Politecnico di Torino

Politecnico di Torino

Master's degree in COMPUTER ENGINEERING

A.a. 2024/2025

Graduation Session March 2026

Multimodal Learning with Missing Data in Healthcare

Supervisors:
Paolo Garza
Maria A. Zuluaga

Candidate:
Mattia Viglino

A Mamma, che mi ha insegnato a credere in me e a distinguermi.
A Papà, che mi ha trasmesso la determinazione e il duro lavoro.
A chi mi ha sostenuto, ogni gesto è stato importante.

Acknowledgements

I would like to thank the AI4Health research group at EURECOM for their continuous support and for the stimulating discussions throughout the development of this thesis. In particular, I would like to thank my supervisor, Maria A. Zuluaga, for the excellent guidance and for believing in me and in my work. Thanks Vincenzo Marcianó for the valuable advice and for being constantly available, patient and kind. Thanks also to Prof. Paolo Garza for his clarifications and availability.

Grazie Mamma e Papà per i vostri preziosi insegnamenti, per avermi sempre sostenuto e per non avermi mai fatto mancare nulla. Senza di voi, nulla sarebbe stato possibile. Un ringraziamento speciale va alla mia ragazza e a tutti i miei amici, che mi hanno motivato e hanno reso questo percorso indimenticabile.

Abstract

Missing data modalities are common in real-world healthcare datasets. They make supervised learning challenging as traditional algorithms cannot be applied directly. Although widely used in practice, imputing missing modalities before supervised learning relies on complex and computationally costly strategies, which can introduce bias in the data and impact subsequent prediction models. Therefore, their use can be risky in certain sensitive applications such as healthcare. To palliate these limitations, this thesis studies the usage of imputation-free techniques and proposes a novel algorithm, MMARE, which consists of a lightweight end-to-end imputation-free strategy designed for supervised learning with missing modalities that can handle inputs of varying dimensions. To achieve this, we introduce a Missing Aware Conditioning module that explicitly conditions the model on the missingness pattern of each patient and a fusion mechanism that efficiently combines available modalities into a unified representation. We experimentally demonstrate the advantages of our method in a 12-month Survival prediction task, a Sex prediction task and the novel Brain Lifespan Epoch prediction task. We demonstrate that our strategy is robust to high rates of missing data and its flexibility across different datasets and tasks.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The integration of **Artificial Intelligence (AI)** and Machine Learning (ML) into clinical workflows represents a significant opportunity to improve healthcare challenges. In many practical settings, AI-based decision support can help clinicians prioritize patients, reduce time-to-diagnosis and enhance consistency in screening, diagnosis and prognosis [1, 2]. In parallel, the growing availability of healthcare data, ranging from medical imaging to electronic health records, laboratory measurements and signals from wearable sensors has accelerated the development of data-driven methods designed to assist clinical decision-making [3]. However, significant concerns remain about the *reliability* and *robustness* of these AI systems in real-world clinical scenarios. Unlike controlled benchmarks, clinical data are heterogeneous, collected under diverse protocols, affected by acquisition noise or artefacts and often incomplete. These issues become more pronounced as datasets become larger, more complex and aggregated across multiple sources and institutions. In such conditions, even high-performing models may fail to generalize, undermining trust and limiting deployment in clinical applications. These factors are especially critical where model failures may lead to harmful decisions. Consequently, developing powerful, robust and trustworthy methodologies is an essential step to fully leverage the benefits of AI in health applications.

1.1 Multimodal Learning in Healthcare

In a real-world setting a diagnosis is rarely the result of an isolated observation. Instead, when assessing a patient condition, clinicians typically rely on multiple complementary information sources, providing context, reducing ambiguity and thereby improving decision quality. Medical images, for example, can be interpreted in light of demographic variables, medical history, symptoms, laboratory results, and vital signs to contextualize their judgment and provide more appropriate treatments. A research study in this direction has shown that 85% of radiologists consider clinical context to be crucial for the accurate interpretation of radiological examinations [4].

Emulating clinical practice and incorporating these heterogeneous sources, multimodal systems can simulate the collaboration between experts from different medical areas, creating synergy that is often difficult to achieve in real-life due to logistical barriers. Multimodal models leverage deep learning's generalization capabilities to find complex relationships and patterns that single modal models might not detect, supporting the development of more accurate Clinical Decision Support Systems (CDSS) from screening to prognosis [1, 2].

Motivated by this, an increasing body of research has focused on multimodal [3], aiming to better mimic clinicians' decision processes and to enhance performance across tasks [1, 5]. From a methodological perspective, the fundamental challenge in multimodal learning is to exploit both *complementarity* and *redundancy* of multiple sources [6]. Complementarity enables models to capture information that may not be visible in a single modality, while redundancy can provide resilience when individual sources are noisy, less informative or corrupted. In practical terms, the availability of additional modalities can lead to more robust, informative and reliable predictions, which are crucial properties for clinical decisions. Currently multimodal techniques are widely used in clinical practice and medical research [6], supporting the development of modern deep leaning architectures that improve this paradigm.

However, multimodal learning is not just a matter of simply aggregating inputs. A multimodal pipeline must integrate feature extraction from heterogeneous sources, feature fusion into a shared representation and decision-making within a unified model. This integration is challenging because modalities differ substantially in structure and representation as well as their relationships are neither uniform nor independent. According to [7] modalities are heterogeneous with diverse qualities, structures, mathematical representations and statistical properties. They are interconnected, sharing complementary information, and interacting in different

way, based on the task, context and integration strategy. Designing models that can effectively summarize multimodal information while respecting these constraints remains a central challenge [6].

1.2 Missing Modalities in Real-World Healthcare

Despite the promise of multimodal learning, a major obstacle to real-world deployment is the pervasive presence of missing modalities. When modalities are missing we can refer to the **missing modality problem**. In clinical practice, having varying numbers of modalities per patient is common, as different patients rarely follow the same medical pathway and therefore do not undergo the exact same set of examinations. Furthermore measurements may be omitted because of lack or malfunction of equipment, because a patient's condition prevents a procedure (contraindications), because of patient's refusal [8, 9] or because of time, cost and privacy constraints. Moreover, hospitals and health centers do not collect identical information due to divergent and different practices, protocols and equipment, enhancing heterogeneous and missingness patterns. This issue is exacerbated when each sample comprises many features, since generally the more data are collected, the more data are missing [10]. In addition datasets can be created by aggregating data from different sources, enhancing heterogeneous and missingness patterns [11]. All those issues makes the missing modality problem more a standard clinical scenario rather than an exception.

This creates a fundamental mismatch between the assumptions of many existing multimodal algorithms and the realities of healthcare data, making the learning process more challenging. A large portion of the literature indeed implicitly assumes the completeness of modalities at both training and inference time, which prevents a straightforward use of such methods on real-world healthcare datasets [9]. Additionally many of the existing approaches rely on imputation, which can be challenging and potentially unreliable to apply in clinical scenarios. For this reason, to leverage the benefits of AI in health applications, there is a need of innovative, robust and reliable learning framework that handle missing modalities natively and manage the complexity of medical data.

1.3 Aim of This Thesis

The main objective of this thesis is to develop a robust and efficient framework for multimodal supervised learning with missing modalities, specifically designed for clinical applications. We want to tackle challenges previously explained in section 1.2 and overcome limitations of existing multimodal missing data approaches. Differently from previous HyperMM work [12], that focused on the problem of predicting with missing images we want to extend the problem to the more general scenario of missing modalities. The goal is to enable learning and inference with a varying number of modalities per patient, while preserving performances across different missingness patterns that naturally arise in clinical data. An important focus will be the development of a deep learning method that ideally preserve reliability and robustness under incomplete multimodal observations, which is essential for trustworthy deployment in real-life healthcare applications [9].

1.4 Contributions

The main contributions of this thesis are the following:

Systematic overview on existing methods We review and organize the main families of approaches, state of art methods and common used strategies offering a comprehensive overview for learning from incomplete multimodal data in healthcare settings. The discussion highlights the practical limitations of common used strategies and motivates missingness-aware design choices.

Novel Solution Proposal We propose **MMARE**, a lightweight end-to-end imputation-free framework for multimodal supervised learning with missing modalities. We offer a imputation-free model, not relying on explicit modality reconstruction models, which can be costly and potentially unreliable in sensitive domains like healthcare. This deep learning framework is designed to handle different modality availability at both training and inference time and importantly does not rely on the presence of a specific modality. In line with the fundamental challenges of multimodal learning [6, 7], the proposed method considers the heterogeneous nature of modalities and exploit both complementary and redundant information. Instead of treating missingness as an afterthought, the proposed approach incorporates modality missingness as informative signal in the modelling process, enabling robust decision-making under incomplete observations. A missing-aware conditioning adapt internal computation based on the missing scenario, improving the informativeness of extracted embeddings while keeping inference discriminative. Instead of just conditioning per-modality feature, we adapt feature extraction to

the specific modality-availability pattern of *each patient*. We then introduce a fusion mechanism based on incremental concatenation and MLP-based merging that accounts modality relevance and efficiently combines available modality embeddings into a unified representation. This results in a robust, model-agnostic, task-agnostic solution. To the best of our knowledge, MMARE is the first approach for multimodal learning with missing modality that explicitly introduces patient-wise conditioning and performs pairwise modality fusion. Section 4.2.6 will explain the advantages of using our method.

Empirical validation on challenging healthcare applications We provide an experimental evaluation of the proposed method on healthcare data, comparing it to the State of the Art (SOTA) imputation-free method [12] and other representative approaches identified in literature. We analyze those methods on different datasets regarding different body parts and evaluate robustness across different missing-modality scenarios. Based on [13] we introduce the novel task of Brain Lifespan Epoch prediction that would be a useful detection task for cognitive, behavioral and mental health outcomes.

1.5 Thesis Outline

This chapter summarize main challenges in multimodal learning with missing modalities with a focus in the healthcare scenario as well as highlighting objectives and contributions of this thesis.

Chapter 2 – Medical Imaging Background provides a brief overview of standard procedures of managing clinical data to understand preprocessing and postprocessing strategies of methods. Then it describes different modalities used in this work, highlighting their characteristics, advantages and limitations to understand their clinical importance in datasets.

Chapter 3 – Related Works reviews the literature on multimodal learning with missing modalities in the healthcare domain discussing their limitations and the gaps that our proposed approach aims to address.

Chapter 4 – Method explains the preliminary HyperMM framework and introduce the proposed MMARE multimodal framework in detail, describing the model components and the missing-modality-aware learning strategy.

Chapter 5 – Experiments and Results demonstrate the advantages and the robustness of the approach in different medical tasks and datasets comparing with

state-of-the-art methods. It presents the experimental setup, baselines, results, as well as ablation studies.

Chapter 6 – Conclusions concludes this thesis by summarizing contributions, discussing limitations as well as possible future works and research directions.

Chapter 2

Medical Imaging Background

2.1 Medical Imaging Background

In this section we introduce the essential medical imaging background required to interpret the data representation and the preprocessing choices adopted in this work. In section 2.1.1 we first review the two most common file formats encountered in clinical environments and research pipelines (**DICOM** and **NIfTI**), clarifying how medical images are represented in terms of geometry and coordinates. Then in section 2.1.2 we analyze the main medical imaging modalities used in this thesis (**CT**, **MRI**, **WSI** and **Clinical Tabular Data**).

2.1.1 Medical Image Formats

DICOM format Most medical imaging data after acquisition are stored as 2D image slices in the Digital Imaging and Communications in Medicine (DICOM) format. It revolutionized medical imaging, defining an international standard format. It enable interoperability across acquisition devices (e.g., CT/MR scanners), archives and visualization workstations, leading to more efficient clinical workflows and improved patient care outcomes. Along with the image data, DICOM files provide a standardized representation of patients' metadata, details regarding the imaging procedure, information about the device used for image acquisition and imaging protocol settings [14]. These slices can be processed and analyzed separately (as 2D images) or collectively (as 3D volumes) to extract valuable

information. Medical 3D volumetric images are typically constructed from a stack of 2D slices with a specified thickness. For conventional radiology imaging, a 3D scan is often represented as a *set of instances* grouped into *Series*. Typically when DICOM is converted into a 3D volume is converted into a single file in NIfTI format.

NifTi format The Neuroimaging Informatics Technology Initiative (NifTi) format, widely adopted in neuroimaging and machine-learning research, is a dedicated medical image analysis format where, along with the image, only essential metadata are stored in the header. NIfTI conveniently stores an entire 3D image in a single file (e.g., `.nii` or `.nii.gz`), simplifying dataset handling in computational pipelines. It has standardized fields to encode image geometry describing how the discrete samples of the volume are placed in physical space. The elementary unit of volumes is the *voxel*, the 3D analogue of a pixel. 3D images are stored as regular grids of voxels.

Image geometry and Coordinates To better understand formats and data representations, it is essential to clarify how medical images are represented in terms of geometry and coordinates. A voxel grid is not only defined by intensity values, but also by a *geometry* that specifies how the discrete indices correspond to physical distances. In image coordinates, voxels are referenced by integer indices along the three spatial dimensions, often denoted as (i, j, k). However, for analysis and multi-subject alignment we need to know where each voxel lies in anatomical real world space (typically measured in millimeters) and how the grid is oriented with respect to physical axes. This relationship is described by an *affine mapping*, commonly represented as a 4×4 matrix. The affine efficiently encodes **spacing, orientation and position** information.

The **voxel spacing** also called resolution, is a 3D vector that describes the physical distance between adjacent voxels along each axis. When the spacing is equal along all axes the grid is **isotropic**, otherwise is called **anisotropic**. A common preprocessing step involves **resampling** volumes to a target spacing which often is the isotropic grid.

The **orientation** describes how the image axes are oriented with respect to an anatomical coordinate system. The anatomical space, also called patient coordinate system, describe the standard anatomical human position. It consists of three planes: The axial, the coronal and the sagittal plane. The **axial** plane is parallel to the ground and separates the head (Superior) from the feet (Inferior), the **coronal** plane is perpendicular to the ground and separates the front from (Anterior) the back (Posterior) and finally the **sagittal** plane separates the Left from the Right [15].

LPS (Left, Posterior, Superior) or RAS (Right, Anterior, Superior) are examples of target orientations commonly used. In neuroimaging is common to define this space with respect to the human whose brain is being scanned [15] meanwhile in radiology usually images are visualized respect to the doctor's prospective. Unlike traditional image formats, DICOM and NifTi files, due to their complex structure, require specialized software for analysis and visualization. In this work we leverage ITK-SNAP [16], MONAI [17], 3D Slicer [18] and FreeSurfer [19].

Finally the **position** (origin) refers to the physical location of the first voxel (index (0,0,0)) in the chosen reference space.

In NIfTI, this voxel-to-world geometry is stored in the header via the *qform*/*sform* fields meanwhile in DICOM, the same information is distributed across different tags (e.g., pixel spacing, image orientation, and image position) and can be assembled into an equivalent affine representation during conversion to NIfTI.

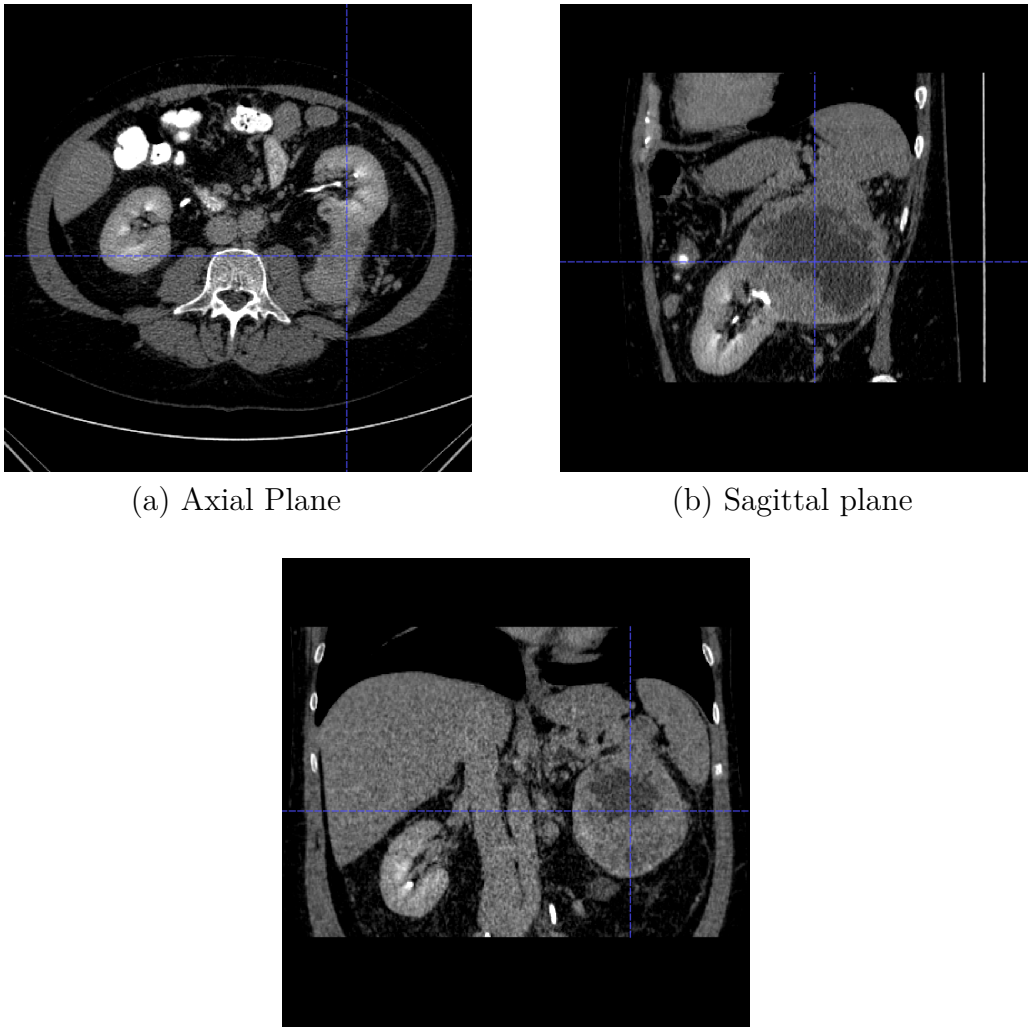
MONAI Framework In this work, preprocessing is implemented using MONAI [17], an open-source and community-supported framework specifically developed for medical imaging. By integrating with PyTorch, MONAI provides utilities and dictionary-based transformations for development and deployment of deep learning models. This results in being a robust, modular and reproducible software becoming an effective choice for advanced computational pipelines in healthcare domain.

2.1.2 Medical Image Modalities

This section reviews the imaging modalities considered in this thesis and clarifies how each modality is represented and preprocessed in the employed datasets.

Computed Tomography (CT) Computed Tomography (CT) is medical imaging technique whose scans provide detailed cross-sectional images of a human body. CT is widely used in routine clinical practice thanks to its high-resolution imaging, wide availability, cost-effectiveness and fast acquisition time. It is a versatile imaging technique primarily employed to identify structural abnormalities, used in respiratory and cardiological studies and plays a useful role in oncology for tumour detection, staging, and treatment planning. Scanners are based on a rotating X-ray source and a corresponding array of detectors that capture the attenuated X-rays after passed through different composition and density body parts (tissues). From the processing of radiographic projections acquired from different angles, multiple subsequent 2D slices are reconstructed and combined to create a 3D image of the scanned area. CT intensities are commonly expressed in Hounsfield Units (HU), which correspond to the measurement of signal attenuation caused by tissue density

with respect to water (which is conventionally set to 0 HU). A key limitation of CT is the exposure of patients to a certain amount of ionising radiation, creating some contraindications for vulnerable subjects such as pregnant women and children. This aspect enhance missingness in real-world datasets, where CT examinations may be unavailable or intentionally avoided. In addition, compared to MRI, CT generally provides lower soft-tissue contrast, which can reduce sensitivity for certain pathologies and anatomical regions. Figure 2.1 shows representative CT views from a clear-cell renal cell carcinoma (ccRCC) patient in the MMIST dataset [11].



(a) Axial Plane

(b) Sagittal plane

(c) Coronal Plane

Figure 2.1: Different CT views of a single CCRCC-ill patient from MMIST dataset [11]. (a) axial, (b) coronal, and (c) sagittal planes are displayed. The blue pointer indicates the location of the tumor.

Magnetic Resonance Imaging (MRI) Magnetic Resonance Imaging (MRI) is one of the most useful techniques used in medical imaging to obtain soft tissues high contrast images in human body [20]. It is effective to analyze a wide range of anatomical structures and for this reason is used variety of tasks. It is particular effective for brain imaging, spinal cord and vascular anatomy, making it extensively used in neuroscience, neuroradiology, as well as in musculoskeletal and cardiovascular applications. It is extensively used to study brain disorders, including Alzheimer's disease, multiple sclerosis and Parkinson's disease. Differently from CT it is a non-ionising imaging technique. The patient is placed in a strong magnetic field which aligns the protons' spin of their body. Radio-frequency pulses disturb the alignment of these protons and realigning with the magnetic field, emits signals that are detected by receiver coils. In practice, MRI acquisition is typically longer than CT and can be perceived as less comfortable due to scanner noise. Additionally it is particularly sensitive to patient motion and to several physics-related artefacts which can degrade image quality or result in unusable sequences. Moreover, MRI may be contraindicated for patients with certain implants such as some pacemakers, cochlear implants or metallic fragments and can be a issue for those suffering of claustrophobia. These factors can lead to missing MRI examinations in clinical cohorts.

There are several MRI type of sequences, each of which exposes a unique feature of the human tissue. They provide complementary information, leading to most accurate diagnoses and treatments. In this work, we consider common clinical sequences described in the following. Figure 2.2 shows an example of axial slices of different MRI modalities from an IXI [21] patent's brain sample.

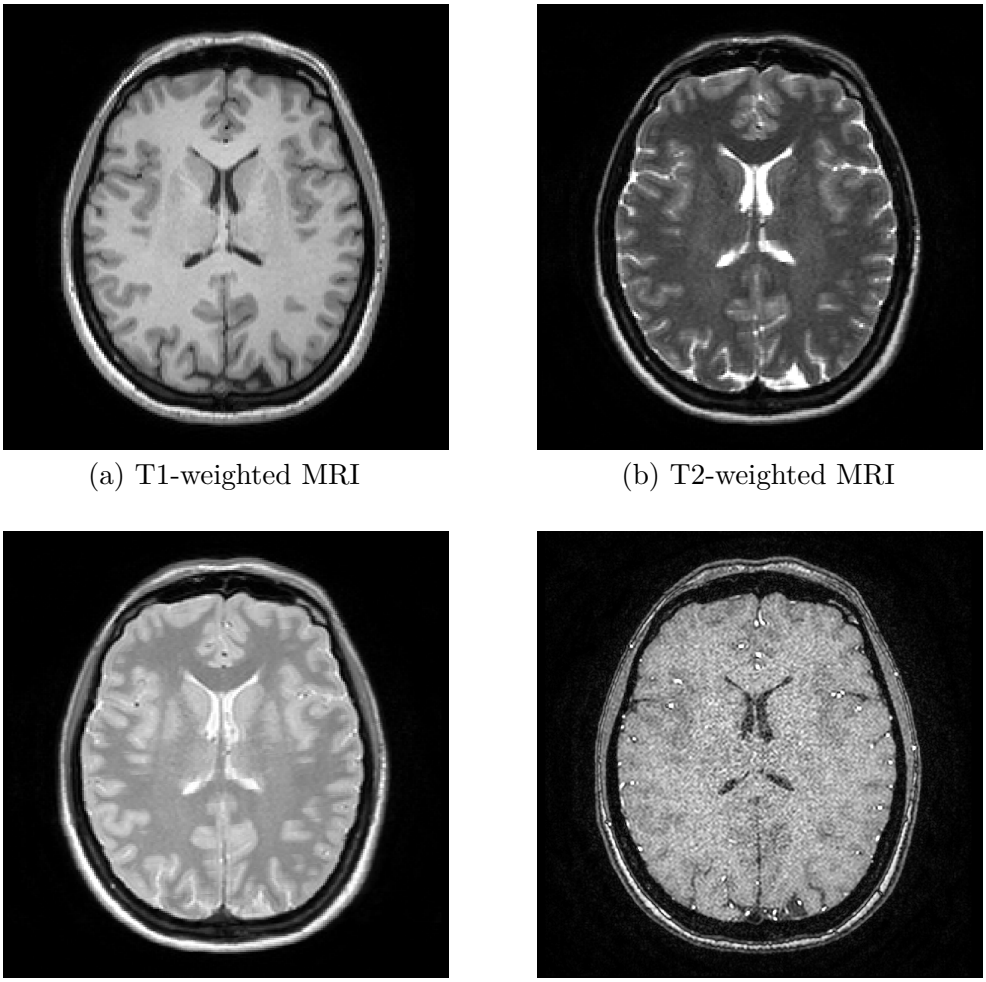
T1-weighted MRI typically highlights fatty tissues and provides clear anatomical detail. Cerebrospinal fluid (CSF) and water appear dark.

T2-weighted MRI enhance signal of fluids. Typically CSF appear bright, which is useful for many clinical findings.

Proton Density(PD)-weighted MRI aims to highlight differences in hydrogen proton density by using acquisition settings that reduce T_1 and T_2 contributions; it is often useful for high-resolution soft-tissue assessment in specific clinical protocols. The signal of water is in the middle between T1 and T2 MRIs.

MR Angiography(MRA) focuses on vascular structures providing detailed images of vessels and blood flow. It may require contrast material depending on the technique and, although this tents to be less toxic than the one used for CT

angiography, it may cause allergic reactions in some patients. For this reason it can be contraindicated in selected patients and may be omitted due to safety considerations, contributing to missingness in datasets.



(a) T1-weighted MRI (b) T2-weighted MRI

(c) PD-weighted MRI (d) MRA

Figure 2.2: Axial views of an healthy subject from the IXI dataset [21]. Different MRI modalities are shown: (a) T1-weighted, (b) T2-weighted, (c) PD-weighted, and (d) MRA. Each modality highlights different tissue characteristics.

WSI Whole Slide Image WSIs are digital representations of histopathology slices containing the tissue samples observed by the pathologist. WSI are typically generated by scanning glass slides containing tissue samples at high resolutions, permitting the exam to be analyzed and shared with computer workstations rather than optical microscopes. These scans capture the entire tissue section at various magnifications, resulting in gigapixel-sized images. The different magnifications are achieved by capturing multiple 2D images at various levels of zoom. These individual 2D images are then stitched together to create a single composite image. These different magnifications are conceptually similar to a 3D image, but processed differently form NifTi images. Figure 2.3 illustrates the pyramid structure of WSIs at different magnification levels. WSIs are commonly stored in specialized file formats, such as SVS, NDPI, or TIFF, which support the storage of large images and multiple resolution levels. These formats also include metadata about the scanning process, magnification levels pyramid layout and other relevant information.

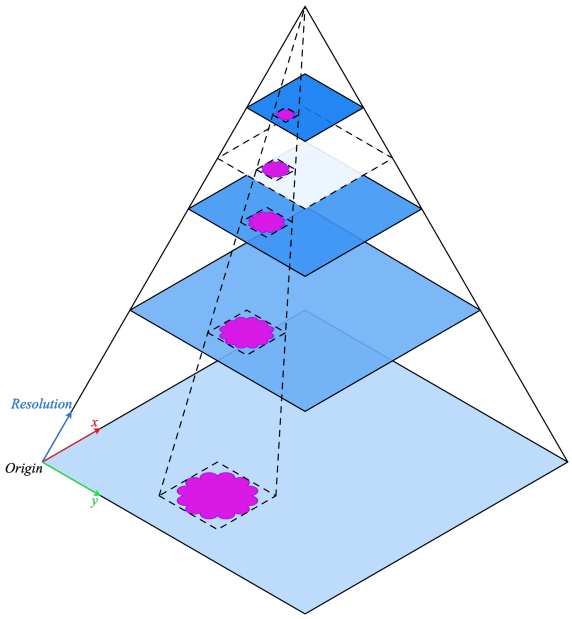


Figure 2.3: Pyramid structure of Whole Slide Images (WSI) at different magnification levels.

Clinical Tabular Data Clinical Tabular Data refers to structured patient information that can be represented in a table. Generally each table row correspond to a subject and each column correspond to a feature. Each entry represents a recorded value for a given patient state at a specific time point, for example at admission or at a follow-up visit. Demographic data, such as patient's age, sex and ethnicity, basic anthropometric information, medication usage, laboratory measurements, genomics data and outcome labels are often provided in this format. Additional entries may include social and behavioral data. These features provide essential context for personalized clinical decision-making and can substantially complement imaging-based evidence. As for other modalities tabular clinical information can be incomplete in real-world cohorts making it not only a complementary modality, but also a realistic source of missingness that must be handled explicitly in multimodal learning settings.

Chapter 3

Related Works

3.1 Multimodal Machine Learning (MML)

Multimodal Machine Learning (MML) it's a machine learning field that aims to build models that process and integrate information from multiple data sources in order to improve predictive performance and robustness [6, 9]. A core challenge in MML is the fusion method that is how modality-specific representations are combined into a single representation for downstream predictions. A common taxonomy organizes fusion strategies into: (i) *early fusion*, where raw inputs or low-level features are merged; (ii) *intermediate fusion*, where latent representations are combined and (iii) *late fusion*, where modality-specific predictors are aggregated [22]. Summation, averaging, max pooling and concatenation are common and straightforward fusion operators. For example [23] shows effectiveness of mean, sum, product and maximum with missing modalities in survival prediction task. In the medical domain, concatenation is widely adopted [2] and has showed its effectiveness in multimodal pipelines [24]. More structured strategies also exist: for instance, [25] proposes a multi-phase iterative fusion procedure that progressively integrates images at different stages. Overall, most practical multimodal methods perform fusion in the latent space, after modality-specific encoding [26].

3.2 Feature Extraction

Recent progress in multimodal clinical prediction is largely enabled by strong modality-specific feature extractors, including convolutional and transformer-based encoders trained on large-scale datasets. In practice, multimodal systems commonly follow a two-stage pattern: each modality is mapped to a latent embedding through a dedicated or partially shared encoder and the resulting embeddings are then fused to optimize a supervised objective. This design accommodates heterogeneous modalities by projecting them into a common latent space where integration becomes feasible. Leveraging pre-trained encoders via transfer learning technique is crucial to mitigate the common data scarcity issues in healthcare applications [11, 1]. However, even with powerful encoders, missing modalities can make fusion challenging and, in some settings, unfeasible. Classical architectures, indeed, are not designed to operate on varying-sized modality subsets and fails to account for missing data.

3.3 MML with Missing Modalities (MLMM)

In clinical practice, multimodal data are frequently incomplete as a patient may lack one or more acquisitions or source. This is mainly due to cost, protocol differences, contraindications or privacy constraints. Missingness may occur only at inference time, only during training or in both phases, with the latter being more realistic for many healthcare scenarios. Straightforward strategies remove missing modalities samples when preprocessing, wasting valuable information of missing patterns and do not address the presence of missing modalities at test time. Solutions that directly handle the missing modality problem can be referred as *multimodal learning with missing modality* (MLMM) approaches, where models operate on *variable-sized* modality subsets rather than full sets inputs [9]. The primary challenge in MLMM is to dynamically and robustly leverage the available modalities while maintaining performance close to the full-modality regime [9].

A practical way to distinguish deep MLMM approaches is to consider *where* missingness is addressed and *how* the overall system adapts to variable modality subsets [9]. From a data processing aspect, some deep MLMM approaches operate directly at the modality level, filling missing information through composition or generation of absent modalities [27]. Other methods operate at the *representation level*, for example enforcing cross-modal alignment, generating missing-modality embeddings from observed ones or directly aggregating available modality representations into a joint embedding [28, 29]. From the strategy design perspective, architecture-focused models adapt internal computation to the observed subset, e.g., via flexible fusion

mechanisms, conditioning on modalities or structured relations between modalities [30, 31]. In contrast, *model-combination* strategies rely on multiple branches trained for different modality cases or combine partial and full predictors according to the available inputs [32, 33].

3.4 Imputation-Based Approaches

Imputation based strategies aim to fill in missing modality in order to reconstruct full modality samples and then apply a downstream predictor. A simple form of imputation is *composition*, where missing modalities are replaced with *dummy* inputs, such as a constant, random values or data copied from similar instances [34, 35, 36]. At representation level, missing information can be approximated by using average embeddings of available modalities [37] or by retrieving nearest neighbors (e.g., via cosine similarity) and transferring features from top-k matching samples [38]. While straightforward to implement, they may introduce distribution shifts, noisy data and easy overfit, degrading performance especially when missingness is severe [9, 34, 35, 36].

More sophisticated approaches rely on generative models to *impute* missing inputs or latent representations [27, 28, 39, 40, 41]. The generated data are then aggregated with available modalities data to form an approximate full-modality sample. Multimodal VAEs for example learn a joint latent space and generate missing modalities by conditioning on observed ones [28]. ShaSpec [42] explicitly models shared and modality-specific components, leveraging modality-shared information for imputation to improve robustness under missingness. The vast majority of existing supervised MLMM solutions first train a generative model using it to impute missing modalities and then train a discriminative model for downstream prediction tasks[12, 43, 44, 45, 27]. Widely used generative paradigms for image synthesis include Generative Adversarial Networks (GANs) [46, 47] which train a generator to produce realistic samples while a discriminator learns to distinguish generated from real data. More recently, diffusion-based have further improved image generation quality [48]. Some work focus on individual modality generation, training one generative model per modality whereas unified generation approaches train a unique model able to produce simultaneously multiple modalities for cross modal completion [9, 27, 49].

Imputation-based approaches have considerable limitations in practice. First, high-capacity generative models can be data-hungry and difficult to train robustly, which is particularly challenging in medical settings where datasets are often limited [50, 9]. Second, they increase storage, computational requirements and system complexity

as the number of modalities grows [9]. Third imputation is not straightforward, since not every modality can be inferred from available ones, especially if the observed sources do not provide enough detailed information [51]. Ill patients for example can be distorted if generated from healthy patients [52]. Generated samples can also introduce artifacts or hallucinations, compromising downstream performance, interpretability and feature importance of subsequent predictors [53], which are crucial aspects to consider in sensitive clinical application. Finally the computational cost and performances of the overall method is heavily linked to the selected imputation strategy; the imputer and the predictor need to be tailored to each other [54, 55], which can be difficult to ensure in real settings [12]. These limitations motivate the research direction through imputation-free approaches that directly operates under missingness.

3.5 Imputation-Free Approaches

Imputation-free MLMM methods aim to perform prediction directly on the observed modality subset, avoiding explicit reconstruction of missing modalities. Those solutions are dedicated and robust to incomplete multimodal data and are particularly relevant in clinical settings. A classic line of work rely on architectures whose fusion mechanism is explicitly defined over the available modalities. HeMIS for example proposes a hetero-modal segmentation framework that aggregates modality-specific statistics (mean and variance), enabling inference under arbitrary modality subsets [56]. However, when multiple modalities are missing, these networks drastically degrade performances [57], motivating more expressive fusion mechanisms.

Beyond pooling, some MLMM methods incorporates attention or gating based mechanisms to dynamically weight or select modality contributions based on relevance or availability [58]. **UniLMMV** encodes each modality into an embedding and uses a masked attention-based aggregator to produce a unified representation that naturally supports variable-sized modality subsets at inference time [36]. While effective, attention-based architectures can be computationally demanding. Mixture-of-experts strategies can also exploit modality-presence patterns to route information through specialized experts [31]. **AdaCoMed** adopts a Mixture-of-Modality-Experts fusion scheme in which different experts specialize in different modality combinations and introduces a collaborative large-small modeling pipeline, aligning representations through contrastive learning and combining predictions with adaptive weighting [59]. Although promising, such routing mechanisms typically require dedicated experts for modality combinations, becoming challenging when modality absence is severe already during training.

Knowledge distillation is another common strategy, transferring knowledge from a larger and sophisticated full-modality network to simpler and faster missing modality one [60]. While can improve missing-modality performance, it often relies on a complex teacher network that is computationally expensive to train and may assume near-complete data. Moreover severe modality absence already during training can limit both teacher quality and the effectiveness of student transfer knowledge [57, 9, 32, 33].

Transformers' sensitivity to missing modalities has also been explored [30, 58], as well as graph based methods such as MUSE [61] which proposes a mutual-consistent graph contrastive learning strategy adaptable to different missing modality patterns [9]. However these approaches tend to have lower efficiency, reduced scalability and higher development complexity compared to simpler alternatives [9]. Mutual information maximization methods optimize similarity metrics between available modalities during training to achieve minimal information loss in missing modality situations [62] but is not so effective with high missing rates since there won't be enough features to reconstruct the missing data [20].

Another important family of MLMM approaches focuses on separating information shared across modalities from modality-specific components. Disentangled representation learning aims to capture independent latent factors that explain different sources of variation and it can be leveraged for modality-level factorization by enforcing independence constraints between modality-specific and shared latent variables [7]. **DRIM** follows this direction by explicitly disentangling shared and unique representations, combining them with a masked fusion mechanism to handle incomplete multimodal clinical data [26]. Similarly, DrFuse [51] factorizes each modality into specific and shared information. Missing modalities are then compensated by the shared component and a dedicated masking matrix ignore modality specific features so that the model do not rely on unobserved information [9]. However, by utilizing common factorized information those models may underutilize modality specific features that capture unique discriminative patterns.

More recently, **HyperMM** [12], an imputation-free state-of-the-art method, introduces an end-to-end framework for supervised multimodal learning with varying-sized inputs, combining a conditional hypernetwork-based feature extractor and a permutation-invariant network operating on the set of extracted embeddings. This permutation invariant design enables the model working under arbitrary modality subsets without requiring imputation or auxiliary strategies described above. However, extending such universal feature extractor to strongly heterogeneous modalities may require non trivial architectural adaptations.

Building on the missing-aware conditioning philosophy of HyperMM [12] and addressing the limitations of existing methods, we propose **MMARE**, a lightweight end-to-end imputation-free framework for multimodal supervised learning with missing modalities. We leverage pre-trained domain-specific encoders to extract rich modality embeddings to handle heterogeneous data types. Since predictive importance of each modality varies across patients and targets [51], our method instead of just conditioning per-modality feature, adapt feature extraction to the specific modality-availability pattern of *each patient*. This *missing-aware conditioning* adapt extracted embeddings on the missing scenario, improving the quality and their inference discriminativeness under missing scenarios. To eliminate the need of imputation or masking, we then introduce a fusion mechanism based on incremental concatenation and MLP-based merging that can handle different number of modalities, considering their relevance for the task and efficiently combining them into a unified representation.

In the next section, we first analyze the methodology of HyperMM as the starting point method and then we present the enhanced version, the MMARE framework.

Chapter 4

Method

4.1 Preliminaries: HyperMM

4.1.1 Overview of HyperMM

HyperMM is an end-to-end multimodal learning framework designed to handle missing modalities at training and test time. It is an imputation-free strategy designed to handle varying-sized inputs. Instead of imputing missing inputs or inserting dummy data and learning to ignore it, HyperMM reformulates multimodal prediction with missing modalities as *learning on sets of variable size* making it more reliable and efficient. This shift makes the model naturally compatible with subjects that have different numbers of available modalities. This work was presented in MICCAI (Medical Image Computing and Computer Assisted Intervention) 2024 workshop and received the *Best Presentation Award*.

We develop a newer code version of the HyperMM framework integrating the MONAI [17] library, aligning the pipeline to modern preprocessing strategies. This version permits to handle more efficiently data, load more image modalities on GPU's RAM space, optimizing it as well as dealing with multi-class tasks.

We introduce a dataset $\mathcal{D}$ of $n \in \mathbb{N}$ independent pairs $\mathcal{D} := \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Given the x, the goal is to predict y. Every sample $x_i := \{x_i^1, \ldots, x_i^j\}$ is a j-modal observation, where each x_i^j represents one of the available modalities. HyperMM introduces a binary missingness indicator $v_i \in \{0,1\}^j$ where $v_i^j = 1$ denotes that

modality j is missing and 0 otherwise. The observed input can be written as

$$x_i^{\text{obs}} = (1 - v_i) \odot x_i + v_i \odot \text{na}, \tag{4.1}$$

where $\odot$ denotes element-wise product and na indicates missing entries. For each subject, HyperMM discards missing entries and represents the observation as a set

$$s_i = \{s_{i,1}, \ldots, s_{i,d}\}, \quad d \leq j, \tag{4.2}$$

where each element is a tuple $s_{i,d} = (x_i^j, m^j)$ containing the observed modality together with its modality identifier. The learning objective becomes predicting y_i from a set s_i whose cardinality varies across subjects.

4.1.2 Two-phase methodology

The framework follows a two-phase strategy. In the first phase a *universal* modality-agnostic feature extractor ϕ is learned and then in the second phase a *permutation-invariant* predictor aggregates the encoded set elements and outputs the final prediction.

Step 1: Universal feature extractor The single universal encoder network ϕ is trained to extract features from any image modality present in the dataset $\mathcal{D}$. HyperMM achieves this by starting from a large pre-trained backbone (e.g., VGG) [63] and adding a trainable final adaptation layer. In this way they obtained general pre-trained features on ImageNet then adapted into medical ones. The weights of the final layer are generate by a conditional hypernetwork h. Hypernetworks are a type of neural network that generate the weights of another network dynamically, conditioned on some input. [64] provides a good overview and their applications use cases. Concretely, given a modality identifier m, $h(m)$ outputs the parameters of the last layer of ϕ, yielding a modality-conditioned encoder $\phi(\cdot \mid m)$. This allows the network to adapt its feature extraction specifically for each modality while still sharing most of the backbone across modalities.

To make the learned representation informative, ϕ is trained jointly with: (i) a unimodal prediction objective using the Cross Entropy (CE) loss and (ii) a feature reconstruction objective to reconstructs the layer features using a Mean Squared Error (MSE) loss. This combination encourages ϕ to produce stable, task-relevant latent features across modalities.
The loss of the Step 1 is then defined as:

$$L_{\text{Step1}} = L_{\text{MSE}} + L_{\text{CE}} \tag{4.3}$$

where:

$$L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} (z_i - z_i^{\text{rec}})^2 \qquad L_{\text{CE}} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{4.4}$$

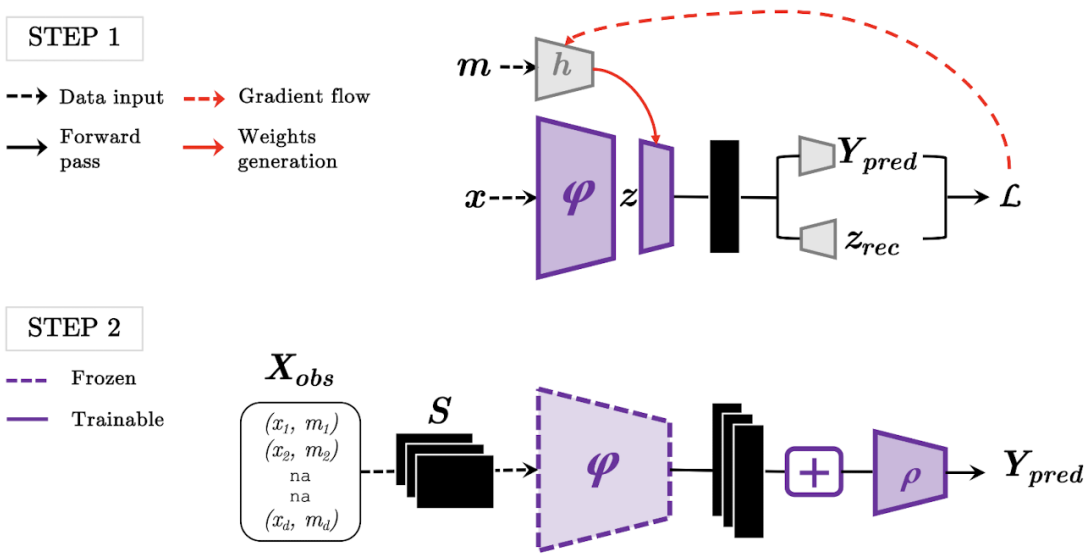


Figure 4.1: HyperMM complete pipeline. The framework consists of two phases: (1) a universal feature extractor is trained through conditional hypernetworks to share information across modalities while still allowing modality-specific adaptation; and (2) a permutation-invariant mid-level fusion architecture that supports varying-sized inputs, enabling robustness to missing modalities.

Step 2: Permutation-invariant fusion and prediction. After Step 1, ϕ is frozen and applied to each observed element $s_{i,d} \in s_i$ independently to obtain embeddings $\phi(s_{i,d}) = \phi(x_i^j \mid m^j)$. These embeddings are then pooled with a permutation-invariant operator (e.g., sum/mean/max) and fed to a classifier ρ. The overall model follows a DeepSets-style [65] decomposition:

$$f(x_i^{\text{obs}}) = \rho\left(\sum_{s_{i,d} \in s_i} \phi(s)\right), \tag{4.5}$$

which maps a variable-size set into a fixed-dimensional representation before prediction. This architecture ensures that the model is invariant to the order of inputs handling missing modalities by construction.
The final learning goal is defined as:

$$\mathcal{L}(\theta) := \mathbb{E}_{(s,y)\in\mathcal{D}}\left[\ell\left(Y, \rho_\theta\left(\sum_{s_d \in s} \varphi(s_d)\right)\right)\right] \tag{4.6}$$

where ℓ is the cross entropy defined as in equation 4.4
Figure 4.1 clarifies the two phases architecture and shows the complete HyperMM pipeline.

4.1.3 Advantages of HyperMM

HyperMM's design yields several practical advantages. By avoiding the imputation, it significantly reduces training time and computational complexity. It eliminates the risk of model degradation caused by poorly generated data or the introduction of noise through dummy inputs. It is agnostic to *which* modality is missing, model agnostic and task agnostic. Finally it generalizes not only with classical missing-modality cases but also to scenarios where the number of views per subject varies.

4.1.4 Limitations

The same design choices imply trade-offs. First, set pooling with simple commutative operators (sum/mean/max) can under-express *structured cross-modal interactions* discarding useful information. Is necessary to fully leverage the modality-specific signals and to adaptively assign modality importance, giving more relevance to most informative modalities in the current context. Second if modalities are extremely heterogeneous, conditioning only the last layer may be insufficient to fully extract meaningful embeddings. Finally, transfer learning is such a powerful and useful technique, but feature extractors like VGG pre-trained on ImageNet can be too generic for medical imaging purposes.

On top of the robustness and efficiency of HyperMM in handling missing data, there is an opportunity to further enhance how the model captures the underlying relationships between modalities. By replacing pooled set fusion with a learned fusion mechanism, it can capture richer inter-modality interactions. An HyperMM evolution aims either to maintain the imputation-free benefits, integrating different data sources that will enable to fully take advantage of rich multimodal datasets. In the next section, build on HyperMM's core insight we describe in detail MMARE framework methodology, explicitly modelling missingness without imputation and targeting the above limitations.

4.2 MMARE Framework

4.2.1 Overview of MMARE Method

We introduce a dataset $\mathcal{D}$ of $n \in \mathbb{N}$ independent pairs $\mathcal{D} := \{(x_1, y_1), \ldots, (x_n, y_n)\}$. The goal is to predict y given x. Each $x_i := \{x_i^1, \ldots, x_i^j\}$ corresponds to a j-modal observation, where each x_i^j represents one of the available modalities. The inter-modality missing vector $m_i \in \{0,1\}^j$ is introduced to identify the positions of missing modalities in x_i, such that $m_i^j = 1$ if x_i^j is missing, and 0 otherwise. The observed data of x_i can be expressed as $x_i^{obs} = (1 - m_i) \odot x_i + m_i \odot \texttt{na}$, where $\odot$ identifies the term-by-term product. The learning goal becomes now the prediction of y given x_i^{obs}. Each observation x_i^{obs} is interpreted as a set of unordered elements where all information available in x_i^{obs} is conserved and no new information is added.

We avoid any form of modalities imputation, resulting in variable-sized samples. However, standard and multimodal machine learning models assume fixed size inputs. To remove this constraint, we model each instance as a *set* of observed modalities and learn a function of the form $f = \rho(\varphi(x_i^{obs}))$, which overcome the fixed-dimensional data requirement.

We propose a framework that we call MMARE to implement our method. Figure 4.2 presents an overview of our strategy. We can decompose the framework in four main steps: 1) first we extract modality specific features from any modality present in x_i^{obs}; 2) then our novel Missing-Aware Conditioning module (MAC) modulates modality embeddings based on the inter-modality missing vector m_i; 3) an iterative pairwise module then permits to aggregate the modalities embeddings into a single fused representation; 4) finally a prediction head processes the generated representation to produce the desired task outcome.
The following MMARE method section reads as follows:

Section 4.2.2 discusses about the modality-sensitive feature extraction step;

Section 4.2.3 explains how features are conditioned with the MAC module;

Section 4.2.4 describe how conditioned features are fused in a single patient representation.

Section 4.2.5 explains the task prediction head.

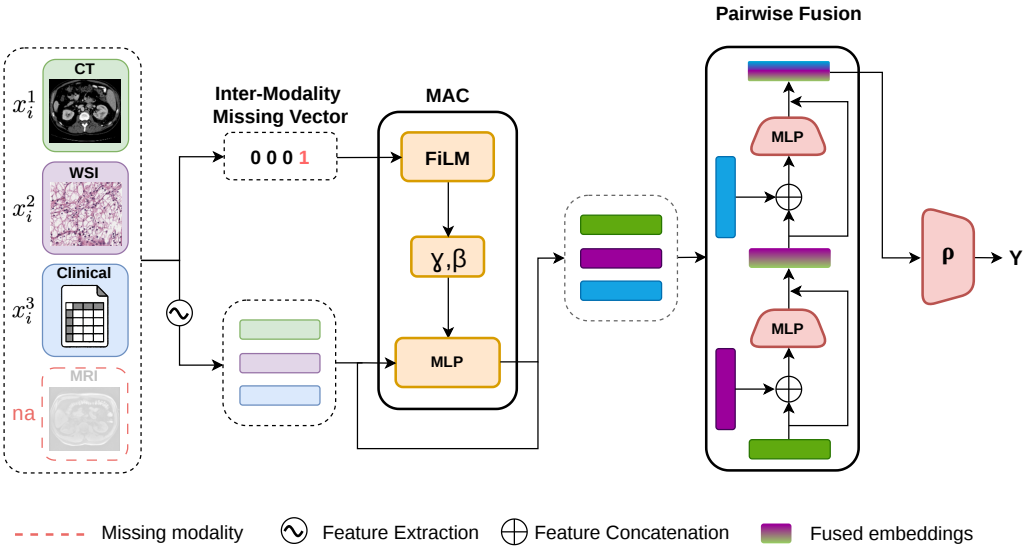


Figure 4.2: MMARE complete pipeline. The framework explicitly condition the extracted features of each patient's modality on the missingness pattern and use a pairwise fusion mechanism to aggregate available modality features into a unified representation for prediction.

4.2.2 Modality Feature Extraction

A fundamental design of our pipeline rely on the modality specific feature extraction before the multimodal fusion. Differently from HyperMM, which relies on a universal feature extractor shared across modality, we adopt strong modality tailored pre-trained encoders. This choice is motivated by two practical considerations.

First, in healthcare, the available modalities can be highly heterogeneous, from radiological volumes to histopathology slides, substantially varying clinical structures. A single universal extractor with just the adapted final layers may struggle to be optimal across such diverse input types. Furthermore by nature some image sources can't be handle by the same backbone base. By contrast, leveraging modality-appropriate pre-trained backbones allows us to exploit training signals that are specific to each data source.

Second, pre-trained encoders provide a strong starting point in medical applications, where labelled data are often limited and costly to obtain [11]. Leveraging pre-trained models we can preserve the general-purpose nature of the representations learned during pre-training on heterogeneous medical data and then reuse

these embeddings for different downstream tasks. It enables the model to reuse robust representations learned from large-scale data, improving sample efficiency, stabilizing training and helping overcoming challenges related to data scarcity. As also highlighted by HyperMM [12] that relies on pre-trained backbones for feature extraction, transfer learning is particularly important in healthcare, where training large models from scratch can easily lead to overfitting [1].

In our setting each available patient modality present in x_i^{obs} is processed by an encoder $\varphi(\cdot)$ producing an embedding

$$\mathbf{e}_i^j = \varphi(\mathbf{x}_i^j) \tag{4.7}$$

where j can be any healthcare modality (e.g., MR and CT volumes, WSI images, clinical data). Different backbones can be used depending on the modality (2D/3D CNNs, ViTs, or pre-trained medical feature extractors). When heterogeneous encoders output different dimensionalities embeddings, we optionally integrate in the pipeline lightweight adapters (e.g., linear projections and a ReLU activation) to map all embeddings to a shared D-dimensional space. When a modality is missing, no embedding is computed and the downstream fusion module operates on the resulting variable-sized set of observed embeddings (see Section 4.2.4). Overall, this design enable flexible integration of modalities with different structures and representations, including non-imaging sources.

Additionally serializing tabular data make MMARE handling intra modality missingness. Our prompt generator permits to handle missing data either by omitting part of the prompt (omit mode) or by explicitly stating the absence of information (explicit mode). In this way Inter modality missingness is handled by the m vector, while Intra modality missingness is handled by nature choosing the prompt type.

Specifically for CT and MRI feature extraction we employ Med3D, a 3D ResNet model pre-trained on 23 different medical imaging datasets (including both CT and MRI) across various tasks [66]. We chose a 3D rather than a 2D backbone to better capture spatial interactions within the volumes. In our framework Med3D is used as a frozen feature extractor to obtain rich clinical feature representations for different tasks. For each input volume, i.e. one modality of one subject, we produce a fixed-dimensional embedding $\mathbf{z} \in \mathbb{R}^{512}$. For Clinical Tabular Data we adopted the text encoder of UniMedCLIP as feature extractor [67]. UniMedCLIP is a transformer-based architecture pre-trained on a large-scale medical dataset, used in various clinical prediction tasks. Furthermore the fact that having a single encoder for tabular and text data as well as a single encoder for imaging data instead of one per modality, allows to have a more efficient and lightweight pipeline (with only 2 encoders in total instead of 4, i.e. one for CT, one for MRI, one for

tabular and one for text).

The framework is naturally extendible without redesigning a global feature extractor. A new modality can easily be integrated defining an encoder that maps that modality into a shared embedding space.

4.2.3 Missing Aware Conditioning Module

As described in Section 4.2.2, each modality embedding $\mathbf{e}_i^j$ is extracted independently using a modality-specific encoder. In contrast to HyperMM, we do *not* perform conditioning within a universal feature extractor. Instead, conditioning is introduced *after* feature extraction, through a dedicated module that operates on the extracted embeddings and explicitly accounts for modality availability.

Explicitly modelling missingness is important in healthcare settings, where missing data are pervasive and can still carry predictive signal. In particular, even variables with substantial missingness may remain informative for downstream prediction when handled appropriately [10]. More broadly, multimodal learning literature highlights that modalities interact in structured ways and leveraging the structure of multimodal observations is central to effective integration [6].

For this reason we introduce a Missing Aware Conditioning module (MAC) that learns to adaptively modulate its activations by applying an affine transformation on the network's modality features. This design is inspired by Conditional Normalization and Feature-wise Linear Modulation (FiLM) [68, 69], where intermediate activations are modulated using learned scaling and shifting parameters conditioned on auxiliary inputs.

The key idea is provides an efficient way to condition the network on the inter modality missing vector m_i, rather than a modality-specific value. This provides the model with structured information about which modalities are observed for the current patient, allowing it to reason not only about the presence of a single modality, but also about the *overall subset* of available sources. The network, being conditioned on m_i, adapts and make features aware to the missing scenario, making the pipeline robust and suitable for disparate missing datasets. Since the conditioning signal depends on m_i rather than on a specific modality input like in [68, 69], the resulting mechanism is not tied to any particular modality. Additionally different modalities can be inconsistent among each other, inducing distribution shifts in the fused representation and giving contradicting information when predicting the target [51]. Conditioning aims also to calibrate modality embeddings and align information between modalities, to make the fusion receiving

more consistent inputs across missingness patterns.

Specifically, the MAC module learns, through neural networks, g and h arbitrary functions. For each x_i^{obs}, those network take the associated inter-modality missing vector m_i as input and output $\gamma_i = g(m_i)$ and $\beta_i = h(m_i)$ respectively. For each observed modality embedding $\mathbf{e}_i^j \in \mathbb{R}^D$, we then apply the following residual modulation:

$$\mathrm{MAC}(\mathbf{e}_i^j \mid \boldsymbol{\gamma}_i, \boldsymbol{\beta}_i) = \mathrm{LN}\Big(\mathbf{e}_i^j + \Big(\boldsymbol{\gamma}_i \odot \mathbf{e}_i^j + \boldsymbol{\beta}_i\Big)\Big), \tag{4.8}$$

where $\odot$ denotes element-wise multiplication and LN is the Layer Normalization applied for stability. The modulation is applied only when modality j is present, excluding missing modality from the subsequent fusion step.

In our implementation g and h are implemented as simple linear layers and we applied a 0.05 scale factor on the resulting beta and gamma parameters.

Role of the missing vector In the context of missing modalities, is important to explicitly model the missingness pattern. As [10] states, even modalities with high amounts of missing values can still contribute significantly to predictive performance. Based on that, since none in literature exploit the strategy of an inter-modality missing vector, we started to evaluate whether the information carried is actually informative for the downstream task. In particular we notice an increase of performances in missing modality scenarios if concatenating the m embedding to the embedding of the fused representation. Starting from that we develop the more sophisticated conditioning module described above.

4.2.4 Pairwise Fusion

The fusion mechanism is crucial to integrate modality embeddings into a single unified representation for prediction. In clinical multimodal deep learning, *concatenation* is the most commonly used fusion operator [2]. In a recent study [2] is shown that it is adopted by the large majority of *single-fusion* architectures (29 out of 35 reviewed works) and in 46% of approaches for *multiple-fusion*. Although this prevalence does not necessarily imply optimality, it motivates focusing on fusion designs that retain the expressive benefits of concatenation while improving flexibility under heterogeneous and incomplete observations.

To handle the challenge of missing modalities, which often leads to variable-length input sets, we implement an iterative pairwise fusion strategy inspired by the multi phase fusion model of [25]. Traditional concatenation requires a fixed set of modalities and struggles with missing data, often requiring zero-padding that can

introduce unwanted biases. Simple pooling methods like mean may underfit complex cross-modal interactions, whereas attention-based fusion although expressive is often more parameter and data-hungry.

Our fusion process iteratively incorporates available patient modalities embeddings into a fused feature $\mathbf{h_i}^{\text{fused}}$. The merge operation is performed by concatenating the current $\mathbf{h_i}^{\text{fused}}$ with the embeddings of another available patient modality and then applying a MLP block to extract a new merged and updated feature. We initialize the fused representation with the first available modality $\mathbf{h_i}^{\text{fused}} = \mathbf{e}_i^1$. Then we iteratively update the representation by concatenating the current state $\mathbf{h}_i^{\text{fused}}$ with the next available modalities in a residual formulation followed by a Layer Normalization:

$$\mathbf{h_i}^{\text{fused}} = \text{LN}\Big(\mathbf{h_i}^{\text{fused}} + \text{MLP}\Big(\mathbf{h_i}^{\text{fused}} \oplus \mathbf{e}_i^{j+1}\Big)\Big), j = 1, \ldots, j-1 \in x_i^{obs} \tag{4.9}$$

where $\oplus$ denotes feature concatenation's that maps $\mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{2D}$ and MLP : $\mathbb{R}^{2D} \to \mathbb{R}^D$. In case just one modality is available, no update is performed and the fused representation simply remains $\mathbf{h_i}^{\text{fused}} = \mathbf{e}_i^1$.

The residual formulation allows to make the aggregation robust, preserving and updating the current fused representation just when a modality is useful and avoiding performance degradation when a modality is noisy. Modalities indeed can bring very different contributions to the fusion degradating the prediction [51]. The LayerNorm controls the scale of activations after each merge, making the iterative fusion less sensitive to the magnitude of individual modality embeddings and to the number of merge steps.

Furthermore during training, we optionally randomize the order of the observed modalities to mitigate dependence on a fixed modality ordering and promote robustness. Compared to attention-based strategies, this approach is more lightweight but it remains more expressive than parameter-free methods like simple averaging, as the model can learn how to prioritize and weigh information during the iterative merge.

The MLP is implemented through two fully connected layers separated by ReLU activations. The first layer maps the concatenated input from $2D$ back to the dimension D, while the second layer projects from D dimensions to D dimensions, matching the original embedding size. The hidden layer is initialized with He (Kaiming) initialization, consistent with ReLU-based networks. This MLP Merge block is shared across all fusion steps, allowing the model to learn a consistent fusion strategy that can be applied regardless of the number of available modalities

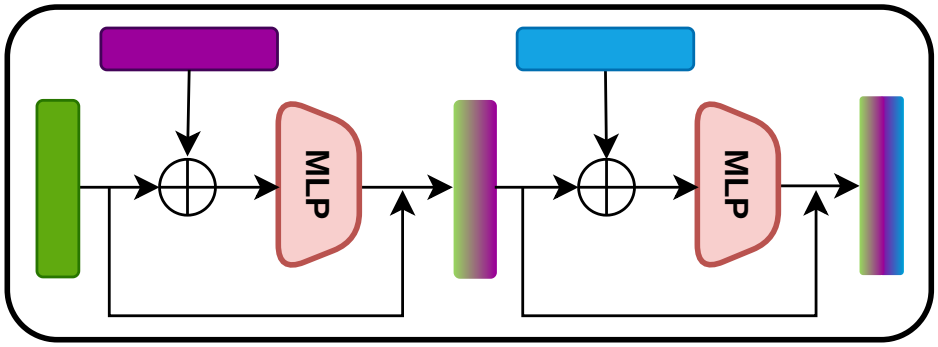


Figure 4.3: Pairwise Fusion Module. Feature modality embeddings are iteratively merged through a concatenation and a learnable MLP block into a single fused representation.

or their order. Figure 4.3 illustrates the architecture of the Pairwise Fusion module.

Since our architecture performs *multiple* fusion steps, our gradual integration strategy [2], instead of merging all available modality embeddings in a single operation, combines them progressively through a sequence of pairwise fusions. By fusing two embeddings at a time, it permits interaction of features between intermediate multimodal representations that would not be possible otherwise. Additionally this procedure can enable hierarchical integration of information step by step that other fusion strategies may struggle to capture, especially when the number of available modalities varies across patients.

4.2.5 Task Prediction Head

After the Pairwise Fusion stage (Section 4.2.4), MMARE yields a single patient fused representation $\mathbf{h_i}^{\text{fused}} \in \mathbb{R}^d$ that integrates information from the available modalities. This representation is fed into a prediction head ρ to produce the final output for the downstream task. It can be represented as:

$$\hat{\mathbf{y}}_i = \rho(\mathbf{h_i}^{\text{fused}}) \tag{4.10}$$

In our classification tasks, the prediction head is a lightweight multilayer perceptron (MLP) that maps the fused representation to C classes logits: $\rho : \mathbb{R}^d \to \mathbb{R}^C$.

Specifically ρ is implemented as two hidden linear layers with ReLU activation functions, followed by dropout and a final linear classifier layer. The dropout probability is fixed to $p = 0.3$ and is applied only during training to mitigate overfitting and

enhance generalization performance. The hidden layers dimensionality are 256 for the first one and 128 for the second one, while the final output dimensionality is C, which corresponds to the number of classes. Changing the output dimensionality C make easy to pass from binary to multi-class tasks without any other architectural changes. To stabilize optimization we initialize hidden layers with Kaiming (He) initialization for ReLU networks. Figure 4.4 illustrates the architecture of the classification task prediction head.

Importantly, the head is *modality-agnostic* since it operates only on the fused embedding and is shared across all missingness patterns. Furthermore the prediction head is *task-agnostic* since the design of the prediction head can be adapted to the desired task. MMARE outputs a generic fused patient representation $\mathbf{z}_i$ after the pairwise fusion stage that can be paired with any standard supervised head such as classification or regression, without modifying the upstream feature extraction, missing-aware conditioning or fusion modules.

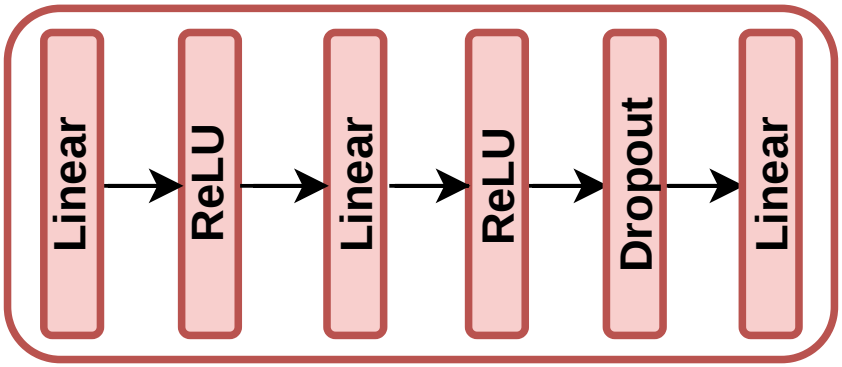


Figure 4.4: MMARE prediction head architecture. The fused patient representation passes through the head and is mapped to class logits for classification tasks.

4.2.6 Advantages of MMARE

The final framework is end-to-end since no manual intervention is required at inference time and importantly is lightweight. It avoids explicit data generation or imputation for missing modalities potentially saving time and reducing some sources of bias. Is not tied to a specific modality ordering thanks to its permutation-invariant design and conditioning strategy based just on the inter modality missing vector. It is task agnostic since can support different types of downstream tasks by choosing the desired prediction head and can handle heterogeneous modalities

of different nature. MMARE is model-agnostic and can flexibly accommodate different types and numbers of modalities by simply adding or changing feature extractors, making it suitable for diverse clinical applications. Our work, explicitly conditions on per-patient modality availability and performs expressive yet efficient fusion over variable-sized modality subsets, improving robustness under realistic missing-modality scenarios.

Chapter 5

Experiments & Results

In this chapter we describe the different multimodal datasets used in this work, the relative performed tasks and different methods' preprocessing steps. We consider two different datasets, corresponding to two different experimental settings:
(i) **IXI**, a fully-observed dataset that enables *controlled* missing-modality simulations via amputation to systematically assess robustness and perform ablation studies.
(ii) **MMIST-ccRCC**, a real-world multimodal dataset with *naturally missing* modalities.
These different experimental setups permit a complete evaluation of our model and comparison methods along datasets regarding different body parts as well as different missing modality rate either natural or simulated.

5.1 IXI dataset

At first we started our experiments on IXI dataset to evaluate performances in a controlled missing modality scenario. IXI dataset is a collection of brain MRI scans of healthy volunteers. Data has been acquired in three different hospitals in London each one with different scanners. In our experiments we focus on four MRI modalities: T1-weighted (T1), T2-weighted (T2), Proton Density (PD) and Magnetic Resonance Angiography (MRA). The demographic information provided for some subjects are sex, height, weight, ethnicity, marital status, occupation, qualification and age. On this dataset we perform two different tasks: the Sex Prediction Task and the Brain Lifespan Epoch Prediction Task.

5.1.1 Sex Prediction Task

The sex prediction task is a binary supervised classification task where the goal is to infer an individual's *biological sex* recorded as male/female. Although some datasets use the term "gender", this variable usually reflects *sex assigned at birth* rather than gender identity. In accordance with the SAGER guidelines [70], this thesis refers to this biological label as *sex*. Sex is a key biological variable that can influence anatomy, physiology and disease mechanisms. For this reason, major biomedical funding and reporting frameworks emphasize the importance of accounting for sex in study design, analysis, and interpretation, to improve the quality and generalizability of results [71]. In neuroimaging specifically, large-scale studies and meta-analyses report measurable average differences across several structural and functional brain measures, while also highlighting substantial distributional overlap between groups [72]. This overlap makes the classification task non-trivial and highlights the need for robust prediction models. Since certain diseases have different prevalence and manifestation across sex, its prediction is important to develop personalized medicine approaches [72]. For example a recent work [73] demonstrates the utility of predicting sex as a meta-feature to enhance the performance of ADHD detection frameworks. Furthermore as noted by [74], the ability of predicting sex is important to avoid bias algorithms and guarantee same clinical precision across male and females. From a methodological perspective, sex prediction provides a well-defined and a stable target to validate whether the pipeline preserve meaningful neuroanatomical information and as it is a biologically grounded task [71], make it a suitable benchmarking procedure.

5.1.2 Brain Lifespan Epoch Prediction Task

The Brain Lifespan Epoch Prediction Task is a multi-class classification task to predict the topological brain epoch of each subject. Brain topological epochs are derived from a recent study [13], which leveraged Diffusion MRI to reconstruct the human structural connectome and then map the structural brain-network organization non-linear evolution across the lifespan. By analyzing the diffusion of water molecules along white matter tracts, the authors were able to model the brain as a complex network (graph) and identify specific "topological turning points" where the global organization of this network undergoes significant transitions. Unlike traditional brain-age estimation, which focuses on chronological age regression, this task investigates whether a model can identify discrete biological phases of brain reorganization. We discretize chronological age into three lifespan epochs, each bounded by a topological turning point. Specifically, we use the turning points at 32 and 66 years to analyze the IXI population in three groups: Adolescence[20-32), Adulthood [32-66), and Early Aging [66-83]. We derive a three-class label `age_bin3` from the Clinical data column AGE, where class 0 corresponds to ages $[20, 32)$,

class 1 to $[32, 66)$, and class 2 to $[66, 83]$. This defines a novel task to test models on a biologically meaningful problem and serves as a as a controlled benchmark to evaluate robustness under different missing-modality settings. Furthermore this task can be useful to assess whether different MRI modalities (e.g., T1-weighted or T2-weighted scans) retain sufficient signal to discriminate between these topologically-defined stages, even when they differ from the DTI modality used in the original study.

5.1.3 Train/Val/Test Split

We create a **6:1:3** train/validation/test rate split (`Split_age_val`) at patient level, stratified by `age_bin3` using a fixed random seed. Tables 5.1 and 5.2 show the distribution of samples across the splits for the Brain Lifespan Epoch and Sex classification tasks respectively.

Table 5.1: IXI split statistics in 6:1:3 rate for the Brain Lifespan Epoch task.

Split	N	Young(0)	Mid(1)	Old(2)
Train	309	70	187	52
Val	51	11	31	9
Test	156	35	95	26

Table 5.2: IXI split statistics in 6:1:3 rate for the Sex Task.

Split	N	Male (0)	Female (1)
Train	309	137	172
Val	51	20	31
Test	156	73	83

5.1.4 MRI Preprocessing

At first we remove all entries associated with duplicated `patient_id` values, ensuring that each subject appears only once in the final dataset. Furthermore we retain all subjects with all available modalities between image and Clinical Tabular ones. Clinical data is considered available if at least one column is not missing. This results in a final dataset of 516 subjects with four MRI image modalities and Clinical Tabular data all available. We perform skull stripping on each volume to

remove non-brain tissue using SynthStrip Harvard tool [75]. Figure 5.1 shows an example of the skull stripping procedure on a T1-MRI scan.

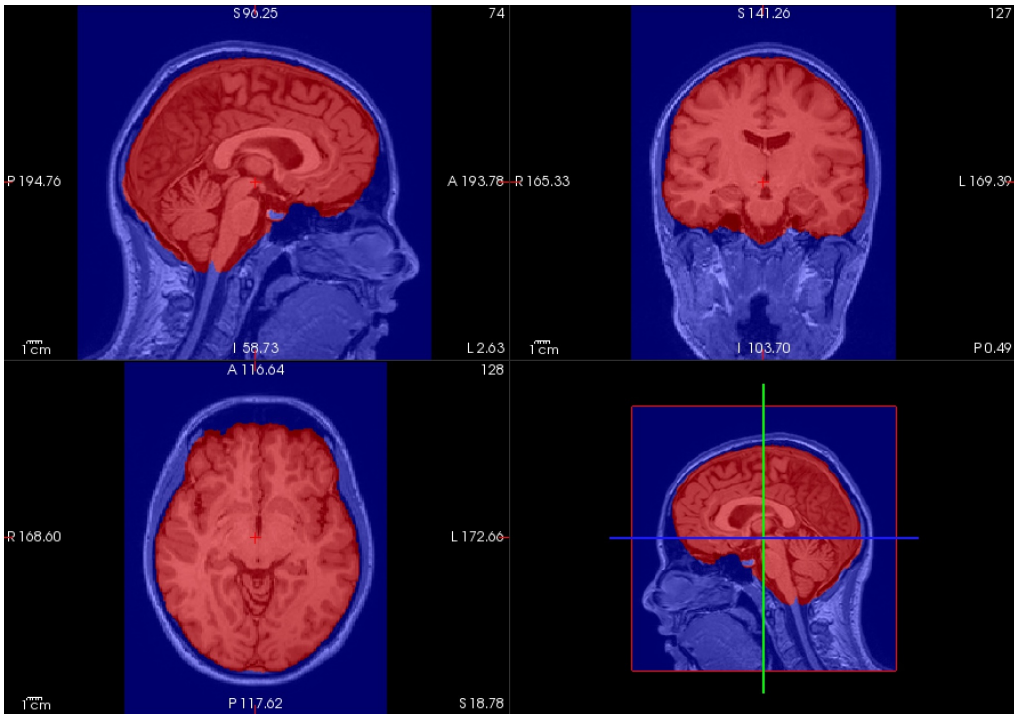


Figure 5.1: ITK-SNAP Visualization: Example of T1-MRI brain stripped volume with SynthStrip tool. In Blue before skull stripping, in Red after skull stripping.

HyperMM preprocessing We apply a unified MONAI [17] preprocessing pipeline to all MRI volumes, resampling to isotropic volumes (1.0, 1.0, 1.0) mm and orienting to the common anatomical convention. We crop foreground independently for each modality and apply a ImageNet-style intensity standardization. To match the 2D VGG pre-trained backbone used in HyperMM, we randomly sample along the z-axes axial slices from each preprocessed volume and resize to 224×224 adapting VGG network input resolution.

MMARE preprocessing For MMARE each volume is reoriented ensuring spatial position coherence and standardized applying z-score volume intensity normalization on foreground voxels. Gaussian noise is used to replace the background as done in [66]. Finally we resize the volumes to obtain a fixed spatial size of 56x448x448, which is required by the 3D Med3D backbone [66]. These pre-processing steps are applied in the same way to all four modalities (T1, T2, PD, MRA) to maintain consistency across the inputs. To reduce scale differences across modalities and improve optimization stability, we apply a per-modality z-score

standardization on resulting embeddings computed only on the training split and using then same statistics on validation and test splits.

5.1.5 Clinical Preprocessing

MMARE preprocessing Unlike HyperMM, MMARE framework as an improvement is able to handle other nature of modalities beyond images. Clinical Tabular Data are serialized into textual prompt and encoded with the pre-trained text-encoder of UniMedCLIP [67]. We create two different prompt structures excluding the 'sex' in the Sex prediction task and excluding the 'age' in the Brain Lifespan Epoch prediction task.

Sex task prompt structure:
"The patient is 'age' years old, with 'White/Black or Black British/Asian or Asian British/Chinese/Other' ethnicity, 'single/married/cohabiting/divorced or separated/widowed'.
Occupation: 'go out to full time employment/go out to part time employment (<25hrs)/study at college or university/full-time housework/retired/unemployed/-work for pay at home/other'; Education: 'no qualifications/O-levels, GCSEs, or CSEs/A-levels/further education (City & Guilds / NVQs)/university or polytechnic degree'.
Body measures include height 'height' and weight 'weight'."

Brain Lifespan Epoch task prompt structure:
"The patient is a 'male/female', with 'White/Black or Black British/Asian or Asian British/Chinese/Other' ethnicity, 'single/married/cohabiting/divorced or separated/widowed'.
Occupation: 'go out to full time employment/go out to part time employment (<25hrs)/study at college or university/full-time housework/retired/unemployed/-work for pay at home/other'; Education: 'no qualifications/O-levels, GCSEs, or CSEs/A-levels/further education (City & Guilds / NVQs)/university or polytechnic degree'.
Body measures include height 'height' and weight 'weight'."

Both omit mode and explicit mode are available. In our experiments we use the omit mode.

5.2 MMIST ccRCC dataset

MMIST ccRCC dataset is a multimodal dataset of 617 patients with clear cell renal carcinoma(ccRCC), curated from TCGA, TCIA and CPTAC [11]. The dataset includes radiology modalities (CT and MRI), histopathology Whole-slide images (WSI) and Genomics along with Clinical Tabular Data. ccRCC is the most common kidney cancer type, amounting to 80% of all renal cell carcinoma cases in adult subjects [11]. Although it is still a challenging task, estimating the prognosis is critical for patient management [76]. All this different modalities' complementary strengths can be leveraged for survival prediction task. For each patient and for each modality we use only the most appropriate scan selected in the MIL stage by the authors [11].

MRI, CT and WSI modalities are provided following the description of general clinical practice as described in Section 2.1.2. Differently clinical and genomic variables all provided as tabular data and need to be better described. Clinical data comprised 11 variables, while genomic data include mutation indicators for three selected genes. Patient clinical information include gender, age, race and cancer history. Gender (male/female) and cancer history (yes/no) are binary variables, age is divided in intervals of 10 years, and race is normalized between: "Black or African American", "White", "Asian", "Hispanic or Latino", or "Other". Additionally Clinical Variables follow AJCC tumor staging system and include tumor descriptors like nodes, metastasis assessments and tumor grade. AJCC considers four relevant categories: i) the extent of the tumor (T); ii) the extent of spread to the lymph nodes (N); iii) the presence of distant metastasis (M); and iv) the assessment of M but based on pathological data (Mp). T is graded in 11 possible levels, while the other categories are all ranked between 0 and 2. The AJCC tumour staging was also considered (4 levels), as well as the tumour grading as assessed by the pathologist (1 to 5) [11]. Genomic data are mutation indicators for selected genes VHL, PBMR1, and TTN.

5.2.1 Vital-12 Survival Task

Vital 12 is a task that predicts patient survival within 12 months, which is an important indicator of long-term healthcare outcomes [59]. This is a relevant binary classification task in the multimodal research field [77]. Multimodal data is collected during the initial hospitalization period to predict whether a patient will survive beyond 12 months after their initial diagnosis [11]. Label 0 identifies dead predicted patients, label 1 survived predicted patients after 12 months.

5.2.2 Train/Val/Test Split

The MMIST-ccRCC partitioning strategy aims to balance both label distribution and modality availability distribution across training and test sets [11]. For robust model selection, we introduce an additional validation split while preserving, as closely as possible, both the label prevalence and the modality-availability across *train/val/test*. Creating a validation set in this setting is non-trivial because a naive stratification on the label alone can yield splits with substantially different modality availability (or vice versa). Moreover, stratifying jointly over label and modality-presence patterns can easily fail due to rare strata (e.g., minority-label patients with MRI available), which may contain too few samples to support standard stratified sampling. To address this, we adopt a hierarchical stratification scheme. For each patient, we define binary presence indicators for MRI(M), CT(C), and Genomics(G) (Clinical and WSI modalities are not included since they are available for all patients). We then build a composite stratification key that combines the label with modality presence (e.g., `label|MCG`), and we progressively fall back to coarser keys whenever some strata are too small for reliable stratification. We first create an 80/20 train/test split, and then split the training portion again to obtain a validation set corresponding to approximately 10% of the total dataset. All splits are generated with a fixed random seed for reproducibility. Compared to the original dataset split, this procedure better align the splits in terms of label prevalence and modality availability, introducing a useful validation set for a reliable validation-based model selection under missing modalities. Table 5.3 shows the comparison with the original split and the final distribution of available modalities for each set. Finally, the statistics highlight that MMIST-ccRCC is characterized by severe class imbalance and highly heterogeneous modality missingness (e.g., MRI is available only for a small fraction of patients), creating a particularly challenging multimodal learning setting.

Table 5.3: Comparison between authors and our label distributions and modality availability for MMIST-ccRCC. WSI and Clinical Tabular data are available for 100% of patients across all subsets.

Split	N	Neg	Pos	Pos(%)	MRI(%)	CT(%)	GEN(%)
Train (authors)	496	58	438	88.3	6.7	36.3	72.8
Test (authors)	121	16	105	86.8	12.4	47.9	83.5
Train (ours)	431	53	378	87.7	7.7	38.5	74.9
Val (ours)	62	7	55	88.7	9.7	38.7	74.2
Test (ours)	124	14	110	88.7	7.3	38.7	75.0
Total Dataset	617	74	543	88.0	7.8	38.6	74.9

5.2.3 MRI & CT Preprocessing

CT and MRI exams are provided as DICOM series. To construct 3D volumes robustly, we grouped DICOM files by `SeriesInstanceUID`. If the reconstructed volume coming from multiple acquisitions contains overlapping slices we remove the duplicates. We use DICOM orientation metadata `ImageOrientationPatient` to ensure correct slice ordering before stacking along the z-axis creating NifTi volumes with the SimpleITK library. Series that fail this volume creation are considered corrupted and discarded.

HyperMM preprocessing To standardize inputs across patients, we apply a unified MONAI [17] preprocessing pipeline to CT and MRI volumes, including orientation normalization, resampling to a target voxel spacing, foreground cropping, and intensity normalization. To calculate the target voxel spacing we perform a data driven approach calculating per-axes median MRI and CT spacing on training set. The final target spacing are respectively (1.87, 1.87, 7.0) mm for MRI and (1.95, 1.95, 4.0) mm for CT. All volumes are reoriented to the common anatomical convention, resampled to the modality-specific target spacing and then foreground cropping is applied independently to each modality to remove background regions. MRI intensities are normalized to [0,1], while CT intensities are at first windowed to [-79, 304] HU to emphasize the kidney [78] and then rescaled to [0,1]. ImageNet-style intensity standardization is applied. To match the 2D VGG pre-trained backbone used in HyperMM, we randomly sample along the z-axes axial slices from each preprocessed volume and resize the resulted slice to 224×224 adapting VGG network input resolution.

Additional Kidney-Centred Preprocessing To investigate whether focusing on organ-relevant regions could improve Vital-12 survival prediction on MMIST-ccRCC, we performed an additional exploratory preprocessing experiment in which CT/MRI volumes were restricted to kidney-centred regions, rather than processing the full abdominal scan. To do so we used TotalSegmentator [79] to obtain left- and right-kidney segmentation masks and to localize the kidneys within each 3D scan. The two separate kidney masks provided by TotalSegmentator are merged into a single binary kidney mask, which is used to define a kidney-centred crop of the volume from where extracting 2D slices to be fed into the HyperMM pipeline. Despite being a interesting strategy especially to make the model focusing on a region and decrease computational load, this kidney-centred preprocessing did not yield consistent improvements in our setting. Therefore, to avoid adding computational complexity, we do not use this variant and retain the original preprocessing protocol.

MMARE preprocessing For the MMARE framework CT and MRI volumes are encoded using the pre-trained 3D ResNet-18 backbone of Med3D [66]. Using this 3D feature extractor instead of analyzing slice by slice like HyperMM does, permits to effectively capture spatial features within volumetric data. Each available volume is loaded and reoriented to common anatomical convention. For the intensity we apply a per-volume z-score normalization on foreground voxels and replace the background with Gaussian noise as done in [66]. To match the requirements of Med3D, we resize each volume to 56x448x448. Feature extraction is performed up to the last convolutional stage (layer4) and then apply global average pooling over the three spatial dimensions to obtain a 512-dimensional embedding per modality as in [11]. Finally we apply z-score normalization per modality to the resulting features across the training set and we use the same normalization parameters to normalize validation and test features.

5.2.4 WSI Preprocessing

HyperMM preprocessing Whole-slide images are provided in .svs format. However they are too large and can't be directly stored in the GPUs. We therefore create an extraction pipeline to extract representative tiles of .svs files at a target magnification to be used as input to the HyperMM feature extractor. At thumbnail level, which is the lowest resolution level of the WSI pyramid, we convert to gray scale and apply Otsu thresholding [80] to identify tissue regions. We then retain only pixels whose color is close to the average tissue color (less than 35 of Euclidean distance) [81]. To cover tissue regions and avoid duplicates, tissue pixels identified on the thumbnail are mapped to a grid of tile indices (t_x, t_y) at a selected source level using the slide-specific downsampling factors provided by OpenSlide [82]. To ensure consistent physical resolution across slides, including cases where the base scan is acquired at 40×, we normalize tile extraction to a target resolution of 0.50 $\mu m/px$ (20×). The tile is read at native target resolution if present, otherwise we read a larger region at finer level and downsample to the source level. Upsampling is always avoided to guarantee same image quality across all slides. To limit the number of patches per slide while maintaining spatial coverage, we apply a grid-based sampling strategy that selects at most 150 tiles uniformly over the tissue extent. Each extracted tile is finally discarded if tissue pixels are below 60%, guaranteeing a minimum patches per slide. Finally tiles are stacked in a .npy volume to make the preprocessing easier, having all image modalities as 2D stacked volumes, ready to be fed in the HyperMM feature extractor. As suggested by authors [11] a new WSI pipeline is implemented, exploring a different strategy to improve task predictions.

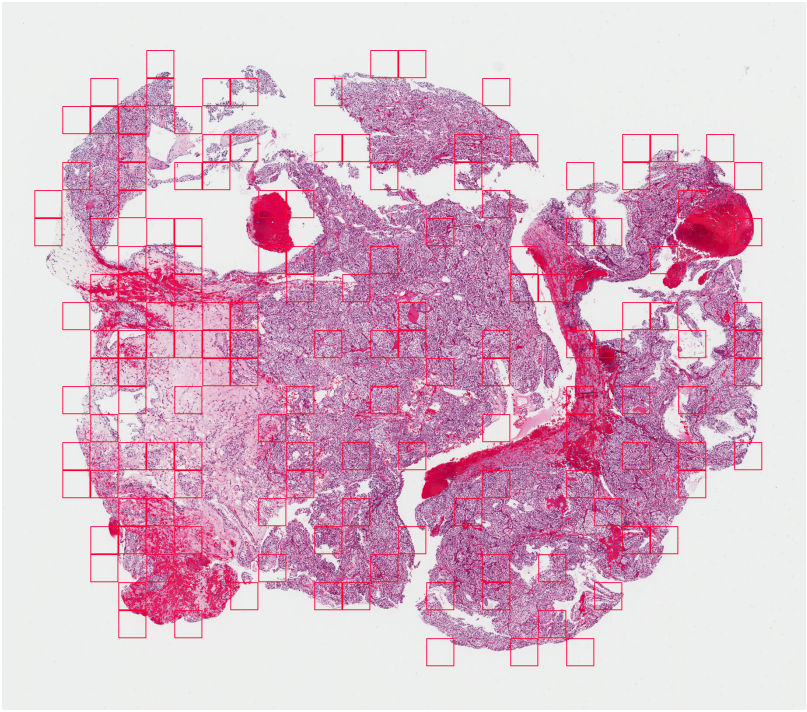


Figure 5.2: Example of the WSI tiling procedure. Tissue-derived tile coordinates are projected onto the slide thumbnail (red rectangles).

MMARE preprocessing In the MMARE experimental setting we do not need to use the same feature extractor across all modalities as is required by HyperMM, so instead of running full WSI extraction pipeline, we use the pre-extracted features provided by the dataset authors [11]. In particular those WSIs representations are obtained using CLAM [83] to isolate tissue areas from the background. Each slide is then divided in 256 x 256 patches and fed to a ResNet50 for feature extraction.Global average pooling is performed across patches and resulting feature vectors are 1024-dimensional. We then add a simple adapter in our model after the feature extraction (see section 4.2.2) which projects these 1024-dimensional features into a 512-dimensional embedding space, aligning them with the dimension of the other modalities. This adapter is a simple linear layer with a ReLU activation, which is trained end-to-end together with the rest of the model. Before training we normalize per-patient WSI features using L2-normalization across all splits.

5.2.5 Clinical & Genomic Preprocessing

MMARE preprocessing Unlike HyperMM, MMARE framework as an improvement is able to handle other nature of modalities beyond images. Clinical and genomic data are serialized into textual prompt simulating electronic health records and then encoded with the pre-trained text-encoder of UniMedCLIP [67]. Following [59] we create the following prompt structure for Clinical Tabular data merging clinical and genomic as the same textual modality:
"The patient is a 'male/female' of 'Asian/Black or African American/Hispanic or Latino/White/other' race, diagnosed at age 'age', who has a/no VHL mutation and a/no PBMR1 mutation with/without a TTN mutation. Tumor characteristics include tumor stage 'pt' with node involvement at stage 'pn' and pathological metastasis at stage 'pm'. The overall tumor stage is 'stage overall' and the tumor grade is 'grade.''
In our experiments we use the omit mode. We furthermore standardize extracted features using L2-normalization in all splits.

5.3 Setup & Implementation Details

All code is developed with Python 3.10, along with Python's MONAI library and PyTorch 1.12 for the implementation of the end-to-end pipeline. Training experiments are run on an NVIDIA TITAN Xp GPU with a 11.4 CUDA version.

We compare MMARE against HyperMM described in 4.1 and other three representative imputation-free baselines: **AdaCoMed** [59], which leverages a mixture-of-modality-experts fusion scheme with adaptive collaboration between large and small models; **DRIM** [26], which disentangles shared and modality-specific representations and fuses them through a masked attention mechanism; and **UniLMMV** [36], which encodes each modality into embeddings and aggregates them using a masked attention-based summarization module that naturally supports variable-sized modality subsets. For all competitors, we relied on the authors' official code implementations when publicly available and followed their recommended training settings whenever applicable.

HyperMM training protocol HyperMM is is trained for a maximum of 150 epochs, with an early stopping strategy, where training stops after 10 epochs without decrease in validation loss. We use a batch size of 1, an Adam optimizer with a learning rate of 1e-4, a weight decay of 0.0005 and a pre-trained VGG11 [63] feature extractor as authors [12].

MMARE training protocol MMARE is trained with an early stopping strategy monitoring balanced accuracy with a patience of 25 epochs. We train the model with an Adam optimizer with a learning rate of 1e-4 a weight decay of 1e-4 and a batch size of 64.

For the Brain Lifespan Epoch prediction task we utilize a weighted random sampler using the inverse of class frequencies as sampling weights to mitigate class imbalance. Sex prediction task is balanced so we do not use any sampling strategy or augmentation. For the ccRCC dataset since the dataset is unbalanced we use a train sampler sampling 6x times the the minority class(0) and perform gaussian noise augmentation with 0.30 probability.

We use Acc (Accuracy), Bacc (Balanced Accuracy), AUC (Area Under the ROC Curve), F1-score, Precision and Recall as evaluation metrics for both datasets and tasks. Since IXI is a full modality dataset, we simulate MCAR (Missing Completely At Random) missing modalities at different rates (20%, 40%, 60%) during training, validation and testing phases. The fact that we drop modalities in each set is to be more realistic in healthcare scenarios. The amputation is performed singularly for each modality and then ensuring that at least one modality is always available for each patient.

5.4 Results

This section reports the experimental results of the proposed MMARE framework and competitors under both *simulated* and *naturally occurring* missing-modality settings. Several observations can be drawn from the results of our experiments on the MMIST-ccRCC and IXI datasets. We first evaluate robustness on the IXI benchmark, where missingness is synthetically introduced at predefined rates (20%, 40%, and 60%) to enable controlled comparisons across methods and to quantify how performance degrades as fewer modalities are observed. We then move to a real-world scenario on the MMIST-ccRCC cohort, where missingness is *natural* and arises from clinical acquisition practice to assess the practical effectiveness of missing-aware learning under heterogeneous data availability. On MMIST-ccRCC we highlight MMARE *parameter efficiency*. Besides reporting predictive performance under naturally missing modalities, we compare methods in terms of trainable parameters to highlight the performance-efficiency trade-off in real clinical settings.

For completeness, we report two variants of our method. **MMARE w/o Clinical** excludes clinical/tabular information and is designed to provide a direct and fair comparison with HyperMM as it can handle only imaging modalities. **MMARE w/ Clinical or MMARE** is the complete version of our model that handle every data nature. It incorporates clinical features, reflecting the full multimodal setting and enabling a comparison against state-of-the-art imputation-free multimodal baselines that leverage both imaging and non-imaging data. Across all experiments, we follow the evaluation protocol described in the previous sections and report standard classification metrics, highlighting both absolute performance and robustness trends across missing-modality regimes.

5.4.1 Sex Task on 20%, 40%, 60% Missing Rates

Table 5.4 reports the Sex classification performance on IXI under increasing missing-modality rates. Overall, MMARE shows considerable improvements over all methods, achieving the best *Accuracy*, *Balanced Accuracy*, and *F1* across all regimes. In particular, MMARE remains strong even in the most challenging 60% setting, suggesting that the conditioning and pairwise fusion still preserve performance when more modalities are unavailable. The clinical variant is consistently slightly stronger than MMARE w/o Clinical, suggesting that clinical tabular information complements MRI features. Interestingly, HyperMM achieves the highest AUC at 20% and 60% missing rates, but it underperforms on other metrics with respect to MMARE w/o Clinical, indicating that does not necessarily translate into better model. Fig 5.3 complements the table by making the robustness trends visually

Method	Acc. (↑)	Bacc. (↑)	AUC (↑)	F1 (↑)	Prec. (↑)	Rec. (↑)
20% Missing Rate						
DRIM	0.7564	0.7455	0.8614	0.8000	0.7103	**0.9157**
UniLMMV	0.7308	0.7264	0.8100	0.7586	0.7253	0.7952
AdaCoMed	0.6923	0.6787	0.7955	0.7551	0.6549	0.8916
HyperMM	0.5833	0.5551	**0.9305**	0.7796	0.6641	0.7166
MMARE w/o Clinical (ours)	0.8333	0.8318	0.9018	0.8458	**0.8381**	0.8554
MMARE w/ Clinical (ours)	**0.8355**	**0.8319**	0.9098	**0.8516**	0.8189	0.8876
40% Missing Rate						
DRIM	0.7179	0.7184	0.7797	0.7284	0.7468	0.7108
UniLMMV	0.6667	0.6504	0.7666	0.7426	0.6303	**0.9036**
AdaCoMed	0.7564	0.7521	0.8064	0.7816	0.7473	0.8193
HyperMM	0.5938	0.5658	0.8658	0.7433	0.6623	0.7311
MMARE w/o Clinical (ours)	0.8079	0.8032	**0.8916**	0.8294	0.7964	0.8683
MMARE w/ Clinical (ours)	**0.8136**	**0.8105**	0.8909	**0.8310**	**0.8139**	0.8496
60% Missing Rate						
DRIM	0.6474	0.6398	0.7033	0.6961	0.6429	0.7590
UniLMMV	0.6538	0.6515	0.7535	0.6786	0.6706	0.6867
AdaCoMed	0.6731	0.6730	0.7405	0.6871	0.7000	0.6747
HyperMM	0.6569	0.6403	**0.9213**	0.7857	0.7411	0.6227
MMARE w/o Clinical (ours)	0.7990	0.7934	0.8825	0.8251	0.7721	0.8889
MMARE w/ Clinical (ours)	**0.8135**	**0.8094**	0.8999	**0.8340**	**0.7824**	**0.8933**

Table 5.4: Performances of the Sex Task on IXI dataset. **Bold** means best, underline means second best.

explicit. The radar plots show that MMARE maintains a balanced profile across metrics in different missing rates. Several methods highly degrade primarily on Bacc/F1 (e.g., DRIM) or exhibit a worse precision-recall trade-off. Either AUC curves highlight substantial deterioration of other method's discriminative performance, meanwhile both MMARE variants show a flatter and more stable trend. All these results suggest that MMARE not only achieves the best overall operating performance but also preserves discriminative information as modalities are removed, supporting its robustness in multimodal missing scenarios.

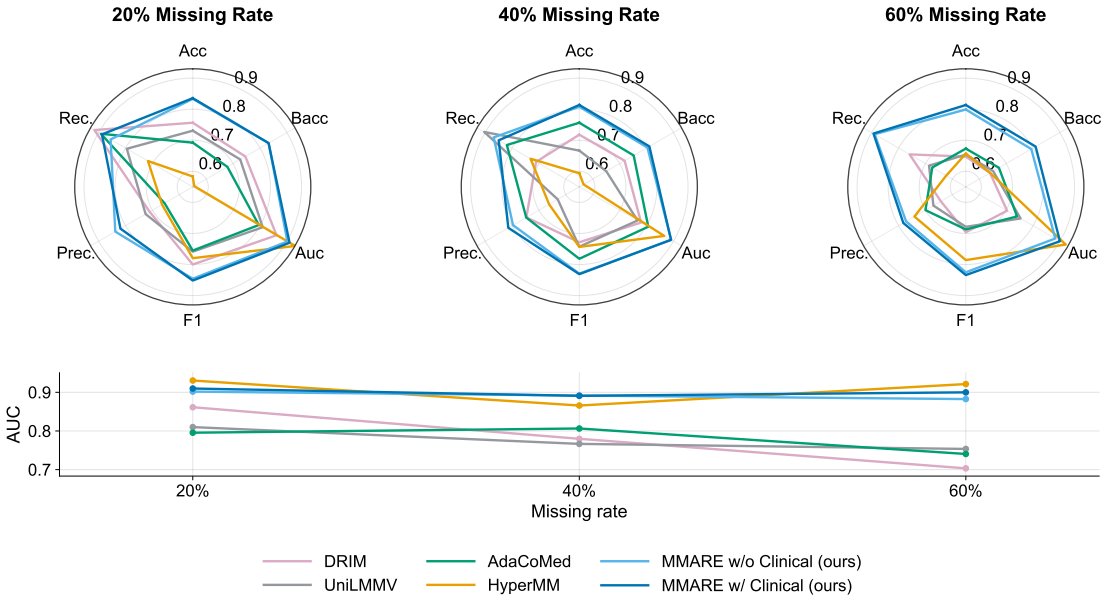


Figure 5.3: Sex classification results on IXI under increasing missing-modality rates. The radar plots show the performance across metrics, while the AUC curves illustrate stability under missingness.

5.4.2 Brain Lifespan Epoch Prediction Task on 20%, 40%, 60% Missing Rates

Method	Acc. (↑)	mF1 (↑)	mPrec. (↑)	mRec. (↑)	AUC (↑)
20% Missing Rate					
DRIM	0.5192	0.4522	0.4711	0.4610	0.6599
UniLMMV	0.4722	0.4537	0.5243	0.5450	0.7336
AdaCoMed	0.5705	0.5453	0.5733	0.5682	0.7446
HyperMM	0.6004	0.5986	<u>0.6115</u>	**0.7010**	<u>0.8139</u>
MMARE w/o Clinical	<u>0.6303</u>	<u>0.6303</u>	0.5850	0.6004	0.7696
MMARE w/ Clinical	**0.6987**	**0.6392**	**0.6791**	<u>0.6153</u>	**0.8248**
40% Missing Rate					
DRIM	0.4316	0.4222	0.4407	0.4886	0.6615
UniLMMV	0.4017	0.4038	0.4641	<u>0.5530</u>	0.6924
AdaCoMed	0.5043	0.4600	0.4651	0.5034	0.6973
HyperMM	0.4614	0.4668	0.4596	**0.6538**	<u>0.7884</u>
MMARE w/o Clinical	<u>0.6115</u>	<u>0.5478</u>	<u>0.5498</u>	0.5525	0.7429
MMARE w/ Clinical	**0.6974**	**0.5879**	**0.7174**	0.5523	**0.8010**
60% Missing Rate					
DRIM	0.3462	0.3164	0.3430	0.3707	0.5196
UniLMMV	0.2521	0.2409	0.2250	0.3940	0.5478
AdaCoMed	0.4188	0.3603	0.3712	0.3945	0.5917
HyperMM	0.4730	0.4506	0.4094	**0.5707**	<u>0.7142</u>
MMARE w/o Clinical	<u>0.5000</u>	<u>0.4580</u>	<u>0.4656</u>	0.4836	0.6532
MMARE w/ Clinical	**0.6177**	**0.5153**	**0.5653**	<u>0.5029</u>	**0.7261**

Table 5.5: Performances on the IXI dataset for the Brain Lifespan Epoch Prediction Task at different missing rates. The best performance for each metric is highlighted in **Bold**. <u>Underlined</u> metrics are the second best.

Table 5.5 shows that **MMARE w/ Clinical** consistently achieves the best overall performance across all missing-modality regimes on the Brain Lifespan Epoch Prediction Task. At **20%** missingness it reaches the highest multi-class AUC, outperforming both HyperMM and imputation-free competitors. Under **40%** missingness, MMARE w/ Clinical remains remarkably stable, while all baselines exhibit a clear drop in accuracy and macro scores, indicating that the proposed framework mitigates the loss of discriminative information when fewer modalities are observed. When missingness becomes **60%**, although performance decreases for every method, MMARE w/ Clinical still maintains the strongest results, suggesting improved robustness of class separation across the three lifespan epochs. HyperMM

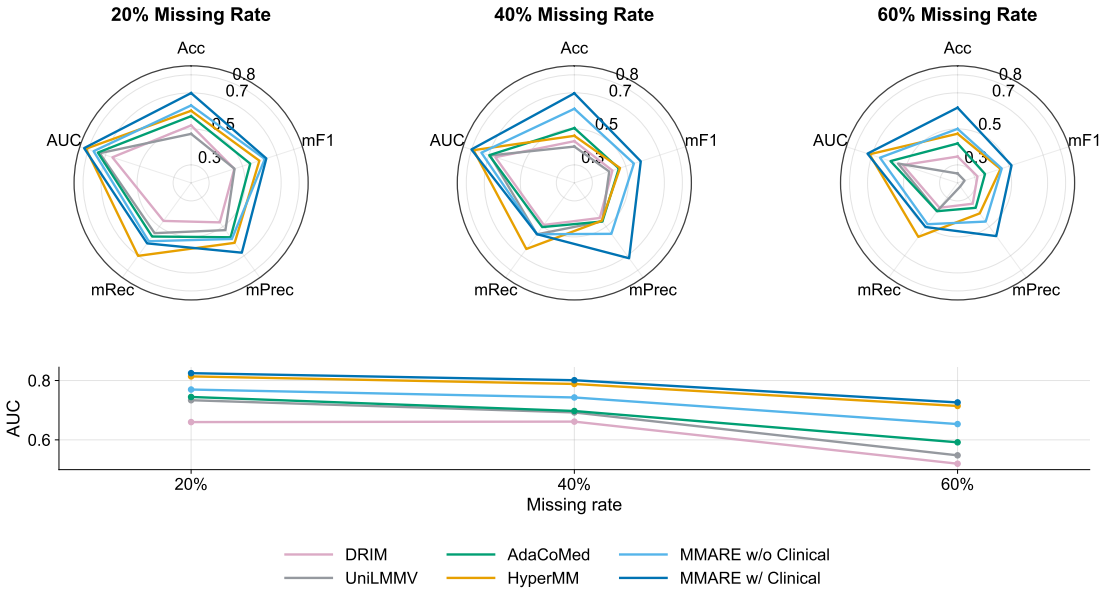


Figure 5.4: Brain Lifespan Epoch Classification results on IXI under increasing missing-modality rates. The radar plots show the performance across metrics, while the AUC curves illustrate stability under missingness.

obtains the highest mRec in each setting, but this higher recall does not translate into better balanced performance (mF1/mPrec./AUC). On the other side, MMARE achieves a stronger and more consistent trade-off across considered metrics.

In addition to the numerical summary, Figure 5.4 provides an intuitive view of performance trend with the increasing missing-modality severity. The radar plots highlight that **MMARE w/ Clinical** achieves the best overall profile across the five metrics (Acc, mF1, mPrec., mRec., and AUC), remaining competitive even when only a small subset of modalities is available. The AUC trend is particularly useful to analyze robustness. **MMARE w/ Clinical** obtains the highest AUC at all missing rates and achieves the smallest overall degradation from 20% to 60% (ΔAUC = 0.0987), indicating a more progressive behavior. While HyperMM often achieves strong mAR values, this advantage does not translate into the best macro scores (mF1/mPrec.) or AUC, suggesting a less balanced trade-off. Overall, the combined evidence from Table 5.5 and Figure 5.4 supports that MMARE better preserves discriminative information across missing-modality regimes, making it a robust choice for the proposed Brain Lifespan Epoch classification benchmark.

5.4.3 Vital-12 on MMIST-ccRCC

We next evaluate MMARE on the MMIST-ccRCC cohort for 12-month survival prediction, a challenging clinical task characterized by *naturally occurring* missing modalities rather than synthetically imposed missingness. In this setting, missingness patterns are not controlled and may correlate with clinical workflow. Given the strong class imbalance, we focus on *Balanced Accuracy* (Bacc), *Macro F1*, and *AUC* metrics, that better reflect clinically meaningful performance under uneven class distributions. In addition to predictive performance, we report the number of *trainable parameters* in the training phase to quantify the efficiency of MMARE with respect to imputation-free competitors.

Table 5.6: Performances on the MMIST ccRCC dataset. **Bold** means best.

Multimodal Model	Bacc.	AUC	mF1	#Params.(M)
DRIM	**0.7799**	0.7987	**0.6932**	16.9
AdaCoMed	0.7390	**0.8500**	0.6242	18.7
UniLMMV	0.7656	<u>0.8104</u>	0.6281	21.0
HyperMM	0.6377	0.6532	0.6377	12.9
MMARE w/o Clinical (ours)	0.6327	0.6636	0.5346	**1.4**
MMARE w/ Clinical (ours)	<u>0.7753</u>	0.7994	<u>0.6847</u>	**1.4**

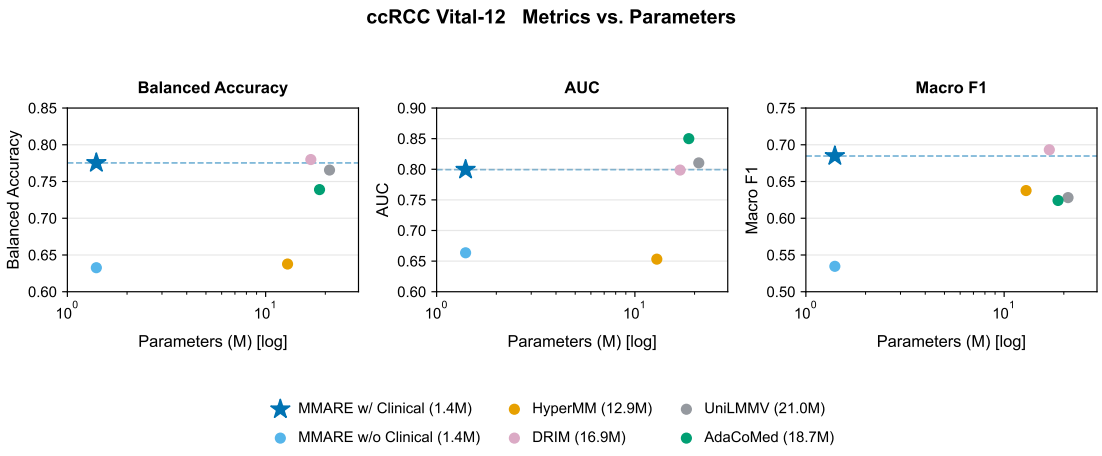


Figure 5.5: MMIST-ccRCC trade-off between performance and trainable parameters.

Figure 5.5 summarizes the trade-off between predictive performance and model size on ccRCC by plotting Bacc, AUC and Macro F1 against the number of trainable parameters (log-scale). Although MMARE is not the top-performing method on

every metric, it achieves *comparable* results to state-of-the-art imputation-free competitors while using a substantially smaller parameter budget. MMARE relies on only **1.4M** trainable parameters, compared to **12.9M** of HyperMM, **16.9M** of DRIM, **18.7M** of AdaCoMed and **21.0M** of the UniLMMV method. This indicates that the proposed pipeline achieves a good accuracy-efficiency balance in a realistic naturally-missing scenario, where bigger models may be harder and slower to train. Overall, the plot supports MMARE as a parameter-efficient alternative that preserves competitive discriminative ability (AUC) and balanced performance (BAcc/Macro F1) under real-world incomplete multimodal data resulting in being an attractive **lightweight** solution for deployment.

5.5 Ablation Studies

To quantify the contribution of the main design choices in MMARE, we perform detailed ablation studies along two axes: (i) the **Missing-Aware Conditioning** module (enabled vs. disabled), and (ii) the **Aggregation Strategy** used to combine the available modality embeddings (**Mean Fusion** vs. **Pairwise Fusion**). This yields a 2×2 comparison: *No cond. & Mean*, *Cond. & Mean*, *No cond. & Pairwise*, and *Cond. & Pairwise* which is the MMARE proposed method. For the **Mean** baseline, we follow the simple permutation-invariant strategy used in HyperMM [12], computing the element-wise mean of the observed modality embeddings. This is feasible because all modalities are mapped to a shared 512-dimensional space despite being produced by different feature extractors. For **Pairwise**, we use our gradual pairwise aggregation, which composes modality information through successive merge operations.

We evaluate the impact of these ablations on both datasets. More in detail on **IXI** (simulated missingness) we focus on the **Sex** task at two representative regimes. We consider the lowest missing rate (20%) and the most challenging missing rate (60%). Tables 5.7 and 5.8 respectively show the results. On **MMIST-ccRCC** (naturally missing), we report results *with* and *without* clinical data in Table 5.9 and Table 5.10 .

Ablation on IXI Sex task

Table 5.7: Ablation study on 20% missing rate WITH clinical data on Sex task.

Setting	Acc.	Bacc.	AUC	F1	Prec.	Rec.
No cond. & Mean	0.7607	0.7550	0.8718	0.7895	0.7421	0.8434
Cond. & Mean	0.7393	0.7270	0.8541	0.7897	0.6924	**0.9197**
No cond. & Pair.	0.8184	0.8156	0.8963	0.8342	0.8113	0.8594
Cond. & Pair.(Ours)	**0.8355**	**0.8319**	**0.9098**	**0.8516**	**0.8189**	0.8876

Table 5.8: Ablation study on 60% missing rate WITH clinical data on Sex task.

Setting	Acc.	Bacc.	AUC	F1	Prec.	Rec.
No cond. & Mean	0.8019	0.7974	0.8848	0.8247	0.7692	0.8889
Cond. & Mean	0.8089	0.8047	0.8804	0.8299	0.7785	0.8889
No cond. & Pair.	0.7855	0.7795	0.8930	0.8155	0.7460	**0.9022**
Cond. & Pair.(Ours)	**0.8135**	**0.8094**	**0.8999**	**0.8340**	**0.7824**	0.8933

At 20% missingness (Table 5.7), Pairwise aggregation already yields a strong gain over Mean and adding the conditioning module further improves *all* key metrics, reaching the best performing combination. Interestingly conditioning with Mean does not consistently improve performance at 20%, but it increases recall, obtaining the highest value across combinations. At 60% missingness (Table 5.8), the effect of conditioning on Mean remains modest, whereas **Cond. & Pairwise (ours)** provides the best overall operating performance while also improving AUC compared to the unconditioned Pairwise variant.

Ablation on MMIST-ccRCC

Table 5.9: Ablation study: Impact of conditioning module and aggregation strategy on ccRCC WITH clinical data

Setting	Acc.	AUC	mF1	mPrec.	mRec.
No cond. & Mean	0.5807	0.7078	0.4912	0.5560	0.6390
Cond. & Mean	0.7338	0.7604	0.5804	0.5806	0.6630
No cond. & Pairwise	**0.8548**	0.7935	0.6587	0.6505	0.6688
Cond. & Pairwise (Ours)	0.8226	**0.7994**	**0.6847**	**0.6577**	**0.7753**

Table 5.10: Ablation study: Impact of conditioning module and aggregation strategy on ccRCC WITHOUT clinical data

Setting	Acc.	AUC	mF1	mPrec.	mRec.
No cond. & Mean	0.4543	0.6396	0.4143	0.5557	0.6301
Cond. & Mean	0.6398	0.6405	0.5218	**0.5678**	0.6411
No cond. & Pairwise	0.5027	0.6413	0.4478	0.5605	**0.6470**
Cond. & Pairwise (Ours)	**0.6801**	**0.6639**	**0.5346**	0.5662	0.6327

With clinical data (Table 5.9), conditioning has a *large* impact when using Mean aggregation as it improves all considered metrics, indicating that conditioning helps the model exploit the clinical/imaging context when the fusion itself is simplistic. Pairwise aggregation substantially strengthens the model even without conditioning. Adding conditioning on top of Pairwise slightly reduces accuracy but improves all other informative metrics, suggesting that conditioning mainly shifts the operating point toward a better balanced performance rather than optimizing raw accuracy.

Without clinical data (Table 5.10), the benefit of conditioning becomes clearer across both aggregation strategies. Conditioning improves the Mean baseline

substantially in Accuracy and mF1. The strongest configuration is **Cond. & Pairwise (ours)**, which achieves the best performance in this setting.

Overall, ablation results indicate that the combinations of modules of our model results in being always the best performing one. **Pairwise Fusion** is the main contributor to stronger performance, consistently improving simple Mean aggregation. The **Missing Aware Conditioning** module provides complementary gains that are most visible when missingness is high or when the available input set is weaker (e.g., without clinical data). Additionally its impact is more evident when the mean aggregation strategy is used and slightly less evident when using pairwise strategy. This supports the claim that the pairwise strategy is more robust and suited in missing modalities scenario. Importantly, conditioning promotes more balanced macro behavior (mF1/mRec/AUC), which is particularly desirable in imbalanced clinical settings.

Chapter 6

Conclusions

This thesis addressed the problem of supervised multimodal learning with missing modalities in healthcare, where incomplete observations are common and standard learning pipelines often rely on explicit imputation procedures that can be computationally expensive and potentially introduce bias. Starting from the limitations of existing methods identified in literature, we contributed proposing **MMARE**, a lightweight end-to-end imputation-free framework for multimodal learning with missing modalities. The method combines two key ideas: (i) a **Missing-Aware Conditioning (MAC)** module that explicitly conditions feature processing on each patient's modality-availability pattern, and (ii) a **Pairwise Fusion** strategy that incrementally merges the available modality embeddings into a unified representation. Together, these components allow the model to exploit complementary information when multiple modalities are available while remaining functional and discriminative under incomplete observations, without requiring explicit modality reconstruction. Our task agnostic framework is able to operate with a variable number of observed modalities at both training and inference time, while preserving robustness across different missingness patterns. Additionally the developed method is model agnostic, since it can be integrated with any neural network-based feature extractor and predictor, making it adaptable to a wide range of multimodal pipelines and tasks in healthcare.

The empirical results support the objectives of this thesis. On the IXI dataset, under *simulated missingness* (20%, 40%, and 60%), MMARE demonstrated strong robustness and consistently strong performance across *Sex* task and the proposed *Brain Lifespan Epoch Prediction* task. In particular, the results showed that MMARE preserves discriminative ability as missingness increases and maintains a

favorable trade-off across metrics. The ablation studies further clarified the role of each component. The Pairwise Fusion mechanism provides the main performance gain, while Missing-Aware Conditioning contributes complementary improvements, especially in more challenging settings as higher missingness or reduced input information.

On the MMIST-ccRCC dataset, which presents *naturally occurring* missing modalities, although MMARE do not always achieve the highest absolute score against all imputation-free competitors, it obtains competitive and comparable performances on clinically relevant metrics while using a substantially smaller number of trainable parameters. This experiment highlights the lightweight characteristic of the proposed framework which is an important practical result of the thesis. In realistic clinical settings, MMARE offers a strong **performance-efficiency trade-off**, making it a promising alternative when model size, training cost, and deployment constraints are relevant.

Beyond the specific benchmarks considered in this work, the proposed framework also provides a broader methodological contribution. MMARE supports the view that missingness should not be treated only as data unavailability to be repaired, but can be explicitly modeled as informative context to improve robustness under incomplete multimodal observations. This would be an inspiring future research methodology in missing-aware learning scenarios. Additionally as MMARE is modular it can be integrated itself with different feature extractors and prediction heads, but also integrated in pre-existing pipelines, making it adaptable to heterogeneous multimodal pipelines and tasks in diverse health applications.

In summary, this thesis demonstrates that it is possible to design an imputation-free multimodal learning framework that is effective, robust to missing modalities, flexible across tasks and datasets and parameter-efficient in realistic healthcare scenarios. We have demonstrated the advantages of MMARE for MLMM eliminating the need of imputation and together with the dependence on which modality is missing. These findings provide a solid basis for future research on trustworthy multimodal learning under incomplete data.

6.1 Limitations and Future Work

Several aspects of our work can be extended and improved in future research, addressing the current limitations.

Extended Experiments and Benchmarks. Public healthcare datasets with multiple modalities, especially with realistic missing-modality patterns, are still limited but with the increasing availability and novel data collections MMARE could be tested on a wider range of multimodal cohorts, tasks and settings. Additionally, introducing more benchmark models would help to compare the proposed framework, improve the reliability of conclusions and place the proposed approach more clearly within the evolving state of the art.

Different Missingness Scenarios. On IXI dataset we have evaluated robustness under simulated missingness at controlled rates, amputating modalities with a completely random process (i.e., *MCAR*). Future work should extend the evaluation to different missingness processes or amputation techniques like Missing At Random *MAR* and Missing Not At Random *MNAR* strategies, studying how different missingness-generation assumptions affect model behavior and robustness claims.

Improve Vital-12 Task. On MMIST-ccRCC, the current setup already provides a realistic naturally missing benchmark, but additional improvements are possible. In particular, future work could exploit a larger fraction of the available patient data to test whether broader coverage improves survival prediction for both MMARE and baseline methods. In our current setup, we follow the dataset protocol that relies on a subset of best scans selected through the authors' MIL pipeline, but multiple available scans per modality for each patient could be included from the dataset, potentially improving the survival prediction. Additionally for HyperMM a related extension would be to explore an alternative multimodal definition, treating different imaging views (axial, coronal, sagittal) as separate complementary modalities.

Generalization beyond missing-modality scenarios. While the MAC module is specifically designed for missing-modality settings, the Pairwise Fusion mechanism is more general and may also be beneficial in standard multimodal learning scenarios where all modalities are available. An interesting direction is therefore to evaluate the fusion component independently in fully observed multimodal tasks, in order to understand whether the benefits observed can extend beyond missing-data settings.

Multimodal training for unimodal inference. Based on observation of [55], a particularly relevant direction would be to investigate the multimodal training

improvement on *unimodal* inference performance, i.e., whether a model trained with multiple modalities can outperform a purely unimodal model even when only one modality is available at test time. This would be highly relevant in real-world applications, where data streams may be intermittently unavailable at inference time. In this setting, the model could leverage cross-modal relationships learned during training to extract richer information from the remaining modality at test time.

Unseen or newly added modalities at test time. The modular structure of MMARE suggests the possibility of extending the framework to settings where unseen or newly added modalities are introduced after training or only at test time as explored by [84]. This direction is promising but non-trivial as it requires compatible feature spaces and robust alignment mechanisms to ensure that new introduced modality encoders can interact coherently with the trained fusion and conditioning components. Future work could investigate modality-agnostic channel integration strategies and adapter-based extensions for this purpose.

Feature disentanglement. Another promising improvement would be to introduce a disentanglement step before the MAC module to better separate specific modalities as well as prepare features to be aligned before the fusion. This may improve the interpretability of the learned representations and allow MAC to act on cleaner latent factors, potentially increasing robustness and improve predictions.

Stronger domain-specific encoders. The current implementation already benefits from pre-trained backbones, but the framework could be strengthened further by integrating more recent and powerful domain-specific multimodal medical encoders and foundation models. For example the recently released MedGemma 1.5 [85], introduced in January 2026, is specifically optimized for medical multimodal tasks and could provide more representative 3D features. Since MMARE is modular at the feature-extractor level, improved encoders can be incorporated without changing the core missing-aware conditioning and fusion logic, making this a practical path for future improvements.

Explainability and model transparency. An important future direction is to complement predictive performance with dedicated explainability analyses, supporting safer deployment in clinical workflows. Integrating explainability methods would help to better understand and validate the decision process of MMARE under missing modalities, improving transparency and trustworthiness for downstream clinical use. Additionally such analyses could be useful for the novel Brain Lifespan Epoch Prediction task introduced in this thesis, to assess which modality most influence the prediction.

Overall, future research should aim to extend the applicability of MMARE beyond the current benchmarks, test it under different missingness mechanisms and further improve representation quality and reliability. These directions would enlarge the contribution of this thesis from missing-modality robustness toward even more general and deployable multimodal learning systems.

Bibliography

[1] Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. «Review of multimodal machine learning approaches in healthcare». In: *Information Fusion* 114 (2025), p. 102690 (cit. on pp. 1, 2, 16, 27).

[2] Valerio Guarrasi, Fatih Aksu, Camillo Maria Caruso, Francesco Di Feola, Aurora Rofena, Filippo Ruffini, and Paolo Soda. «A systematic review of intermediate fusion in multimodal deep learning for biomedical applications». In: *Image and Vision Computing* (2025), p. 105509 (cit. on pp. 1, 2, 15, 29, 31).

[3] Zhaoyi Sun, Mingquan Lin, Qingqing Zhu, Qianqian Xie, Fei Wang, Zhiyong Lu, and Yifan Peng. «A scoping review on multimodal deep learning in biomedical images and texts». In: *Journal of Biomedical Informatics* 146 (2023), p. 104482. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2023.104482. URL: https://www.sciencedirect.com/science/article/pii/S1532046423002034 (cit. on pp. 1, 2).

[4] Yuyin Zhou et al. «Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr». In: *arXiv preprint arXiv:2111.11665* (2021) (cit. on p. 2).

[5] Luis R Soenksen et al. «Integrated multimodal artificial intelligence framework for healthcare applications». In: *NPJ digital medicine* 5.1 (2022), p. 149 (cit. on p. 2).

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. «Multimodal machine learning: A survey and taxonomy». In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018), pp. 423–443 (cit. on pp. 2–4, 15, 28).

[7] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. «Foundations & trends in multimodal machine learning: Principles, challenges, and open

questions». In: *ACM Computing Surveys* 56.10 (2024), pp. 1–42 (cit. on pp. 2, 4, 19).

[8] Kapil Joshi, Mohit Kumar, Amrendra Tripathi, Anuj Kumar, Jitender Sehgal, and Archana Barthwal. «Latest Trends in Multi-modality Medical Image Fusion: A Generic Review». In: *Rising Threats in Expert Applications and Solutions*. Ed. by Vijay Singh Rathore, Subhash Chander Sharma, Joao Manuel R.S. Tavares, Catarina Moreira, and B. Surendiran. Singapore: Springer Nature Singapore, 2022, pp. 663–671 (cit. on p. 3).

[9] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. *Deep Multimodal Learning with Missing Modality: A Survey*. 2024. arXiv: 2409.07825 [cs.CV]. URL: https://arxiv.org/abs/2409.07825 (cit. on pp. 3, 4, 15–19).

[10] Alexandre Perez-Lebel, Gaël Varoquaux, Marine Le Morvan, Julie Josse, and Jean-Baptiste Poline. «Benchmarking missing-values approaches for predictive models on health databases». In: *GigaScience* 11 (2022), giac013 (cit. on pp. 3, 28, 29).

[11] Tiago Mota, M Rita Verdelho, Diogo J Araújo, Alceu Bissoto, Carlos Santiago, and Catarina Barata. «Mmist-ccrcc: A real world medical dataset for the development of multi-modal systems». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 2395–2403 (cit. on pp. 3, 10, 16, 26, 39, 40, 42, 43).

[12] Hava Chaptoukaev, Vincenzo Marcianó, Francesco Galati, and Maria A. Zuluaga. *HyperMM : Robust Multimodal Learning with Varying-sized Inputs*. 2024. arXiv: 2407.20768 [cs.LG]. URL: https://arxiv.org/abs/2407.20768 (cit. on pp. 4, 5, 17–20, 27, 44, 53).

[13] Alexa Mousley, Richard AI Bethlehem, Fang-Cheng Yeh, and Duncan E Astle. «Topological turning points across the human lifespan». In: *Nature communications* 16.1 (2025), p. 10055 (cit. on pp. 5, 35).

[14] Mario Mustra, Kresimir Delac, and Mislav Grgic. «Overview of the DICOM standard». In: *2008 50th international symposium ELMAR*. Vol. 1. IEEE. 2008, pp. 39–44 (cit. on p. 7).

[15] 3D Slicer Community. *Coordinate Systems*. https://www.slicer.org/wiki/Coordinate_systems. Accessed: 2026-02-15. 2024 (cit. on pp. 8, 9).

[16] Paul A Yushkevich, Yang Gao, and Guido Gerig. «ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images». In: *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2016, pp. 3342–3345 (cit. on p. 9).

[17] M Jorge Cardoso et al. «Monai: An open-source framework for deep learning in healthcare». In: *arXiv preprint arXiv:2211.02701* (2022) (cit. on pp. 9, 21, 37, 41).

[18] Steve Pieper, Michael Halle, and Ron Kikinis. «3D Slicer». In: *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)*. IEEE. 2004, pp. 632–635 (cit. on p. 9).

[19] Bruce Fischl. «FreeSurfer». In: *Neuroimage* 62.2 (2012), pp. 774–781 (cit. on p. 9).

[20] Reza Azad, Nika Khosravi, Mohammad Dehghanmanshadi, Julien Cohen-Adad, and Dorit Merhof. «Medical image segmentation on mri images with missing modalities: A review». In: *arXiv preprint arXiv:2203.06217* (2022) (cit. on pp. 11, 19).

[21] Imperial College London. *IXI Dataset: Information eXtraction from Images*. `https://brain-development.org/ixi-dataset/`. Dataset collected as part of EPSRC grant GR/S21533/02. 2006 (cit. on pp. 11, 12).

[22] Peng Xu, Xiatian Zhu, and David A Clifton. «Multimodal learning with transformers: A survey». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023) (cit. on p. 15).

[23] Luís A Vale-Silva and Karl Rohr. «Long-term cancer survival prediction using multimodal deep learning». In: *Scientific Reports* 11.1 (2021), p. 13505 (cit. on p. 15).

[24] Sandra Steyaert, Yeping Lina Qiu, Yuanning Zheng, Pritam Mukherjee, Hannes Vogel, and Olivier Gevaert. «Multimodal deep learning to predict prognosis in adult and pediatric brain tumors». In: *Communications Medicine* 3.1 (2023), p. 44 (cit. on p. 15).

[25] Wei Huang et al. «LIDIA: Precise Liver Tumor Diagnosis on Multi-Phase Contrast-Enhanced CT via Iterative Fusion and Asymmetric Contrastive Learning». In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 394–404 (cit. on pp. 15, 29).

[26] Lucas Robinet, Ahmad Berjaoui, Ziad Kheil, and Elizabeth Cohen-Jonathan Moyal. « DRIM: Learning Disentangled Representations from Incomplete Multimodal Healthcare Data ». In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Vol. LNCS 15003. Springer Nature Switzerland, Oct. 2024 (cit. on pp. 15, 19, 44).

[27] Yue Zhang, Chengtao Peng, Qiuli Wang, Dan Song, Kaiyan Li, and S Kevin Zhou. «Unified multi-modal image synthesis for missing modality imputation». In: *IEEE Transactions on Medical Imaging* 44.1 (2024), pp. 4–18 (cit. on pp. 16, 17).

[28] Tongzhou Wu and Max Goodman. *Multimodal Variational Autoencoder*. 2018. DOI: 10.48550/arXiv.1802.05335. arXiv: 1802.05335 [stat.ML] (cit. on pp. 16, 17).

[29] Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan. «Latent correlation representation learning for brain tumor segmentation with missing MRI modalities». In: *IEEE Transactions on Image Processing* 30 (2021), pp. 4263–4274 (cit. on p. 16).

[30] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. «Are Multimodal Transformers Robust to Missing Modality?» In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 18177–18186 (cit. on pp. 17, 19).

[31] Sijie Li, Chen Chen, and Jungong Han. «Simmlm: A simple framework for multi-modal learning with missing modality». In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 24068–24077 (cit. on pp. 17, 18).

[32] Reza Azad, Nika Khosravi, and Dorit Merhof. *SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities*. 2022. arXiv: 2204.02961 [cs.CV]. URL: https://arxiv.org/abs/2204.02961 (cit. on pp. 17, 19).

[33] Junjie Shi, Caozhi Shang, Zhaobin Sun, Li Yu, Xin Yang, and Zengqiang Yan. «Passion: Towards effective incomplete multi-modal medical image segmentation with imbalanced missing rates». In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, pp. 456–465 (cit. on pp. 17, 19).

[34] Srinivas Parthasarathy and Shiva Sundaram. «Training strategies to handle missing modalities for audio-visual expression recognition». In: *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 2020, pp. 400–404 (cit. on p. 17).

[35] Qian Zhou, Hua Zou, Haifeng Jiang, and Yong Wang. «Incomplete Multimodal Learning for Visual Acuity Prediction After Cataract Surgery Using Masked Self-Attention». In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 735–744 (cit. on p. 17).

[36] Boqi Chen, Junier Oliva, and Marc Niethammer. «A unified model for longitudinal multi-modal multi-view prediction with missingness». In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 410–420 (cit. on pp. 17, 18, 44).

[37] Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. «Learnable cross-modal knowledge distillation for multi-modal learning with missing modality». In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 216–226 (cit. on p. 17).

[38] Yuhang Sun, Zhizhong Liu, Quan Z Sheng, Dianhui Chu, Jian Yu, and Hongxiang Sun. «Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities». In: *Information Fusion* 110 (2024), p. 102454 (cit. on p. 17).

[39] Nikhilanand Arya and Sriparna Saha. «Generative Incomplete Multi-View Prognosis Predictor for Breast Cancer: GIMPP». In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.4 (2022), pp. 2252–2263. DOI: 10.1109/TCBB.2021.3090458 (cit. on p. 17).

[40] Qianqian Chen, Jiadong Zhang, Runqi Meng, Lei Zhou, Zhenhui Li, Qianjin Feng, and Dinggang Shen. «Modality-Specific Information Disentanglement From Multi-Parametric MRI for Breast Tumor Segmentation and Computer-Aided Diagnosis». In: *IEEE Transactions on Medical Imaging* 43.5 (2024), pp. 1958–1971. DOI: 10.1109/TMI.2024.3352648 (cit. on p. 17).

[41] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. «Smil: Multimodal learning with severely missing modality». In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2021, pp. 2302–2310 (cit. on p. 17).

[42] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. «Multi-modal learning with missing modality via shared-specific feature modelling». In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 15878–15887 (cit. on p. 17).

[43] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. «Deep adversarial learning for multi-modality missing data completion». In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 1158–1166 (cit. on p. 17).

[44] Joo-Chang Kim and Kyungyong Chung. «Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data». In: *IEEE Access* 8 (2020), pp. 104933–104943 (cit. on p. 17).

[45] Wangbin Sun, Fei Ma, Yang Li, Shao-Lun Huang, Shiguang Ni, and Lin Zhang. «Semi-supervised multimodal image translation for missing modality imputation». In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 4320–4324 (cit. on p. 17).

[46] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. «Image-to-image translation with conditional adversarial networks». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134 (cit. on p. 17).

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. «Unpaired image-to-image translation using cycle-consistent adversarial networks». In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232 (cit. on p. 17).

[48] Yuanzhi Wang, Yong Li, and Zhen Cui. «Incomplete Multimodality-Diffused Emotion Recognition». In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 17117–17128. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/372cb7805eaccb2b7eed641271a30eec-Paper-Conference.pdf (cit. on p. 17).

[49] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. «Missing Modalities Imputation via Cascaded Residual Autoencoder». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017 (cit. on p. 17).

[50] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. «Training generative adversarial networks with limited data». In: *Advances in neural information processing systems* 33 (2020), pp. 12104–12114 (cit. on p. 17).

[51] Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. «Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency». In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. 15. 2024, pp. 16416–16424 (cit. on pp. 18–20, 28, 30).

[52] Biting Yu, Luping Zhou, Lei Wang, Jurgen Fripp, and Pierrick Bourgeat. «3D cGAN based cross-modality MR image synthesis for brain tumor segmentation». In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 626–630 (cit. on p. 18).

[53] Tolou Shadbahr et al. «The impact of imputation quality on machine learning classifiers for datasets with missing values». In: *Communications Medicine* 3.1 (2023), p. 139 (cit. on p. 18).

[54] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. «What's a good imputation to predict with missing values?» In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11530–11540 (cit. on p. 18).

[55] Zhou Lu. «A Theory of Multimodal Learning». In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on pp. 18, 58).

[56] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. «Hemis: Hetero-modal image segmentation». In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 469–477 (cit. on p. 18).

[57] Yixin Wang, Yang Zhang, Yang Liu, Zihao Lin, Jiang Tian, Cheng Zhong, Zhongchao Shi, Jianping Fan, and Zhiqiang He. «Acn: Adversarial co-training network for brain tumor segmentation with missing modalities». In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2021, pp. 410–420 (cit. on pp. 18, 19).

[58] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. «Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks». In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019. arXiv: 1810.00825 [cs.LG] (cit. on pp. 18, 19).

[59] Wanyi Chen, Zihua Zhao, Jiangchao Yao, Ya Zhang, Jiajun Bu, and Haishuai Wang. «Multi-modal Medical Diagnosis via Large-small Model Collaboration». In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 30763–30773 (cit. on pp. 18, 39, 44).

[60] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. «Multimodal learning with incomplete modalities by knowledge distillation». In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1828–1838 (cit. on p. 19).

[61] Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. «Multimodal Patient Representation Learning with Missing Modalities and Labels». In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=Je5SHCKpPa (cit. on p. 19).

[62] Tristan Sylvain, Francis Dutil, Tess Berthier, Lisa Di Jorio, Margaux Luck, Devon Hjelm, and Yoshua Bengio. «Cross-modal information maximization for medical imaging: Cmim». In: *arXiv preprint arXiv:2010.10593* (2020) (cit. on p. 19).

[63] Karen Simonyan and Andrew Zisserman. «Very deep convolutional networks for large-scale image recognition». In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on pp. 22, 44).

[64] Vinod Kumar Chauhan, Jiandong Zhou, Ping Lu, Soheila Molaei, and David A Clifton. «A Brief Review of Hypernetworks in Deep Learning». In: *arXiv preprint arXiv:2306.06955* (2023) (cit. on p. 22).

[65] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. «Deep Sets». In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. arXiv: 1703.06114 [cs.LG] (cit. on p. 23).

[66] Sihong Chen, Kai Ma, and Yefeng Zheng. «Med3d: Transfer learning for 3d medical image analysis». In: *arXiv preprint arXiv:1904.00625* (2019) (cit. on pp. 27, 37, 42).

[67] Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. «Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities». In: *arXiv preprint arXiv:2412.10372* (2024) (cit. on pp. 27, 38, 44).

[68] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. «Modulating early visual processing by language». In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 28).

[69] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. «FiLM: Visual Reasoning with a General Conditioning Layer». In: *International Conference on Learning Representations (ICLR)*. 2018. arXiv: 1709.07871 [cs.LG] (cit. on p. 28).

[70] Shirin Heidari, Thomas F Babor, Paola De Castro, Sera Tort, and Mirjam Curno. «Sex and gender equity in research: rationale for the SAGER guidelines and recommended use». In: *Research integrity and peer review* 1.1 (2016), p. 2 (cit. on p. 35).

[71] Srikanth Ryali, Yuan Zhang, Carlo de Los Angeles, Kaustubh Supekar, and Vinod Menon. «Deep learning models reveal replicable, generalizable, and behaviorally relevant sex differences in human functional brain organization». In: *Proceedings of the National Academy of Sciences* 121.9 (2024), e2310012121 (cit. on p. 35).

[72] Stuart J Ritchie et al. «Sex differences in the adult human brain: evidence from 5216 UK biobank participants». In: *Cerebral cortex* 28.8 (2018), pp. 2959–2975 (cit. on p. 35).

[73] Srushti Honnangi, Anushri Kajagar, Shashank Shetgeri, Tanvi Korgaonkar, Salma Shahapur, and Rajashri Khanai. «Gender-Aware ADHD Detection Framework Combining XGBoost and FLAML Models: Exploring Predictive Features in Women Advancing Personalized ADHD Diagnosis». In: *Computer Sciences & Mathematics Forum*. Vol. 12. 1. MDPI. 2025, p. 6 (cit. on p. 35).

[74] Mahsa Dibaji, Johanna Ospel, Roberto Souza, and Mariana Bento. «Sex differences in brain MRI using deep learning toward fairer healthcare outcomes». In: *Frontiers in Computational Neuroscience* 18 (2024), p. 1452457 (cit. on p. 35).

[75] Andrew Hoopes, Jocelyn S Mora, Adrian V Dalca, Bruce Fischl, and Malte Hoffmann. «SynthStrip: skull-stripping for any brain image». In: *NeuroImage* 260 (2022), p. 119474 (cit. on p. 37).

[76] Andreas M Hötker, Christoph A Karlo, Junting Zheng, Chaya S Moskowitz, Paul Russo, Hedvig Hricak, and Oguz Akin. «Clear cell renal cell carcinoma: associations between CT features and patient survival». In: *American Journal of Roentgenology* 206.5 (2016), pp. 1023–1030 (cit. on p. 39).

[77] Kaoutar Ben Ahmed, Lawrence O Hall, Dmitry B Goldgof, and Robert Gatenby. «Ensembles of convolutional neural networks for survival time estimation of high-grade glioma patients from multimodal MRI». In: *Diagnostics* 12.2 (2022), p. 345 (cit. on p. 39).

[78] Fabian Isensee and Klaus H Maier-Hein. «An attempt at beating the 3D U-Net». In: *arXiv preprint arXiv:1908.02182* (2019) (cit. on p. 41).

[79] Jakob Wasserthal et al. «TotalSegmentator: robust segmentation of 104 anatomic structures in CT images». In: *Radiology: Artificial Intelligence* 5.5 (2023), e230024 (cit. on p. 41).

[80] Nobuyuki Otsu. «A Threshold Selection Method from Gray-Level Histograms». In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. DOI: 10.1109/TSMC.1979.4310076 (cit. on p. 42).

[81] Martim Afonso, Praphulla MS Bhawsar, Monjoy Saha, Jonas S Almeida, and Arlindo L Oliveira. «Multiple Instance Learning for WSI: A comparative analysis of attention-based approaches». In: *Journal of Pathology Informatics* 15 (2024), p. 100403 (cit. on p. 42).

[82] Adam Goode, Benjamin Gilbert, Jan Harkes, David Jukic, and Milind Satyanarayanan. «OpenSlide: A vendor-neutral software foundation for digital pathology». In: *Journal of Pathology Informatics* 4.1 (2013), p. 27. DOI: 10.4103/2153-3539.119005 (cit. on p. 42).

[83] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. «Data-efficient and weakly supervised computational pathology on whole-slide images». In: *Nature biomedical engineering* 5.6 (2021), pp. 555–570 (cit. on p. 43).

[84] Anthony P Addison, Felix Wagner, Wentian Xu, Natalie Voets, and Konstantinos Kamnitsas. «Modality-Agnostic Input Channels Enable Segmentation of Brain lesions in Multimodal MRI with Sequences Unavailable During Training». In: *arXiv preprint arXiv:2509.09290* (2025) (cit. on p. 59).

[85] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, et al. «MedGemma Technical Report». In: *arXiv preprint arXiv:2507.05201* (2025). URL: `https://arxiv.org/abs/2507.05201` (cit. on p. 59).