Multimodal approaches in healthcare Big Data Analytics for precision medicine

Department of Electrical and Information Engineering
ELECTRICAL AND INFORMATION ENGINEERING
Ph.D. Program

SSD: IINF-05/A - INFORMATION PROCESSING SYSTEMS
SSD: IBIO-01/A - BIOENGINEERING

Final Dissertation

Multimodal Approaches in Healthcare Big Data Analytics for Precision Medicine

---

by

Francesco Berloco

Supervisors:
Prof. Simona Colucci, Ph.D.
Prof. Vitoantonio Bevilacqua, Ph.D.

Coordinator of Ph.D. Program:
Prof. Mario Carpentieri, Ph.D.

Course n°37, 01/11/2021 - 31/10/2024

Politecnico di Bari

LIBERATORIA PER L'ARCHIVIAZIONE DELLA TESI DI DOTTORATO

Al Magnifico Rettore
del Politecnico di Bari

Il sottoscritto Berloco Francesco, nato a Altamura il 15/01/1995

residente a Altamura (BA) in via Rodi 28, e-mail francesco.berloco@poliba.it

iscritto al 3° anno di Corso di Dottorato di Ricerca in Ingegneria Elettrica e dell'Informazione, ciclo XXXVII

ed essendo stato ammesso a sostenere l'esame finale con la prevista discussione della tesi dal titolo:

Multimodal Approaches in Healthcare Big Data Analytics for Precision Medicine

DICHIARA

1) di essere consapevole che, ai sensi del D.P.R. n. 445 del 28.12.2000, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni;
2) di essere iscritto al Corso di Dottorato di ricerca Corso di Dottorato di Ricerca in Ingegneria Elettrica e dell'Informazione ciclo XXXVII, corso attivato ai sensi del "*Regolamento dei Corsi di Dottorato di ricerca del Politecnico di Bari"*, emanato con D.R. n.286 del 01.07.2013;
3) di essere pienamente a conoscenza delle disposizioni contenute nel predetto Regolamento in merito alla procedura di deposito, pubblicazione e autoarchiviazione della tesi di dottorato nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica;
4) di essere consapevole che attraverso l'autoarchiviazione delle tesi nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica del Politecnico di Bari (IRIS-POLIBA), l'Ateneo archivierà e renderà consultabile in rete (nel rispetto della Policy di Ateneo di cui al D.R. 642 del 13.11.2015) il testo completo della tesi di dottorato, fatta salva la possibilità di sottoscrizione di apposite licenze per le relative condizioni di utilizzo (di cui al sito http://www.creativecommons.it/Licenze), e fatte salve, altresì, le eventuali esigenze di "embargo", legate a strette considerazioni sulla tutelabilità e sfruttamento industriale/commerciale dei contenuti della tesi, da rappresentarsi mediante compilazione e sottoscrizione del modulo in calce (Richiesta di embargo);
5) che la tesi da depositare in IRIS-POLIBA, in formato digitale (PDF/A) sarà del tutto identica a quelle **consegnate**/inviate/da inviarsi ai componenti della commissione per l'esame finale e a qualsiasi altra copia depositata presso gli Uffici del Politecnico di Bari in forma cartacea o digitale, ovvero a quella da discutere in sede di esame finale, a quella da depositare, a cura dell'Ateneo, presso le Biblioteche Nazionali Centrali di Roma e Firenze e presso tutti gli Uffici competenti per legge al momento del deposito stesso, e che di conseguenza va esclusa qualsiasi responsabilità del Politecnico di Bari per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;
6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto il Politecnico di Bari ed i suoi funzionari sono in ogni caso esenti da responsabilità di qualsivoglia natura: civile, amministrativa e penale e saranno dal sottoscritto tenuti indenni da qualsiasi richiesta o rivendicazione da parte di terzi;
7) che il contenuto della tesi non infrange in alcun modo il diritto d'Autore né gli obblighi connessi alla salvaguardia di diritti morali od economici di altri autori o di altri aventi diritto, sia per testi, immagini, foto, tabelle, o altre parti di cui la tesi è composta.

Bari, 11/12/2024 Firma Francesco Berloco

Il/La sottoscritto, con l'autoarchiviazione della propria tesi di dottorato nell'Archivio Istituzionale ad accesso aperto del Politecnico di Bari (POLIBA-IRIS), pur mantenendo su di essa tutti i diritti d'autore, morali ed economici, ai sensi della normativa vigente (Legge 633/1941 e ss.mm.ii.),

CONCEDE

- al Politecnico di Bari il permesso di trasferire l'opera su qualsiasi supporto e di convertirla in qualsiasi formato al fine di una corretta conservazione nel tempo. Il Politecnico di Bari garantisce che non verrà effettuata alcuna modifica al contenuto e alla struttura dell'opera.
- al Politecnico di Bari la possibilità di riprodurre l'opera in più di una copia per fini di sicurezza, back-up e conservazione.

Bari,11/12/2024 Firma Francesco Berloco

Department of Electrical and Information Engineering
ELECTRICAL AND INFORMATION ENGINEERING
Ph.D. Program

SSD: IINF-05/A - INFORMATION PROCESSING SYSTEMS
SSD: IBIO-01/A - BIOENGINEERING

Final Dissertation

Multimodal Approaches in Healthcare Big Data Analytics for Precision Medicine

---

by

Francesco Berloco

Referees:
Prof. Sara Moccia, Ph.D.
Prof. Giuseppe Jurman, Ph.D.

Supervisors:
Prof. Simona Colucci, Ph.D.

Prof. Vitoantonio Bevilacqua, Ph.D.

Coordinator of Ph.D. Program:
Prof. Mario Carpentieri, Ph.D.

Course n°37, 01/11/2021 - 31/10/2024

Abstract

The primary purpose of this thesis is to present several pipelines for developing multimodal Decision Support Systems that leverage omics and healthcare Big Data analytics, contributing to the advancement in precision medicine field.

Healthcare Big Data are analyzed using Machine Learning and Deep Learning models which are implemented in prototypal form, known as biomedical Decision Support Systems, across different healthcare domains such as medical image analysis, bioinformatics, natural language processing and survival analysis.

Deep Learning models play a crucial role in medical imaging and bioinformatics fields. In the first one, Deep Learning models find application in extracting features from medical images and making prediction about diseases status or genetic mutations. Within the bioinformatics field, Deep Learning plays a pivotal role in extracting actionable insights from omics data clusters, facilitating a deeper understanding of biological systems (e.g., a patient). Such kinds of data are heterogeneous and generated in a large number, during constants time periods. Concerning survival analysis, Machine Learning and Deep Learning are widely used for assessing and categorizing the severity of pathologies over time, aiding personalized treatment strategies.

Notably, most of medical and clinical examinations are provided with free-text reports; Machine Learning and Deep Learning can be exploited for extracting useful information from them, in the context of natural language processing.

In such scenarios, this thesis objective is to develop and validate several pipelines for heterogeneous healthcare Big Data analytics. Specifically, two sets of multimodal and unimodal pipelines are presented. The former includes the multimodal pipelines that integrate medical imaging data with omics to study Pancreatic Ductal Adenocarcinoma disease from different perspectives. The latter includes pipelines for medical image classification, survival analysis, and natural language processing in different use cases.

Technical contributions of this work include designing novel algorithms, improving existing workflows, designing multimodal algorithms for analyzing heterogeneous data coming

from different sources and incorporating Explainable Artificial Intelligence algorithms for interpreting the decision of investigation models.

In order to develop and validate the proposed pipelines, several heterogeneous case studies have been examined, using either public or private datasets. Regarding the multimodal pipelines, proposed applications focus on pancreatic cancer, including: (i) multi-omics analysis (Radiomics, Genomics and clinical) for overall survival and recurrence prediction; (ii) multimodal analyis based on pathomics and transcrittomics for gene mutation prediction. In unimodal analysis pipelines, proposed applications include: (i) enhancing model selection in survival analysis using time-dependent explainability algorithms for Obstructive Sleep Apnea; (ii) Deep Learning approaches for medical image classification for IgA nephropathy; (iii) shape based breast lesion classification using digital tomosynthesis images; (iv) diagnosis standardization from free-text reports.

Table of contents

List of figures

List of tables

List of Acronyms

Acronym	Definition
AE	Autoencoder
AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Average Precision
ARGO	Automatic Record Generator for Onco-Hematology
AUPRC	Area Under Precision-Recall Curve
AUROC	Area Under ROC Curve
BD	Big Data
C-Index	Concordance Index
C/D AUC	Cumulative-Dynamic Area Under Curve
CLAM	Clustering-Constrained-Attention Multiple-Instance Learning
CNN	Convolutional Neural Network
CPH	Cox Proportional Hazard
CT	Cox Time
DAM	Diagnosis Assignment Manager
DBT	Digital Breast Tomosynthesis
DEG	Differentially Expressed Genes

DH	Deep Hit
DL	Deep Learning
DLBCL	Diffuse Large B-Cell
DNN	Deep Neural Network
DP	Digital Pathology
DS	Deep Surv
DSS	Decision Support System
DT	Decision Tree
FC	Fold Change
FCL	Follicular Lymphoma
GBM	Gradient Boosted Model
GLCM	Gray Level Co-occurrence Matrix
HL	Hodgkin's Lymphoma
HR	Hazard Ratio
IgAN	IgA Nephropathy
LH	Logistic Hazard
LIME	Local Interpretable Model-agnostic
LOOCV	Leave-One-Out Cross-Validation
MCL	Mantle Cell Lymphoma
MIL	Multiple Instance Learning
ML	Machine Learning
MR	Matching Rate
MSE	Mean-Squared-Error

NER	Named Entity Recognition
NLP	Natural Language Processing
OS	Overall Survival
OSA	Obstructive Sleep Apnea
PCH	Piecewise Constant Hazard
PDA/PDAC	Pancreatic Ductal Adenocarcinoma
REC	Recurrence
RF	Random Forest
ROC	Receiver Operating Characteristic
SA	Survival Analysis
SHAP	Shapley Additive Explanations
SML	Survival Machine Learning
SRF	Survival Random Forest
SSVM	Survival Support Vector Machine
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
ViT	Vision Transformer
WSI	Whole Slide Image
XAI	Explainable Artificial Intelligence
XGB	Extreme Gradient Boost

Chapter 1

Introduction

In recent years, the term "Big Data" (BD) has rapidly gained popularity among IT researchers and professionals. Although numerous definitions of BD exist, the one perhaps encapsulates them all is offered by De Mauro et al. [1]: "*Big Data refers to the information asset characterized by such a High Volume, Velocity, and Variety that necessitates specific Technologies and analytical methods to transform it into Value*."

The concepts involved in this definition pertain to: (i) the characteristics of the managed information, "Volume", (ii) the demands for processing this information, "Velocity and Variety", (iii) its economic "Value," which is one of the key factors influencing the societal impact of BD. Such a definition also highlights the necessity of employing specialized technologies and methods specifically tailored for BD to address diverse analytical objectives.

Big Data analytics encompasses three primary branches: statistical analysis (both descriptive and inferential), Machine Learning (ML), and Deep Learning (DL) techniques. The last two have an essential role in BD analytics, as they offer a range of techniques for extracting meaningful insights from vast amounts of data, making predictions, processing unstructured data, and recognizing as well as detecting patterns and anomalies.

In medical domain, BD analytics contribute to improve healthcare systems by supporting specialists in improving the diagnoses accuracy, predicting patient outcomes, optimizing and enhancing patient care, providing personalizing treatment plans. Such set of innovative approaches is defined with the term "Precision Medicine", formally defined by König et al. [2] as a cyclical process in which patient data are used for developing clinical models and every development loop leads to models improvement.

This can achieved by developing clinical Decision Support Systems (DSSs) and integrating them in clinical practice. The advance of technology achieved in the last years has led to an exponential increase of available heterogeneous medical data, such as clinical data, free-

text clinical reports, Magnetic Resonance Imaging (RMI) data, Computed Tomography (CT) data, multi-omics data, Electromyography (EMG) and Electrocardiography (ECG) signals, and so on. In these contexts, DSSs empower clinical decision processes by integrating all patient data, providing a personalized treatment according to the subject characteristics, and assessing and monitoring the health status over time [3].

Artificial Intelligence (AI), particularly Machine Learning and Deep Learning, is highly valued in the development of clinical Decision Support Systems (DSSs) [4, 5].

Radiomics analysis [6] involves the use of complex mathematical models for extracting quantitative features from images, aiming at revealing information that is not directly visible from a human perspective; these features can be mainly classified in first order features (voxel intensities distribution), 2D and 3D region shape features, grey-level features and pixel relationships. ML is then exploited to develop classification and survival analysis models for diagnosis, treatment and prognosis [7–9].

In Computational Imaging [10], DL is well appreciated due to its capability in learning a hierarchical features representation, spanning from low-level features (*e.g.* borders, angles, colors, etc.) to high-level features (also called deep features) obtained by combining the lower-level ones for representing of an entire entity; thus, deep features consist in a global abstract representation of entities in images. The adoption of DL approaches deletes the necessity of design handcrafted features, reduces the dependency from feature engineering and enhances the models generalization on new data due to the higher level of abstraction. Similarly to ML for radiomics, DL models find application in diagnosis, treatment and prognosis by deep feature extraction, lesions classification, lesions and instances segmentation [11–15].

Omic sciences focus on the generation and analysis of genomic, pathomic, transcriptomic, proteomic, and metabolomic data, which are essential for describing biological systems from a global point of view, considering multiple characteristics simultaneously. From a dimensionality standpoint, omics data are characterized by high dimensionality [16], necessitating specific pre-processing steps to reduce and aggregate the data while preserving their informational content. ML and DL offer solutions for dimensionality reduction, such as feature selection algorithms [17–20] or low-dimensional encoding representations using Autoencoders [21–23], as well as data integration and prediction algorithms, enabling the development of multi-omics models [24–27].

Moreover, most clinical examinations are accompanied by free-text reports, that include crucial information like patients data, type of clinical test, diagnosis, medical opinion, potential marker predictors status and so on [28]. One of the issues in medical reports is the lack of standardization in diagnosis term definitions. Named Entity Recognition (NER)

is a branch of NLP dealing with the extraction of key words from free texts and it can be exploit for diagnosis standardization. Despite the rise of DL models such as Transformers [29], many existing systems still rely on traditional Natural Language Processing (NLP) techniques, such as regular expressions, and the integration of simple ML models can help to improve systems performance. [30–32].

1.1 Motivations and Contributions of the Work

Biomedical field indeed benefits from the advance in AI for the design and implementation of clinical DSSs but, nevertheless, several challenges need to be addressed.

From a BD point of view, AI models rely on high data quality but the real-word data contain missing values, outliers and noise that affect the performance of the model; additionally, the integration of heterogeneous data is required for describing a biological system from a global perspective [33]. Such issues can be faced by using BD analytics approaches [34], in particular those based on ML and DL paradigms.

Notably, AI models are inherently complex and are often perceived as "black boxes" by healthcare operators. This opacity can impact trustworthiness [35], as operators may view such systems unfavorably due to the lack of transparency in the AI decision-making process. In fact, Shortliffe et al. [36] state that a clinical DSS should not work as black-box and the operators need to be aware of the underlying process. Here, Explainable Artificial Intelligence (XAI) algorithms come in help, aiming at making AI-based models more transparent and interpretable for users [37].

The aforementioned scenarios raise several opportunities and challenges for AI in biomedical field, this thesis partially addresses by pursuing three main objectives: *(i)* design, develop and implement AI-based pipelines for the processing of healthcare big data, integrating and analyzing heterogeneous data for improving decision-making process in healthcare; *ii* show the flexibility of the methods adopted across different case studies; *(ii)* advance the state-of-art of ML or DL methods in the examined domain.

More specifically, this work focus on developing multimodal and unimodal analytics pipelines in medical imaging, omic sciences and NLP.

The developed multimodal pipelines focus on Pancreatic Adenocarcinoma, with the following tasks:

- **Overall Survival (OS) and Recurrence (REC) prediction in Pancreatic Ductal Adenocarcinoma cohort (CPTAC) using Radiomics, clinical and genomics data.** This study, presented in "A Time-Dependent Explainable Radiomic Analysis from

the Multi-Omic Cohort of CPTAC-Pancreatic Ductal Adenocarcinoma" [27], aims to develop a time-dependent, explainable survival model for patients with pancreatic ductal adenocarcinoma (PDAC). The model integrates radiomic, clinical, and mutational features to predict OS and REC. Four survival machine learning (SML) classifiers are designed, trained, and validated using data from the CPTAC, which includes annotated CT images, clinical information, and mutational profiles. The study also employs SurvSHAP(t) to explore the decision-making mechanisms of the survival algorithms, providing insights into the most significant contributing features and their impact on survival probabilities over time.

- **Genetic mutations prediction in PDAC using histopathological images and trascriptomic data with Deep Learning approaches.** The research work "A Multimodal Framework for Assessing the Link Between Pathomics, Transcriptomics, and Pancreatic Cancer Mutation" [38] focuses on designing an explainable multimodal pipeline to predict genetic mutations in PDAC using transcriptomic and pathomic data. The target genes include the most commonly mutated ones in PDAC: *KRAS*, *TP53*, *SMAD4*, and *CDKN2A* [39]. Two configurations of the CLAM model and three feature extractors are applied for image analysis. For transcriptomic data (RNA-seq), a panel of 60,660 transcripts is pre-processed through two approaches: (i) Differentially Expressed Genes analysis and (ii) an unsupervised deep learning approach using three autoencoder architectures (small, medium, large). The processed transcript panels are then input into three machine learning models, i.e. Random Forest, XGBoost, and Multilayer Perceptron for gene mutation classification (wild-type vs. mutated). A fusion layer combines the outputs of the unimodal models (pathomic and transcriptomic) to produce a multimodal prediction. For each gene, the study compares the performance of the combined models against their unimodal counterparts in terms of AUROC and AUPRC. Explainability is achieved through attention maps and SHAP analysis, providing insights into the most influential features from both pathomic and transcriptomic models.

Concerning the unimodal pipelines, proposed applications cover:

- **Enhancing Survival Model Selection Using Time-Dependent XAI in Obstructive Sleep Apnea.** The work "Enhancing Survival Analysis Model Selection through XAI(t) in Healthcare" [40] focuses on improving the selection of ML and DL models for survival analysis through time-dependent explainable AI. An end-to-end pipeline is developed to estimate OS in patients with obstructive sleep apnea using various ML

and DL survival models. Model evaluation is conducted using metrics such as the C-Index, time-dependent AUC , and the Brier Score. The survSHAP algorithm is applied to the best-performing models, demonstrating how explainability can aid in distinguishing between models with similar performances, ultimately supporting more informed model selection.

- **Glomeruli detection and lesions classification using a supervised Deep Learning approach in histopathology.** The study "Performance and Limitations of a Supervised Deep Learning Approach for the Histopathological Oxford Classification of Glomeruli with IgA Nephropathy" [41] presents a pipeline for glomeruli detection and lesions classification in histopathology. The approach consists of two main components: (i) segmentation block: Whole-slide images are divided into tiles, followed by glomeruli segmentation using object detection models; (ii) classification block: Convolutional Neural Networks classify the segmented glomeruli. The classification results are reported at both the glomerular and biopsy levels. To assess performance, intraclass correlation coefficients (ICCs) and Cohen's Kappa statistics are used to evaluate agreement between the model's predictions and expert pathologist annotations.

- **Shape-Based Breast Lesion Classification Using Digital Tomosynthesis Images.** The work entitled "Shape-Based Breast Lesion Classification Using Digital Tomosynthesis Images: The Role of Explainable Artificial Intelligence" [42] presents a mathematically and visually interpretable deep learning framework for multiclass, shape-based classification of breast lesion images. Eight pretrained CNN architectures are employed to classify previously extracted regions of interest containing lesions. To address the black-box nature of deep learning models, two XAI techniques, Grad-CAM and LIME, are used for visual interpretability. Additionally, t-SNE and UMAP are applied as mathematical interpretability methods to analyze multiclass feature clustering and the behavior of the pretrained models.

- **Diagnosis standardiziation and Named Entity Recognition in hematological free-text report using hybrid NLP and ML approach.** The work "ARGO 2.0: A Hybrid NLP/ML Framework for Diagnosis Standardization" [43] introduces an enhanced version of the Automatic Record Generator for Onco-Hematology (ARGO). This framework extracts key fields from oncological free-text reports and standardizes diagnoses based on definitions from the National Institute of Health, aligned with the International Classification of Diseases, 10th Edition (ICD-10) and Oncology (ICD-O) [44]. The enhancement incorporates a machine learning model to support

the existing architecture, along with a decision heuristic for classifying extracted diagnoses. Preliminary results using a transformer-based architecture, replacing the previous system, are presented for the diagnosis Named Entity Recognition (NER) task.

Notably, most of the methods and models employed in the unimodal pipelines were re-employed in the development of the two multimodal ones, showing their flexibility w.r.t. the case studies. Nevertheless, the multimodal pipelines are presented first in Chapter 3, since they are the main focus of this work. The methodologies employed, as well as the various case studies, are thoroughly examined in the following chapters.

1.2 Thesis Outline

The thesis is structured as follows: Chapter 2 provides a comprehensive overview of Machine Learning and Deep Learning methodologies and models employed, defining all key topics for each pipeline developed. Chapter 3 outlines the contributions of two multimodal pipelines in pancreatic cancer studies. Chapter 4 includes the development of time-dependent Explainable Artificial Intelligence algorithms that enhance model selection in survival analysis and the application of deep learning techniques to improve diagnostic accuracy in medical image analysis and diagnosis standardization in medical free-text reports. The taxonomy related to the aforementioned chapters in depicted in Figure 1.1. Finally, Chapter 5 synthesizes the key results achieved in this thesis, highlighting their contribution to advancing healthcare analytics and presents potential future works.

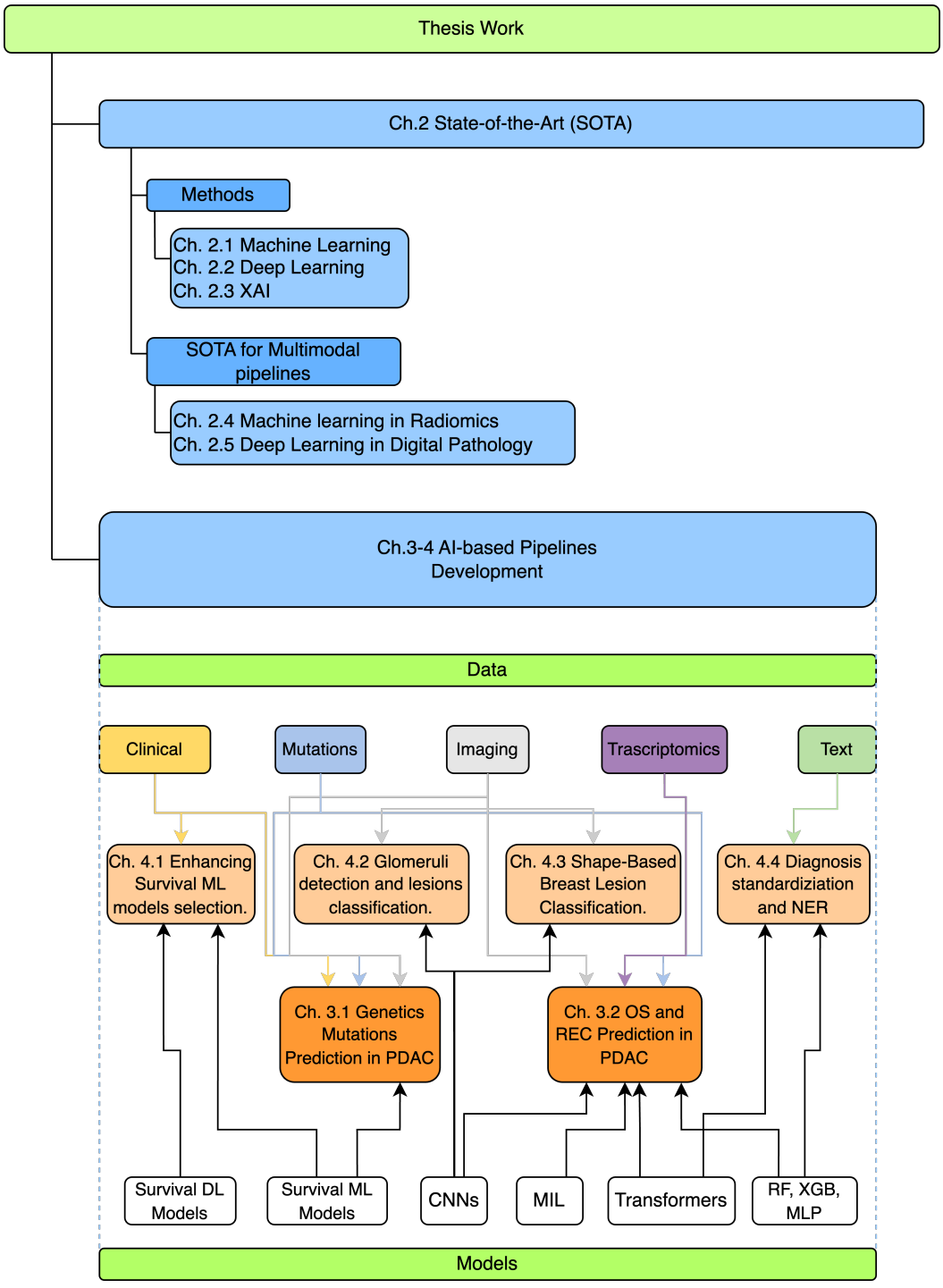


Fig. 1.1 Taxonomy of the thesis work related to chapters 2,3, and 4.

Chapter 2

State Of The Art

This chapter provides a comprehensive overview about the State-of-the-Art methodologies adopted in this thesis. Section 2.1 gives an overview of AI tasks and introduces classical Machine Learning models. In Section 2.2, Deep Learning models are presented with related applications. Finally, Section 2.3 illustrates the state-of-the-art related to XAI algorithms.

2.1 Machine Learning

Machine Learning [45] is a branch of Artificial Intelligence aiming at building computational models able to learn patterns from data and making predictions. ML techniques can be classified into three main branches, according to the kind of analysis performed:

- **Supervised Learning.** It consists in developing models capable of mapping input observations to corresponding learned outputs; in this scenario, models training is performed according to a labeled dataset given as input.
- **Unsupervised Learning.** It produces models able to find patterns in unlabeled data.
- **Reinforcement Learning**: produces models interacting with the environment to accomplish a specific task; the model training is performed according to a rewards-penalties rules.

Furthermore, there are also other sub-types of learning algorithms such as weakly-supervised learning, semi-supervised learning, self-supervised learning, active learning, transfer learning and multi-task learning [46, 47].

Supervised learning algorithms primarily address three main tasks: Regression, Classification, and Survival Analysis (SA).

All three tasks involve estimating the relationship between a dependent variable, often referred to as the target variable, and a set of independent variables (features). The distinction between such type of algorithms lies in the nature of the target variable: regression focuses on predicting numerical targets, classification concerns the prediction of categorical variables, and survival analysis aims to estimate the time until a specific event occurs, along with the associated probability.

The next paragraphs focus on Classification and SA models.

2.1.1 Classification

2.1.1.1 Logistic Regression

Logistic Regression is an algorithm developed for estimating the probability that a sample belongs to a specific class. In case of binary classification, the estimated probability concerns the likelihood that a sample belongs to one of the two classes, typically represented as class 0 (negative) or class 1 (positive). The relation between dependent variable and the feature set is modeled through a hypothesis function $h_\Theta(X)$, in particular a logistic function defined as:

$$\hat{Y} = h_\Theta(X) = \frac{1}{1 + e^{-\Theta^T \cdot X}} \tag{2.1}$$

where:

1. X indicates the feature set.
2. $\hat{Y}$ is the output of the model (the probability estimated)
3. Θ is the model parameters vectors that will be learned during the training phase.

The hypothesis function represents a *sigmoid* and lies between $]0,1[$; the choice of such a function is justified by its probabilistic interpretation. A probability threshold is set to define the binary class according to the model output. The probability estimated for a specific sample i (assumed that belongs to the positive class, 1) is defined as:

$$p(y^{(i)} = 1 | x^{(i)}; \Theta) = h_\Theta(x^{(i)}) \tag{2.2}$$

while for class 0 is is:

$$p(y^{(i)} = 0 | x^{(i)}; \Theta) = 1 - h_\Theta(x^{(i)}) \tag{2.3}$$

The classification aims at maximize the *Likelihood*, meaning the probability to obtain a certain output y, giving x features with Θ parameters. The Likelihood is formally defined as:

$$L(\Theta) = p(Y|X;\Theta) = \prod_{i=1}^{n} p(y^{(i)}|x^{(i)};\Theta) \tag{2.4}$$

If the Likelihood is combined definition with *2.2* and *2.3* and transformed to a logarithmic form, given m samples, it can be written as:

$$\begin{aligned} l(\Theta) &= \log \prod_{i=1}^{m} \left(h_\Theta(x^{(i)})^{y^{(i)}} \cdot (1 - h_\Theta(x^{(i)}))^{1-y^{(i)}} \right) \\ &= \sum_{i=1}^{m} \left(y^{(i)} \log \left(h_\Theta(x^{(i)}) \right) + (1 - y^{(i)}) \log \left(1 - h_\Theta(x^{(i)}) \right) \right) \end{aligned} \tag{2.5}$$

In such a scenario, the model training is led by the minimization of a cost function retrieved by the previous equation:

$$J(\Theta) = -\frac{1}{m} \sum_{i=1}^{m} (y^{(i)} \log(h_\Theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\Theta(x^{(i)})) \tag{2.6}$$

called *Cross Entropy* (CE) function.

From a geometrical perspective, binary classification problem with logistic regression consists in find the best decision boundary that separates two classes (defined as the product between Θ parameters and X features).

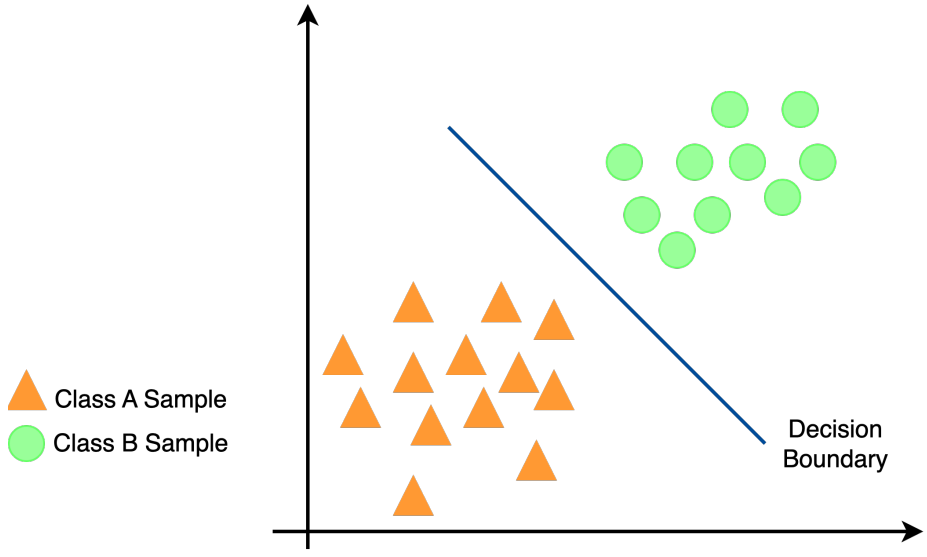


Fig. 2.1 Binary classification problem with two classes linearly separable.

2.1.1.2 Tree-based Models

A tree model is a type of decision model based on tree-graph structure; it is composed by a root node, with arches and leaf nodes. The key-idea is to exploit the tree structure for splitting data by several decision rules defined according to features characteristics. The main tree-based models are Decision Tree (DT), Random Forest (RF) and Boosted Tree (BT).

Decision Tree. It is the simplest model, composed by a *root* node, containing the full dataset, *decision* nodes (also called intermediate nodes), containing a subset of data previously split by a decision rule (e.g. color types, sex, shape and so on) and *leaf* nodes where the decision is made. The goal of splitting is to improve the homogeneity of the resulting sub-groups, i.e. the data within each child node should be more similar to each other than the parent node. The number of decision nodes determines the *depth* of the model.

There are several split criteria that can be adopted but the most common are the Gini Index and the Information Gain [48]. Figure 2.2 represents an example of DT for heart failure risk estimation.

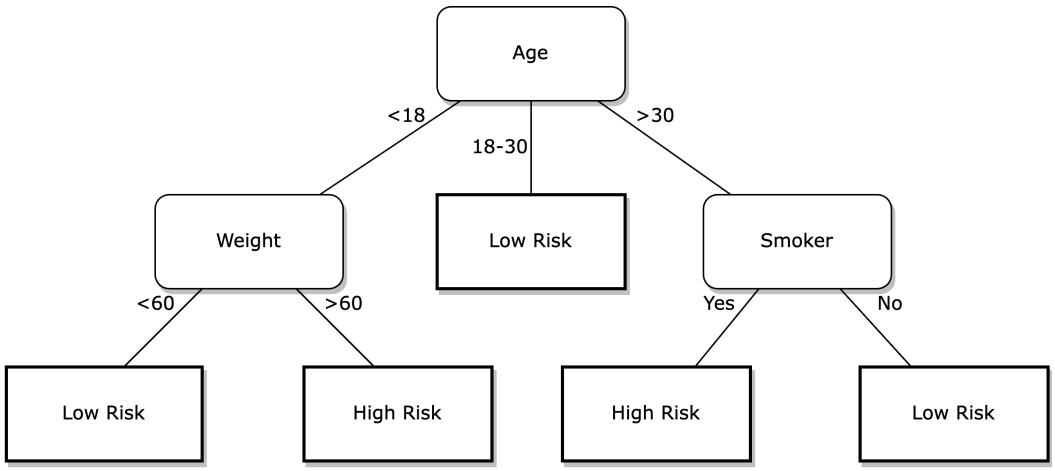


Fig. 2.2 Visualization of a Decision Tree model for heart failure risk estimation.

On one hand the main benefit of DTs is their high interpretability, due to the visualization of the decision model, along with splitting rules; on the other hand, they suffer of overfitting problems, leading to misclassifications on new datasets.

Random Forest. As the name suggests, a RF is a set of DTs trained using the *Bagging* ensemble method. The Bagging is type of model training technique, in which several models are trained on a sub-set of data and a sub-set of features (bootstrap). The models are trained independently and paralleling and the prediction is made by the

majority voting of all sub-models. In this way it is possible to reduce overfitting, while decrease variance and improving the accuracy.

Boosted Models. They are models trained with *Boosting* ensamble: several models (called weak learners) are trained sequentially and each model try to correct the errors of the previous models by fine tuning its weights according to the classification error of the previous model. There are two main type of boosting:

- Adaptive Boosting - Also called AdaBoost, each weak learner is trained sequentially, and after each iteration, the weights of misclassified data points are increased. The final prediction is a weighted vote of all weak learners, where the weight of each learner depends on its accuracy.
- Gradient Boosting - Aims at minimizing the errors of the previous model by training the next model to predict the difference between the model output and the ground truth (residual).

The models trained with boosting can be DTs or other types.

2.1.1.3 Multilayer Perceptron Classifier

The Multilayer Perceptron Classifier (MLP), also called a Feedforward Neural Network, is a type of Artificial Neural Network (ANN), structured into layers as follows:

- Input Layer: The first layer, which receives the input data.
- Output Layer: The final layer, which provides the prediction.
- Hidden Layers: The set of layers between the Input Layer and the Output Layer.

Each layer contains one or more nodes (neurons), depending on the type of network, and each node is connected to at least one neuron in the subsequent layer. If all neurons in one layer are connected to all neurons in the next layer, the network is referred to as fully connected. Additionally, each neuron has an Activation Function, that help the network learn complex patterns; common functions include: ReLU (Rectified Linear Unit), Sigmoid, Tanh and Softmax (usually in the output layer for multi-class classification).

Each connection between neurons is associated with a specific parameter Θ, learned during the training process, called *weight*; in particular, the computation of Θ parameters is made using the *Forward and Backward Propagation* algorithm [49]. Finally, each layer

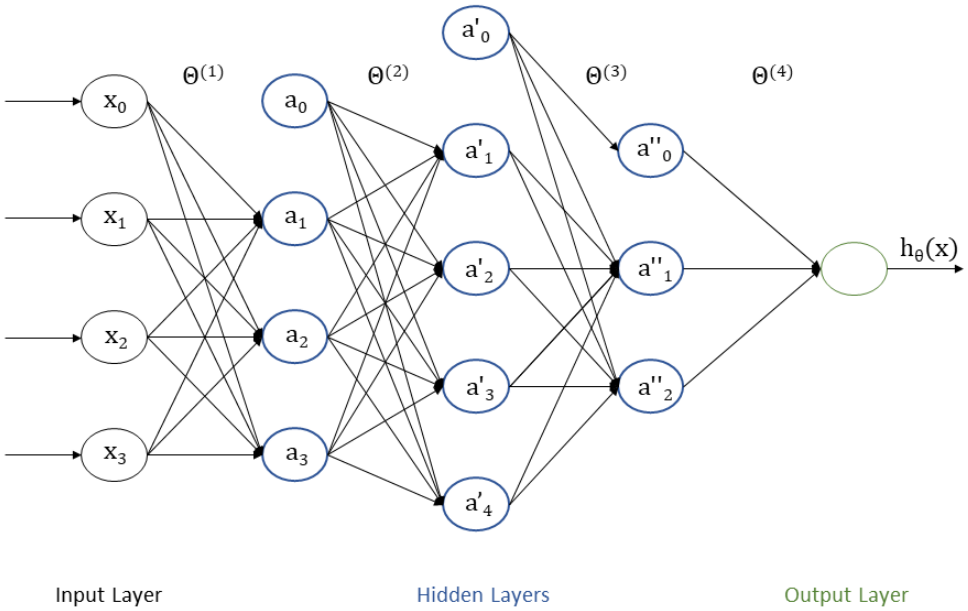


Fig. 2.3 Example of Artificial Neural Network fully connected.

(except for the Input and Output layers) includes a Bias Unit, which serves as an additional neuron to improve the network's learning ability.

Where:

- a_0, a'_0, a''_0 are Bias units.
- $\Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}, \Theta^{(4)}$ are the vectors of parameters associated to the connections from the neurons of one layer to the ones of the next one; these are initialized randomically at the first iteration.
- a_i, is the generic neuron.
- x_i, is the input data.
- $h_\Theta(x)$ is the hypothesis function of the output layer.

The cost function for classification task is an extension of Cross-entropy (eq. 2.6):

$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K}\left(y_k^{(i)}\log\left(h_\Theta(x^{(i)})_k\right) + (1-y_k^{(i)})\log\left(1-h_\Theta(x^{(i)})_k\right)\right)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(\Theta_{ji}^{(l)}\right)^2 \tag{2.7}$$

Where:

- m is the number of samples.
- K is the number of classes of target variable.
- L is the number of layers.

2.1.2 Survival Analysis

2.1.2.1 Cox Regression

The Cox Regression model, also known as the Cox Proportional Hazards (CPH) model, is a specialized multiple regression technique used to examine the relationship between a predictor (often a risk factor) and the likelihood of a specific outcome, while adjusting for one or more confounding factors. It is a proportional hazards model, meaning that covariates have a multiplicative effect on the hazard rate, which remains constant over time for a given factor.

Such a model can be divided into two parts: the underlying baseline hazard function, often denoted as $H_0(t)$, that describes how the risk of event per time unit changes over time at baseline levels of covariates, and the effect parameters, that describes how the hazard varies in response to explanatory covariates. Let $X_i = (X_{i1}, \ldots, X_{ip})$ the p covariates for subject i. The hazard function of a CPH model is:

$$H(t \mid X_i) = H_0(t)^{\left(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right)} = H_0(t)^{(X_i \cdot \beta)} \tag{2.8}$$

where β are the effect parameters. Therefore, the Cox model returns the regression coefficients for each co-variate. These coefficients β indicate the risk that the event occurs and they are called *Hazard Ratio* (HR).

2.1.2.2 Survival Random Forest

Survival Random Forests (SRF) were introduced in [50] to extend RF to the setting of right-censored survival data. The implementation of SRF follows the same general principles as RF: survival trees are grown by using bootstrapped data, then a random feature selection is used when splitting tree nodes. Trees are generally grown deeply, and the survival forest ensemble is calculated by averaging terminal node statistics.

Let h be a terminal node of a forest tree and N be the instances that, in the process of building the tree, fall into it. Every instance is represented by X_i that is the vector of the

features and its survival information, involving the time event $T_{i,h}$ and related status $\delta_{i,h}$). Thus for the node h there are $T_{i,h} = t_0, \ldots, t_N$ time events and $\delta_{i,h} = 0, 1$ status.

For every instant of time $t_{j,h}$ in the tree node h, it is possible to define the number of events $d_{j,t}$ and the number of the number of individuals at risk $Y_{j,t}$. Then Cumulative Hazard Function (CHF) estimated for leaf h is defined as:

$$CHF_h = \hat{H}_h(t) = \sum_{t_{j,h} \leq t} \frac{d_{j,h}}{Y_{j,h}} \tag{2.9}$$

The SRF prediction of the tree for subject i, denoted by $H(t \mid \mathbf{X_i})$, coincides with the CHF estimate of the leaf end node:

$$H(t \mid X_i) = \hat{H}_h(t) \tag{2.10}$$

The final Hazard function is mediated on all trees:

$$\overline{H}(t \mid \mathbf{X}) = \frac{1}{n_{\text{tree}}} \sum_{n_{\text{tree}}} H_b(t \mid \mathbf{X}) \tag{2.11}$$

2.1.2.3 Survival Gradient Boosting Model

Following the idea of boosted models, according to the loss function employed for a Gradient Boosted Model (GBM) it is possible to adapt such a model according to a specific use case, even for survival analysis.

The general problem in GBM models is to learn a functional mapping Φ, in the form $y = F(X;\beta)$ from data $\{\mathbf{X_i}, y_i\}_{i=1}^n$ where β is the set of parameters of F, such that the following cost function:

$$(CF) = \sum_{i=1}^{n} \Phi(y_i, F(X_i;\beta)) \tag{2.12}$$

is minimized. $F(x)$ follows an "additive" expansion form and incorporates all the base learners with a weight ρ and a parameter set τ. The function is dependent on the family of the model chosen When GBM is used in survival analysis (Survival GBM - SGBM), $\Phi(y,F)$ and $F(X)$ can be adapted to the Cox model as made by Ridgeway [51].

2.1.2.4 Survival Support Vector Machine

Support Vector Machines (SVMs) found a widespread application in standard classification problem, due to their promising performance and linear training times. They are able to account for complex, non-linear relationships between features and the target variable through

the so-called kernel trick: a *kernel* function implicitly maps the input features in a high-dimensional feature spaces, to a lower dimensional space where the target can be represented by a hyperplane. This makes SVMs highly versatile and applicable to a wide range of data. The Survival SVM (SSVM) aims at dealing with survival problems, by approaching them in two main different ways [52]:

- A ranking problem - the model learns to assign samples with shorter survival times a lower rank by considering all possible pairs of samples in the training data. This approach is based on the idea of support vector regression (SVR), which aims at finding a function estimating observed survival times as continuous outcome values y_i by using covariates X_i.
- A regression problem - the model learns to directly predict the (log) survival time, considering the survival problem a classification problem with an ordinal target variable; one of the widespread ranking formulation is the *Van Belle* [53].

2.2 Deep Learning

Deep Learning [49] is a branch of Machine Learning, referring to complex ANN-based models, called Deep Neural Networks (DNNs), built and employed for hierarchical feature extraction and raw data processing; the complexity of DL models lies in the elevate number of layers and parameters. This section focuses on DL models for images classification, SA and Dimensionality Reduction with Autoencoders.

2.2.1 Classification

2.2.1.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a popular type of DNNs optimized for processing matrix data. They are widely used in computer vision tasks like image classification, object detection, and segmentation, but they can also be applied to other types of data, such as signals and time series.

Differently from classical ANNs, such models are characterized by convolutional layers, Pooling layers, Activation Functions, and Fully Connected layers. For computer vision tasks, the input is an image represented by a 3D matrix, representing the image height, width, and its three color channels - RGB.

The convolutional layers apply a series of filters (called *kernels*) that slide over the input matrix (according to preset stride and padding parameters) to detect features like edges, textures, or more complex patterns. Each filter is learned during training, and the result of applying the filters is a set of feature maps.

After, the activation functions are applied to the output of convolutional layer to introduce non-linarity in model (often a ReLu). Then the data are processed by pooling layers, for reducing spatial dimensions of the feature maps while retaining important information. The most common form is max pooling, where the maximum value from a set of values in a feature map is selected but alternatively it is also possible to use an average pooling. Lastly, the output of the final pooling layer is flattened ad processed by several fully connected layers to combine the learned features for making prediction.

Figure 2.4 shows an example of CNN architecture, in particular VGG-16 [54].

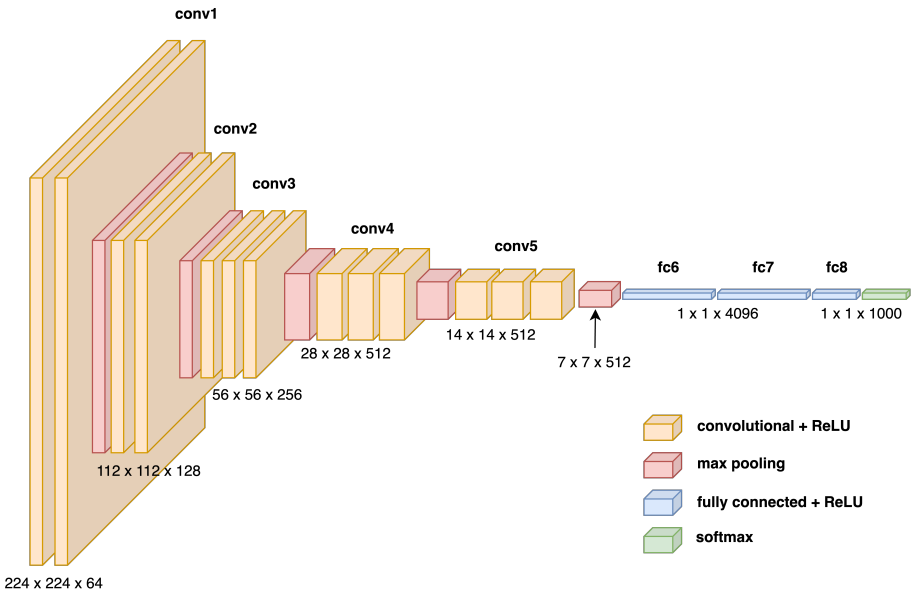


Fig. 2.4 Example of CNN architecture - VGG16 Architecture.

However, training a CNN from scratch is an expensive task, especially the convolutional layers. The key approach lies in *transfer learning*, consisting in the use of models with pre-trained parameters, freezing the convolutional layers (used for feature extraction) and training only the fully convolutional ones according to the task involved.

This approach is especially useful in medical scenarios where data availability is a challenge, and it is enabled by the fact that the feature extraction layers of a CNN capture important structural information from the images. Typically, CNN models trained on ImageNet [55] are leveraged for this purpose.

Based on the network architecture, there are several models that might be employed for different tasks. The ones used in this thesis are the following:

- **MobileNetV2** [56], is a compact CNN architecture known for its efficiency. The model starts with a low-dimensional input that is expanded into a high-dimensional representation. The high-dimensional data is then processed through a lightweight depth-wise convolution layer and, subsequently, the features are projected back into a low-dimensional space using a linear convolution operation. This process leverages an inverted residual structure with a linear bottleneck block.

- **SqueezeNet** [57], is a model or system designed for tasks with limited computational resources. Although this characteristic may result in reduced accuracy, which is less ideal for medical tasks, it can act as a safeguard against overfitting when data availability is limited.

- **DenseNet** [58], differently from the other CNNs, this model uniquely incorporates the feature maps from all preceding layers into each subsequent layer using the concatenation operation.

- **ResNet** [59], introduces the concept of skip connections. In this approach, the input of a convolutional layer block is combined with its output. This allows the model to learn a residual function, expressed as $f(x) = h(x) - x$, which is easier to model than the original function $h(x)$. This innovation not only accelerates the training process but also enables the development of deeper models.

- **EfficientNetV2** [60], builds upon the MobileNetV2 architecture. It employs a uniform scaling approach for all network dimensions using a compound coefficient. Compared to its previous version, EfficientNet, this one significantly enhances training speed and parameter efficiency by reducing memory access and the total number of parameters. EfficientNetV2 offers three variants, each differing in the number of parameters: EfficientNetV2-S (22.10 million parameters), EfficientNetV2-M (55.30 million parameters), and EfficientNetV2-L (119.36 million parameters).

- **VGG** [54] comes in two famous versions with 16 and 19 layers comprising 144 million parameters. This study considers the earlier i.e. VGG-16, which consists of several number of channels, 3x3 receptive fields, and a stride of 1. This model is composed of convolutions layers, max pooling layers, fully connected layers with 5 blocks and each

block with a max pooling layer, and extra convolutional layers contained in last three blocks.

- **ResNeXt** [61] is counterpart of ResNet, is a specifically designed image classification network with very few tuneable parameters. It contains a series of blocks with a set of aggregations of similar topology with an additional dimension called cardinality. This cardinality, which creates major difference between its brother networks, competes with the depth and width of network. The simpler architecture based on VGG and ResNet with fewer parameters yields better accuracy on ImageNet classification dataset. The word *NeXt* in the name of the network refers to next dimension which surpasses ResNet-101, ResNet-152, ResNet-200, Inception-v3, and ResNet-v2 on the ImageNet dataset in accuracy.

2.2.1.2 Transformer and Vision Transformer

The transformer architecture, proposed by Vaswani et al. [62], is a DL model designed for NLP tasks, including an Encoder-Decoder architecture and exploiting the so-called *self-attention mechanism*.

The encoder takes the input sequence and transforms it in a new internal representation while the decoder takes the input encoding and returns an output. The self-attention it is used to weights parts of the input according to the context of the entire sequence.

The encoder can be composed by stacking multiple blocks, each one composed by the self-attention block and a feed-forward net. In particular, the input sequence is transformed into a numerical vector with an *embedding* operation and, using a positional encoding operation, the order of the original tokens in the sequence is saved along the numerical vector. The innovative part concern the self-attention, described by the following equation:

$$Attention(Q,K,V) = Softmax(\frac{Q * K^T}{\sqrt{d_k}}) * V \tag{2.13}$$

Where d_k represents the dimensionality of the key vectors, while Q, K, and V are called respectively *queries*, *keys*, and *values*. They are different representation of the tokens: Q is used for comparing the tokens among themselves (what are we looking for), K is used by a token for focusing on the other tokens (how a token can be relevant) and V represents the information used for computing the output. The dot product between Q and K is computed to determinate the importance (weight) of a token with respect to the others. Then the attention is computed according to Equation 2.13. The key-idea lies in the possibility of each input token to "pay attention" to the other ones in the sequence. Once obtained the token weights,

for each token it is possible to obtain a new token representation by computing the weighted sum of values. Such an operation is performed several times, processing several parts of the input sequence with a Multi-head attention.

The new representation is given as input to the feed-forward net, introducing the non-linearity. The decoder works in a similar way, using a masked self attention and an encoder-decoder attention, looking the tokens generated by the encoder and using them for generating the next tokens. At the end of the model, the token is transformed with a linear function and the softmax is applied to the result. The transformer architecture is portrayed in Figure 2.5.

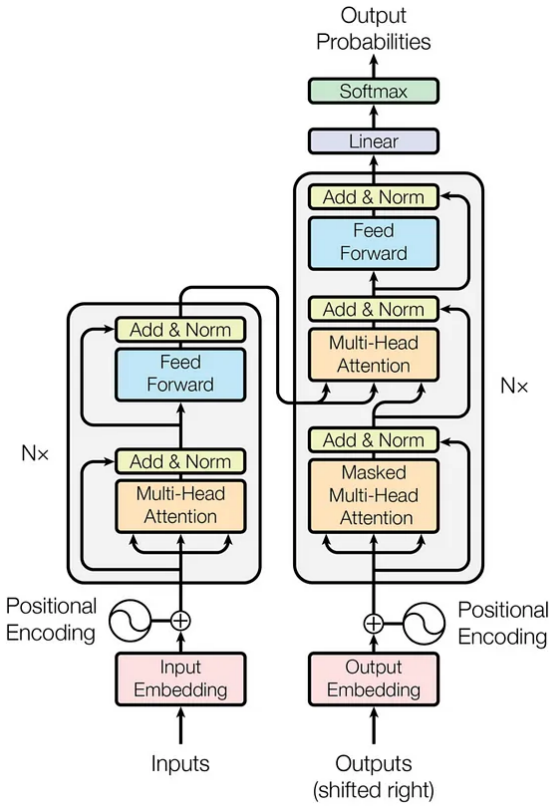


Fig. 2.5 Transformer architecture introduced by Vaswani et al. [62].

Vision Transformers (ViTs) are a class models that have gained significant traction in the field of computer vision. These models adapt the transformer architecture for image processing tasks [63].

Differently from traditional transformers, in ViT the input is an image, divided in sequence of patches. Each patch is flattened and processed using an embedding matrix with a positional encoding (for retaining the positional information of the patch) and given as input to the transformer. Another difference lies in the token of class - CLS, added to the begin of

the sequence and used for aggregating the information learned from patches through the attention.

The development of ViTs represents a shift in a field traditionally dominated by CNNs, due to their promising performance comparable to CNNs, absence of convolution, kernel independence, and global attention (the model focuses on all patches of all images); on the other hand, transformers are computationally expensive and require large dataset for being trained effectively.

The ViT Architecture schema is depicted in Figure 2.6.

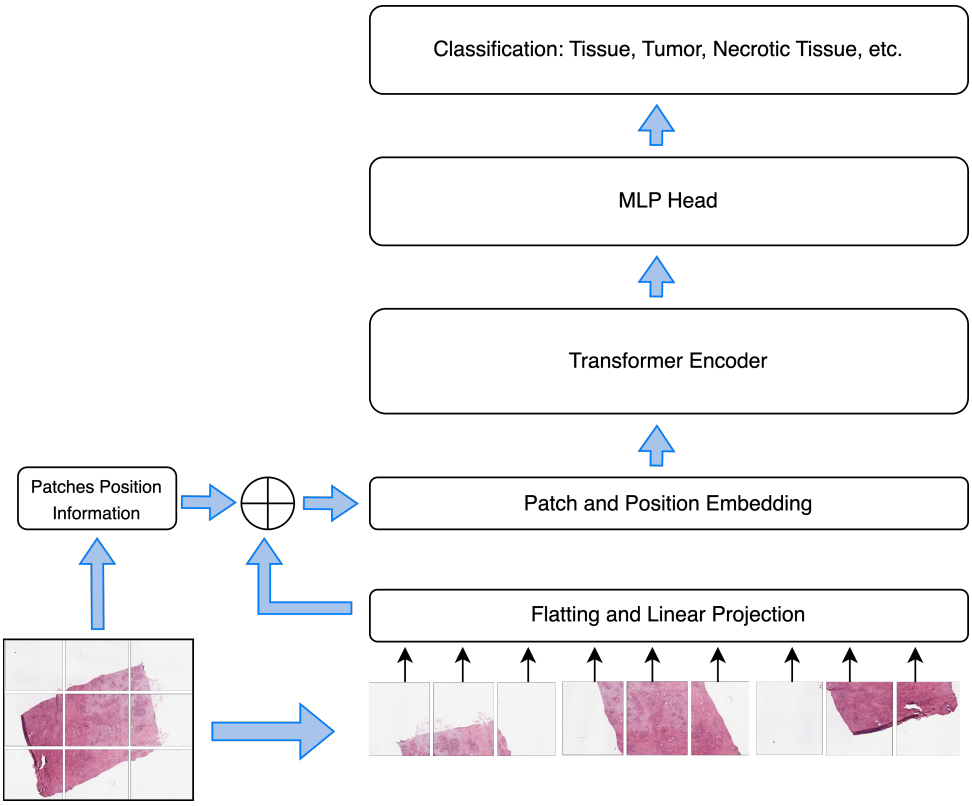


Fig. 2.6 Vision Transformer (ViT) Schema

One of the key strength of traditional transformers and ViT is the use of the attention scores for explainability purposes (Section 2.3).

In this thesis, two ViT-based foundation models are employed as feature extractor, UNI and CONCH.

UNI Presented by Chen et al. [64], UNI is ad hoc vision-based foundation model for histopathology, trained over 100 millions of images derived from hundred of thousands WSIs, coming from Brigham and Women's Hospital and Massachusetts General Hospital; the use of private datasets reduces the risk of data contamination in benchmarking.

It is based on ViT Large architecture, leveraging DINOv2 (DIstillation of knowledge with NO labels) for pre-training [65]. Specifically, the distillation process consists in two networks, called teacher and student, in which the teacher networks is a large, complex model that has been trained on a task and achieved high performance and the latter is a smaller simple model, trained to mimic the output of the teacher. The student network is trained with output probabilities of the teacher, instead of hard labels; this allows the student model to learn how the teacher generalized the input data. This is achieved by weighting the student loss function with the output probabilities of the teacher, minimizing the cross-entropy between the student's predictions and the teacher's soft targets. DINO is self-supervised learning framework designed by Facebook AI Research, designed for learning image representations without needing labeled data; this is achieved through contrastive learning, where the model learns by comparing different augmented views of the same image and ensuring that their representations are similar, while representations of different images are distinct.

CONCH Released by Lu et al [66], CONCH is a multimodal ViT-based vision-language foundation model specifically designed for histopathology, trained on over 1.17 million image-caption pairs. It uses an image-text encoder-decoder model trained with Contrastive Captioning (CoCa) loss, which helps align the representations of images and their corresponding captions. This enables the model to perform well on cross-modality tasks, where it needs to retrieve images based on text or generate captions based on images. For computer vision tasks it is possible to extract and use only the image encoder.

2.2.1.3 Multiple-Instance Learning

In last years, Multiple Instance Learning (MIL) [67] approaches gained attention across pathomic field, especially for Whole-Slide-Images (WSIs) classification. Traditional approaches consist in patch-based methods [41]: the WSI is first divided into smaller sub-images (patches) and each patch is analyzed and classified as independent instance; the WSI classification is derived from the classification of the individual instance.

This approach requires an expensive manual labeling of all instances present in the WSI by experts. Another approach consists in analyzed the entire WSI by down-sampling; however this approach may cause an information loss that heavily affects model's performance.

In MIL-based classification paradigm, the input WSI is treated as a bag of instances, where only the label of the bag is known; this is a form of weakly-supervised learning.

Differently from patch-based approaches, instead of classifying the single patch, the model learns to classify the entire WSI by treating it as a bag of instances, aiming at finding a relationship between the single instances and the global WSI label.

This allows the model to identify the most critical regions in the WSI that contribute most to the final classification. The identification of the most important patches is often accomplished with the support of attention-based mechanism, for weighting patches importance, as made by CLAM.

Clustering-constrained Attention Multiple Instance Learning - CLAM Introduced by Lu et al. [68] CLAM is a multi-class, weakly-supervised, and attention-based model for WSIs classification. Specifically, it identifies the WSI regions of high diagnostic value and exploits an instance-level clustering approach over such regions for constraining and refining the feature space. During the training and inference, for a specific class, the models analyzes and ranks all patches of the tissue regions, assigning an attention score about their importance at slide-level representation.

At slide-level, there is an attention-based pooling layer, which produce a slide representation as the average of all patches in the WSI, weighted by their attention. Moreover, CLAM has several parallel attention branches for calculating different unique WSI representation, for each class. The choice of use a trainable attention-based pooling layer instead of max pooling (or other classical aggregation operation) is justified by the fact that, in the latter, when only the WSI-level label is known the gradient signal for updating network parameters comes from a single instance, causing an inefficient use of the others WSI instances during the training.

Due to the presence of multiple attention branches, the model produces several sets of attention scores specific for each class; this allows the model to unambiguously learn for each class which morphological features should be considered as positive and negative. Moreover, for improving the learning of class-specific features, a binary clustering layer is placed after the first fully connected layer, which use the slide-level attention score coming from the previous layer to further divide the patches according to their contribution for the specific class. These *clustering constraints* encourages the model to group similar patches and focus on informative clusters. The final classification for the entire slide is derived by pooling the attention-weighted patch-level features and predicting the WSI's label. Another key strength of CLAM is its interpretability; in fact the Attention-Maps can be retrieved by projecting the attention scores on the input WSI, without any additional algorithm. This transparency makes CLAM particularly

valuable in clinical settings, where understanding the rationale behind predictions is essential.

A simplification of CLAM workflow is depicted in Figure 2.7

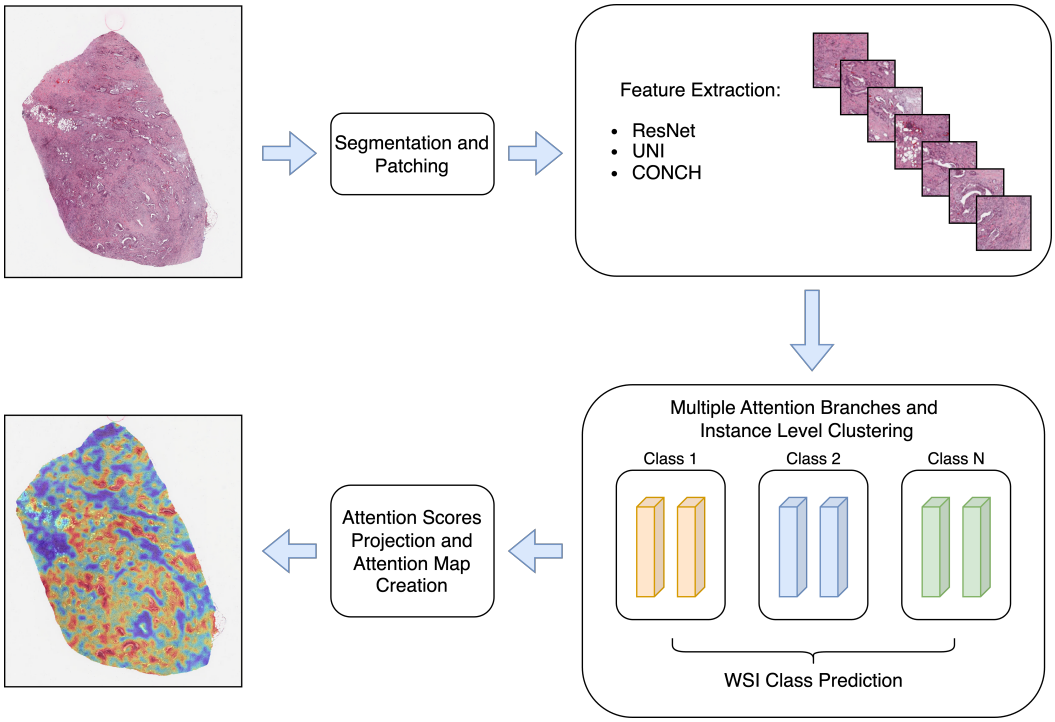


Fig. 2.7 CLAM Workflow.

2.2.2 Survival Analysis

In survival analysis, DL techniques have been increasingly utilized for modeling complex relationships between covariates and survival outcomes. The models included in this thesis are described in the following sections.

2.2.2.1 DeepSurv

Deep Surv (DS) [69] is a deep feed-forward neural network based on CPHs for modeling interactions between a patient's covariates and an event. The network propagates the inputs through several hidden which layers which consist of fully connected nonlinear activation functions followed by dropout. The final layer is a single node performing a linear combination of the hidden features, thus returning the output of the network that is taken as the predicted log-risk function.

Differently from CPH, the risk function in the Cox equation in 2.8 is replaced by the output from neural network $W_{h,\beta} \cdot X_i$, where β is the weight for the last hidden layer and h is the weight for other hidden layers of the neural network.

2.2.2.2 Nnet-Survival

Nnet-Survival [70] is a fully parametric survival model that discretizes survival time, so that the follow-up time is divided in n intervals. In these cases the survival estimates are a step function with steps at the grid points. This method was proposed to improve two main aspects of the neural network model that are adapted from Cox model: computational speed and the violation of the proportional hazard assumption. Here, hazard is defined as the conditional probability of surviving time interval j given the event is not yet verified at the beginning of interval j. There are various approaches for mapping input data to hazard probabilities [71, 72].

Log Hazard. The flexible version of Nnet-Survival is called Logistic Hazard[72] (LH). In this case the output layer has n neurons, where n is the number of time intervals, since the final hidden layer is densely connected to the output layer with a sigmoid activation function, in which log odds are converted to the conditional probability of surviving this interval. Every output neuron represents the survival probability at the specific time interval given that an individual is alive at the beginning of the time interval.

PC Hazard. Piecewise Constant Hazard (PCH) [72] has the same concept of Nnet-Survival but is a continuous-time model, meaning that it assumes that the continuous-time hazard function is constant in predefined intervals. It is similar to the Piecewise Exponential Models [72] but with a softplus activation instead of the exponential function.

2.2.2.3 Cox-Time

Cox-Time (CT) [73] is a deep neural network based on the Cox Model but relaxes the proportionally assumption. This is a parametric model that does not need to apply a stratified version of the Cox model. This becomes possible because the time is treated by model as a regular covariate and not as an output feature.

2.2.2.4 Deep Hit

Deep Hit (DH) [74] is a deep neural network that aims at learning directly the distribution of survival times. DH is a discrete-times model and makes no assumptions about the underlying stochastic process, thus allowing for a relation between covariates and risk that changes over time. The architecture of DH consists of a single shared sub-network and a family of cause-specific sub-networks. The network is trained by using a loss function that exploits both survival times and relative risks with the aim of learning $\hat{P}$, i.e. the estimate of the joint distribution of the first hitting time and competing events. As the network is constructed, each sub-net takes input via a residual connection from both the event-associated covariate vector and a latent representation produced by the shared sub-network. This gives the sub-networks access to the learned representation while still allowing them to learn non-common part of the representation as well. The output of the softmax layer i.e. the output layer, is a probability distribution $y = [y_{1,1}, \ldots, y_{1,t_{max}}, \ldots, y_{K,1}, \ldots, y_{K,t_{max}}]$: given a patient with covariates X, an output element $y_{k,t}$ is the probability that the patient will experience the event k at time t.. This innovative architecture drives the network to learn potentially non-linear and even non-proportional relationships between covariates and risks.

2.2.3 Autoencoders

Autoencoders (AEs) [75] are a class of ANNs used in the context of unsupervised learning, aiming to produce a compressed representation of input data for dimensionality reduction, data compression or feature extraction. Such models are made up of two main components:

- **Encoder**: It takes the input data and compress it into a lower-dimensional space representation, called *latent space*; this process aims to capture the most important information and compress them, while discharging the less useful ones.
- **Decoder**: It takes the latent representation and tries to reconstruct the input as closely as possible.

The latent space is the the most important part of the network, in which the compressed information is stored in a smaller in size w.r.t. the original input. For such a reason, it is also called *bottleneck)*. An example of AE architecture is depicted in Figure 2.8.

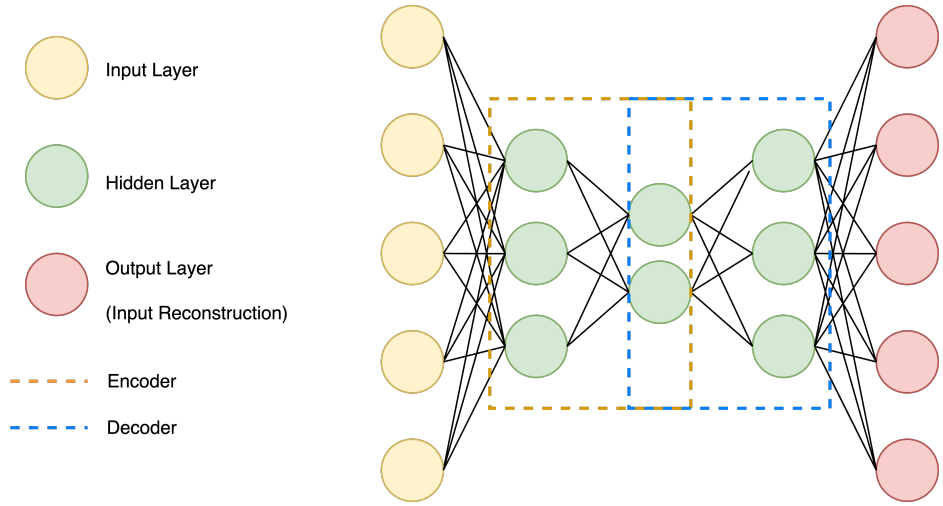


Fig. 2.8 Vanilla AE Architecture

The training process is the same of a typical ANN, aiming to minimize the difference between input data and the respective reconstruction by computing the loss function as Mean-Squared-Error (MSE) or binary CE, according to the nature of input data. According to the net architecture and the application, it is possible to distinguish several type of AE:

- **Vanilla Autoencoder.** It consists of one encoder and one decoder, where both are typically fully connected layers.
- **Convolutional Autoencoder.** For image data where spatial hierarchies are important, this type of AEs use convolutional layers to better capture spatial features.
- **Denoising Autoencoder.** The input data is intentionally corrupted (e.g., by adding noise), and the autoencoder is trained to reconstruct the uncorrupted original data, by filter out noise and focus on the essential features.
- **Sparse Autoencoder.** A regularization term is added to the loss function to enforce sparsity in the latent representation, leading to more efficient feature extraction.
- **Variational Autoencoder.** It adds a probabilistic aspect to the latent space, i.e. instead of learning a fixed compressed representation, VAEs learn to encode the input as a probability distribution. This is especially useful in data generation tasks, as the VAE to generate new data by sampling from the learned distribution.

2.3 Explainable Artificial Intelligence

AI models, especially DL ones, are often seen as black-box model as understanding the logic behind a model's decision making process is a challenging task; moreover, they also require a compromise among performance and interpretability, introducing issues in AI models trustworthiness [76].

To tackle these problems, a set of techniques called Explainable Artificial Intelligence, have been proposed for make AI models more transparent. XAI techniques may be categorized based on several criteria:

Interpretability Level

- Global methods offer an overall understanding of the relationships that a model learns from data, providing an explanation about the entire behavior of the model.
- Local methods provide insights into why the model made a specific decision for a specific instance (or a set of instance).

Approach to the explanation

- Post-Hoc explanations are generated after the model training phase, trying to understand the decision making process by examining the decisions made by the model.
- Intrinsic explanations, i.e. the explanations are easily retrievable from the model, as this last one has been designed to be understandable (e.g., Decision Trees).

Agnosticism

- Model-based methods are tailored to a specific type of model and cannot be applied to other types of models.
- Model-Agnostic methods treat the model as a black-box and try to provide an explanation regardless the underlying model architecture.

2.3.1 Shapley Additive Explanations

SHapley Additive exPlanations (SHAP), introduced by Lundberg et al.[77], is an explanation method derived from the Shapley values of the cooperative game theory [78] that quantify the contribution of the single player to the overall result generated by the entire set of players.

Lundberg et al. adapts such a concept to ML, by considering each data feature as a cooperative player that contributes to the prediction of the target variable.

The Shapley values are computed according to the following equation:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!}(f(S \cup \{i\}) - f(S)) \tag{2.14}$$

Where N is the total number of players, $f : 2^N \to \mathbb{R}$ a characteristic function, with $f(\emptyset) = 0$. Given a set of players P, $f(P)$ represents the total payoff expected from the coalition P. The Shapley value quantifies the contribution of i-th player calculating the difference of the values of f with a set including the i-th player and a set excluding the i-th player. Furthermore, it is possible to retrieve the feature importance for a set of data by averaging the SHAP values of the single samples.

Notably, SHAP can be classified as post hoc, model-agnostic, and local XAI method. Figures 2.10 2.9 show the visual explanation returned by SHAP for gene (*KRAS)* mutation prediction with trascriptomic data. Figure2.10 shows the global explanation for a set of patients related to the features importance in genetic mutation prediction, while Figure 2.9 shows the SHAP decision plot for a single patient, highlight the contribution of the features to the outcome.

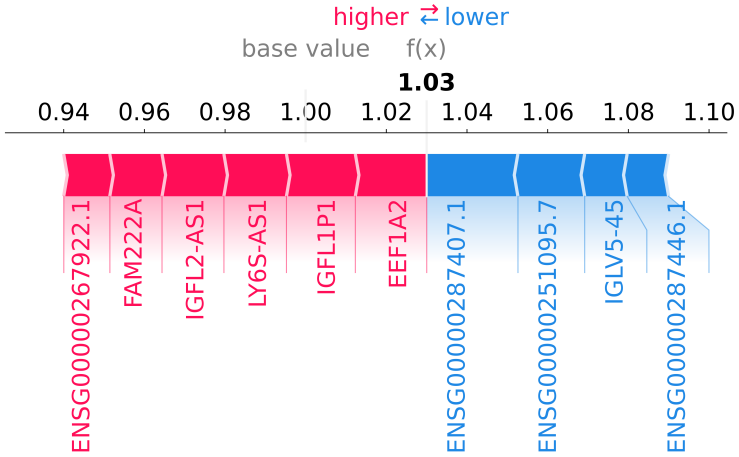


Fig. 2.9 Example of SHAP Decision Plot for a single subject. The feature importance is indicated according to the bar magnitude, while the color indicates the type of contribution (positive-red, negative-blue) to the prediction outcome.

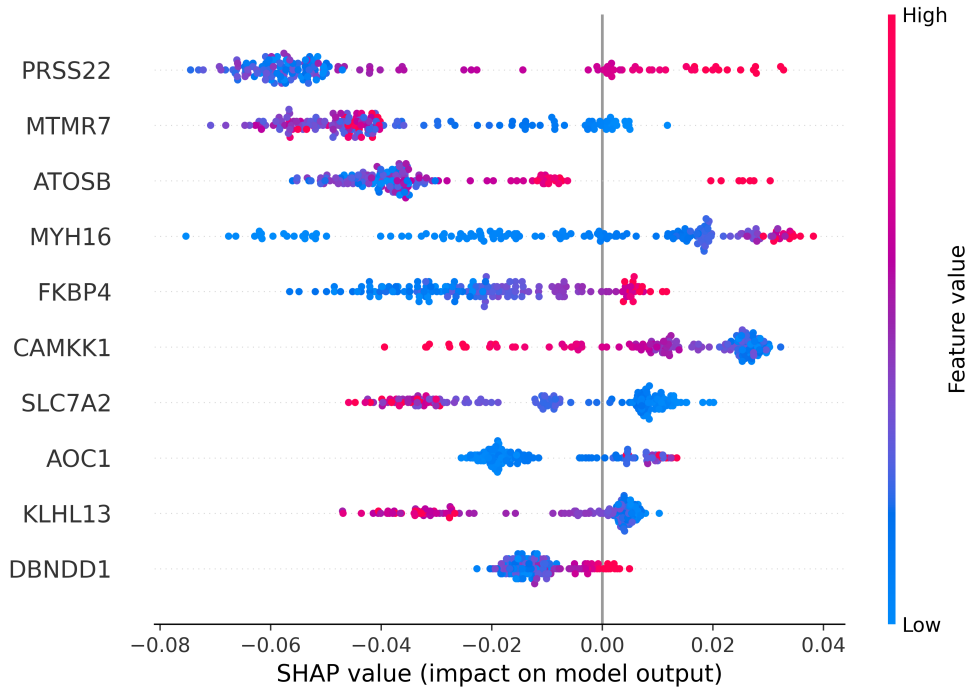


Fig. 2.10 Example of SHAP global explanation for a set of data, related to the impact of trascriptomic features on a classifier model in genetic mutation prediction. The features are sorted according to their importance, while the color indicates the feature value of a specific sample. The feature importance for classification is directly proportional to the distance of points from the origin.

2.3.2 SurvSHAP

SurvSHAP [79] generalizes SHAP to survival models, giving an explanation about the overall behavior of the model over time.

It captures variable contributions across the entire time period under study, allowing the detection of variables with time-dependent effects. Its aggregation method enhances the determination of each variable's importance for a prediction, offering a more effective approach compared to other methods. Given the data $\mathbb{D} = \{(\mathbf{X_i}, y_i, t_i)\}$ and assuming that $\mathbb{D}$ contains m unique time instants $t_m > t_{m-1} > \cdots > t_1$, with y_i as the event and t_i as the time point of interest. Therefore, each sample will be represented by covariates n, a status y, and an instant of time t. For each individual described by a variable vector $\mathbf{X}$, the model returns the individual's survival distribution $\hat{S}(t, \mathbf{X})$. For the observation of interest $\mathbf{X}_*$ at any selected time point t, the algorithm assigns an importance value $\phi_t(\mathbf{X}_*, c)$ to the value of each variable $\mathbf{X}(c)$ included in the model with $c \in \{1, 2, \ldots, n\}$ where n is the variable number. To

calculate every value of SurvSHAP function, it is necessary to define the expected value for the survival function conditioned by the values of the features:

$$e^{c}_{t,\mathbf{X}_*} = \mathbb{E}[\hat{S}(t,\mathbf{X})|\mathbf{X}^c = \mathbf{X}^c_*] \tag{2.15}$$

defining P(c,π) as the precedence subset of c in a permutation π, i.e. $\mathscr{P}(c,\pi) = \{x \in \mathbf{X} \mid x \text{ precedes } c \text{ in } \pi\}$,the contribution of the variable c to the model is calculated as:

$$\phi_t(\mathbf{X}_*, c) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} e^{\mathscr{P}(c,\pi)\cup c}_{\mathbf{X}_*,t} - e^{\mathscr{P}(c,\pi)}_{\mathbf{X}_*,t} \tag{2.16}$$

where Π is a set of all permutations of n variables, and the apexes indicate which variable contributes to the estimation of survival function. The first term of 2.16 refers to the contribution of the model with the preset up to the variable c (c included), while the second brings only the contribution of the subset without the variable c. In this way, it is possible to obtain the contribution of each single variable to the prediction. For an easier comparison among different models and time points, this value can be normalized to obtain values on a common scale from -1 to 1, so that the contribution becomes:

$$\phi^*_t(\mathbf{X}_*, c) = \frac{\phi_t(\mathbf{X}_*, c)}{\sum_{j=1}^{p} |\phi_t(\mathbf{X}_*, j)|} \tag{2.17}$$

To calculate global variable importance, the time-dependent contributions are aggregated, achieving the following Average Aggregate SurvSHAP value:

$$\Psi(\mathbf{X}, c) = \int_0^{t_m} |\phi^*_t(\mathbf{X}_*, c)| \tag{2.18}$$

Notably, the average aggregate SurvSHAP value corresponds to the average aggregated SHAP value. An example of a SurvSHAP plot is depicted in Figure 2.11.

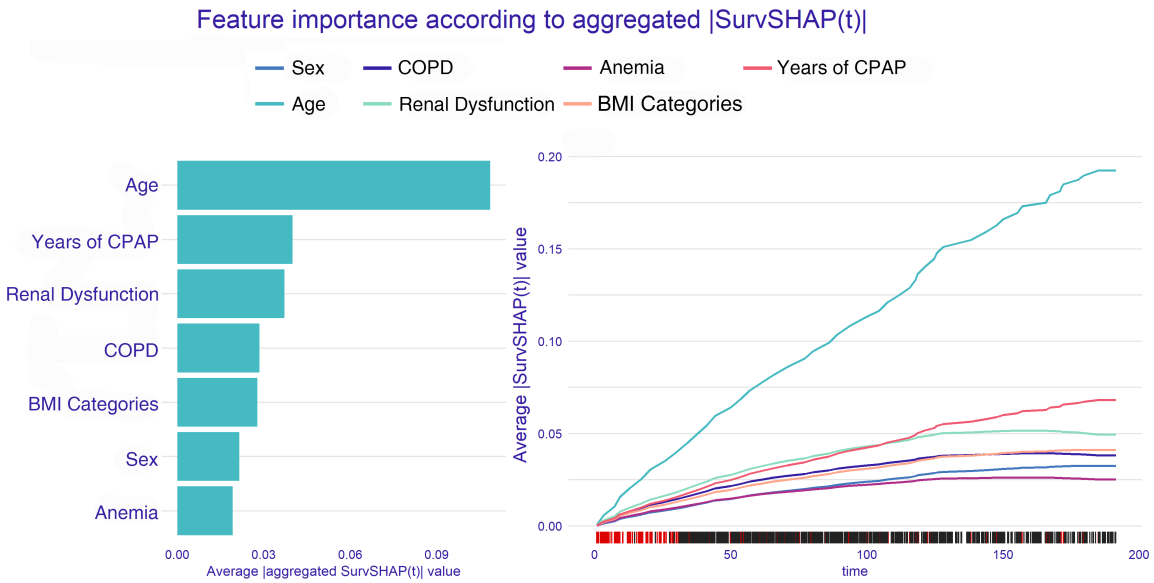


Fig. 2.11 Example of SurvSHAP explanation for CPH Model: on the left the features importance ranking according to the average of absolute shapley values. On the right the feature importance according to the observation time.

2.3.3 LIME

Local Interpretable Model-agnostic Explanations (LIME) is a well-known explanation technique based on model-agnostic phenomena, that can be applied to any DL model. Specifically:

- **Local**: states that LIME explains the behaviour of the model by approximating its local behaviour;
- **Interpretable**: emphasizes on the ability of the LIME to provide an output useful to understand the behaviour of the model from a human point of view;
- **Model-Agnostic**: means that LIME is not dependent on the model used; all models are treated as a black-box.

For imaging classification tasks, LIME takes the superpixels (a patch of pixels) of the original input image after generating a linear model, and generates several samples by exploiting the superpixels. The quick-shift algorithm is responsible for the computation of superpixels of an image. Thereafter, the perturbation images are generated and the final prediction is made.

Afterwards, a heatmap appears over the image that highlights the important pixels, i.e. regions that contribute in classification. The positively contributing features are highlighted in green while the negatively contributing superpixels are colored in red. The LIME also allows to pick a threshold value to select the number of top contributing pixels, either positively or negatively.

2.3.4 Mathematically Explained XAI

Mathematical interpretability techiniques, such as t-SNE and UMAP, aims to represent high-dimensional data in a lower-dimensional space while preserving the clustering structure.

Although both t-SNE and UMAP are primarily used for visualization, they differ in how they interpret distances between clusters. t-SNE preserves only the local structure of the data, whereas UMAP can maintain both local and global structures. This means that, unlike UMAP, t-SNE does not allow for interpreting dissimilarities and distances between clusters.

t-distributed Stochastic Neighbor Embedding. t-SNE [80] is a variation of the SNE technique that enables the visualization of high-dimensional data by mapping each data point to a location in a lower-dimensional space (typically two or three dimensions). It was developed to address two key issues in the SNE technique:

1. Optimization of the cost function, achieved by using a symmetrized version of the SNE cost function and employing a Student-t distribution to calculate the similarity between data points in the lower-dimensional space.
2. The "crowding problem," mitigated by using a heavy-tailed distribution in the low-dimensional space.

Uniform Manifold Approximation and Projection. UMAP [81] is a nonlinear technique for dimensionality reduction based on three main assumptions:

1. The data is uniformly distributed across an existing manifold.
2. The topological structure of the manifold should be preserved.
3. The manifold is locally connected.

The UMAP method consists of two primary phases: learning a manifold structure in a high-dimensional space and representing it in a lower-dimensional space. In the first phase, the nearest neighbors of each data point are identified using the Nearest-Neighbor-Descent algorithm.

Next, UMAP constructs a graph by connecting these identified neighbors. Since the data is uniformly distributed across the manifold, the spacing between data points varies based on regions of higher or lower density. This assumption allows for the introduction of *edge weights*: for each point, the distance to its nearest neighbors is calculated. The edge weights between data points are then computed, although there may be issues with conflicting edges.

2.3.5 Class Activation Maps and Attention Maps

For imaging data it is possible visualize the regions of the image in which the model focused the most for prediction, according to the type of classification model. This can be achieved by the use of post-hoc, model-based XAI method such as Gradient-weighted Class Activation Mapping (Grad-CAM) [82] for CNNs, that uses model's gradients to produce a class activation map highlighting the regions of an input image that are most influential in the model's decision-making process for a particular class.

Concerning attention-based model like transformers, it is possible to use the projection of the attention weights on the input image for visualizing the distribution of the attention over the input image. Differently from saliency maps (like Grad-CAM maps) that highlight the importance of specific image pixels/region w.r.t. the prediction, the attention maps focus on the entire input, showing how different parts affect each other.

Figure 2.12 illustrates a histological WSI along the respective attention map retrieved by a classifier model, highlighting the most influent regions for a classifier in genetic mutation prediction task.

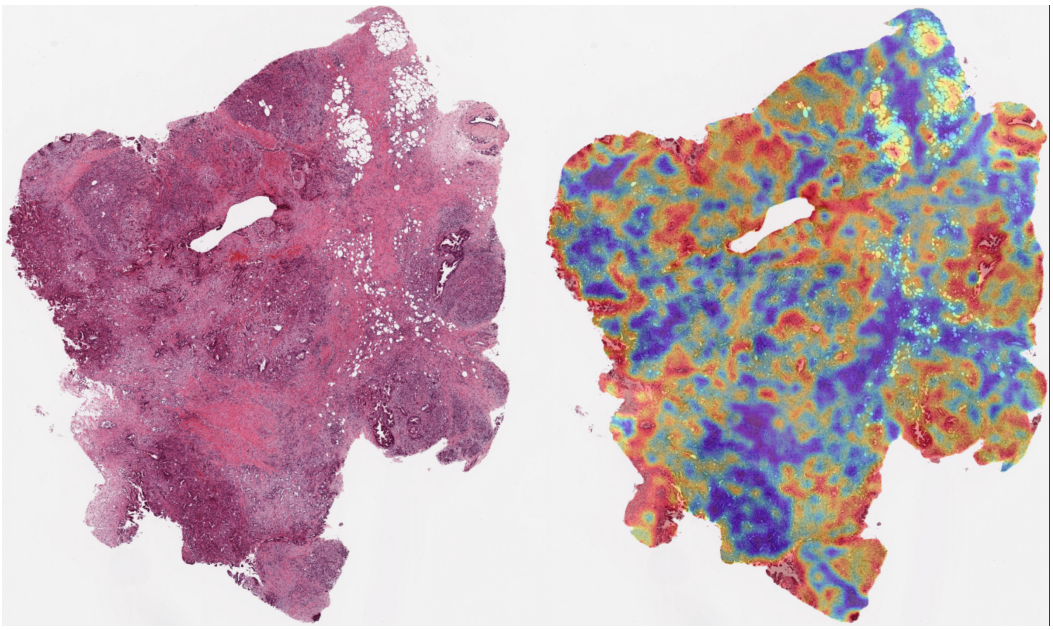


Fig. 2.12 Example of Attention-Map for CLAM model in genetic mutation prediction. The most influent regions that contribute positively to the outcome are highlighted in red.

2.4 Evaluation Metrics

A crucial aspect of model training and validation consists in the assessment of a model performance according to the task involved. This is accomplished using several metrics depending on the type of problem, like classification and survival analysis.

2.4.1 Classification

A standard method for evaluating the performance of a classifier is given by the *confusion matrix*, depicted in Figure 2.13.

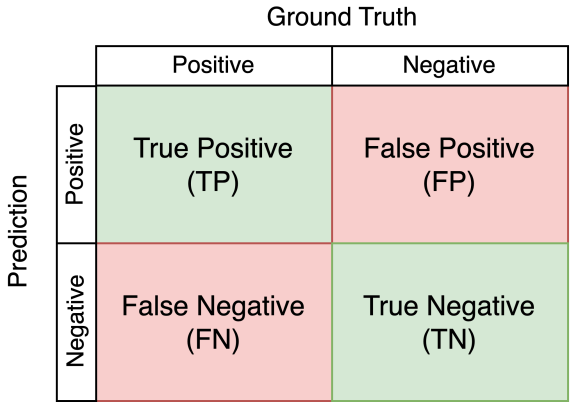


Fig. 2.13 Confusion Matrix for binary classification problem.

As shown in Figure 2.13, the confusion matrix is composed by True Positive (TP) and True Negative (TN) as the number of samples correctly predicted for the positive and negative class, respectively, and by False Positive (FP) as the number of samples incorrectly predicted as positive class when the actual class was negative, and by False Negative (FN) as the number of samples incorrectly predicted as negative class when the actual class was positive.

Starting from such values it is possible to calculate several performance index:

- Accuracy: it measures the proportion of correct predictions out of the total predictions.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.19}$$

- Precision: it measures the proportion of true positive predictions out of all predicted positives.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.20}$$

- Recall: it measures the proportion of actual positives that are correctly predicted; It is also known as Sensitivity or True Positive Rate (TPR)

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2.21}$$

- F1-Score: It is the harmonic mean of precision and recall, providing a score that balances both metrics.

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.22}$$

- Specificity: It indicates the proportions of actual negatives that are correctly predicted; it is known also as True Negative Rate (TNR)

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{2.23}$$

As stated in Section 2.1.1, the predicted value of a classification model is a probability score and, setting a threshold, it is possible to decided whether belongs to a class or not. Setting several thresholds it is possible to obtain several models, with different classification results. Based on this assumption, it is possible to plot the Receiver Operating Characteristic (ROC) curve, representing the proportion between the Recall (TPR) and the False Positive Rate (FPR), computes as:

$$\text{FPR} = \frac{FP}{FP+TN} \tag{2.24}$$

the ROC curves compares the Recall with the FPR, by varying different probability thresholds. The ideal ROC curve has a $TPR = 1$ and $FPR = 0$. The model performance is retrieved by computing the Area Under Curve (ROC-AUC).

Notably, the ROC curve suffers from data unbalancing problem leading to biased performance metrics. For mitigating such a problem, the Precision-Recall (PR) curve with related PR-AUC should be computed, along with ROC curve. Both curves are depicted in Figure 2.14

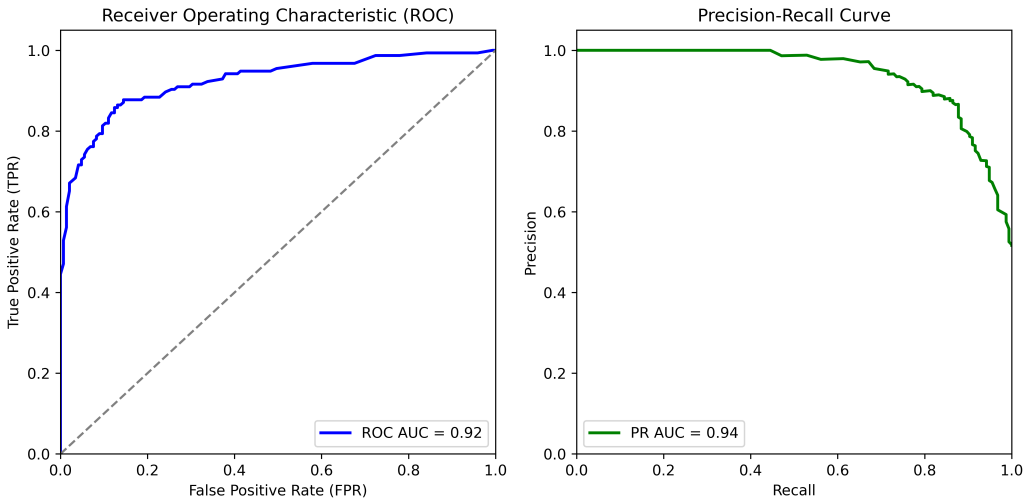


Fig. 2.14 Example of ROC curve and PR curve for binary classification problem, on the left ROC curve with related AUC, on the right PR curve with related AUC.

2.4.2 Survival Analysis

For SA tasks the models can be evaluated in terms of three different performance metrics:

- Harrell's C-index [83] – Also known as the Concordance Index, it assesses the proportion of all observation pairs for which the model's predicted survival order corresponds to the actual survival order in the data. A higher C-Index (ideally 1) reflects better concordance between the model's predictions and the ground truth, whereas a C-Index of 0.5 indicates a random prediction. The formula for the C-index is as follows:

$$CI = \frac{N_{CP}}{N_{CP} + N_{DP}} \tag{2.25}$$

  where N_{CP} is the number of concordant pairs and N_{DP} is the number of discordant pairs.

- Integrated Cumulative-Dynamic Area Under the Curve (C/D AUC) [84] – This metric evaluates the area under the ROC curve at various points throughout the observation period. The C/D AUC is calculated by integrating the AUC over time:

$$\text{C/D AUC} = \frac{1}{T}\int_0^T \text{ROC-AUC}(t), dt \tag{2.26}$$

- Brier Score [85] – This score measures the difference between the model's predictions and the actual outcome. A lower Brier score indicates better model accuracy, while a higher score suggests performance decline. The Brier score is given by:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{2.27}$$

where $p_i$is the predicted probability of the event for observation i-th ant, y_i is the actual outcome (0 or 1); a value of 0.5 reflects random predictions. Notably, the Brier-Score is similar to the MSE, as they both measures the squared difference between ground truth and predicted value:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.28}$$

2.5 Machine Learning in Radiomics

Radiomics is a quantitative approach applied to medical images (such as CT, MRI, or PET scans), aimed at extracting quantitative data from images by using sophisticated and non-intuitive mathematical analyses. The radiomic approach is based on the assumption that lesions possess phenotypic traits, often imperceptible to the human eye, but rich in valuable information that allows for the characterization of various lesions [86].

According to the type of the information content, radiomic features can be classified in the following categories: intensity-based, morphological, and textural features. Intensity-based features, also referred to as first-order features, describe the distribution of pixel/voxel intensity values within the Region-Of-Interest (ROI) or Volume-Of-Interest (VOI) by analyzing the intensity histogram. Morphological features capture the geometric attributes of the region, taking into account spatial relationships between pixels or voxels. Textural features, which are second-order statistics, provide insights into the spatial arrangement of intensity values. These textural features are computed using various data structures, often based on Gray-Level Matrices, such as Gray-Level Co-occurrence Matrix (GLCM) [87], Gray-Level Size Zone Matrix (GLSZM) [88, 89], Gray-Level Length Matrix (GLRLM) [90], the Gray-Level Neighboring Gray Tone Difference Matrix (NGTDM) [91], and the Gray-Level Dependence Matrix (GLDM) [92].

Noteworthy, radiomic analysis can produce a high dimensional feature vector, which includes potential redundant information. Additionally, the number of extracted features can exceed the sample size, which may reduce the study's statistical power and generalizability

(curse of dimensionality). Thus, employing feature selection algorithms or dimensionality reduction techniques is recommended to create an informative, reproducible, and non-redundant feature vector [93], also known as *radiomc signature*.

After radiomics feature extraction and dimensionality reduction [94], the radiomic signature is given as input to ML models for tumor lesions analysis. This process showed promising results in lesions classification [95, 96] and Overall Survival (OS) and Recurrence (REC) prediction [97].

Radiomics in Pancreatic Cancer Studies Radiomics-based approaches proved to be effective in characterization and analysis of several cancer types, such as lung [98], breast [99], gastric [100], and pacreatic cancers.

Focusing on the last one, a radiomics-based signature has been proposed to differentiate between PDAC and pancreatic adenosquamous carcinoma (PASC) with high accuracy [101]. Qiu et al. [102], exploits radiomics approaches have for discriminating PDA histological subtypes, obtaining 0.77 of Accuracy and a ROC AUC of 0.79. In the context of prognostic analyses for PDA, radiomics has proven to be an effective tool for predicting lymph node metastasis, which in many studies is considered an independent risk factor for overall survival (OS) due to its high prevalence [103]. Parr and colleagues developed a prognostic model based on seven first- and second-order radiomic features extracted from wavelet-transformed images. Their study demonstrated that a model combining clinical and radiomic predictors outperformed a purely clinical model in predicting disease recurrence, achieving a concordance index (C-index) of 0.78 compared to 0.66 for the clinical model [97]. Similarly, Xie et al. integrated a radiomic score into a clinical nomogram to predict both disease-free survival (DFS) and OS. For their test set, the radiomic nomogram achieved C-indexes of 0.70 and 0.73 for DFS and OS, respectively, outperforming the tumor, nodal, and metastatic (TNM) staging system, which is the established clinical prognosticator [104].

Additionally, Khalvati and colleagues identified Sum Entropy and Cluster Tendency as the most robust radiomic features in predicting OS, with HR of 1.56 ($p = 0.005$) and 1.35 ($p = 0.022$), respectively, for their test set [105]. Despite the potential for radiomic features to correlate with clinical and biological patterns, the clinical translation of radiomics-based pipelines remains limited. Finally, Keyl et al. [106] utilized SHAP in a multimodal survival prediction model for advanced pancreatic cancer. While the authors incorporated three radiomic features into their analysis, they did not evaluate their impact over time. Consequently, one of the primary limitations of existing

radiomics studies is the absence of a framework capable of explaining the contribution of these radiomic prognosticators when they are integrated into survival models with censored data.

Table 2.1 Summary of related works on radiomics-based studies on pancreatic cancer.

Study	**Objective**	**Results**
Qiu et al. [102]	Discrimination of PDA histological subtypes using radiomics aprroaches	0.77 of Accuracy and a ROC AUC of 0.79.
Ren et al. [101]	Radiomics signatore for differentiation between PDA and PASC	0.94 of Accuracy, 0.98 of Sensitivity, 0.90 of Specificity .
Li et al. [103]	Prediction of lymph node metastasis in PDA	ROC AUC of 0.912.
Parr et al [97]	Prognostic multimodal model for PDA recurrence (REC)	C-Index of 0.78 .
Xie et al. [104]	Prediction of DFS and OS in PDA	C-index of 0.70 and 0.73 for DFS and OS.
Khalvati et al. [105]	Identification of robust radiomic features for predicting OS	Sum Entropy and Cluster Tendency were identified as robust radiomic features for predicting OS, with HR of 1.56 (p = 0.005) and 1.35 (p = 0.022), respectively.

2.6 Deep Learning in Digital Pathology

Digital Pathology refers to the process of digitizing pathology slides, typically derived from biopsies or tissue sections, for diagnostic and research use. High-resolution scanners are used to convert histological slides into digital images WSI, which can then be analyzed using DL approaches. Pathomics is an extension of DP that focuses on the extraction of quantitative features, known as *pathomic features*, from digital tissue images. Such features can be derived from various aspects of the WSI, such as morphology, texture, cellular structures, and the spatial organization of cells within the sample. Under this aspect, DL models for

WSIs processing such as CNNs and ViTs play a pivotal role in the realization of biomedical DSS [107].

Notably, WSIs are characterized by:

- Resolution: They capture tissue at very high resolutions, which means they offer detailed views at cellular levels. These images can be zoomed in and out, enabling fine-grained analysis, but their large size presents challenges for computational processing.
- Size: They are extremely large files, ranging from gigabytes to tens of gigabytes depending on the size of the tissue and the magnification level.
- Format: They are usually stored in specialized formats (e.g., SVS, NDPI, TIFF) that allow for pyramidal representations, enabling efficient navigation through different levels of magnification.

In light of this, classification tasks of WSIs is challenging due to their large size and complexity. Such a task may be accomplished in two ways:

1. Dividing it in two sub-tasks, a segmentation and classification task. The former can be accomplished by a segmentation model for object detection and instance segmentation within the WSI, while the latter by a classification model for the instances extracted. Then the WSI-level score is computed according the single instance classification.
2. Approaching it as a MIL problem, by treating the WSI as bag of instances, aiming at finding a relationship between the single instances and the global WSI label.

Deep Learning for Genetic Mutations Prediction in Pancreatic Cancer Pancreatic cancer is associated with genetic mutations that affect both tumor suppressor genes, such as *TP53*, *SMAD4*, and *CDKN2A*, and oncogenes, such as *KRAS* [108–111]. Mutations in these genes are present in more than 50% of the cancer cases, and have earned the name of "four mountains" [39, 112].

KRAS mutations are remarkably prevalent in pancreatic tumors, occurring in 90–95% of pancreatic adenocarcinomas, making *KRAS* the most commonly mutated gene in this form of cancer [113]. Oncogenic *KRAS* mutations are critical in the initiation and progression of PDAC by promoting the generation of reactive oxygen species (ROS) through metabolic alterations. The elevated ROS levels activate key signaling pathways that drive PDAC development [114].

SMAD4 functions as a key mediator in the Transforming Growth Factor Beta (TGF-β) signaling pathway, regulating essential cellular processes such as cell growth, differentiation, apoptosis, and migration [115]. In the context of tumorigenesis, *SMAD4* plays a vital role in triggering cell-cycle arrest and apoptosis, which are essential mechanisms for regulating cell proliferation and removing damaged cells [115].

TP53 mutations, appearing in a range from 50% to 90% of PDAC cases, have a strong impact on carcinogenesis, prognosis and response to treatment [111]. The *TP53* gene, situated on chromosome 17, serves as a tumor suppressor by managing cell division. Hence, mutations affecting this gene result in uncontrolled cell division [111].

CDKN2A mutations are highly significant in pancreatic tumors, with somatic mutations found in up to 95% of cases and a genetic predisposition seen in familial instances. There is a clear link to an increased risk of pancreatic cancer, and families with *CDKN2A* germline mutations may show signs of a pancreatic cancer-melanoma syndrome [116].

In last years, the task of predicting genetic mutations from histopathology images gained significant attention due to advancements in Deep Learning (DL) and Computational Pathology. Such approaches aim at extracting genetic information without the need for genetic sequencing.

By leveraging CNN and multimodal learning approaches, researchers have made progress in integrating image features with genomic data. On one side, histopathology images can be used for complex tasks such as predicting genetic alterations of key genes, tumor composition, and prognosis [117]. On the other hand, complementing different omics (including those derived from images, e.g. radiomics and pathomics [118]) can improve predictive accuracy of models employed on tasks such as prognosis [119**?**].

Although several pan-cancer studies concerned mutation status prediction, there are very few ones that include PDAC cases [120]. In 2020, Kather et al. [121], exploited weakly supervised learning approaches with CNN models for inferring a wide range of genetic mutations, molecular tumor subtypes, gene expression signatures and pathology biomarkers from histological images. They reported Area Under ROC curve (AUROC) scores of 0.67 for *KRAS*, 0.45 for *SMAD4*, 0.51 for *TP53*, and 0.24 for *CDKN2A*, with corresponding Area Under Precision-Recall curve (AUPRC) scores of 0.70, 0.17, 0.58, and 0.12, respectively. Komura et al. [122] proposed a deep texture representations for predicting several combinations of genomic features and cancer

types from hematoxylin-and-eosin-stained (H&E) images. Their model achieved AUROC scores of 0.61 for *KRAS*, 0.51 for *SMAD4*, 0.60 for *TP53*, and 0.54 for *CDKN2A*. Recently, Saldanha et al. [123] provided a self-supervised feature extraction method followed by an attention-based multiple instance learning model for pan-cancer mutation prediction from histopathology. Their results showed AUROC scores of 0.58 for *KRAS*, 0.47 for *SMAD4*, 0.44 for *TP53*, and 0.61 for *CDKN2A*.

Another interesting aspect concerns the adoption of other omics to infer the mutational status, which can allow biological discovery among different omic layers, reaching a more comprehensive knowledge of tumor biology. According to Crawford et al. [124], transcriptomic data is the most effective omic to predict the mutation status. Furthermore, they stated that adding other omics into a multi-omic model does not improve the prediction of mutation status. Hence, we also considered transcriptomic data, to provide a broader perspective on the task of predicting the mutation status, and to have another modality comparison for the pathomic models. Furthermore, this allowed us to investigate a multimodal fusion based on pathomics and transcriptomics.

Chapter 3

Multimodal Pipelines for Pancreatic Ductal Adenocarcinoma Analysis

This chapter introduces the first pipelines developed: two multimodal big data analytics pipelines for the analysis of Pancreatic Ductal Adenocarcinoma (PDAC) cases.

According to the methods investigated in the previous chapter, the first section deals with the role of ML in radiomics with particular focus on PDAC, for Overall Survival and Recurrence prediction. The developed multimodal pipeline includes multi-omics features, i.e. Radiomics, Clinical and Genomics and provides several survival analysis models for OS and REC estimation; the achieved performance are assessed in terms of C-index. Finally, an innovative time-dependent XAI method (survSHAP) is applied to the best survival models for investigating their behavior, finding the most relevant features useful for patients clinical evaluation.

The second section, investigates the application of DL methods for genetic mutations prediction in PDAC cases using pathomic and trascriptomic models. Specifically, an attention-based MIL approach is exploited for imaging classification, while Differentially Expressed Genes (DEG) analysis and deep AEs are exploited and compared for mutations prediction with transcriptomic data using classical ML models. Then the two unimodal approaches are combined, obtaining multimodal predictions. Finally, the use of attention-maps and SHAP method allows for retrieving an explanation at both pathomic and transcriptomic level.

3.1 Multimodal analysis from the multi-omic cohort of CPTAC-PDA

3.1.1 Contribution

The primary objective of the work "A time-dependent explainable radiomic analysis from the multi-omic cohort of CPTAC-Pancreatic Ductal Adenocarcinoma" [27] was to develop a time-dependent, explainable survival model for patients with pancreatic ductal adenocarcinoma by integrating radiomic, clinical, and mutational features. Four survival machine learning (SML) classifiers were designed, trained, and validated to predict overall survival (OS) and recurrence (REC) in PDAC patients. The data used for this study were obtained from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) project, which provides multiple public multi-omic datasets, including annotated CT images, clinical, and mutational data. Additionally, the mechanisms behind the decision-making processes of the survival algorithms were explored using SurvSHAP(t), allowing for a deeper understanding of the most significant contributing features and their impact on survival probability over different time intervals.

In summary, this work offered three main contributions:

1. A comprehensive investigation of the clinical, mutational, and radiomic data from the CPTAC-PDA project, aimed at developing accurate and explainable multi-omic prognostic models for PDAC.

2. A systematic comparison of various SML classifiers built on multi-omic predictors, utilizing a leave-one-out cross-validation (LOOCV) approach to study OS and REC.

3. An innovative explainability analysis for translational purposes, focusing on the radiomic determinants featured in the best risk prediction models, achieved through the use of the time-dependent, model-agnostic explainability algorithm, SurvSHAP(t).

3.1.2 Datasets

Patients' data were obtained from the National Cancer Institute's (NCI) CPTAC-PDA cohort [125]. This cohort comprises 170 patients with PDAC, including treatment-naive and surgically resected tissue samples, collected from multiple translational research centers to promote advancements in proteomics.

Radiology Images. Radiology CT images were collected from 98 patients enrolled across six international radiology departments participating in the CPTAC-PDA project. Tumor-annotated CT images were provided by clinical investigators for 87 of these patients [126]. Radiology images from the CPTAC-PDA cohort were collected and made publicly available by The Cancer Imaging Archive (TCIA). Of the 489 CT series available, 298 were labeled with anatomical structures such as the pancreas and its ducts. Among these, 134 series included tumor annotations based on a clinical annotation protocol; the scans were obtained using a multi-slice CT system, and the images were downloaded in DICOM format, while their corresponding annotations were provided in RTSTRUCT (DICOM Radiotherapy Structure Sets) format.

Clinical and Mutational Data. The data included clinical outcomes (vital status, disease recurrence or progression, and follow-up [FU]), demographic data (age at diagnosis and gender), and pathology and clinical data (including tumor stage, tumor grade, and the American Joint Commission on Cancer [AJCC] TNM status according to the 8th edition, clinical response, and residual disease). Frailty data, such as the Eastern Cooperative Oncology Group performance status (ECOG) and the Karnofsky performance status, were also incorporated. Features with more than 50% missing values were excluded from the analysis.

For standardization, age at diagnosis, Body Mass Index (BMI), and the largest tumor diameter were categorized based on clinical criteria: age was categorized at 65 years (high vs. low), BMI at 25 (overweight vs. underweight), and tumor diameter at 3.5 cm (larger vs. smaller). Discrete variables such as the number of lymph nodes involved, ECOG score, Karnofsky score, and clinical response were dichotomized into two classes: at least one involved lymph node vs. none, ECOG score of 3–5 vs. 0–2, Karnofsky score $> 0\%$ vs. 0%, and responders (complete and partial) vs. non-responders (progressors and stable disease). Surgical residual disease was categorized as totally resected (R0), partially resected (R1/R2), or resection not defined (RX).

3.1.3 Proposed Approach

The workflow involved three main stages and can be summarized as follow:

1. Data preparation: from the whole dataset, radiomics, clinical, and mutational features have been retrieved to prepare single omics and the combinations of multi-omic datasets. Then a dichotomizing phase of continuous radiomic features according to the clinical outcomes has made for translational purposes.

2. Training and Validation: the survival models considered have been trained with LOOCV method. Then, predicted responses have been globally evaluated in a vector combining every risk of occurrence of either OS or REC through the C-index assessment.

3. The best models for OS and REC prediction were selected according to C-index and their behavior has been investigated using a time-dependent XAI algorithm, i.e. SurvSHAP(t).

The full processing pipeline is depicted in Figure 3.1

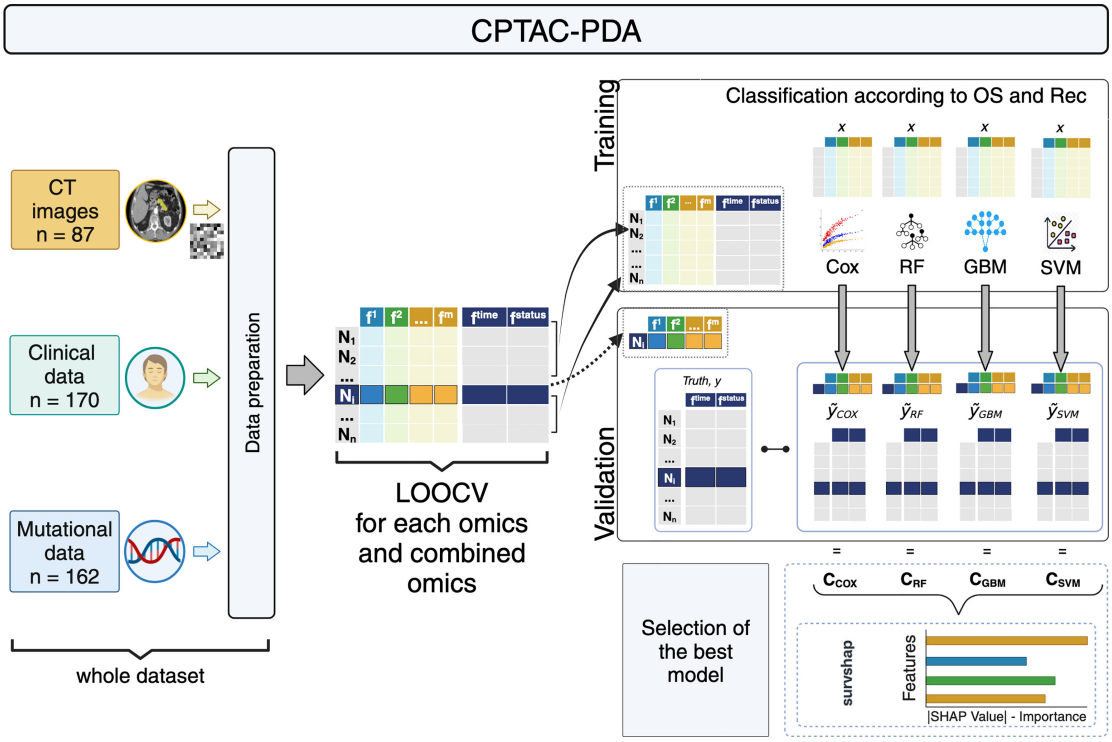


Fig. 3.1 Multimodal Processing Pipeline for OS and REC prediction in PDAC.

3.1.4 Data Preparation

Mutational Data. Based on whole-exome sequencing (WXS) analysis, 251 out of 270 tumor samples exhibited genetic alterations. As shown in the oncoplot in Figure 3.2A, the top 10 mutated genes were KRAS (81% of samples), TP53 (64%), CDKN2A (19%), SMAD4 (19%), TTN (11%), MUC16 (7%), CSMD1 (6%), RYR2 (6%), KMT2D (6%), and RYR1 (5%). Mutations with a frequency below 5% were excluded from further analyses.

Radiomics Features. A total of 87 patients with annotated tumor CT scans were included, comprising 134 series acquired during the enhancement phases of arterial (AR), portal venous (PV), and delayed (De) contrast medium phases. For the analysis, CT scans from the PV phase were used when available (n = 75 patients). In cases where the PV phase was unavailable, the AR phase was selected (n = 11 patients). In one case where both PV and AR phases were unavailable, the De phase was chosen (n = 1 patient). Figures 3.2B and 3.2C provide two examples of series acquired during the AR and PV phases, shown in the axial, coronal, and sagittal planes.

Feature extraction was performed from the original images, those filtered with the Laplacian of Gaussian (LoG), and the wavelet-transformed VOIs. The sigma parameter for the LoG filter ranged from 1 to 5, with increments of 1 [mm]. For the wavelet-transformed images, eight decompositions (LLL, LLH, LHL, LHH, HLL, HLH, HHL, HHH) were obtained by applying combinations of low-pass (L) and high-pass (H) filters, using a Coiflet 1 mother wavelet on the 3D volumes. The values for the sigma parameter and the wavelet decompositions were selected based on prior research [98, 127]. All images were resampled to a spacing of 1 [mm] in each direction using the sitkBSpline interpolator, and discretized with a bin width of 25, which has shown good reproducibility and performance in previous studies [128].

PyRadiomics tool was utilized to extract all shape, first-order, and second-order features from GLRLM, GLSZM, GLDM, and Neighboring Gray Tone Difference Matrix (NGTDM). The software requires that both the images and the tumor masks be converted to the NIfTI (Neuroimaging Informatics Technology Initiative) format. Consequently, the SimpleITK library was employed to convert the original DICOM series into NIfTI images, and the rt-utils package was used to transform the DICOM-RTSTRUCT annotations into NIfTI masks.

A total of 1,288 quantitative radiomic features were extracted. After performing a correlation analysis (Fig. 2D), a subset of 16 features for OS and 14 features for REC were retained for further analysis.

A correlation analysis was applied for dimensionality reduction of radiomic features eligible for the next steps. Such step was performed by computing the Pearson Correlation for each feature and setting a cut-off of $|0.3|$ to select features with the lowest inter-correlations. The selected features are reported in Table 3.1.

Table 3.1 GLCM radiomics features considered for analyses.

GLCM Feature	**Description**
Autocorrelation	Measure of the magnitude of the fineness and coarseness of texture
JointAverage	Returns the mean gray level intensity of each distribution
ClusterProminence	Measure of the skewness and asymmetry of the GLCM
ClusterShade	Measure of the skewness and uniformity of the GLCM
ClusterTendency	Measure of groupings of voxels with similar gray-level values
Contrast	Measure of the local intensity variation
Correlation	Value between 0 (uncorrelated) and 1 (perfectly correlated) showing the linear dependency of gray level values to their respective voxels in the GLCM
DifferenceAverage	Measures the relationship between occurrences of pairs with similar intensity values and pairs with differing intensity values
DifferenceEntropy	Measure of the randomness/variability in neighborhood intensity value differences
DifferenceVariance	Measure of heterogeneity that places higher weights on differing intensity level pairs that deviate more from the mean
JointEnergy	Measure of homogeneous patterns in the image
JointEntropy	Measure of the randomness/variability in neighborhood intensity values
IMC1	Informational Measures of Correlation-1 quantifies the complexity of the texture
IMC2	Informational Measures of Correlation-2 quantifies the complexity of the texture
IDM	Inverse Difference Moment, a measure of the local homogeneity of an image
IDMN	IDM normalized
ID	Inverse Difference, another measure of the local homogeneity of an image
IDN	ID normalized
InverseVariance	Inverse Variance
MaximumProbability	Occurrences of the most predominant pair of neighboring intensity values
SumEntropy	Sum of neighborhood intensity value differences
SumSquares	Measure of the distribution of neighboring intensity level pairs about the mean intensity level in the GLCM

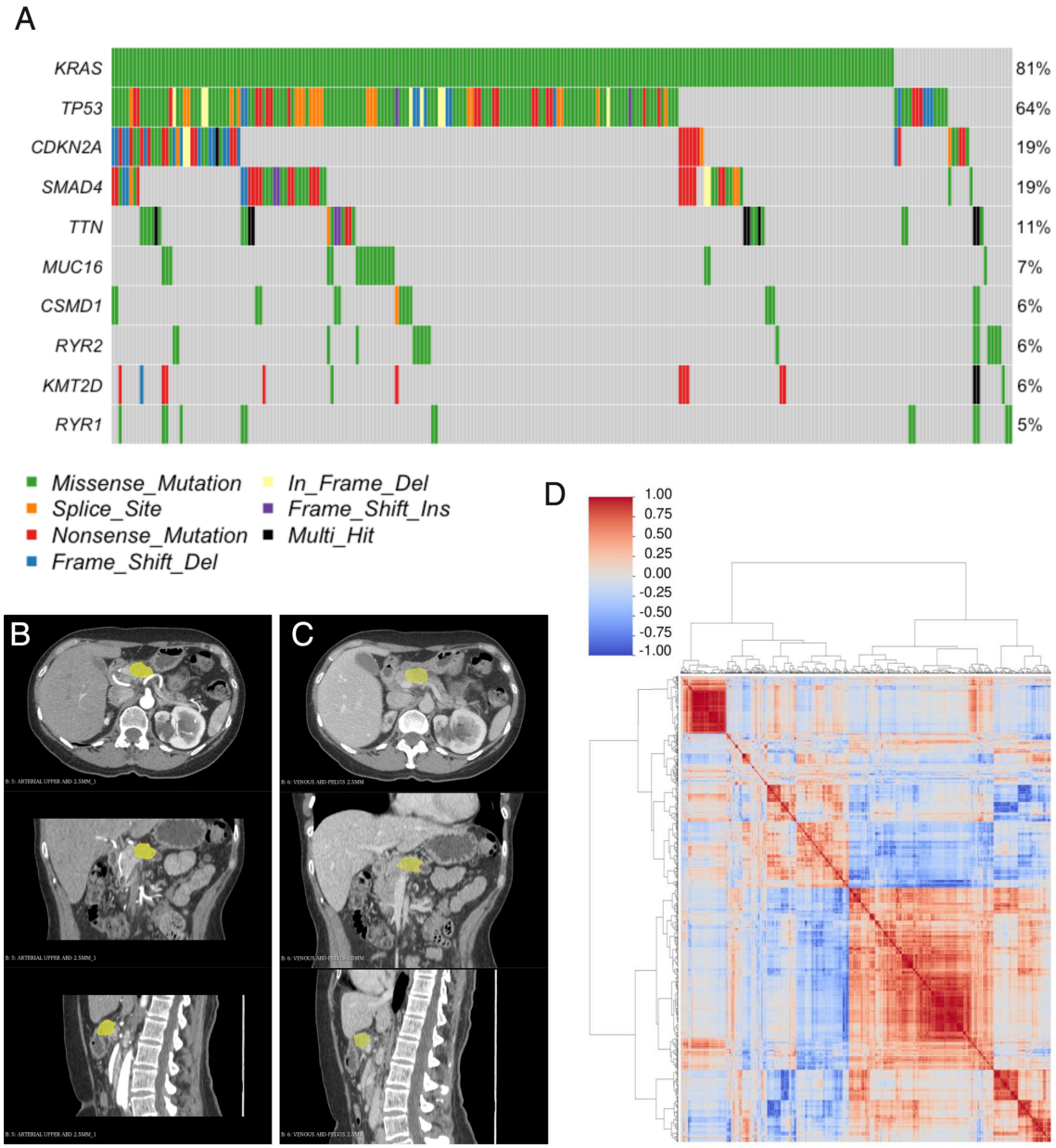


Fig. 3.2 Data preparation for mutational data and radiology images. (A) Oncoplot of top ten mutated genes. (B-C) Example of AR/PV phase series shown in the axial, coronal, and sagittal planes. (D) Correlation Matrix heatmap of the initial 1,288 radiomic features.

The integration of clinical, mutational data, and annotated CT images allowed the identification of two distinct cohorts of patients:

- A cohort of 60 patients with complete data and available OS.

- A cohort of 49 patients with complete data and available REC, after excluding 4 stage IV patients for clinical reasons and 7 patients due to the unavailability of REC data in the clinical database.

3.1.5 Feature Selection Through Survival Analysis

Quantitative levels of each radiomic feature were dichotomized into two groups (high or low) based on a cutoff identified by maximally selected rank statistics, using the surv_cutpoint function from the survminer R package, as applied in various translational studies [129, 130]. Each feature was then evaluated for significance after applying the univariate (UV) survival model for overall survival (OS) or recurrence (REC). Only features that significantly discriminated clinical outcomes with a p-value < 0.15 in the UV analysis were considered eligible for multivariate (MV) analysis, conducted using CPH models. Features that discriminated clinical outcomes with a p-value < 0.10 were considered eligible for subsequent steps. The p-values for the univariate (UV) and MV analyses were derived from pairwise comparisons, using log-rank statistics for the UV analysis and z statistics for the MV analysis.

Overall Survival Based on the univariate (UV) analysis of patients who experienced overall survival (OS), 13 out of 16 radiomic features were included in the multivariate (MV) analysis: Original first-order Kurtosis, Original first-order 90 Percentile, Original shape Elongation, Original GLCM Joint Energy, LoG.sigma.1.0.mm.3D GLDM Dependence Variance, LoG.sigma.1.0.mm.3D first-order Skewness, LoG.sigma.2.0.mm.3D first-order Median, Wavelet.LLH GLCM Cluster Tendency, Wavelet.HLL GLCM Inverse Variance, Wavelet.HLH GLCM IMC1, Wavelet.HHL first-order Mean, Wavelet.HHL GLCM IMC1, and Wavelet.HHL GLSZM Large Area Emphasis. Following the MV analysis, 6 out of the 13 radiomic predictors were selected for inclusion in multi-omic models: Original first-order 90 Percentile (high vs. low levels, HR = 2.45, p-val = 0.090), Original shape Elongation (high vs. low, HR = 0.27, p-val = 0.005), Original GLCM Joint Energy (high vs. low, HR = 11.98, p-val = 0.010), LoG.sigma.1.0.mm.3D first-order Skewness (high vs. low, HR = 0.50, p-val = 0.078), LoG.sigma.2.0.mm.3D first-order Median (high vs. low, HR = 0.18, p-val < 0.001), and Wavelet.HLH GLCM IMC1 (high vs. low, HR = 0.24, p-val = 0.018).

For the clinical features, based on both UV and MV analyses, 3 out of 11 predictors were selected for the multi-modal models: gender (M vs. F, HR = 0.46, p-val = 0.013), tumor grade (G3 vs. G1/G2, HR = 3.11, p-val = 0.001), and residual disease (R0 vs. R1/R2, HR = 0.73, p-val = 0.076).

Regarding mutational features, 2 out of 10 genes were included in the MV analysis based on the UV analysis: KRAS and TTN (Table 3). Ultimately, the MV analysis identified the TTN gene as retaining a prognostic impact, with HR = 3.32, p-val = 0.008, for mutated patients versus non-mutated patients.

Recurrence Based on the univariate (UV) analysis of patients who experienced recurrence (REC), 10 out of 14 radiomic features were included in the multivariate (MV) analysis: Original first-order Kurtosis, Original GLCM Correlation, Original shape Elongation, Original GLCM Joint Energy, Original shape Least Axis Length, LoG.sigma.2.0.mm.3D GLSZM Small Area Low Gray Level Emphasis (SALGLE), LoG.sigma.2.0.mm.3D first-order Median, Wavelet.LLH first-order Variance, Wavelet.HHL first-order Mean, and Wavelet.HHL GLSZM Large Area Emphasis (LAE). Following the MV analysis, 2 out of 10 radiomic predictors were selected for inclusion in the multi-omic models: Original shape Elongation (high vs. low, HR = 4.23, p-val = 0.044) and LoG.sigma.2.0.mm.3D first-order Median (high vs. low, HR = 0.35, p-val = 0.024).

For the clinical features, based on both UV and MV analyses, only residual disease was selected to be included with the other determinants in the multi-modal models (R0 vs. R1/R2, HR = 0.66, p-val = 0.075).

Regarding the mutational features, following UV and MV analyses, 2 out of 10 genes were selected for inclusion in the multi-omic models: patients with mutated KRAS (HR = 2.83, p-val = 0.056) and SMAD4 (HR = 0.47, p-val = 0.094) compared to non-mutated patients.

The Kaplan-Meier curves related to the feature selected are represented in Figure 4.11A for OS cohort and Figure 4.11B for REC cohort.

3.1.6 OS and REC Prediction

The selected features were subsequently combined in a multi-omic approach. Seven different models were compared: radiomic features only (1), clinical features only (2), mutational features only (3), a combination of radiomic and clinical features (4), radiomic and mutational features (5), clinical and mutational features (6), and a comprehensive model integrating radiomic, clinical, and mutational features (7). Due to the significant dataset imbalance, both overall survival (OS) and recurrence (REC) analyses were conducted using LOOCV. The SML models involved are the following: CPH, SRF, Survival SVM and Survival GB model (see Chapter 2, Section 2.1.2); the models hyperparameters are reported in Table 3.2.

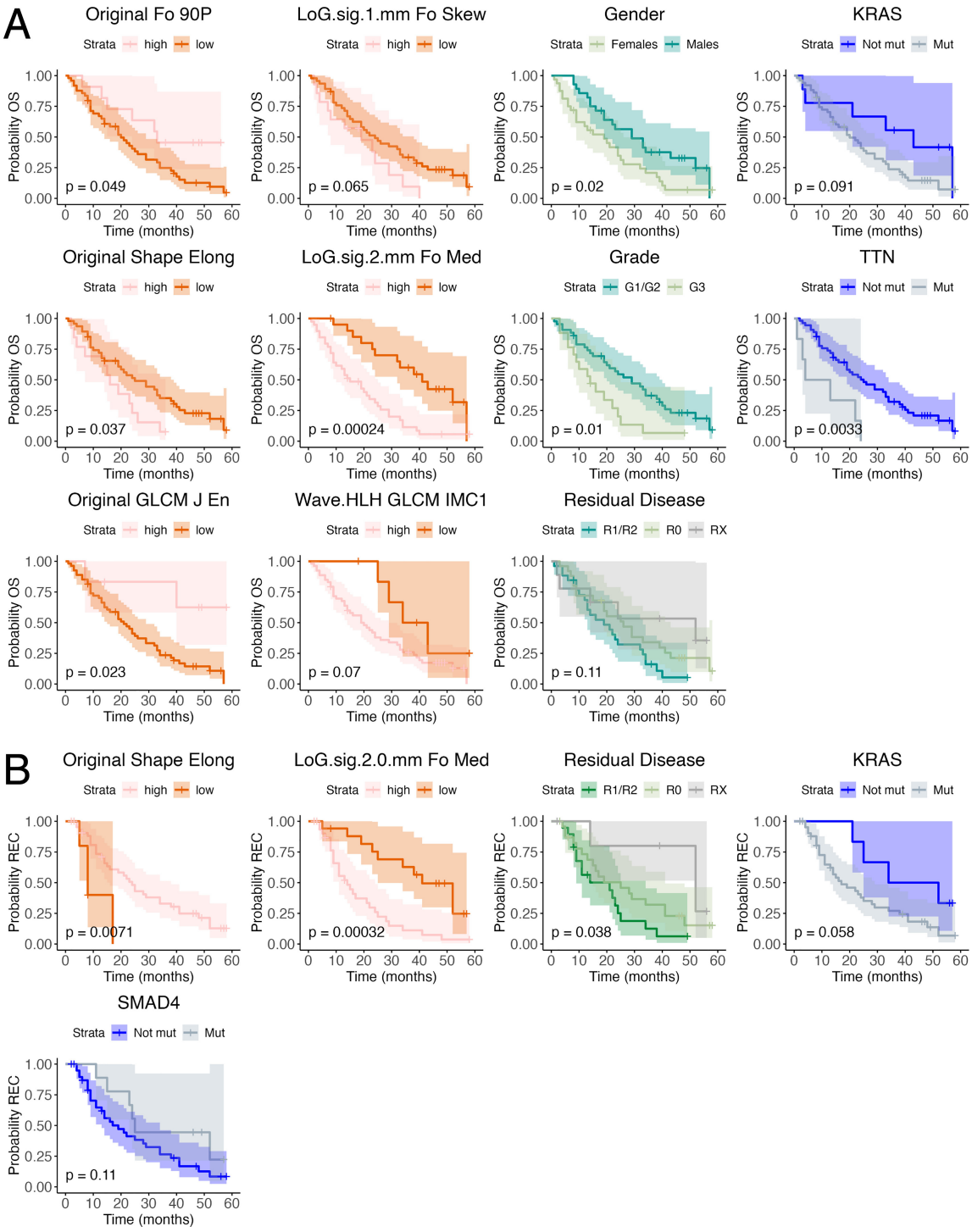


Fig. 3.3 Survival curves related to the feature selected. (A-B) Univariate Kaplan-Meier curves (OS-REC) for features retained after the multivariate analysis.

Table 3.2 List of hyper-parameter adopted for each classifier. Abbreviations: RF, random forest; GB, generalized boosted; SVM, support vector machine.

Classifier	Parameters
COX	Convergence tolerance = 1e-09 Max Iterations = 20 Tolerance for infinite parameters = 1e-09 Tolerance for Cholesky decomposition = 1e-10
RF	Number of trees = 500 Subset of features for split = floor() with p = number of features Node size = 15 Max Trees Depth = Not prefixed limit Split rule = logrank
GBM	Number of trees = 100 Interaction Depth = 10 Bag Function = 0.9 Shrinkage = 0.001
SVM	Gamma = 0.2 Coefficient type estimation = Vanbelle2 Time difference method = makediff3

3.1.7 Results

Using the available data retained after pre-processing operations and considering the following:

- The classifiers employed (Cox, SRF, survival GB, and survival SVM);
- Each individual omic and various combinations of omics;
- The clinical outcomes predicted (OS and REC);

a total of 56 different models were compared based on their C-index. Overall, for each classifier, models that included radiomic predictors in single-omic models or combined with other omics, outperformed models without radiomic features for both OS and REC. For instance, in the COX model for OS, adding radiomic features to the TTN feature increased the C-index from 46% to 75% (Fig. 3.4A). For REC, the combination of radiomics with clinical features in the GB model increased the C-index from 52% to 66% (Fig. 3.4B).

For OS, Cox classifiers outperformed SRF, survival GB, and survival SVM, with an average C-index of 66.6%, compared to 63.1%, 60.4%, and 51.1%, respectively. The best-performing COX classifiers, achieving a C-index of 75%, combined (i) radiomics and mutational data, (ii) radiomics, clinical, and mutational data, (iii) radiomics and clinical data (Fig. 3.4A). CPH combining radiomics, clinical, and mutational data was selected as best model for XAI analysis.

For REC, Cox, SRF, and survival SVM classifiers outperformed survival GB, with mean C-indexes of 59.0%, 59.9%, and 58.4%, compared to 52.3%, respectively (Fig. 3.4A). Notably, SVM classifiers incorporating only radiomic features achieved the highest performance for both OS and REC, with C-indexes of 57% and 68%, respectively. This last model was selected as best model for XAI analysis.

The comparison among survival model performance is depicted in Figure 3.4.

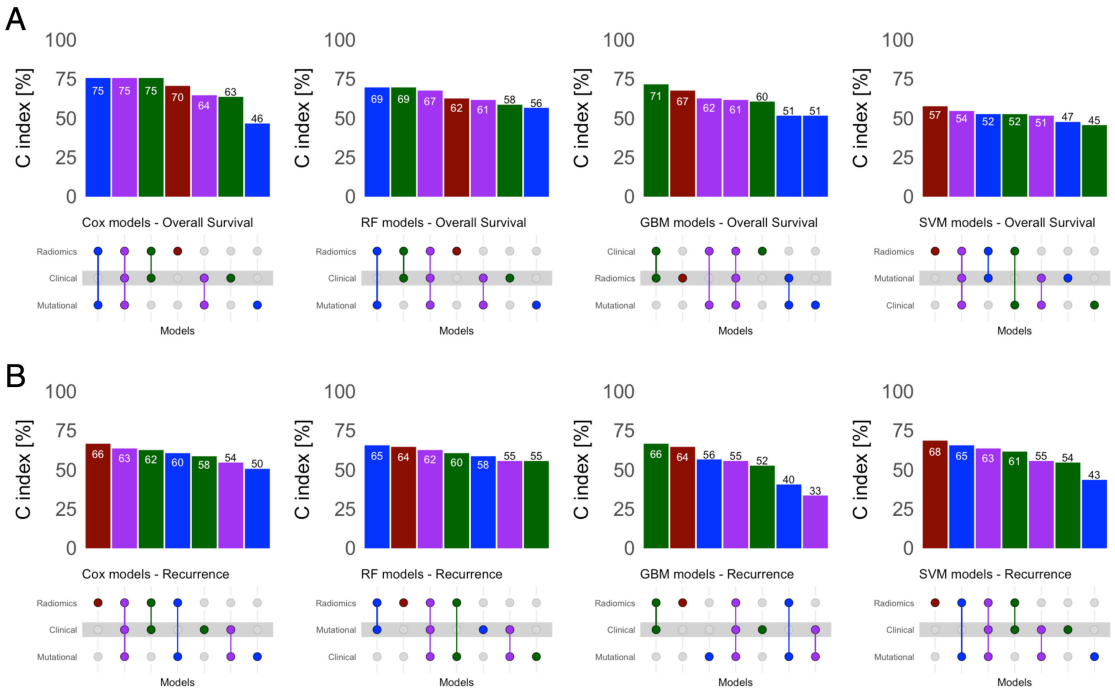


Fig. 3.4 Model performance comparison among survival models, in terms of C-index. (A) Performance comparison with multi-omic approach for OS prediction. (B) Performance comparison with multi-omic approach for REC prediction.

3.1.8 Time-Dependent Explainability

The results from the validation set of the explainability analysis for the evaluated COX and SVM classifiers, based on OS and REC, are shown in Figure 3.5, with Absolute Average SHAP values on the left part and the SHAP values over time on the right part.

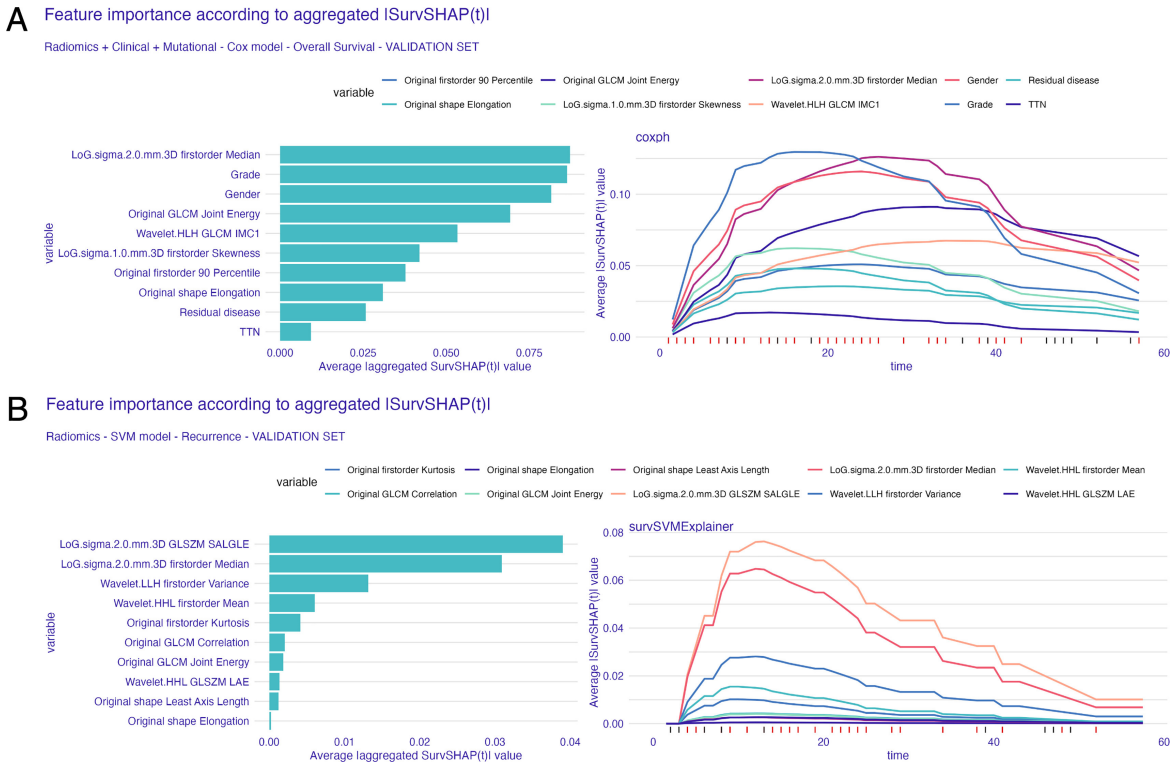


Fig. 3.5 Feature Importance for OS (A) and REC(B).

For the observation of 60 patients as a function of their follow-up, the most important variables were identified as follows (Figure 3.5A):

- LoG.sigma.2.0.mm.3D first-order Median, which reached the highest average |SurvSHAP(t)| of 0.13 at 26 months.
- Grade, with the highest average |SurvSHAP(t)| of 0.13 at 16 months.
- Gender, with the highest average |SurvSHAP(t)| of 0.12 at 24 months.
- Wavelet.HLH GLCM IMC1, which achieved the highest average |SurvSHAP(t)| of 0.10 at 34 months.
- Original GLCM Joint Energy, with the highest average |SurvSHAP(t)| of 0.09 at 32 months.
- LoG.sigma.1.0.mm.3D first-order Skewness, which recorded the highest average |SurvSHAP(t)| of 0.06 at 16 months.
- Original first-order 90 percentile, which reached the highest average |SurvSHAP(t)| of 0.05 at 24 months.

- Original shape Elongation, with the highest average |SurvSHAP(t)| of 0.05 at 25 months.
- Residual disease, which reached the highest average |SurvSHAP(t)| of 0.04 at 21 months.
- TTN, with the highest average |SurvSHAP(t)| of 0.02 at 13 months.

Patients with high values of LoG.sigma.2.0.mm.3D first-order Median, with grade G3, low values of Wavelet.HLH GLCM IMC1, females, and high values of Original GLCM Joint Energy were associated with a higher risk of overall survival (OS) compared to their complementary values. Interestingly, the TTN gene did not contribute as significantly as the other determinants.

In the second case, for recurrence (REC) observed in 49 patients, the explainability of the SVM classifier, which included radiomic features, identified the following as the most important (Figure 3.5B):

- LoG.sigma.2.0.mm.3D GLSZM SALGLE, with the highest average |SurvSHAP(t)| of 0.08 at 13 months.
- LoG.sigma.2.0.mm.3D first-order Median, which reached the highest average |SurvSHAP(t)| of 0.07 at 12 months of follow-up.
- Wavelet.LLH first-order Variance, with the highest average |SurvSHAP(t)| of 0.03 at 12 months.

Patients with low values of LoG.sigma.2.0.mm.3D GLSZM SALGLE, LoG.sigma.2.0.mm.3D first-order Median, and Wavelet.LLH first-order Variance had a higher risk of recurrence compared to patients with higher values (Figure S2B and Figure S4). Finally, according to SurvSHAP(t), the contributions from the remaining radiomic variables were negligible.

3.1.9 Discussion

In the context of PDAC, recognized clinical and biological prognosticators are insufficient for accurately stratifying patients, highlighting the need for additional tools. When effectively combined with existing tools, radiomics can offer clinicians alternative and more precise methods for prognosis. The release of PDAC-annotated CT images by the CPTAC project has contributed to the uniqueness of this cohort, enabling multi-omic studies of the disease. This study compared four different classifiers and seven multi-omic datasets to explore

the role of radiomics in PDAC prognosis. Using a LOOCV strategy, OS and REC were analyzed, with feature selection based on UV and MV survival analyses. The approach demonstrated that models incorporating radiomic signatures outperformed those using only clinical and mutational predictors. The best models were further subjected to time-dependent explainability analysis using SurvSHAP(t) to assess the local contributions of radiomic features, either independently (in the REC analysis) or in combination (in the OS analysis) with other omics.

For OS, the radiomic signature identified elongated (Original shape Elongation) and asymmetric (LoG.sigma.1.0.mm.3D first-order Skewness) tumor masses with high gray-level intensities (Original first-order 90 Percentile and LoG.sigma.2.0.mm.3D first-order Median). These tumors were spatially characterized by homogeneous (Original GLCM Joint Energy) and complex (Wavelet.HLH GLCM IMC1) textures. This signature was refined by including gender and grade variables, both of which proved significant in the explainability process. Not surprisingly, grade, a well-established prognostic factor for PDAC, was included in the model.

In contrast, for REC, the radiomic signature (Fig. 5B) identified tumor masses with high gray-level intensities (LoG.sigma.2.0.mm.3D first-order Median, Wavelet.HHL first-order Mean) and spread gray-level intensities (Wavelet.LLH first-order Variance and Original first-order Kurtosis). These tumors were spatially characterized by the joint distribution of smaller size zones with lower gray-level values, independent of VOI rotation (LoG.sigma.2.0.mm.3D GLSZM Small Area Low Gray Level Emphasis).

SurvSHAP(t) is a novel, standardized, and easy-to-apply tool designed to explain survival models with censored observations. Notably, in the multi-omic model used for OS analysis, between 0 and 20 months, the most important variables were grade and LoG.sigma.1.0.mm.3D first-order Skewness. Between 20 and 40 months of follow-up, LoG.sigma.2.0.mm.3D first-order Median and gender became more significant, while the importance of grade decreased. Second-order radiomic features (Original GLCM Joint Energy and Wavelet.HLH GLCM IMC1) reached their peak importance later. This generalization of SHAP could assist researchers in designing future prospective studies to analyze diverse patient multi-omics cohorts, considering the limitations of both standard Cox and machine learning-based survival models [79]. Additionally, the results from SurvSHAP(t) analysis confirmed those from the feature selection phase. For OS, features with the highest |SurvSHAP(t)| values also had the most significant hazard ratios (HR) in the MV analyses, as seen with LoG.sigma.2.0.mm.3D first-order Median, grade, gender, Original GLCM Joint Energy, and Wavelet.HLH GLCM IMC1. In contrast, SurvSHAP(t) indicated that the

contribution of the TTN mutation was negligible for predicting OS. Interestingly, while TTN is frequently mutated in other studies, its prognostic impact in PDAC remains controversial and is not yet scientifically recognized [131, 132].

Limitations Although the CPTAC-PDA cohort is both multi-centric and multi-omic, integrating CT images with clinical and mutational datasets significantly reduced the number of complete cases available for analysis. This limitation necessitated the use of a LOOCV strategy to mitigate the risk of overfitting. LOOCV allows the use of all available data for evaluation and provides a more reliable accuracy estimate compared to other methods, such as k-fold cross-validation, particularly when the sample size is small. Additionally, since the radiomic models were trained on CT series acquired at different phases (75 in the PV phase, 11 in the AR phase, and 1 in the De phase), this heterogeneity may have slightly impacted their stability. Nevertheless, an external test set would be necessary to further validate these findings. For the survival analysis, only dichotomized features were included, which, while simplifying clinical translation, may also reduce the precision of the prognostic predictions.

3.2 Pathomics and Transcriptomics for Genetic Mutation Prediction in PDAC

3.2.1 Contribution

The research goal of the study "A Multimodal Framework for Assessing the Link between Pathomics, Transcriptomics, and Pancreatic Cancer Mutation" [38] was to design and develop an explainable multimodal pipeline for genetic mutation predictions in PDAC cases, from transcriptomic and pathomic data. The target genes considered are the most mutated ones in PDAC cases, *KRAS*, *TP53*, *SMAD4* and *CDKN2A* [39]. Specifically, two CLAM model configurations, as well as three different feature extractors were employed for image analysis. Concerning the transcriptomics (RNA-seq), a panel of 60,660 different transcripts was pre-processed with two different pipelines: (i) a Differentially Expressed Genes (DEG) analysis; (ii) an unsupervised DL approach based on three autoencoders (AE) architectures (small, medium, big). The pre-processed transcript panels were given as input to three ML models: a RF, XGB and MLP for gene mutation classification (wild-type vs mutated). A fusion layer followed the output of unimodal models (pathomics and transcriptomics), combining the output of such models and obtaining a multimodal prediction. Then, for each gene, a performance comparison (in terms of AUROC and AUPRC) among the combined models has been made in reference to the corresponding unimodal models. Finally, Attention-maps and SHAP methods have been employed for the models' explainability, allowing for a deeper understanding of the most contributing features, from both the pathomic and transcriptomic models, respectively.

3.2.2 Datasets

For this study two public datasets were used as training set and independent test set:

1. the TCGA-PAAD dataset includes newly-diagnosed Pancreatic Adenocarcinoma (PAAD) patients' data from The Cancer Genome Atlas (TCGA) project [133].

2. the CPTAC-PDA includes newly diagnosed PDAC patients' data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) project [125].

For both of them, WSI data were aligned with the transcriptomic data and genetic alteration labels. This step ensured that only the samples having histopathology, transcriptomic, and genetic data at the same time were retained, ensuring consistency for the further analyses.

Considering that a relation one-to-many exists between patients and slides/samples (*e.g*, one patient can be associated to more WSIs and transcriptomic samples) the final dataset is reported in Table 3.3:

Table 3.3 TCGA-PAAD and CPTAC-PDA Datasets Summary.

Project	**Data Type**	**Number of Subjects**		**Number of Samples**		**Mutations Frequency On Final Dataset**			
		Original	**After Filtering**	**Original**	**After Filtering**	**KRAS**	**SMAD4**	**TP53**	**CDKN2A**
TCGA	WSI	183	168	209	188	0.59 (111)	0.21 (41)	0.55 (105)	0.16 (31)
	RNA-Seq	162	162	162	162	0.60 (97)	0.21 (35)	0.57 (93)	0.17 (29)
CPTAC	WSI	147	124	489	367	0.85 (314)	0.15 (58)	0.72 (264)	0.16 (61)
	RNA-Seq	151	128	192	161	0.83 (143)	0.15 (24)	0.70 (114)	0.20 (32)

3.2.3 Proposed Approach

Once retrieved data from GDC data portal, the workflow comprised of several stages, summarized as follows:

1. Data Preparation. As first step, both imaging and transcriptomic samples were filtered on the availability of each gene mutation status assumed as target (*KRAS*, *TP53*, *SMAD4*, and *CDKN2A*), leading to the creation of the datasets reported in Table 3.3.

2. Data Processing. Since we have built a multimodal pipeline, this module is composed of two sub-blocks:

   (a) Histopathology Data Feature Extraction. Three feature extractors were used for image processing: ResNet50, UNI, and CONCH (using only the image encoder). The tissue were first segmented, then patched and provided as input to the feature extractors.

   (b) Transcriptomic Data Feature Extraction. Transcriptomic data were processed in two ways: (i) using a DEG analysis to retain only the most differentially expressed genes, and (ii) using three Deep Vanilla AEs [75] with latent space dimensions of 64, 128, and 256, to extract compact latent representations from the whole transcriptome.

3. Models Training and Prediction Ensemble. For WSI data classification, two single-branch CLAM models (large and small versions, see Section 3.2.5) were trained using Monte Carlo 10-fold cross-validation. For RNA-Seq data classification, RF, XGB, and MLP models were trained with a 10-fold cross-validation using both DEG and AE-processed data, independently. The predicted probabilities were ensembled between

each model-fold and at the subject level. Performance assessments were conducted using AUROC and AUPRC metrics.

4. Model Ensemble. After retaining the unimodal models, the output from the softmax function of each classifier was aggregated to obtain a combined prediction.

5. Model Explanations. For the best performing multimodal models, attention maps and SHAP plots were generated to investigate their behavior at both global and local levels.

The full pipeline is depicted in Figure 3.6.

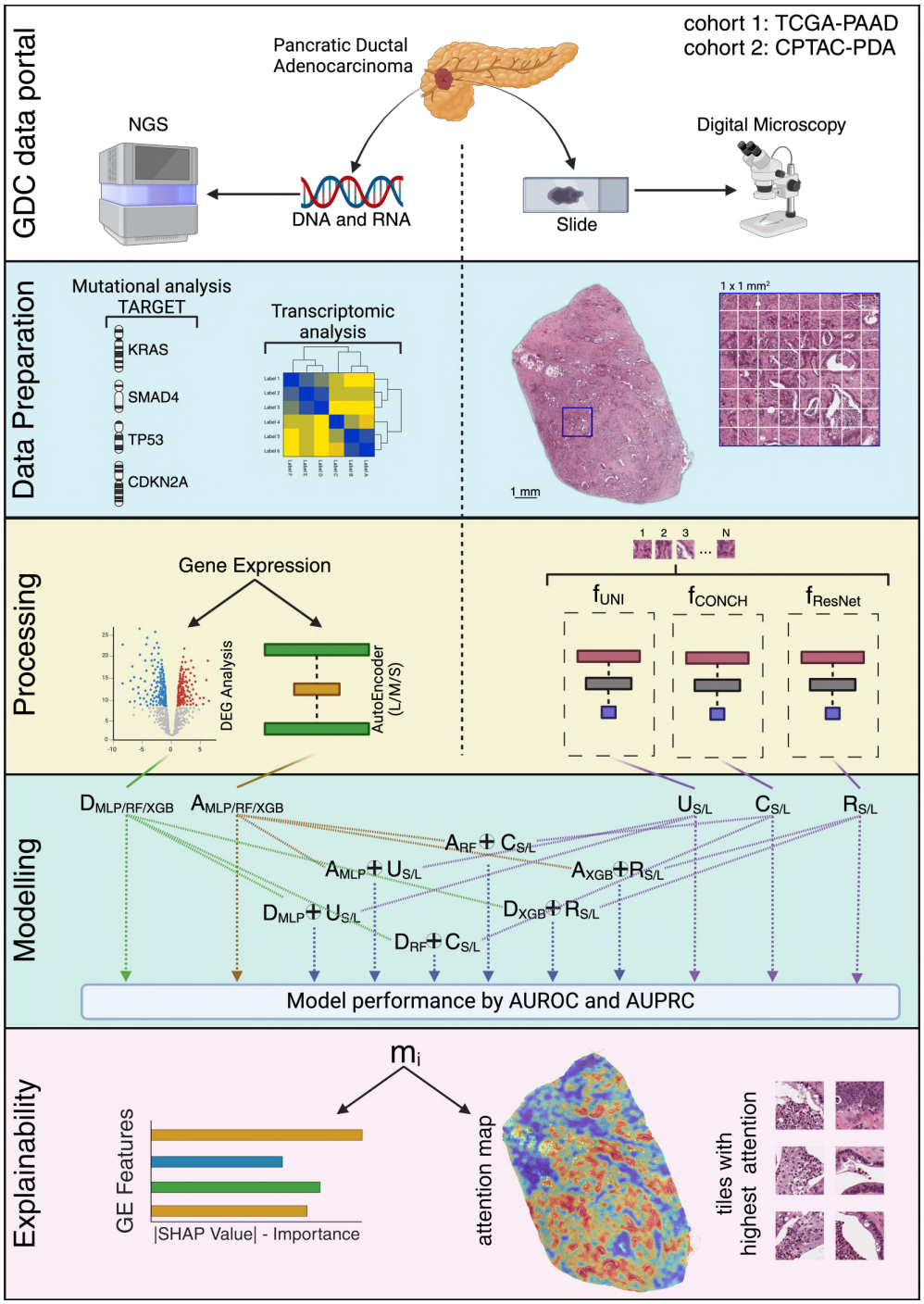


Fig. 3.6 Overview of Multimodal Processing Pipeline. Imaging data were analyzed using a CNN and two foundation models for feature extraction, followed by a CLAM model for classification. Two distinct dimensionality reduction techniques, namely DEG analysis and Deep AEs, were considered for transcriptomic. Finally, both transcriptomic and pathomic data were classified using ML models. The predictions from both branches were combined to produce a multimodal output, and attention-maps and SHAP values were visualized to interpret the logic behind model's predictions.

3.2.4 Data Preparation

As first step, the original imaging and transcriptomic datasets were filtered based on the availability of labels for all target mutations to ensure consistency across the subsets with respect to targets. Subjects missing any target mutation label were excluded from the analysis. After samples filtering, the whole datasets were pre-processed in the following way:

WSI. A total of 555 WSIs were retained, 188 for training and 367 for test, respectively. A segmentation operation was performed as first step, detecting the tissue and separating it from the background. TCGA image data were acquired at $\times 40$ magnification, while CPTAC image data were scanned at $\times 20$ magnification. In order to uniform the region sizes considered, since only pyramid levels of 1, 4, 16 and 32 were available for TCGA images, during the patching phase TCGA data patches were extracted with a size of 512×512 (at maximum resolution, corresponding to $\times 40$ magnification), and then resized to 256×256 (corresponding to $\times 20$ magnification) while CPTAC patches were extracted at 256×256. The patches were then used to feed the feature extractors: ResNet50, UNI, and CONCH, obtaining a total of 6 different feature vectors for each image (considering 3 feature extractors and 2 CLAM configurations).

Transcriptomics. The whole panel of gene expression data was composed by $60,660$ different transcripts and a dimensionality reduction operations was required. Specifically, two different modalities were tested and compared:

- For each target gene, a DEG analysis was performed on TCGA-PAAD data with DESeq2 [134]. DeSeq2 normalizes the absolute counts of reads using the median ratio method, where the counts for each gene are divided by the geometric mean of the gene across all samples, and the median of these ratios is used to normalize each sample. Then, a generalized linear model (GLM) is fitted to the data, which incorporates the dispersion estimates. Next, a negative binomial model is fitted to each gene's expression data across samples. Finally, the Wald test is performed for calculating the statical significance of each feature. Resulting p-values are then adjusted (padj) for multiple testing using the Benjamini-Hochberg correction [135], which controls the false discovery rate; this ensures that the reported DEGs are statistically reliable. Only the feature with $padj < 0.05$ were considered differentially expressed genes. The Log2 fold change (FC) was computed to determine of if a gene is up-regulated ($FC > 1$) or down-regulated ($FC < -1$).

- An unsupervised approach using three Deep Vanilla AEs with latent space dimensions of 64, 128 and 256 was implemented. For readability purposes, the three AEs were defined according to their latent space dimension, *i.e.*, AE-Small (AE-S, 64), AE-Medium (AE-M, 128), and AE-Large (AE-L, 256). Due to the high dimensionality of transcriptomic feature data, the transcripts were first reduced to $5,000$ by extracting those with the highest Median Absolute Deviation (MAD), before feeding the AEs. Given a dataset $D = x_1, x_2, \ldots, x_n$ composed of n transcripts $x_i, i = 1, ..., n$, where each x_i contains m observations $x_{ij}, j = 1, ..., m$, the median of each transcript was computed

$$M_i = \text{median}(x_{i1}, x_{i2}, \ldots, x_{im}) \tag{3.1}$$

Followed by the computation of absolute deviations from the median:

$$|x_{ij} - M_i| \quad \text{for } j \in 1, ..., m \tag{3.2}$$

Finally the MAD for each transcript $i = 1, ..., n$ was computed as:

$$MAD_i = \text{median}(|x_{i1} - M_i|, |x_{i2} - M_i|, \ldots, |x_{im} - M_i|) \tag{3.3}$$

The features were ranked according to the MAD and only the top $5,000$ were retained; this approach is very effective as also demonstrated by other authors [136]. Before feeding the models, the features were normalized using a Robust Scaler, *i.e.* subtracting the respective medians and dividing them by the interquartile range. The AEs were trained for reducing the input dimensionality to the aforementioned dimensions and reconstructing it to the original one; the Mean-Squared-Error (MSE) function was chosen as loss function used for models training. The use of such a metric allowed for evaluating the average offset between original and reconstructed data. Each model was trained by setting *adam* as optimizer and a total number of epochs set to 500, with an EarlyStopping criterion using a tolerance of 100 epochs, according to the decrease of validation loss. TCGA data was split into training and validation sets in proportions of 80% and 20%, respectively. After AEs training and inference phases, three new datasets were obtained with a features number of 64, 128 and 256, respectively.

A straightforward representation of both procedures is reported in Figure 3.7.

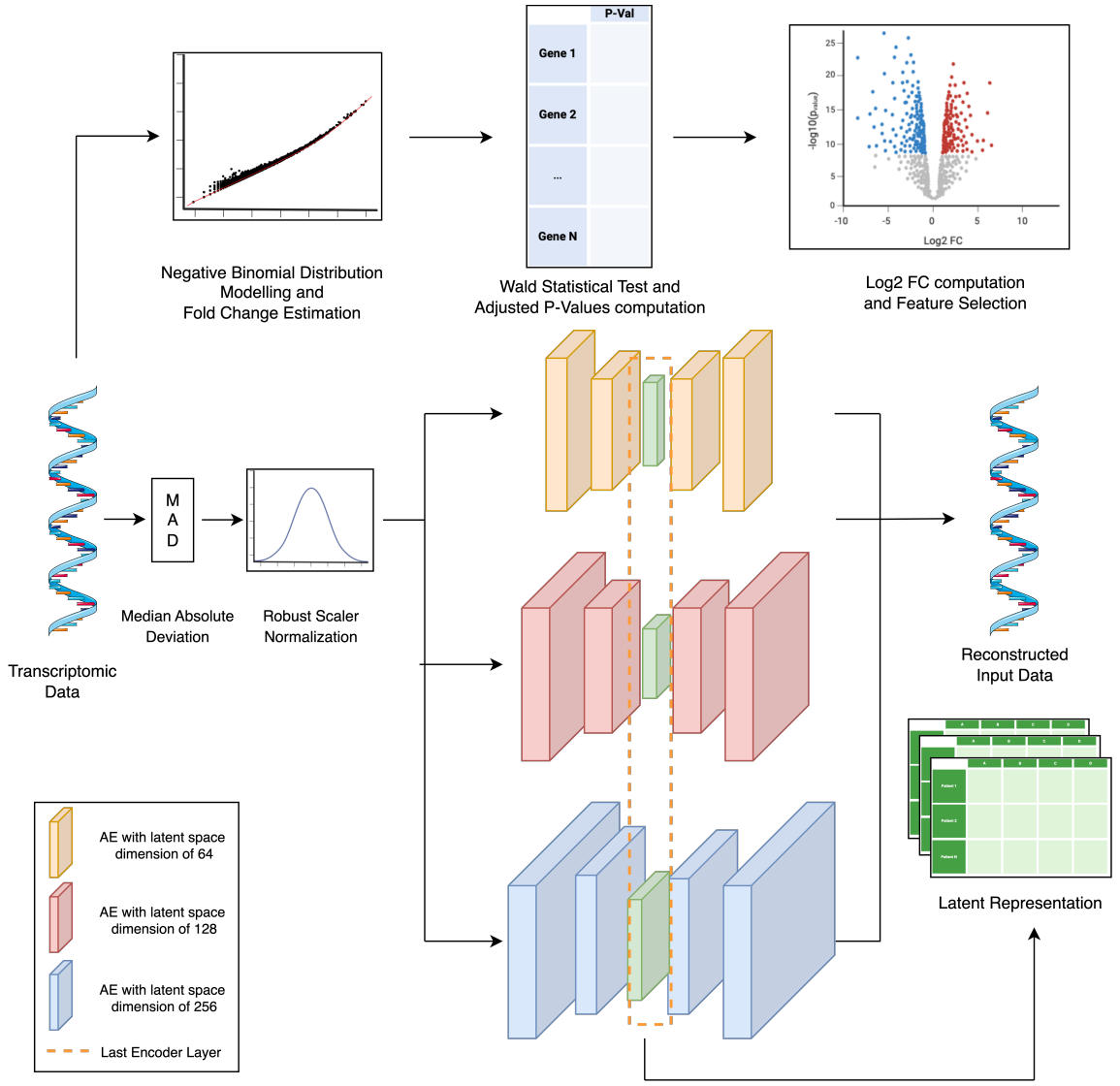


Fig. 3.7 Trascriptomic data pre-processing Pipeline. The original trascriptomic data are processed in parallel with a DEG analysis (top branch), selecting the genes up/down-regulated, and with a AE-based pipeline, retrieving the latent representation of the tree AE model, according to their size.

Finally, for CPTAC dataset, imaging samples were intersected with transcriptomic samples, allowing only subjects with labels for both data types to be retained, ensuring consistency across the data for the multi-modal validation. After the intersection operation, the number of subjected retained and used for constructing the test set was equal to 122.

3.2.5 Methods

Classification Models. The genetic mutations classification task with transcriptomic data was handled with three machine learning models: RF, XGB, and MLP classifiers. The WSIs classification task was approached with a MIL paradigm, using CLAM Deep Learning model. Notably, the performance of CLAM model relies of the quality of feature retrieved by the feature extractors. In its first version, CLAM used ResNet50 [59], pre-trained on ImageNet. However, recently, CLAM implemented the possibility of using of UNI and CONCH, two foundation models pre-trained on specific histopathology private datasets. In this study all three feature extractors were included.

Explainable Artificial Intelligence. The last step involved the use of eXplainable Artificial Intelligence (XAI) methods for retrieving an explanation at both imaging-level and transcriptomics.

Concerning the imaging-level XAI, a key strength of CLAM is the interpretability, due to its attention mechanism. In fact, a visual explanation was obtained by projecting the model attention scores on the input WSI, without any additional algorithm. In this way, it was possible to obtain attention-maps that can be used to investigate which image regions are crucial for model predictions. This transparency makes CLAM particularly valuable in clinical settings, where understanding the rationale behind predictions is essential [137, 138].

Shifting to transcriptomics XAI, the SHAP algorithm was used for retrieving global and local explanations of models behavior. Introduced by Lundberg et al. [139], SHAP is an explanation methods derived from the Shapley values of the cooperative game theory [78], that quantifies the contribution of the single player to the overall result generated by the entire set of players. Lundberg et al. adapts such a concept to ML, by considering each data feature as a cooperative player that contributes to the prediction of the target variable.

3.2.6 Results

3.2.6.1 Dimensionality Reduction of transcriptomic data

For transcriptomic data, the feature extraction according to DEG analysis and dimensionality reduction with AEs were evaluated for assessing the quality of the feature retrieved in both approaches.

DEG Analysis. A total of 117, 105, 41, and 134 DEGs were obtained for *KRAS*, *SMAD4*, *TP53*, and *CDKN2A*, respectively. As depicted in the Figure 3.8, according to *KRAS*, genes with lowest p-value and higher Log_2FC were *C6orf58* and *GAST* as up-regulated and *STYXL2* as down-regulated. According to *SMAD4*, top genes were *LDB3*, *FENDRR*, *COL9A1*, *MASP1*, *GIP*, and *ACTG2* as up-regulated and *STYXL2*, *ZFP57*, *UPK2*, and *SCGB2A2* as down-regulated. Again, according to *TP53*, top down-regulated genes were *GAST*, *C6orf58* and *CFAP47*, and *DEFA5* whereas the best down-regulated gene was *CSF3*. According to *CDKN2A*, top up-regulated genes were *HORMAD1* and *SLC52A1*, and down-regulated genes were *AMY2A*, again *STYXL2*, and *RPL3L*.

The differentially expressed genes retained, along with relative mean, $\log_2 FC$, Log FC Standard Error, and $p-values$ are reported in supplementary materials S1.

Autoencoders. The AEs performance was evaluated in terms of MSE value, since this was chosen as loss function for model training. Table 3.4 shows the MSE achieved on the training and validation set along with the one achieved on the test set (CPTAC dataset), according to the AE latent space dimension.

Table 3.4 MSE values achieved on validation set (TCGA data sub-set) and the test set (CPTAC dataset). In bold, the lowest values of MSE. Abbreviations: Mean Squared Error (MSE), AE-L (Autoencoder with latent space dimension of 256), AE-M (Autoencoder with latent space dimension of 128), AE-S (Autoencoder with latent space dimension of 64).

Model	Total Epochs	Training Loss	Validation Loss	Test Loss
AE-S	187/500	0.0019	0.0457	0.0314
AE-M	170/500	0.0006	0.0452	0.0305
AE-L	175/500	**0.0002**	**0.0441**	**0.0300**

3.2.6.2 Classification

As described in Section 3.2.5, a total of 72 models were trained, 24 for pathomics and 48 for transcriptomics. Performance assessment was made in terms of AUROC and AUPRC, due to the high unbalancing for *SMAD4* and *CDKN2A* labels. The best models were selected by choosing the best compromise among the two metrics, since AUROC was used as indicator of general performance while AUPRC was used to assess how well the model performed on the positive class predictions.

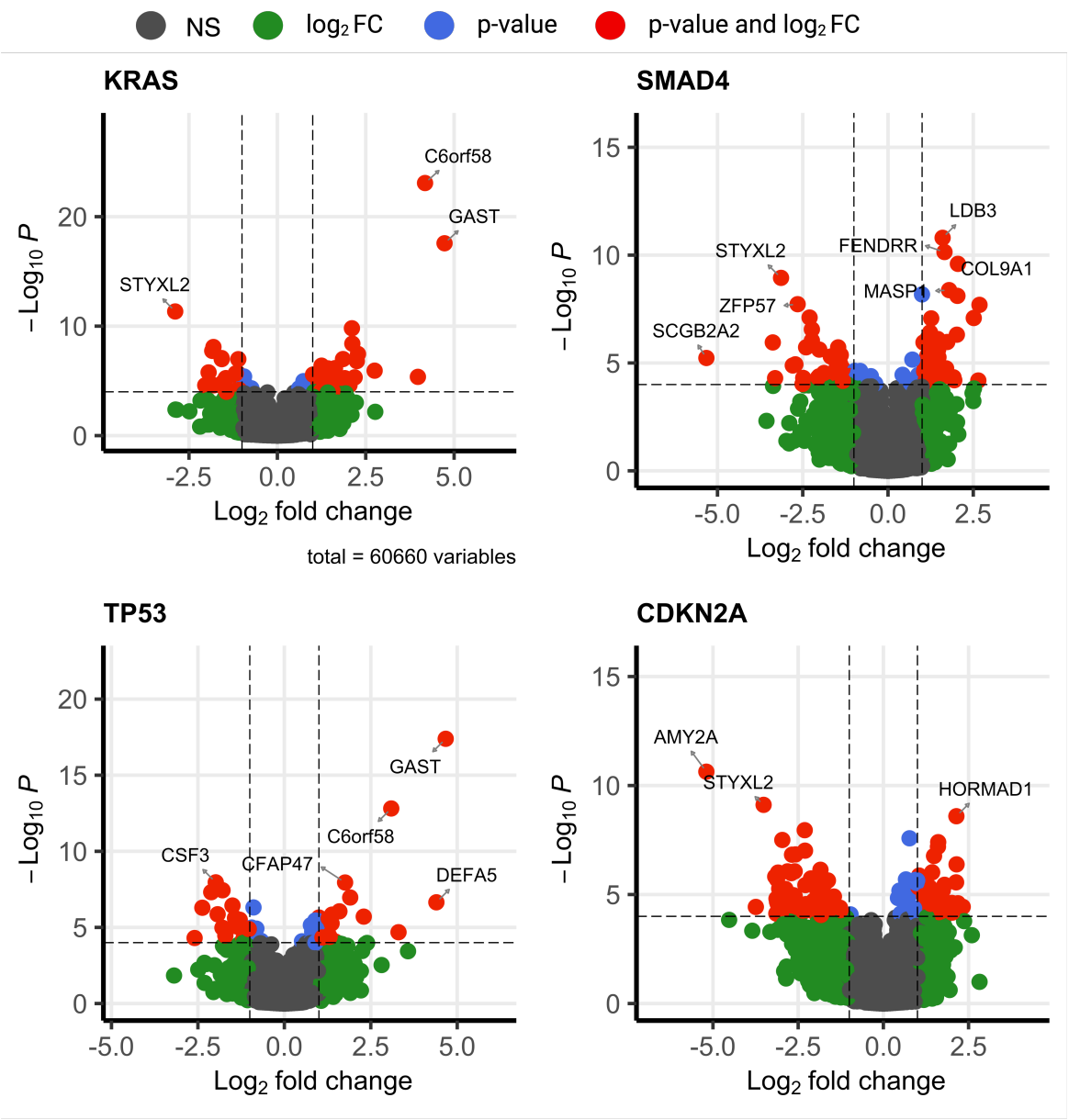


Fig. 3.8 Volcano Plot of Transcriptomic Data for Differential Gene Expression. The X-axis represents log2 fold change (log2FC) in gene expression, with positive values for up-regulated genes and negative values for down-regulated genes. The Y-axis shows -log10 p-value (logP), indicating statistical significance; genes further from the origin in either direction are significantly up- or down-regulated.

Unimodal Results. Figure 3.9 presents AUROCs and AUPRCs in the form of radar plots, emphasizing the pair feature extractor-CLAM model, for imaging data, and data pre-processing-classifier, for transcriptomics.

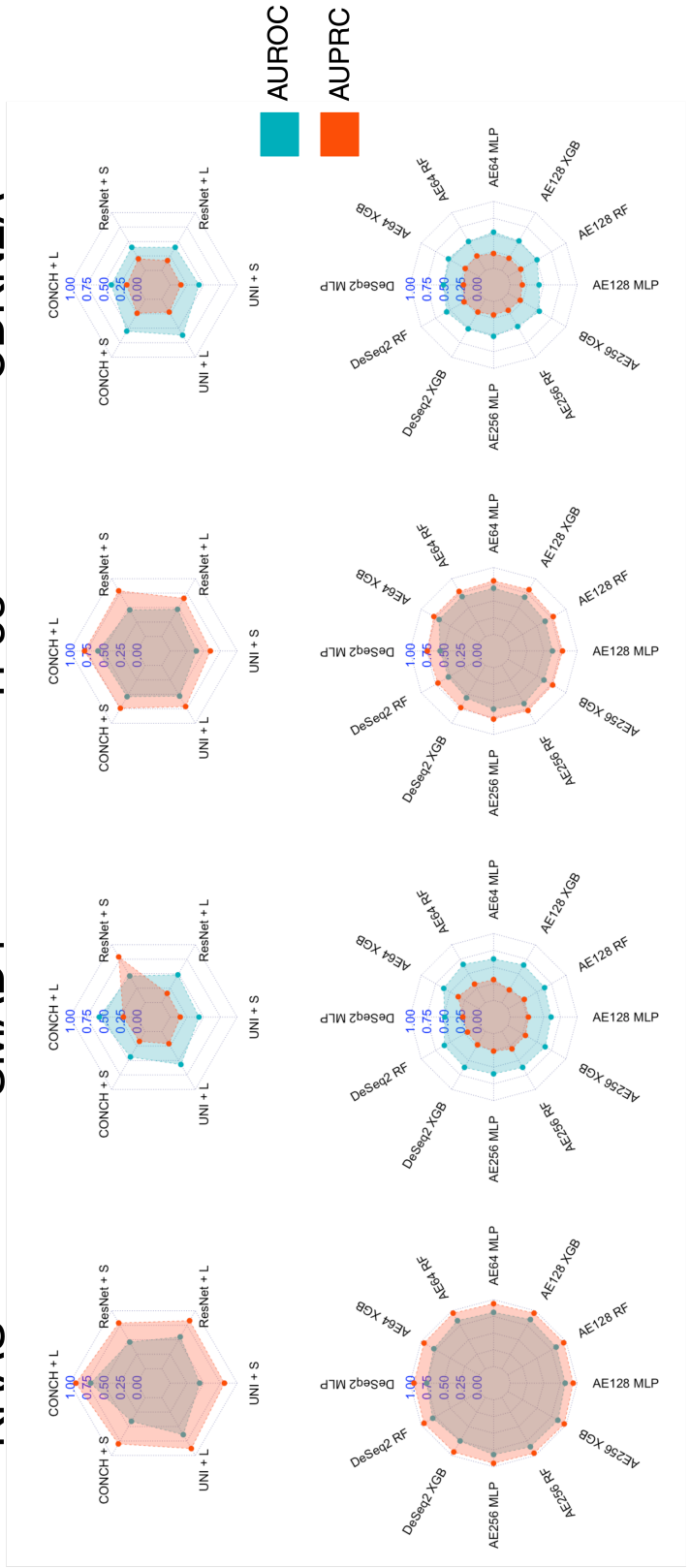


Fig. 3.9 AUROC and AUPRC metrics for imaging and transcriptomic models for *KRAS*, *SMAD4*, *TP53*, and *CDKN2A* genetic mutation predictions. S - CLAM small, L - CLAM large.

As highlighted by Figure 3.9, the models performance achieved on *KRAS* and *TP53* were generally higher than the ones on *SMAD4* and *CDKN2A*. This difference was justified by the high unbalancing for the latter mutations, that introduced underfitting during the model training. Imaging-based models that were trained on ResNet-based features were outperformed by imaging models trained with either UNI or CONCH-based features. The best imaging models are listed as follow:

- ***KRAS***: CONCH features and CLAM Large, with AUROC = 0.69 and AUPRC = 0.91.
- ***SMAD4***: UNI features and with CLAM Large, with AUROC = 0.57 and AUPRC = 0.21.
- ***TP53***: CONCH features and CLAM Large, with AUROC = 0.58 and AUPRC = 0.77.
- ***CDKN2A***: UNI features and CLAM Large and Small, with AUROC = 0.62 and AUPRC = 0.22.

Figure 3.9B depicted the perfomance achieved with transcriptomic models. At first glance, the overall performance achieved by such models outperformed the corresponding imaging models. The best transcriptomic models, with related performance, are listed as follow:

- ***KRAS***: XGB, with AUROC = 0.86 and AUPRC = 0.96 on data processed with AE-L.
- ***SMAD4***: RF, with AUROC = 0.66 and AUPRC = 0.32 on data processed with AE-S.
- ***TP53***: MLP, with AUROC = 0.69 and AUPRC = 0.79 on data processed with AE-S.
- ***CDKN2A***: RF, with AUROC = 0.56 and AUPRC = 0.26, on data processed with DeSeq2.

The metric values for trascriptomic models reflected the ones for the imaging models, with a *KRAS* and *TP53* emerging as the best classified mutations.

Multimodal Results. The output of imaging and transcriptomics models were then combined to obtain multimodal models. Figure 3.10 depicts a grid of radar plots where the

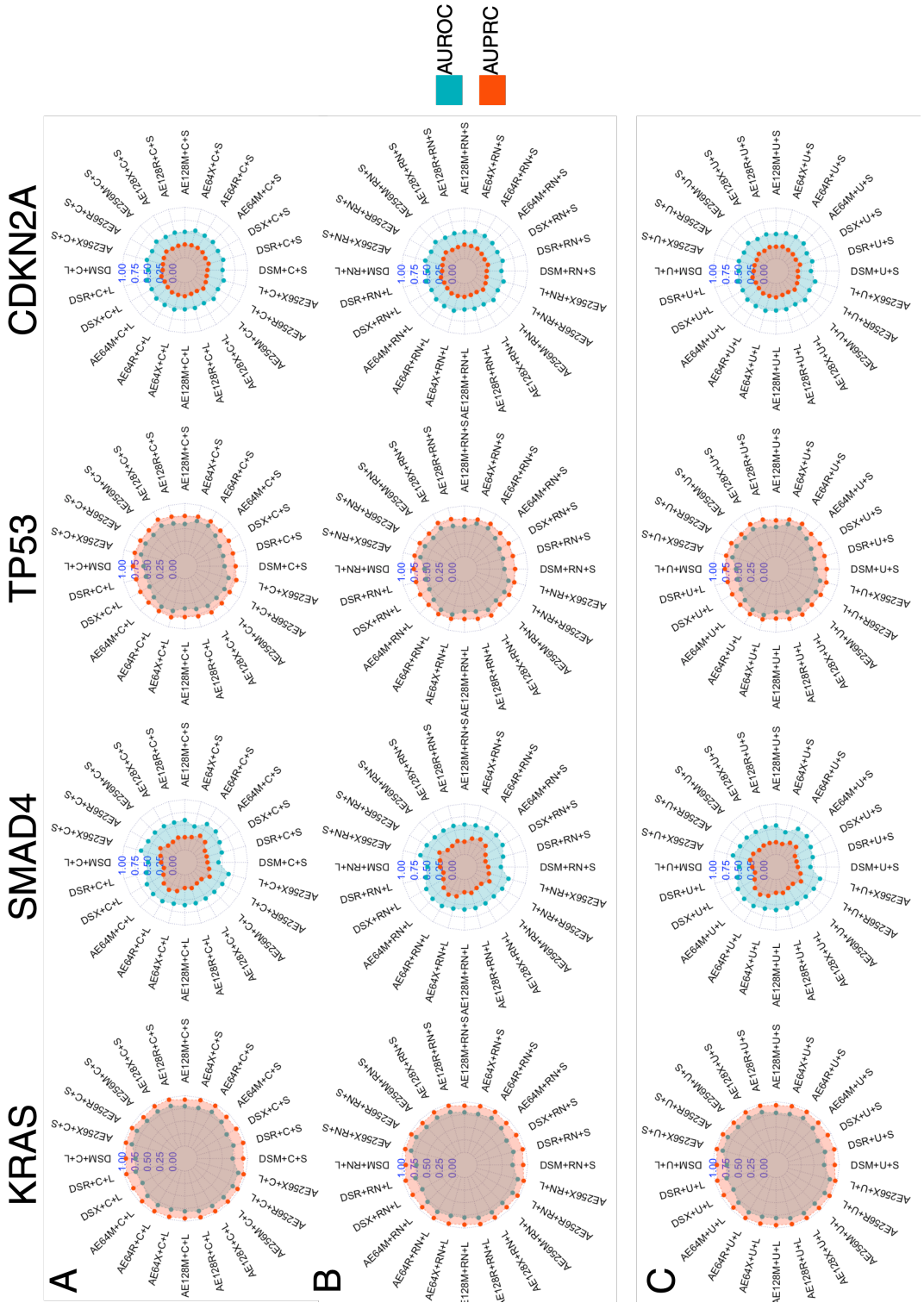


Fig. 3.10 AUROC and AUPRC metrics for multimodal predictions for *KRAS*, *SMAD4*, *TP53* and *CDKN2A* genetic mutation predictions. S - CLAM small, L - CLAM large, C - CONCH, U - UNI and RN - ResNet.

rows correspond to different imaging feature extractor and the columns to the target gene mutations.

The best multimodal models were as follows:

- ***KRAS***: (UNI features) + (CLAM Large) and RF + AE-L, with AUROC = 0.87 and AUPRC = 0.97.
- ***SMAD4***: (UNI features) + (CLAM Large) and XGB + AE-L, with AUROC = 0.65 and AUPRC = 0.25.
- ***TP53***: (UNI features) + (CLAM Large) and XGB + AE-L, with AUROC = 0.69 and AUPRC = 0.78.
- ***CDKN2A***: (UNI features) + (CLAM Large) and MLP + AE-L, with AUROC = 0.56 and AUPRC = 0.27.

The comparison with related works is reported in Table 3.5, showing that the adoption of foundation models with MIL approaches in pathomics outperforms the existing approaches for genetic mutation prediction in PDAC.

Table 3.5 Comparison of the proposed approach with related works for genetic mutations predictions in PDAC.

Work	Methods	Target Mutations							
		KRAS		SMAD4		TP53		CDKN2A	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Kather et al. [121]	Weakly supervised learning with CNN models.	0.67	0.7	0.45	0.17	0.51	0.58	0.24	0.12
Komura et al. [122]	Supervised Learning with DTRs and CNN models.	0.61	-	0.51	-	0.60	-	0.54	-
Saldanha et al. [123]	Self-supervised feature extraction and attention-based MIL.	0.58	-	0.47	-	0.44	-	0.61	-
Proposed Approach (Pathomics)	Foundation models for feature extraction and MIL.	0.69	0.91	0.57	0.21	0.58	0.77	0.62	0.22
Proposed Approach (Transcriptomics)	DeSeq2 and AEs with ML classifiers.	0.88	0.97	0.63	0.37	0.69	0.82	0.55	0.25
Proposed Approach (Multimodal)	Combination of Pathomics and Trascriptomics models.	0.86	0.97	0.65	0.25	0.69	0.78	0.62	0.27

3.2.7 XAI

Example of local XAI results for all considered genes are portrayed in Figure 3.11 and a global explanation was also provided in Figure 3.12. Global XAI suggested the top genes which contributed positively to the prediction (right side) and those which contributed negatively (left side). Among top genes, *FAM222A* and *SFTPA2* genes importantly discriminated the *KRAS* and *SMAD4* mutational status, respectively.

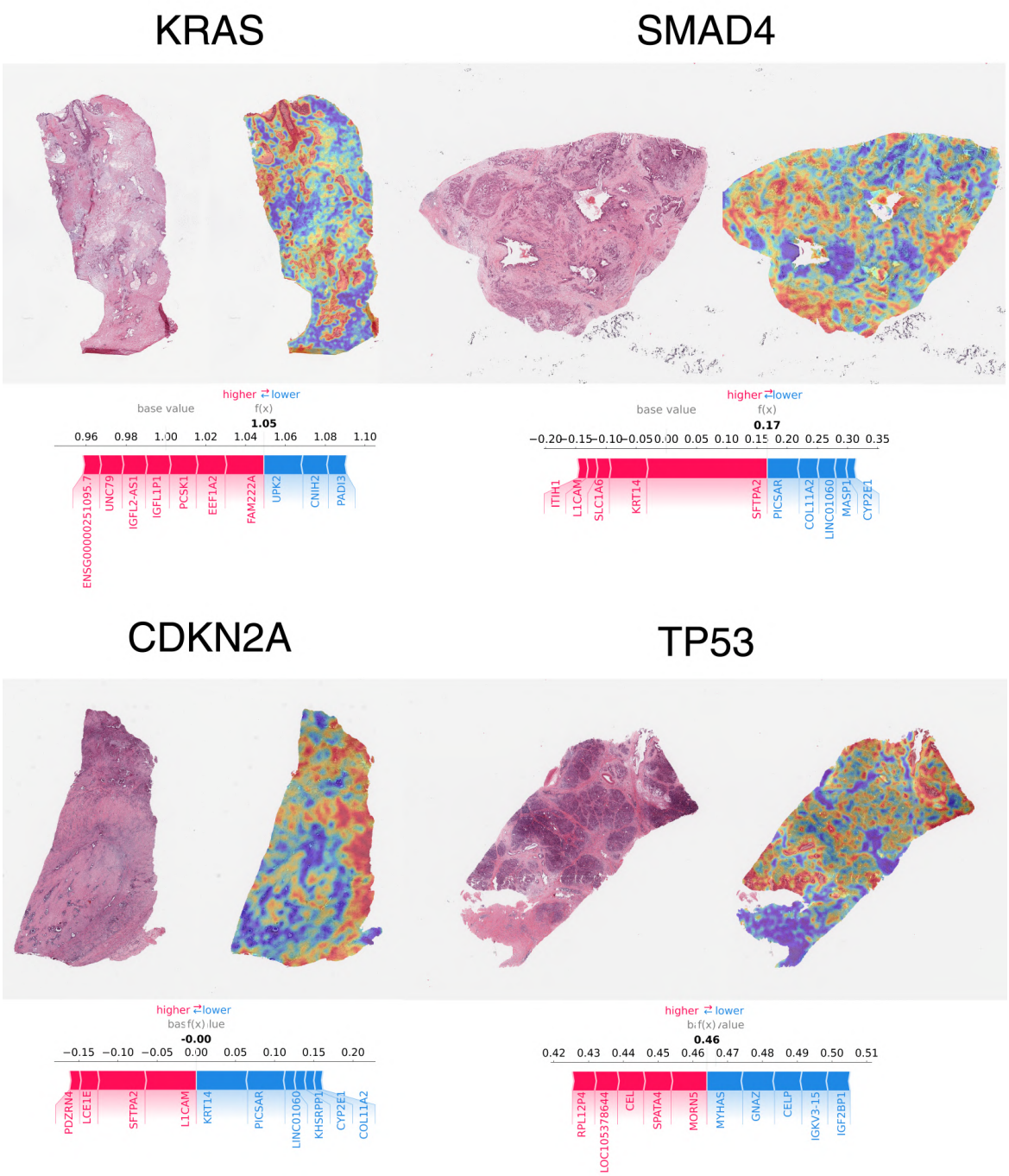


Fig. 3.11 Attention maps and SHAP decision plots for each considered target gene. For the KRAS, SMAD4, CDKN2A, TP53 mutations, the C3L-00277-23, C3N-02585-22, C3L-01598-22, and C3N-03190-22 cases from the CPTAC project are shown, respectively. In the WSIs, red regions indicated features that contributed positively to the prediction, while blue regions indicated negative contributions. The SHAP decision plots followed the same color scheme, with feature importance reflected by the width of each bar.

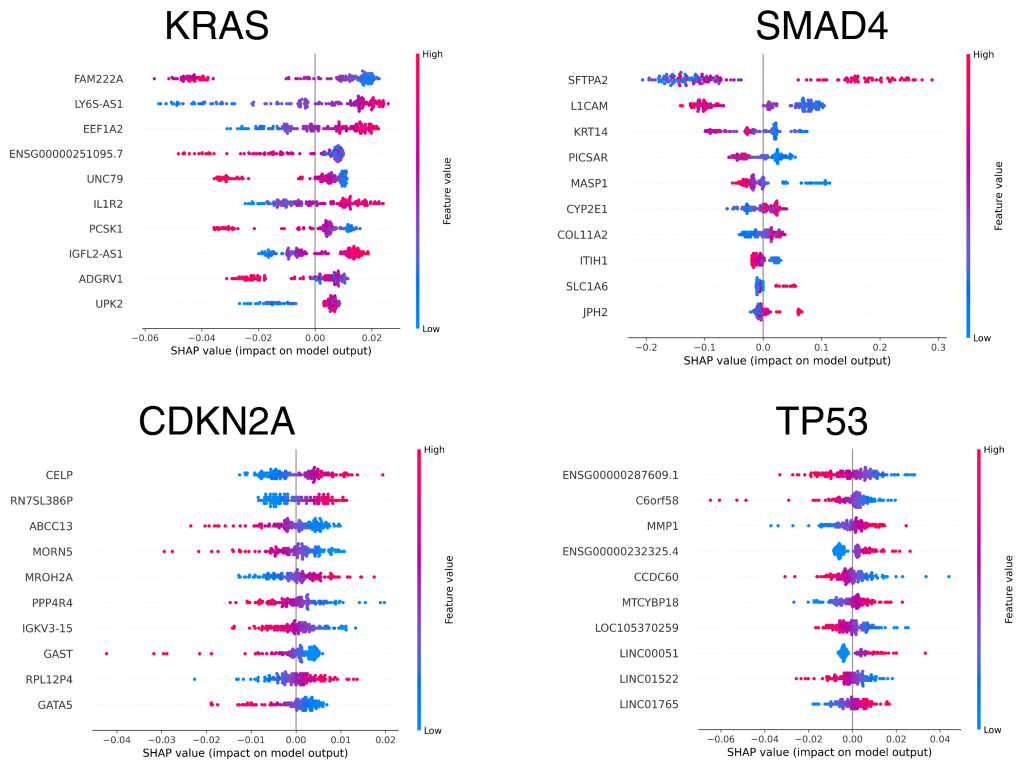


Fig. 3.12 SHAP beeswarm plots for each target gene. According to the legend, the color of each dot represents the feature value for a specific sample, while the dot's distance from the origin indicates its importance.

3.2.8 Discussion

Pre-Processing and Classification. Predicting genetic mutations from histopathology data is a challenging task due to heterogeneity of the tumor microenvironment, variability in tissue samples, and the complexity of linking histopathology features to the underlying genetic alterations. This work studied the mutational status in PDAC combining classical transcriptomics with pathomics on two independent series from TCGA-PAAD and CPTAC-PDA. Pathomic features were extracted by applying three different feature extractors, ResNet50, UNI, and CONCH. Classification models were designed with two CLAM architectures, small and large. Each model was then evaluated in terms or AUROC and AUPRC. AUPRC helped in a better understanding of case-studies with imbalanced classes. The results achieved were superimposable to those from the state-of-the-art, confirming that foundation models are able to extract higher quality determinants than those which use ResNet50 [64, 66]. As observed in Figure 3.9A,

the pathomic models achieved a good performance on *KRAS* and *TP53* achieving a best AUROC score of 0.70 and 0.58, respectively, and a best AUPRC value of 0.91 and 0.77, respectively. On the other hand, those models suffered in providing reliable predictions on *SMAD4* and *CDKN2A*. For the *SMAD4*, the CLAM Large model trained on UNI features, which resulted as the best model, achieved an AUROC of 0.57 and an AUPRC of 0.20. Again for the *SMAD4* target, the CLAM Small model on ResNet50 features achieved an AUPRC score of 0.79, which was very high. However, the AUROC score of 0.46 evidenced a potential bias in the assessment of the AUPRC score. For *CDKN2A*, again, the best model achieved an AUROC of 0.62 and an AUPRC of 0.22, making it worthless for predicting *CDKN2A* mutations. Overall for pathomics, the results obtained slightly outperformed those from the state-of-the-art, as reported in Table 3.5. Notably, while Kather et al. and Komura et al. achieved comparable metrics, they reported results on the TCGA-PAAD dataset, using the same cohort for both training and validation, which could potentially inflate performance. In contrast, the evaluation proposed relied on an external test set, offering a more rigorous measure of generalization. Interestingly, Komura et al. excluded those WSIs with poor staining quality and did not evaluate performance on datasets where individual subjects had multiple WSIs. In general, transcriptomic-based models outperformed those pathomic-based. This is likely due for the nature of data. Comparing AE-based models with DEGs, the models trained on the latent data representations achieved a better performance w.r.t. than those trained on DEG dataset; DEG analysis approach outperformed AE-based approach only for *CDKN2A* classification. For other mutations, AE-based models retained higher predictive power than those DEG-based. Thus, AE approaches were very effective in reducing data dimensionality, while retaining the information content.

As for imaging, the models were able to better fit data for both *KRAS* and *TP53* targets achieving good performances for AUROC score (0.86 for *KRAS* and 0.69 for *TP53*) and AUPRC score (0.96 for *KRAS* and 0.79 for *TP53*). Conversely, for *SMAD4*, although the XGB achieved a good AUROC, its AUPRC was too low for considering it as a reliable predictive model. This is confirmed in using RF to predict the *CDKN2A* status.

Looking at multimodal predictions obtained by leveraging both pathomic and transcriptomic models, the overall performance achieved is acceptable. Interestingly, the best metrics for target mutation were not obtained by the combination of the best unimodal models. In general, the performances achieved by combining the unimodal predictions were comparable to those from transcriptomic models. Hence, although

pathomic-based models are emerging tools to predict genetic mutations, the results carefully suggest that transcriptomics stills remain preferable in predicting mutational status for PDA. Despite that, the integration of pathomic models may adds value by incorporating insights from histopathology images, which can offer additional context or understanding of the tumor environment and genetic mutations. Thus, the combined approach provided richer information for interpreting the biological factors underlying the predictions.

Models Explainability. The attention maps offered by transformer models are an useful way of portraying saliency regions on the input images, displaying which image regions were most useful to make the classification. As is possible to see from Figure 3.11, where local XAI is executed on four samples for the four considered target genes, the imaging models tend to concentrate on abnormal glands and patterns in the WSI tissue, showing the capacity of the attention mechanism to discover likely tumor regions that could reflect the most patterns induced by the specific genetic mutations.

In the same figure, a decision plot obtained by SHAP for the transcriptomic model is also shown. As we can see, this technique has the potential to uncover which transcript expression is more important as link to the genetic mutation. Specifically, SHAP *TreeExplainer* and *GradientExplainer* algorithm were applied to Tree-based and ANN-based models, respectively. For the cascade combination of AE and Tree-based model the explanation was retrieved by using *KernelSHAP*.

Noteworthy, although the combination of AE and ML models performed better, the use on SHAP on it did not lead to a meaningful result, since the algorithm computed the feature importance for the prediction as the sum of infinitesimal contributions of a large feature set. In light of this, from a XAI perspective, the use of DEG-based data remains a preferable approach. However, the questions rising about the trade-off between model's performance and interpretability are still debated in literature [140–142].

In Figure 3.12, beeswarm SHAP plots are shown, where the contributions for each sample in the datasets are considered together to devise a feature importance at dataset level and not just at instance level. This technique can show which transcript expressions are more linked to genetic mutations not just at patient-level, but on the whole cohort, allowing to uncover eventually more complex relationships between genome and transcriptome.

3.3 Summary of Findings

Multimodal analysis from the multi-omic cohort of CPTAC-PDA.

Radiomic, clinical, and mutational features that correlate with OS and REC are identified in this study. The findings indicate that radiomics, when effectively combined with established clinical and biological determinants, has the potential to enhance patient risk stratification for PDA. The results show that the Cox model outperforms SML models for OS prediction, while the SVM model proves most effective for REC prediction. Additionally, this work represents the first application of time-dependent explanations of radiomic features within a PDA cohort. Future efforts focus on validating the proposed signature using "real-life" data. The CPTAC project provides additional mutational data, including copy number variations and methylation data, as well as histopathology images of biopsied tumors collected at diagnosis. To fully utilize these public datasets, future research is directed towards radiogenomics and pathomics analyses. Specifically, the development of an integrated radiopathomic prognostic model is prioritized to improve the accuracy of patient prognosis predictions. Furthermore, the identified signatures have potential applications in accelerating clinical workflows, such as predicting genetic alterations directly from imaging data.

Pathomics and Transcriptomics for Genetic Mutation Prediction in PDAC.

Predicting genetic mutations from histopathology data presents significant challenges due to the heterogeneity of the tumor microenvironment, variability in tissue samples, and the complexity of associating histopathological features with underlying genetic alterations. Although various approaches have been proposed in recent years to link genetic mutations to histopathology data, few have focused on mutation prediction in PDAC. To date, none of these methods have proven sufficiently reliable for routine use in this context. This study demonstrates the potential of foundation models for digital pathology to enhance performance in predicting genetic mutations. Notably, the pathomic models achieved promising results in predicting KRAS mutations, highlighting their applicability in this domain. Consistent with prior research, transcriptomic data emerged as the strongest predictor for genetic mutations. Multimodal models combining transcriptomic and pathomic data were also tested; however, they did not show significant advantages over purely transcriptomic models. Nevertheless, the integration of pathomic models contributes additional insights by leveraging histopathology images, offering valuable context for understanding the tumor environment and genetic alterations. This combined approach enriches the interpretation of the biological factors

underlying the predictions. Furthermore, the explainability provided by the attention mechanisms in CLAM models and the post-hoc interpretations of transcriptomic models using SHAP offers crucial insights into the relationships between transcriptomic, genomic, and histopathological data. These explainability tools hold promise for unveiling the complex interplay of these biological layers, advancing understanding and interpretation in this field.

Chapter 4

Unimodal Big Data Analytics Pipelines

This chapter examines the four unimodal analytics pipelines developed with some of the methods included in the two multimodal pipelines. Such methods are applied to other case studies, showing their flexibility.

The first pipeline exploits a time-dependent XAI method for enhancing the model selection process in the survival analysis [40]. In particular, an end-to-end pipeline is developed for estimating the OS in patients affected by Obstructive Sleep Apnea (OSA) using several ML and DL survival models. C-Index, C/D AUC and Brier Score are considered as evaluation metrics for survival models. Finally, survSHAP algorithm is applied to the best performing models showing how explainability can support the model selection process for models with similar performances.

The second section presents another contribution of this thesis work to digital pathology. In particular a pipeline is tailored to the segmentation and classification of glomerular lesions according to the Oxford classification for IgA nephropathy (IgAN) cases [41]. The pipeline consists of two main components: (i) a segmentationblock, for dividing WSIs into tiles, followed by glomeruli segmentation using object detection models. (ii) a classification Block, with CNNs employed for the classification of the segmented glomeruli. The classification outcomes are reported at both the glomerular and biopsy levels. To evaluate the pipeline's performance, intraclass correlation coefficients and Cohen's Kappa statistics are calculated to measure agreement between the model's predictions and expert pathologist labels at the glomerular and biopsy levels.

The third pipeline proposes a mathematically and visually interpretable deep learning-based framework for multiclass, shape-based classification of tomosynthesis breast lesion images Ėight pretrained CNN architectures are utilized for the classification task on previously extracted regions of interest containing lesions. The black-box nature of the deep

learning models is further explored using two well-known explainable AI (XAI) techniques: Grad-CAM and LIME. Additionally, two mathematical-structure-based interpretability methods, t-SNE and UMAP, are applied to analyze the behavior of the pretrained models in multiclass feature clustering.

The last pipeline deals with NER approach for medical oncological free-text report. This is another important aspect related to clinical examinations. While this thesis has focused on algorithms that support the diagnostic and decision-making phases, it is also important to support the data collection phase from unstructured clinical reports. This can be achieved through NLP tasks such as NER. In particular, a framework called Automatic record generator for Onco-Hematology (ARGO) was enhanced for extracting key-fields from oncological free-text reports and for standardizing the diagnosis associated, according to definitions coming from the National Institute of Health in accordance with the International Classification of Diseases, 10th (ICD-10) and oncology (ICD-O) versions [44]. The enhancement involves the inclusion of Machine Learning model, supporting the existing architecture and a decisional heuristic for classifying the extracted diagnosis. Finally, a preliminary results with transformer-based architecture, replacing the existing one, are shown for diagnosis NER task.

4.1 Enhancing Survival Analysis Model Selection Through XAI(t) in Healthcare

Survival analysis is particularly valuable in healthcare for assessing patient outcomes, identifying risk factors, and tailoring treatments to individual patients. By incorporating variables like comorbidities, age, and treatment responses, survival analysis helps clinicians determine which patients are at higher risk and adjust care plans accordingly.

When applied to OSA, SA can play a crucial role in better assessing and categorizing disease severity. OSA, characterized by recurrent airway collapse during sleep, often leads to oxygen deprivation and interrupted sleep cycles [143–145]. Moreover, the presence of comorbidities, particularly metabolic, cardiovascular, and renal diseases, complicates the prognosis and treatment of OSA [143, 146].

Moreover, SA can help identify high-risk individuals by incorporating both the comorbid conditions and the symptoms typical of OSA. This approach allows for a more personalized treatment plan, especially during rehabilitation and follow-up phases, ultimately improving the survival probability of patients by tailoring care to their specific needs.

Notwithstanding, AI algorithms addressing survival analysis are known to be characterized by the poor interpretability of their results, since determining which feature impacts on the model prediction is not a simple task. Moreover, the role of comorbidities and risk factors is often evaluated only according to Hazard Ratios and Odds Ratios [147]. This limits their applicability in decision support systems for clinical purposes [79].

A gap has been found in the existing literature about XAI methods for SA tasks. No prior works applies time-dependent XAI - XAI(t) methods to survival DL models, while offering a model comparison in terms of data and model explanation. Most of existing XAI-oriented researches focus on standard XAI methodology applied to classification tasks, possibly facing the survival analysis as a separated task. Such a gap can be attributed to the novelty of XAI(t) approaches as well as to challenges in producing a simple and reliable explanation when the time is involved.

Focusing on the OSA clinical scenario, the work named "Enhancing Survival Analysis Model Selection through XAI(t) in Healthcare" [40] a survival analysis pipeline aiming at:

- training and validating different Machine Learning and Deep Learning survival models, selecting the best performing ones according to the metrics used in survival tasks;
- investigating the role of comorbidities in OSA from XAI(t) perspective;
- performing a model comparison, selecting the most reliable models according to the explanations retrieved by XAI(t) algorithms.

4.1.1 Related Works

In Artificial intelligence field results are difficult to be interpreted, especially when dealing with deep models that are “black-box” where it is difficult to understand how the model got to the prediction. XAI has been recently extended to this context to improve explainability, interpretability and transparency for modeling results [148–150]. XAI algorithms’ goal is then to convert unexplained ML and DL predictions into more interpretable “white-box” glass ones.

Notably, in the realm of survival analysis, to be able to understand which characteristics have a more important influence on the prediction, that means to understand how it "weighs" such features for the prognosis or diagnosis, is fundamental because it allows clinicians to choose the type of treatment on the patient (preventive or curative).

Most of existing works applies XAI techniques like SHAP and LIME on AI-based frameworks addressing SA. Qi et al. [151] used SHAP to improve the explainability of a

ML-based framework about the role of mithochondrial regulatory genes on the evolution of renal clear cell carcinoma . Zaccaria et al. [152] adopted approximated SHAP values to build an interpretable transcriptomics-based prognostic system for Diffuse Large B-Cell Lymphoma (DLBCL). Srinidhi and Bhargavi [153] embedded SHAP and LIME in a ML- and DL-based framework to support the prediction of survival rates of patients with pancreatic cancer. Zuo et al. [154] employed both SHAP and LIME to explain the survival predictions of many ML methods fed with a radiomics feature set that is extracted from tomographic images. Chadaga et al. [155] used SHAP and LIME for explaining the ML-based estimations about the survival probability of children after bone marrow transplantation. Both SHAP and LIME were included in Alabi's study [156] to better interpret the predictions of ML algorithms addressing SA on clinical data from people with nasopharyngeal carcinoma. Peng and colleagues [157] used SHAP and LIME to corroborate the outcomes of ML models about hepatitis diagnosis and prognosis.

Even fewer works employed either SurvSHAP or SurvLIME for performing SA with ML-based workflows. Zhu and colleagues [158] employed SurvSHAP to investigate the efficacy of adjuvant chemotherapy starting from the prediction of both ML and DL models about the survival probability of breast cancer patients. Passera et al. [159] explained the outcomes of survival models fed by demographic and clinical features by means of both SurvSHAP(t) and SurvLIME. Such XAI methods were oriented to global and local explaination by analyzing data from the whole cohort or a single patient. Baniecki et al. [160] performed SA to estimate the hospitalization time by training ML algorithms with a multimodal dataset, and also explained the results of time-to-event models for a single patient through SurvSHAP as well. Remarkably, as far as we know, no studies that utilizes XAI(t) to elucidate the predictions generated by a deep learning model. In fact, such architectures often play a supporting role in SA workflows: they are used not to predict mortality risks, but to determine additional metrics that are then fed in a classical survival pipeline for performing SA. [144, 161, 162]. In addition, these works do not include XAI strategies to better interpret how the model achieves survival predictions.

Related work shows a paucity in the literature about survival analysis with either ML algorithms or DL models embedding XAI techniques - e.g., SurvSHAP and SurvLIME - to increase the interpretability of the predictions of mortality risk. Table 4.1 reports a summary of related work involving classical XAI and XAI(t) in SA.

Table 4.1 Related Works of XAI and time-dependent XAI.

Author	Year	Task	Input data	Survival analysis model	Explainability model
Zaccaria et al. [152]	2023	Prognosis of DLBCL	Transcriptomic data	AutoEncoders	DeepSHAP
Alabi et al. [156]	2023	Prognosis of NPC	CT images, clinical data	Linear Regression, KNN, Support Vector Machines, Naive Bayes, Tree-based models,	SHAP, LIME
Srinidhi et al. [153]	2023	Prognosis of pancreatic cancer	CT images, clinical data	Convolutional Neural Networks, Support Vector Machines	SHAP, LIME
Chadaga et al. [155]	2023	Prediction of BMT efficacy	Clinical data	Tree-base models, Linear Regression, KNN, AdaBoost, CartBoost	SHAP, LIME
Peng et al. [157]	2021	Prognosis of hepatitis	Clinical and demographic data	Linear Regression, CART, KNN, Tree-based models, Naive Bayes	SHAP, LIME
Qi et al. [151]	2023	Prognosis of RCC	Genomic data	LASSO-Cox	SHAP, LIME
Zuo et al. [154]	2023	Identification of EGFR in lung adenocarcinoma	CT images	Light GBM, Linear Regression, Tree-based models	SHAP, LIME
Zhu et al. [158]	2024	Prognosis of breast cancer	Clinical and demographic data	Cox Mixtures, DeepSurv, Cox PH, Survival Random Forest	SurvSHAP
Baniecki et al. [160]	2023	Prediction of hospital LoS	Text data, tabular data, X-ray images	Tree-based models, CoxPH, DeepSurv, DeepHit	SurvSHAP, SurvLIME
Passera et al. [159]	2023	Test XAI on SA for BMT	Clinical and demographic data	CoxPH, Survival Random Forest	SurvSHAP, SurvLIME

4.1.2 Material and Methods

Initially, pre-processing operations were conducted to clean and prepare the data. Statistical tests and correlation analyses were then performed to aid in feature selection. Subsequently, the data was divided into training and test sets, and survival analysis models were trained, with the best-performing ones selected based on evaluation metrics from the test set. Finally, the chosen models were interpreted and compared using SurvSHAP. Figure 4.1 provides an overview of the pipeline used for the analysis.

Dataset

The dataset utilized in this study was collected by the Istituti Clinici Scientifici (ICS) Maugeri Hospital sleep laboratory in Bari, during the rehabilitation phase of patients diagnosed with Obstructive Sleep Apnea (OSA). The diagnosis was established for all patients through an in-laboratory overnight polysomnography (PSG). The original dataset contained 1,592 samples and 45 features, encompassing clinical data and information retrieved from PSG exams. In addition to demographic variables like age, follow-up duration, and gender, the dataset included medical parameters such as Body Mass Index (BMI), Glomerular Filtrate Rate (GFR), Ejection Fraction (EF), Oxygen Desaturation Index (ODI), minimum blood oxygen saturation (SaO_2), Apnea Hypopnea Index (AHI), and details on comorbidities such as heart disease and diabetes.

Data Pre-Processing

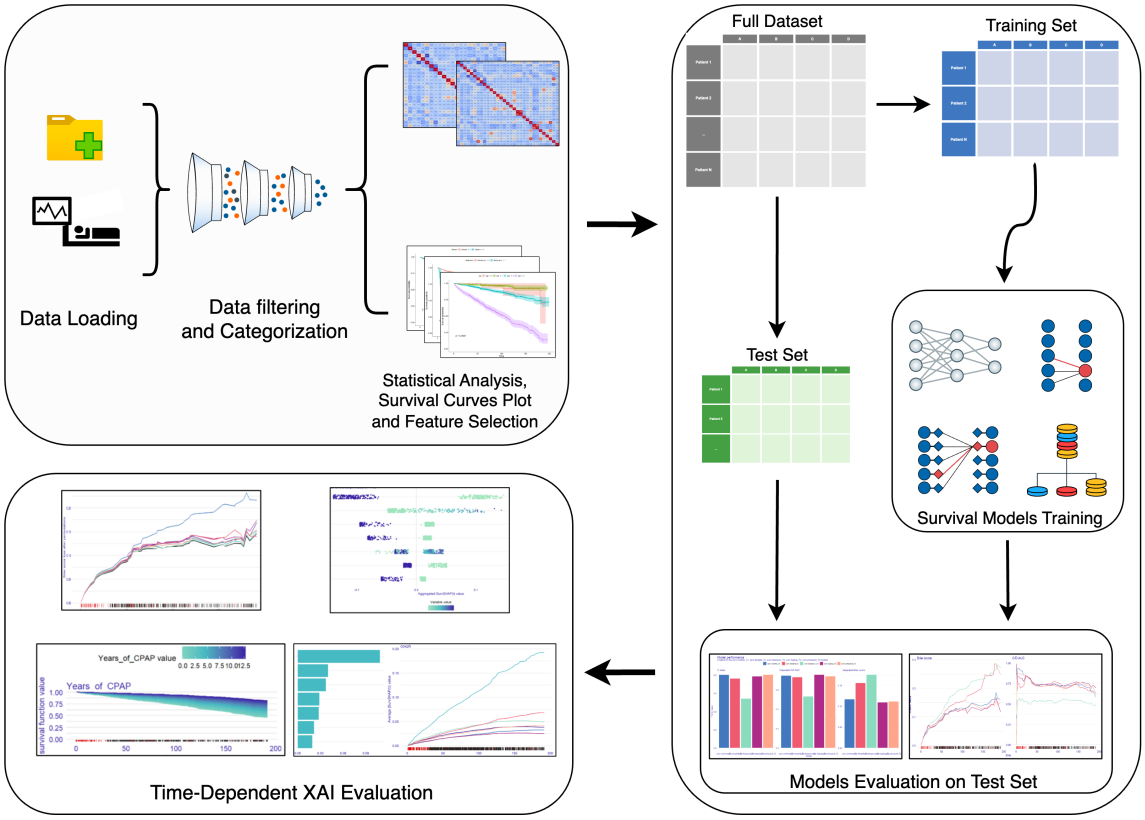


Fig. 4.1 Processing Pipeline followed.

Initially, irrelevant features, including *patient I.D.*, *N CC* (number of health records), *Admission Date*, *Discharge Date*, and *profession*, were removed from the dataset. The survival time was then converted from days to months to simplify the analysis and interpretation. Features with high levels of missing data, such as *EF* and *ODI* (which had 90% of their values missing), were dropped since no reliable imputation strategy was available. The *Anemia* feature was retained, and missing values were imputed based on hemoglobin levels, considering the patient's gender.

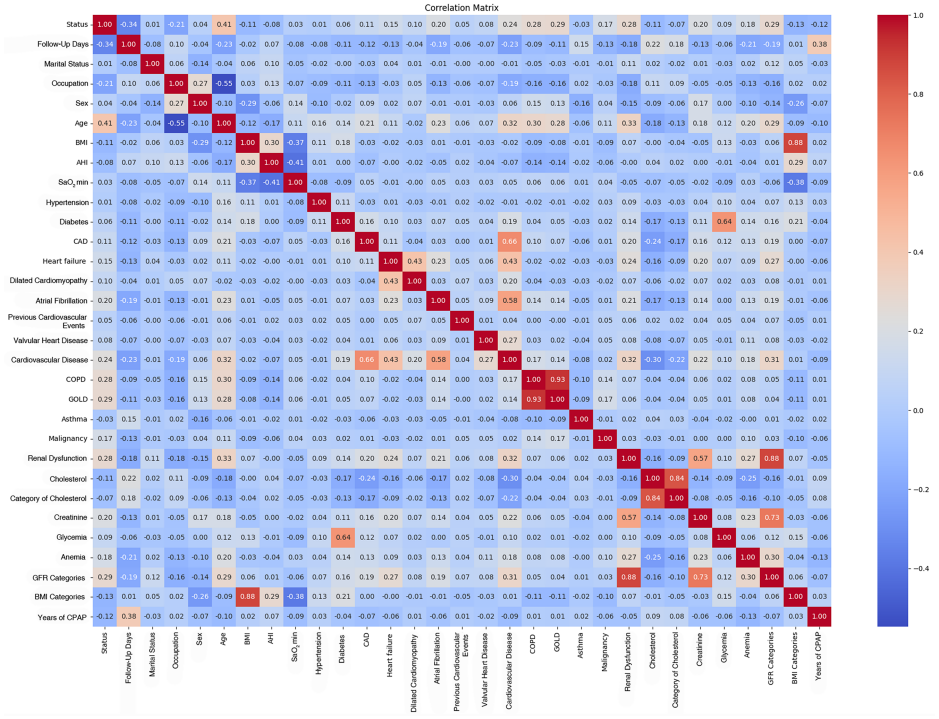
Next, several variables were discretized: *Age* was categorized as 0 for subjects aged 65 or younger, and 1 for subjects older than 65; *BMI* and *GFR* were binned according to reference values from medical literature [163, 164]. The age cut-off was chosen based on medical relevance [165, 166], ensuring better generalization of the results. Additionally, the dataset included two new features: *Continuous Positive Airway Pressure (CPAP)* treatment (a binary categorical feature indicating whether the patient had undergone CPAP treatment) and the corresponding duration of treatment in years.

Outlier removal was not performed, as in the medical domain, outliers may represent important clinical conditions. Finally, to avoid bias caused by the COVID-19 pandemic, samples with a follow-up after 2020 (198 samples) were excluded. After this pre-processing, 1,394 samples were retained for analysis.

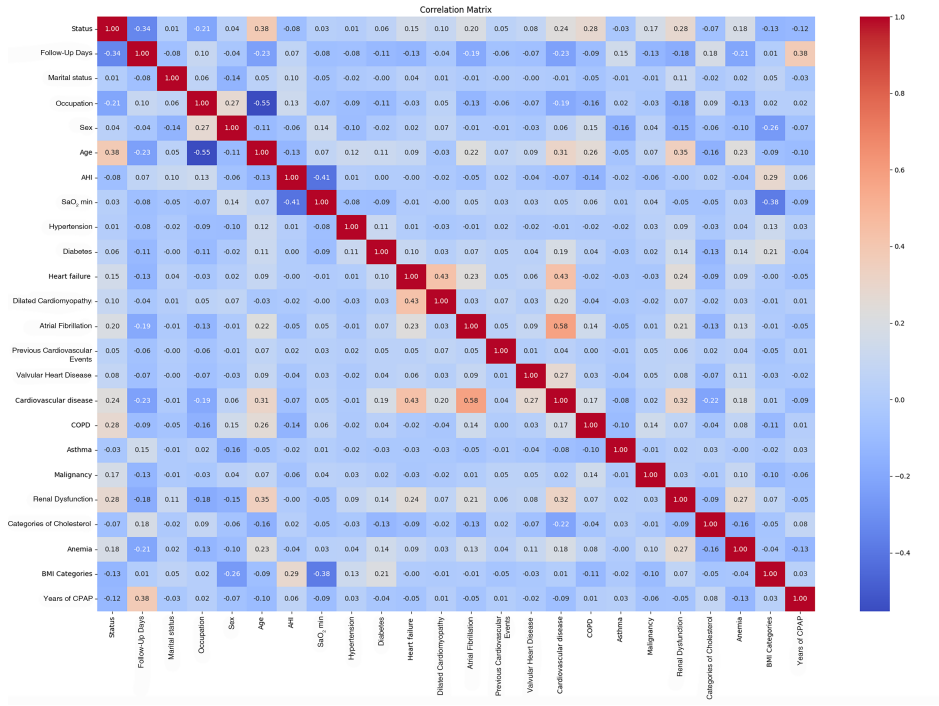
Statistical Analysis and Feature Selection

Before training the survival analysis models, feature selection was conducted based on the training set. Pearson's correlation coefficient (PC) was calculated for each feature to identify and remove those with high correlations (absolute $|PC|$ greater than 0.6). The correlation matrices before and after filtering are shown in Figure 4.2a and Figure 4.2b. In cases of high correlation between numerical and categorical features, the numerical variables were retained.

Following the pre-processing and feature selection, 23 features from 1,394 records were used for survival analysis. The *Status* variable served as the event indicator, and the *Follow-up Days* (converted to months) represented the survival time.



(a) Correlation Matrix before filtering.



(b) Correlation Matrix after filtering.

Fig. 4.2 Correlation Matrix after filtering.

Table 4.2 Final dataset with related statistics.

Feature	Type	Description	Number	Mean±Std	P-Value
Number of patients = 1394					
Demographics					
Status	Categorical	Indicates if the patient is dead or alive at follow-up			Reference
Dead			363	-	
Alive			1031	-	
Sex	Categorical	Sex of the patient			0.182
Male			997	-	
Female			397	-	
Age	Categorical	Indicates whether or not the patient is over 65			<0.001
Under 65 years			700	-	
Over 65 years			694	-	
Marital Status	Categorical	Marital Status of the patient			0.842
Married			262	-	
Not Married			1132	-	
Comorbidities					
Hypertension	Categorical	Presence of hypertension	752	-	0.652
Diabetes	Categorical	Presence of diabetes	413	-	0.033
Heart Failure	Categorical	Indicates whether patients have a history of heart failure	79	-	<0.001
Dilated cardiomyopathy	Categorical	Presence of dilated cardiomyopathy	17	-	<0.001
Atrial Fibrillation	Categorical	Indicates whether patients have a history of atrial fibrillation	135	-	<0.001
Previous Cardiovascular Events	Categorical	Indicates whether patients have a history of Previous CV Events	32	-	0.089
Valvular Heart Disease	Categorical	Presence of Valvular Heart Disease	33	-	0.006
Cardiovascular Disease	Categorical	Presence of Cardiovascular Disease	339	-	<0.001
Chronic obstructive pulmonary disease (COPD)	Categorical	Presence of COPD	288	-	<0.001
Asthma	Categorical	Presence of Asthma	70	-	0.297
Malignancy	Categorical	Indicates whether patients have a history or Presence of malignancy	26	-	<0.001
Renal Dysfunction	Categorical	Presence of Renal Dysfunction in Patient	308	-	<0.001
Anemia	Categorical	Presence of Anemia	263	-	<0.001
Cholesterol Category	Categorical	Categorical column binning the Cholesterol in 3 categories			0.030
Value $\leq$ 200 [mg/dL]			870	-	
Value in 200-239 [mg/dL]			371	-	
Value $\geq$ 240 [mg/dL]			153	-	
Weight Categories	Categorical	Categories of weight based on BMI value			<0.001
Normal Weight			75	-	
Overweight			291	-	
Obesity Class I			397	-	
Obesity Class II			327	-	
Morbid Obesity			304	-	
Polysomnographic data					
AHI	Numeric	Apnea-Hypopnea Index	-	57.01±19.16	0.002
SaO_2 min [%]	Numeric	Minimum oxygen saturation	-	70.94±13.63	0.335
Treatment Info					
CPAP	Categorical	Received CPAP treatment	555	-	0.006
Years of CPAP	Numeric	Duration of CPAP usage (years)	-	4.44±3.25	<0.001
Follow-Up Days	Numeric	Days from admission to follow-up (months)	-	98.62±49.76	<0.086

Survival models

For performing the experiments the following models were included:

- *Machine Learning* - Cox Proportional Hazard (CPH), Survival Random Forest (SRF), Survival Gradient Boosting Model, Survival Support Vector Machine

- *Deep Learning* - Cox-Time (CT), Deep-Hit (DH), DeepSurv (DS) , NNet-Survival (Logistic Hazard (LH) and Piecewise Constant Hazard (PCH)).

Explanation Methods

The explainability phase was conducted exploiting survSHAP. As mentioned in Chapter 2 (Section 2.3.1), this method generalizes SHAP to survival models, giving a global explanation about the overall behavior of the model over the time. SurvSHAP can reveal the importance of comorbidities affects the prognosis over follow-up, offering valuable insights into the progression of OSA and what to do to improve the patient's prognosis.

The model explanation methods can be classified in two categories: Dataset explanation and Model explanation. The former category aims at analyzing the dataset characteristics in order to understand their impact on the event prediction; the latter is focused on ranking and highlights the features that the model considers important for the prediction.

Intuitively both category methods present some limitations:

- **Dataset-level explanation limitations**: SHAP is designed to return a local-explanation, i.e. gives an explanation for a single sample; consequently, SurvSHAP behaves the same way. When used on a dataset, its resulting explanation depends on the samples distribution. In fact, if the data are unbalanced for specific features, their contribution will be minimal, but this conclusion cannot be generalized to other data. Hence, the ideal scenario could be to use a large dataset for the sake of a higher generalization of the results. However, the computation of features contribution for a single sample is computational expensive, because of the model complexity and the operations involved (e.g. multiple features permutations, predictions and performance computations for SHAP values, local-samples generation, local-model training and prediction for LIME). In XAI(t), this is worsened since the feature contributions are computed also for different time instants.
- **Model-level explanation limitations**: Although the model-level explanation is computationally less expensive than the data-level one, it returns explanation information at populational level and cannot be used for explanation at single prediction level. Moreover, the explanation methods based on permutation can lead to misinterpretations when the independent variables are strongly correlated [167].

To overcome the limitations of both explanation methods, they can be exploited in a complementary way: while the model-level explanation can be used to identify the *most important* features affecting the prediction, the dataset-level explanation can be used to investigate *how* they affect the prediction.

Obviously, both methods strongly depend on the model performance: if a model is not reliable, then the model explanation will not be able to identify some features that can be important for the event prediction, while the data explanation can lead to misinterpretations of features behavior.

4.1.3 Experimental Pipeline

The experiments were conducted using both *R* and *Python*. The dataset was randomly split using the holdout method, with 70% of the data allocated for the training set and the remaining 30% for the test set, while ensuring stratification by the *event* feature. Additionally, it was verified that the observation period in the test set did not exceed that in the training set.

The experimental pipeline workflow is illustrated in Figure 4.3.

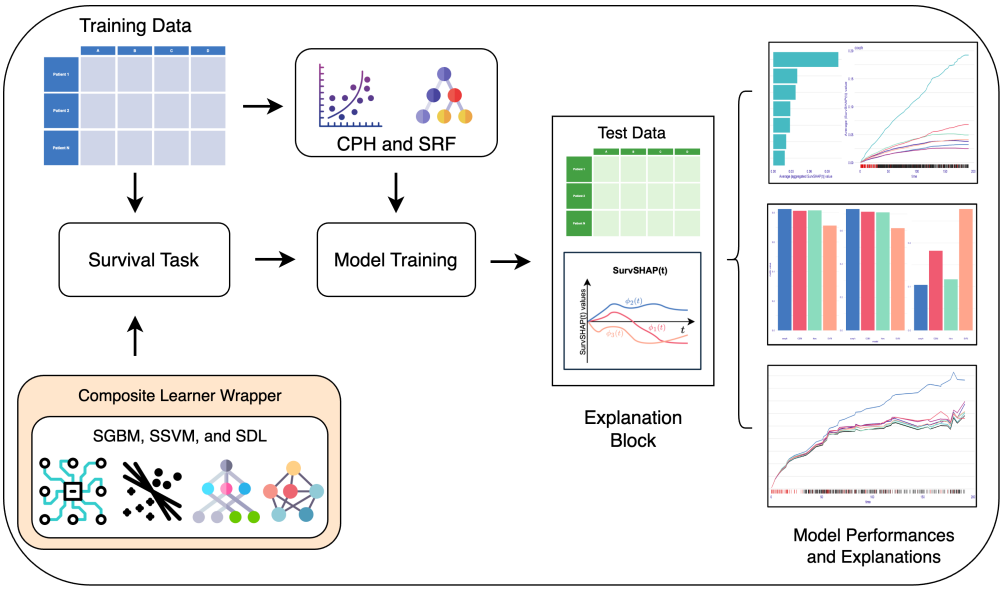


Fig. 4.3 Experimental Pipeline. For each trained model the related explainer object was created, with relative performance and explanation computation. Cox Regression (CPH), Survival Random Forest (SRF), Survival SVM (SSVM), Survival Generalized Boosted Model (SGBM), Survival Deep Learning Models (SDL).

4.1.4 Results

The performance of the models on the test set was evaluated using the following metrics: C-Index, Brier-Score and C/D AUC. Greater emphasis was placed on the C-index and Brier Score, as these were the most commonly used and interpretable metrics for survival tasks, offering a clearer understanding than the C/D AUC.

The evaluation metrics on test set are showed in Table 4.3 with a graphical comparison in Figure 4.4 and Figure 4.5, for ML and DL models, respectively.

Table 4.3 Survival models metrics computed on test set. Cox Regression (CPH), Survival Random Forest (SRF), Survival SVM (SSVM), Survival Generalized Boosted Model (SGBM).

Family	Model	C-Index	Integrated C/D AUC	Integrated Brier Score
Machine Learning	CPH	**0.81**	**0.72**	**0.10**
	SRF	**0.81**	0.70	0.12
	SSVM	0.71	0.61	0.15
	SGBM	0.79	0.69	0.14
Deep Learning	CoxTime	**0.78**	**0.73**	0.12
	DeepHit	0.73	0.70	0.13
	DeepSurv	0.57	0.60	0.16
	LogHazard	0.77	0.70	**0.11**
	PCHazard	**0.78**	0.71	0.13

ML Models Results As a general observation, SSVM emerged as the poorest performing model, while all other models demonstrated good results. The differences between Cox Regression, SGBM, and SRF were minimal in terms of C-index and Integrated C/D AUC, although Cox Regression exhibited a lower Brier Score.

Consequently, CPH was selected for the explainability step. Additionally, this model provided the Hazard Ratio for each feature, which could be utilized for data explainability alongside the XAI techniques discussed in Section 4.1.5.

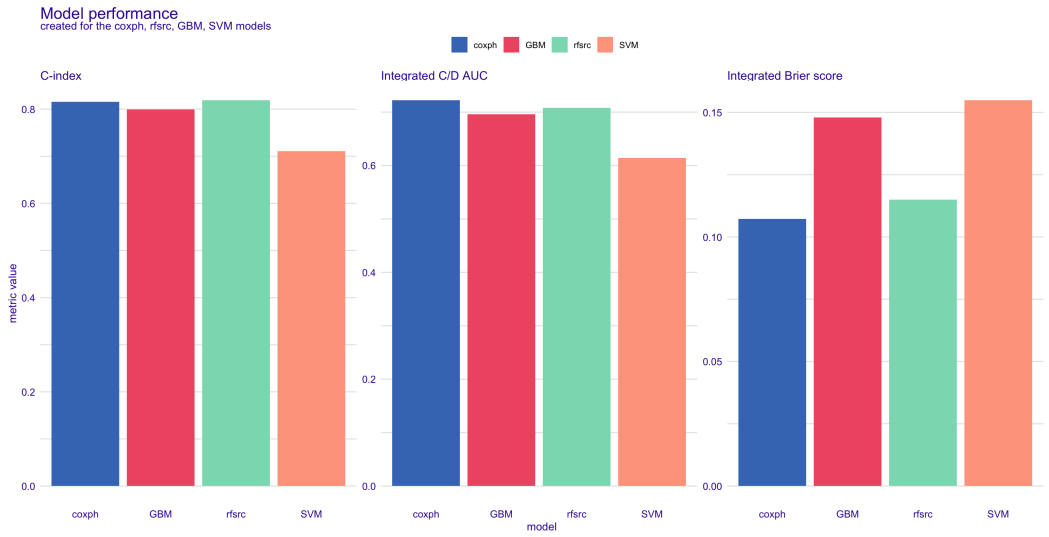


Fig. 4.4 Survival Machine Learning models metrics computed on test set; Cox Regression (CPH), Survival Random Forest (SRF), Survival SVM (SSVM), Survival Gradient Boosted Model (SGBM)

DL Models Results As illustrated in Figure 4.5, CT, PCH, and LH demonstrated comparable performances, positioning them as the best-performing models. LH was chosen for the explainability phase, with further details provided in Section 4.1.5.

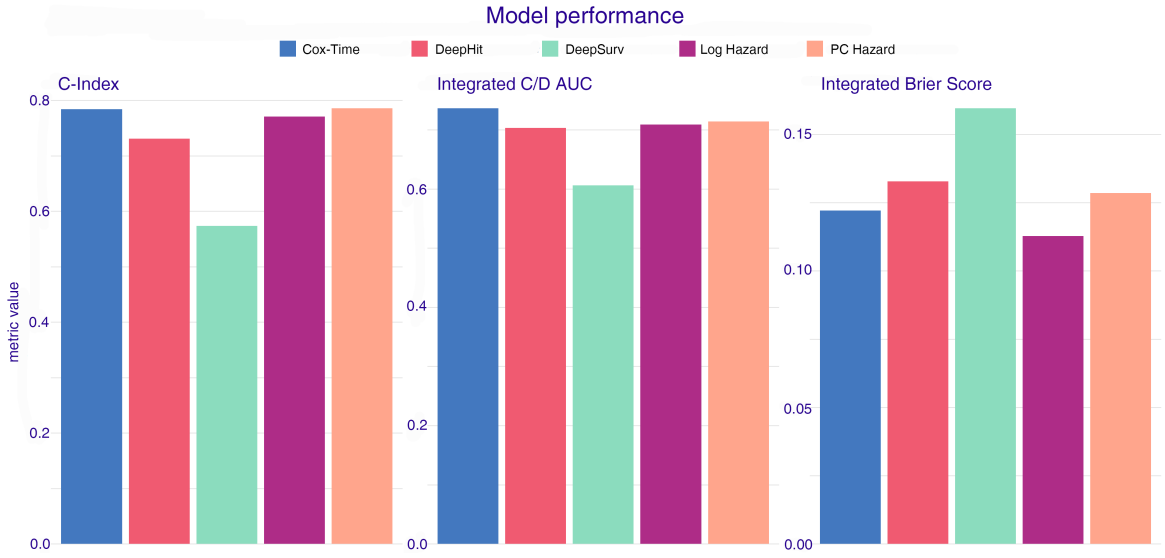


Fig. 4.5 Survival Deep Learning models metrics computed on test set.

The time variant Brier Score and C/D AUC for CPH, CT and LH models are depicted in Figure 4.6

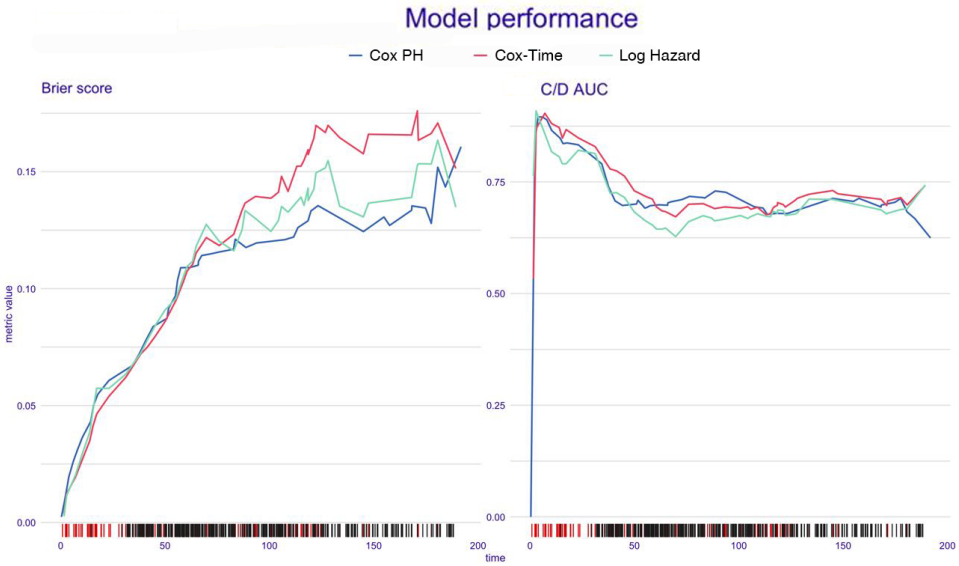


Fig. 4.6 Time variant models comparison for CPH, Cox-Time and Log Hazard models. The x-axis represents the event time expressed in months, where each tick represent the event (black - 0, red - 1).

4.1.5 Discussion

Both data and model-level time-dependent explanations were performed on the test set (419 samples) by comparing the survival models trained using ML and DL approaches. SurvSHAP was used to interpret and analyze the results, all of which are presented in relation to the survival function.

It was noted that generating explanations for deep learning models across all test samples can be computationally expensive, with minimal performance differences observed between the CT and LH models. For this reason, the LH model was selected for the explanation phase, due to its computational efficiency, enabling the generation of detailed explanations in a reasonable time frame.

To mitigate issues associated with the permutation method used for model-level explanations, features were filtered based on correlation coefficients, as explained in Section 4.1.2.

Dataset-level Explanation SurvSHAP values on test data were computed, retrieving and ranking the most important features.

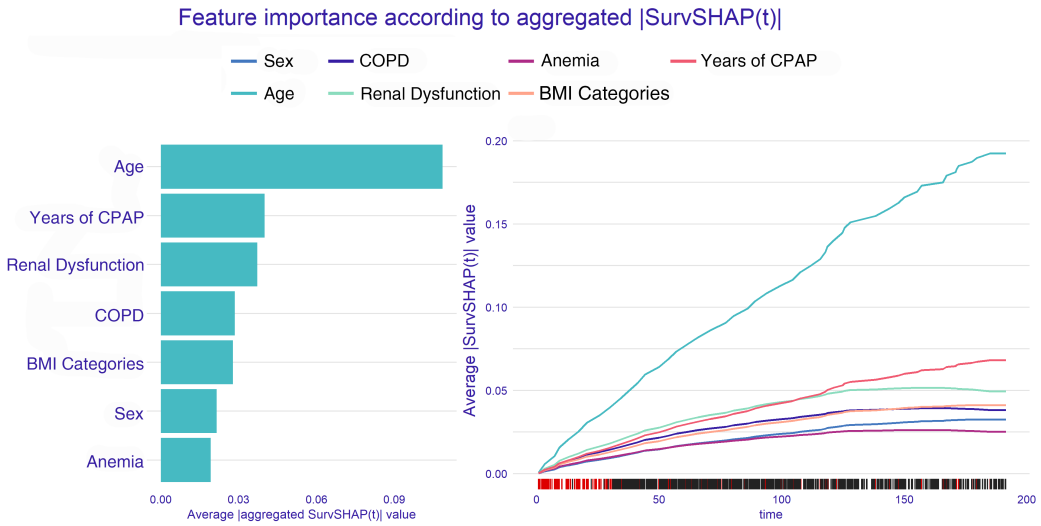


Fig. 4.7 Dataset level explanation for Cox Regression Model: on the left the features importance ranking according to the average of absolute shapley values. On the right the feature importance according to the observation time.In the right part of the figure, the x-axis represents the event time expressed in months, where each tick represent the event (black - 0, red - 1).

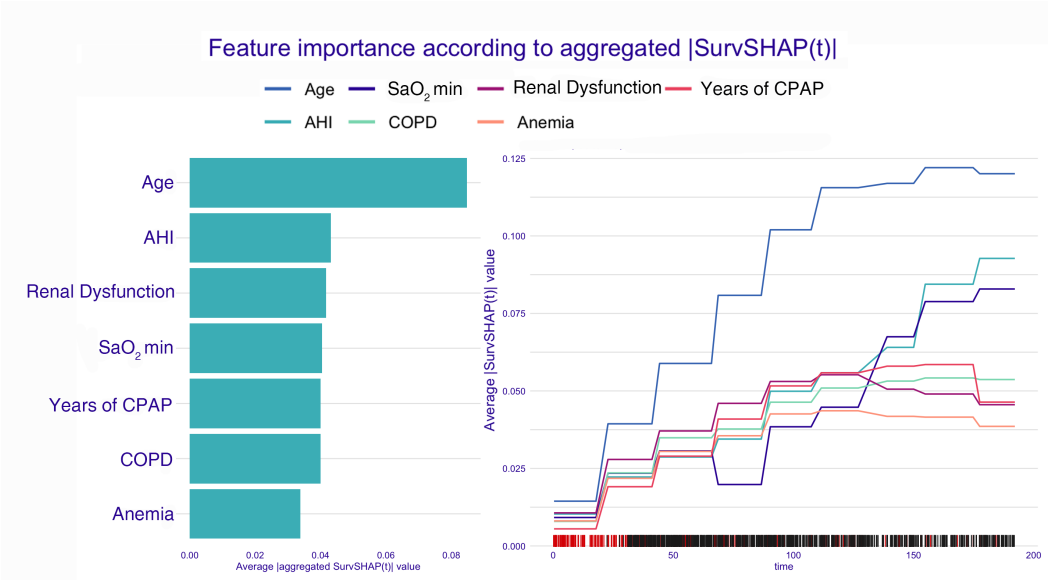


Fig. 4.8 Dataset level explanation for Log Hazard Model: on the left the features importance ranking according to the average of absolute shapley values. On the right the feature importance according to the observation time. In the right part of the figure, the x-axis represents the event time expressed in months, where each tick represent the event (black - 0, red - 1).

As depicted on the left side of Figure 4.7 relative to CPH, Age emerged as the most important feature, followed by Years of CPAP, Renal Dysfunction, COP, BMI Categories, Sex and Anemia.

Similarly, in Figure 4.8 Age remained a predominant feature, although the LH model assigned nearly equal importance to other features, with only minimal differences. More specifically, AHI ranked second, followed by Renal Dysfunction, SaO_2 minimum, Years of CPAP, COPD, and Anemia. In the LH model, these features demonstrated greater significance in terms of their average contribution to the prediction compared to the CPH model.

The discrepancy in feature importance could be attributed to the non-linear relationships between features and targets, uncovered by the LH model, which in certain cases prioritized numeric features over categorical ones. The temporal trends associated with the model outcomes are illustrated on the right side of Figures 4.7 and 4.8. nterestingly, in the CPH scenario, after approximately 9 years (around 110 months), the importance of Years of CPAP, Sex, and BMI Categories increased compared to Renal Dysfunction, Anemia, and COPD, respectively.

A similar pattern was observed in the LH model, where the discretization effect on predictions became noticeable. Although SaO_2 was initially ranked fourth in terms of feature contribution and appeared to be less important, its significance grew over time, emerging as the third most crucial feature in the latter part of the observation period.

Finally, while in the CPH model, all features contributed similarly in the final observation period, the LH model highlighted two distinct feature groups: one consisting of AHI and SaO_2, and the other comprising Renal Dysfunction, COPD, and Years_of_CPAP.

In contrast to classical XAI methods, where variable importance is presented as a static measure reflecting overall relevance, SurvSHAP allowed for variable importance to shift over time, offering a more dynamic evaluation of feature impact at any given time t.

Model-level Explanation

The second main evaluation step consisted in investigating the feature importance from the models perspective. This was accomplished by computing the difference between the loss function of the trained model and the loss function of the model with permutations. Specifically, the loss function (i.e. Brier Score) was computed

multiple times by changing the values of each single feature at time, while keeping the others unchanged. Features that lead to greater fluctuations in the difference are considered more influential according to the model perspective. The results are depicted in Figure 4.9 for CPH and Figure 4.10 for LH.

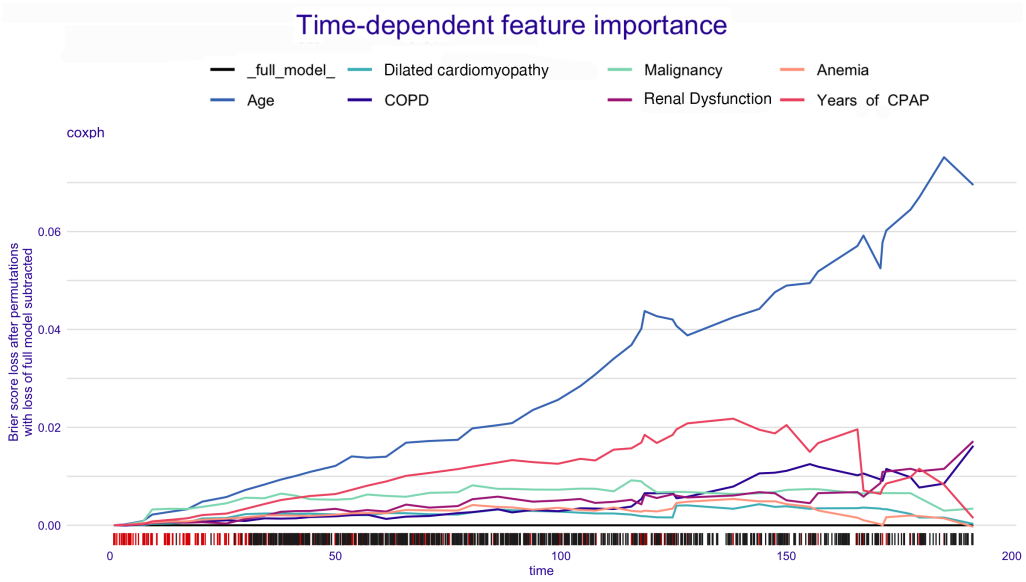


Fig. 4.9 Time-dependent feature importance for Cox Regression model, obtained by subtracting full model Brier Score from the Brier Score after single feature permutations.

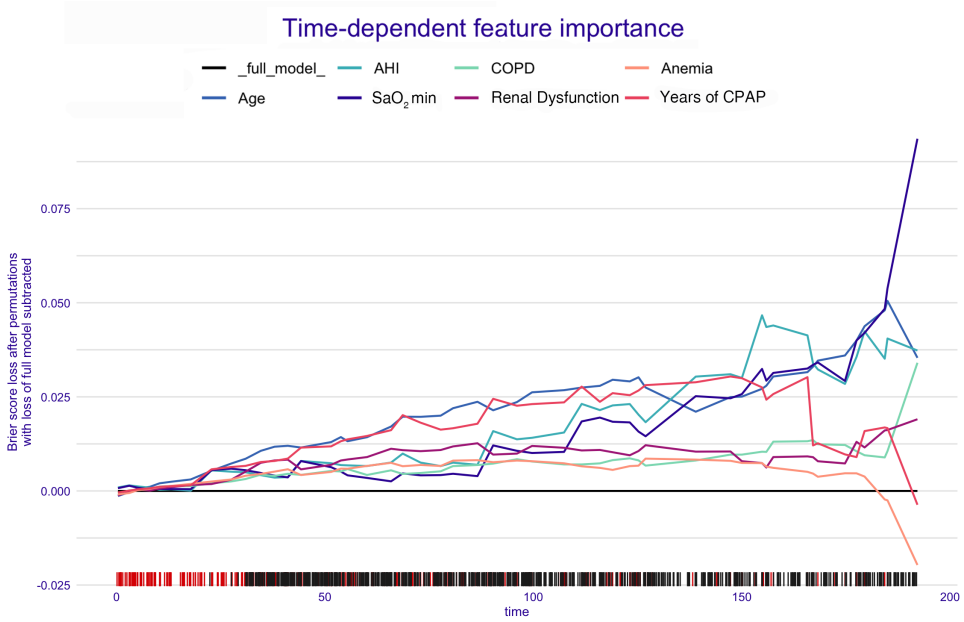


Fig. 4.10 Time-dependent feature importance for Log Hazard model, obtained by subtracting full model Brier Score from the Brier Score after single feature permutations.

The relationships between OSA and the most important features retrieved by SA models found confirmation in the medical literature [168–172]. Such comorbities revealed to be predictors of a lower survival probability for people with OSA in a previous work [143]. The most important features identified for the CPH model aligned with the features highlighted by computing the SurvSHAP values on the test set. Remarkably, additional features such as Malignancy and Dilated Heart Disease were not reported in the data-level explanations. This suggested that, while certain features may not stand out prominently in the individual data instances (i.e., they are underrepresented in data), they still hold a significant weight when considered from CPH model-level perspective. Such conclusion was also confirmed by looking the related Hazard Ratios in Table 4.4.

Table 4.4 Cox proportional hazards matrix with features sorted by Hazard Ratio in descending order.

Variable	**Coef**	**Exp.coef**	**Se.coef**	**Z**	**Pr...z..**
Malignancy	1.83	6.21	0.29	6.30	2.89E-10
Idiopathic dilated cardiomyopathy	1.41	4.08	0.48	2.92	0.00
COPD	0.53	1.70	0.14	3.81	1.37E-4
Renal dysfunction	0.56	1.75	0.15	3.81	1.37E-4
Age	1.37	3.94	0.18	7.75	9.28E-15
Anemia	0.40	1.49	0.15	2.73	0.01
Atrial fibrillation	0.37	1.45	0.23	1.64	0.10
Heart failure	0.35	1.41	0.28	1.24	0.22
Diabetes	0.11	1.12	0.14	0.78	0.43
Sex	0.30	1.35	0.16	1.81	0.07
Ipertension	0.02	1.02	0.13	0.15	0.88
Cardiovascular disease	-0.01	0.99	0.19	-0.06	0.95
BMI Categories	-0.07	0.93	0.06	-1.12	0.26
Cholesterol categories	-0.08	0.92	0.10	-0.83	0.41
Valvular disease	-0.02	0.98	0.42	-0.05	0.96
SaO2 min	-0.01	0.99	0.01	-1.16	0.25
AHI	-0.01	0.99	0.00	-1.54	0.12
Years of CPAP	-0.13	0.87	0.03	-5.06	4.11E-7

However, although the presence of malignancy and the dilated cardiomyopathy condition had a strong influence on the individual survival, such features are not always available since they are not common in population and cannot be used in clinical practice.

In addition, unlike the data-level explanations where the contribution of CPAP treatment period increased over time, in the general model it loss its importance in the final part of the observation period, thus lessening the contributions related to Renal Dysfunction and COPD. As concerns LH model (Figure 4.10), the features identified were the same retrieved by applying SurvSHAP to the test set. Differently from CPH model, the contribution of Age was almost equal to the other features. Surprisingly, AHI gave the greatest contribution from $\sim$150 to $\sim$170 months ($\sim$ 12.5 to 14 years); besides, starting from $\sim$125 months the SaO_2 min started to gain more importance until ending to be the most important feature. Comparing Figures 4.9 with 4.10, LH model computed the feature contribution in a more balanced way, w.r.t. CPH.

Notably, while having similar performance, the selected models focus on different feature sets. Specifically, CPH identified several features that seem related to the mortality in a general way, such as dilated heart disease and the presence of malignancies, while giving a greater importance to the Age. On the other hand, LH focuses on features that are related to the OSA pathology, such as minimum oxygenation level (SaO_2 min) Apnea and Hypopnea Index (AHI), which gains importance over the observation time. In light of this, the use of SurvSHAP lead clinicians to assess that, to parity of performances, LH model results more reliable here. Ultimately, LH model results also more useful, since AHI and SaO_2 min are more useful features compared to the presence of malignancy or dilated cardiomyopathy, because the latter are rarer conditions in the population, as demonstrated with survival curves depicted in Figure 4.11, (thus making such data not always applicable) and because the former two are directly derived from polysomnography.

The summary of the differences in the retrieved models from XAI(t) perspective is depicted in Table 4.5.

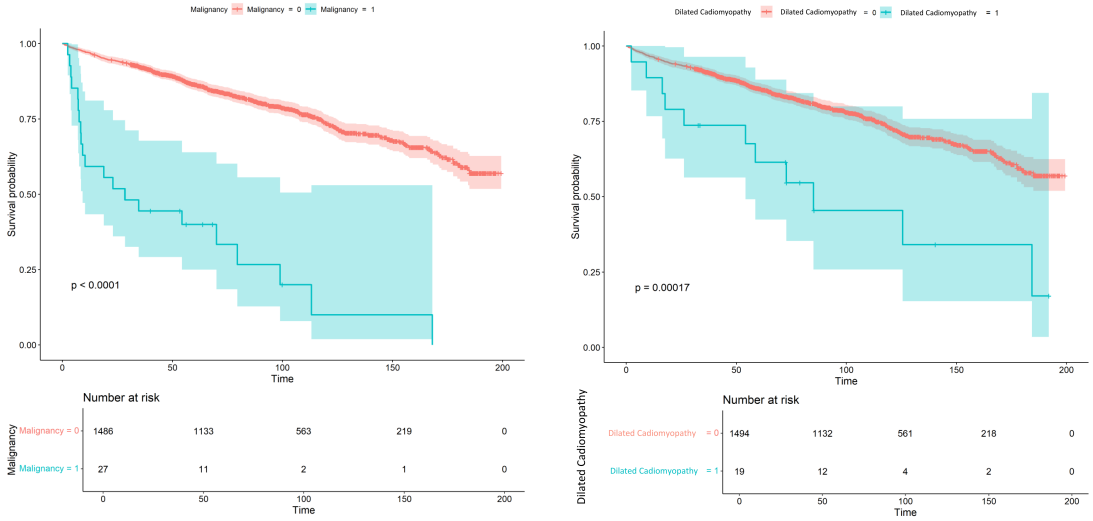


Fig. 4.11 Survival curves for malignancy and dilated cardiomyopathy features.

Table 4.5 Summary of differences between CPH and LH model. DCM - Dilated Cardiomyopathy.

	Performance Metrics		**Data-Level Explanation (419 test samples)**			**Model-Level Explanation**		
	C-Index	**Brier Score**	**Relevant Features**	**Prevailing Feature**	**Observations**	**Relevant Features**	**Prevailing Features**	**Observations**
CPH	0.81	0.10	Age, Years of CPAP, Renal Dysfunction, COPD, BMI, Sex, Anemia	Age	Huge gap in Age contribution w.r.t other features; the features contributions have few variations over the time.	Age, Years of CPAP, Renal Dysfunction, COPD, Anemia, Malignancy, DCM.	Age, followed by Years of CPAP	Age still prevails on other features; Malignancy and DCM are not strictly related to the mortality and they are not so common in population.
LH	0.77	0.11	Age, AHI, Renal Dysfunction, SaO2 min,Years of CPAP, COPD, Anemia	Age	Moderated gap between Age contribution and other features; these ones provide the same contribution, but it variates over the time.	Age, Years of CPAP, Renal Dysfunction, COPD, Anemia, AHI, SaO2min.	All features give the same contribution	The relevant features give almost the same contribution. AHI and SaO2min are more useful and accessible in OSA context.

4.2 A deep learning approach for Oxford Classification of glomeruli lesions

Another contribution of DL in pathomics of this thesis work relies in its application to the Oxford Classification system, which is a highly valuable tool in nephropathology for assessing IgA Nephropathy (IgAN) [173–175]. The Oxford Classification evaluates lesions in two main tissue compartments: the glomeruli and the cortical tubulointerstitium. From this evaluation, three binary components (M: mesangial hypercellularity, E: endocapillary hypercellularity, and S: segmental glomerulosclerosis) and two ordinal components (C: crescent formation and T: tubular atrophy/interstitial fibrosis) are derived.

By integrating these histopathological features with clinical parameters, the IgAN risk prediction score can be computed, which helps predict the progression of the disease . Machine learning could support nephropathologists by automating the interpretation of histological features and clinical data, thus offering more consistent and accurate risk predictions, potentially leading to better patient outcomes.

This work introduced MESCnn [41] (MESC classification by neural network), a novel decision support system for nephropathology that incorporates instance-level segmentation of glomeruli for PAS-stained sections in IgAN cases. This system adheres closely to the Oxford Classification's standards for glomerular lesion classification. This distinction stems from the fact that it supports prognosis within the framework of the International Risk Prediction Tool for IgAN (including the Oxford Classification), but only after IgAN diagnosis has already been confirmed through other nephropathological methods like electron microscopy and immunostaining.

The focus was placed on four key glomerular score components—mesangial hypercellularity (M), endocapillary hypercellularity (E), segmental sclerosis (S), and active crescents (C) adhering strictly to the original definitions of the Oxford Classification. This was done using periodic acid-Schiff (PAS) sections to ensure alignment with established guidelines. Unlike previous studies, which either extended beyond IgAN classification, employed different staining techniques like trichrome, or focused on only a subset of the components such as mesangial hypercellularity or hypercellularity alone, this work concentrated on implementing the full spectrum of these components according to PAS staining protocols.

A custom glomerular segmentation module was developed for MESC classification, specifically trained on a large and diverse dataset from three different institutions, encompassing 11 common classes of glomerulonephritis. This allowed for the creation of a comprehensive end-to-end pipeline designed to classify glomerular lesions from whole slide

images (WSIs). The related works for glomerular classification and segmentation of MEST-C lesions are reported in Table 4.6 and Table 4.7

Table 4.6 Related works for the classification of MEST-C lesions.

Author	Year	MEST-C	Method	Data	Stain
Chagas et al. [176]	2020	M, E	CNN+SVM	FIOCRUZ	H&E, PAS
Yang et al. [**?**]	2021	E, S, C	Mask R-CNN + LSTM & ResNeXT-101	LKCGMH, KSCGMH, KMUMH	H&E, PAS, PAM, trichrome
Purwar et al. [**?**]	2020	M	KNN, SVM vs CNN (TL)	138 glomeruli (×20) from IgAN patients (AIIMS, Delhi)	PAS
Weis et al. [**?**]	2022	M, S	CNN + CAM (XAI)	12,253 images + 11,142 images + 180 consensus images	PAS
Sato et al. [**?**]	2021	(unsupervised)	CNN (NAS-Net) + Clustering + Score-CAM, Grad-CAM	68 patients with IgAN	H&E
Jaugey et al. [**?**]	2023	M, E, S, T, C	CNN	42 biopsies (train) + 66 biopsies (test) + 88 biopsies (application)	Masson's trichrome
Uchino et al. [177]	2020	M, E, S, C	fine-tuning of InceptionV3	283 kidney biopsies with 15,888 glomerular images	PAS, PAM

Table 4.7 Related works for glomerular segmentation.

Author	Year	Acquisition Magnification	Resolution [µm/pixel]	Patients (Biopsies)	Staining	Glomeruli	Method	Tile Size	Downsampling Magnification	Downsampling Factor
Zeng et al. [?]	2020	40×	N/A	400 slides from IgAN patients (360 training; 40 test)	PAS	12,418 glomeruli (train: 10,935; 1483 test)	U-Net (structural similarity loss)	1024 × 1024	N/A	N/A
Bueno et al. [178]	2020	20×	N/A	47 PAS WSIs (38 training, 9 test)	PAS	1245 glomeruli (303 sclerosed; 942 normal)	SegNet	400 × 400	4×	5 times from 20×
Altini et al. [?]	2020	20×	0.50	26 PAS WSIs (19 training, 7 test)	PAS	2344 non-sclerotic glomeruli (1852 training, 492 test); 428 sclerotic glomeruli (341 training, 87 test)	Faster R-CNN; Mask R-CNN	500 × 500	5×	4 times from 20×
Jiang et al. [179]	2021	40×	0.12	348 WSIs from 148 patients (training: 296 WSIs from 118 patients; test: 52 WSIs from 30 patients)	PASM, PAS, MT	8665 glomeruli (7193 training, 1472 test). PAM: 3248; PAS: 2525; MT: 2892.	Cascade Mask R-CNN	2048 × 2048	N/A	N/A
Salvi et al. [?]	2021	10×	0.934	83 patients; 50 patients train glomeruli; 11 patients test glomeruli	PAS	587 glomeruli (473 train; 114 test)	RENTAG	512 × 512	N/A	N/A
Yang et al. [?]	2021	40×	0.23	Used 1379 kidney biopsies from LKCGMH. 60 cases for testing detection model; LKCGMH: 20; KSCGMH: 20; KMUMH: 20	H&E,PAS, PAM, MT	Annotated 15,298, 5649, 5641, 5679 glomeruli from H&E, PAS, PAM, MT. 8633 glomeruli for testing detection model; LKCGMH: 1585; KSCGMH: 4211; KMUMH: 2837	Mask R-CNN	512 × 512	0.625×	64 times from 40×
Kawazoe et al. [?]	2022	40×	0.23	600 WSIs for detection model (6-fold cross-validation, Facility T: 200 WSIs train, 50 validation, 50 test; Facility K: 200 WSIs train, 50 validation, 50 test)	PAS	Only average number of glomeruli per WSI reported in the paper. Roughly 18000 glomeruli for detection models. Roughly $24*300 = 10200$ glomeruli form Facility T; $26*300 = 7800$ glomeruli from Facility K.	Faster R-CNN	500 × 500	5×	8 times from 40×
Jaugey et al. [?]	2023	20×	0.454	196 IgAN (42 training, 66 test, 88 application)	MT	Only average number of glomeruli per WSI reported in the paper. Roughly $13*196+3*196 = 3546$ glomeruli. Training: roughly $16*42+3*42 = 798$. Test: roughly $8*66+3*66 = 726$. Application: roughly $16*88+3*88 = 1672$.	Mask R-CNN Inception ResNet V2	N/A	2.5×	8 times from 20×

4.2.1 Materials and Methods

A total of 386 WSIs from various kidney biopsies were annotated using QuPath software [180] by a specialized nephropathologist. The dataset was divided into two groups: 102 biopsies diagnosed with IgAN and 284 biopsies diagnosed with other forms of glomerulonephritis (referred to as Other GN). The Other GN cases covered a broad spectrum of glomerulonephritis variations, representing all morphological forms of IgAN observed in PAS-stained WSIs.

The Other GN cohort was employed in developing the segmentation models, while the IgAN cohort was used for the classification models. This two-step approach ensured that the models were well-trained for both segmentation and classification tasks, using a diverse and representative set of biopsy samples.

Annotations covered both sclerosed and non-sclerosed glomeruli, with the exception of empty Bowman's capsules. When feasible, the annotations followed the Bowman's capsule outline. For dislodged glomerular tufts, all cellular and matrix components were annotated according to the Bowman's capsule trajectory. The biopsies used for these annotations were sourced from three institutions: the Institute of Pathology in Cologne, the University of Szeged, and the Nephrology Department in Bari. PAS-stained sections from these biopsies were scanned at a resolution of either approximately 0.23 $\mu m/pixel$ or 0.12 $\mu m/pixel$, using different imaging equipment. In Cologne and Bari, a NanoZoomer Scanner (Hamamatsu, Herrsching am Ammersee, Germany) with a $40\times$ objective was used, while in Szeged, a Pannoramic Midi Slide Scanner (3DHISTECH, Budapest, Hungary) was employed. Each biopsy included for analysis met the Oxford scoring system's minimum criterion of at least eight glomeruli. The focus on PAS staining was critical since the study aimed to develop a precise Oxford scoring system, which, by definition, is restricted to PAS-stained samples [174].

102 renal biopsies diagnosed with IgAN were involved for the Oxford classification [174], ensuring the exclusion of IgA vasculitis (Henoch-Schönlein Purpura) through clinical assessment [176].

All glomerular sections were uploaded to the Labelbox platform (www.labelbox.com(for expert nephropathologist annotation. The total count of labeled glomerular sections reached 6206.

Segmentation Dataset The dataset used for developing the glomerular segmentation model included 284 biopsies, corresponding to 748 whole slide images (WSIs). These biopsies came from the PanGN cohort, which covers 11 different classes of glomerulonephritis (GN). The GN classes represented in this cohort include anti-glomerular basement membrane antibody GN, anti-neutrophil cytoplasmic antibody GN, C3-GN, cryoglobulinemic GN, dense deposit disease, infection-associated GN, membranous nephropathy, idiopathic membranoproliferative GN, proliferative GN with monoclonal immunoglobulin deposits, and systemic lupus erythematosus GN class IV. Importantly, the 102 biopsies related to IgA nephropathy (IgAN), accounting for 308 WSIs, were excluded from the dataset to prevent overlap with the target dataset.

The dataset was divided into training and test sets for segmentation model development. The training set consisted of 227 biopsies (587 WSIs), while the test set included 57 biopsies (161 WSIs).

In total, the dataset for both glomerular segmentation and MESC classification models consisted of 386 biopsies, of which 102 were specific to IgA-GN. All biopsies were stained with PAS and were sourced from three institutions: the University Hospital of Cologne (Germany), the Department of Emergency and Organ Transplantations (DETO) at Bari University Hospital (Italy), and the Szeged University (Hungary).

Classification Dataset The dataset used for developing the classification model consisted of 102 biopsies, representing 308 WSIs, and a total of 6206 glomerular crops. It was divided into a training set of 67 biopsies (207 WSIs, 4298 glomerular crops) and a test set of 35 biopsies (101 WSIs, 1908 glomerular crops). These subsets were used for training the classification models aimed at predicting the Oxford labels: M (mesangial hypercellularity), E (endocapillary hypercellularity), S (segmental glomerulosclerosis), and C (active crescents).

After predicting the Oxford labels, they were translated into biopsy-level Oxford scores as M0/M1, E0/E1, S0/S1, and C0/C1/C2, according to the methodology described previously. An expert nephropathologist independently assigned biopsy-level Oxford scores based on the same WSIs, following the Oxford classification guidelines. Importantly, the dataset only had one biopsy with a C2 score (indicating more than 25% active crescents), highlighting the rarity of C2 cases in clinical practice. For the purpose of statistical analysis, C1 and C2 categories were merged.

The distribution of ground truth label classes is shown in Figure 4.12, while Figure 4.13 provides sample images from the dataset. Notably, for E and C lesions, it is possible to observe and imbalance for positive labels with respect to the negative ones. This dataset was used to develop the MESCnn pipeline, which facilitates the segmentation of glomeruli and the classification of lesions according to the Oxford M, E, S, and C scoring system. Noteworthy is the distinction between the annotations applied to individual glomerular level and the biopsy-level Oxford scores. This study focused only on the glomerular components of the Oxford score (MESC), while the T score (cortical Tubular atrophy) and interstitial fibrosis fell outside the scope of this research.

- Mesangioproliferation (**M**): Mesangial cells are situated at the core of the glomerular tuft, extending between capillary loops. According to the Oxford Classifi-

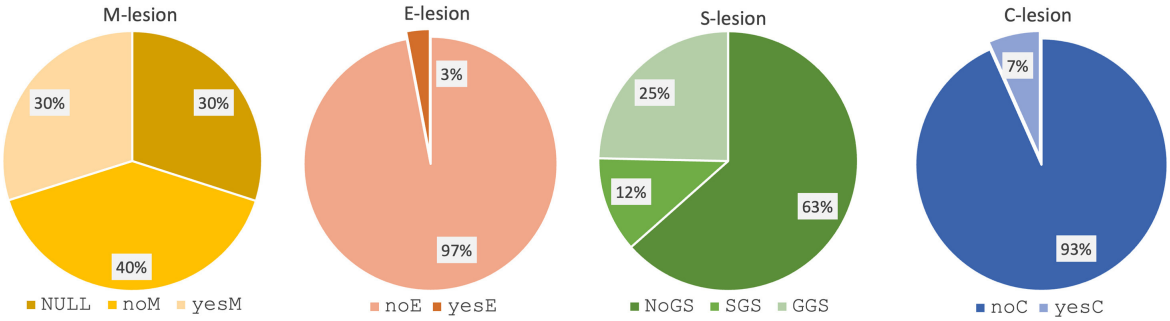


Fig. 4.12 *Lesions distribution*. Pie charts representing the M, E, S, C lesion distribution in the dataset collected from the three cohorts. The number of samples for each lesion is reported in Table 4.8.

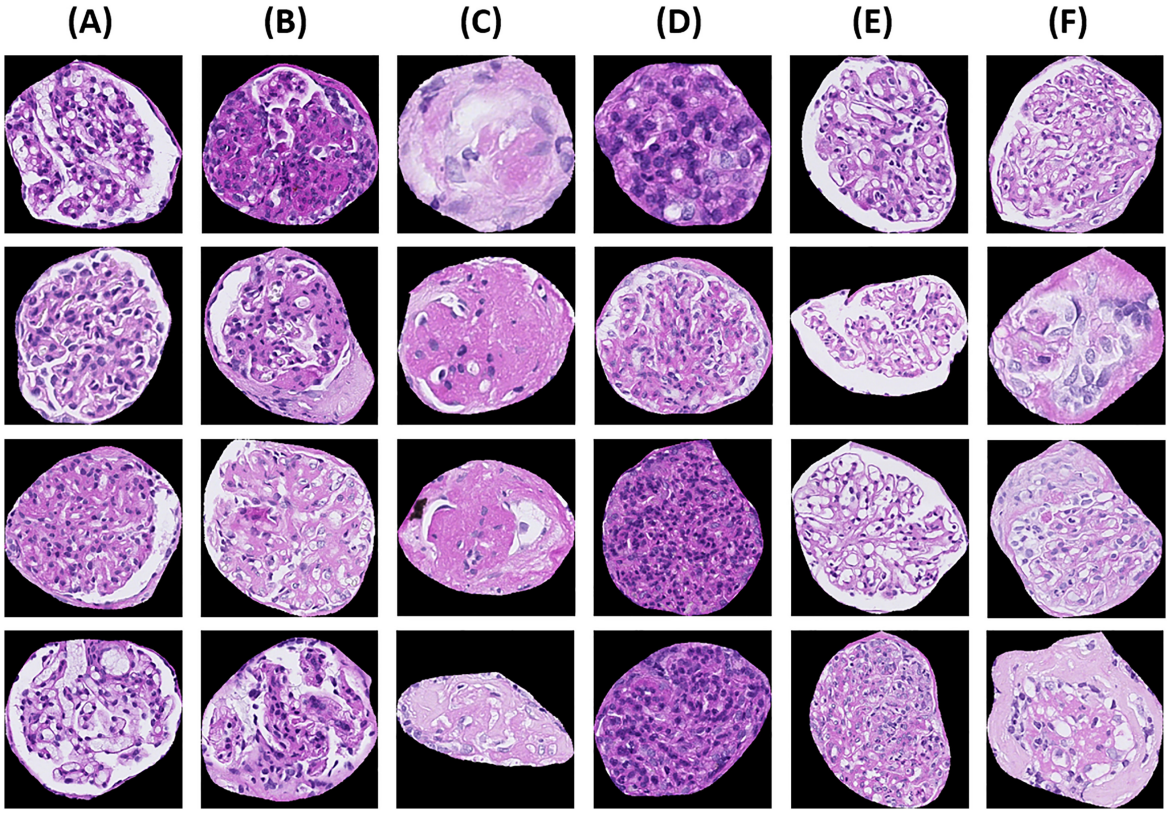


Fig. 4.13 *Sample images from the dataset used for classification*. (A) Masked (extraglomerular background set to black) glomerular crops without any M (mesangioproliferation), E (endocapillary hypercellularity), S (segmental glomerulosclerosis) or C (active crescent) lesion coded in the Oxford Classification; (B) Glomerular crops with ground truth S label applied by an expert nephropathologist; (C) Global Glomerulosclerosis; (D) Mesangioproliferation; (E) Endocapillary Hypercellularity; (F) Cellular or Fibrocellular Crescent.

cation, the term "mesangioproliferation" is assigned to glomeruli displaying clusters of more than three mesangial cells in a mesangial area, excluding the stalk region. Based on the Oxford Classification scoring system, the label "M" is categorized as "noM" (no mesangioproliferation), "yesM" (mesangioproliferation), or "indeterminate" (NULL). If more than 50% of the glomeruli exhibit mesangioproliferation, the patient is assigned a score of M1; otherwise, the score is M0 [174].

- Endocapillary Hypercellularity (**E**): Endocapillary hypercellularity refers to an increase in leukocytes within the glomerular capillaries. The Oxford Classification uses a binary system for labeling: "noE" for absence and "yesE" for presence of endocapillary hypercellularity. The patient receives an E1 score if at least one glomerular section exhibits endocapillary hypercellularity, and an E0 score if none does.
- Segmental Glomerulosclerosis (**S**): Segmental glomerulosclerosis (SGS) describes a condition where scarring affects less than the entirety of the glomerular capillary loop. The Oxford Classification includes a binary labeling system: "noGS" for no glomerulosclerosis, "SGS" for segmental glomerulosclerosis, and "GGS" for global glomerulosclerosis. If any glomerular section in a biopsy shows segmental glomerulosclerosis, the score is S1; if none shows such lesions, the score is S0.
- Active Crescent (**C**): This lesion involves extracapillary crescent formation with a cellular content of at least 10% compared to matrix content. The Oxford Classification assigns labels of "noC" or "yesC" based on the presence of active crescents. Biopsies are scored C1 if they contain up to 25% of active crescents, C2 if more than 25%, and C0 if none [175].

The data distribution according to the labels is reported in Table 4.8

4.2.2 Experimental Pipeline

Segmentation A comprehensive evaluation of several Mask R-CNN variants was performed to identify the optimal architecture for the glomerular segmentation task. Initially, segmentation results were generated tile by tile and then mapped onto WSIs for further processing. This was integrated into QuPath, under the name "QuPath Interface for Glomeruli Segmentation" (QIGS), which facilitates automatic glomeruli segmentation

Table 4.8 Sample data distribution according to MESC labels.

Cohort	Classification									
	M lesion			E lesion		S lesion			C lesion	
	nan	noM	yesM	noE	yesE	GGS	NoGS	SGS	noC	yesC
Bari	90	241	34	360	5	40	307	18	336	29
Cologne	1233	1922	1370	4417	108	1000	2969	556	4250	275
Szeged	537	329	450	1242	74	491	662	163	1206	110
Train	1277	1771	1250	4191	107	1063	2739	496	3994	304
Validation	583	721	604	1828	80	468	1199	241	1798	110
Total	**1860**	**2492**	**1854**	**6019**	**187**	**1531**	**3938**	**737**	**5792**	**414**

of PAS-stained WSIs, enabling easier use of these results for downstream tasks, such as the Oxford classification or export via Python scripts interfacing with QuPath projects.

The models were trained using Stochastic Gradient Descent with Momentum (SGDM), adhering to default Detectron2 settings, with a learning rate of 3e-4, running up to 300,000 iterations, and using batch sizes of two tiles per iteration. For model training and validation, WSIs were split into 1024×1024 pixel tiles, taken at a 10× magnification. Overlaps of 512 pixels per axis were included to ensure glomerular regions were fully represented in at least one tile.

During inference, results were initially generated at the tile level. Overlap between adjacent tiles helped prevent missing glomeruli near tile edges. To address duplicate detections, the Non-Max-Area-Suppression (NMAS) [181] algorithm was used on detections projected back into WSI space. Results were stored in a QuPath project, with a Python-based interface using the PAQUO library to link Detectron2 and QuPath.

Classification For the image classification tasks, two types of deep learning models were predominantly utilized: CNNs and ViT. In preparing the glomerular images for training and validation datasets, several pre-processing steps were applied to ensure the images were in optimal form:

1. Mask Application: A mask was applied to isolate the glomerular regions, eliminating irrelevant pixels outside the target area based on expert-guided glomerular segmentation.

2. Zero Padding: This process ensured the glomerular crops became square-shaped, which is often required by CNNs and ViTs to maintain consistency in input dimensions.
3. Resizing: Images were resized to 256×256 pixels, while maintaining their aspect ratio to avoid distortion or stretching.

To enhance variability in training data and minimize overfitting, an augmentation technique based on the Pytorch AutoAugmentPolicy.IMAGENET policy was implemented, introducing random variations to the images. For model training, the Adam optimizer was employed with an initial learning rate set to 1e-5. Training lasted for 50 epochs. The inherent imbalance in lesion class distributions (e.g., the E lesion, representing only 2.49% of glomeruli in the training set) was addressed using PyTorch's WeightedRandomSampler. This approach assigned a weight to each class inversely proportional to its frequency, ensuring more balanced class representation during the model's learning process. Additionally, the multi-class cross-entropy loss function was employed to train the classifiers effectively.

Evaluation To assess the performance of instance segmentation models, the metrics employed were defined as by the COCO dataset evaluation framework, which is widely used in object detection tasks. The main metrics considered include:

- Average Precision (AP): This metric calculates the area under the precision-recall curve, averaged over Intersection over Union (IoU) thresholds ranging from 0.5 to 0.95, in increments of 0.05. It provides a balanced overview of the segmentation accuracy across different IoU thresholds, which accounts for both object localization and segmentation quality.
- AP50 and AP75: These specific variants of AP focus on fixed IoU thresholds. AP50 measures precision when the IoU threshold is set at 50%, while AP75 is based on a stricter 75% IoU threshold. These thresholds provide additional insights into how well the model performs at different levels of overlap between predicted and ground truth regions.
- Aggregated Jaccard Index (AJI): This metric combines both detection and segmentation accuracy into a single measure, making it suitable for assessing the overall performance of instance segmentation models. AJI measures the similarity between the predicted and actual segmented regions, offering a comprehensive evaluation of both object detection and pixel-wise segmentation.

- Dice Coefficient: This metric, commonly used in medical imaging, is computed as 2TP/(2TP + FP + FN), where TP, FP, and FN stand for true positives, false positives, and false negatives, respectively. The Dice coefficient focuses on the overlap between the predicted and actual segmented areas, and is particularly effective for evaluating pixel-level segmentation performance.

For the classification of M, E, S, and C lesions, various CNNs and ViT architectures were leveraged to classify these glomerular lesions. AUROC and AUPRC metrics were used to assess classification performance. Finally, to visually inspect the quality of features extracted by the classification models, UMAP was exploited. This technique helps in visualizing high-dimensional data by projecting the feature representations of the model into a lower-dimensional space. UMAP plots allowed to compare the feature distributions of models trained from scratch on the dataset with those using pretrained models (such as those pretrained on ImageNet), providing insights into how well the models differentiate between the lesion classes.

The segmentation and classification modules were assembled to create an end-to-end pipeline, utilizing PAS WSIs as input, along with Oxford M, E, S, and C labels for individual glomerular crops. The pipeline then produced Oxford M, E, S, and C classifications at the biopsy level. This pipeline, named MESCnn, is depicted in Figure 4.14 which outlines the comprehensive workflow used during its development and execution.

The "QuPath Interface for Glomeruli Segmentation" (QIGS) was deployed by integrating it with the QuPath software using the PAQUO library. This integration enabled the visualization of segmentation results and allowed for their export to subsequent classification stages. Additionally, pathologists could review the segmentation results, providing expert oversight where needed.

The end-to-end pipeline, MESCnn, which generated a spreadsheet report as output and used WSIs as input, was made available through a repository https://github.com/Nicolik/MESCnn. This repository includes the QIGS module, which is responsible for creating the QuPath project with segmented glomerular masks. The weights of the trained models and example WSIs were also shared for reference and testing purposes. Links to both the pipeline and the QIGS module were provided for public access.

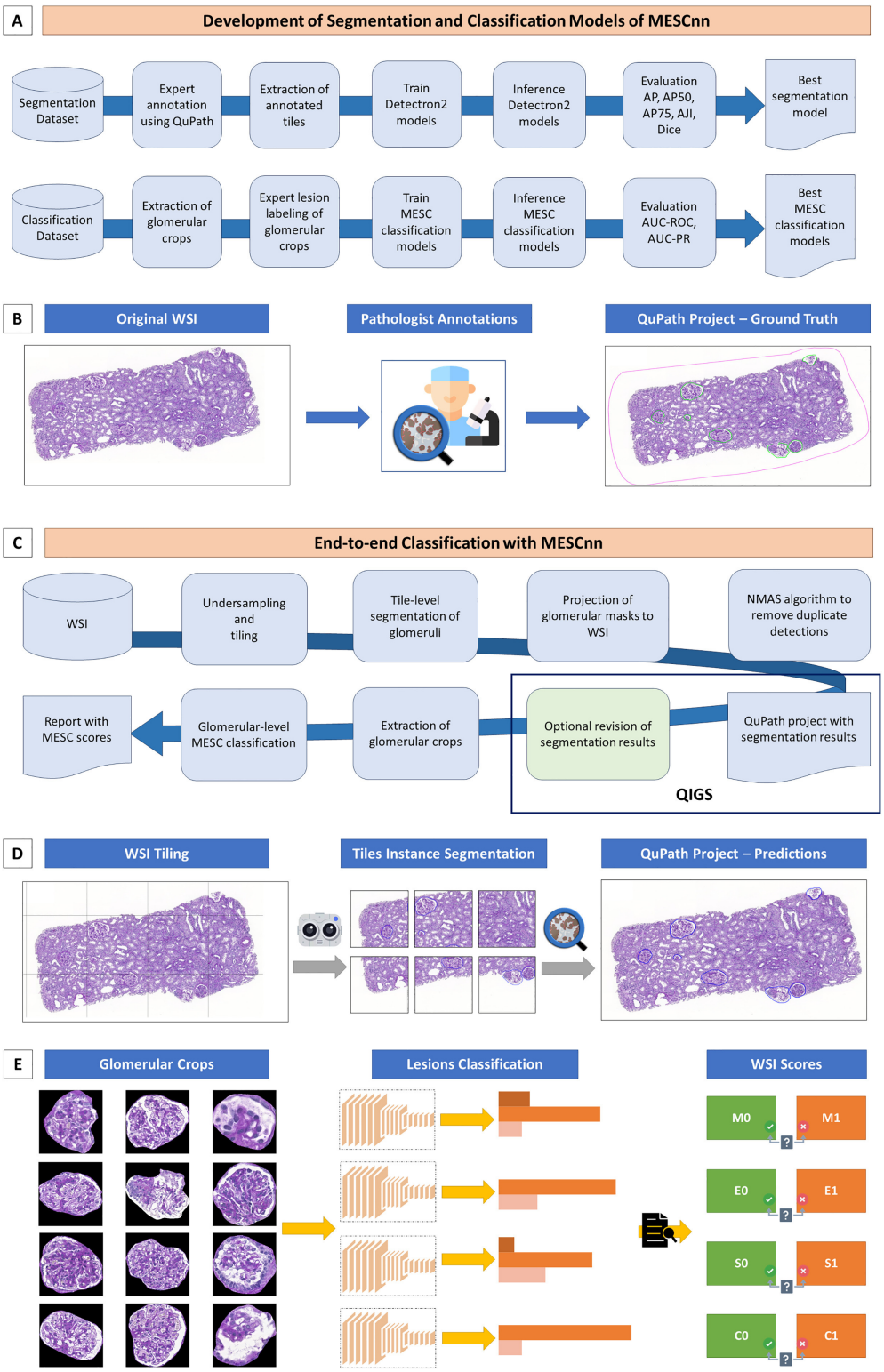


Fig. 4.14 *End-to-end workflow for glomeruli segmentation and classification of M, E, S, C lesions with the proposed MESCnn pipeline*. (A) Models development stage. (B) Glomerular annotation generation by pathologists using QuPath. (C) End-to-end usage of the proposed pipeline. (D) Instance segmentation of glomeruli. (E) Classification of M, E, S, C lesions taking advantage of convolutional neural networks and vision transformers. Finally, WSI scores are determined by applying decision rules as defined in the Oxford Classification.

4.2.3 Results

Segmentation Results The segmentation results obtained using Mask R-CNN and its variants demonstrated impressive performance, the AP50 reached about 80% on the validation set and roughly 78% on the test set for the instance segmentation task. The AJI peaked around 76% on the validation set and approximately 73% on the test set. The detection models exhibited robust performance across several metrics, reflecting their effectiveness in the instance segmentation task. The best detection model was chosen based on the average performance across evaluation metrics on the validation set. On the test set, AP values were ranging from 61.2% to 62.8%, with an average of 62.1%. AP50 ranged between 75.1% to 77.7%, averaging 76.5%, while the AP75 reached a mean of 71.0%. For AJI, the the test set values ranged between 69.1% and 73.4%, with an average of 72.2%. Pixelwise segmentation, measured using the Dice coefficient, produced results between 79.0% and 80.7% (with an average of 79.8%) on the test set. These results highlight the effectiveness of the model in both identifying and segmenting glomeruli across the dataset. Qualitative results of the segmentations produced by the trained detection model are reported in Figure 4.15.

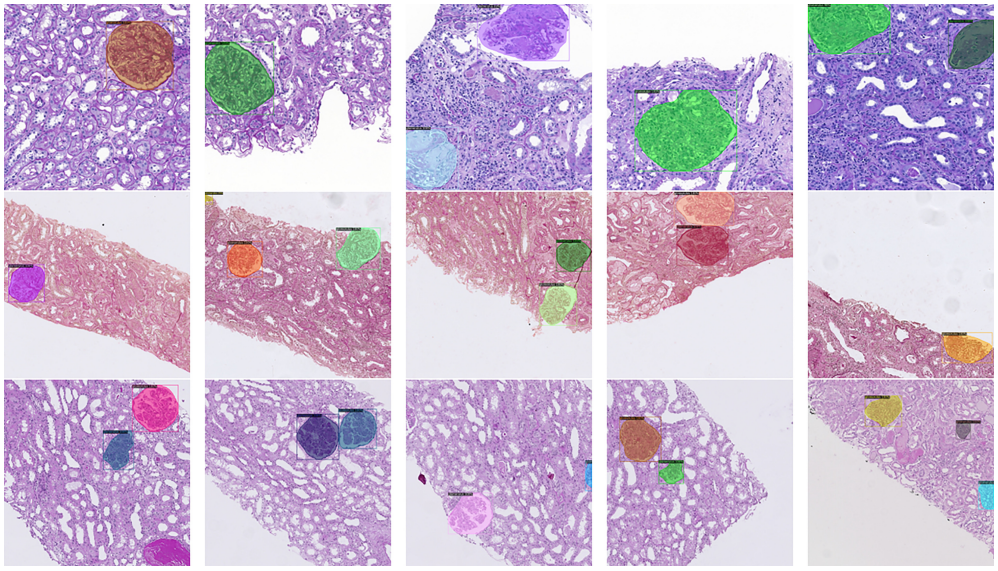


Fig. 4.15 *Qualitative results of the glomeruli segmentation process.* The top row exhibits examples from the Szeged cohort, the middle row from the Bari cohort, and the bottom row from the Cologne cohort. Notably, observe the impressive segmentation performance despite the distinct color variations in the PAS stainings across these three different institutions.

Classification The classification results at the glomerular level for CNN and ViT models were summarized in terms of ROC-AUC and PR-AUC across four types of lesions (M, E, S, and C). The best-performing models varied by lesion type:

- M lesions: EfficientNetV2-L achieved the highest performance with a mean ROC-AUC of 90.2% and a mean PR-AUC of 81.8%.
- E lesions: MobileNetV2 achieved the top ROC-AUC of 94.7%, while ResNet50 obtained the highest PR-AUC at 75.8%.
- S lesions: EfficientNetV2-M demonstrated the best results, with a ROC-AUC of 92.7% and a PR-AUC of 78.6%.
- C lesions: EfficientNetV2-L delivered the best ROC-AUC of 92.3%, and EfficientNetV2-S scored the highest PR-AUC at 54.7%.

Figure 4.16 visually compares these models' ROC-AUC and PR-AUC performance.

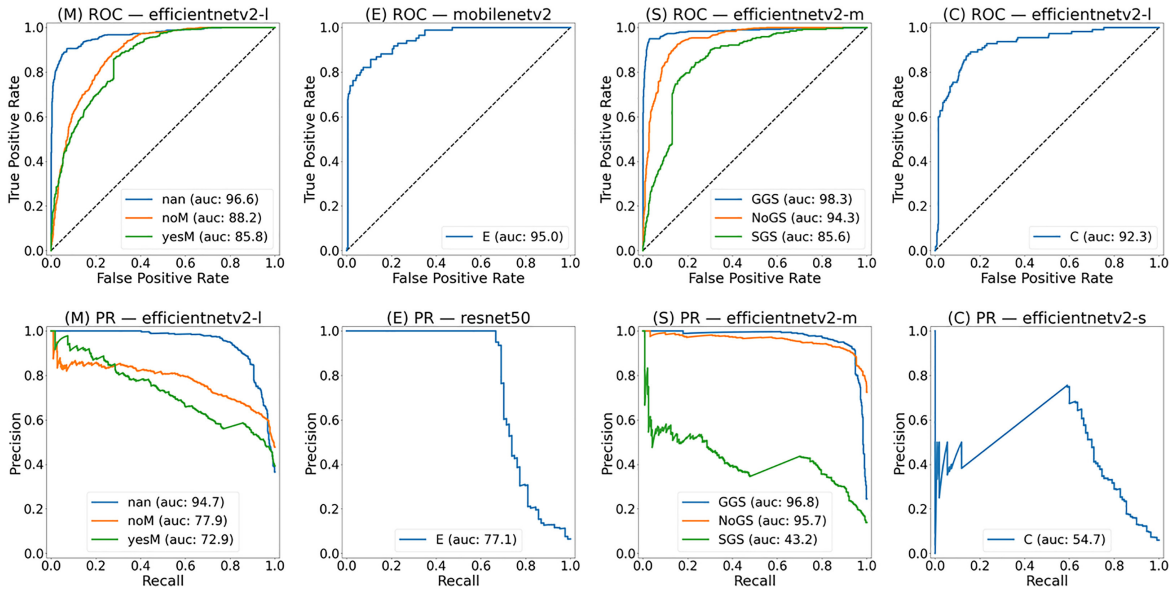


Fig. 4.16 *ROC and PR curves on the test set for the best-performing models regarding M, E, S, C lesions on glomerular level.*

The classification results are detailed in Table 4.9 for M lesions, Table 4.10 for E lesions, Table 4.11 for S lesions, and Table 4.12 for C lesions.

Table 4.9 Classification results for M lesion on the test set. The best-performing architecture is highlighted in bold typeface.

Architecture	M Label	ROC Curve		PR Curve	
		AUC	Mean AUC	AUC	Mean AUC
EfficientNetV2-L	nan_label	**96.6**	**90.2**	**94.7**	**81.8**
	noM	**88.2**		**77.9**	
	yesM	**85.8**		**72.9**	
EfficientNetV2-M	nan_label	95.0	89.0	92.5	80.5
	noM	87.1		77.3	
	yesM	84.7		71.8	
EfficientNetV2-S	nan_label	96.9	89.2	95.2	80.4
	noM	87.0		78.8	
	yesM	83.8		67.3	
DenseNet161	nan_label	97.0	89.7	95.1	81.1
	noM	87.8		78.8	
	yesM	84.2		69.5	
DenseNet121	nan_label	97.0	88.6	95.1	78.6
	noM	87.4		77.6	
	yesM	81.3		63.2	
ResNet50	nan_label	96.6	89.5	94.6	80.0
	noM	88.3		78.8	
	yesM	83.7		66.6	
ResNet34	nan_label	96.7	87.5	94.9	76.6
	noM	86.2		76.0	
	yesM	79.6		59.0	
MobileNetV2	nan_label	96.9	88.9	94.6	80.4
	noM	87.2		78.9	
	yesM	82.6		67.6	
SqueezeNet	nan_label	93.0	80.0	88.8	66.8
	noM	80.9		69.2	
	yesM	66.0		42.3	
PretrainedViTB32	nan_label	93.2	82.5	90.2	69.4
	noM	80.7		66.8	
	yesM	73.6		51.3	
PretrainedViTL32	nan_label	94.4	83.1	91.3	69.7
	noM	81.2		66.7	
	yesM	73.7		51.0	

Table 4.10 Classification results for E lesion on the test set The best-performing architecture is highlighted in bold typeface.

Architecture	**E Label**	**ROC Curve**		**PR Curve**	
		AUC	**Mean AUC**	**AUC**	**Mean AUC**
EfficientNetV2-L	noE	91.4	91.5	99.4	87.1
	yesE	91.5		74.7	
EfficientNetV2-M	noE	92.0	92.0	99.4	85.0
	yesE	92.0		70.5	
EfficientNetV2-S	noE	92.5	92.5	99.6	83.1
	yesE	92.4		66.6	
DenseNet161	noE	92.9	92.9	99.6	81.9
	yesE	92.8		64.3	
DenseNet121	noE	93.0	92.2	99.6	69.8
	yesE	91.4		40.1	
ResNet50	noE	93.6	93.4	**99.6**	**87.7**
	yesE	93.3		**75.8**	
ResNet34	noE	91.9	91.7	99.5	77.1
	yesE	91.6		54.6	
MobileNetV2	**noE**	**94.9**	**94.8**	99.7	83.2
	yesE	**94.7**		66.6	
SqueezeNet	noE	89.3	57.6	99.4	51.6
	yesE	25.9		3.9	
PretrainedViTB32	noE	82.8	82.8	98.6	69.8
	yesE	82.8		40.9	
PretrainedViTL32	noE	74.6	74.6	98.4	55.7
	yesE	74.6		13.0	

Table 4.11 Classification results for S lesion on the test set. The best-performing architecture is highlighted in bold typeface.

Architecture	S Label	ROC Curve		PR Curve	
		AUC	Mean AUC	AUC	Mean AUC
EfficientNetV2-L	GGS	98.7	92.4	97.3	76.9
	NoGS	93.8		95.5	
	SGS	84.8		38.1	
EfficientNetV2-M	GGS	**98.3**	**92.7**	**96.8**	**78.6**
	NoGS	**94.3**		**95.7**	
	SGS	**85.6**		**43.2**	
EfficientNetV2-S	GGS	98.7	90.6	97.2	75.0
	NoGS	94.0		95.6	
	SGS	79.0		32.0	
DenseNet161	GGS	99.0	90.1	97.6	75.0
	NoGS	94.0		95.7	
	SGS	77.3		31.6	
DenseNet121	GGS	99.1	88.6	98.0	74.1
	NoGS	94.4		96.4	
	SGS	72.1		27.8	
ResNet50	GGS	98.9	87.7	97.4	72.3
	NoGS	93.4		95.9	
	SGS	70.8		23.7	
ResNet34	GGS	98.6	88.0	96.8	72.2
	NoGS	93.2		95.1	
	SGS	72.4		24.7	
MobileNetV2	GGS	98.9	88.6	97.4	73.8
	NoGS	93.6		95.6	
	SGS	73.3		28.3	
SqueezeNet	GGS	97.3	79.8	91.7	66.3
	NoGS	88.1		91.1	
	SGS	54.1		16.1	
PretrainedViTB32	GGS	94.7	85.5	91.6	69.2
	NoGS	87.6		89.1	
	SGS	74.2		26.8	
PretrainedViTL32	GGS	97.4	85.7	94.4	69.2
	NoGS	88.7		89.7	
	SGS	70.9		23.4	

Table 4.12 Classification results for C lesion on the test set. The best-performing architecture is highlighted in bold typeface.

Architecture	C Label	ROC Curve		PR Curve	
		AUC	Mean AUC	AUC	Mean AUC
EfficientNetV2-L	noC	**92.0**	**92.1**	99.3	75.5
	yesC	**92.3**		51.7	
EfficientNetV2-M	noC	89.0	88.9	99.1	71.9
	yesC	88.7		44.7	
EfficientNetV2-S	noC	89.9	90.2	**99.0**	**76.9**
	yesC	90.6		**54.7**	
DenseNet161	noC	92.4	92.3	99.4	73.9
	yesC	92.2		48.4	
DenseNet121	noC	90.1	90.7	99.3	75.0
	yesC	91.3		50.7	
ResNet50	noC	89.0	88.9	99.0	71.7
	yesC	88.9		44.3	
ResNet34	noC	88.9	89.2	99.1	70.1
	yesC	89.5		41.1	
MobileNetV2	noC	90.2	90.4	99.2	71.8
	yesC	90.6		44.3	
SqueezeNet	noC	69.8	56.1	97.1	52.7
	yesC	42.4		8.4	
PretrainedViTB32	noC	52.1	52.1	95.6	51.1
	yesC	52.1		6.7	
PretrainedViTL32	noC	37.9	37.9	93.1	49.5
	yesC	37.9		5.9	

UMAP plots in Figure 4.17 depict how well the models grouped glomeruli according to the lesion types, indicating distinct feature clusters.

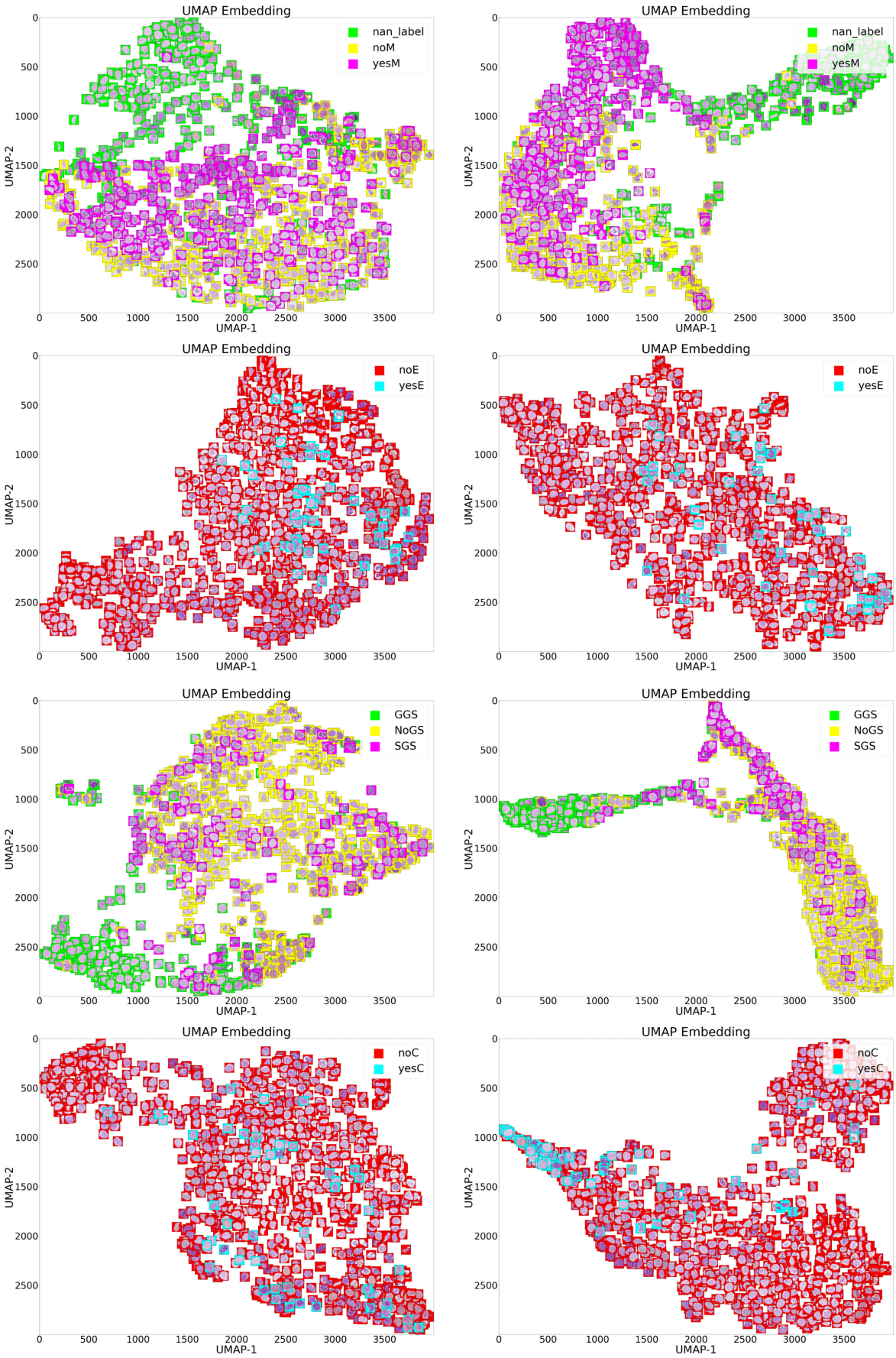


Fig. 4.17 *Embedding plots for M, E, S, C lesions classification by best-performing models*. Specific training for Oxford M, E, S, C labels improved the separation of clusters obtained from CNN features compared to the pretrained baseline.

4.2.4 Discussion

In this study, a comprehensive pipeline referred to as MESCnn was developed for a computer-aided nephropathology system, focusing on glomeruli segmentation and the classification of M, E, S, and C lesions in line with the Oxford Classification for IgA nephropathy biopsies.

The segmentation stage of the pipeline was achieved using variations of the Mask R-CNN architecture, trained on a large, diverse dataset obtained from three different institutions, covering 11 types of glomerulonephritis. The Mask R-CNN models demonstrated strong performance, surpassing comparable models from previous studies [178, 179, 182]. Various configurations of Mask R-CNN were explored, with detailed results shown in the study's tables, affirming the architecture's effectiveness in accurately identifying glomerular regions in IgAN biopsies.

Following segmentation, the pipeline's classification stage was dedicated to classifying M, E, S, and C lesions within each glomerulus and providing biopsy-level classifications. This study is one of the few that strictly adheres to the Oxford Classification using only PAS-stained sections and focuses exclusively on IgAN biopsies. Although Uchino et al. [177] have tackled similar lesions in their study, their performance fell considerably short of with respect to the ones obtained here across all scores, despite having access to a larger dataset. The models tested include various CNN architectures ViTs, with EfficientNet and ResNet consistently yielding superior results across the classification tasks. The study highlights that, while ViTs have yet to match CNN performance in glomerular classification tasks, future advancements may lead to improved performance. Additionally, the challenge in replicating human pathologist assessments of E and S lesions is noted, which may be due to their low reproducibility in clinical settings. Finally, this study also addresses the issue of stain color variation by training models on multicentric data with diverse stain color characteristics. Previous research supports this strategy as a way to mitigate the challenges of stain color variability in histopathology data analysis [183, 184]. Future studies could explore the benefits of stain color normalization techniques to further enhance the pipeline's accuracy in lesion scoring.

4.3 Shape-based Breast Lesions Classification using Digital Tomosynthesis Images

Breast cancer, the second most common cancer among women worldwide, has become a global public health issue due to its complex intrinsic etiology [185]. Early diagnosis and cancer monitoring significantly reduce death risks, improve prognosis and treatment outcomes, and lower treatment costs.

Mammography is considered the gold standard among various imaging modalities as it offers the potential for early pathology detection [186]. However, as a 2D method, it has limitations in visualizing lesions, particularly in dense breasts with a prevalent glandular component. Additionally, mammography provides a 2D projection of a 3D structure, resulting in superimposition of tissues from different planes in the radiographic image.

Other imaging techniques, such as magnetic resonance, computed tomography, and digital breast tomosynthesis (DBT), are strong alternatives when in-depth analysis of high-risk cases is required. Among these, DBT has been shown to have greater accuracy compared to 2D imaging methods [187]. By acquiring multiple thin, high-resolution images, the DBT system creates a quasi-three-dimensional representation of breast images, reducing the effects of tissue superimposition.

Additionally, DBT requires a lower radiation dose than conventional imaging techniques, while producing images with higher resolution and contrast [188]. DBT provides a more precise diagnostic tool than 2D imaging for assessing morphological features, such as the shape and margins of various breast cancer immunophenotypes, allowing it to play a critical role in molecular imaging and prognosis [189–193].

Over the last decade, DL has emerged as a promising computational approach for the automatic detection, classification and segmentation of cancerous masses thorough the analysis of diagnostic medical images, thus enabling the computer-aided diagnosis (CAD) and clinical decision support systems [194–197]. The DL methods along with the traditional image processing techniques have already been established as an effective approach to automatically analyse diagnostic images for the breast cancer diagnosis and monitoring [187, 198, 199]. Numerous studies have been dedicated for the automatic detection, segmentation and classification of the breast lesions that achieved considerably moderate to high performances [200–209].

However, the automatic classification of the breast lesions according to the shape, size and physical appearance remains a challenging task due to the varying shape that refers to different type and stage of the cancer [210] (see Figure 4.18). The breast cancer is

morphologically categorised into several varying shapes based on cancer growth pattern, named as round, oval, lobulated, irregular, and architectural distortion [211, 212].

Numerous existing studies deal with the shape based breast cancer classification [210, 213, 214], however, most of these consider the mammogram instead of the DBT that offers several advantages as discussed above.

In this work named "Shape based Breast Lesion Classification using Digital Tomosynthesis Images: the role of Explainable Artificial Intelligence" [42], a CNN-based deep learning framework was developed and validated for classifying breast lesions based on shape by analyzing the ROIs on DBT images. The shapes of cancerous masses were considered according to the Breast Imaging Reporting and Data System (BIRADS) classification from the American College of Radiology, which is widely used in clinical and digital breast tomosynthesis settings [215]. This classification includes the following three categories (see Figure 4.19):

- regular opacity (Oro), encompassing round, oval, and lobulated shapes;
- irregular opacity (Ori);
- architectural distortion shape (Ost).

The clinical importance of these three BIRADS categories lies in the ability to distinguish between regular masses and irregular masses/architectural distortions, which is crucial for early breast cancer diagnosis. It is known that *Oro* lesions are typically benign, while *Ori* and *Ost* lesions are malignant. Additionally, a '*no lesion*' category, containing images without any lesions, was included in this study (see Figure 4.19).

Furthermore, eight state-of-the-art pretrained CNN architectures were employed, and their performance was compared with and without fine-tuning (using XAI techniques). Two

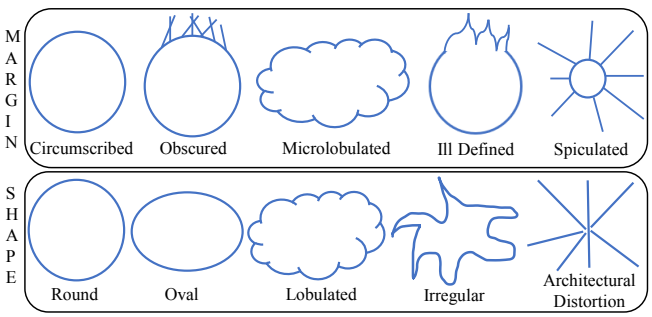


Fig. 4.18 The morphological division of the breast cancer shapes according to the growth pattern [211].

different online data augmentation routines were tested to evaluate the impact of various augmentation methods on the model's performance. The dataset used in this study was derived from the authors' previous research and included 39 breast DBT exams from 16 patients. For more details on data acquisition and composition, readers are referred to the previous study [99].

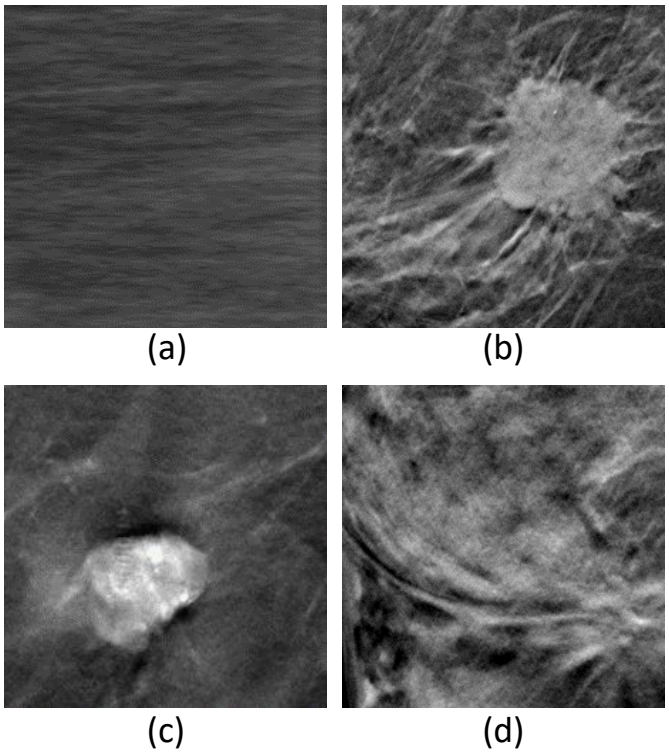


Fig. 4.19 The ready to classify RoIs on the images. (a) Example of image with no lesions (None); (b) Example of image with irregular opacity (Ori); (c) Example of image with regular opacity (Oro); and (d) Example of image with stellar opacity (Ost).

The trained DL models and related results have been further interpreted employing two different methodologies for each of the two explanation mechanisms. Grad-CAM method and LIME have been used to visually interpret the results, whereas t-SNE and UMAP techniques have been utilized to study the mathematical interpretability of the features automatically extracted by all eight CNN architectures.

4.3.1 Materials and Methods

This study uses the RoI-level images generated in a previous study [99], aimed at building a dataset of RoIs suitable for feeding into deep learning models for shape-based classification.

A total of 16 patients participated in breast tomosynthesis examinations. The average age of the participants was 49.8 years, with a standard deviation of 9.2 years. The youngest patient was 35, while the oldest was 65. Since some patients underwent multiple exams, the total number of examinations amounted to 39. Machine learning algorithms were used to generate tiles from the original images.

Figure 4.19 displays the RoIs after the segmentation phase. For the None class (i.e., the no lesion class), random images were selected from breast areas without any lesions.

A radiologist from the University of Bari Medical School, with fifteen years of experience in breast imaging, labeled the images. To verify labeling accuracy, all radiological reports were reviewed, including histological reports for detected lesions and a two-year follow-up with DBT for negative cases. The images were labeled and categorized into four classes: no lesions (None); irregular opacity (Ori); regular opacity (Oro); and stellar opacity (Ost). The None class contains 1000 images, while the Ori, Oro, and Ost classes contain 391, 654, and 480 images of lesions, respectively, resulting in a total of 2525 samples.

The CNN model architectures employed for classification task were the following: VGG, ResNet, ResNeXt, DenseNet, SqueezeNet and MobileNet-v2.

4.3.2 Experimental Pipeline

Figure 4.20 illustrates the overall flow diagram of the experimental approach. As shown, the experimental setup began by fine-tuning the selected pretrained networks with three different datasets: the original dataset and two augmented versions created using different data augmentation techniques. The CNN models were trained using 5-Fold cross-validation. Subsequently, the features extracted from the feature maps of all versions of the fine-tuned and pretrained networks were analyzed using both t-SNE and UMAP. Finally, Grad-CAM and LIME were applied to the RoI images to provide interpretability.

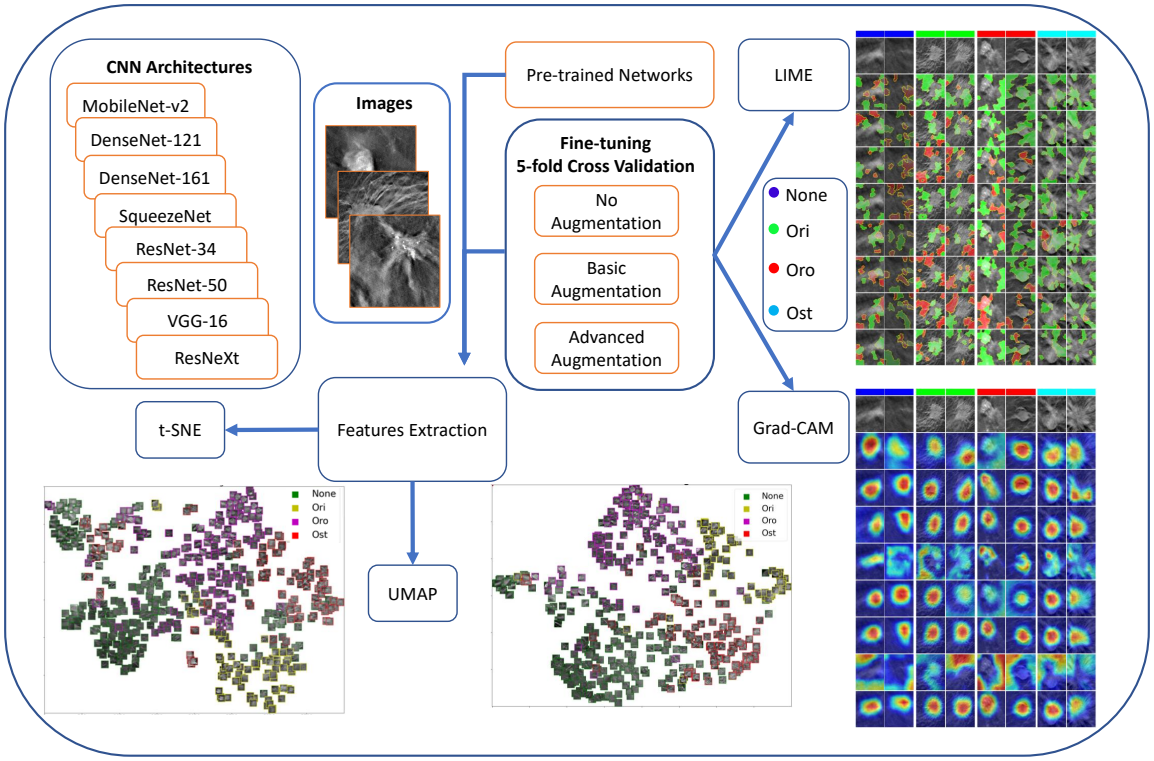


Fig. 4.20 The overall flow diagram of the experiments. The experimental setup starts by fine tuning the considered pretrained networks with three different datasets, i.e the original one and two datasets obtained with two different data augmentation procedures. Thereafter, the features extracted by the feature maps of all versions of fine-tuned and pretrained networks have been analyzed with both t-SNE and UMAP. Finally, Grad-CAM and LIME have been applied to the RoI images.

Data Augmentation. Due to the limited dataset size, two types of augmentation were considered: basic and advanced. The basic augmentation included rotation and flipping, while the advanced augmentation also incorporated color jittering. Various configurations of data augmentation were tested, as shown in Table 4.13 and described below. Using the *transforms.Compose* interface provided by PyTorch, the augmentations were applied sequentially on-the-fly, each with a probability set to 0.25.

No Aug involved no augmentation other than normalization, which was performed by rescaling the image intensity values from integer values in the range $[0, 255]$ to floating-point values in $[0, 1]$. **Basic Aug** included random rotation in multiples of 90 degrees, as well as random horizontal and vertical flips. **Adv Aug** extended the basic augmentation by adding `ColorJitter` transformations, introducing random variations in brightness, contrast, saturation, and hue. Normalization was performed in the same way as in **No Aug**.

Table 4.13 Data augmentation summary. All augmentations are done on-the-fly with 0.25 probability in the order they are presented in the table. Normalization is always performed at the end after all other augmentations. `ColorJitter` refers to the random alterations of the *brightness*, range: $[0.8, 1.2]$; *contrast*, range: $[0.8, 1.2]$; *saturation*, range: $[0.8, 1.2]$; and *hue*, range: $[-0.2, 0.2]$

Transform	**No Aug**	**Basic Aug**	**Adv Aug**
`RandomRotation90`	✗	✓	✓
`RandomRotation180`	✗	✓	✓
`RandomRotation270`	✗	✓	✓
`RandomHorizontalFlip`	✗	✓	✓
`RandomVerticalFlip`	✗	✓	✓
`ColorJitter`	✗	✗	✓
`Normalization`	✓	✓	✓

4.3.3 Results

The results of all pretrained and fine-tuned nets were analysed based on ROC AUC. The mean and standard deviation of AUC was computed for each classifier among 5-fold results. The AUC and the standard deviation were also computed for each individual class against all architectures in three augmentation configurations.

In case of **No Aug** configuration, it can be observed from the Table 4.14, DenseNet-161 is the architecture with the highest mean AUC of 96.3%. The ResNeXt and ResNet-50 networks are slightly behind with the AUC of 96.2% and 96.1%, respectively. The MobileNet-v2, ResNet-34, and VGG-16 collectively form a third cluster with an AUC of around 95%. Conversely, the SqueezeNet is the least performing model in our experimental setup, managing to achieve merely 72.4% of the AUC.

In the case of **Basic Aug** configuration, all architectures performed considerably better than the previous **No Aug** configuration. The results reveal that ResNeXt obtains the highest AUC of 97.9% beating all other architectures. The DenseNet-161 and ResNet-50 achieve similar performances with the AUC of 97.7% and 97.6%, respectively. Once again, the performance of the SqueezeNet failed to present significant outcomes, thus abiding by the **No Aug** configuration.

The second augmentation setup, called **Adv Aug**, emerged to be even better than both previously conceived **No Aug** and **Basic Aug** setups. The DenseNet-161 reached the top

AUC of 98.2%. The ResNet-50 appeared to be the second best model with a slightly less AUC of 98.0%.

Finally, as noted during the **No Aug** and **Basic Aug** configurations, the SqueezeNet was the model which offers least reliability with the largest inter-fold variability; however, the AUC achieved by such a model improved from the previous setups.

Therefore, it can be summed up that the ResNeXt and DenseNet-161 remained the top performing models and the augmentation configurations considerably improved the performance of all CNN architectures. Notably, SqueezeNet failed to produce convinceable results.

Table 4.14 The summary of the results obtained for **No Aug**, **Basic Aug**, and **Adv Aug** configurations is provided hereunder. The bold text represents the best value of the corresponding parameter among all CNN models, that is mean over all four classes

Architecture	**Area Under the Curve (AUC)**		
	No Aug (None, Ori, Oro, Ost)	**Basic Aug** (None, Ori, Oro, Ost)	**Adv Aug** (None, Ori, Oro, Ost)
MobileNet-v2	91.9 ± 1.1	92.4 ± 0.9	93.6 ± 1.2
	97.4 ± 0.4	98.0 ± 0.6	97.6 ± 0.9
	95.2 ± 1.3	95.9 ± 1.1	96.3 ± 0.9
	95.8 ± 0.7	96.6 ± 0.5	96.5 ± 0.7
	95.1	95.7	96.0
DenseNet-121	90.1 ± 1.2	93.9 ± 1.9	94.5 ± 1.3
	94.2 ± 1.4	98.5 ± 0.6	98.2 ± 0.8
	89.9 ± 1.7	95.5 ± 0.6	96.7 ± 0.8
	92.9 ± 1.8	97.1 ± 0.8	97.2 ± 1.2
	91.8	96.2	96.6
DenseNet-161	94.8 ± 0.9	95.8 ± 1.0	96.4 ± 0.5
	97.6 ± 1.4	99.1 ± 0.7	99.4 ± 0.2
	95.8 ± 1.3	97.8 ± 1.0	98.7 ± 0.7
	97.0 ± 0.9	98.2 ± 0.3	98.0 ± 0.7
	96.3	97.7	**98.2**
SqueezeNet	50.9 ± 3.0	56.6 ± 5.6	62.7 ± 8.1
	85.9 ± 3.2	84.3 ± 1.4	86.4 ± 2.9
	68.9 ± 5.6	67.6 ± 3.8	71.7 ± 7.2
	83.8 ± 2.6	86.2 ± 3.7	87.6 ± 3.1
	72.4	73.7	77.1
ResNet-34	92.0 ± 0.8	94.5 ± 1.0	95.4 ± 0.6
	96.2 ± 0.8	98.6 ± 0.5	98.9 ± 0.5
	94.7 ± 1.7	97.6 ± 0.4	97.4 ± 1.0
	96.1 ± 1.3	97.6 ± 0.7	97.7 ± 0.7
	94.8	97.1	97.3
ResNet-50	93.8 ± 1.1	95.3 ± 1.2	96.2 ± 0.6
	98.0 ± 0.5	99.4 ± 0.3	99.3 ± 0.3
	95.8 ± 0.8	97.8 ± 0.6	97.9 ± 0.7
	97.0 ± 1.0	97.8 ± 0.9	98.5 ± 0.4
	96.1	97.6	98.0
VGG-16	90.6 ± 1.7	92.5 ± 1.4	93.6 ± 1.3
	98.1 ± 0.6	98.9 ± 0.6	97.7 ± 0.7
	96.1 ± 0.7	96.7 ± 0.7	97.2 ± 0.7
	96.6 ± 0.4	97.7 ± 0.6	98.1 ± 0.6
	95.3	96.4	96.6
ResNeXt	94.1 ± 1.0	96.1 ± 0.7	95.8 ± 0.7
	97.7 ± 0.7	99.3 ± 0.2	99.0 ± 0.7
	96.0 ± 0.8	97.9 ± 0.7	98.2 ± 0.8
	97.1 ± 0.5	98.3 ± 0.3	98.2 ± 0.9
	96.2	**97.9**	97.8

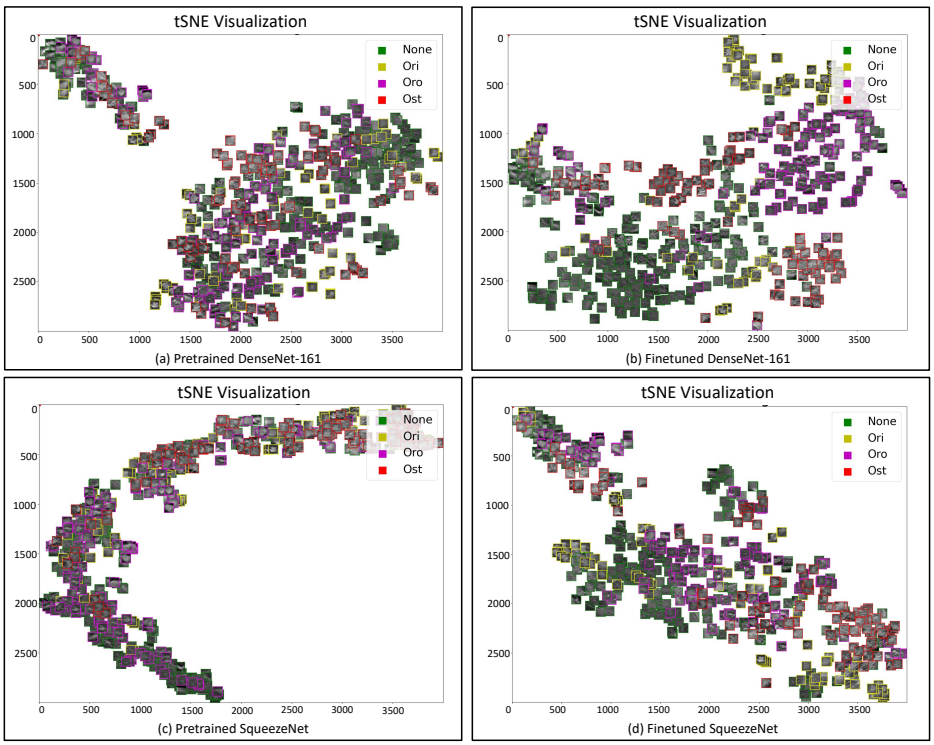
4.3.4 XAI Interpretation

UMAP and t-SNE The extracted features from both pretrained and fine-tuned networks were visualized to understand the patterns that emerged in low-dimensional spaces after applying nonlinear dimensionality reduction techniques, such as t-SNE and UMAP.

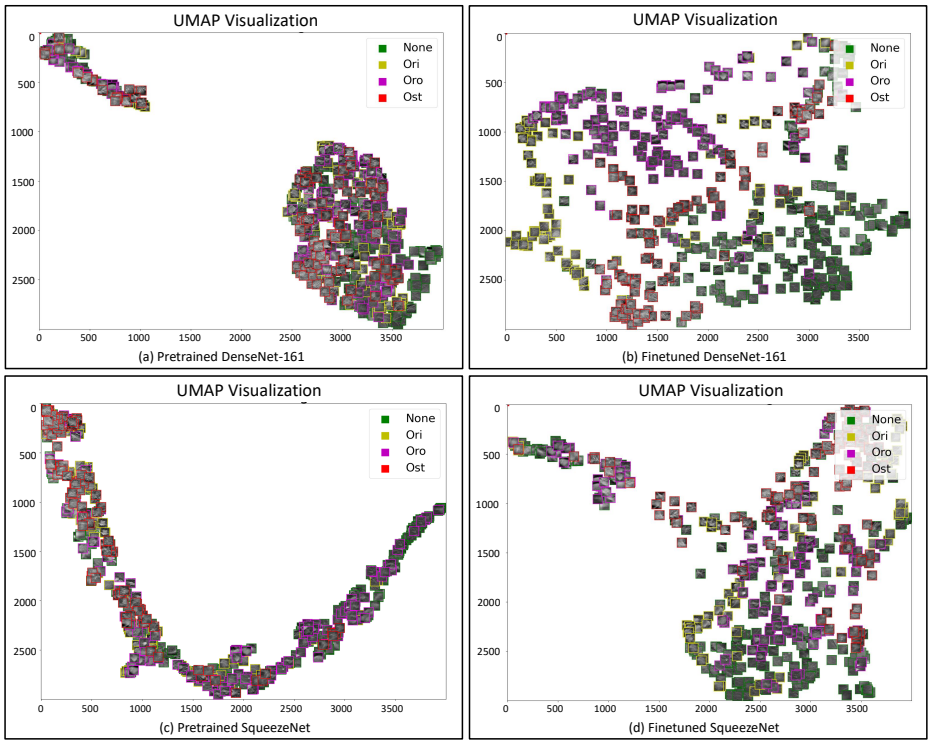
In Figure 4.21a, the t-SNE embedding plots for both pretrained and fine-tuned DenseNet-161 and SqueezeNet architectures are presented. Similarly, Figure 4.21b shows the UMAP embedding plots for both pretrained and fine-tuned DenseNet-161 and SqueezeNet models. In the pretrained versions, no clear patterns emerged in either embedding plot, indicating that the features learned from the ImageNet dataset were not necessarily well-suited for discriminative tasks in radiological image applications.

However, after 50 epochs of fine-tuning on the designated training set, the clusters became more distinctive. With fine-tuned CNN features, both UMAP and t-SNE allowed for the visualization of separate clusters corresponding to the four classes: *None*, *Ori*, *Oro*, and *Ost*.

Notably, the distances between clusters in the t-SNE visualizations cannot be directly interpreted. For instance, the proximity of clusters in Figure 4.21a does not imply similarity; rather, points closer to each other within a cluster represent more similar objects compared to those farther apart. In contrast, Figure 4.21b demonstrates UMAP's ability to represent both local and global feature structures, providing a clearer distinction between clusters and more interpretable positioning of points.



(a) The t-SNE embedding plots of the features extracted from pretrained, (a) and (c), and fine-tuned (b) and (d), DenseNet-161 and SqueezeNet, respectively, on the validation set of 1^{st} $fold$.



(b) The UMAP embedding plots of the features extracted from the pretrained (a) and (c) and fine-tuned (b) and (d) DenseNet-161 and SqueezeNet, respectively, on the validation set of 1^{st} $fold$.

Fig. 4.21 T-sne and UMAP visualization of extracted features from DenseNet-161 and SqueezeNet.

Grad-CAM. The visual explanation of all eight fine-tuned networks is shown in Figure 4.22, using Grad-CAM as the reference method. In this figure, two sample images for each class are displayed, along with the corresponding saliency maps for each network. Only images for which all networks made correct predictions were included, allowing for visualization of the relationship between highlighting of the lesion area and network performance. The saliency maps of the approximate features were generated based on the ground-truth/predicted class view.

Interestingly, the CNN architectures that struggled to correctly identify the lesion areas also showed lower performance in the classification task. For instance, SqueezeNet, which was the least effective network in terms of AUC, and VGG-16, failed to highlight the relevant lesion areas. This trade-off implies that an increasing number of parameters did not consistently result in higher AUC. In contrast, DenseNet-161, DenseNet-121, and ResNet-50 successfully highlighted the lesion areas in the images. Consequently, this XAI-based CAD system revealed the potential applicability of reliable and less reliable models for use in CAD applications.

LIME. The superpixel perturbations performed by LIME are shown in Figure 4.23. Observations regarding the performance of the LIME technique were similar to those obtained with the Grad-CAM method. The figure presents the same images compared in Figure 4.22 for Grad-CAM, allowing for a robust and clear comparison. The class used for LIME perturbations was the ground truth class, which in this case also matched the prediction made by all CNNs. Regions positively correlated with the CNN's decision are highlighted in green, while those negatively correlated are shown in red.

However, it should be noted that reasoning in terms of superpixels may result in explanations that are visually less intuitive than those provided by CAM-based methods. Comparing Figures 4.22 and 4.23, some superpixels identified as relevant to the prediction according to LIME were not highlighted in the corresponding Grad-CAM activation maps. Therefore, it is recommended to use both methods when developing an explanation for a CAD system, as complementary information from both sources can provide a broader understanding of how the model operates.

4.3.5 Discussion

This study proposed a novel visually interpretable DL framework for multiclass, shape-based classification of breast lesions in tomosynthesis images. For morphological classification,

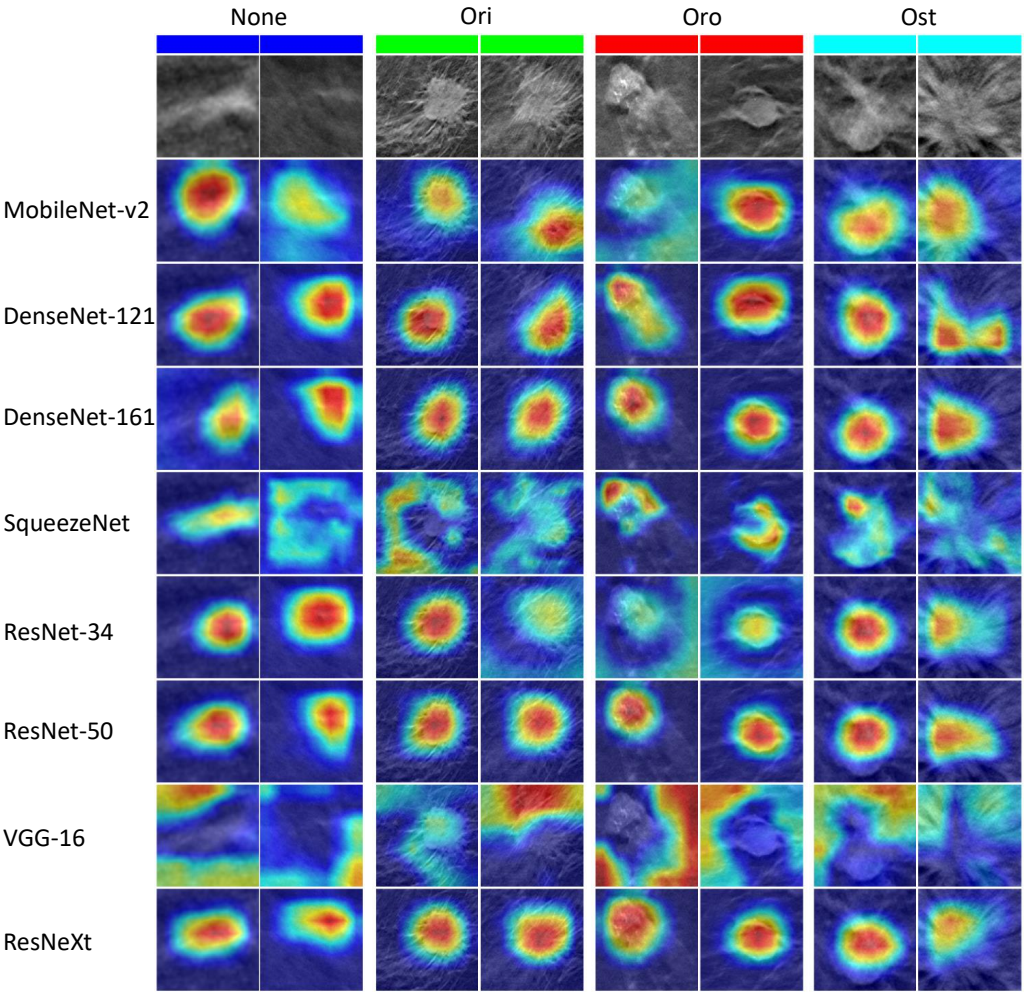


Fig. 4.22 The visualization of the Grad-CAM method with the eight different CNN architectures considered throughout the study. To illustrate the better view, two examples for each class are portrayed and the ground truth class label is provided above the set of each image. As the jet color scheme is employed for depicting saliency zones, the red color represents the higher intensity, i.e. pixels on which the network is focusing more for performing the classification, whereas, the tendency towards the blue color represents the lower intensity of focus. The header bar is used to distinguish among several classes and colored uniquely. The similar color of header for two images represents the sample chosen from same class.

eight DL models were employed on tomosynthesis breast images, and two families of XAI methods—perceptive interpretability and mathematical interpretability—were incorporated to explain the results obtained during the validation study, aiming to build trust between clinicians and AI.

The perceptive interpretability models visually explained the top features contributing to classification, while the mathematical interpretability methods revealed the clustering

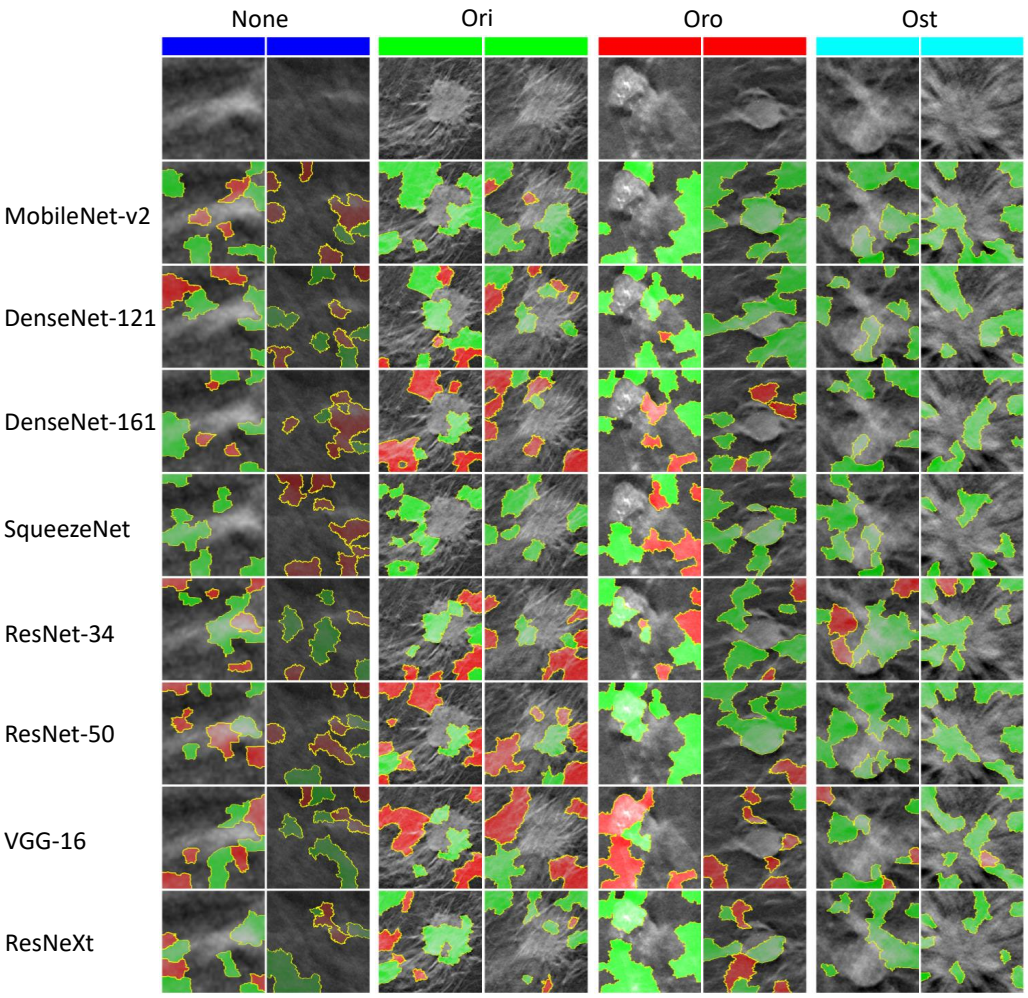


Fig. 4.23 The visualization of LIME superpixels positive and negative regions with the eight different CNN architectures considered throughout this study. To illustrate the better view, two examples for each class are portrayed and the ground truth class label is provided above the set of each image. The red color highlights the negatively contributing superpixels, whereas, the green represents otherwise. The header bar is used to distinguish among several classes and colored uniquely. The similar color of header for two images represents the sample chosen from same class.

capabilities of the DL architectures. The CAD system developed in this study was capable of identifying potential growth patterns of tumorous regions in DBT images, resulting in improved diagnostic and prognostic performance. Its successful implementation also enhanced trust in high-accuracy DL models within the clinical field.

Quantitative analysis of extensive experimental results was performed using pretrained DL models with and without data augmentation. The mean AUC values of the models improved with data augmentation. DenseNet-161 emerged as the best-performing algorithm,

consistently achieving an AUC higher than 96.0% across **No Aug**, **Basic Aug**, and **Adv Aug** setups.

In particular, DenseNet-161 demonstrated a 1.45% and 1.97% increase in mean AUC from **No Aug** to **Basic Aug** and **Adv Aug**, respectively. It outperformed SqueezeNet by 33.01%, 33.28%, and 27.10% in comparative configuration, respectively. In **Basic Aug**, ResNeXt outperformed other architectures, with a 33.56% improvement over the least performing model.

While results were comparable, the best-performing model in terms of AUC did not always perform optimally across all aspects, due to the primitive learning and weight updating mechanisms of CNN models. For example, in the **No Aug** phase, three out of four individual AUC values of ResNeXt among classes were higher than those of DenseNet-161, despite having the same mean AUC.

Nevertheless, both augmentation types and three execution setups (10, 30, and 50 epochs) showed clear improvement with augmentation. Basic augmentation improved performance compared to no augmentation, while advanced augmentation further increased AUC. High visual similarity between training and validation data, along with state-of-the-art architectures, clinical data, and RoI-level cropped images, were likely contributing factors.

Regarding mathematical explanations, feature embeddings from both t-SNE and UMAP effectively extracted meaningful relationships in low-dimensional spaces when features were representative of underlying patterns. In Figures 4.21a and 4.21b, four clusters were visible for DenseNet-161. For less accurate models, like the lightweight SqueezeNet, cluster formation varied, with UMAP yielding more compact representations. Therefore, the study recommended using mathematical XAI techniques to visualize feature relevance for the given problem.

For perceptive XAI techniques, the performance of CNN models aligned with complementary insights from Grad-CAM and LIME. Grad-CAM highlighted regions with relevant gradients for classification, while LIME identified superpixels as positively or negatively correlated with predictions. LIME allowed adjustment of the number of top contributing features, and Grad-CAM saliency maps provided intensity values. Positively correlated regions were marked in green, and negatively correlated regions in red, allowing for an intuitive understanding of significant regions.

Interestingly, CNN architectures that struggled to identify lesion areas also had lower AUC scores. The high AUC of certain models could be explained using XAI methods. For example, SqueezeNet, the least performing network in terms of AUC, and VGG-16, failed to highlight relevant lesions, as shown in Figure 4.22. In contrast, DenseNet-161,

DenseNet-121, ResNeXt, and ResNet-50, with higher AUC values, accurately highlighted lesion areas using Grad-CAM.

In summary, this study demonstrated the applicability of CNN models for DBT lesion classification at the RoI level. By leveraging transfer learning, the framework achieved efficient results with fine-tuning of parameters. The black-box nature of DL models was effectively explained, building radiologists' trust in reliable CAD systems for diagnostic tasks.

4.4 Supervised Diagnosis Standardization in Free-Text Reports

In clinical practice, routine examinations are typically accompanied by textual reports, written by experts to document findings and observations. These reports provide crucial information for diagnosis and treatment planning, capturing nuances that are often specific to the individual case. However, since these reports are often composed manually, they vary in structure and terminology, depending on the physician's style and expertise. This lack of standardization poses challenges for consistent interpretation and integration into digital health records, especially as healthcare systems increasingly rely on data-driven insights to support clinical decision-making.

Standardizing the diagnosis related to clinical cases represents a critical challenge in medicine, one that can be addressed through digital solutions designed to automatically support physicians. These solutions often need to process free-text reports and identify clinically relevant terms, such as the diagnosis itself and terminology used for diagnostic formulation. Despite the increased adoption of electronic reporting systems, many clinical institutions still rely on manual collection and processing of reports. If not adequately aligned with shared protocols and standards, the recognition of relevant terms can suffer from a lack of standardization.

Nowadays, the availability of standard vocabularies for clinically relevant terms, including diagnoses, is a standing reality. The challenge of automated systems for diagnosis standardization is taking full benefit from such a shared knowledge, to reach the highest possible level of standardization in assigning a diagnosis.

In this section ARGO 2.0 is presented ("ARGO 2.0: a Hybrid NLP/ML Framework for Diagnosis Standardization" [43]), a framework aimed at diagnosis standardization based on a hybrid approach, in which both NLP standard methods and MLP techniques cooperate to the standardization. Recent research has focused heavily on NLP-based automated digitalization of medical reports [216–221] and the prediction of diagnoses using ML models [222–225]. However, many of these solutions are highly domain-specific and fail to integrate the dual goals of automatically digitalizing reports and inferring diagnoses from structured digital datasets. ARGO 2.0 addresses both challenges in a unified approach, summarized as follow:

- it includes a component for automating the digitalization of medical reports, integrated within a comprehensive tool known as ARGO Core. This tool leverages Optical Character Recognition (OCR) and NLP to transform free-text reports into digital content, which is then used to populate a relational database modeled in RedCap [226];

this database serves as a data source for training a machine learning model capable of learning diagnoses for incoming patients.

- it was designed to be flexible w.r.t. the template used in the reports and agnostic to the medical field. The system was evaluated in the domain of hemo lympho-pathology, processing 502 heterogeneous textual reports from various institutions.

Finally, preliminary results with a deep learning approach for the diagnosis NER were evaluated, proposing a migration from the hybrid NLP/ML architecture to a deep learning one.

4.4.1 Materials and Methods

Dataset A set of 502 reports including diagnoses of lymphomas from 2014 to 2021 from 9 Italian centers was collected from 9 italians centers and organized as follows:

- An internal series of n. 353 reports collected from the Research Institution Istituto Tumori 'Giovanni Paolo II' from Bari; this was used as training set and validation set.
- An external series including n. 149 reports coming from 8 collaborative sites; this was used as test set.

The reports collected were stored in the data model, supporting 9 heterogeneous templates, adopted in reports coming from different Italian institutions:

1. Hematology and Cell Therapy Unit, I.R.C.C.S. Istituto Tumori 'Giovanni Paolo II', Bari.
2. Division of Hematology 1, AOU "Città della Salute e della Scienza di Torino", Turin.
3. Hematology, AUSL/IRCCS, Reggio Emilia.
4. Division of Hematology, Azienda Ospedaliero-Universitaria Maggiore della Carità di Novara, Novara.
5. Unit of Hematology, Azienda Ospedaliero-Universitaria Policlinico Umberto I, Rome.
6. Department of Medicine, Section of Hematology, University of Verona, Verona.
7. Department of laboratory diagnostics, "ASST Degli Spedali Civili di Brescia", Brescia.

8. Division of Diagnostic Hematopathology, IRCCS European Institute of Oncology, Milan.
9. Histologic Pathology and Molecular diagnostic, Azienda Ospedaliero-Universitaria Careggi, Firenze.

ML Models The NER task with diagnosis standardization was approached as a classification task. The ML models supporting the ARGO Core were chosen among the classical ones in ML, in particular a RF, XGB and MLP classifiers. The best model was chosen by

The markers related to the internal series have been manually extracted and reviewed, to reduce the error related to wrong data input during the training phase. Instead, the markers in the external series were extracted from the input reports, through OCR and NLP techniques. Several factors, including the slight imprecisions of OCR algorithm during the text acquisition and the non-optimal image quality, negatively affect the performance of the classification model on the external series.

Table 4.15 shows the percentual occurrence of most frequent biomarkers in the pathology reports for definition of every aforementioned diagnosis, for both internal and external series.

Collected reports refer to five prevalent diagnosis classes of lymphoma: diffuse large b-cell (DLBCL), follicular (FCL), mantle cell (MCL), Hodgkin (HL), and T-cell. Label distribution for test set is reported in Table 4.16

4.4.2 Experimental Pipeline

The architecture of ARGO 2.0 was made up by four components:

Data Model was the underlying source of information, designed for the integration of heterogeneous sources of data ingestion. It automatically collected information coming from medical reports by modelling them as personalized electronic Case Report Forms (eCRFs) [20]. Adopted eCRFs have been designed to match the knowledge model shared in RedCap, set on the clinically relevant variables identified by recognized investigators. In the addressed use case, eCRFs were designed according to the requirements provided and approved by the College of American Pathologists [21], [22]. Included features were: Report ID, Report Data, Sample Type, Markers, Type of Diagnosis.

ARGO core was the tool responsible for converting free-text reports into a digital format for storage in the data model. It exhibited three key features: (i) independence from the

Table 4.15 Percentual occurrence of most frequent biomarkers retrieved from both internal and external series of pathology reports. Abbreviations: NA, not available. : BCL2 assessed by in situ hybridization.

Most frequent biomarkers	**Occurrence of Biomarker, %**	
	Internal series	**External series**
CD20	88.95	91.87
CD3	84.99	56.10
CD5	84.42	42.28
BCL6	75.92	60.16
BCL2	71.10	67.48
CD10	69.69	68.29
CD30	60.62	21.14
Ki-67	53.54	65.04
IRF4/MUM1	50.42	30.08
CD23	42.49	20.33
CD15	33.99	19.51
CD45/LCA	28.90	NA
PAX5	27.76	20.33
Cyclin-D1	23.51	35.77
CD79alpha	22.38	5.69
EBV/LMP1	16.43	4.07
IgM	13.31	NA
IgD	12.75	NA
MYC	10.76	33.33
BCL2*	10.48	NA
EMA	8.22	0.81

Table 4.16 Label distribution for external test set: diffuse large b-cell (DLBCL), follicular (FCL), mantle cell (MCL), Hodgkin (HL), and T-cell.

Diagnosis	**N.**	**Frequency (%)**
DLBCL	59	39.60
FCL	32	21.48
HL	30	20.12
MCL	26	17.45
T-CELL	2	1.35

report template, (ii) independence from the medical field, and (iii) the application of a NER approach to selecting the most appropriate diagnosis. In the addressed use case, ARGO core analyzed the incoming reports, retrieving all relevant sub-sections from the acquired template. Using an OCR algorithm to retrieve the full report text, ARGO

core chunked it into sub-sections using a basic set of regular expression. The section report is then processed by a DL-based translation service API (DeepL [1], detecting the report source langue and translating it into english. For the diagnosis sub-section, it extracted each marker mentioned in the report and used them to query the SEER database for potential diagnoses associated with each marker value. ARGO core then computed a degree of correspondence, termed the Matching Rate (MR), as defined by the following formula:

$$MR = \frac{counter}{DiagnosisLength} \tag{4.1}$$

where the *counter* represented the number of occurrences of each label in the diagnosis section, normalized by the length of that section (*DiagnosisLength*). Finally, the retrieved sub-sections, marker list, and diagnosis list (with corresponding MR) were sent to the DAM. The retrieval of sections and the NER for markers were based on regular expressions.

ML Model was the machine learning model trained to predict diagnoses based on the available data model. In the addressed use case, the training phase involved the implementation of several model categories, including Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, and Artificial Neural Network. Each model category was trained and fine-tuned using the Grid Search approach. Each combination of parameters generated a model with specific performance characteristics; only the Random Forest model was retained, as it outperformed the others. The remaining models were discarded due to lower performance. The ML model received a list of markers as input and returned a predicted diagnosis, which was subsequently sent to the DAM.

Diagnosis Assignment Manager DAM module was responsible for managing the suggested diagnosis to optimize assignment accuracy, following a strategy tailored to the specific medical field. No decision was needed when the diagnosis retrieved by both ARGO core and the ML Model coincided, as DAM validated this single suggestion. Similarly, if ARGO core was unable to return a result, DAM suggested the only available diagnosis— the one provided by the ML Model. However, when both diagnoses were available but conflicting, an optimization strategy was required. In this case, ARGO core computed the Matching Rate (MR) for the diagnosis predicted by the ML Model. The rationale behind this strategy was to trust the ML Model's prediction

[1] https://www.deepl.com/it/translator

if it was supported by ARGO core with a sufficient MR. A threshold value for MR, which was domain-dependent and heuristically set to 0.67 in the addressed use case, was employed to measure the adequacy of this support. Finally, the chosen diagnosis and the report sub-sections were sent to the Data Model, that used them to create the corresponding eCRF.

As first step, the data were pre-processed as follows: drop of non-relevant features for the prediction (i.e., Report ID, Data and those containing all missing values), drop one of the same features expressed in both numerical and categorical way, categorical feature encoding, Z-Score Normalization of numerical feature. Figure 4.24 depicted the full framework developed.

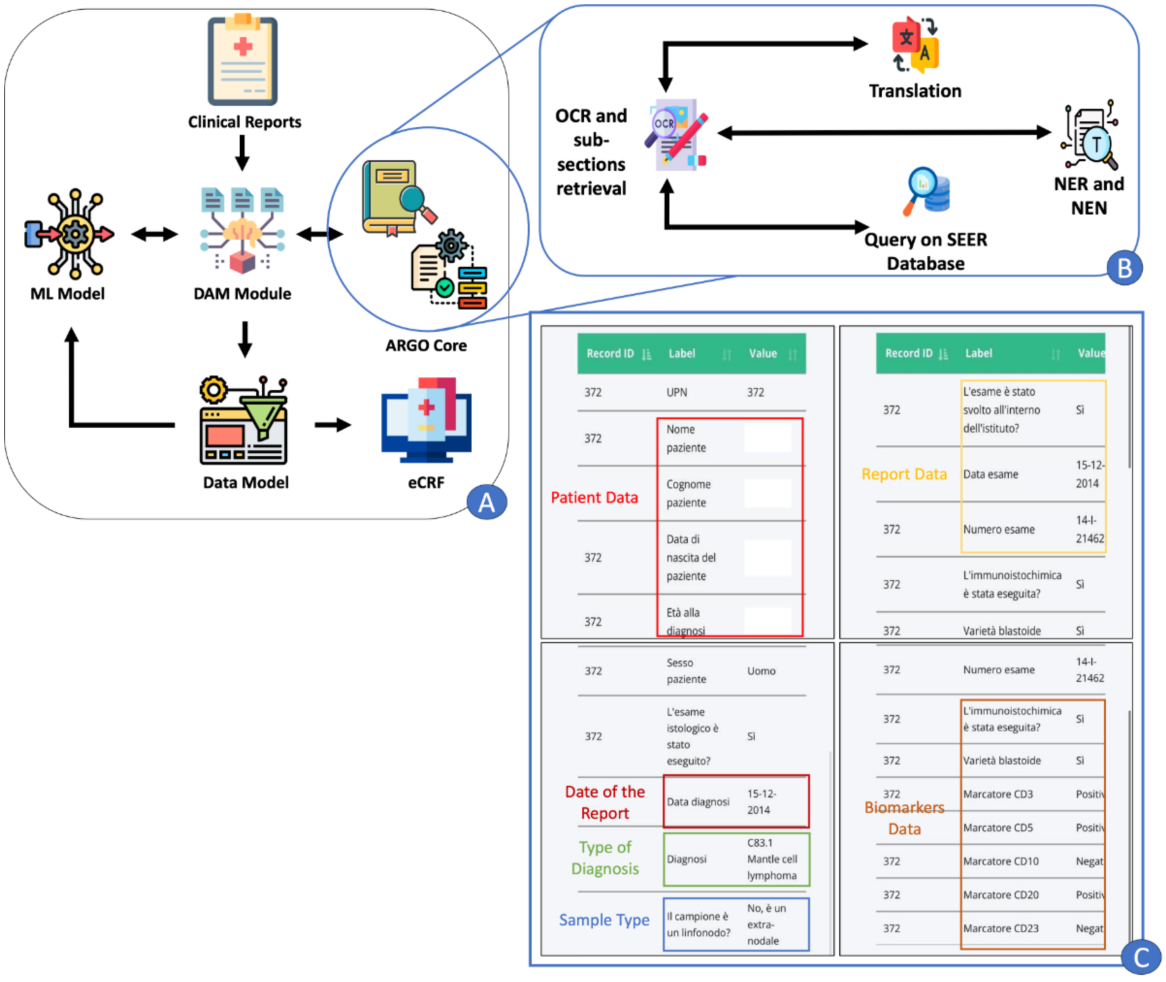


Fig. 4.24 ARGO 2.0 Architecture: (A), ARGO core services (B) and an example of eCRF generated by ARGO 2.0 (C). DAM – Diagnosis Assignment Manager, ML Model – Machine Learning Model, OCR – Optical Character Recognition, SEER - Surveillance, Epidemiology, and End Results, NER – Named Entity Recognition, NEN – Named Entity Normalization.

4.4.3 Results

The results achieved were divided into three categories: ARGO Core only, ML only and ARGO 2.0 (the combined approach). For this task, the framework performance was assessed in terms of Accuracy (A), Precision(P), Recall (R) and average F-Score.

ARGO Core Results According to the internal series, ARGO core achieved 86.4% of A, average R, and average F-Score, and 100.0% of average P in the process of assigning a diagnosis, showing higher performance than the one presented in its previous version. Table 4.17 reported the performance achieved on test set.

Table 4.17 Performance achieved with ARGO Core. Abbreviations: DLBCL, diffuse large b cell lymphoma; FCL, follicular lymphoma; HL, Hodgkin lymphoma; MCL, mantle cell lymphoma

	Precision %	***Recall%***	***F-Score%***
DLBCL	100	87.27	93.20
FCL	93.93	96.87	95.38
HL	100	89.65	95.45
MCL	96.15	96.15	96.10
Average Metrics	97.52	92.48	94.82
Accuracy %	91.56		
Not Available daignosis	9		

ML Results As mentioned before, the best overall accuracy was obtained by RF on the validation set, resulting as the best model for making predictions in most diagnosis. Then, the performance of the such a model was evaluated on the test set.

Table 4.18 Performance achieved with ML Model. Abbreviations: DLBCL, diffuse large b cell lymphoma; FCL, follicular lymphoma; HL, Hodgkin lymphoma; MCL, mantle cell lymphoma

	Precision %	***Recall%***	***F-Score%***
DLBCL	90.47	69.09	78.35
FCL	75	75	75
HL	57.14	96.55	71.79
MCL	89.47	65.38	75.55
Average Metrics	78.02	76.50	75.17
Accuracy %	75.35		

ARGO 2.0 Results As a final step the performance of ARGO 2.0 was evaluated, which embedded the DAM module to improve the reliability of diagnosis assignment by combining ARGO core with the RF model. Table presented performance by class of ARGO 2.0 with reference to the test set.

Table 4.19 Performance achieved with ARGO 2.0.Abbreviations: DLBCL, diffuse large b cell lymphoma; FCL, follicular lymphoma; HL, Hodgkin lymphoma; MCL, mantle cell lymphoma

	Precision %	***Recall%***	***F-Score%***
DLBCL	100	89.09	94.23
FCL	94.11	100	96.96
HL	85.29	100	92.06
MCL	100	96.15	98.03
Average Metrics	94.85	96.31	95.32
Accuracy %	95.07		

In this study, ARGO core was enhanced with a DAM that incorporated a RF model, resulting in the development of ARGO 2.0. The upgraded framework demonstrated promising results when applied to hemo lympho-pathology. ARGO 2.0 outperformed both the standalone ARGO core and the ML model across most metrics, with the exception of precision, which was lower than that of ARGO core. The cause of this reduced precision was explored by evaluating the model's performance by diagnosis class. It became evident that precision

was particularly low for the HL class. While ARGO core achieved 100% precision for HL diagnosis (as indicated in Table I), the ML model performed poorly, with a precision of 57.14% (as shown in Table 4.18). Consequently, the overall precision of ARGO 2.0 decreased due to the DAM module prioritizing the ML model's prediction over ARGO core's assignment. This misclassification occurred when ARGO core produced a list of possible diagnoses and the ML model incorrectly selected a diagnosis from that list with a MR exceeding the threshold. In such cases, ARGO 2.0 would favor the ML model's selection, even though ARGO core alone would have discarded the wrong diagnosis.

A potential technical solution to this issue lies in fine-tuning the MR threshold, ideally using machine learning techniques to learn the optimal threshold dynamically, rather than relying on empirical setting. This will be addressed in future work. Nonetheless, the fundamental issue remains the low precision of the ML model in classifying the HL class, which had a direct impact on ARGO 2.0's overall performance. Such analysis revealed that descriptions of HL cases lacked standardization. HL is a common diagnosis that physicians typically further specify, which may lead to variability in descriptions. To address this, future work will involve training the ML model on a larger number of classes, including HL subclasses, to improve prediction specificity. A direct comparative analysis of ARGO 2.0 results with related tools is challenging, as these tools do not perform exactly the same tasks. However, they tackle similar objectives, such as NER, information extraction (IE), and pathology classification. The results achieved are in line with similar ones in the same field [220, 221].

4.4.4 Alternative approach with transformer model

An alternative approach for performing NER of diagnosis in free-text report is given by the use DL models, in particular with transformers.

In such a scenario, the adoption of a DL model may introduce several vantages: (i) total independence from regular expressions, allowing the framework to be generalized to every pathology with only fine-tuning process; (ii) possibility of performing NER on other targets (markers, diagnosis, examination data, and so on) at the same time with a single model; (iii) replace the handcraft decisional heuristic of the DAM module with a straightforward approach, since the decision depends only from the model; (iv) framework architecture simplification, replacing the DAM module and ML module with DL model; (v) explainability od the decisions retrieved directly from the attention mechanism.

For the use case aforementioned, a BERT model was fine-tuned and tested against ARGO 2.0 DAM. Specifically, BERT was employed in its biomedical version, Bio-Bert.

BERT Introduced by Delvin et al. [227], the Bidirectional Encoder Representations from Transformers (BERT) model is a pre-trained language model developed to process and understand large amounts of text data. BERT's unique Transformer architecture allows it to encode bidirectional word contexts, meaning it considers both the words preceding and following a given word, considering also the intra-word patterns. This bidirectional context helps BERT produce high-quality representations of words in a text. In NER, BERT can be effectively used as a pre-trained model for extracting entities from text. To apply BERT to the specific task of NER, a process called fine-tuning is performed. Fine-tuning involves training a classification model using a labeled dataset containing examples of text with annotated entities. The BERT model encodes the context of each word in the text and generates a sequence of word representations. These representations are then passed to a classifier, which determines for each word whether it belongs to an entity, and if so, classifies the type of entity (e.g., person, place, organization). During the fine-tuning process, the weights of the BERT model are adjusted to better suit the NER dataset, enabling it to learn how to correctly identify and classify entities in the text. BERT's ability to consider both the preceding and following contexts of words gives it an advantage over other NER models based on older techniques, such as RNNs. This results in improved entity recognition accuracy.

BERT uses two main training paradigms: pre-training and fine-tuning. During the pre-training phase, BERT is trained on a large dataset to learn general language patterns and representations. This phase is typically conducted in an unsupervised manner, where the model is exposed to unlabeled data, allowing it to learn from large corpora like English Wikipedia and BooksCorpus. In this stage, BERT learns to capture linguistic structures such as grammar, semantics, and syntax without needing labeled examples, involving two two techniques:

- Masked Language Modeling : it predicts hidden (masked) words in a sentence based on their surrounding context. This improves the model's ability to learn semantic relationships between words.
- Next Sentence Prediction: it predicts whether a given sentence follows a previous one, improving text comprehension and context prediction.

In the fine-tuning phase, BERT is then adapted to perform specific tasks, such as classification, text generation, language translation, question answering, and more. This involves training the model on task-specific datasets in a supervised manner, where it uses labeled data to learn how to predict desired outputs for a given task.

BERT leverages the Transformer architecture to process text sequences and generate contextualized language representations. The Transformer architecture consists of encoding and decoding blocks, but BERT only uses the encoding block. This encoding block is composed of multiple stacked layers, each tasked with encoding the input text sequence. These layers encode both semantic and syntactic information into dense vector representations (embeddings) that are highly useful for a wide range of downstream NLP tasks. Moreover, BERT uses tokenization to break text into smaller units, called tokens, for processing. Specifically, it employs *WordPiece* tokenization, which splits words into subtokens based on frequency in the training corpus, allowing BERT to handle words more efficiently. During this process, BERT adds two special tokens: [CLS] at the beginning of a sentence, representing the entire sentence, and [SEP] to separate different sentences. The type of embedding used are the following:

- Token Embedding – represents each token in the sentence.
- Segment Embedding – separates multiple sentences within the input.
- Position Embedding – captures the position of tokens in a sentence, helping BERT understand the structure.

The output of BERT model in NER task is a representation following the BIO tagging scheme. Given a sentence, each entity (word) is labeled as follows: **B**-Beginning of the entity, **I**-Inside, tokens that are part of the same entity but not the first one and **O**-Outside, tokens that do not belong to any entity. An overview of BERT architecture with BIO tagging scheme is depicted in Figure 4.25

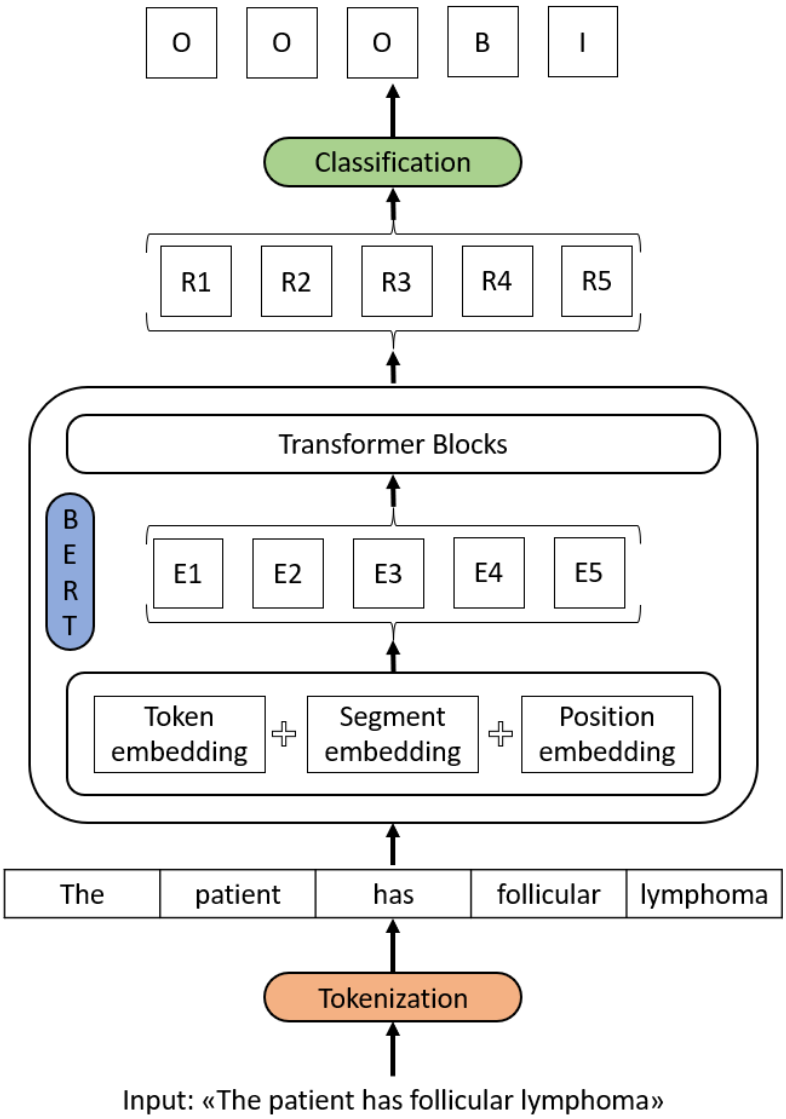


Fig. 4.25 BERT Architecture with tagging scheme. B-Beginning of the entity, I-Inside, O-Outside.

In this work Bio-BERT model [228], a pre-trained BERT on medical corpora, was fine-tuned and used for NER on the external test set. Such a choice was justified by its superior performance compared with other biomedical language representation models, such as PubMedBERT [229] and SciBERT [230], on a wide range of biomedical text processing tasks.

4.4.5 Preliminary Results

Bio-BERT was fine tuned on CancerMine public dataset [231], chosen as a source of information. , CancerMine is a high-quality text-extracted database that catalogs more than 856 genes as drivers, 2,421 as oncogenes, and 2,037 as tumor suppressors in 426 cancer. This resource was created through analysis of a wide range of scientific publications on oncology and identification of important genes involved in cancer pathogenesis and is an automatically updated dataset.

It was decided, therefore, to process the datasets by extracting 25,000 cases of lymphomas out of 100,000 total cases. Then the model was tested on the external test set. Three experiments were conducted by setting a number of epoch of 10, 20, and 40 respectively and setting a maximum length of sentence processed of 256, 512 and 512.

The preliminary performances achieved on the test are shown in Table 4.20.

Table 4.20 Bio-BERT NER metrics on test set.

Experiment Configuration	**Precision**	**Recall**	**F1-score**
CancerMine fine-tuning with 10 epochs	66.9%	66.7%	66.8%
CancerMine fine-tuning with 20 epochs	81.3%	80.6%	80.9%
CancerMine fine-tuning with 40 epochs	75.1%	71.3%	73.2%

4.4.6 Discussion

Bio-BERT fine-tuned on CancerMine achieved promising performance on the external test set. Notably the trends of performance in Table 4.20 shows an overfitting behavior after 20 epochs, with the second experiment resulting as the most promising one. Comparing the model with ARGO 2.0, although the acceptable performance, the metrics achieved still remain lower w.r.t. Argo Core and ARGO 2.0 but higher than ML module. On the other hand, the following consideration should be taken into account:

- The results achieved by Bio-BERT do not consider partial matching, i.e. if a diagnosis is partially retrieved with a missing word, it is classified as false negative; this leads to a performance decrease. Such issue can be mitigate by adding another layer that matches the diagnosis retrieved with a list of possible diagnosis, returning the most similar one.

- ARGO 2.0 performance metrics were considered after the standardization. The support of ARGO 2.0 DAM module inevitably increase the performance, since partial

diagnosis sequences are converted into a standardization format and the framework can misclassify an input sequence only if the partial diagnosis sequence is fully wrong.

- Bio-BERT was fine-tuned using an external public training set while ARGO 2.0 was tuned on a private dataset, specific for the target lymphoma cases and supervised by expert operators; the use of public dataset that includes other type of lymphomas can introduce noise during the training phase. Moreover, transformers demonstrate their full potential when trained on very large datasets, as their architecture excels at capturing complex patterns and relationships in vast amounts of data. In light of this, the model can benefit from the adoption of other public datasets [232] during the fine-tuning phase.

- Bio-BERT was fine-tuned only considering the training epochs and the maximum sequence processed by the model; a more elaborate fine-tuning strategies can further increase the performance achieved.

- Bio-BERT can perform the NER task, even if the report image is cut (see Table 4.17, Not Available diagnosis), assuming that the diagnosis sequence is contained in the cut image. In case of Not Available diagnosis, the ML model still produces an output but not reliable without the support of the heuristic incorporate into DAM.

Considering the aforementioned considerations, a transformer-based approach indeed showed a promising performance and proved to be a valuable asset for accomplishing NER on free-text report diagnosis. Noteworthy is the absence of a heuristic strategy and a straightforward approach in performing NER w.r.t. ARGO 2.0 architecture. Indeed, more experiments and more model training strategies are required for achieving higher performance and, possibly, for outperforming ARGO 2.0.

4.5 Summary of Findings

Enhancing Survival Analysis Model Selection Through XAI(t) in Healthcare.

In this study, a comprehensive pipeline for training and comparing survival ML and DL models was developed, incorporating SurvSHAP to enhance the understanding of model predictions over time. The work addressed two key challenges in survival analysis: selecting the most reliable model among similarly performing ones and analyzing variable importance based on observation time. The study focused on OSA, emphasizing the impact of comorbidities, but the proposed approach is applicable

to other pathologies. Performance and reliability of survival models were evaluated considering data quality, feature relevance, and model architecture. Complex models demonstrated their capability to identify features that accurately describe the examined pathology. After data preparation, four ML models and five DL models were trained, and the best-performing models were selected using test-set metrics. SurvSHAP was applied to classify explanations into dataset- and model-level categories, illustrating differences in feature importance between ML and DL models over time. Results highlighted how models identified key variables and their time-dependent impact on predictions. While the CPH model showed slightly better performance, the LH model emerged as more reliable and clinically valuable for OSA patient follow-up. The time-dependent explainability provided by XAI(t) proved critical for understanding model behavior and feature contributions to survival predictions. This dynamic evaluation may supports clinical decision-making by helping physicians assign varying weights to patient-related features as their importance evolves, ultimately improving follow-up strategies in rehabilitation.

Supervised deep learning approach for the histopathological Oxford Classification of glomeruli with IgA nephropathy.

The Oxford Classification for IgAN represents a successful example of an evidence-based nephropathology system. This study aimed to develop a deep learning model for the automatic analysis of large biopsy cohorts, ensuring perfect reproducibility and minimal human effort. The results demonstrated that features extracted from modern deep networks pretrained on ImageNet effectively replicate expert labels for the glomerular components of the Oxford Classification. Expanding the availability of larger datasets for glomerular lesion classification is expected to further enhance the model's performance.

Shape-based Breast Lesion Classification using Digital Tomosynthesis Images: the role of Explainable Artificial Intelligence.

Breast cancer is a leading cause of mortality in women, and timely diagnosis can significantly improve outcomes by limiting cancer progression. DL models have shown success in detecting and classifying breast lesions using medical imaging, but their black-box nature hinders trust among clinicians. XAI techniques address this issue by uncovering model mechanisms, fostering confidence in DL-based CAD systems. This study proposed an explainable DL framework for multiclass shape classification of tomosynthesis breast lesions using eight pretrained CNN models. With

data augmentation, the best model achieved a mean AUC of 98.2%, compared to 96.3% without augmentation. XAI methods, including Grad-CAM, LIME, t-SNE, and UMAP, were used to explain classification results visually and mathematically, aligning model performance with interpretability and building trust in clinical applications. Future efforts will focus on quantifying individual feature contributions, testing on external datasets, and incorporating novel DL models to enhance robustness and generalizability.

Supervised Diagnosis Standardization in Free-Text Reports.

The study addresses the challenge of standardizing diagnosis assignment in hemo lympho-pathology, a task complicated by the heterogeneity of clinical descriptions and the lack of uniformity in pathology reports. ARGO 2.0, an enhanced framework integrating a DAM with a RF model, was developed to tackle this issue. The framework demonstrated improved performance over its individual components, ARGO core and the standalone ML model, across most metrics. However, lower precision in the classification of Hodgkin's Lymphoma highlighted limitations in the ML model and the need for more refined tuning of parameters, such as threshold learning. Through this study, the integration of NLP techniques and ML methods in ARGO 2.0 provided novel insights and underscored the value of a cooperative approach. Notably, the test of BIO-BERT transformer model yielded promising results, potential enhancing the capacity to process and interpret unstructured textual data. Future work aims to refine the framework further by including HL subclasses in predictions and extending its application to other medical domains, demonstrating its potential as a versatile tool for diagnosis standardization and clinical decision support.

Chapter 5

Conclusions

This thesis introduces several Big Data analytics pipelines designed as clinical Decision Support Systems, assisting physicians and operators in biomedical field. After the introduction of all methods included in this work in Chapter 2, including Machine Learning, Deep Learning and Explainable Artifical Intelligence algorithms, the contribution chapters open with the two multimodal pipelines, Chapter 3, and the unimodal pipelines in Chapter 4.

The use cases covered a wide range of applications, bringing contributions in the fields of Radiomics, Digital Pathology, Clinical and biomedical Natural Language Processing.

The two multimodal pipelines presented in Chapter 3 focus on the study of Pancreatic Ductal Adenocarcinoma from different perspectives: (i) according to the target, the former deals with Overall Survival and Recurrence prediction while the latter deals with genetic mutation prediction; (ii) according to the methods, the former exploits multi-omcs (radiomics, clinical and genomics) data with survival analysis methods while the latter exploits pathomic and transcriptomic data with Deep Learning and Machine Learning methods.

Delving into the first pipeline, the Cox Proportional Hazard model achieved good performance metrics with a C-Index of 75% of Overall Survival, while Survival SVM achieved 68% of C-Index score, resulting the best models. The use of time-dependent XAI (survSHAP) allows for retrieving the feature importance over time, highlighting how the combination of radiomics, clinical and genomics features contributes to the model predictions.

The second pipeline exploits DL and ML models for prediction the genomic mutations of the four principal genes (*KRAS*, *TP53*, *SMAD4*, and *CDKN2A*) that play a pivotal role in PDAC. Concerning the image processing side, CLAM and foundation models were employed for feature extraction and Whole Slide Images classification, outperforming the state-of-the-art, in terms of Area Under ROC Curve and Area Under PR Curve. *KRAS* and *TP53* revealed to be easy prediction targets w.r.t. the other two genes. Shifting to the transcriptomics side,

ML models achieved higher performances w.r.t. to the imaging models. The fusion layer obtained by combining the two type of predictions, achieved the same performance level as transcriptomic models on *KRAS* and *TP53* and a slight higher performance on the other two genes. The combined adoption of attention-maps and SHAP shows how it is possible to investigate a prediction from both imaging and trascriptomics perspective.

Chapter 4 illustrates the developed unimodal pipelines. The first section shows how time-dependent XAI can be exploited in survival model selection process. The trained models belong to both ML and DL survival model families and the dataset involved is related to Obstructive Sleep Apnea cases, with Overall Survival as target. The performance achieved lead to select Cox Proportional Hazard model, CoxTime and LogHazards as the best models, with a C-Index scores of 81%, 78% and 78% respectively; the Integrated Brier score achieved by such models is of 0.10, 0.12 and 0.11. LogHazard and Cox Proportional Hazard are examined from an explainability perspective. Although the two models show comparable performance metrics, the use of SurvSHAP enables clinician to assess that LogHazard is more reliable and useful in clinical context, according to the feature importance over time. Notably, Cox Proportional Hazard metrics are slight higher w.r.t. LogHazard, underlying that performance and usefulness are not strictly bound to each other.

The second unimodal pipeline deals with IgAN classification from Whole Slide Images according to Oxford Score, resulting a first end-to-end pipeline accomplishing such a task. This pipeline involves two key tasks. First, glomeruli segmentation is carried out using object detection models, with the segmentation results stored in a Qupath project for pathologists to review manually. The most successful segmentation model, Cascade Mask R-CNN, achieved a Dice score of 80.7% on the external test set. Differently from the second multimodal pipeline, in which the classification problem was approached as Multiple Instance Learning problem, here the slide-level classification relies on the independent single-glomerulus classification. The set of glomerulus classifications is used for assign a MESC score to the Whole Slide Image.The classification of MESC lesions, compares the performance of various CNNs and ViT architectures. For the M lesion, the best model is EfficientNetV2-L, with an ROC AUC of 90.2%. For the E lesion, MobileNetV2 performes best, achieving an ROC AUC of 94.8%. Similarly, EfficientNetV2-M is the top model for the S lesion, with a ROC AUC of 92.7%. For the C lesion, EfficientNetV2-L excelled, attaining a ROC AUC of 92.1%. Notably, these results surpass the current state-of-the-art techniques.

The third pipeline presents a visually interpretable deep learning framework for multiclass, shape-based classification of breast lesions in tomosynthesis images. Eight pretrained DL models are used for morphological classification, supported by two explainable AI

(XAI) methods—perceptive and mathematical interpretability—to provide insight into model decisions and build clinician trust. Perceptive models visually highlight the top features for classification, while mathematical models reveal clustering in low-dimensional spaces. DenseNet-161 demonstrates the best performance, achieving over 96% AUC with data augmentation. Augmentation further improves mean AUC across configurations, while high similarity between training and validation data and RoI-level images enhance model accuracy. Using Grad-CAM and LIME, the framework explains the model's focus on lesion areas, with high AUC models like DenseNet-161 and ResNeXt accurately highlighting lesions. This approach effectively demystifies DL models, fostering trust in CAD systems for clinical applications.

The last unimodal pipeline provides a framework for Named Entity Recognition and diagnosis standardization in free-text reports of lymphoma cases. In particular, starting from an existing framework named ARGO Core, an important enhancement was made by supporting ARGO Core with a ML module and a Diagnosis Assignment Manager (DAM) module (resulting into a new framework version, called ARGO 2.0). The adoption of ML with a decision heuristic of DAM module increases ARGO performance achieving a Precision score of 94.85%, a Recall score of 96.31% and F-Score of 95.31%. Finally a migration to a DL model is proposed, achieving promising preliminary results.

In conclusion, this work showcases a wide range of big data computational methodologies across different biomedical heterogeneous fields. The pipelines developed bring advancements in precision medicine-oriented decision support systems.

Apart from the discussions and future enhancements regarding each pipeline (faced at the end of each pipeline section), noteworthy is the flexibility of the models employed across different tasks and target. Moreover, this work highlights how explainable AI methods play a crucial role in understanding a model behavior, increasing the trustworthy in AI.

In light of this, future steps can involve the integration of all pipelines in a single distributed framework, making them agnostic to the target and to the data information nature, considering only the main task (i.e. a classification, a Named Entity Recognition, a survival analysis, and so on) and the type of input data (generic images and generic tabular data). In this way the models involved can be re-used to cover use cases in fields different from the biomedical one, abstracting them at a higher level.

My Publications

1. **Francesco Berloco**, Vitoantonio Bevilacqua*, Simona Colucci, *Distributed Analytics For Big Data: A Survey*, Neurocomputing, Volume 574, 2024, 127258, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2024.127258.

2. **Francesco Berloco**, Pietro Maria Marvulli, Vladimiro Suglia, Simona Colucci, Gaetano Pagano*, Lucia Palazzo, Maria Aliani, Giorgio Castellana, Patrizia Guido, Giovanni D'Addio, Vitoantonio Bevilacqua, *Enhancing Survival Analysis Model Selection through XAI(t) in Healthcare*, Applied Sciences, Volume 14, Issue 14, 2024, Article 6084, https://doi.org/10.3390/app14146084.

3. **Francesco Berloco**, Gian Maria Zaccaria, Nicola Altini*, Simona Colucci, Vitoantonio Bevilacqua, *A Multimodal Framework for Assessing the Link between Pathomics, Transcriptomics, and Pancreatic Cancer Mutation*, (submitted to Computerized Medical Imaging And Graphics - Under Review).

4. Gian Maria Zaccaria, **Francesco Berloco***, Domenico Buongiorno, Antonio Brunetti, Nicola Altini, Vitoantonio Bevilacqua, *A time-dependent explainable radiomic analysis from the multi-omic cohort of CPTAC-Pancreatic Ductal Adenocarcinoma*, Computer Methods and Programs in Biomedicine, Volume 257, 2024, 108408, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2024.108408.

5. Nicola Altini, Michele Rossini, Sándor Turkevi-Nagy, Francesco Pesce, Paola Pontrelli, Berardino Prencipe, **Francesco Berloco**, Surya Seshan, Jean-Baptiste Gibier, Aníbal Pedraza Dorado, Gloria Bueno, Licia Peruzzi, Mattia Rossi, Albino Eccher, Feifei Li, Adamantios Koumpis, Oya Beyan, Jonathan Barratt, Huy Quoc Vo, Chandra Mohan, Hien Van Nguyen, Pietro Antonio Cicalese, Angela Ernst, Loreto Gesualdo, Vitoantonio Bevilacqua*, Jan Ulrich Becker, *Performance and limitations of a supervised deep learning approach for the histopathological Oxford Classification of glomeruli*

with IgA nephropathy, Computer Methods and Programs in Biomedicine, Volume 242, 2023, 107814, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2023.107814.

6. Sardar Mehboob Hussain, Domenico Buongiorno, Nicola Altini, **Francesco Berloco**, Berardino Prencipe, Marco Moschetta, Vitoantonio Bevilacqua* and Antonio Brunetti, *Shape-Based Breast Lesion Classification using Digital Tomosynthesis Images: the role of Explainable Artificial Intelligence*, Applied Sciences, Volume 12, Issue 12, 2022, Article 6230, https://doi.org/10.3390/app12126230.

7. **Francesco Berloco***, Sabino Ciavarella, Simona Colucci, Luigi Alfredo Grieco, Attilio Guarini, Gian Maria Zaccaria, *ARGO 2.0: a Hybrid NLP/ML Framework for Diagnosis Standardization*, 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, 2023, pp. 1-4, https://doi.org/10.1109/EMBC40787.2023.10340022.

8. **Francesco Berloco***, Vitoantonio Bevilacqua, Simona Colucci, *A Systematic Review of Distributed Deep Learning Frameworks for Big Data*, in: D.S. Huang, K.H. Jo, J. Jing, P. Premaratne, V. Bevilacqua, A. Hussain (eds), Intelligent Computing Methodologies. ICIC 2022. Lecture Notes in Computer Science, vol. 13395. Springer, Cham, https://doi.org/10.1007/978-3-031-13832-4_21.

* Corresponding Author.

References

[1] Andrea De Mauro, Marco Greco, and Michele Grimaldi. What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings*, 1644(1):97, feb 2015. ISSN 0094-243X. doi: 10.1063/1.4907823. URL https://aip.scitation.org/doi/abs/10.1063/1.4907823.

[2] Inke R. König, Oliver Fuchs, Gesine Hansen, Erika von Mutius, and Matthias V. Kopp. What is precision medicine? *European Respiratory Journal*, 50:1700391, 10 2017. ISSN 0903-1936. doi: 10.1183/13993003.00391-2017. URL https://erj.ersjournals.com/content/50/4/1700391https://erj.ersjournals.com/content/50/4/1700391.abstract.

[3] Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine 2020 3:1*, 3:1–10, 2 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0221-y. URL https://www.nature.com/articles/s41746-020-0221-y.

[4] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2022.

[5] Lu Wang, Xinyi Chen, Lu Zhang, Long Li, YongBiao Huang, Yinan Sun, and Xianglin Yuan. Artificial intelligence in clinical decision support systems for oncology. *International Journal of Medical Sciences*, 20(1):79, 2023.

[6] Janita E. van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging*, 11:1–16, 12 2020. ISSN 18694101. doi: 10.1186/S13244-020-00887-2/TABLES/3. URL https://insightsimaging.springeropen.com/articles/10.1186/s13244-020-00887-2.

[7] Michele Avanzo, Lise Wei, Joseph Stancanello, Martin Vallières, Arvind Rao, Olivier Morin, Sarah A. Mattonen, and Issam El Naqa. Machine and deep learning methods for radiomics. *Medical Physics*, 47:e185–e202, 5 2020. ISSN 2473-4209. doi: 10.1002/MP.13678. URL https://onlinelibrary.wiley.com/doi/full/10.1002/mp.13678https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13678https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.13678.

[8] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies. Radiomics: the process and the challenges. *Magnetic resonance imaging*, 30:1234–1248, 11 2012. ISSN 1873-5894. doi: 10.1016/J.MRI.2012.06.010.

[9] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G.P.M. Van Stiphout, Patrick Granton, Catharina M.L. Zegers, Robert Gillies, Ronald Boellard, André Dekker, and Hugo J.W.L. Aerts. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48:441–446, 3 2012. ISSN 0959-8049. doi: 10.1016/J.EJCA.2011.11.036.

[10] George Barbastathis, Aydogan Ozcan, and Guohai Situ. On the use of deep learning for computational imaging. *Optica*, 6:921, 8 2019. ISSN 23342536. doi: 10.1364/OPTICA.6.000921.

[11] Heang Ping Chan, Ravi K. Samala, Lubomir M. Hadjiiski, and Chuan Zhou. Deep learning in medical image analysis. *Advances in experimental medicine and biology*, 1213:3, 2020. ISSN 22148019. doi: 10.1007/978-3-030-33128-3_1.

[12] Dinggang Shen, Guorong Wu, and Heung Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221, 6 2017. ISSN 15454274. doi: 10.1146/ANNUREV-BIOENG-071516-044442. URL /pmc/articles/PMC5479722//pmc/articles/PMC5479722/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5479722/.

[13] Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, Francescomaria Marino, Maria Teresa Rocchetti, Silvia Matino, Umberto Venere, Michele Rossini, Francesco Pesce, Loreto Gesualdo, and Vitoantonio Bevilacqua. Semantic segmentation framework for glomeruli detection and classification in kidney histological sections. *Electronics*, 9 (3):503, 3 2020. ISSN 2079-9292. doi: 10.3390/electronics9030503.

[14] Nicola Altini, Berardino Prencipe, Giacomo Donato Cascarano, Antonio Brunetti, Gioacchino Brunetti, Vito Triggiani, Leonarda Carnimeo, Francescomaria Marino, Andrea Guerriero, Laura Villani, Arnaldo Scardapane, and Vitoantonio Bevilacqua. Liver, kidney and spleen segmentation from ct scans and mri with deep learning: A survey. *Neurocomputing*, 490:30–53, 6 2022. ISSN 0925-2312. doi: 10.1016/J.NEUCOM.2021.08.157.

[15] Berardino Prencipe, Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, Andrea Guerriero, and Vitoantonio Bevilacqua. Focal dice loss-based v-net for liver segments classification. *Applied Sciences 2022, Vol. 12, Page 3247*, 12:3247, 3 2022. ISSN 2076-3417. doi: 10.3390/APP12073247.

[16] Dylan Feldner-Busztin, Panos Firbas Nisantzis, Shelley Jane Edmunds, Gergely Boza, Fernando Racimo, Shyam Gopalakrishnan, Morten Tønsberg Limborg, Leo Lahti, and Gonzalo G. de Polavieja. Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics*, 39, 2 2023. ISSN 13674811. doi: 10.1093/BIOINFORMATICS/BTAD021.

[17] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[18] Aleix M Martinez and Avinash C Kak. Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):228–233, 2001.

[19] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003. doi: 10.5555/944919.944968.

[20] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. doi: doi.org/10.1080/01621459.2024.2412464.

[21] Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, 4 2016. ISSN 0925-2312. doi: 10.1016/J.NEUCOM.2015.08.104.

[22] Mohammad Sultan Mahmud, Joshua Zhexue Huang, and Xianghua Fu. Variational autoencoder-based dimensionality reduction for high-dimensional small-sample data classification. *International Journal of Computational Intelligence and Applications*, 19 (01):2050002, 2020.

[23] Gian Maria Zaccaria, Nicola Altini, Giuseppe Mezzolla, Maria Carmela Vegliante, Marianna Stranieri, Susanna Anita Pappagallo, Sabino Ciavarella, Attilio Guarini, and Vitoantonio Bevilacqua. Surviae: Survival prediction with interpretable autoencoders from diffuse large b-cells lymphoma gene expression data. *Computer Methods and Programs in Biomedicine*, 244:107966, 2 2024. ISSN 0169-2607. doi: 10.1016/J.CMPB.2023.107966.

[24] Ili Nadhirah Jamil, Juwairiah Remali, Kamalrul Azlan Azizan, Nor Azlan Nor Muhammad, Masanori Arita, Hoe-Han Goh, and Wan Mohd Aizat. Systematic multi-omics integration (moi) approach in plant systems biology. *Frontiers in plant science*, 11:944, 2020.

[25] Nam D Nguyen and Daifeng Wang. Multiview learning for understanding functional multiomics. *PLoS computational biology*, 16(4):e1007677, 2020. doi: doi:10.1371/journal.pcbi.1007677.PMID:32240163.

[26] Lianhe Zhao, Qiongye Dong, Chunlong Luo, Yang Wu, Dechao Bu, Xiaoning Qi, Yufan Luo, and Yi Zhao. Deepomix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. *Computational and Structural Biotechnology Journal*, 19:2719–2725, 1 2021. ISSN 2001-0370. doi: 10.1016/J.CSBJ.2021.04.067.

[27] Gian Maria Zaccaria, Francesco Berloco, Domenico Buongiorno, Antonio Brunetti, Nicola Altini, and Vitoantonio Bevilacqua. A time-dependent explainable radiomic analysis from the multi-omic cohort of CPTAC-Pancreatic Ductal Adenocarcinoma. *Computer Methods and Programs in Biomedicine*, 257:108408, December 2024. ISSN 0169-2607. doi: 10.1016/j.cmpb.2024.108408.

[28] Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015, 2016.

[29] Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[30] Frank Po Yen Lin, Adrian Pokorny, Christina Teng, and Richard J. Epstein. Tepapa: a novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. *Scientific Reports 2017 7:1*, 7:1–13, 7 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-07111-0.

[31] Yichi Zhang, Tianrun Cai, Sheng Yu, Kelly Cho, Chuan Hong, Jiehuan Sun, Jie Huang, Yuk Lam Ho, Ashwin N. Ananthakrishnan, Zongqi Xia, Stanley Y. Shaw, Vivian Gainer, Victor Castro, Nicholas Link, Jacqueline Honerlaw, Sicong Huang, David Gagnon, Elizabeth W. Karlson, Robert M. Plenge, Peter Szolovits, Guergana Savova, Susanne Churchill, Christopher O'Donnell, Shawn N. Murphy, J. Michael Gaziano, Isaac Kohane, Tianxi Cai, and Katherine P. Liao. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nature Protocols 2019 14:12*, 14:3426–3444, 11 2019. ISSN 1750-2799. doi: 10.1038/s41596-019-0227-6.

[32] Guhan Ram Venkataraman, Arturo Lopez Pineda, Oliver J. Bear Don't Walk, Ashley M. Zehnder, Sandeep Ayyar, Rodney L. Page, Carlos D. Bustamante, and Manuel A. Rivas. Fastag: Automatic text classification of unstructured medical narratives. *PloS one*, 15, 6 2020. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0234647.

[33] Juliana KF Bowles, Juan Mendoza-Santana, Andreas F Vermeulen, Thais Webber, and Euan Blackledge. Integrating healthcare data for enhanced citizen-centred care and analytics. In *Integrated Citizen Centered Digital Health and Social Care*, pages 17–21. IOS Press, 2020. doi: 10.3233/SHTI200686.

[34] Ashwin Belle, Raghuram Thiagarajan, SM Reza Soroushmehr, Fatemeh Navidi, Daniel A Beard, and Kayvan Najarian. Big data analytics in healthcare. *BioMed research international*, 2015(1):370194, 2015.

[35] Daniel Schwabe, Katinka Becker, Martin Seyferth, Andreas Klaß, and Tobias Schaeffter. The metric-framework for assessing data quality for trustworthy ai in medicine: a systematic review. *npj Digital Medicine 2024 7:1*, 7:1–30, 8 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01196-4.

[36] Edward H. Shortliffe and Martin J. Sepúlveda. Clinical decision support in the era of artificial intelligence. *JAMA*, 320:2199–2200, 12 2018. ISSN 0098-7484. doi: 10.1001/JAMA.2018.17163.

[37] Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424, 2021.

[38] Berloco Francesco, Zaccaria Gian Maria, Altini Nicola, Colucci Simona, and Bevilacqua Vitoantonio. A multimodal framework for assessing the link between pathomics, transcriptomics, and pancreatic cancer mutation. *Submitted to - Computerized Medical Imaging And Graphics - Under Review*, 2024.

[39] Elizabeth D. Thompson, Nicholas J. Roberts, Laura D. Wood, James R. Eshleman, Michael G. Goggins, Scott E. Kern, Alison P. Klein, and Ralph H. Hruban. The genetics of ductal adenocarcinoma of the pancreas in the year 2020: Dramatic progress, but far to go. *Modern Pathology*, 33(12):2544–2563, December 2020. ISSN 08933952. doi: 10.1038/s41379-020-0629-6.

[40] Francesco Berloco, Pietro Maria Marvulli, Vladimiro Suglia, Simona Colucci, Gaetano Pagano, Lucia Palazzo, Maria Aliani, Giorgio Castellana, Patrizia Guido, Giovanni D ' Addio, and Vitoantonio Bevilacqua. Enhancing survival analysis model selection through xai(t) in healthcare. *Applied Sciences 2024, Vol. 14, Page 6084*, 14:6084, 7 2024. ISSN 2076-3417. doi: 10.3390/APP14146084.

[41] Nicola Altini, Michele Rossini, Sándor Turkevi-Nagy, Francesco Pesce, Paola Pontrelli, Berardino Prencipe, Francesco Berloco, Surya Seshan, Jean Baptiste Gibier, Aníbal Pedraza Dorado, Gloria Bueno, Licia Peruzzi, Mattia Rossi, Albino Eccher, Feifei Li, Adamantios Koumpis, Oya Beyan, Jonathan Barratt, Huy Quoc Vo, Chandra Mohan, Hien Van Nguyen, Pietro Antonio Cicalese, Angela Ernst, Loreto Gesualdo, Vitoantonio Bevilacqua, and Jan Ulrich Becker. Performance and limitations of a supervised deep learning approach for the histopathological oxford classification of glomeruli with iga nephropathy. *Computer Methods and Programs in Biomedicine*, 242:107814, 12 2023. ISSN 0169-2607. doi: 10.1016/J.CMPB.2023.107814.

[42] Sardar Mehboob Hussain, Domenico Buongiorno, Nicola Altini, Francesco Berloco, Berardino Prencipe, Marco Moschetta, Vitoantonio Bevilacqua, and Antonio Brunetti. Shape-based breast lesion classification using digital tomosynthesis images: The role of explainable artificial intelligence. *Applied Sciences 2022, Vol. 12, Page 6230*, 12:6230, 6 2022. ISSN 2076-3417. doi: 10.3390/APP12126230.

[43] Francesco Berloco, Sabino Ciavarella, Simona Colucci, Luigi Alfredo Grieco, Attilio Guarini, and Gian Maria Zaccaria. Argo 2.0: a hybrid nlp/ml framework for diagnosis standardization. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4, 2023. doi: 10.1109/EMBC40787.2023.10340022.

[44] World Health Organization. Icd-10 : international statistical classification of diseases and related health problems : tenth revision, 2004.

[45] Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021.

[46] Zehui Zhao, Laith Alzubaidi, Jinglan Zhang, Ye Duan, and Yuantong Gu. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*, page 122807, 2023. doi: doi.org/10.1016/j.eswa.2023.122807.

[47] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

[48] Suryakanthi Tangirala. Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2):612–619, 2020.

[49] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[50] Hemant Ishwaran, Udaya Kogalur, Eugene Blackstone, and Michael Lauer. Random survival forests. *The Annals of Applied Statistics*, 2, 12 2008. doi: 10.1214/08-AOAS169.

[51] Greg Ridgeway. The state of boosting. *Comp Sci Stat*, 31, 12 2001.

[52] C.J.K. Fouodo, Inke König, Claus Weihs, Andreas Ziegler, and Marvin Wright. Support vector machines for survival analysis with r. *R Journal*, 10:412–423, 07 2018. doi: 10.32614/RJ-2018-005.

[53] Vanya Van Belle, Kristiaan Pelckmans, Sabine Huffel, and Johan Suykens. Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial intelligence in medicine*, 53:107–18, 08 2011. doi: 10.1016/j.artmed.2011.06.006.

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[55] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[56] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[57] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[58] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. pages 2261–2269, 7 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.243.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778, 12 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.90.

[60] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[61] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[63] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[64] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[65] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[66] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[67] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

[68] Ming Y. Lu, Drew F.K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering 2021 5:6*, 5:555–570, 3 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w.

[69] Jared Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18, 02 2018. doi: 10.1186/s12874-018-0482-1.

[70] Michael Francis Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7, 2018. URL https://api.semanticscholar.org/CorpusID:53875100.

[71] yu Deng, Lei Liu, Hongmei Jiang, Yifan Peng, Yishu Wei, Zhiyang Zhou, Yizhen Zhong, Yun Zhao, Xiaoyun Yang, Jingzhi Yu, Zhiyong Lu, Abel Kho, Hongyan Ning, Norrina Allen, John Wilkins, Kiang Liu, Donald Lloyd-Jones, and Lihui Zhao. Comparison of state-of-the-art neural network survival models with the pooled cohort equations for cardiovascular disease risk prediction. *BMC Medical Research Methodology*, 23, 01 2023. doi: 10.1186/s12874-022-01829-w.

[72] Antonio Eleuteri, Min Aung, Azzam Taktak, Bertil Damato, and P.j.g Lisboa. Continuous and discrete time survival analysis: Neural network approaches. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2007:5420–3, 08 2007. doi: 10.1109/IEMBS.2007.4353568.

[73] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. 2019.

[74] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, 04 2018. doi: 10.1609/aaai.v32i1.11842.

[75] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374, 2023.

[76] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2):1–38, 2022.

[77] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[78] Lloyd S Shapley. Notes on the n-person game—ii: The value of an n-person game. 1951.

[79] Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. SurvSHAP(t): Time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262:110234, 2023. ISSN 09507051. doi: 10.1016/j.knosys.2022.110234.

[80] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[81] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[82] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.

[83] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387, 2 1996. ISSN 0277-6715. doi: 10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4.

[84] Jérôme Lambert and Sylvie Chevret. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent roc curves. *Statistical methods in medical research*, 25:2088–2102, 10 2016. ISSN 1477-0334. doi: 10.1177/0962280213515571.

[85] Charles Sawyer, F W Reichelderfer, James E Editor, J R Caskey, and Glenn W Brie. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78: 1–3, 1 1950. ISSN 1520-0493. doi: 10.1175/1520-0493(1950)078.

[86] Janita E Van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into imaging*, 11(1):1–16, 2020.

[87] Robert M Haralick, Its'hak Dinstein, and K Shanmugam. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6):610–621, 1973. ISSN 21682909. doi: 10.1109/TSMC.1973.4309314.

[88] Mary M Galloway. Texture analysis using gray level run lengths. *Computer graphics and image processing*, 4(2):172–179, 1975.

[89] A Chu, Chandra M Sehgal, and James F Greenleaf. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters*, 11(6):415–419, 1990.

[90] Guillaume Thibault, Bernard Fertil, Claire Navarro, Sandrine Pereira, Pierre Cau, Nicolas Levy, Jean Sequeira, and Jean-luc Mari. Texture indexes and gray level size zone matrix application to cell nuclei classification. *Pattern Recognition and Information Processing*, (May 2014):140–145, 2009.

[91] Chengjun Sun and William G Wee. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*, 23(3):341–352, 1983. ISSN 0734-189X. doi: https://doi.org/10.1016/0734-189X(83)90032-4.

[92] M Amadasun and R King. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(5):1264–1274, 1989. doi: 10.1109/21.44046.

[93] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies. Radiomics: The process and the challenges. *Magnetic Resonance Imaging*, 30(9):1234–1248, 2012. ISSN 0730725X. doi: 10.1016/j.mri.2012.06.010.

[94] Marius E Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to radiomics. *Journal of Nuclear Medicine*, 61(4):488–495, 2020.

[95] Hishan Tharmaseelan, Abhinay K Vellala, Alexander Hertel, Fabian Tollens, Lukas T Rotkopf, Johann Rink, Piotr Woźnicki, Isabelle Ayx, Sönke Bartling, Dominik Nörenberg, et al. Tumor classification of gastrointestinal liver metastases using ct-based radiomics and deep learning. *Cancer Imaging*, 23(1):95, 2023.

[96] Anuj Kumar, Ashish Kumar Jha, Jai Prakash Agarwal, Manender Yadav, Suvarna Badhe, Ayushi Sahay, Sridhar Epari, Arpita Sahu, Kajari Bhattacharya, Abhishek Chatterjee, et al. Machine-learning-based radiomics for classifying glioma grade from magnetic resonance images of the brain. *Journal of Personalized Medicine*, 13(6):920, 2023.

[97] Elsa Parr, Qian Du, Chi Zhang, Chi Lin, Ahsan Kamal, Josiah McAlister, Xiaoying Liang, Kyle Bavitz, Gerard Rux, Michael Hollingsworth, et al. Radiomics-based outcome prediction for pancreatic cancer following stereotactic body radiotherapy. *Cancers*, 12 (4):1051, 2020.

[98] Antonio Brunetti, Nicola Altini, Domenico Buongiorno, Emilio Garolla, Fabio Corallo, Matteo Gravina, Vitoantonio Bevilacqua, and Berardino Prencipe. A machine learning and radiomics approach in lung cancer for predicting histological subtype. *Applied Sciences*, 12(12):5829, 2022.

[99] Vitoantonio Bevilacqua, Antonio Brunetti, Andrea Guerriero, Gianpaolo Francesco Trotta, Michele Telegrafo, and Marco Moschetta. A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. *Cognitive Systems Research*, 53:3–19, 2019.

[100] Wei Wang, Ying Peng, Xingyu Feng, Yan Zhao, Sharvesh Raj Seeruttun, Jun Zhang, Zixuan Cheng, Yong Li, Zaiyi Liu, and Zhiwei Zhou. Development and validation of a computed tomography–based radiomics signature to predict response to neoadjuvant chemotherapy for locally advanced gastric cancer. *JAMA network open*, 4(8):e2121143–e2121143, 2021.

[101] Shuai Ren, Rui Zhao, Wenjing Cui, Wenli Qiu, Kai Guo, Yingying Cao, Shaofeng Duan, Zhongqiu Wang, and Rong Chen. Computed tomography-based radiomics signature for the preoperative differentiation of pancreatic adenosquamous carcinoma from pancreatic ductal adenocarcinoma. *Frontiers in Oncology*, 10:1618, 2020.

[102] Jia-Jun Qiu, Jin Yin, Wei Qian, Jin-Heng Liu, Zi-Xing Huang, Hao-Peng Yu, Lin Ji, and Xiao-Xi Zeng. A novel multiresolution-statistical texture analysis architecture: Radiomics-aided diagnosis of pdac based on plain ct images. *IEEE Transactions on Medical Imaging*, 40(1):12–25, 2021. doi: 10.1109/TMI.2020.3021254.

[103] Ke Li, Qiandong Yao, Jingjing Xiao, Meng Li, Jiali Yang, Wenjing Hou, Mingshan Du, Kang Chen, Yuan Qu, Lian Li, et al. Contrast-enhanced ct radiomics for predicting lymph node metastasis in pancreatic ductal adenocarcinoma: a pilot study. *Cancer Imaging*, 20:1–10, 2020.

[104] Tiansong Xie, Xuanyi Wang, Menglei Li, Tong Tong, Xiaoli Yu, and Zhengrong Zhou. Pancreatic ductal adenocarcinoma: a radiomics nomogram outperforms clinical model and tnm staging for survival estimation after curative resection. *European Radiology*, 30: 2513–2524, 2020.

[105] Farzad Khalvati, Yucheng Zhang, Sameer Baig, Edrise M Lobo-Mueller, Paul Karanicolas, Steven Gallinger, and Masoom A Haider. Prognostic value of ct radiomic features in resectable pancreatic ductal adenocarcinoma. *Scientific reports*, 9(1):5449, 2019.

[106] Julius Keyl, Stefan Kasper, Marcel Wiesweg, Julian Götze, Martin Schönrock, Marianne Sinn, Aron Berger, Enrico Nasca, Karina Kostbade, Brigitte Schumacher, et al. Multimodal survival prediction in advanced pancreatic cancer using machine learning. *ESMO open*, 7(5):100555, 2022.

[107] The emergence of pathomics. *Current Pathobiology Reports*, 7:73–84, 9 2019. ISSN 2167485X. doi: 10.1007/S40139-019-00200-X/FIGURES/4. URL https://link.springer.com/article/10.1007/s40139-019-00200-x.

[108] Paola Ghiorzo. Genetic predisposition to pancreatic cancer. *World Journal of Gastroenterology*, 20(31):10778, 2014. ISSN 1007-9327. doi: 10.3748/wjg.v20.i31.10778.

[109] Yuriko Saiki, Can Jiang, Masaki Ohmuraya, and Toru Furukawa. Genetic mutations of pancreatic cancer and genetically engineered mouse models. *Cancers*, 14(1):71, 2021.

[110] Aamir Ali Khan, Xinhui Liu, Xinlong Yan, Muhammad Tahir, Sakhawat Ali, and Hua Huang. An overview of genetic mutations and epigenetic signatures in the course of pancreatic cancer progression. *Cancer and Metastasis Reviews*, 40(1):245–272, March 2021. ISSN 0167-7659, 1573-7233. doi: 10.1007/s10555-020-09952-0.

[111] Dimitrios Stefanoudakis, Maximos Frountzas, Dimitrios Schizas, Nikolaos V. Michalopoulos, Alexandra Drakaki, and Konstantinos G. Toutouzas. Significance of TP53, CDKN2A, SMAD4 and KRAS in Pancreatic Cancer. *Current Issues in Molecular Biology*, 46(4):2827–2844, March 2024. ISSN 1467-3045. doi: 10.3390/cimb46040177.

[112] Siân Jones, Xiaosong Zhang, D. Williams Parsons, Jimmy Cheng-Ho Lin, Rebecca J. Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, Hirohiko Kamiyama, Antonio Jimeno, Seung-Mo Hong, Baojin Fu, Ming-Tseh Lin, Eric S. Calhoun, Mihoko Kamiyama, Kimberly Walter, Tatiana Nikolskaya, Yuri Nikolsky, James Hartigan, Douglas R. Smith, Manuel Hidalgo, Steven D. Leach, Alison P. Klein, Elizabeth M. Jaffee, Michael Goggins, Anirban Maitra, Christine Iacobuzio-Donahue, James R. Eshleman, Scott E. Kern, Ralph H. Hruban, Rachel Karchin, Nickolas Papadopoulos, Giovanni Parmigiani, Bert Vogelstein, Victor E. Velculescu, and Kenneth W. Kinzler. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, 321(5897):1801–1806, 2008. doi: 10.1126/science.1164368.

[113] Samuel Amintas, Benjamin Fernandez, Alexandre Chauvet, Laurence Chiche, Christophe Laurent, Geneviève Belleannée, Marion Marty, Etienne Buscail, and Sandrine Dabernat. Kras gene mutation quantification in the resection or venous margins of pancreatic ductal adenocarcinoma is not predictive of disease recurrence. *Scientific Reports*, 12(1):2976, 2022. doi: 10.1038/s41598-022-07004-x.

[114] Peter Storz. Kras, ros and the initiation of pancreatic cancer. *Small GTPases*, 8(1): 38–42, 2017. doi: 10.1080/21541248.2016.1192714.

[115] Ming Zhao, Lopa Mishra, and Chu-Xia Deng. The role of tgf-β/smad4 signaling in cancer. *International journal of biological sciences*, 14(2):111, 2018. doi: 10.7150/ijbs.23230.

[116] Robert R McWilliams, Eric D Wieben, Kari G Rabe, Katrina S Pedersen, Yanhong Wu, Hugues Sicotte, and Gloria M Petersen. Prevalence of cdkn2a mutations in pancreatic cancer patients: implications for genetic counseling. *European Journal of Human Genetics*, 19(4):472–478, 2011. doi: 10.1038/ejhg.2010.198.

[117] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature cancer*, 1(8):800–810, 2020. doi: 10.1038/s43018-020-0085-8.

[118] Lili Feng, Zhenyu Liu, Chaofeng Li, Zhenhui Li, Xiaoying Lou, Lizhi Shao, Yunlong Wang, Yan Huang, Haiyang Chen, Xiaolin Pang, et al. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. *The Lancet Digital Health*, 4(1):e8–e17, 2022. doi: 10.1016/S2589-7500(21)00215-6.

[119] Arsen Osipov, Ognjen Nikolic, Arkadiusz Gertych, Sarah Parker, Andrew Hendifar, Pranav Singh, Darya Filippova, Grant Dagliyan, Cristina R. Ferrone, Lei Zheng, Jason H. Moore, Warren Tourtellotte, Jennifer E. Van Eyk, and Dan Theodorescu. The Molecular Twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients. *Nature Cancer*, January 2024. ISSN 2662-1347. doi: 10.1038/s43018-023-00697-7.

[120] Michaela Unger and Jakob Nikolas Kather. A systematic analysis of deep learning in genomics and histopathology for precision oncology. *BMC Medical Genomics*, 17:1–10, 12 2024. ISSN 17558794. doi: 10.1186/S12920-024-01796-9/FIGURES/2.

[121] Jakob Nikolas Kather, Lara R. Heij, Heike I. Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M. Niehues, Kai A.J. Sommer, Peter Bankhead, Loes F.S. Kooreman, Jefree J. Schulte, Nicole A. Cipriani, Roman D. Buelow, Peter Boor, Nadina Ortiz-Brüchle, Andrew M. Hanby, Valerie Speirs, Sara Kochanny, Akash Patnaik, Andrew Srisuwananukorn, Hermann Brenner, Michael Hoffmeister, Piet A. van den Brandt, Dirk Jäger, Christian Trautwein, Alexander T. Pearson, and Tom Luedde. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer 2020 1:8*, 1:789–799, 7 2020. ISSN 2662-1347. doi: 10.1038/s43018-020-0087-6.

[122] Daisuke Komura, Akihiro Kawabe, Keisuke Fukuta, Kyohei Sano, Toshikazu Umezaki, Hirotomo Koda, Ryohei Suzuki, Ken Tominaga, Mieko Ochi, Hiroki Konishi, Fumiya Masakado, Noriyuki Saito, Yasuyoshi Sato, Takumi Onoyama, Shu Nishida, Genta Furuya, Hiroto Katoh, Hiroharu Yamashita, Kazuhiro Kakimi, Yasuyuki Seto, Tetsuo Ushiku, Masashi Fukayama, and Shumpei Ishikawa. Universal encoding of pan-cancer histology by deep texture representations. *Cell Reports*, 38:110424, 3 2022. ISSN 2211-1247. doi: 10.1016/J.CELREP.2022.110424.

[123] Oliver Lester Saldanha, Chiara M.L. Loeffler, Jan Moritz Niehues, Marko van Treeck, Tobias P. Seraphin, Katherine Jane Hewitt, Didem Cifci, Gregory Patrick Veldhuizen, Siddhi Ramesh, Alexander T. Pearson, and Jakob Nikolas Kather. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *npj Precision Oncology 2023 7:1*, 7:1–5, 3 2023. ISSN 2397-768X. doi: 10.1038/s41698-023-00365-0.

[124] Jake Crawford, Brock C Christensen, Maria Chikina, and Casey S Greene. Widespread redundancy in-omics profiles of cancer mutation states. *Genome Biology*, 23(1):137, 2022. doi: 10.1186/s13059-022-02705-y.

[125] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium pancreatic ductal adenocarcinoma collection (cptac-pda). The Cancer Imaging Archive, 2018. URL https://www.cancerimagingarchive.net/collection/cptac-pda/.

[126] M Rozenfeld and P Jordan. Annotations for the clinical proteomic tumor analysis consortium pancreatic ductal adenocarcinoma collection (cptac-pda-tumor-annotations)(version 1)[data set]. *The Cancer Imaging Archive*, 2023.

[127] Berardino Prencipe, Claudia Delprete, Emilio Garolla, Fabio Corallo, Matteo Gravina, Maria Iole Natalicchio, Domenico Buongiorno, Vitoantonio Bevilacqua, Nicola Altini, and Antonio Brunetti. An explainable radiogenomic framework to predict mutational

status of kras and egfr in lung adenocarcinoma patients. *Bioengineering*, 10(7):747, 2023.

[128] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[129] Gian Maria Zaccaria, Maria Carmela Vegliante, Giuseppe Mezzolla, Marianna Stranieri, Giacomo Volpe, Nicola Altini, Grazia Gargano, Susanna Anita Pappagallo, Antonella Bucci, Flavia Esposito, et al. A decision-tree approach to stratify dlbcl risk based on stromal and immune microenvironment determinants. *HemaSphere*, 7(4):e862, 2023.

[130] Gian Maria Zaccaria, Nicola Altini, Giuseppe Mezzolla, Maria Carmela Vegliante, Marianna Stranieri, Susanna Anita Pappagallo, Sabino Ciavarella, Attilio Guarini, and Vitoantonio Bevilacqua. Surviae: survival prediction with interpretable autoencoders from diffuse large b-cells lymphoma gene expression data. *Computer Methods and Programs in Biomedicine*, 244:107966, 2024. doi: doi.org/10.1016/j.cmpb.2023.107966.

[131] Yonggang He, Wen Huang, Yichen Tang, Yuming Li, Xuehui Peng, Jing Li, Jing Wu, Nan You, Ling Li, Chuang Liu, et al. Clinical and genetic characteristics in pancreatic cancer from chinese patients revealed by whole exome sequencing. *Frontiers in Oncology*, 13:1167144, 2023.

[132] Bin Baek and Hyunju Lee. Prediction of survival and recurrence in patients with pancreatic cancer by integrating multi-omics data. *Scientific reports*, 10(1):18951, 2020.

[133] Benjamin J. Raphael, Ralph H. Hruban, Andrew J. Aguirre, Richard A. Moffitt, Jen Jen Yeh, and Chip Stewart et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell*, 32:185, 8 2017. ISSN 18783686. doi: 10.1016/J.CCELL.2017.07.007. URL /pmc/articles/PMC5964983//pmc/articles/PMC5964983/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5964983/.

[134] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:1–21, 12 2014. ISSN 1474760X. doi: 10.1186/S13059-014-0550-8/FIGURES/9. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

[135] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[136] Gregory P Way and Casey S Greene. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pages 80–91. World Scientific, 2018.

[137] Sardar Mehboob Hussain, Domenico Buongiorno, Nicola Altini, Francesco Berloco, Berardino Prencipe, Marco Moschetta, Vitoantonio Bevilacqua, and Antonio Brunetti. Shape-Based Breast Lesion Classification Using Digital Tomosynthesis Images: The Role of Explainable Artificial Intelligence. *Applied Sciences*, 12(12):6230, June 2022. ISSN 2076-3417. doi: 10.3390/app12126230.

[138] Nicola Altini, Emilia Puro, Maria Giovanna Taccogna, Francescomaria Marino, Simona De Summa, Concetta Saponaro, Eliseo Mattioli, Francesco Alfredo Zito, and Vitoantonio Bevilacqua. Tumor cellularity assessment of breast histopathological slides via instance segmentation and pathomic features explainability. *Bioengineering*, 10(4): 396, 2023. doi: 10.3390/bioengineering10040396.

[139] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *A Unified Approach to Interpreting Model Predictions*, NIPS'17, pages 4768–4777, Long Beach, California, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. doi: 10.5555/3295222.3295230.

[140] Barnaby Crook, Maximilian Schlüter, and Timo Speith. Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 316–324, 2023. doi: 10.1109/REW57809.2023.00060.

[141] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 2019 1:5*, 1:206–215, 5 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://www.nature.com/articles/s42256-019-0048-x.

[142] Zachary C. Lipton. The mythos of model interpretability, 2017. URL https://arxiv.org/abs/1606.03490.

[143] Domenico Scrutinio, Pietro Guida, Maria Aliani, Giorgio Castellana, Patrizia Guido, and Mauro Carone. Age and comorbidities are crucial predictors of mortality in severe obstructive sleep apnoea syndrome. *European Journal of Internal Medicine*, 90 (December 2020):71–76, 2021. ISSN 18790828. doi: 10.1016/j.ejim.2021.04.018.

[144] Margaux Blanchard, Mathieu Feuilloy, Abdelkebir Sabil, Chloé Gervès-Pinquié, Frédéric Gagnadoux, and Jean-Marc Girault. A deep survival learning approach for cardiovascular risk estimation in patients with sleep apnea. *IEEE Access*, 10:133468–133478, 2022. doi: 10.1109/ACCESS.2022.3231743.

[145] Gaetano Pagano, Maria Aliani, Maddalena Genco, Armando Coccia, Vito Proscia, Mario Cesarelli, and Giovanni D'Addio. Rehabilitation outcome in patients with obstructive sleep apnea syndrome using wearable inertial sensor for gait analysis. *2022 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2022 - Conference Proceedings*, pages 1–6, 2022. doi: 10.1109/MeMeA54994.2022.9856405.

[146] Eun Yeol Ma, Jeong Whun Kim, Youngmin Lee, Sung Woo Cho, Heeyoung Kim, and Jae Kyoung Kim. Combined unsupervised-supervised machine learning for phenotyping complex diseases with its application to obstructive sleep apnea. *Scientific Reports*, 11 (1):1–15, 2021. ISSN 20452322. doi: 10.1038/s41598-021-84003-4.

[147] Janez Stare and Delphine Maucort-Boulch. Odds ratio, hazard ratio and relative risk. *Metodoloski zvezki*, 13(1):59, 2016.

[148] Jinwei Hu, Kewei Zhu, Sibo Cheng, Nina M. Kovalchuk, Alfred Soulsby, Mark J.H. Simmons, Omar K. Matar, and Rossella Arcucci. Explainable ai models for predicting drop coalescence in microfluidics device. *Chemical Engineering Journal*, 481:148465, 2024. ISSN 1385-8947. doi: https://doi.org/10.1016/j.cej.2023.148465.

[149] Dayou Chen, Sibo Cheng, Jinwei Hu, Matthew Kasoar, and Rossella Arcucci. Explainable global wildfire prediction models using graph neural networks. *arXiv preprint arXiv:2402.07152*, 2024. doi: 10.48550/arXiv.2402.07152.

[150] Nicola Altini, Emilia Puro, Maria Giovanna Taccogna, Francescomaria Marino, Simona De Summa, Concetta Saponaro, Eliseo Mattioli, Francesco Alfredo Zito, and Vitoantonio Bevilacqua. Tumor cellularity assessment of breast histopathological slides via instance segmentation and pathomic features explainability. *Bioengineering*, 10(4), 2023. ISSN 2306-5354. doi: 10.3390/bioengineering10040396.

[151] Xiaochen Qi, Yangyang Ge, Ao Yang, Yuanxin Liu, Qifei Wang, and Guangzhen Wu. Potential value of mitochondrial regulatory pathways in the clinical application of clear cell renal cell carcinoma: a machine learning-based study. *Journal of Cancer Research and Clinical Oncology*, (0123456789), 2023. ISSN 14321335. doi: 10.1007/s00432-023-05393-8.

[152] Gian Maria Zaccaria, Nicola Altini, Giuseppe Mezzolla, Maria Carmela Vegliante, Marianna Stranieri, Susanna Anita Pappagallo, Sabino Ciavarella, Attilio Guarini, and Vitoantonio Bevilacqua. Surviae: Survival prediction with interpretable autoencoders from diffuse large b-cells lymphoma gene expression data. *Computer Methods and Programs in Biomedicine*, 244:107966, 2024. ISSN 0169-2607. doi: 10.1016/j.cmpb.2023.107966.

[153] B Srinidhi and M S Bhargavi. An XAI Approach to Predictive Analytics of Pancreatic Cancer. *2023 International Conference on Information Technology (ICIT)*, pages 343–348, 2023. doi: 10.1109/ICIT58056.2023.10225991.

[154] Yan Zuo, Qiufang Liu, Nan Li, Panli Li, Jianping Zhang, and Shaoli Song. Optimal 18f-fdg pet/ct radiomics model development for predicting egfr mutation status and prognosis in lung adenocarcinoma: a multicentric study. *Frontiers in Oncology*, 13, 2023. ISSN 2234-943X. URL https://www.doi.org/10.3389/fonc.2023.1173355.

[155] Krishnaraj Chadaga, Srikanth Prabhu, Niranjana Sampathila, and Rajagopala Chadaga. Healthcare Analytics A machine learning and explainable artificial intelligence approach

for predicting the efficacy of hematopoietic stem cell transplant in pediatric patients. *Healthcare Analytics*, 3(February):100170, 2023. ISSN 2772-4425. doi: 10.1016/j.health.2023.100170.

[156] Rasheed Omobolaji Alabi, Mohammed Elmusrati, Ilmo Leivo, Alhadi Almangush, and Antti A Mäkitie. Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Scientific Reports*, pages 1–14, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-35795-0.

[157] Junfeng Peng, Kaiqiang Zou, Mi Zhou, Yi Teng, Xiongyong Zhu, Feifei Zhang, and Jun Xu. An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. *Journal of Medical Systems*, 45(5):61, Apr 2021. ISSN 1573-689X. doi: 10.1007/s10916-021-01736-5.

[158] Enzhao Zhu, Linmei Zhang, Jiayi Wang, Chunyu Hu, Huiqing Pan, Weizhong Shi, Ziqin Xu, Pu Ai, Dan Shan, and Zisheng Ai. Deep learning-guided adjuvant chemotherapy selection for elderly patients with breast cancer. *Breast Cancer Research and Treatment*, Jan 2024. ISSN 1573-7217. doi: 10.1007/s10549-023-07237-y.

[159] Roberto Passera, Sofia Zompi, Jessica Gill, and Alessandro Busca. Explainable Machine Learning (XAI) for Survival in Bone Marrow Transplantation Trials: A Technical Report. *BioMedInformatics*, 3(3):752–768, 2023. doi: 10.3390/biomedinformatics3030048.

[160] Hubert Baniecki, Bartlomiej Sobieski, Przemysław Bombiński, Patryk Szatkowski, and Przemysław Biecek. Hospital Length of Stay Prediction Based on Multi-modal Data Towards Trustworthy Human-AI Collaboration in Radiomics. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13897 LNAI:65–74, 2023. ISSN 16113349. doi: 10.1007/978-3-031-34344-5_9.

[161] Henri Korkalainen, Timo Leppanen, Brett Duce, Samu Kainulainen, Juhani Aakko, Akseli Leino, Laura Kalevo, Isaac O. Afara, Sami Myllymaa, and Juha Toyras. Detailed Assessment of Sleep Architecture with Deep Learning and Shorter Epoch-to-Epoch Duration Reveals Sleep Fragmentation of Patients with Obstructive Sleep Apnea. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2567–2574, 2021. ISSN 21682208. doi: 10.1109/JBHI.2020.3043507.

[162] Riku Huttunen, Timo Leppänen, Brett Duce, Arie Oksenberg, Sami Myllymaa, Juha Töyräs, and Henri Korkalainen. Assessment of obstructive sleep apnea-related sleep fragmentation utilizing deep learning-based sleep staging from photoplethysmography. *Sleep*, 44(10):1–10, 2021. ISSN 15509109. doi: 10.1093/sleep/zsab142.

[163] Pierre Delanaye, Elke Schaeffner, Natalie Ebert, Etienne Cavalier, Christophe Mariat, Jean-Marie Krzesinski, and Olivier Moranne. Normal reference values for glomerular filtration rate: what do we really know? *Nephrology Dialysis Transplantation*, 27(7):2664–2672, 07 2012. ISSN 0931-0509. doi: 10.1093/ndt/gfs265.

[164] A healthy lifestyle - who recommendations. URL https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations.

[165] Mortality in obstructive sleep apnea syndrome (osas) and overlap syndrome (os): The role of nocturnal hypoxemia and cpap compliance. *Sleep Medicine*, 112:96–103, 12 2023. ISSN 1389-9457. doi: 10.1016/J.SLEEP.2023.10.011.

[166] Paul E. Peppard, Terry Young, Jodi H. Barnet, Mari Palta, Erika W. Hagen, and Khin Mae Hla. Increased Prevalence of Sleep-Disordered Breathing in Adults. *American Journal of Epidemiology*, 177(9):1006–1014, 04 2013. ISSN 0002-9262. doi: 10.1093/aje/kws342.

[167] Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31:1–16, 11 2021. ISSN 15731375. doi: 10.1007/S11222-021-10057-Z/FIGURES/9.

[168] Zamarrón E., Jaureguizar A., García-Sánchez A. Díaz-Cambriles T., Alonso-Fernández A., Lores V., Mediano O., Rodríguez-Rodríguez P., Cabello-Pelegrín S., Morales-Ruíz E., Ramírez-Prieto MT., Valiente-Díaz MI., Gómez-García T., and García-Río F. Obstructive sleep apnea is associated with impaired renal function in patients with diabetic kidney disease. *Scientific Reports 2021 11:1*, 11:1–11, 3 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-85023-w.

[169] Walter T. McNicholas. Copd-osa overlap syndrome: Evolving evidence regarding epidemiology, clinical consequences, and management. *Chest*, 152:1318–1326, 12 2017. ISSN 19313543. doi: 10.1016/j.chest.2017.04.160.

[170] Amir M. Khan, Santoro Ashizawa, Violetta Hlebowicz, and David W. Appel. Anemia of aging and obstructive sleep apnea. *Sleep & breathing = Schlaf & Atmung*, 15:29–34, 1 2011. ISSN 1522-1709. doi: 10.1007/S11325-010-0326-7.

[171] Linjie Cheng, Hai Guo, Zhenlian Zhang, Yangyang Yao, and Qiaoling Yao. Obstructive sleep apnea and incidence of malignant tumors: a meta-analysis. *Sleep Medicine*, 84: 195–204, 8 2021. ISSN 1389-9457. doi: 10.1016/J.SLEEP.2021.05.029.

[172] Shazia Jehan, Alyson K Myers, Ferdinand Zizi, Seithikurippu R Pandi-Perumal, Girardin Jean-Louis, and Samy I McFarlane. Obesity, obstructive sleep apnea and type 2 diabetes mellitus: Epidemiology and pathophysiologic insights. *Sleep medicine and disorders : international journal*, 2:52, 6 2018. ISSN 2577-8285. doi: 10.15406/smdij.2018.02.00045.

[173] Daniel C. Cattran, Rosanna Coppo, H. Terence Cook, John Feehally, Ian S.D. Roberts, Stéphan Troyanov, Charles E. Alpers, Alessandro Amore, Jonathan Barratt, Francois Berthoux, Stephen Bonsib, Jan A. Bruijn, Vivette D'Agati, Giuseppe D'Amico, Steven Emancipator, Francesco Emma, Franco Ferrario, Fernando C. Fervenza, Sandrine

Florquin, Agnes Fogo, Colin C. Geddes, Hermann Josef Groene, Mark Haas, Andrew M. Herzenberg, Prue A. Hill, Ronald J. Hogg, Stephen I. Hsu, J. Charles Jennette, Kensuke Joh, Bruce A. Julian, Tetsuya Kawamura, Fernand M. Lai, Chi Bon Leung, Lei Shi Li, Philip K.T. Li, Zhi Hong Liu, Bruce MacKinnon, Sergio Mezzano, F. Paolo Schena, Yasuhiko Tomino, Patrick D. Walker, Haiyan Wang, Jan J. Weening, Nori Yoshikawa, and Hong Zhang. The oxford classification of iga nephropathy: Rationale, clinicopathological correlations, and classification. *Kidney International*, 76:534–545, 9 2009. ISSN 00852538. doi: 10.1038/ki.2009.243.

[174] I.S.D. Roberts, H.T. Cook, S. Troyanov, C.E. Alpers, A. Amore, J. Barratt, F. Berthoux, S. Bonsib, J.A. Bruijn, D.C. Cattran, R. Coppo, V. D'Agati, G. D'Amico, S. Emancipator, F. Emma, J. Feehally, F. Ferrario, F.C. Fervenza, S. Florquin, A. Fogo, C.C. Geddes, H.-J. Groene, M. Haas, A.M. Herzenberg, P.A. Hill, R.J. Hogg, S.I. Hsu, J.C. Jennette, K. Joh, B.A. Julian, T. Kawamura, F.M. Lai, L.-S. Li, P.K.T. Li, Z.-H. Liu, B. MacKinnon, S. Mezzano, F.P. Schena, Y. Tomino, P.D. Walker, H. Wang, J.J. Weening, N. Yoshikawa, and H. Zhang. The oxford classification of iga nephropathy: Pathology definitions, correlations, and reproducibility. 76(5):546–556, 2009. ISSN 00852538. doi: 10.1038/ki.2009.168. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-68949206247&doi=10.1038%2fki.2009.168&partnerID=40&md5=a1479e9a2db6d4374a19545e45368472.

[175] H. Trimarchi, J. Barratt, D.C. Cattran, H.T. Cook, R. Coppo, M. Haas, Z.-H. Liu, I.S.D. Roberts, Y. Yuzawa, H. Zhang, J. Feehally, C.E. Alpers, A.M. Asunis, S. Barbour, J.U. Becker, J. Ding, G. Espino, F. Ferrario, A. Fogo, M. Hladunewich, K. Joh, R. Katafuchi, J. Lv, K. Matsuzaki, K. Nakanishi, A. Pani, R. Perera, A. Perkowska-Ptasinska, H. Reich, Y. Shima, M.F. Soares, Y. Suzuki, K. Takahashi, S. Troyanov, J.C. Verhave, S. Wang, J. Weening, R. Wyatt, N. Yoshikawa, and C. Zeng. Oxford classification of iga nephropathy 2016: an update from the iga nephropathy classification working group. 91(5):1014–1021, 2017. ISSN 00852538. doi: 10.1016/j.kint.2017.02.003. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85015786890&doi=10.1016%2fj.kint.2017.02.003&partnerID=40&md5=f8da8098fdc3cb67fa61b45646a74276.

[176] P. Chagas, L. Souza, I. Araújo, N. Aldeman, A. Duarte, M. Angelo, W.L.C. dos Santos, and L. Oliveira. Classification of glomerular hypercellularity using convolutional features and support vector machine. 103, 2020. ISSN 09333657. doi: 10.1016/j.artmed.2020.101808. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85078406097&doi=10.1016%2fj.artmed.2020.101808&partnerID=40&md5=7f189111ae7de5001cfcd19a3e0bc3e9.

[177] E. Uchino, K. Suzuki, N. Sato, R. Kojima, Y. Tamada, S. Hiragi, H. Yokoi, N. Yugami, S. Minamiguchi, H. Haga, M. Yanagita, and Y. Okuno. Classification of glomerular pathological findings using deep learning and nephrologist–ai collective intelligence approach. 141, 2020. ISSN 13865056. doi: 10.1016/j.ijmedinf.2020.104231. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85087958098&doi=10.1016%2fj.ijmedinf.2020.104231&partnerID=40&md5=4b71b4c8b97b2da5e1a3a85b69014fa2.

[178] G. Bueno, M.M. Fernandez-Carrobles, L. Gonzalez-Lopez, and O. Deniz. Glomerulosclerosis identification in whole slide images using semantic segmentation. 184, 2020. ISSN 01692607. doi: 10.1016/j.cmpb.2019.105273. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076980250&doi=10.1016%2fj.cmpb.2019.105273&partnerID=40&md5=6aea847777012324282eaf2c7a80a766.

[179] L. Jiang, W. Chen, B. Dong, K. Mei, C. Zhu, J. Liu, M. Cai, Y. Yan, G. Wang, L. Zuo, and H. Shi. A deep learning-based approach for glomeruli instance segmentation from multistained renal biopsy pathologic images. 191(8): 1431–1441, 2021. ISSN 00029440. doi: 10.1016/j.ajpath.2021.05.004. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110683344&doi=10.1016%2fj.ajpath.2021.05.004&partnerID=40&md5=738308575be013c17f062035cc838023.

[180] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

[181] Nicola Altini, Giacomo Donato Cascarano, Antonio Brunetti, Irio De Feudis, Domenico Buongiorno, Michele Rossini, Francesco Pesce, Loreto Gesualdo, and Vitoantonio Bevilacqua. A deep learning instance segmentation approach for global glomerulosclerosis assessment in donor kidney biopsies. *Electronics*, 9(11):1768, 2020. ISSN 2079-9292. doi: 10.3390/electronics9111768.

[182] A. Jha, H. Yang, R. Deng, M.E. Kapp, A.B. Fogo, and Y. Huo. Instance segmentation for whole slide imaging: End-to-end or detect-then-segment. 8(1), 2021. ISSN 23294302. doi: 10.1117/1.JMI.8.1.014001. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101923751&doi=10.1117%2f1.JMI.8.1.014001&partnerID=40&md5=e769628f3d872bed9502e2f7b8d8ca87.

[183] Massimo Salvi, Filippo Molinari, U Rajendra Acharya, Luca Molinaro, and Kristen M Meiburger. Impact of stain normalization and patch selection on the performance of convolutional neural networks in histological breast and prostate cancer classification. *Computer Methods and Programs in Biomedicine Update*, 1:100004, 2021.

[184] T. de Bel, J. M. Bokhorst, J. van der Laak, and G. Litjens. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Medical Image Analysis*, 70(10200):34–41, 2021.

[185] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahttps Jemal. Cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 68(6):394–424, 2018.

[186] Marzieh Esmaeili, Seyed Mohammad Ayyoubzadeh, Zohreh Javanmard, and Sharareh R Niakan Kalhori. A systematic review of decision aids for mammography screening:

Focus on outcomes and characteristics. *International Journal of Medical Informatics*, 149:104406, 2021.

[187] Zahra Rezaei. A review on image-based approaches for breast cancer detection, segmentation, and classification. *Expert Systems with Applications*, 182:115204, 2021.

[188] Supriya Kulkarni, Vivianne Freitas, and Derek Muradali. Digital breast tomosynthesis: potential benefits in routine clinical practice. *Canadian Association of Radiologists Journal*, page 08465371211025229, 2021.

[189] Mingxiang Wu and Jie Ma. Association between imaging characteristics and different molecular subtypes of breast cancer. *Academic Radiology*, 24(4):426–434, 2017.

[190] Si-Qing Cai, Jian-Xiang Yan, Qing-Shi Chen, Mei-Ling Huang, and Dong-Lu Cai. Significance and application of digital breast tomosynthesis for the bi-rads classification of breast cancer. *Asian Pacific Journal of Cancer Prevention*, 16(9):4109–4114, 2015.

[191] E Sickles, CJ D'Orsi, and LW Bassett. Acr bi-rads® mammography. acr bi-rads® atlas, breast imaging reporting and data system. american college of radiology 2013.

[192] Su Hyun Lee, Jung Min Chang, Sung Ui Shin, A Jung Chu, Ann Yi, Nariya Cho, and Woo Kyung Moon. Imaging features of breast cancers on digital breast tomosynthesis according to molecular subtype: association with breast cancer detection. *The British Journal of Radiology*, 90(1080):20170470, 2017.

[193] Siqing Cai, Miaomiao Yao, Donglu Cai, Jianxiang Yan, Meiling Huang, Lisheng Yan, and Huirong Huang. Association between digital breast tomosynthesis and molecular subtypes of breast cancer. *Oncology Letters*, 17(3):2669–2676, 2019.

[194] Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, and Qingling Sun. Deep learning for image-based cancer detection and diagnosis- a survey. *Pattern Recognition*, 83:134–149, 2018.

[195] Vitoantonio Bevilacqua. Three-dimensional virtual colonoscopy for automatic polyps detection by artificial neural network approach: New tests on an enlarged cohort of polyps. *Neurocomputing*, 116:62–75, 2013.

[196] Vitoantonio Bevilacqua, Antonio Brunetti, Gianpaolo Francesco Trotta, Giovanni Dimauro, Katarina Elez, Vito Alberotanza, and Arnaldo Scardapane. A novel approach for hepatocellular carcinoma detection and classification based on triphasic ct protocol. In *2017 IEEE congress on evolutionary computation (CEC)*, pages 1856–1863. IEEE, 2017.

[197] Vitoantonio Bevilacqua, Nicola Altini, Berardino Prencipe, Antonio Brunetti, Laura Villani, Antonello Sacco, Chiara Morelli, Michele Ciaccia, and Arnaldo Scardapane. Lung segmentation and characterization in covid-19 patients for assessing pulmonary thromboembolism: An approach based on deep learning and radiomics. *Electronics*, 10 (20):2475, 2021.

[198] Gunjan Chugh, Shailender Kumar, and Nanhay Singh. Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cognitive Computation*, 13(6): 1451–1470, 2021.

[199] Essam H Houssein, Marwa M Emam, Abdelmgeid A Ali, and Ponnuthurai Nagaratnam Suganthan. Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*, 167:114161, 2021.

[200] Jiande Wu and Chindo Hicks. Breast cancer type classification using machine learning. *Journal of personalized medicine*, 11(2):61, 2021.

[201] SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, and Joel JP C Rodrigues. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 125:1–6, 2019.

[202] Samir S Yadav and Shivajirao M Jadhav. Thermal infrared imaging based breast cancer diagnosis using machine learning techniques. *Multimedia Tools and Applications*, pages 1–19, 2020.

[203] Dina A Ragab, Omneya Attallah, Maha Sharkas, Jinchang Ren, and Stephen Marshall. A framework for breast cancer classification using multi-dcnns. *Computers in Biology and Medicine*, 131:104245, 2021.

[204] Mohammad M Ghiasi and Sohrab Zendehboudi. Application of decision tree-based ensemble learning in the classification of breast cancer. *Computers in Biology and Medicine*, 128:104089, 2021.

[205] Yu-Dong Zhang, Suresh Chandra Satapathy, David S Guttery, Juan Manuel Górriz, and Shui-Hua Wang. Improved breast cancer classification through combining graph convolutional network and convolutional neural network. *Information Processing & Management*, 58(2):102439, 2021.

[206] Raouia Mokni, Norhene Gargouri, Alima Damak, Dorra Sellami, Wiem Feki, and Zeineb Mnif. An automatic computer-aided diagnosis system based on the multimodal fusion of breast cancer (mf-cad). *Biomedical Signal Processing and Control*, 69:102914, 2021.

[207] Jiaqiao Shi, Aleksandar Vakanski, Min Xian, Jianrui Ding, and Chunping Ning. Emt-net: Efficient multitask network for computer-aided diagnosis of breast cancer. *arXiv preprint arXiv:2201.04795*, 2022.

[208] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S Gene Kim, Linda Moy, Kyunghyun Cho, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical image analysis*, 68:101908, 2021.

[209] Nasibeh Saffari, Hatem A Rashwan, Mohamed Abdel-Nasser, Vivek Kumar Singh, Meritxell Arenas, Eleni Mangina, Blas Herrera, and Domenec Puig. Fully automated breast density segmentation and classification using deep learning. *Diagnostics*, 10(11): 988, 2020.

[210] Neeraj Shrivastava and Jyoti Bharti. Breast tumor detection and classification based on density. *Multimedia Tools and Applications*, 79(35):26467–26487, 2020.

[211] DB Kopans. Mammography, breast imaging. *JB Lippincott Company, Philadelphia*, 30:34–59, 1989.

[212] Pavel Kisilev, Eli Sason, Ella Barkan, and Sharbell Hashoul. Medical image description using multi-task-loss cnn. In *Deep learning and data labeling for medical applications*, pages 121–129. Springer, 2016.

[213] Vivek Kumar Singh, Hatem A Rashwan, Santiago Romani, Farhan Akram, Nidhi Pandey, Md Mostafa Kamal Sarker, Adel Saleh, Meritxell Arenas, Miguel Arquez, Domenec Puig, et al. Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network. *Expert Systems with Applications*, 139:112855, 2020.

[214] Seong Tae Kim, Hakmin Lee, Hak Gu Kim, and Yong Man Ro. Icadx: interpretable computer aided diagnosis of breast masses. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 1057522. International Society for Optics and Photonics, 2018.

[215] Edward A Sickles, Carl J D'Orsi, Lawrence W Bassett, Catherine M Appleton, Wendie A Berg, Elizabeth S Burnside, et al. Acr bi-rads® atlas, breast imaging reporting and data system. *Reston, VA: American College of Radiology*, pages 39–48, 2013.

[216] Frank Po-Yen Lin, Adrian Pokorny, Christina Teng, and Richard J Epstein. Tepapa: a novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. *Scientific Reports*, 7(1):6918, 2017. doi: 10.1038/s41598-017-07111-0.

[217] Yichi Zhang, Tianrun Cai, Sheng Yu, Kelly Cho, Chuan Hong, Jiehuan Sun, Jie Huang, Yuk-Lam Ho, Ashwin N Ananthakrishnan, Zongqi Xia, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nature protocols*, 14(12):3426–3444, 2019. doi: doi.org/10.1038/s41596-019-0227-6.

[218] Julliette M Buckley, Suzanne B Coopey, John Sharko, Fernanda Polubriaginof, Brian Drohan, Ahmet K Belli, Elizabeth MH Kim, Judy E Garber, Barbara L Smith, Michele A Gadd, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. *Journal of pathology informatics*, 3(1):23, 2012. doi: 10.4103/2153-3539.97788.

[219] Borim Ryu, Eunsil Yoon, Seok Kim, Sejoon Lee, Hyunyoung Baek, Soyoung Yi, Hee Young Na, Ji-Won Kim, Rong-Min Baek, Hee Hwang, et al. Transformation of pathology reports into the common data model with oncology module: use case for colon cancer. *Journal of medical Internet research*, 22(12):e18526, 2020. doi: 10.2196/18526.

[220] Alexander P Glaser, Brian J Jordan, Jason Cohen, Anuj Desai, Philip Silberman, and Joshua J Meeks. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. *JCO clinical cancer informatics*, 2:1–8, 2018. doi: 10.1200/CCI.17.00128.

[221] Anobel Y Odisho, Mark Bridge, Mitchell Webb, Niloufar Ameli, Renu S Eapen, Frank Stauf, Janet E Cowan, Samuel L Washington III, Annika Herlemann, Peter R Carroll, et al. Automating the capture of structured pathology data for prostate cancer clinical care and research. *JCO clinical cancer informatics*, 3:1–8, 2019. doi: 10.1200/CCI.18.00084.

[222] S. K. Srivatsa G. Parthiban. Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems*, 3(7):25–30, August 2012. ISSN 2249-0868. doi: 10.5120/ijais12-450593. URL https://www.ijais.org/archives/volume3/number7/244-0593/.

[223] Aiswarya Iyer, Jeyalatha S, and Ronak Sumbaly. Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process*, 5(1):01–14, January 2015. ISSN 2230-9608. doi: 10.5121/ijdkp.2015.5101. URL http://dx.doi.org/10.5121/ijdkp.2015.5101.

[224] Vasilii Khammad, Jose Javier Otero, Yolanda Cabello Izquierdo, Francisco Garagorry Guerra, Aline P Becker, Nataliy Kharchenko, and Gadzhimurad Zapirov. Application of machine learning algorithms for the diagnosis of primary brain tumors., 2020.

[225] Valentina Gaidano, Valerio Tenace, Nathalie Santoro, Silvia Varvello, Alessandro Cignetti, Giuseppina Prato, Giuseppe Saglio, Giovanni De Rosa, and Massimo Geuna. A clinically applicable approach to the classification of b-cell non-hodgkin lymphomas with flow cytometry and machine learning. *Cancers*, 12(6), 2020. ISSN 2072-6694. doi: 10.3390/cancers12061684. URL https://www.mdpi.com/2072-6694/12/6/1684.

[226] Paul A. Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G. Conde. Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, 42:377–381, 4 2009. ISSN 1532-0464. doi: 10.1016/J.JBI.2008.08.010.

[227] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[228] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: A pre-trained biomedical language representation model

for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019. doi: 10.1093/bioinformatics/btz682.

[229] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

[230] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained language model for scientific text. In *EMNLP*, 2019.

[231] Jake Lever, Eric Y. Zhao, Jasleen Grewal, Martin R. Jones, and Steven J. Jones. Cancermine: A literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. 2018. doi: 10.1101/364406.

[232] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014. doi: 10.1016/j.jbi.2013.12.006.