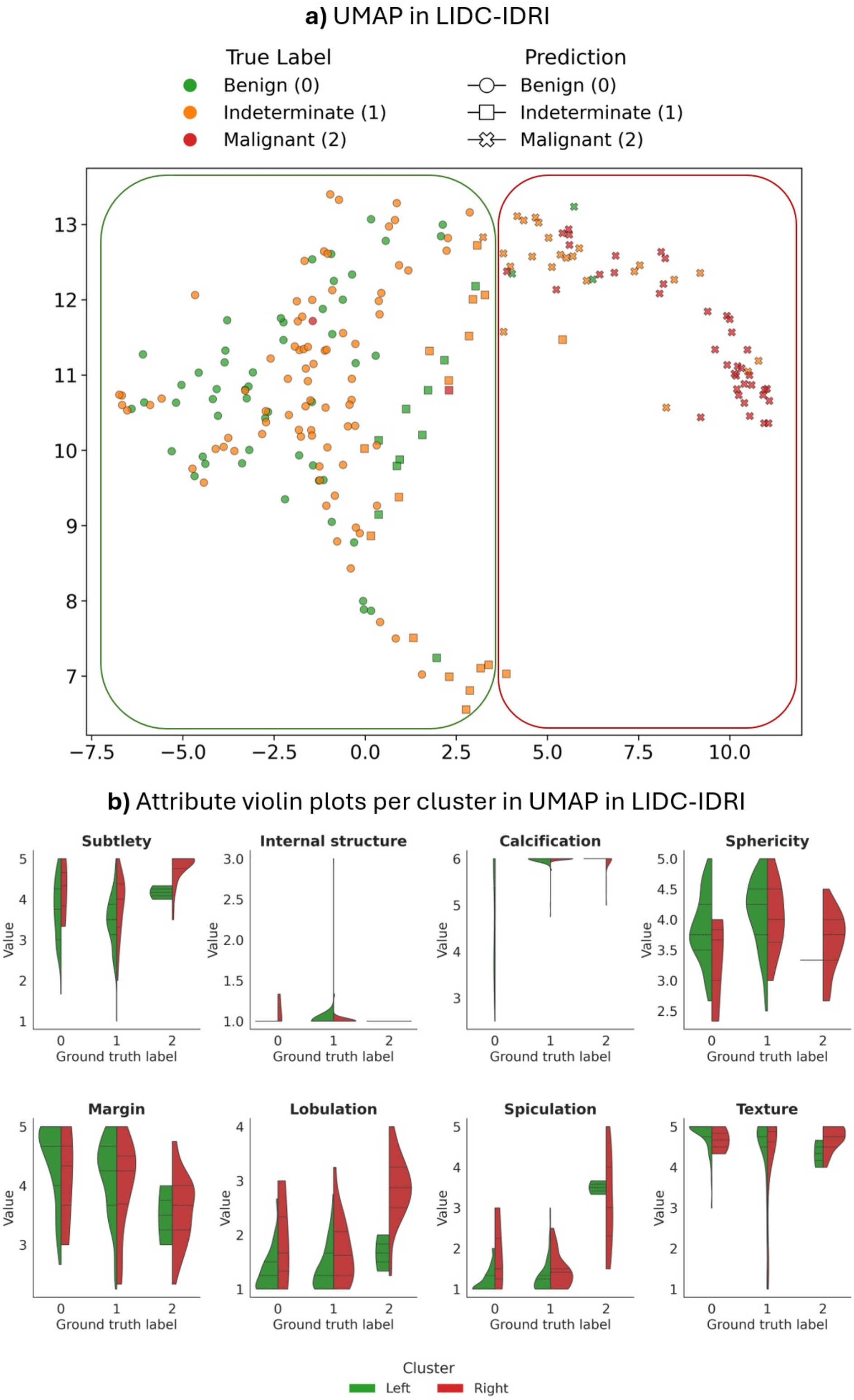


Figure 8.4: UMAP visualization of **a)** LIDC-IDRI dataset on the test subset, and **b)** corresponding attributes per cluster and ground truth label in LIDC-IDRI.

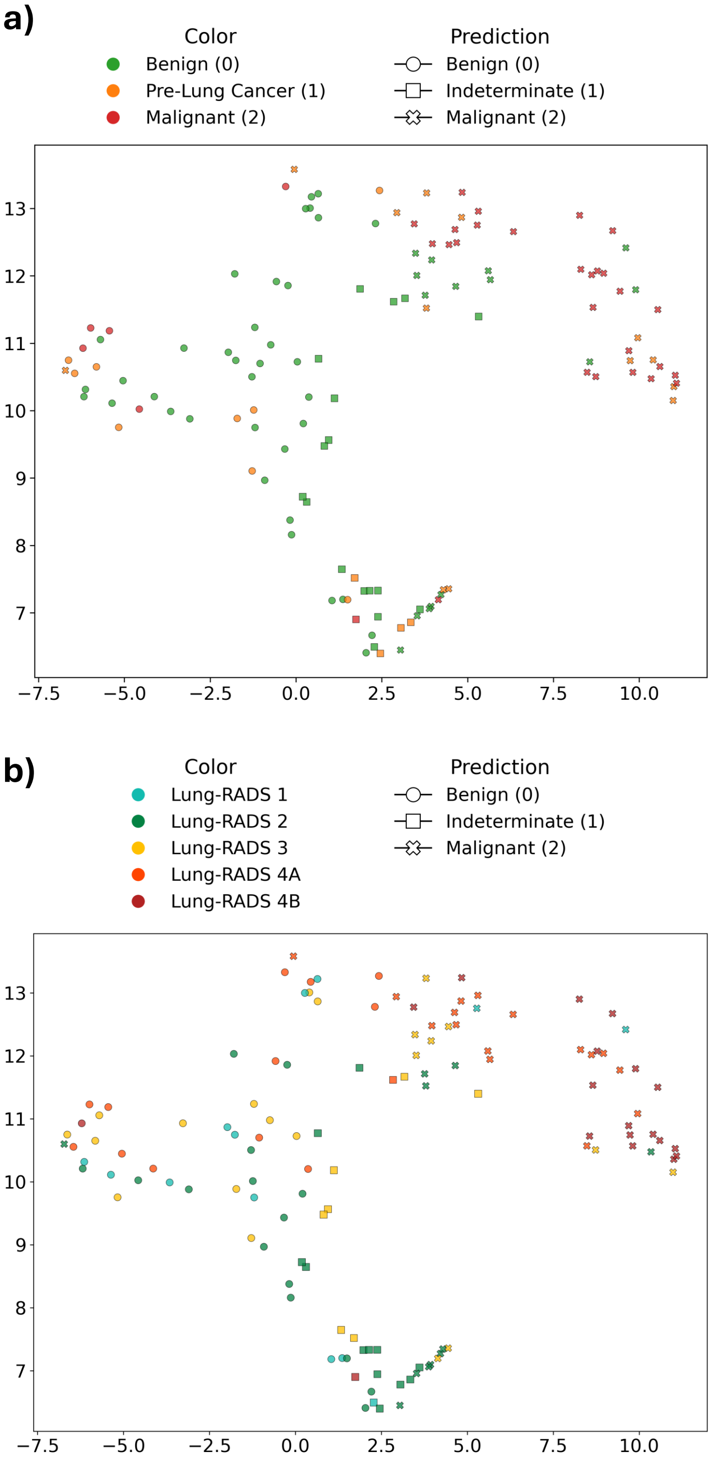


Figure 8.5: UMAP visualization of P-ELCAP **a)** with true labels, and **b)** with LungRADS categories.

malignancy [329, 330]. Interestingly, Figure 8.5 shows a similar distribution of benign and malignant IPNs in the external validation dataset P-ELCAP, supporting the generalizability of the 3D CapsNet model to our screening cohort.

8.4.5 Evaluation on pre-lung cancer and false positive nodules

Table 8.7 shows predictions in P-ELCAP dataset for lung nodules belonging to pre-LC cases at years 1, 2 and 3 before diagnosis, as well as to the imaging false positives (FP) group.

Table 8.7: CapsNet-3D predictions in pre-lung cancer (Pre-LC) nodules at years 1, 2 and 3 previous to diagnosis, and in imaging false positive nodules (FP).

Group	**Pre-LC year 1**	**Pre-LC year 2**	**Pre-LC year 3**	**FP**
# Pred benign	4	2	3	0
# Pred indet.	0	3	1	1
# Pred malig.	5	3	5	4

8.5 Discussion and conclusion

An evaluation of the generalizability of SOTA DL models in lung nodule risk assessment resembling clinical practice with a three class ordinal approach was performed. The findings indicate that high performance on internal test data may not consistently translate to the external dataset. The best performance on the external dataset is achieved by the proposed 3D CapsNet. This architecture yields the highest specificity and the lowest number of indeterminate predictions, with only one corresponding to a true malignant, as presented in Table 8.6. This high performance, along with the lowest number of indeterminate predictions, is also evident in the internal evaluation on LIDC-IDRI that excludes indeterminate cases, reported in Table 8.5. In lung cancer screening, high specificity is essential to minimize patient anxiety and avoid unnecessary procedures. While the 2D ViT also achieves high performance in binary classification, it predicts a notably larger number of indeterminates compared to the 3D CapsNet. The remaining models exhibit very low specificity in external validation, limiting their suitability for clinical translation. Both Transformers and CapsNets were designed to address CNNs' limitations in preserving spatial relationships [6], as discussed in Chapter 2. Our results support the superior generalization of CapsNets, which explicitly model viewpoint invariance within the network's weights [6]. Although Transformers integrate both local and global context, their performance is limited when trained on insufficient data, due to the absence of inductive biases inherent to CNNs. This limitation is particularly evident in the 3D ViT model, which is trained on fewer samples compared to its 2D counterpart.

The 2D CapsNet in previous work [3, 228] ignores the 3D nature of nodules, and does not subsequently aggregate per-slice scores into a global prediction for the entire nodule. Additionally, if there are middle slides where the nodule is not visible and not segmented, the 2D approach loses this spatial context, which is relevant for the analysis of the nodule. In contrast with the 2D architecture [3, 228], the reconstruction module was not incorporated in the 3D version, as external validation in 2D showed suboptimal reconstruction quality based on the Dice similarity coefficient (DSC) [331, 332] (DSC: 0.66 in LIDC-IDRI, 0.55 in P-ELCAP).

We further demonstrated the feasibility of aligning LIDC-IDRI radiological labels with a clinical dataset, highlighting their utility despite the lack of pathological confirmation and addressing previous limitations [303]. This alignment is essential for adapting publicly available datasets to smaller clinical cohorts, where expert annotations are costly and scarce. Our work explores the problem of *sequential overfitting* [4], demonstrating the importance of understanding clinical practice and differences in dataset's distributions to effectively solve unmet clinical needs. Future work could focus on validating the properties and interpretability of the learned attributes in the external dataset, particularly with a radiologist. Further investigation on the 3D CapsNet's explainability and generalizability to additional datasets constitutes another direction of interest.

Prospect of Application: The proposed framework holds significant potential for enhancing IPN's characterization in clinical settings, ultimately contributing to earlier lung cancer diagnosis and providing clinically meaningful feedback to radiologists. Additionally, this study emphasizes the often underestimated importance of addressing clinical needs and performing validation in screening cohorts.

8.6 Future work

The beginning of this chapter introduced Sybil [301], a DL algorithm that analyzes single volumetric LDCT scans to predict lung cancer occurring 1-6 years after the screening scan. This algorithm was assessed on P-ELCAP cohort, raising major concerns about the clinical applicability of the originally reported evaluation metrics and the generalizability of the algorithm itself. These shortcomings motivated the development and validation of the medical informed ordinal framework for lung nodule assessment described in the remainder of this chapter. Additional external validation will be performed in the Liverpool Lung cancer Project screening cohort.

Although Sybil [301] holds potential for enhancing early lung cancer diagnosis, it does not sufficiently address clinical requirements. The approach is innovative but underestimates the importance of considering standard nodule management practices in routine screening, which are essential for clinical translation. To address these limitations in future work, a project proposal was submitted to request access to the NLST dataset [299], used in Sybil's development, and access was later granted through the Cancer Data Access System. This collection contains approximately 28 000 LDCT images corresponding to 3700 participants. Future work may explore retraining Sybil with a methodology that explicitly accounts for key aspects relevant to clinical practice.

As a long-term objective, DL image-based early diagnosis of lung cancer is envisioned as a multi-scale approach that integrates local nodule assessment with global evaluation of the entire LDCT scan. The combination of both levels of abstraction would enable the model to capture both fine-grained discriminative features and high-level contextual information, including the nodule's location within the LDCT scan, lung parenchymal characteristics, and comorbid respiratory conditions related to lung cancer, such as emphysema [333]. Ultimately, this multi-scale strategy seeks to reflect the radiological diagnostic process, integrating global scan-level assessment with detailed analysis of the nodules at a local scale. The clinical foundations underpinning this goal are established by the ordinal lung nodule assessment framework described in this chapter. Furthermore, the proposed medical informed approach could be integrated into a DL pipeline that automatically segments nodules detected in LDCT scans and provides malignancy assessments to assist radiologists in lung cancer diagnosis and clinical decision-making. Numerous DL-based algorithms for pulmonary nodule detection and segmentation have been proposed in the literature. For instance, Gao *et al.* [334] recently performed a comprehensive systematic review on this topic. Integrating the work in this chapter in an end-to-end pipeline for lung nodule detection, segmentation and assessment holds significant potential for future research.

Complementary to image-based DL assessment of LDCT images, the integration of molecular biomarkers in a multimodal approach to enhance lung cancer screening is a promising direction of research. This methodology is explored in the next chapter of this thesis, analyzing the combination of plasma protein biomarkers with the image embedding extracted by the 3D CapsNet model proposed in this chapter.

Chapter 9

Multimodal early lung cancer diagnosis integrating imaging and proteomics

9.1 Introduction

Early detection of lung cancer is critical for improving patient outcomes, as survival rates significantly decline once the disease progresses to advanced stages. Diagnosis of lung cancer often occurs at a late stage, after symptoms become evident. Consequently, treatment options are limited and survival rates are substantially lower. Low dose computed tomography (LDCT) has been developed in the last decades as a successful tool for lung cancer screening in high-risk individuals [79, 287, 288]. Several risk models have been proposed to support inclusion criteria for lung cancer screening programs. High-risk individuals participating in lung cancer screening undergo annual LDCT, which commonly reveals the presence of lung nodules, most of which present the typical features of a benign nodule. Nodules that are more suspicious of malignancy are monitored more closely for growth, typically through a follow-up LDCT after three months, or they are biopsied to confirm malignancy, which usually leads to surgical resection of the nodule. Moreover, a proportion of LDCT-detected nodules remain indeterminate (neither clearly benign, nor malignant). There are models based on clinical and epidemiological data, as will be explained subsequently, which provide some additional information to *rule in* or *rule out* the indeterminate nodules from the malignant classification. However, advanced image analysis, initially through hand-crafted radiomics and, more recently, through DL techniques, has emerged as a powerful strategy for characterizing indeterminate nodules identified during screening. Beyond imaging, other factors may further improve the risk assessment capabilities of AI models, particularly through multimodal approaches. By integrating diverse data sources (imaging, clinical data, circulating protein biomarkers) within a unified AI model, multimodal AI [335] has the potential to advance clinical decision-making and improve diagnostic outcomes in lung cancer screening. This chapter introduces a multimodal DL approach incorporating image-based nodule assessment and protein biomarkers to enhance early lung cancer detection.

As aforementioned, LDCT has emerged as the standard screening method for high-risk populations, including heavy smokers and individuals with a family history of lung cancer. LDCT has been shown to reduce lung cancer mortality by enabling the detection of early-stage tumors [79], as discussed in Chapter 7. The current United States Preventive Service Task Force (USPSTF) guidelines for LDCT screening recommend screening individuals aged 50 to 80 years with a smoking history of more than 20 pack-years. However, epidemiological studies have shown that over 40% of patients diagnosed with lung cancer fall outside these criteria, based on age and smoking exposure. For instance, in Spain, more than one-third of lung cancer patients do not meet the USPSTF eligibility requirements for inclusion in LDCT screening programs [336]. These findings highlight an important unmet clinical need: the refinement of risk assessment criteria to enable more personalized and inclusive screening strategies.

Moreover, a second critical clinical challenge lies in managing pulmonary nodules of indeterminate malignancy. In the context of screening, pulmonary nodules with indeterminate (neither low nor high) malignancy risk are observed in approximately 20-40% of scans [291], posing a significant challenge for clinical decision-making. Risk management of indeterminate pulmonary nodules (IPNs) found during lung cancer screening relies on estimating the probability of malignancy. Several risk assessment scores based on the probability of malignancy upon the detection of a lung nodule have been developed. The Brock model is a risk prediction tool that estimates the probability of malignancy in pulmonary nodules through the calculation of a quantitative risk score which combines several variables such as nodule size, location, and patient characteristics (e.g., smoking history, age, family history of lung cancer) to provide a quantitative risk score [337]. The Lung Imaging Reporting and Data System (Lung-RADS), developed by the American College of Radiology, assigns categories to lung nodules based on their size and appearance, with each category associated with a corresponding risk level that provides clinical management recommendations for patients (as discussed in Chapter 8). The Mayo Lung Nodule Model integrates LDCT scan-based nodule characteristics such as location, diameter, and spiculation, with clinical variables such as age and smoking history, and, when available, Positron Emission Tomography (PET) imaging [338]. However, these risk score models alone do not reach optimal levels of specificity and sensitivity, particularly in cases where nodules exhibit atypical features.

Multimodal DL approaches that combine LDCT imaging with clinical data and blood-based biomarkers represent a promising strategy for improving lung cancer diagnosis. Building on this premise, Gao *et al.* [339] proposed a multimodal framework to estimate the malignancy risk of IPNs integrating both imaging and clinical information. The clinical variables included age, body mass index, smoking status, personal cancer history, pack-years of smoking, nodule size, spiculation, and location, along with one blood-based circulating protein biomarker, hs-CYFRA 21–1. The method employed an image-based module to detect and segment the top five most suspicious nodules in LDCT scans. Features were then extracted from each nodule's bounding box and concatenated with clinical variables processed through fully connected layers. This combined representation was used to generate a malignancy risk score for each patient. The proposed approach outperformed existing risk prediction tools (Mayo and Brock models) and an image-only baseline. Thus, the integration of additional biological information has the potential to further improve early lung cancer diagnosis by capturing complementary aspects of tumor behavior.

Protein biomarkers present in blood offer a minimally invasive and promising approach for improving current practices in the early detection and diagnosis of lung cancer. Circulating proteins can be used as a surrogate indicator of underlying pathological processes occurring during lung cancer development and progression. Thus, the altered levels or the presence of specific tumor-associated proteins can serve as early *warning* signals, potentially even before clinical symptoms manifest or LDCT imaging can detect lung cancer. The identification of individuals at high risk or those with early-stage disease through blood tests could significantly improve patient outcome by enabling timely intervention and personalized treatment strategies [340]. In addition, the use of panels combining several protein biomarkers constitutes a robust approach to enhance diagnostic accuracy, sensitivity and specificity [341]. This strategy offers complementary information to refine current LDCT image-based screening procedures.

From a biological standpoint, the relevance of blood protein biomarkers lies in their direct connection to the cancer development and the body response to it. As tumors grow, they shed proteins, genetic content and exosomes into the bloodstream. Some of these blood-based biomarkers include proteins involved in critical cellular processes that become dysregulated in cancer [313]. Importantly, these circulating proteins often play an active role in tumor development, metastasis and modulation of the microenvironment, including interactions with the patient immune system; thereby offering areas for research of exploitable vulnerabilities for the development of targeted therapies. Therefore, the detection and quantification of these blood-based biomarkers constitutes a direct insight into the molecular landscape of lung cancer. This information is critical not only for early diagnosis, but also for clinical management and the development of personalized medicine strategies. Ultimately, integrating blood protein markers with nodule features evaluated in LDCT scans and clinical information represents a promising strategy to differentiate malignant from benign IPNs in early stages of lung cancer.

This chapter presents a proof-of-concept multimodal AI joint fusion model that integrates image-based pulmonary nodule assessment with a selected panel of circulating protein biomarkers relevant to lung cancer diagnosis.

9.2 Materials and methods

9.2.1 Dataset and experimental design

The multimodal approach was trained and evaluated on P-ELCAP dataset, described in Chapter 7. We considered LDCT images from controls with benign nodules (68 cases), as well as lung cancer participants diagnosed in the same year as the LDCT acquisition (31 cases). This procedure enabled the study of circulating protein markers relevant for imminent lung cancer diagnoses within one year.

Data was divided into 3 fold cross-validation (CV) train and test splits, including a validation subset representing 10% of training data, following the same data splits described in Section 7.4.2. Performance was evaluated using binary classification metrics: area under the curve (AUC), accuracy, balanced accuracy, precision (i.e., positive predictive value, PPV), recall (i.e., sensitivity), F1-score, specificity and negative predictive value (NPV). The threshold was calculated for a fixed sensitivity of 0.8, 0.9, and Younden's index, in accordance with prior work in the field [342].

9.2.2 Method

The multimodal DL model incorporated image embeddings extracted from the nodule's patch using the medical informed DL model described in Chapter 8. The protein embeddings were derived from 33 biomarkers normalized using a min–max scaler fitted on the training subset. This scaling approach was preferred over standard scaling (used in Chapter 7) to align the range of protein values with that of the image embeddings, ensuring a more balanced contribution from both modalities during training. Among the 33 circulating protein markers, 29 belonged to the set of 36 biomarkers previously identified as potentially informative for distinguishing malignant from benign nodules, reported by Khodayari *et al.* [2]. This prior work is part of the INTEGRAL consortium [312], comprising four lung cancer screening cohorts, including P-ELCAP. Seven proteins (KPNA1, PLXDC1, RASA1, NOS3, CXCL17, LY9, ALDH3A1) from the original list of 36 were excluded due to quality control (QC) issues in P-ELCAP's Olink assay, including QC errors and expression levels falling below the detection threshold. Furthermore, four additional proteins (AP-N, PVRL-4, TCL1A, EGFR) were incorporated due to their predictive importance, as identified by the protein models presented in Table 7.5.

The model was designed following a joint fusion approach, as depicted in Figure 9.1. The image embeddings were extracted by the ordinal 3D CapsNet model described in Chapter 8. Subsequently, the image embeddings were concatenated with the scaled protein markers and processed through a series of fully connected (FC) layers. During training, only the FC layers were updated, while the weights of the 3D CapsNet model were kept frozen.

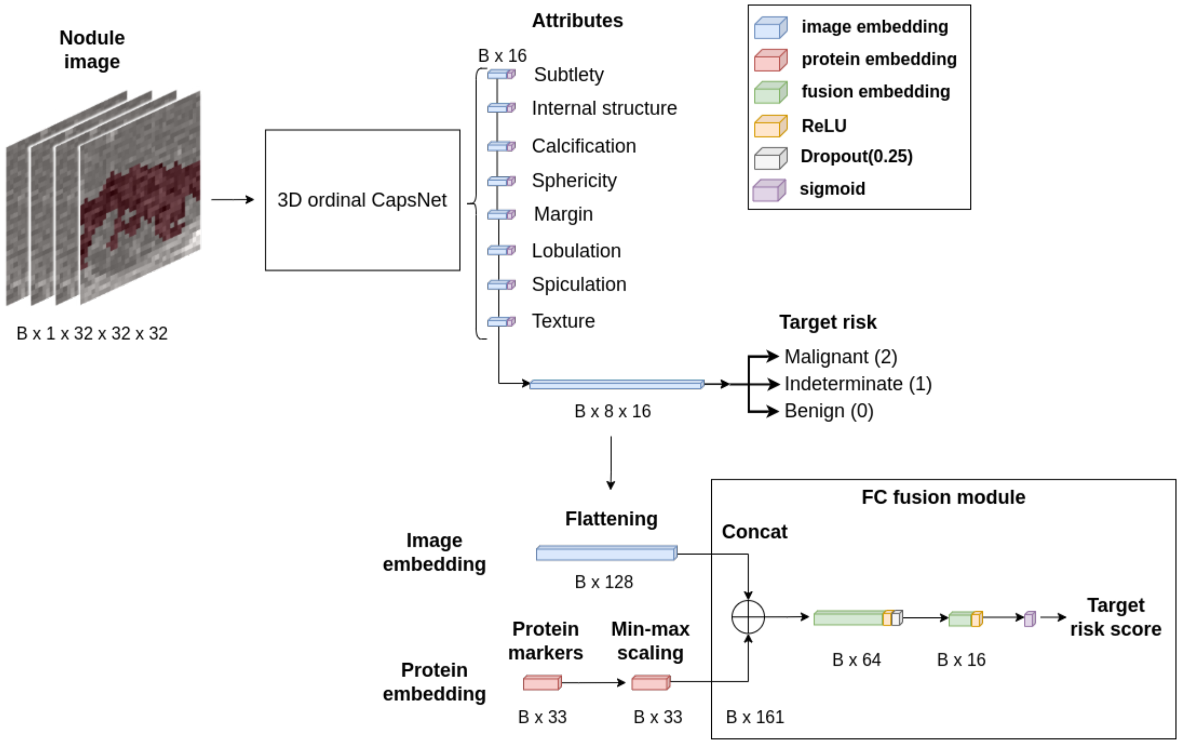


Figure 9.1: Multimodal architecture integrating image and protein embeddings for early lung cancer diagnosis.

In the validation and test sets, for patients with more than one segmented nodule (a single case in our cohort), each nodule's image features were concatenated with the same protein embedding, and the maximum predicted risk score across all nodules was selected as the final prediction.

The model was trained using the binary cross-entropy loss, incorporating a weighting factor to account for class imbalance. Specifically, the positive class (lung cancer) was weighted by the inverse class frequency ratio (negative-to-positive sample count) computed from the training subset. In each fold, the model was trained for a maximum of 50 epochs, with early stopping based on validation accuracy and a patience of 20 epochs. All models were coded in Pytorch 2.5.1 [268] and trained on NVIDIA A30 GPU (24 GB VRAM), with a batch size of 64, and Adam optimizer [343].

An ablation study was conducted by training the model with the same configuration using only the image embeddings or the protein features, in order to evaluate the individual contribution of each modality to the final prediction.

9.2.3 Interpretability with SHAP values

SHAP (SHapley Additive exPlanations) is a model-agnostic interpretability framework that assigns an importance value to each input feature based on its contribution to a given prediction [266]. SHAP values, introduced as an additive feature attribution method from game theory, allow for the local interpretation of model predictions by attributing the model's prediction to the individual input features.

In the proposed multimodal approach, which integrates both image-based embeddings and protein expression markers, SHAP values provide insights into how each modality, and each individual feature, contributes to the model's decision-making process. By analyzing SHAP values, the most influential image features and protein markers that drive predictions can be identified. This is particularly important in the case of circulating protein markers, as their clinical validation requires targeted assays such as ELISA (Enzyme-Linked ImmunoSorbent Assay) [344, 345], which are commonly used in clinical diagnostics. In contrast, Olink proteomics platform using proximity extension assay (PEA) is primarily employed for biomarker discovery [317], as it is a costly technique not routinely available in hospital settings. Prioritizing the most relevant markers can therefore streamline experimental validation and support their potential translation into clinical practice.

9.2.4 Existing lung cancer risk prediction tools

The multimodal approach was compared with three widely used and well-known risk prediction tools: Mayo Clinic model, Mayo Clinic model incorporating positron emission tomography (Mayo-PET), and Brock University model. Mayo-PET model could not be applied to all patients, as it is specifically used when a PET scan is available, and takes precedence over the standard Mayo model in such cases; otherwise, the standard Mayo score is considered. From a clinical perspective, when PET imaging is available, it is incorporated into the risk score because it provides crucial information on whether the pulmonary nodule exhibits increased metabolic activity, a hallmark of malignant tumors. Patient age, nodule size, smoking history, extra-thoracic cancer diagnosis $\geq$ 5-year prior to nodule presentation, upper lobe location and nodule spiculation parameters were incorporated into the Mayo Clinic model [338]. In the Brock University model, gender, age, nodule size, family history of cancer, emphysema, number of nodules, nodule type (nonsolid or with ground-glass opacity, part-solid, solid), nodule location (upper vs. middle or lower lobe) and nodule spiculation parameters were considered [337].

9.3 Results

Table 9.1 presents the performance of the multimodal model on P-ELCAP cohort, including benign nodule participants and lung cancer cases diagnosed in the same year as the blood collection. Sensitivity was calculated at different operating points (0.8, 0.9, and Youden's index) by determining the corresponding thresholds on the receiver operating characteristic (ROC) curve and applying them to the get the binary prediction scores. These thresholds were selected to reflect distinct clinical trade-offs relevant in a screening context, where prioritizing high sensitivity is essential to reduce the likelihood of missed cancer cases, while Youden's index (Equation 8.2) offers an optimal balance between sensitivity and specificity. The use of different random seeds had no statistically significant impact on model performance.

Table 9.1: Multimodal model performance on P-ELCAP cohort for predicting benign participants and patients diagnosed with lung cancer in the same year as the plasma collection, evaluated using 3-fold cross-validation test sets. YI: Youden's index.

Metric	**Sensitivity at 0.8**	**Sensitivity at 0.9**	**Sensitivity at YI**
AUC	0.86 ± 0.05	0.86 ± 0.05	0.86 ± 0.05
Threshold	0.55 ± 0.07	0.46 ± 0.02	0.47 ± 0.01
Accuracy	0.85 ± 0.05	0.80 ± 0.03	0.82 ± 0.06
Balanced Accuracy	0.84 ± 0.03	0.84 ± 0.03	0.84 ± 0.04
Precision/PPV	0.75 ± 0.09	0.62 ± 0.04	0.67 ± 0.13
Recall/Sensitivity	0.81 ± 0.02	0.94 ± 0.05	0.91 ± 0.09
F1-score	0.77 ± 0.05	0.75 ± 0.03	0.76 ± 0.05
Specificity	0.87 ± 0.07	0.74 ± 0.04	0.78 ± 0.12
NPV	0.91 ± 0.01	0.96 ± 0.03	0.95 ± 0.04

Figure 9.2 depicts normalized SHAP values averaged over the three folds evaluated in test subsets.

Table 9.2 shows the number of predicted lung cancer and benign nodules by the multimodal model in the false positives group of P-ELCAP cohort.

Table 9.2: Multimodal model predictions in the false positive group of P-ELCAP cohort at Youden's index threshold.

Fold	**# Lung cancer**	**# Benign nodules**
1	4	1
2	4	1
3	4	1

9.3.1 Ablation study

An ablation study was performed to assess the impact of each modality on the final result separately. Receiver Operating Characteristic (ROC) curves for each fold and model are shown in Figure 9.3. Figure 9.4 shows the average performance metrics for each model across the three folds evaluated at YI.

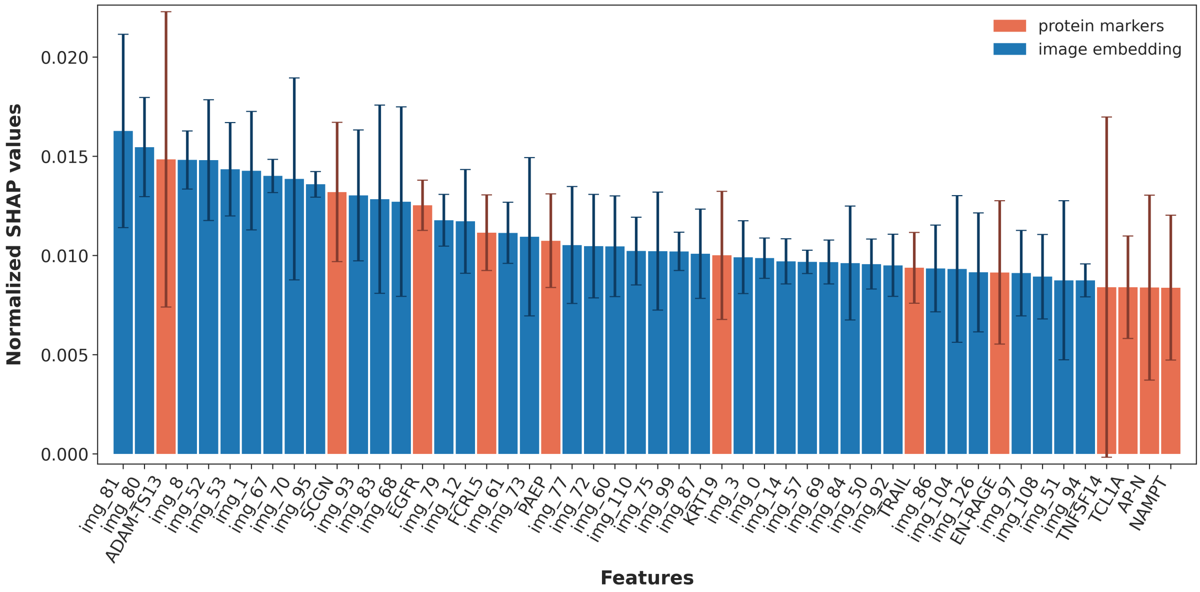


Figure 9.2: Normalized SHAP values averaged across the three folds in the multimodal approach integrating protein markers and image embeddings in the P-ELCAP cohort (mean and standard deviation).

Figure 9.5 depicts normalized SHAP values averaged over the three folds evaluated in test subsets for the image and protein models.

9.3.2 Comparison with existing risk prediction tools

ROC curves for each fold comparing existing risk prediction tools (Mayo, Mayo-PET and Brock) and the multimodal model are shown in Figure 9.6. Figure 9.7 presents the average performance metrics for the multimodal and the aforementioned risk prediction models across the three folds evaluated at YI.

9.4 Discussion and conclusion

This chapter presents a first proof-of-concept of a joint fusion approach integrating image-based feature embeddings with circulating protein markers to enhance early lung cancer diagnosis within a screening cohort. The proposed fully connected (FC) architecture constitutes a simple and effective strategy to combine both modalities, achieving high performance at different sensitivity thresholds. The manual selection of relevant protein markers was based on findings from a previous study [2] and prior analyses conducted on P-ELCAP dataset (Chapter 7). This step aimed to mitigate overfitting risks associated with the small sample size and the high dimensionality of the full Olink protein panel. Biologically informed neural networks (BINNs) [157] could be employed to incorporate pathway-level information directly into the model architecture, enabling the extraction of biologically meaningful embeddings from the full set of protein expression levels prior to integration. However, preliminary analysis on protein markers presented in Section 7.4.2 showed that BINNs achieved lower predictive performance compared to simpler models such as LASSO. In their original work, Ma *et al.* [157] highlighted that BINNs' performance is dependent on the

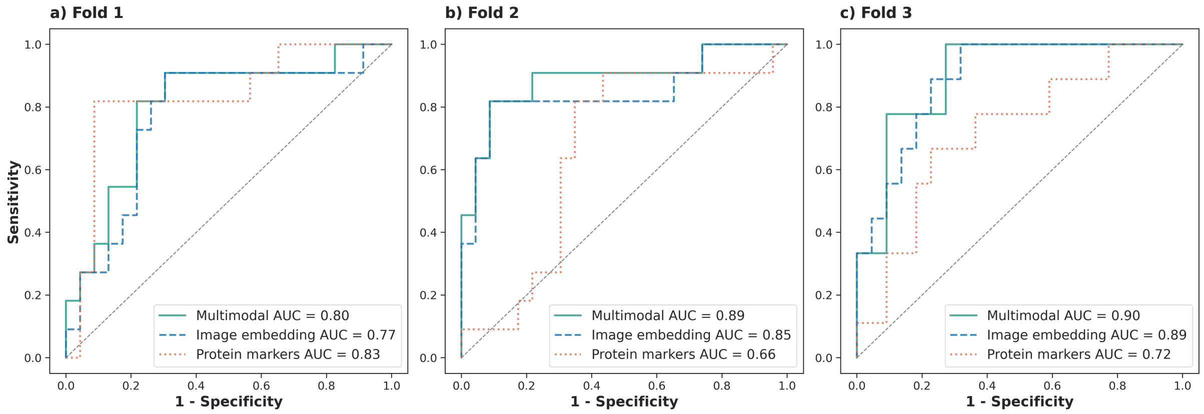


Figure 9.3: ROC curves for each cross-validation fold on the test subset of each model in the ablation study (multimodal, image embedding, protein markers) in the P-ELCAP cohort.

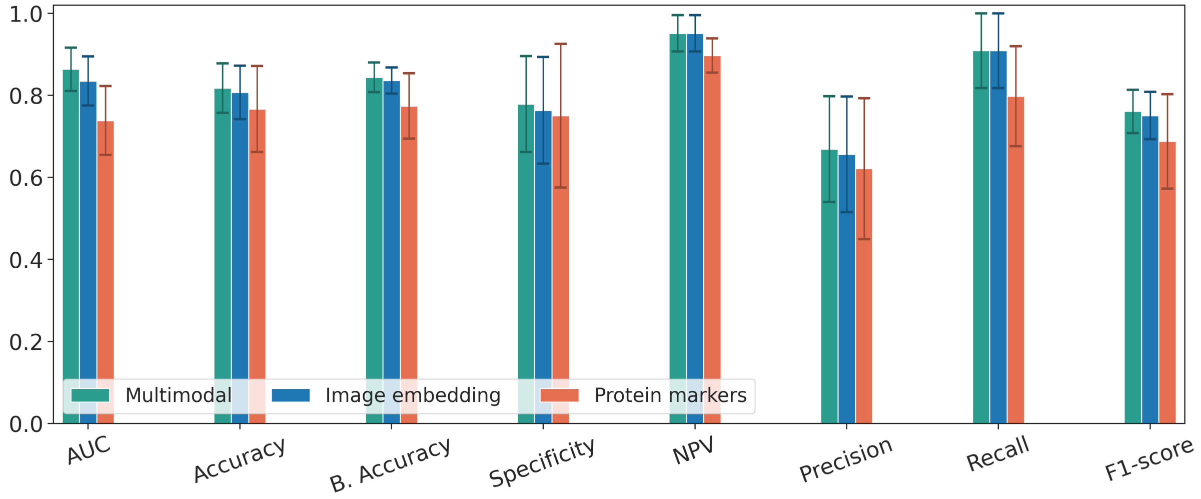
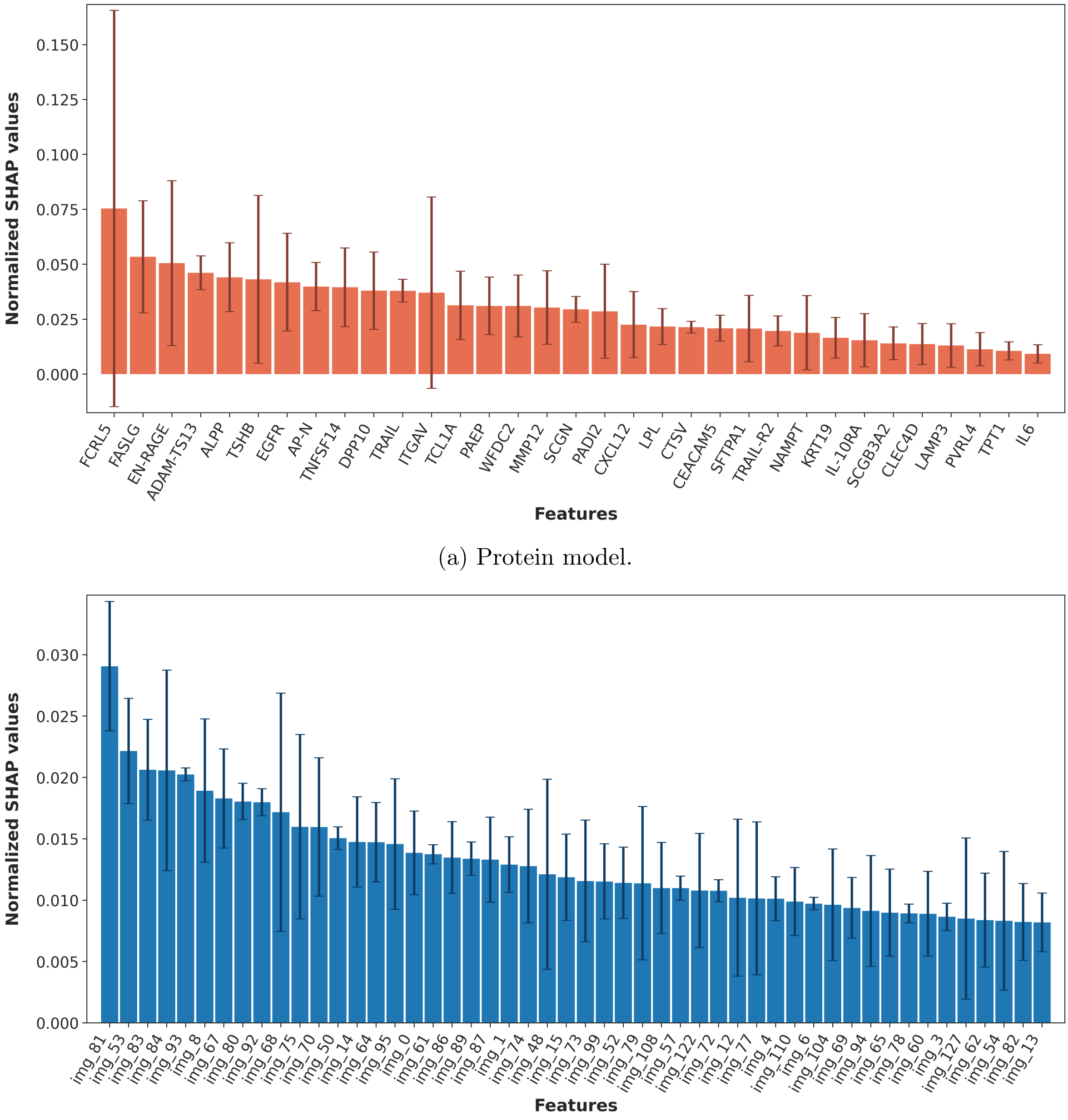


Figure 9.4: Performance on test for the multimodal, image and protein models trained on P-ELCAP for 3 fold cross-validation (mean and standard deviation) optimizing the threshold at Younden's index.

quality of the prior knowledge graph derived from the Reactome database, the specific characteristics of the dataset, and the degree of alignment between both. In small-scale clinical datasets such as the one used in this project, regularized linear models like LASSO may offer more robust performance than more complex neural network architectures.

The most significant contribution to the multimodal model comes from the image embeddings extracted by the ordinal CapsNet architecture. This result is consistent with the strong generalization performance exhibited by CapsNets on the P-ELCAP dataset, presented in Chapter 8. The protein model still achieves compatible performance with the multimodal and image-based models; however, it presents systematically higher variance and lower metrics at YI threshold, as depicted in Figure 9.4.

A key novelty of this research lies in exploring the complementary and overlapping nature of the two modalities. While image embeddings from lung nodules may be sufficient to indicate



(a) Protein model.

(b) Image model.

Figure 9.5: Normalized SHAP values averaged across the three folds in the ablation study in the P-ELCAP cohort (mean and standard deviation).

potential lung cancer, this model is refined with circulating protein markers that may be closely related to the patient's overall physiological or pathological state in response to lung cancer. Future research could investigate integrating clinical information with image embeddings and protein markers to further enhance the multimodal approach. Furthermore, an additional interesting finding emerged from the ablation study. When using only protein biomarkers, the model achieved an AUC of 0.73 (average across the 3 folds) with a NPV close to 1. This is clinically relevant because, in the context of lung cancer screening, LDCT imaging may not be accessible, individuals may decline the procedure, or they may not meet

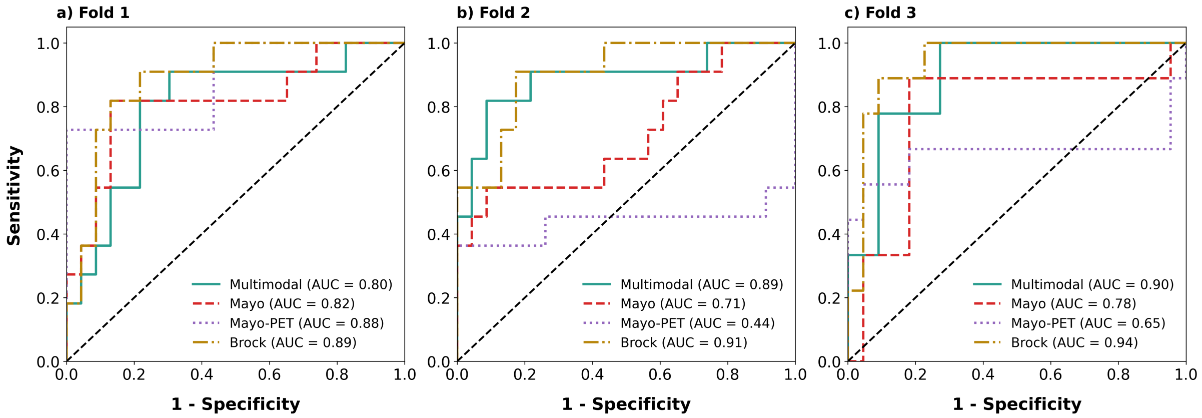


Figure 9.6: ROC curves for each cross-validation fold on the test subset comparing the multimodal model with existing lung cancer risk prediction tools (Mayo, Mayo-PET and Brock models) in the P-ELCAP cohort.

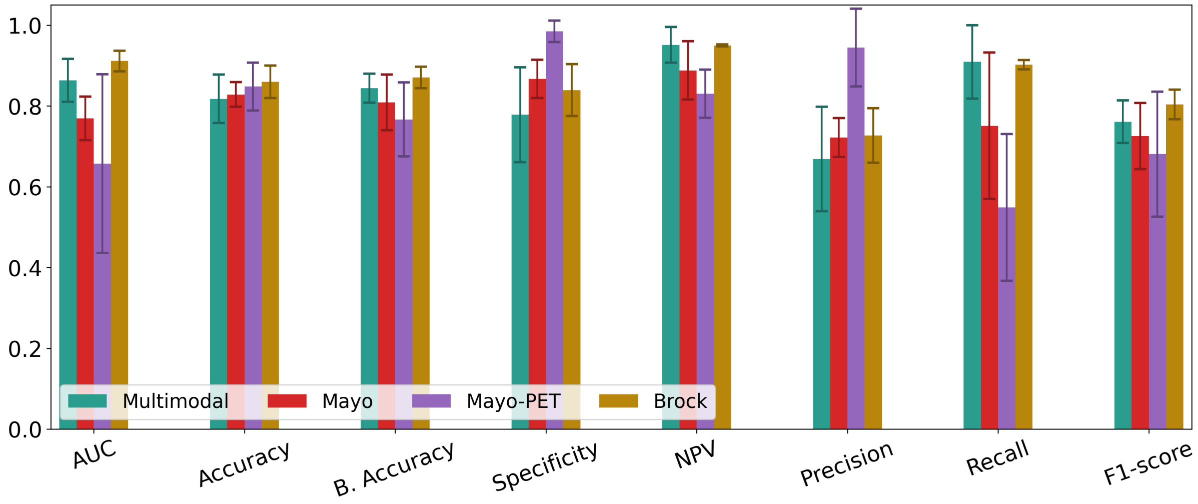


Figure 9.7: Performance on test for the multimodal, Mayo, Mayo-PET and Brock models on P-ELCAP for 3 fold cross-validation (mean and standard deviation) optimizing the threshold at Younden's index.

the established eligibility criteria for screening. In such cases, a simple blood extraction followed by protein quantification could be sufficient for assessing lung cancer risk.

The multimodal model exhibits comparable performance to the Mayo and Brock risk models. In particular, Brock model outperforms the multimodal approach, exhibiting lower variance. Both Mayo and Brock models, currently considered the gold standard for assessing the risk of IPNs, rely on the manual input of nodule characteristics together with clinical and demographic data. This process is inherently subjective and often requires interpretation by a multidisciplinary team, making it resource intensive. In contrast, the proposed multimodal model leverages the automated extraction of imaging features from the nodule's bounding box alongside circulating protein markers measured in plasma. This approach does not yet incorporate clinical information, although such integration is planned in future work to enhance predictive performance. The majority of variables employed by the Brock and Mayo tools are independent of the multimodal model. Thus, including clinical variables from Brock model in the multimodal framework will very likely improve risk stratification, and increase the model's interpretability by evaluating the shared and modality-specific contributions of imaging features, plasma proteomics, and clinical factors in lung cancer assessment.

Additionally, future work could explore alternative fusion strategies such as early and late fusion. Regarding early fusion methods, in which modalities are combined prior to feature encoding [335], a promising direction for future work involves using transformer-based architectures to model the interactions between modalities with more flexibility, integrating cross-modal relationships and intra-modality attention. In cancer prognosis, previous research has explored the integration of whole-slide imaging (WSI) and bulk transcriptomics through multimodal transformers [38]. These architectures could be adapted to LDCT images, circulating protein markers and clinical data. The main limitation of this approach is that transformers typically need to be trained on large enough datasets to generalize and avoid overfitting, whereas the joint fusion FC architecture employed in this study is better suited to the limited sample size available. Moreover, foundation models for WSI are more established in the literature than those based on CT images.

Late fusion methods combine the predicted scores for each individual modality (image, proteomics and clinical data) by applying a (weighted) average over the predictions for each modality or by training a separate model on top of the unimodal outputs (e.g., logistic regression, gradient techniques, random forest, Cox models) [335]. This modular strategy allows for robust integration of heterogeneous data sources and can enhance predictive performance while retaining interpretability for each individual modality. In late fusion, modalities are processed independently during training, enabling the integration of unpaired data and facilitating the handling of missing modalities. However, the absence of inter-modal interaction may constrain the expressiveness of the model [335].

Overall, investigating different fusion techniques to enhance early lung cancer diagnosis may offer complementary approaches that support a more comprehensive understanding of both individual and shared information across modalities.

9.5 Validation in future work

The multimodal AI algorithm will be externally evaluated in the Liverpool Lung cancer Project (LLP) screening cohort. This cohort will be validated under blinded conditions. LLP consists of 137 lung cancer screening cases, including LDCT image data and the corresponding segmentation of the nodule, that was performed by our collaborators in CUN. Circulating proteomics data was collected under the same conditions as P-ELCAP, as part of INTEGRAL consortium.

Future work could also explore mapping the most relevant image features to the corresponding nodule attributes extracted by the CapsNet model, to further interpret which characteristics are more strongly associated with lung nodule malignancy.

Conclusions

Artificial intelligence is significantly shaping the future of healthcare and medical diagnosis. Beyond automating processes, AI systems are driving innovation in clinical practice through the integration of heterogeneous data modalities to reveal hidden patterns in highly dimensional spaces, that enable the discovery of novel biomarkers, and contribute to more personalized clinical diagnoses and treatments. The present thesis focused on enhancing the robustness of DL methods for computed tomography-based diagnostic systems through the incorporation of prior knowledge and explainable AI techniques. Building on this objective, the research in this thesis explores existing limitations from both methodological and applied perspectives.

In the first part of the thesis, special emphasis was placed on informed ML to enhance the learning of AI systems in medical imaging, particularly with physics informed constraints. Furthermore, best practices were proposed to improve the standardization of clinical workflows in medical imaging, and to promote FAIR reporting of preprocessing pipelines, which prepare medical data for input into ML algorithms. In the second part of the thesis, two applications were investigated: intracranial hemorrhage prognosis and lung cancer early diagnosis. In the former, clinical knowledge embedded in CT images was leveraged through a multitask learning approach, achieving superior performance and higher interpretability. The latter focused on developing a medical informed DL ordinal approach for lung nodule risk malignancy assessment, evaluating the generalization to the P-ELCAP screening cohort. Additionally, this research explored integrating plasma protein biomarkers with image-based DL representations of screening-detected lung nodules to develop an individualized risk prediction model aimed at improving current screening practices. Ultimately, this work involved the design and curation of the P-ELCAP dataset, a novel lung cancer screening cohort including low-dose CT imaging, proteomics, clinical and demographic data, which is planned to be made publicly available to support the development and validation of personalized diagnostic tools in early lung cancer research.

This thesis has been done in an interdisciplinary environment in close collaboration with clinicians, radiologists and biologists, who have provided very valuable feedback to understand the clinical needs and the practical challenges in real-world medical applications. Overall, this thesis shows that developing AI systems in healthcare requires a comprehensive understanding of the clinical scenario and the standard protocols in medical practice. This knowledge is essential to produce clinically relevant algorithms that can subsequently be translated into real-world settings.

Summary of contributions

This thesis addresses challenges and opportunities in medical diagnostic systems with deep learning techniques, focusing on CT imaging. Particularly, it provides novel insights into developing more robust and explainable DL methods for CT-based lung cancer early diagnosis and intracranial hemorrhage prognosis. The contributions can be summarized as follows:

1. **The role of prior knowledge and physics in machine learning.** Chapter 3 emphasizes the relevance of prior knowledge to enhance the learning process of AI algorithms in medical imaging, focusing on physics informed ML. Domain knowledge in the form of physics, biology, or medical expertise can improve diagnostic accuracy and contribute to more personalized and reliable clinical decision-making. This chapter is intended to serve as a reference point for both newcomers to the field and experienced practitioners seeking to deepen their understanding of the underlying physical principles that inform image-based AI applications. Ultimately, domain expertise provides complementary information to purely data-driven methods, and is especially useful in medical applications, where data and annotations are limited.

2. **Best practices for standardization of medical imaging workflows.** Chapter 4 examines current limitations to the safe adoption of radiomics and AI algorithms, emphasizing the critical need for standardized protocols and workflows in medical imaging. Guidelines are presented to standardize clinical workflows in medical imaging, with references to the different levels at which homogenization is required and the hospital personnel involved in each phase. Overall, the clinical translation of AI systems in medical imaging relies on building trustworthy algorithms that capture the complexity of real world data.

3. **Best practices for preprocessing clinical informatics data.** In chapter 5, best practices to support standardized and reusable preprocessing pipelines are presented to foster the advancement of scientific research, promoting reproducibility, accessibility and validity of ML results. The current underreporting of preprocessing workflows is identified as a key challenge to the development of trustworthy AI systems in healthcare, as systematic documentation in accordance with FAIR principles is essential for detecting and preventing common reproducibility issues that threaten ML research, such as data leakage.

4. **Application of a multitask learning approach to enhance image-based intracranial hemorrhage prognosis.** Chapter 6 introduces a multi-task DL image-based approach for ICH prognosis, integrating the primary prognostic task with complementary predictions of Glasgow Coma Scale (GCS) and age. These variables guide decision-making in tabular models based on clinical records and demographic information relevant to ICH prognosis. Overall, the proposed multi-task methodology leverages clinical information embedded in GCS and age outputs to regularize the loss and learn more robust feature representations than state-of-the-art approaches.

5. **Design and curation of a screening cohort for personalized lung cancer diagnosis.** Chapter 7 describes the design and curation of the P-ELCAP cohort, in addition to technical validation experiments that demonstrate the quality and potential of the dataset. To the best of our knowledge, this is the first pulmonary nodule dataset to incorporate annotated LDCT images together with molecular protein biomarkers,

clinical and demographic variables. This dataset constitutes a valuable resource for advancing multimodal approaches to early lung cancer diagnosis. P-ELCAP holds significant value for external validation in other lung cancer screening cohorts, and represents an important step towards transforming the field of personalized medicine through AI-driven multimodal models in the context of LDCT-based lung cancer screening. Notably, P-ELCAP will be publicly released in Zenodo for the research community to support the development and validation of novel multimodal strategies in early lung cancer screening.

6. **Development and validation of a medical informed deep learning ordinal approach for lung nodule malignancy assessment.** Chapter 8 focuses on clinical relevance to introduce a medical informed ordinal framework for lung nodule malignancy assessment. The generalizability of state-of-the-art 2D and 3D DL architectures is explored to enhance indeterminate pulmonary nodule's characterization in LDCT scans. This chapter emphasizes that addressing real-world challenges in medical diagnosis requires both a solid understanding of clinical workflows and context-specific adaptation of existing publicly available datasets to the particular task. Ultimately, this framework contributes to earlier diagnosis of lung cancer, providing clinically meaningful feedback to radiologists, and highlights the critical role of external validation in the clinical translation of AI systems.

7. **Application of a multimodal approach to enhance early lung cancer diagnosis.** Chapter 9 presents a multimodal joint fusion model integrating LDCT image-based feature embeddings extracted from lung nodules with circulating protein markers in prediagnostic plasma samples. This approach leverages the complementary information in the two modalities to enhance early lung cancer diagnosis in the P-ELCAP cohort. The multimodal model exhibits high performance across various sensitivity thresholds, outperforming individual unimodal approaches, and underscoring its potential to refine current lung cancer screening practices through clinical validation of the most informative protein markers.

Future directions

Future validation and deeper exploration of robust and explainable DL systems for CT–based diagnosis are essential to support clinical translation. Expert knowledge will continue to play a key role in ensuring robustness, clinical relevance, and alignment with clinical practices. Future advances will facilitate real-world deployment and support more personalized diagnostic tools, ultimately contributing to the progress of precision medicine. Building upon the work presented in this thesis, several promising directions for future research can be identified:

1. **Extension of the multitask approach for intracranial hemorrhage prognosis to new datasets.** Future work on ICH prognosis will evaluate the generalizability of the multi-task model using two independent datasets from the reference hospitals in Oviedo and León (Spain). This approach will incorporate uncertainty quantification techniques to assess the reliability and robustness of the model's predictions in external data.

2. **Multi-scale approach to LDCT-based lung cancer diagnosis.** The proposed medical informed DL ordinal approach for lung nodule assessment will be externally validated on another LDCT screening cohort, the Liverpool Lung cancer Project.

Besides, future work could validate the properties and interpretability of the learned nodule attributes in external datasets, assessing the findings with a radiologist. Further investigation on the 3D CapsNet's explainability and generalizability to new datasets constitutes another direction of interest.
The next generation of DL algorithms for early lung cancer diagnosis using LDCT imaging is foreseen as a multi-scale approach, which integrates local nodule assessment with global evaluation of the entire LDCT scan. In the future, the proposed medical informed approach could be incorporated into an end-to-end DL pipeline for lung nodule detection, segmentation and malignancy risk analysis.

3. **Personalized multimodal approaches for early lung cancer diagnosis.** Future work will incorporate clinical variables, evaluate new architectures and validate the results in the Liverpool lung cancer cohort. Next steps could involve mapping the most relevant image features to the corresponding nodule attributes extracted by the CapsNet model, to further interpret which characteristics are more strongly associated with lung nodule malignancy, and the complementary information between them and the protein markers. Moreover, future research could investigate different fusion techniques as complementary approaches that support a more comprehensive understanding of both individual and shared information across modalities to enhance lung cancer early diagnosis.

Final remarks

The future of AI in medical imaging relies on high quality external validation, interdisciplinary collaborative environments and effective prior knowledge integration to build trustworthy AI systems. AI in healthcare is envisioned as a collaborative effort among clinicians, physicists, biologists, and AI developers, where AI algorithms support the workflow of medical professionals. This interdisciplinary synergy will enable AI to move from a tool of analysis to a true collaborative partner in clinical decision-making. Ultimately, the adoption of these technologies will promote the advancement of personalized medicine through more targeted and informed clinical decision-making.

Conclusiones

La inteligencia artificial (IA) está transformando profundamente el futuro de la atención sanitaria y el diagnóstico médico. Más allá de automatizar procesos, los sistemas de IA están impulsando la innovación en la práctica clínica a través de la integración de modalidades de datos heterogéneas con el fin de encontrar patrones ocultos en espacios de alta dimensionalidad, que permiten el descubrimiento de nuevos biomarcadores, y contribuyen a diagnósticos y tratamientos clínicos más personalizados. La presente tesis se centra en la mejora de la robustez de los métodos de aprendizaje profundo (DL) para sistemas de diagnóstico basados en tomografía computarizada (TC), mediante la incorporación de conocimientos previos y técnicas de IA explicable. Partiendo de este objetivo, la investigación de esta tesis examina las limitaciones existentes desde perspectivas tanto metodológicas como aplicadas.

En la primera parte de la tesis, se hizo especial hincapié en el aprendizaje automático (ML) informado para mejorar el aprendizaje de los sistemas de IA en imagen médica, en particular aplicando restricciones fundamentadas en la física de las imágenes. Además, se propusieron buenas prácticas para mejorar la estandarización de los flujos de trabajo clínicos en imagen médica, así como para promover la documentación siguiendo los principios FAIR de las metodologías de preprocesado, que preparan los datos médicos para ser usados por algoritmos de ML. En la segunda parte de la tesis, se investigaron dos aplicaciones: el pronóstico de hemorragia intracraneal y el diagnóstico precoz del cáncer de pulmón. En la primera, se aprovechó el conocimiento clínico integrado en las imágenes de TC con un método de aprendizaje multitarea, logrando un rendimiento superior y una mayor interpretabilidad. La segunda se centró en desarrollar un modelo ordinal de DL informado desde el punto de vista médico para la evaluación del riesgo de malignidad de nódulos pulmonares, evaluando su capacidad de generalización a la cohorte de cribado P-ELCAP. Asimismo, esta investigación exploró la integración de biomarcadores proteicos plasmáticos con representaciones de DL obtenidas a partir de la imagen de nódulos pulmonares detectados en cribado, con el objetivo de desarrollar un modelo de predicción de riesgo individualizado destinado a mejorar las prácticas de diagnóstico precoz actuales. Finalmente, este trabajo implicó el diseño y la curación del conjunto de datos P-ELCAP, una nueva cohorte de cribado de cáncer de pulmón que incluye imágenes de TC de baja dosis, proteómica, datos clínicos y demográficos, que se publicará en abierto para apoyar el desarrollo y la validación de nuevas herramientas de diagnóstico personalizadas en la investigación en cáncer de pulmón.

Esta tesis se ha realizado en un entorno interdisciplinar en estrecha colaboración con profesionales médicos, radiólogos y biólogos, que han aportado información muy valiosa para comprender las necesidades clínicas y los retos prácticos de las aplicaciones médicas en el mundo real. En conjunto, esta tesis pone de manifiesto que el desarrollo de sistemas de IA en medicina requiere una comprensión exhaustiva del escenario clínico y de los protocolos sanitarios. Este conocimiento es esencial para desarrollar algoritmos relevantes desde el punto de vista clínico que puedan ser trasladados posteriormente al mundo real.

Resumen de contribuciones

Esta tesis aborda retos y oportunidades en el diagnóstico médico con técnicas de aprendizaje profundo, enfocándose en sistemas basados en TC. En particular, aporta nuevas perspectivas para desarrollar métodos de DL más robustos y explicables en TC para el diagnóstico precoz del cáncer de pulmón y el pronóstico de hemorragias intracraneales. Las contribuciones pueden resumirse como sigue:

1. **El papel del conocimiento previo y la física en el aprendizaje automático.** El capítulo 3 hace hincapié en la relevancia del conocimiento a priori para mejorar el proceso de aprendizaje de los algoritmos de IA en imagen médica, centrándose en el ML informado por la física. El conocimiento de la física, la biología o la experiencia médica puede mejorar la precisión del diagnóstico y contribuir a la adopción de decisiones clínicas más personalizadas y fiables. Este capítulo sirve como punto de referencia tanto para los recién llegados a este campo como para los profesionales con experiencia que buscan profundizar en su comprensión de los principios físicos subyacentes que informan las aplicaciones de IA basadas en imágenes médicas. En definitiva, el conocimiento experto proporciona información complementaria a los métodos puramente basados en los datos, y es especialmente útil en aplicaciones médicas, donde los datos y las anotaciones son limitados.

2. **Buenas prácticas para la estandarización de los procesos y flujos de trabajo en imagen médica.** El capítulo 4 examina las limitaciones actuales para la adopción segura de la radiómica y los algoritmos de IA, destacando la necesidad crítica de protocolos y flujos de trabajo estandarizados en imagen médica. Se presentan pautas para estandarizar los procesos clínicos en imagen médica, con referencias a los diferentes niveles en los que se requiere homogeneización y al personal sanitario implicado en cada fase. En general, la traslación clínica de los sistemas de IA en imagen médica radica en el desarrollo de algoritmos fiables que capturen la complejidad de los datos del mundo real.

3. **Buenas prácticas para el preprocesado de datos informáticos clínicos.** En el capítulo 5, se presentan buenas prácticas para fomentar técnicas de preprocesado estandarizadas y reutilizables que impulsen el avance de la investigación científica, promoviendo la reproducibilidad, accesibilidad y validez de los resultados del ML. La actual documentación incompleta de los métodos de preprocesado se identifica como un desafío clave para el desarrollo de sistemas de IA fiables en la asistencia sanitaria. Una documentación sistemática conforme a los principios FAIR es esencial para detectar y prevenir problemas comunes de reproducibilidad que amenazan la investigación en ML, tales como la filtración de datos.

4. **Aplicación de un modelo de aprendizaje multitarea para mejorar el pronóstico de hemorragia intracraneal basado en imagen.** El capítulo 6 presenta un modelo de aprendizaje profundo multitarea basado en imagen TC para el pronóstico de hemorragia intracraneal (HIC), integrando la tarea principal de pronóstico con las predicciones complementarias de la Escala de Coma de Glasgow (GCS) y la edad. Estas

variables guían la toma de decisiones en los modelos tabulares basados en registros clínicos e información demográfica relevantes para el pronóstico de HIC. En resumen, la metodología multitarea propuesta aprovecha la información clínica integrada en las predicciones de la GCS y la edad para regularizar la función de pérdida y aprender representaciones en el espacio de características más robustas que las de los modelos existentes.

5. **Diseño y curación de una cohorte de cribado para diagnóstico personalizado del cáncer de pulmón.** El capítulo 7 describe el diseño y la curación de la cohorte P-ELCAP, así como los experimentos de validación técnica que demuestran la calidad y el potencial de este conjunto de datos. Hasta donde nuestro conocimiento alcanza, se trata del primer conjunto de datos con nódulos pulmonares que incorpora imágenes de TC de baja dosis (LDCT) anotadas junto con biomarcadores moleculares de proteínas, variables clínicas y demográficas. Este conjunto de datos es muy valioso para avanzar en técnicas multimodales de diagnóstico precoz del cáncer de pulmón. P-ELCAP es de gran utilidad para llevar a cabo validaciones externas de metodologías desarrolladas en otras cohortes de cribado de cáncer de pulmón. Asimismo, constituye un paso importante para impulsar el campo de la medicina personalizada con modelos multimodales de IA en el contexto de cribado de cáncer de pulmón con LDCT. Cabe destacar que P-ELCAP se publicará en Zenodo para toda la comunidad investigadora con el fin de promover el desarrollo y la validación de nuevas estrategias multimodales en el diagnóstico temprano del cáncer de pulmón.

6. **Desarrollo y validación de un modelo ordinal de DL informado desde el punto de vista médico para evaluar la malignidad de nódulos pulmonares.** El capítulo 8 se centra en la relevancia clínica para introducir un método ordinal de DL informado desde el punto de vista médico para la evaluación de la malignidad de nódulos pulmonares. Se examina la generalizabilidad de arquitecturas de DL 2D y 3D avanzadas para mejorar la caracterización de nódulos pulmonares indeterminados en LDCT. Este capítulo pone de manifiesto que abordar los retos del mundo real en el diagnóstico médico requiere una sólida comprensión de los protocolos clínicos, y una adaptación específica del contexto de los conjuntos de datos en abierto disponibles a la tarea en particular. En última instancia, este método contribuye a un diagnóstico más temprano del cáncer de pulmón, proporcionando resultados útiles desde el punto de vista clínico a los radiólogos, y resalta el papel fundamental de la validación externa en la traslación clínica de los sistemas de IA.

7. **Aplicación de un modelo multimodal para la mejora del diagnóstico precoz del cáncer de pulmón.** El capítulo 9 presenta un modelo de fusión multimodal que integra características extraídas por DL de la imagen LDCT de nódulos pulmonares con marcadores proteicos circulantes en muestras de plasma prediagnóstico. Este método aprovecha la información complementaria de ambas modalidades para mejorar el diagnóstico precoz del cáncer de pulmón en la cohorte P-ELCAP. El modelo multimodal presenta un alto rendimiento en distintos umbrales de sensibilidad, superando los modelos unimodales individuales, y evidenciando su potencial para refinar las prácticas actuales de cribado de cáncer de pulmón a través de la validación clínica de los marcadores proteicos más informativos.

Perspectivas futuras

La validación y la investigación en técnicas de DL explicables y robustas para sistemas de diagnóstico basados en tomografía computarizada son esenciales para lograr la traslación clínica en el futuro. El conocimiento experto seguirá desempeñando un papel clave para garantizar la robustez, la relevancia clínica y el alineamiento con las prácticas sanitarias. Los avances futuros facilitarán el despliegue en entornos reales, y contribuirán a la creación de herramientas de diagnóstico más personalizadas, impulsando en última instancia el avance de la medicina de precisión. A partir del trabajo presentado en esta tesis, se identifican varias líneas prometedoras para investigaciones futuras:

1. **Extensión del modelo de aprendizaje multitarea para el pronóstico de hemorragia intracraneal a nuevos conjuntos de datos.** El trabajo futuro en el pronóstico de hemorragia intracraneal evaluará la generalización del modelo multitarea utilizando dos conjuntos de datos independientes de los hospitales de referencia en Oviedo y León (España). Este enfoque incorporará técnicas de cuantificación de la incertidumbre para evaluar la fiabilidad y la robustez de las predicciones del modelo en datos externos.

2. **Sistema multiescala para el diagnóstico del cáncer de pulmón con tomografía computarizada de baja dosis.** El modelo ordinal de DL informado desde el punto de vista médico propuesto para la evaluación de nódulos pulmonares se validará externamente en otra cohorte de cribado de LDCT, el Proyecto de cáncer de pulmón de Liverpool. Además, en el futuro se podrían validar las propiedades y la interpretabilidad de los atributos de los nódulos aprendidos por el modelo en conjuntos de datos externos, evaluando los resultados con un radiólogo. Otra línea de interés es la investigación en explicabilidad y generalizabilidad del modelo 3D CapsNet a nuevos conjuntos de datos.
   Se prevé que la próxima generación de algoritmos de DL para el diagnóstico precoz del cáncer de pulmón a través de LDCT consistirá en sistemas multiescala, que integren la evaluación a nivel local de los nódulos con el análisis global de toda la imagen LDCT. En el futuro, el modelo informado propuesto podría incorporarse en un sistema de DL integral para la detección, la segmentación y el análisis del riesgo de malignidad de los nódulos pulmonares.

3. **Modelos multimodales personalizados para el diagnóstico precoz del cáncer de pulmón.** El trabajo futuro incorporará variables clínicas, evaluará nuevas arquitecturas y validará los resultados en la cohorte de cribado de cáncer de pulmón de Liverpool. Los próximos pasos podrían consistir en asignar las características más relevantes de la imagen a los atributos correspondientes del nódulo extraídos por el modelo CapsNet, para interpretar qué características están más asociadas con la malignidad de los nódulos pulmonares, así como la información complementaria entre estas y los marcadores proteicos.
   Además, futuras investigaciones podrían estudiar distintas técnicas de fusión como métodos complementarios que permitan una comprensión más completa de la información individual y compartida entre las diferentes modalidades, con el fin de mejorar el diagnóstico precoz del cáncer de pulmón.

Comentarios finales

El futuro de la IA en imagen médica depende de una validación externa de alta calidad, entornos colaborativos interdisciplinarios, y una integración eficaz de los conocimientos previos existentes para diseñar sistemas de IA fiables. La IA en la atención sanitaria se concibe como un esfuerzo colaborativo entre profesionales médicos, físicos, biólogos y desarrolladores de IA, donde los algoritmos apoyan el flujo de trabajo de los profesionales sanitarios. Esta sinergia interdisciplinar permitirá que la IA pase de ser una herramienta de análisis a un verdadero aliado en la toma de decisiones. En última instancia, la adopción de estas tecnologías promoverá el avance de la medicina personalizada a través de decisiones clínicas más precisas e informadas.

References

[1] A. P. del Barrio, P. M. Fernández-Miranda, P. S. Bellón, A. E. Domínguez, L. L. Iglesias, E. M. Fraguela, and D. R. González, "Head-ct 2d/3d images with and without ich prepared for deep learning," 2022.

[2] E. Khodayari Moez, M. T. Warkentin, Y. Brhane, S. Lam, J. K. Field, G. Liu, J. J. Zulueta, K. Valencia, M. Mesa-Guzman, A. P. Nialet *et al.*, "Circulating proteome for pulmonary nodule malignancy," *JNCI: Journal of the National Cancer Institute*, vol. 115, no. 9, pp. 1060–1070, 2023.

[3] L. Gallée, M. Beer, and M. Götz, "Interpretable medical image classification using prototype learning and privileged information," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2023, pp. 435–445.

[4] M. A. Lones, "Avoiding common machine learning pitfalls," *Patterns*, vol. 5, no. 10, 2024.

[5] M. Cobo, F. Pérez-Rojas, C. Gutiérrez-Rodríguez, I. Heredia, P. Maragaño-Lizama, F. Yung-Manriquez, L. Lloret Iglesias, and J. A. Vega, "Novel deep learning method for coronary artery tortuosity detection through coronary angiography," *Scientific Reports*, vol. 13, no. 1, p. 11137, 2023.

[6] F. De Sousa Ribeiro, K. Duarte, M. Everett, G. Leontidis, and M. Shah, "Object-centric learning with capsule networks: A survey," *ACM Computing Surveys*, vol. 56, no. 11, pp. 1–291, 2024.

[7] V. Kaul, S. Enslin, and S. A. Gross, "History of artificial intelligence in medicine," *Gastrointestinal endoscopy*, vol. 92, no. 4, pp. 807–812, 2020.

[8] B. Bhinder, C. Gilvary, N. S. Madhukar, and O. Elemento, "Artificial intelligence in cancer research and precision medicine," *Cancer discovery*, vol. 11, no. 4, pp. 900–915, 2021.

[9] E. Fountzilas, T. Pearce, M. A. Baysal, A. Chakraborty, and A. M. Tsimberidou, "Convergence of evolving artificial intelligence and machine learning techniques in precision oncology," *npj Digital Medicine*, vol. 8, no. 1, p. 75, 2025.

[10] M. Graziani, L. Dutkiewicz, D. Calvaresi, J. P. Amorim, K. Yordanova, M. Vered, R. Nair, P. H. Abreu, T. Blanke, V. Pulignano *et al.*, "A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences," *Artificial intelligence review*, vol. 56, no. 4, pp. 3473–3504, 2023.

[11] E. H. Shortliffe, *Computer-Based Medical Consultations: MYCIN*, 1st ed. New York: Elsevier, 1976, eBook edition.

[12] U.S. Food and Drug Administration, "Premarket approval (pma) for imagechecker m1000 (p970058)," 1998, accessed: 2025-05-12. [Online]. Available: https://www.accessdata.fda.gov/cdrh_docs/pdf/p970058.pdf

[13] S. S. Gill, H. Wu, P. Patros, C. Ottaviani, P. Arora, V. C. Pujol, D. Haunschild, A. K. Parlikad, O. Cetinkaya, H. Lutfiyya *et al.*, "Modern computing: Vision and challenges," *Telematics and Informatics Reports*, vol. 13, p. 100116, 2024.

[14] S. Pawan and J. Rajan, "Capsule networks for image classification: A review," *Neurocomputing*, vol. 509, pp. 102–120, 2022.

[15] B. Widrow and M. A. Lehr, "30 years of adaptive neural networks: perceptron, madaline, and backpropagation," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1415–1442, 1990.

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[17] C. Chen, N. A. M. Isa, and X. Liu, "A review of convolutional neural network based methods for medical image classification," *Computers in Biology and Medicine*, vol. 185, p. 109507, 2025.

[18] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, p. 99, 2024.

[19] L. Zhao and Z. Zhang, "A improved pooling method for convolutional neural networks," *Scientific Reports*, vol. 14, no. 1, p. 1589, 2024.

[20] P. M. Fernández-Miranda, E. M. Fraguela, M. Á. de Linera-Alperi, M. Cobo, A. P. Del Barrio, D. R. González, J. A. Vega, and L. L. Iglesias, "A retrospective study of deep learning generalization across two centers and multiple models of x-ray devices using covid-19 chest-x rays," *Scientific Reports*, vol. 14, no. 1, p. 14657, 2024.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019. [Online]. Available: https://arxiv.org/abs/1912.01703

[26] M. Consortium, “Monai: Medical open network for ai,” Oct. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8436376

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[28] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning.* PMLR, 2019, pp. 6105–6114.

[29] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[30] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[32] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[34] M. E. Consens, C. Dufault, M. Wainberg, D. Forster, M. Karimzadeh, H. Goodarzi, F. J. Theis, A. Moses, and B. Wang, “Transformers and genome language models,” *Nature Machine Intelligence*, pp. 1–17, 2025.

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

[36] J.-B. Cordonnier, A. Loukas, and M. Jaggi, “On the relationship between self-attention and convolutional layers,” *arXiv preprint arXiv:1911.03584*, 2019.

[37] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, “Transformers in medical imaging: A survey,” *Medical image analysis*, vol. 88, p. 102802, 2023.

[38] G. Jaume, A. Vaidya, R. J. Chen, D. F. Williamson, P. P. Liang, and F. Mahmood, “Modeling dense multimodal interactions between biological pathways and histology for survival prediction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 579–11 590.

[39] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” *Advances in neural information processing systems*, vol. 30, 2017.

[40] R. Shi and L. Niu, "A brief survey on capsule network," in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2020, pp. 682–686.

[41] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," in *International conference on learning representations*, 2018.

[42] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[43] Y. Bengio, Y. Lecun, and G. Hinton, "Deep learning for ai," *Communications of the ACM*, vol. 64, no. 7, pp. 58–65, 2021.

[44] X. Hu, M. Shi, W. Wang, S. Wu, L. Xing, W. Wang, X. Zhu, L. Lu, J. Zhou, X. Wang *et al.*, "Demystify transformers & convolutions in modern image deep networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[45] M. Tran, V.-K. Vo-Ho, K. Quinn, H. Nguyen, K. Luu, and N. Le, "Capsnet for medical image segmentation," in *Deep Learning for Medical Image Analysis*. Elsevier, 2024, pp. 75–97.

[46] R. LaLonde, Z. Xu, I. Irmakci, S. Jain, and U. Bagci, "Capsules for biomedical image segmentation," *Medical image analysis*, vol. 68, p. 101889, 2021.

[47] F. Long, J.-J. Peng, W. Song, X. Xia, and J. Sang, "Bloodcaps: A capsule network based model for the multiclassification of human peripheral blood cells," *Computer methods and programs in biomedicine*, vol. 202, p. 105972, 2021.

[48] A. Zador, S. Escola, B. Richards, B. Ölveczky, Y. Bengio, K. Boahen, M. Botvinick, D. Chklovskii, A. Churchland, C. Clopath *et al.*, "Catalyzing next-generation artificial intelligence through neuroai," *Nature communications*, vol. 14, no. 1, p. 1597, 2023.

[49] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024.

[50] F. Li, X. Lu, and J. Yuan, "Mha-corocapsule: multi-head attention routing-based capsule network for covid-19 chest x-ray image classification," *IEEE Transactions on Medical Imaging*, vol. 41, no. 5, pp. 1208–1218, 2021.

[51] J. T. Bushberg and J. M. Boone, *The essential physics of medical imaging.* Wolters Kluwer, 2021.

[52] M. Cobo, P. Menéndez Fernández-Miranda, G. Bastarrika, and L. Lloret Iglesias, "Enhancing radiomics and deep learning systems through the standardization of medical imaging workflows," *Scientific data*, vol. 10, no. 1, p. 732, 2023.

[53] N. Konz and M. A. Mazurowski, "The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images," *arXiv preprint arXiv:2401.08865*, 2024.

[54] R. Qureshi, M. Irfan, H. Ali, A. Khan, A. S. Nittala, S. Ali, A. Shah, T. M. Gondal, F. Sadak, Z. Shah *et al.*, “Artificial intelligence and biosensors in healthcare and its clinical relevance: A review,” *IEEE access*, vol. 11, pp. 61 600–61 620, 2023.

[55] G. Varoquaux and V. Cheplygina, “Machine learning for medical imaging: methodological failures and recommendations for the future,” *NPJ digital medicine*, vol. 5, no. 1, p. 48, 2022.

[56] S. N. Saw and K. H. Ng, “Current challenges of implementing artificial intelligence in medical imaging,” *Physica Medica*, vol. 100, pp. 12–17, 2022.

[57] O. S. Pianykh, *Digital imaging and communications in medicine (DICOM): a practical introduction and survival guide.* Springer, 2012, vol. 10.

[58] DICOM Standard Committee, “About dicom: Overview,” accessed: 2024-11-26. [Online]. Available: https://www.dicomstandard.org/about

[59] S. L. Schneider, I. Kohli, I. H. Hamzavi, M. L. Council, A. M. Rossi, and D. M. Ozog, “Emerging imaging technologies in dermatology: Part ii: Applications and limitations,” *Journal of the American Academy of Dermatology*, vol. 80, no. 4, pp. 1121–1131, 2019.

[60] D. Chahal and M. F. Byrne, “A primer on artificial intelligence and its application to endoscopy,” *Gastrointestinal endoscopy*, vol. 92, no. 4, pp. 813–820, 2020.

[61] N. Kumar, R. Gupta, and S. Gupta, “Whole slide imaging (wsi) in pathology: current perspectives and future directions,” *Journal of digital imaging*, vol. 33, no. 4, pp. 1034–1040, 2020.

[62] C. D. Bahadir, M. Omar, J. Rosenthal, L. Marchionni, B. Liechty, D. J. Pisapia, and M. R. Sabuncu, “Artificial intelligence applications in histopathology,” *Nature Reviews Electrical Engineering*, vol. 1, no. 2, pp. 93–108, 2024.

[63] J. Zhang and A. Zhang, “Deep learning-based multi-model approach on electron microscopy image of renal biopsy classification,” *BMC nephrology*, vol. 24, no. 1, p. 132, 2023.

[64] A. Grzybowski, K. Jin, J. Zhou, X. Pan, M. Wang, J. Ye, and T. Y. Wong, “Retina fundus photograph-based artificial intelligence algorithms in medicine: A systematic review,” *Ophthalmology and Therapy*, vol. 13, no. 8, pp. 2125–2149, 2024.

[65] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito *et al.*, “Optical coherence tomography,” *science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[66] J. Wang, S. Nolen, W. Song, W. Shao, W. Yi, A. Kashani, and J. Yi, “A dual-channel visible light optical coherence tomography system enables wide-field, full-range, and shot-noise limited human retinal imaging,” *Communications Engineering*, vol. 3, no. 1, p. 21, 2024.

[67] D. Kalupahana, N. S. Kahatapitiya, D. Kamalathasan, R. E. Wijesinghe, B. N. Silva, and U. Wijenayake, “State-of-the-art of deep learning in multidisciplinary optical coherence tomography applications,” *IEEE Access*, vol. 12, pp. 164 462–164 490, 2024.

[68] L. Álvarez-Rodríguez, A. Pueyo, J. de Moura, E. Vilades, E. Garcia-Martin, C. I. Sánchez, J. Novo, and M. Ortega, "Fully automatic deep convolutional approaches for the screening of neurodegeneratives diseases using multi-view oct images," *Artificial Intelligence in Medicine*, p. 103006, 2024.

[69] Q. Lyu, R. Neph, and K. Sheng, "Tomographic detection of photon pairs produced from high-energy x-rays for the monitoring of radiotherapy dosing," *Nature Biomedical Engineering*, vol. 7, no. 3, pp. 323–334, 2023.

[70] I. Różyło-Kalinowska, "Panoramic radiography in dentistry," *Clinical Dentistry Reviewed*, vol. 5, no. 1, p. 26, 2021.

[71] S. S. Alharbi and H. F. Alhasson, "Exploring the applications of artificial intelligence in dental image detection: A systematic review," *Diagnostics*, vol. 14, no. 21, p. 2442, 2024.

[72] J. A. Brink, J. P. Heiken, G. Wang, K. W. McEnery, F. J. Schlueter, and M. Vannier, "Helical ct: principles and technical considerations." *Radiographics*, vol. 14, no. 4, pp. 887–893, 1994.

[73] M. Van Straten, H. Venema, C. Majoie, N. Freling, C. Grimbergen, and G. Den Heeten, "Image quality of multisection ct of the brain: thickly collimated sequential scanning versus thinly collimated spiral scanning with image combining," *American journal of neuroradiology*, vol. 28, no. 3, pp. 421–427, 2007.

[74] Y. Salimi, I. Shiri, A. Akhavanallaf, Z. Mansouri, A. Sanaat, M. Pakbin, M. Ghasemian, H. Arabi, and H. Zaidi, "Deep learning-based calculation of patient size and attenuation surrogates from localizer image: Toward personalized chest ct protocol optimization," *European Journal of Radiology*, vol. 157, p. 110602, 2022.

[75] J. Wang and D. Fleischmann, "Improving spatial resolution at ct: development, benefits, and pitfalls," pp. 261–262, 2018.

[76] M. J. Willemink and P. B. Noël, "The evolution of image reconstruction for ct—from filtered back projection to artificial intelligence," *European radiology*, vol. 29, pp. 2185–2195, 2019.

[77] R. Singh, W. Wu, G. Wang, and M. K. Kalra, "Artificial intelligence in image reconstruction: the change is here," *Physica Medica*, vol. 79, pp. 113–125, 2020.

[78] S. L. Brady, "Implementation of ai image reconstruction in ct—how is it validated and what dose reductions can be achieved," *The British Journal of Radiology*, vol. 96, no. 1150, p. 20220915, 2023.

[79] N. L. S. T. R. Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, 2011.

[80] W. Xia, H. Shan, G. Wang, and Y. Zhang, "Physics-/model-based and data-driven methods for low-dose computed tomography: A survey," *IEEE signal processing magazine*, vol. 40, no. 2, pp. 89–100, 2023.

[81] J. Greffier, A. Viry, A. Robert, M. Khorsi, and S. Si-Mohamed, "Photon-counting ct systems: A technical review of current clinical possibilities," *Diagnostic and Interventional Imaging*, 2024.

[82] D. Odedra, S. Narayanasamy, S. Sabongui, S. Priya, S. Krishna, and A. Sheikh, "Dual energy ct physics—a primer for the emergency radiologist," *Frontiers in radiology*, vol. 2, p. 820430, 2022.

[83] J. Jeong, A. Wentland, D. Mastrodicasa, G. Fananapazir, A. Wang, I. Banerjee, and B. N. Patel, "Synthetic dual-energy ct reconstruction from single-energy ct using artificial intelligence," *Abdominal Radiology*, vol. 48, no. 11, pp. 3537–3549, 2023.

[84] M. J. Willemink, M. Persson, A. Pourmorteza, N. J. Pelc, and D. Fleischmann, "Photon-counting ct: technical principles and clinical prospects," *Radiology*, vol. 289, no. 2, pp. 293–312, 2018.

[85] X. Yu, Q. Wu, W. Qin, T. Zhong, M. Su, J. Ma, Y. Zhang, X. Ji, G. Quan, Y. Chen *et al.*, "Material decomposition in photon-counting ct: A deep learning approach driven by detector physics and asic modeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2024, pp. 457–466.

[86] O. Díaz, A. Rodríguez-Ruíz, and I. Sechopoulos, "Artificial intelligence for breast cancer detection: Technology, challenges, and prospects," *European journal of radiology*, p. 111457, 2024.

[87] V. Sorin, M. Sklair-Levy, B. S. Glicksberg, E. Konen, G. N. Nadkarni, and E. Klang, "Deep learning for contrast enhanced mammography-a systematic review," *medRxiv*, pp. 2024–05, 2024.

[88] R. M. Mann, A. Athanasiou, P. A. Baltzer, J. Camps-Herrero, P. Clauser, E. M. Fallenberg, G. Forrai, M. H. Fuchsjäger, T. H. Helbich, F. Killburn-Toppin *et al.*, "Breast cancer screening in women with extremely dense breasts recommendations of the european society of breast imaging (eusobi)," *European radiology*, vol. 32, no. 6, pp. 4036–4045, 2022.

[89] P. D. Lopez, "Fluoroscopy history, evolution, and technological advancements: A narrative review," *Journal of Medical Imaging and Radiation Sciences*, 2024.

[90] P. Glielmo, S. Fusco, S. Gitto, G. Zantonelli, D. Albano, C. Messina, L. M. Sconfienza, and G. Mauri, "Artificial intelligence in interventional radiology: state of the art," *European Radiology Experimental*, vol. 8, no. 1, p. 62, 2024.

[91] W. E. Kwok, "Basic principles of and practical guide to clinical mri radiofrequency coils," *RadioGraphics*, vol. 42, no. 3, pp. 898–918, 2022.

[92] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan, *Magnetic resonance imaging: physical principles and sequence design.* John Wiley & Sons, 2014.

[93] J. A. Fessler, "Model-based image reconstruction for mri," *IEEE signal processing magazine*, vol. 27, no. 4, pp. 81–89, 2010.

[94] R. Heckel, M. Jacob, A. Chaudhari, O. Perlman, and E. Shimron, "Deep learning for accelerated and robust mri reconstruction," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 37, no. 3, pp. 335–368, 2024.

[95] S. Kim, H. Park, and S.-H. Park, "A review of deep learning-based reconstruction methods for accelerated mri using spatiotemporal and multi-contrast redundancies," *Biomedical Engineering Letters*, pp. 1–22, 2024.

[96] Z. Lu, J. Wang, Z. Li, S. Ying, J. Wang, J. Shi, and D. Shen, "Two-stage self-supervised cycle-consistency transformer network for reducing slice gap in mr images," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3337–3348, 2023.

[97] R. Shaul, I. David, O. Shitrit, and T. R. Raviv, "Subsampled brain mri reconstruction by generative adversarial neural networks," *Medical Image Analysis*, vol. 65, p. 101747, 2020.

[98] A. Okolie, T. Dirrichs, L. C. Huck, S. Nebelung, S. T. Arasteh, T. Nolte, T. Han, C. K. Kuhl, and D. Truhn, "Accelerating breast mri acquisition with generative ai models," *European Radiology*, pp. 1–9, 2024.

[99] B. Xin, M. Ye, L. Axel, and D. N. Metaxas, "Fill the k-space and refine the image: Prompting for dynamic and multi-contrast mri reconstruction," in *International Workshop on Statistical Atlases and Computational Models of the Heart.* Springer, 2023, pp. 261–273.

[100] W. Peng, L. Feng, G. Zhao, and F. Liu, "Learning optimal k-space acquisition and reconstruction using physics-informed neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 794–20 803.

[101] B. Shah, S. W. Anderson, J. Scalera, H. Jara, and J. A. Soto, "Quantitative mr imaging: physical principles and sequence design in abdominal imaging," *Radiographics*, vol. 31, no. 3, pp. 867–880, 2011.

[102] N. Weiskopf, L. J. Edwards, G. Helms, S. Mohammadi, and E. Kirilina, "Quantitative magnetic resonance imaging of brain anatomy and in vivo histology," *Nature Reviews Physics*, vol. 3, no. 8, pp. 570–588, 2021.

[103] H. Eichhorn, V. Spieker, K. Hammernik, E. Saks, K. Weiss, C. Preibisch, and J. A. Schnabel, "Physics-informed deep learning for motion-corrected reconstruction of quantitative brain mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2024, pp. 562–571.

[104] Y.-D. Xiao, R. Paudel, J. Liu, C. Ma, Z.-S. Zhang, and S.-K. Zhou, "Mri contrast agents: Classification and application," *International journal of molecular medicine*, vol. 38, no. 5, pp. 1319–1326, 2016.

[105] G. Crișan, N. S. Moldovean-Cioroianu, D.-G. Timaru, G. Andrieș, C. Căinap, and V. Chiș, "Radiopharmaceuticals for pet and spect imaging: a literature review over the last decade," *International journal of molecular sciences*, vol. 23, no. 9, p. 5023, 2022.

[106] E. Enlow and S. Abbaszadeh, "State-of-the-art challenges and emerging technologies in radiation detection for nuclear medicine imaging: A review," *Frontiers in Physics*, vol. 11, p. 1106546, 2023.

[107] S. R. Cherry, "Multimodality imaging: Beyond pet/ct and spect/ct," in *Seminars in nuclear medicine*, vol. 39, no. 5. Elsevier, 2009, pp. 348–353.

[108] S. M. Rathmann, Z. Ahmad, S. Slikboer, H. A. Bilton, D. P. Snider, and J. F. Valliant, "The radiopharmaceutical chemistry of technetium-99m," *Radiopharmaceutical chemistry*, pp. 311–333, 2019.

[109] T. E. Peterson and L. R. Furenlid, "Spect detectors: the anger camera and beyond," *Physics in Medicine & Biology*, vol. 56, no. 17, p. R145, 2011.

[110] T. Mannarino, R. Assante, A. D'Antonio, E. Zampella, A. Cuocolo, and W. Acampa, "Radionuclide tracers for myocardial perfusion imaging and blood flow quantification," *Cardiology Clinics*, vol. 41, no. 2, pp. 141–150, 2023.

[111] J. R. Ballinger, "Radiopharmaceuticals in clinical diagnosis and therapy," in *Basic Sciences of Nuclear Medicine.* Springer, 2021, pp. 103–118.

[112] M. Casali, C. Lauri, C. Altini, F. Bertagna, G. Cassarino, A. Cistaro, A. P. Erba, C. Ferrari, C. G. Mainolfi, A. Palucci *et al.*, "State of the art of 18f-fdg pet/ct application in inflammation and infection: a guide for image acquisition and interpretation," *Clinical and Translational Imaging*, vol. 9, no. 4, pp. 299–339, 2021.

[113] H. Tan, Y. Gu, H. Yu, P. Hu, Y. Zhang, W. Mao, and H. Shi, "Total-body pet/ct: current applications and future perspectives," *American Journal of Roentgenology*, vol. 215, no. 2, pp. 325–337, 2020.

[114] D. H. Hagos, R. Battle, and D. B. Rawat, "Recent advances in generative ai and large language models: Current status, challenges, and perspectives," *IEEE Transactions on Artificial Intelligence*, 2024.

[115] S. J. Patey and J. P. Corcoran, "Physics of ultrasound," *Anaesthesia & Intensive Care Medicine*, vol. 22, no. 1, pp. 58–63, 2021.

[116] D. Zander, S. Hüske, B. Hoffmann, X.-W. Cui, Y. Dong, A. Lim, C. Jenssen, A. Löwe, J. B. Koch, and C. F. Dietrich, "Ultrasound image optimization ("knobology"): B-mode," *Ultrasound international open*, vol. 6, no. 01, pp. E14–E24, 2020.

[117] A. Anvari, F. Forsberg, and A. E. Samir, "A primer on the physical principles of tissue harmonic imaging," *Radiographics*, vol. 35, no. 7, pp. 1955–1964, 2015.

[118] H. Yusefi and B. Helfield, "Ultrasound contrast imaging: Fundamentals and emerging technology," *Frontiers in Physics*, vol. 10, p. 791145, 2022.

[119] G. Cloutier, F. Destrempes, F. Yu, and A. Tang, "Quantitative ultrasound imaging of soft biological tissues: a primer for radiologists and medical physicists," *Insights into Imaging*, vol. 12, pp. 1–20, 2021.

[120] H. Hasegawa, "Advances in ultrasonography: Image formation and quality assessment," *Journal of Medical Ultrasonics*, vol. 48, no. 4, pp. 377–389, 2021.

[121] M. Domínguez, Y. Velikova, N. Navab, and M. F. Azampour, "Diffusion as sound propagation: Physics-inspired model for ultrasound image generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2024, pp. 613–623.

[122] B. Huang, F. Yang, M. Yin, X. Mo, and C. Zhong, "A review of multimodal medical image fusion techniques," *Computational and mathematical methods in medicine*, vol. 2020, no. 1, p. 8279342, 2020.

[123] D. Hussain, N. Abbas, and J. Khan, "Recent breakthroughs in pet-ct multimodality imaging: Innovations and clinical impact," *Bioengineering*, vol. 11, no. 12, p. 1213, 2024.

[124] F. S. Furtado, M. Hesami, S. Mcdermott, H. Kulkarni, A. Herold, and O. A. Catalano, "The synergistic effect of pet/mri in whole-body oncologic imaging: an expert review," *Clinical and Translational Imaging*, vol. 11, no. 4, pp. 351–364, 2023.

[125] T. Beyer and B. Pichler, "A decade of combined imaging: from a pet attached to a ct to a pet inside an mr," *European journal of nuclear medicine and molecular imaging*, vol. 36, pp. 1–2, 2009.

[126] V. P. Sudarshan, U. Upadhyay, G. F. Egan, Z. Chen, and S. P. Awate, "Towards lower-dose pet using physics-based uncertainty-aware multimodal learning with robustness to out-of-distribution data," *Medical Image Analysis*, vol. 73, p. 102187, 2021.

[127] N. Liang, "Medical image fusion with deep neural networks," *Scientific Reports*, vol. 14, no. 1, p. 7972, 2024.

[128] W. Zhao, T. Lv, Y. Chen, and L. Xing, "Dual-energy ct imaging using a single-energy ct data via deep learning: A contrast-enhanced ct study," *International Journal of Radiation Oncology, Biology, Physics*, vol. 108, no. 3, p. S43, 2020.

[129] Z. Li, Y. Long, and I. Y. Chun, "An improved iterative neural network for high-quality image-domain material decomposition in dual-energy ct," *Medical physics*, vol. 50, no. 4, pp. 2195–2211, 2023.

[130] A. Khorasani, N. Dadashi serej, M. Jalilian, A. Shayganfar, and M. B. Tavakoli, "Performance comparison of different medical image fusion algorithms for clinical glioma grade classification with advanced magnetic resonance imaging (mri)," *Scientific Reports*, vol. 13, no. 1, p. 17646, 2023.

[131] J. F. Barrett and N. Keat, "Artifacts in ct: recognition and avoidance," *Radiographics*, vol. 24, no. 6, pp. 1679–1691, 2004.

[132] Y. Xu, S. Hu, and Y. Du, "Research on optimization scheme for blocking artifacts after patch-based medical image reconstruction," *Computational and Mathematical Methods in Medicine*, vol. 2022, no. 1, p. 2177159, 2022.

[133] O. Lang, D. Yaya-Stupp, I. Traynis, H. Cole-Lewis, C. R. Bennett, C. R. Lyles, C. Lau, M. Irani, C. Semturs, D. R. Webster *et al.*, "Using generative ai to investigate medical imagery models and datasets," *EBioMedicine*, vol. 102, 2024.

[134] M. Amirian, D. Barco, I. Herzig, and F.-P. Schilling, "Artifact reduction in 3d and 4d cone-beam computed tomography images with deep learning-a review," *IEEE Access*, 2024.

[135] F. Shomal Zadeh, A. Pooyan, E. Alipour, N. Hosseini, P. C. Thurlow, F. Del Grande, M. Shafiei, and M. Chalian, "Dynamic contrast-enhanced magnetic resonance imaging (dce-mri) in differentiation of soft tissue sarcoma from benign lesions: a systematic review of literature," *Skeletal Radiology*, pp. 1–15, 2024.

[136] Y. Lee, W.-H. Jee, Y. S. Whang, C. K. Jung, Y.-G. Chung, and S.-Y. Lee, "Benign versus malignant soft-tissue tumors: differentiation with 3t magnetic resonance image textural analysis including diffusion-weighted imaging," *Investigative Magnetic Resonance Imaging*, vol. 25, no. 2, pp. 118–128, 2021.

[137] S. Joutard, M. Pietsch, and R. Prevost, "Hyperspace: Hypernetworks for spacing-adaptive image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2024, pp. 339–349.

[138] Y. Liu, R. Li, Y. Fan, Đ. Antonijević, P. Milenković, Z. Li, M. Djuric, and Y. Fan, "The influence of anisotropic voxel caused by field of view setting on the accuracy of three-dimensional reconstruction of bone geometric models," *AIP Advances*, vol. 8, no. 8, 2018.

[139] R. P. Cabeen, M. E. Bastin, and D. H. Laidlaw, "A comparative evaluation of voxel-based spatial mapping in diffusion tensor imaging," *Neuroimage*, vol. 146, pp. 100–112, 2017.

[140] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Frontiers in Public Health*, vol. 11, p. 1273253, 2023.

[141] D. Qiu, Y. Cheng, and X. Wang, "Medical image super-resolution reconstruction algorithms based on deep learning: A survey," *Computer Methods and Programs in Biomedicine*, vol. 238, p. 107590, 2023.

[142] Z. Chen, K. Pawar, M. Ekanayake, C. Pain, S. Zhong, and G. F. Egan, "Deep learning for image enhancement and correction in magnetic resonance imaging—state-of-the-art and challenges," *Journal of Digital Imaging*, vol. 36, no. 1, pp. 204–230, 2023.

[143] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[144] Á. Planchuelo-Gómez, M. Descoteaux, H. Larochelle, J. Hutter, D. K. Jones, and C. M. Tax, "Optimisation of quantitative brain diffusion-relaxation mri acquisition protocols with physics-informed machine learning," *Medical Image Analysis*, vol. 94, p. 103134, 2024.

[145] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.

[146] C. Banerjee, K. Nguyen, C. Fookes, and K. George, "Physics-informed computer vision: A review and perspectives," *ACM Computing Surveys*, vol. 57, no. 1, pp. 1–38, 2024.

[147] C. Banerjee, K. Nguyen, O. Salvado, T. Tran, and C. Fookes, "Pinns for medical image analysis: A survey," *arXiv preprint arXiv:2408.01026*, 2024.

[148] E. Moya-Sáez, Ó. Peña-Nogales, R. de Luis-García, and C. Alberola-López, "A deep learning approach for synthetic mri based on two routine sequences and training with synthetic data," *Computer Methods and Programs in Biomedicine*, vol. 210, p. 106371, 2021.

[149] L. R. Koetzier, J. Wu, D. Mastrodicasa, A. Lutz, M. Chung, W. A. Koszek, J. Pratap, A. S. Chaudhari, P. Rajpurkar, M. P. Lungren *et al.*, "Generating synthetic data for medical imaging," *Radiology*, vol. 312, no. 3, p. e232471, 2024.

[150] L. D. McClenny and U. M. Braga-Neto, "Self-adaptive physics-informed neural networks," *Journal of Computational Physics*, vol. 474, p. 111722, 2023.

[151] C. Sirocchi, A. Bogliolo, and S. Montagna, "Medical-informed machine learning: integrating prior knowledge into medical decision systems," *BMC Medical Informatics and Decision Making*, vol. 24, no. Suppl 4, p. 186, 2024.

[152] F. Leiser, S. Rank, M. Schmidt-Kraepelin, S. Thiebes, and A. Sunyaev, "Medical informed machine learning: A scoping review and future research directions," *Artificial Intelligence in Medicine*, vol. 145, p. 102676, 2023.

[153] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy *et al.*, "Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 614–633, 2021.

[154] E. Juralewicz and U. Markowska-Kaczmar, "Capsule network versus convolutional neural network in image classification: comparative analysis," in *International Conference on Computational Science.* Springer, 2021, pp. 17–30.

[155] H. Xu, L. Xiang, H. Ye, D. Yao, P. Chu, and B. Li, "Permutation equivariance of transformers and its applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5987–5996.

[156] D. A. Selby, M. Sprang, J. Ewald, and S. J. Vollmer, "Beyond the black box with biologically informed neural networks," *Nature Reviews Genetics*, pp. 1–2, 2025.

[157] E. Hartman, A. M. Scott, C. Karlsson, T. Mohanty, S. T. Vaara, A. Linder, L. Malmström, and J. Malmström, "Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis," *Nature Communications*, vol. 14, no. 1, p. 5359, 2023.

[158] T. S. Toh, F. Dondelinger, and D. Wang, "Looking beyond the hype: applied ai and machine learning in translational medicine," *EBioMedicine*, vol. 47, pp. 607–615, 2019.

[159] B. S. Kelly, C. Judge, S. M. Bollard, S. M. Clifford, G. M. Healy, A. Aziz, P. Mathur, S. Islam, K. W. Yeom, A. Lawlor *et al.*, "Radiology artificial intelligence: a systematic review and evaluation of methods (raise)," *European radiology*, vol. 32, no. 11, pp. 7998–8007, 2022.

[160] H. Hricak, M. Abdel-Wahab, R. Atun, M. M. Lette, D. Paez, J. A. Brink, L. Donoso-Bach, G. Frija, M. Hierath, O. Holmberg *et al.*, "Medical imaging and nuclear medicine: a lancet oncology commission," *The Lancet Oncology*, vol. 22, no. 4, pp. e136–e172, 2021.

[161] N. Elshafeey, A. Kotrotsou, A. Hassan, N. Elshafei, I. Hassan, S. Ahmed, S. Abrol, A. Agarwal, K. El Salek, S. Bergamaschi *et al.*, "Multicenter study demonstrates radiomic features derived from magnetic resonance perfusion images identify pseudoprogression in glioblastoma," *Nature communications*, vol. 10, no. 1, p. 3170, 2019.

[162] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy ai: From principles to practices," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–46, 2023.

[163] K. Lekadir, A. F. Frangi, A. R. Porras, B. Glocker, C. Cintas, C. P. Langlotz, E. Weicken, F. W. Asselbergs, F. Prior, G. S. Collins *et al.*, "Future-ai: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *bmj*, vol. 388, 2025.

[164] C. Scapicchio, M. Gabelloni, A. Barucci, D. Cioni, L. Saba, and E. Neri, "A deep look into radiomics," *La radiologia medica*, vol. 126, no. 10, pp. 1296–1311, 2021.

[165] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali, "From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 132–160, 2019.

[166] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications*, vol. 5, no. 1, p. 4006, 2014.

[167] N. Chen, R. Li, M. Jiang, Y. Guo, J. Chen, D. Sun, L. Wang, and X. Yao, "Progression-free survival prediction in small cell lung cancer based on radiomics analysis of contrast-enhanced ct," *Frontiers in Medicine*, vol. 9, p. 833283, 2022.

[168] J. M. Murray, B. Wiegand, B. Hadaschik, K. Herrmann, and J. Kleesiek, "Virtual biopsy: just an ai software or a medical procedure?" *Journal of Nuclear Medicine*, vol. 63, no. 4, p. 511, 2022.

[169] L. J. Grimm and M. A. Mazurowski, "Breast cancer radiogenomics: current status and future directions," *Academic Radiology*, vol. 27, no. 1, pp. 39–46, 2020.

[170] M. R. Tomaszewski and R. J. Gillies, "The biological meaning of radiomic features," *Radiology*, vol. 298, no. 3, pp. 505–516, 2021.

[171] J. E. Van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—"how-to" guide and critical reflection," *Insights into imaging*, vol. 11, no. 1, pp. 1–16, 2020.

[172] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature biomedical engineering*, vol. 2, no. 3, pp. 158–164, 2018.

[173] L. Fournier, L. Costaridou, L. Bidaut, N. Michoux, F. E. Lecouvet, L.-F. de Geus-Oei, R. Boellaard, D. E. Oprea-Lager, N. A. Obuchowski, A. Caroli *et al.*, "Incorporating radiomics into clinical trials: expert consensus endorsed by the european society of radiology on considerations for data-driven compared to biologically driven quantitative biomarkers," *European radiology*, vol. 31, pp. 6001–6012, 2021.

[174] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[175] L. Hadjiiski, K. Cha, H.-P. Chan, K. Drukker, L. Morra, J. J. Näppi, B. Sahiner, H. Yoshida, Q. Chen, T. M. Deserno *et al.*, "Aapm task group report 273: Recommendations on best practices for ai and machine learning for computer-aided diagnosis in medical imaging," *Medical Physics*, vol. 50, no. 2, pp. e1–e24, 2023.

[176] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi, "Radiomics: the facts and the challenges of image analysis," *European radiology experimental*, vol. 2, no. 1, pp. 1–8, 2018.

[177] F. Orlhac, F. Frouin, C. Nioche, N. Ayache, and I. Buvat, “Validation of a method to compensate multicenter effects affecting ct radiomics,” *Radiology*, vol. 291, no. 1, pp. 53–59, 2019.

[178] J. W. Son, J. Y. Hong, Y. Kim, W. J. Kim, D.-Y. Shin, H.-S. Choi, S. H. Bak, and K. M. Moon, “How many private data are needed for deep learning in lung nodule detection on ct scans? a retrospective multicenter study,” *Cancers*, vol. 14, no. 13, p. 3174, 2022.

[179] Y. Soleymani, Z. Valibeiglou, M. F. Ghaziani, A. Jahanshahi, and D. Khezerloo, “Radiomics reproducibility in computed tomography through changes of roi size, resolution, and hounsfield unit: A phantom study,” *Radiography*, vol. 30, no. 6, pp. 1629–1636, 2024.

[180] M. Vallieres, D. Visvikis, and M. Hatt, “Dependency of a validated radiomics signature on tumor volume and potential corrections,” 2018.

[181] J.-P. Fortin, D. Parker, B. Tunç, T. Watanabe, M. A. Elliott, K. Ruparel, D. R. Roalf, T. D. Satterthwaite, R. C. Gur, R. E. Gur *et al.*, “Harmonization of multi-site diffusion tensor imaging data,” *Neuroimage*, vol. 161, pp. 149–170, 2017.

[182] R. F. Cabini, F. Brero, A. Lancia, C. Stelitano, O. Oneta, E. Ballante, E. Puppo, M. Mariani, E. Alì, V. Bartolomeo *et al.*, “Preliminary report on harmonization of features extraction process using the combat tool in the multi-center “blue sky radiomics” study on stage iii unresectable nsclc,” *Insights into Imaging*, vol. 13, no. 1, p. 38, 2022.

[183] M. Ligero, O. Jordi-Ollero, K. Bernatowicz, A. Garcia-Ruiz, E. Delgado-Muñoz, D. Leiva, R. Mast, C. Suarez, R. Sala-Llonch, N. Calvo *et al.*, “Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis,” *European radiology*, vol. 31, pp. 1460–1470, 2021.

[184] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical bayes methods,” *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[185] F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, and I. Buvat, “A postreconstruction harmonization method for multicenter radiomic studies in pet,” *Journal of Nuclear Medicine*, vol. 59, no. 8, pp. 1321–1328, 2018.

[186] R. Mahon, M. Ghita, G. D. Hugo, and E. Weiss, “Combat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets,” *Physics in Medicine & Biology*, vol. 65, no. 1, p. 015010, 2020.

[187] D. Leithner, H. Schöder, A. Haug, H. A. Vargas, P. Gibbs, I. Häggström, I. Rausch, M. Weber, A. S. Becker, J. Schwartz *et al.*, “Impact of combat harmonization on pet radiomics-based tissue classification: a dual-center pet/mri and pet/ct study,” *Journal of Nuclear Medicine*, vol. 63, no. 10, pp. 1611–1616, 2022.

[188] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard *et al.*, “The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping,” *Radiology*, vol. 295, no. 2, pp. 328–338, 2020.

[189] T. E. Yankeelov, “The quantitative imaging network: a decade of achievement,” p. 0, 2019.

[190] A. R. Guimaraes, “Quantitative imaging biomarker alliance (qiba): Protocols and profiles,” *Quantitative Imaging in Medicine: Background and Basics*, 2021.

[191] N. M. deSouza, E. Achten, A. Alberich-Bayarri, F. Bamberg, R. Boellaard, O. Clément, L. Fournier, F. Gallagher, X. Golay, C. P. Heussel *et al.*, “Validated imaging biomarkers as decision-making tools in clinical trials and routine practice: current status and recommendations from the eiball* subcommittee of the european society of radiology (esr),” *Insights into imaging*, vol. 10, no. 1, pp. 1–16, 2019.

[192] M. L. Welch, C. McIntosh, B. Haibe-Kains, M. F. Milosevic, L. Wee, A. Dekker, S. H. Huang, T. G. Purdie, B. O’Sullivan, H. J. Aerts *et al.*, “Vulnerabilities of radiomic signature development: The need for safeguards,” *Radiotherapy and Oncology*, vol. 130, pp. 2–9, 2019.

[193] A. Traverso, M. Kazmierski, I. Zhovannik, M. Welch, L. Wee, D. Jaffray, A. Dekker, and A. Hope, “Machine learning helps identifying volume-confounding effects in radiomics,” *Physica Medica*, vol. 71, pp. 24–30, 2020.

[194] L. Lu, F. S. Ahmed, O. Akin, L. Luk, X. Guo, H. Yang, J. Yoon, A. A. Hakimi, L. H. Schwartz, and B. Zhao, “Uncontrolled confounders may lead to false or overvalued radiomics signature: a proof of concept using survival analysis in a multicenter cohort of kidney cancer,” *Frontiers in Oncology*, vol. 11, p. 638185, 2021.

[195] Z. Shi, I. Zhovannik, A. Traverso, F. J. Dankers, T. M. Deist, P. Kalendralis, R. Monshouwer, J. Bussink, R. Fijten, H. J. Aerts *et al.*, “Distributed radiomics as a signature validation study using the personal health train infrastructure,” *Scientific data*, vol. 6, no. 1, p. 218, 2019.

[196] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[197] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj

[198] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, “Adversarial attacks on medical machine learning,” *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[199] A. Ziller, D. Usynin, R. Braren, M. Makowski, D. Rueckert, and G. Kaissis, “Medical imaging deep learning with differential privacy,” *Scientific Reports*, vol. 11, no. 1, p. 13524, 2021.

[200] E. Wu, K. Wu, R. Daneshjou, D. Ouyang, D. E. Ho, and J. Zou, “How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals,” *Nature Medicine*, vol. 27, no. 4, pp. 582–584, 2021.

[201] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman *et al.*, “The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans,” *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[202] J. Wang, N. Sourlos, S. Zheng, N. van der Velden, G. J. Pelgrim, R. Vliegenthart, and P. van Ooijen, "Preparing ct imaging datasets for deep learning in lung nodule analysis: Insights from four well-known datasets," *Heliyon*, 2023.

[203] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[204] F. Prior, K. Smith, A. Sharma, J. Kirby, L. Tarbox, K. Clark, W. Bennett, T. Nolan, and J. Freymann, "The public cancer radiology imaging collections of the cancer imaging archive," *Scientific data*, vol. 4, no. 1, pp. 1–7, 2017.

[205] K. J. Gorgolewski, G. Varoquaux, G. Rivera, Y. Schwarz, S. S. Ghosh, C. Maumet, V. V. Sochat, T. E. Nichols, R. A. Poldrack, J.-B. Poline *et al.*, "Neurovault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain," *Frontiers in neuroinformatics*, vol. 9, p. 8, 2015.

[206] R. Informatics, "RadLex," https://radlex.org/ Accessed 15 Feb 2023, 2016.

[207] J. M. Nobel, E. M. Kok, and S. G. Robben, "Redefining the structure of structured reporting in radiology," *Insights into imaging*, vol. 11, pp. 1–5, 2020.

[208] A. Carré, G. Klausner, M. Edjlali, M. Lerousseau, J. Briend-Diop, R. Sun, S. Ammari, S. Reuzé, E. Alvarez Andres, T. Estienne *et al.*, "Standardization of brain mr images across machines and protocols: bridging the gap for mri-based radiomics," *Scientific reports*, vol. 10, no. 1, p. 12340, 2020.

[209] P. S. Sharma and A. M. Saindane, "Standardizing magnetic resonance imaging protocols across a large radiology enterprise: barriers and solutions," *Current Problems in Diagnostic Radiology*, vol. 49, no. 5, pp. 312–316, 2020.

[210] Y. Wang, P. Chu, T. P. Szczykutowicz, C. Stewart, and R. Smith-Bindman, "Ct acquisition parameter selection in the real world: impacts on radiation dose and variation amongst 155 institutions," *European Radiology*, pp. 1–9, 2023.

[211] C. McCollough and S. Leng, "Use of artificial intelligence in computed tomography dose optimisation," *Annals of the ICRP*, vol. 49, no. 1_suppl, pp. 113–125, 2020.

[212] A. Midya, J. Chakraborty, M. Gönen, R. K. Do, and A. L. Simpson, "Influence of ct acquisition and reconstruction parameters on radiomic feature reproducibility," *Journal of Medical Imaging*, vol. 5, no. 1, pp. 011 020–011 020, 2018.

[213] R. Lacson, M. Eskian, A. Licaros, N. Kapoor, and R. Khorasani, "Machine learning model drift: predicting diagnostic imaging follow-up as a case example," *Journal of the American College of Radiology*, vol. 19, no. 10, pp. 1162–1169, 2022.

[214] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[215] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1424, 2021.

[216] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[217] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[218] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–42, 2023.

[219] R. Marcinkevičs and J. E. Vogt, "Interpretable and explainable machine learning: A methods-centric overview with concrete examples," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 3, p. e1493, 2023.

[220] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa, "Explainable ai in medical imaging: An overview for clinical practitioners–saliency-based xai approaches," *European journal of radiology*, p. 110787, 2023.

[221] M. Sahakyan, Z. Aung, and T. Rahwan, "Explainable artificial intelligence for tabular data: A survey," *IEEE access*, vol. 9, pp. 135 392–135 422, 2021.

[222] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information fusion*, vol. 99, p. 101805, 2023.

[223] European Commission, "Ethics Guidelines for Trustworthy AI," 2019, retrieved March 5, 2025. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[224] J. Bernett, D. B. Blumenthal, D. G. Grimm, F. Haselbeck, R. Joeres, O. V. Kalinina, and M. List, "Guiding questions to avoid data leakage in biological machine learning applications," *Nature Methods*, vol. 21, no. 8, pp. 1444–1453, 2024.

[225] A. AbuHalimeh, "Improving data quality in clinical research informatics tools," *Frontiers in Big Data*, vol. 5, p. 871897, 2022.

[226] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 2023.

[227] A. P. Sanner, N. F. Grauhan, M. A. Brockmann, A. E. Othman, and A. Mukhopadhyay, "Voxel scene graph for intracranial hemorrhage," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 519–529.

[228] R. LaLonde, D. Torigian, and U. Bagci, "Encoding visual attributes in capsules for explainable medical diagnoses," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 294–304.

[229] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot *et al.*, "Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation," *Medical image analysis*, vol. 63, p. 101694, 2020.

[230] W. Khan, S. Leem, K. B. See, J. K. Wong, S. Zhang, and R. Fang, "A comprehensive survey of foundation models in medicine," *IEEE Reviews in Biomedical Engineering*, 2025.

[231] C. Fang, C. Sandino, B. Mahasseni, J. Minxha, H. Pouransari, E. Azemi, A. Moin, and E. Zippi, "Promoting cross-modal representations to improve multimodal foundation models for physiological signals," *arXiv preprint arXiv:2410.16424*, 2024.

[232] S. Seoni, A. Shahini, K. M. Meiburger, F. Marzola, G. Rotunno, U. R. Acharya, F. Molinari, and M. Salvi, "All you need is data preparation: A systematic review of image harmonization techniques in multi-center/device studies for medical support systems," *Computer Methods and Programs in Biomedicine*, p. 108200, 2024.

[233] S. Masoudi, S. A. Harmon, S. Mehralivand, S. M. Walker, H. Raviprakash, U. Bagci, P. L. Choyke, and B. Turkbey, "Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research," *Journal of Medical Imaging*, vol. 8, no. 1, pp. 010 901–010 901, 2021.

[234] S. Pereira, R. Meier, V. Alves, M. Reyes, and C. A. Silva, "Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and IMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1*. Springer, 2018, pp. 106–114.

[235] P. Afshar, A. Oikonomou, F. Naderkhani, P. N. Tyrrell, K. N. Plataniotis, K. Farahani, and A. Mohammadi, "3d-mcn: a 3d multi-scale capsule network for lung nodule malignancy prediction," *Scientific reports*, vol. 10, no. 1, p. 7948, 2020.

[236] M. Aiello, G. Esposito, G. Pagliari, P. Borrelli, V. Brancato, and M. Salvatore, "How does dicom support big data management? investigating its use in medical imaging community," *Insights into Imaging*, vol. 12, no. 1, p. 164, 2021.

[237] Y. Wang, F. Ye, Y. Chen, C. Wang, C. Wu, F. Xu, Z. Ma, Y. Liu, Y. Zhang, M. Cao *et al.*, "A multi-modal dental dataset for semi-supervised deep learning image segmentation," *Scientific Data*, vol. 12, no. 1, p. 117, 2025.

[238] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[239] K. Yan, X. Wang, L. Lu, and R. M. Summers, "Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of medical imaging*, vol. 5, no. 3, pp. 036 501–036 501, 2018.

[240] A. Jiménez-Sánchez, N.-R. Avlona, S. de Boer, V. M. Campello, A. Feragen, E. Ferrante, M. Ganz, J. W. Gichoya, C. González, S. Groefsema *et al.*, "In the picture: Medical imaging datasets, artifacts, and their living review," *arXiv preprint arXiv:2501.10727*, 2025.

[241] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, “Multi-omics data integration, interpretation, and its application,” *Bioinformatics and biology insights*, vol. 14, p. 1177932219899051, 2020.

[242] Y. Ramakrishnaiah, N. Macesic, G. I. Webb, A. Y. Peleg, and S. Tyagi, “Ehr-qc: A streamlined pipeline for automated electronic health records standardisation and preprocessing to predict clinical outcomes,” *Journal of Biomedical Informatics*, vol. 147, p. 104509, 2023.

[243] W. K. Michener, “Ten simple rules for creating a good data management plan,” *PLoS computational biology*, vol. 11, no. 10, p. e1004525, 2015.

[244] M. Johns, T. Meurers, F. N. Wirth, A. C. Haber, A. Müller, M. Halilovic, F. Balzer, and F. Prasser, “Data provenance in biomedical research: scoping review,” *Journal of medical Internet research*, vol. 25, p. e42289, 2023.

[245] G. Fraga-González, H. van de Wiel, F. Garassino, W. Kuo, D. de Zélicourt, V. Kurtcuoglu, L. Held, and E. Furrer, “Affording reusable data: recommendations for researchers from a data-intensive project,” *Scientific Data*, vol. 12, no. 1, p. 258, 2025.

[246] G. Vandewiele, I. Dehaene, G. Kovács, L. Sterckx, O. Janssens, F. Ongenae, F. De Backere, F. De Turck, K. Roelens, J. Decruyenaere *et al.*, “Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling,” *Artificial Intelligence in Medicine*, vol. 111, p. 101987, 2021.

[247] M. Akhtar, O. Benjelloun, C. Conforti, L. Foschini, J. Giner-Miguelez, P. Gijsbers, S. Goswami, N. Jain, M. Karamousadakis, M. Kuchnik *et al.*, “Croissant: A metadata format for ml-ready datasets,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 82 133–82 148, 2024.

[248] B. Sahiner, W. Chen, R. K. Samala, and N. Petrick, “Data drift in medical machine learning: implications and potential remedies,” *The British Journal of Radiology*, vol. 96, no. 1150, p. 20220878, 2023.

[249] S. Kapoor, E. M. Cantrell, K. Peng, T. H. Pham, C. A. Bail, O. E. Gundersen, J. M. Hofman, J. Hullman, M. A. Lones, M. M. Malik *et al.*, “Reforms: Consensus-based recommendations for machine-learning-based science,” *Science Advances*, vol. 10, no. 18, p. eadk3452, 2024.

[250] J. Witsch, B. Siegerink, C. H. Nolte, M. Sprügel, T. Steiner, M. Endres, and H. B. Huttner, “Prognostication after intracerebral hemorrhage: a review,” *Neurological Research and Practice*, vol. 3, pp. 1–14, 2021.

[251] T. Gregorio, S. Pipa, P. Cavaleiro, G. Atanasio, I. Albuquerque, P. C. Chaves, and L. Azevedo, “Assessment and comparison of the four most extensively validated prognostic scales for intracerebral hemorrhage: systematic review with meta-analysis,” *Neurocritical Care*, vol. 30, pp. 449–466, 2019.

[252] M. Bahrami, M. Keyhanifard, and M. Afzali, “Spontaneous intracerebral hemorrhage, initial computed tomography (ct) scan findings, clinical manifestations and possible risk factors,” *American Journal of Nuclear Medicine and Molecular Imaging*, vol. 12, no. 3, p. 106, 2022.

[253] J. Magid-Bernstein, R. Girard, S. Polster, A. Srinath, S. Romanos, I. A. Awad, and L. H. Sansing, "Cerebral hemorrhage: pathophysiology, treatment, and future directions," *Circulation research*, vol. 130, no. 8, pp. 1204–1229, 2022.

[254] M. C. T. Predispose, "Unmet needs and challenges in clinical research of intracerebral hemorrhage," *Stroke*, vol. 49, pp. 00–00, 2018.

[255] A. Perez del Barrio, A. S. Esteve Domínguez, P. Menéndez Fernández-Miranda, P. Sanz Bellón, D. Rodríguez González, L. Lloret Iglesias, E. Marques Fraguela, A. A. González Mandly, and J. A. Vega, "A deep learning model for prognosis prediction after intracranial hemorrhage," *Journal of Neuroimaging*, vol. 33, no. 2, pp. 218–226, 2023.

[256] J. Wang, H. Zhu, S.-H. Wang, and Y.-D. Zhang, "A review of deep learning on medical image analysis," *Mobile Networks and Applications*, vol. 26, pp. 351–380, 2021.

[257] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *NPJ digital medicine*, vol. 3, no. 1, p. 136, 2020.

[258] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical Image Analysis*, vol. 69, p. 101985, 2021.

[259] W. Tang, Z. Yang, and Y. Song, "Disease-grading networks with ordinal regularization for medical imaging," *Neurocomputing*, vol. 545, p. 126245, 2023.

[260] J. Barbero-Gómez, P.-A. Gutiérrez, V.-M. Vargas, J.-A. Vallejo-Casas, and C. Hervás-Martínez, "An ordinal cnn approach for the assessment of neurological damage in parkinson's disease patients," *Expert Systems with Applications*, vol. 182, p. 115271, 2021.

[261] X. Shan, X. Li, R. Ge, S. Wu, A. Elazab, J. Zhu, L. Zhang, G. Jia, Q. Xiao, X. Wan *et al.*, "Gcs-ichnet: Assessment of intracerebral hemorrhage prognosis using self-attention with domain knowledge integration," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 2217–2222.

[262] S. Jain and L. M. Iverson, "Glasgow coma scale," 2018.

[263] W. Ma, C. Chen, J. Abrigo, C. H.-K. Mak, Y. Gong, N. Y. Chan, C. Han, Z. Liu, and Q. Dou, "Treatment outcome prediction for intracerebral hemorrhage via generative prognostic model with imaging and tabular data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 715–725.

[264] Y.-J. Zhou, W. Liu, Y. Gao, J. Xu, L. Lu, Y. Duan, H. Cheng, N. Jin, X. Man, S. Zhao *et al.*, "A novel multi-task model imitating dermatologists for accurate differential diagnosis of skin diseases in clinical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 202–212.

[265] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "Synthstrip: Skull-stripping for any brain image," *NeuroImage*, vol. 260, p. 119474, 2022.

[266] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett,

Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[267] G. Li, M. Zhang, J. Li, F. Lv, and G. Tong, “Efficient densely connected convolutional neural networks,” *Pattern Recognition*, vol. 109, p. 107610, 2021.

[268] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[269] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.

[270] W. Silva, J. R. Pinto, and J. S. Cardoso, “A uniform performance index for ordinal classification with imbalanced classes,” in *2018 international joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[271] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[272] R. Artstein and M. Poesio, “Inter-coder agreement for computational linguistics,” *Computational linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[273] K. Gotkowski, C. Gonzalez, A. Bucher, and A. Mukhopadhyay, “M3d-cam: A pytorch library to generate 3d attention maps for medical deep learning,” in *Bildverarbeitung für die Medizin 2021: Proceedings, German Workshop on Medical Image Computing, Regensburg, March 7-9, 2021*. Springer, 2021, pp. 217–222.

[274] J. C. Hemphill III, D. C. Bonovich, L. Besmertis, G. T. Manley, and S. C. Johnston, “The ich score: a simple, reliable grading scale for intracerebral hemorrhage,” *Stroke*, vol. 32, no. 4, pp. 891–897, 2001.

[275] E. Serrano, A. López-Rueda, J. Moreno, A. Rodríguez, L. Llull, C. Zwanzger, L. Oleaga, and S. Amaro, “The new hematoma maturity score is highly associated with poor clinical outcome in spontaneous intracerebral hemorrhage,” *European Radiology*, vol. 32, pp. 290–299, 2022.

[276] A. Mata Agudo, “Dynamic spot sign predicts hematoma expansion in acute intraparenchymatous hemorrhage: a perfusion ct study,” 2016.

[277] H. Nakaguchi, T. Tanishima, and N. Yoshimasu, “Factors in the natural history of chronic subdural hematomas that influence their postoperative recurrence,” *Journal of neurosurgery*, vol. 95, no. 2, pp. 256–262, 2001.

[278] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren, “Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection,” *Scientific reports*, vol. 10, no. 1, p. 22147, 2020.

[279] T. Gregório, S. Pipa, P. Cavaleiro, G. Atanásio, I. Albuquerque, P. C. Chaves, and L. Azevedo, "Prognostic models for intracerebral hemorrhage: systematic review and meta-analysis," *BMC Medical Research Methodology*, vol. 18, pp. 1–17, 2018.

[280] H. M. la Roi-Teeuw, F. S. van Royen, A. de Hond, A. Zahra, S. de Vries, R. Bartels, A. J. Carriero, S. van Doorn, Z. S. Dunias, I. Kant *et al.*, "Don't be misled: Three misconceptions about external validation of clinical prediction models," *Journal of Clinical Epidemiology*, p. 111387, 2024.

[281] S. Sugeir and S. Naylor, "Critical care and personalized or precision medicine: who needs whom?" *Journal of Critical Care*, vol. 43, pp. 401–405, 2018.

[282] A. Gorini and G. Pravettoni, "P5 medicine: a plus for a personalized approach to oncology," *Nature Reviews Clinical Oncology*, vol. 8, no. 7, pp. 444–444, 2011.

[283] S. Schleidgen, C. Klingler, T. Bertram, W. H. Rogowski, and G. Marckmann, "What is personalized medicine: sharpening a vague term based on a systematic literature review," *BMC medical ethics*, vol. 14, pp. 1–12, 2013.

[284] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[285] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA: a cancer journal for clinicians*, vol. 74, no. 1, pp. 12–49, 2024.

[286] T. B. Kratzer, P. Bandi, N. D. Freedman, R. A. Smith, W. D. Travis, A. Jemal, and R. L. Siegel, "Lung cancer statistics, 2023," *Cancer*, vol. 130, no. 8, pp. 1330–1348, 2024.

[287] H. J. de Koning, C. M. van Der Aalst, P. A. de Jong, E. T. Scholten, K. Nackaerts, M. A. Heuvelmans, J.-W. J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg *et al.*, "Reduced lung-cancer mortality with volume ct screening in a randomized trial," *New England journal of medicine*, vol. 382, no. 6, pp. 503–513, 2020.

[288] C. I. Henschke, R. Yip, D. Shaham, S. Markowitz, J. Cervera Deval, J. J. Zulueta, L. M. Seijo, C. Aylesworth, K. Klingler, S. Andaz *et al.*, "A 20-year follow-up of the international early lung cancer action program (i-elcap)," *Radiology*, vol. 309, no. 2, p. e231988, 2023.

[289] Centers for Medicare & Medicaid Services, "Screening for lung cancer with low dose computed tomography (ldct)," 2025, accessed March 22, 2025. [Online]. Available: https://www.cms.gov/medicare-coverage-database/view/ncacal-decision-memo.aspx?proposed=N&NCAId=304

[290] X. Feng, P. Goodley, K. Alcala, F. Guida, R. Kaaks, R. Vermeulen, G. S. Downward, C. Bonet, S. M. Colorado-Yohar, D. Albanes *et al.*, "Evaluation of risk prediction models to select lung cancer screening participants in europe: a prospective cohort consortium analysis," *The Lancet Digital Health*, vol. 6, no. 9, pp. e614–e624, 2024.

[291] P. P. Massion and R. C. Walker, "Indeterminate pulmonary nodules: risk for having or for developing lung cancer?" *Cancer prevention research*, vol. 7, no. 12, pp. 1173–1178, 2014.

[292] T. L. Leong, A. McWilliams, and G. M. Wright, "Incidental pulmonary nodules: an opportunity to complement lung cancer screening," *Journal of Thoracic Oncology*, vol. 19, no. 4, pp. 522–524, 2024.

[293] J. Christensen, A. E. Prosper, C. C. Wu, J. Chung, E. Lee, B. Elicker, A. R. Hunsaker, M. Petranovic, K. L. Sandler, B. Stiles *et al.*, "Acr lung-rads v2022: assessment categories and management recommendations," *Journal of the American College of Radiology*, vol. 21, no. 3, pp. 473–488, 2024.

[294] C. E. Rydzak, S. G. Armato, R. S. Avila, J. L. Mulshine, D. F. Yankelevitz, and D. S. Gierada, "Quality assurance and quantitative imaging biomarkers in low-dose ct lung cancer screening," *The British journal of radiology*, vol. 91, no. 1090, p. 20170401, 2018.

[295] R. S. Avila, K. Krishnan, N. Obuchowski, A. Jirapatnakul, R. Subramaniam, and D. Yankelevitz, "Calibration phantom-based prediction of ct lung nodule volume measurement performance," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 9, p. 6193, 2023.

[296] C. Wang, J. Shao, Y. He, J. Wu, X. Liu, L. Yang, Y. Wei, X. S. Zhou, Y. Zhan, F. Shi *et al.*, "Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography," *Nature Medicine*, pp. 1–12, 2024.

[297] K. Chen, Y. Nie, S. Park, K. Zhang, Y. Zhang, Y. Liu, B. Hui, L. Zhou, X. Wang, Q. Qi *et al.*, "Development and validation of machine learning–based model for the prediction of malignancy in multiple pulmonary nodules: Analysis from multicentric cohorts," *Clinical Cancer Research*, vol. 27, no. 8, pp. 2255–2265, 2021.

[298] Y. Lei, Y. Tian, H. Shan, J. Zhang, G. Wang, and M. K. Kalra, "Shape and margin-aware lung nodule classification in low-dose ct images via soft activation mapping," *Medical Image Analysis*, vol. 60, p. 101628, 2020.

[299] N. L. S. T. R. Team, "The national lung screening trial: overview and study design," *Radiology*, vol. 258, no. 1, pp. 243–253, 2011.

[300] D. Cherezov, S. H. Hawkins, D. B. Goldgof, L. O. Hall, Y. Liu, Q. Li, Y. Balagurunathan, R. J. Gillies, and M. B. Schabath, "Delta radiomic features improve prediction for lung cancer incidence: A nested case–control analysis of the national lung screening trial," *Cancer medicine*, vol. 7, no. 12, pp. 6340–6356, 2018.

[301] P. G. Mikhael, J. Wohlwend, A. Yala, L. Karstens, J. Xiang, A. K. Takigami, P. P. Bourgouin, P. Chan, S. Mrah, W. Amayri *et al.*, "Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography," *Journal of Clinical Oncology*, vol. 41, no. 12, pp. 2191–2200, 2023.

[302] S. G. Armato III, K. Drukker, F. Li, L. Hadjiiski, G. D. Tourassi, R. M. Engelmann, M. L. Giger, G. Redmond, K. Farahani, J. S. Kirby *et al.*, "Lungx challenge for computerized lung nodule classification," *Journal of Medical Imaging*, vol. 3, no. 4, pp. 044 506–044 506, 2016.

[303] Y. Shao, M. Wang, J. Mai, X. Fu, M. Li, J. Zheng, Z. Diao, A. Yin, Y. Chen, J. Xiao *et al.*, "Lidp: A lung image dataset with pathological information for lung cancer screening," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2022, pp. 770–779.

[304] M. Jian, H. Chen, Z. Zhang, N. Yang, H. Zhang, L. Ma, W. Xu, and H. Zhi, "a lung nodule dataset with histopathology-based cancer type annotation," *Scientific Data*, vol. 11, no. 1, p. 824, 2024.

[305] M. Jian, H. Zhang, M. Shao, H. Chen, H. Huang, Y. Zhong, C. Zhang, B. Wang, and P. Gao, "A cross spatio-temporal pathology-based lung nodule dataset," *Scientific Data*, vol. 11, no. 1, p. 1007, 2024.

[306] F. Song, Q. Yang, T. Gong, K. Sun, W. Zhang, M. Liu, and F. Lv, "Comparison of different classification systems for pulmonary nodules: a multicenter retrospective study in china," *Cancer Imaging*, vol. 24, no. 1, p. 15, 2024.

[307] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi, "The limits of fair medical imaging ai in real-world generalization," *Nature Medicine*, vol. 30, no. 10, pp. 2838–2848, 2024.

[308] L. M. Seijo, N. Peled, D. Ajona, M. Boeri, J. K. Field, G. Sozzi, R. Pio, J. J. Zulueta, A. Spira, P. P. Massion *et al.*, "Biomarkers in lung cancer screening: achievements, promises, and challenges," *Journal of Thoracic Oncology*, vol. 14, no. 3, pp. 343–357, 2019.

[309] M. Rodríguez, D. Ajona, L. M. Seijo, J. Sanz, K. Valencia, J. Corral, M. Mesa-Guzmán, R. Pío, A. Calvo, M. D. Lozano *et al.*, "Molecular biomarkers in early stage lung cancer," *Translational lung cancer research*, vol. 10, no. 2, p. 1165, 2021.

[310] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton *et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[311] F. Cardoso, L. J. van't Veer, J. Bogaerts, L. Slaets, G. Viale, S. Delaloge, J.-Y. Pierga, E. Brain, S. Causeret, M. DeLorenzi *et al.*, "70-gene signature as an aid to treatment decisions in early-stage breast cancer," *New England Journal of Medicine*, vol. 375, no. 8, pp. 717–729, 2016.

[312] H. A. Robbins, K. Alcala, E. K. Moez, F. Guida, S. Thomas, H. Zahed, M. T. Warkentin, K. Smith-Byrne, Y. Brhane, D. Muller *et al.*, "Design and methodological considerations for biomarker discovery and validation in the integrative analysis of lung cancer etiology and risk (integral) program," *Annals of epidemiology*, vol. 77, pp. 1–12, 2023.

[313] "The blood proteome of imminent lung cancer diagnosis," *Nature communications*, vol. 14, no. 1, p. 3042, 2023.

[314] P. Sanchez-Salcedo, J. Berto, J. P. de Torres, A. Campo, A. B. Alcaide, G. Bastarrika, J. C. Pueyo, A. Villanueva, J. I. Echeveste, M. D. Lozano *et al.*, "Lung cancer screening: fourteen year experience of the pamplona early detection program (p-ielcap)," *Archivos de Bronconeumología (English Edition)*, vol. 51, no. 4, pp. 169–176, 2015.

[315] M. Mesa-Guzmán, J. González, A. B. Alcaide, J. Bertó, J. de Torres, A. Campo, L. Seijo, M. Ocón, J. Pueyo, G. Bastarrika *et al.*, "Surgical outcomes in a lung cancer-screening program using low dose computed tomography," *Archivos de Bronconeumología (English Edition)*, vol. 57, no. 2, pp. 101–106, 2021.

[316] M. Cobo Cano, D. Serrano, J. Barranco, A. Pasquier, J. P. de Torres, J. J. Zulueta, J. I. Echeveste, A. Ezponda, A. Argueta Morales, J. Sanz-Ortega, J. Berto, A. B. Alcaide, M. Di Frisco, C. Felgueroso Rodero, A. Campo, A. de la Fuente Añó, A. Escobar, K. Valencia, D. Orive Mauleón, M. d. M. Ocón, H. B. Globacka, M. A. Fortuño, V. Perna, M. Rodriguez, M. D. Lozano, A. Calvo, R. Pio, R. Hung, L. Seijo, G. Bastarrika, L. Lloret Iglesias, and L. M. Montuenga, "P-elcap/ccun," Apr. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15120062

[317] O. Proteomics, "Measuring protein biomarkers with olink-technical comparisons and orthogonal validation," 2020.

[318] M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong *et al.*, "The reactome pathway knowledgebase 2022," *Nucleic acids research*, vol. 50, no. D1, pp. D687–D692, 2022.

[319] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[320] R. Zhang, Y. Wei, D. Wang, B. Chen, H. Sun, Y. Lei, Q. Zhou, Z. Luo, L. Jiang, R. Qiu *et al.*, "Deep learning for malignancy risk estimation of incidental sub-centimeter pulmonary nodules on ct images," *European Radiology*, vol. 34, no. 7, pp. 4218–4229, 2024.

[321] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[322] N. Hartman, S. Kim, K. He, and J. D. Kalbfleisch, "Pitfalls of the concordance index for survival outcomes," *Statistics in medicine*, vol. 42, no. 13, pp. 2179–2190, 2023.

[323] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[324] J. Y. Verbakel, E. W. Steyerberg, H. Uno, B. De Cock, L. Wynants, G. S. Collins, and B. Van Calster, "Roc curves for clinical prediction models part 1. roc plots showed no added value above the auc when evaluating the performance of clinical prediction models," *Journal of Clinical Epidemiology*, vol. 126, pp. 207–216, 2020.

[325] B. Wu, X. Sun, L. Hu, and Y. Wang, "Learning with unsure data for medical image diagnosis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 590–10 599.

[326] V. M. Vargas, P. A. Gutiérrez, and C. Hervás-Martínez, "Unimodal regularisation based on beta distribution for deep ordinal regression," *Pattern Recognition*, vol. 122, p. 108310, 2022.

[327] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek *et al.*, "Metrics reloaded: recommendations for image analysis validation," *Nature methods*, vol. 21, no. 2, pp. 195–212, 2024.

[328] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[329] C. Zwirewich, S. Vedal, R. Miller, and N. Müller, “Solitary pulmonary nodule: high-resolution ct and radiologic-pathologic correlation.” *Radiology*, vol. 179, no. 2, pp. 469–476, 1991.

[330] F. Girvin and J. P. Ko, “Pulmonary nodules: detection, assessment, and cad,” *American Journal of Roentgenology*, vol. 191, no. 4, pp. 1057–1069, 2008.

[331] A. A. Taha and A. Hanbury, “Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool,” *BMC medical imaging*, vol. 15, no. 1, p. 29, 2015.

[332] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[333] C. Leduc, D. Antoni, A. Charloux, P.-E. Falcoz, and E. Quoix, “Comorbidities in the management of patients with lung cancer,” *European Respiratory Journal*, vol. 49, no. 3, 2017.

[334] C. Gao, L. Wu, W. Wu, Y. Huang, X. Wang, Z. Sun, M. Xu, and C. Gao, “Deep learning in pulmonary nodule detection and segmentation: a systematic review,” *European radiology*, vol. 35, no. 1, pp. 255–266, 2025.

[335] D. Schouten, G. Nicoletti, B. Dille, C. Chia, P. Vendittelli, M. Schuurmans, G. Litjens, and N. Khalili, “Navigating the landscape of multimodal ai in medicine: a scoping review on technical challenges and clinical applications,” *Medical Image Analysis*, p. 103621, 2025.

[336] C. Candal-Pedreira, A. Ruano-Ravina, V. C. de Juan, M. Cobo, J. M. Trigo, E. Carcereny, M. Cucurull, R. L. Castro, E. S. García, A. Sánchez-Gastaldo *et al.*, “Addressing lung cancer screening eligibility in spain using 2013 and 2021 us preventive service task force criteria: cross-sectional study,” *ERJ Open Research*, vol. 9, no. 6, 2023.

[337] A. McWilliams, M. C. Tammemagi, J. R. Mayo, H. Roberts, G. Liu, K. Soghrati, K. Yasufuku, S. Martel, F. Laberge, M. Gingras *et al.*, “Probability of cancer in pulmonary nodules detected on first screening ct,” *New England journal of medicine*, vol. 369, no. 10, pp. 910–919, 2013.

[338] S. J. Swensen, M. D. Silverstein, D. M. Ilstrup, C. D. Schleck, and E. S. Edell, “The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules,” *Archives of internal medicine*, vol. 157, no. 8, pp. 849–855, 1997.

[339] R. Gao, T. Li, Y. Tang, K. Xu, M. Khan, M. Kammer, S. L. Antic, S. Deppen, Y. Huo, T. A. Lasko *et al.*, “Reducing uncertainty in cancer risk estimation for patients with indeterminate pulmonary nodules using an integrated deep learning model,” *Computers in biology and medicine*, vol. 150, p. 106113, 2022.

[340] D. Orive, M. Echepare, F. Bernasconi-Bisio, M. F. Sanmamed, A. Pineda-Lucena, C. de la Calle-Arroyo, F. Detterbeck, R. J. Hung, M. Johansson, H. A. Robbins *et al.*, “Protein biomarkers in lung cancer screening: technical considerations and feasibility assessment,” *Archivos de Bronconeumología*, 2024.

[341] M. P. Davies, T. Sato, H. Ashoor, L. Hou, T. Liloglou, R. Yang, and J. K. Field, "Plasma protein biomarkers for early prediction of lung cancer," *EBioMedicine*, vol. 93, 2023.

[342] B. Hunter, M. Chen, P. Ratnakumar, E. Alemu, A. Logan, K. Linton-Reid, D. Tong, N. Senthivel, A. Bhamani, S. Bloch *et al.*, "A radiomics-based decision support tool improves lung cancer diagnosis in combination with the herder score in large lung nodules," *EBioMedicine*, vol. 86, 2022.

[343] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[344] E. Engvall and P. Perlmann, "Enzyme-linked immunosorbent assay (elisa) quantitative assay of immunoglobulin g," *Immunochemistry*, vol. 8, no. 9, pp. 871–874, 1971.

[345] S. Aydin, E. Emre, K. Ugur, M. A. Aydin, İ. Sahin, V. Cinar, and T. Akbulut, "An overview of elisa: a review and update on best laboratory practices for quantifying peptides and proteins in biological fluids," *Journal of International Medical Research*, vol. 53, no. 2, p. 03000605251315913, 2025.