Computerized Medical Imaging and Graphics 133 (2026) 102794

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Multimodal clinical data integration for prognosis of pulmonary embolism: A comparative study☆

Domenico Paolo [a], Paolo Soda [a,c,*], Matteo Tortora [b], Alessandro Bria [d], Rosa Sicilia [e]

[a] Unit of Computer Systems and Artificial Intelligence, Department of Engineering, University Campus Bio-Medico of Rome, Italy
[b] Department of Naval, Electrical, Electronics and Telecommunications Engineering, University of Genoa, Italy
[c] Department of Diagnostics and Intervention, Radiation Physics, Biomedical Engineering, Umeå University, Sweden
[d] Department of Electrical and Information Engineering, University of Cassino and Southern Latium, Italy
[e] UniCamillus - Saint Camillus International University of Health Sciences, Italy

ARTICLE INFO

Dataset link: https://github.com/som-shahlab/INSPECT_public

Keywords:
Pulmonary embolism risk prediction
Multimodal data integration
Electronic health records
Medical imaging
Fusion strategies

ABSTRACT

Pulmonary embolism (PE) is a life-threatening condition for which accurate risk stratification is essential for clinical decision-making. Most existing prognostic models rely on a single data modality, typically structured electronic health records (EHR), limiting their ability to capture the heterogeneous nature of the disease. Multimodal approaches combining EHR, computed tomography (CT) imaging, and free-text clinical reports provide complementary information, but the optimal fusion strategy for prognosis prediction remains unclear. In this study, we present a comprehensive comparative study of multimodal fusion strategies for PE prognosis using the INSPECT dataset, which includes structured EHR, CT images, and clinical text with mortality labels. We develop unimodal baselines and multiple bimodal and trimodal models using early, intermediate, late, and cross-fusion strategies within a unified framework. Models are evaluated for 1-, 6-, and 12-month mortality prediction. Multimodal models consistently outperform unimodal baselines across all prediction horizons, with MCC improvements up to +0.13 for 1-month mortality (0.399 vs 0.269) and +0.11 for 6-month mortality (0.479 vs 0.367). The best overall performance is achieved by late fusion models, reaching an MCC of 0.497 for 12-month prediction, compared to 0.454 for the best unimodal model. Interestingly, bimodal combinations often outperform trimodal fusion in short-term prediction, suggesting that improper integration of multiple modalities may introduce noise and reduce effectiveness. These results highlight the importance of fusion design in multimodal clinical AI and provide practical guidance for developing robust prognosis models for pulmonary embolism. Code is publicly available at https://github.com/arco-group/INSPECT-CS.

1. Introduction

Pulmonary embolism (PE) is a potentially life-threatening condition characterized by obstruction of the pulmonary arteries, typically due to thromboembolic events. Despite advances in diagnosis and treatment, PE remains associated with significant short- and medium-term mortality, emphasizing the importance of timely risk stratification and prognostic assessment (Barco et al., 2021). Accurate prediction of patient outcomes is crucial for guiding therapeutic decisions, optimizing resource allocation, and ultimately improving survival rates.

Traditional prognostic models for PE, such as the Pulmonary Embolism Severity Index (PESI) and its simplified versions (Aujesky et al., 2005; Jiménez et al., 2010), rely primarily on demographic and clinical variables, such as age, comorbidities, and vital signs, and have been validated for predicting short-term mortality. Although these tools provide valuable risk estimates, they are limited by their reliance on a narrow set of features and may not fully capture the multidimensional nature of PE. In clinical practice, PE outcomes are influenced not only by demographic and clinical variables but also by imaging findings, textual clinical information, and structured data recorded in electronic health records (EHR). This limitation underscores the need for more comprehensive approaches capable of integrating these heterogeneous data modalities to improve prognostic accuracy.

Recent advances in artificial intelligence (Sun et al., 2023; Huang et al., 2020; Cahan et al., 2023; Zhong et al., 2024) have enabled the integration of multimodal clinical data across various medical contexts. These approaches combine imaging studies, textual reports, and structured EHR data, such as laboratory values, vital signs, and comorbidities, providing complementary information that can improve

☆ This article is part of a Special issue entitled: 'Data fusion in healthcare' published in Computerized Medical Imaging and Graphics.

* Corresponding author at: .
E-mail address: paolo.soda@umu.se (P. Soda).

https://doi.org/10.1016/j.compmedimag.2026.102794
Received 15 December 2025; Received in revised form 20 April 2026; Accepted 28 June 2026
Available online 2 July 2026
0895-6111/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

predictive models' ability to assess patient risk more accurately than unimodal approaches. Multimodal fusion is the process of combining these diverse data sources into a unified representation for predictive modeling. However, although multimodal learning has gained significant attention, the optimal fusion strategy, including when to integrate information, which modalities to combine, and how to design the fusion mechanism, remains an open and task-dependent challenge (Jandoubi and Akhloufi, 2025b,a; Liu et al., 2025a,b; Simon et al., 2025; Stahlschmidt et al., 2022).

Despite the growing interest in multimodal learning, current approaches in PE still face several important limitations. Most studies focus on short-term outcomes or diagnostic tasks rather than clinically actionable prognostic horizons (1-, 6-, and 12-month mortality). They often evaluate only subsets of available modalities, leaving the incremental value of additional data sources unexplored. Finally, guidance on optimal fusion strategies remains limited, hindering clinical translation.

To address these gaps, our main contributions are summarized as follows:

- **Systematic multimodal comparison:** We present the first comprehensive evaluation of unimodal (CT, radiology reports, structured EHR) and multimodal models for PE prognosis across multiple clinically relevant time horizons (1-, 6-, 12-months) using the INSPECT dataset (Huang et al., 2023).
- **Addressing prognostic scope:** We focus on outcomes beyond disease detection, systematically evaluating short- and medium-term mortality.
- **Comprehensive modality coverage and clinical insight:** We analyze all three key modalities (CT, text, EHR) and their combinations, quantifying the incremental value of each through changes in predictive performance metrics. We also highlight scenarios in which specific data types contribute most to prediction, providing interpretable guidance for clinical decision-making.
- **Fusion strategies:** We empirically investigate *when*, *how*, and *which* modalities should be fused to maximize predictive performance.
- **Benchmarking and reproducibility:** We establish baselines on INSPECT, offering a reproducible framework for future multimodal PE studies.

Together, these contributions provide both a methodological and practical framework for developing multimodal prognostic models in pulmonary embolism and related cardiovascular conditions.

2. Related work

Risk stratification in PE traditionally relies on validated clinical scores. The Pulmonary Embolism Severity Index (PESI) and its simplified version (sPESI) are the most widely used tools for estimating short-term mortality; they combine demographic and readily available clinical variables to stratify patients into risk classes and have been externally validated in large cohorts (Aujesky et al., 2005; Jiménez et al., 2010). Whilst these scores remain useful in practice, their reliance on a small set of structured inputs limits their ability to capture the full clinical complexity of many PE presentations, especially information embedded in imaging studies and free-text radiology reports.

In recent years, the advent of machine learning (ML) and deep learning (DL) has led to a surge of research efforts aimed at enhancing PE assessment. Many of these approaches focus on single modalities, such as structured clinical data, imaging studies, or radiology text reports. More recently, multimodal deep learning (MDL) has emerged as a strategy to integrate complementary information from multiple sources, aiming to overcome the limitations of unimodal models and provide more comprehensive and robust predictions.

In this section, we review representative studies on PE, organized into two complementary perspectives. First, we discuss unimodal approaches that focus on a single data source, including structured clinical variables, imaging studies, or radiology text reports. Next, we examine MDL approaches that integrate information across multiple modalities, leveraging the complementary strengths of each. Although the cited studies are representative of the state of the art, they are not exhaustive and serve to illustrate key trends and challenges in the field.

2.1. Unimodal approaches

Unimodal approaches focus on a single data source (clinical, imaging, or textual) and have demonstrated promise to improve PE diagnostic accuracy and risk prediction beyond traditional scoring systems.

EHR-based models, for instance, rely on classical statistical and ML techniques, such as logistic regression, gradient-boosted trees, and deep multilayer perceptrons, trained on structured EHR data encompassing demographics, vital signs, laboratory results, and comorbidities (Eini et al., 2025; Sadegh-Zadeh et al., 2023). These models have demonstrated the ability to predict short-term outcomes and have been used to augment or recalibrate clinical scores. Their strength lies in leveraging longitudinal data; however, they may miss important imaging-derived markers.

Imaging-only approaches, in contrast, apply deep learning to CT pulmonary angiography (CTPA) for detection and characterization of emboli as well as prognostic tasks (Diaz-Lorenzo et al., 2024). These models can extract morphological features, such as clot burden or proxies for right-to-left ventricular ratio, that are not captured in structured EHR. Nevertheless, their prognostic accuracy can be limited when clinical context, including comorbidities or laboratory values, is decisive for patient outcomes.

Natural language processing (NLP) applied to radiology reports represents a third modality for PE research. Early rule-based systems have progressively evolved into transformer-based architectures capable of contextual semantic understanding at scale. Most studies have focused on PE detection and classification (Rashedi et al., 2024; Hassan et al., 2025), but recent efforts have begun to explore prognostic applications (O'Gallagher et al., 2025; Amin et al., 2024). Radiology reports encapsulate a wealth of qualitative indicators, such as clot burden, vascular obstruction index, and right-heart strain, that are increasingly recognized as potential sources of prognostic insight not available in structured EHR or imaging data.

2.2. Multimodal deep learning

Multimodal deep learning (MDL) has been proposed as a way to aggregate complementary information and improve predictive performance because each unimodal approach offers distinct strengths, but individually they may not fully account for the heterogeneity of PE patients across short- and medium-term horizons. In MDL, the integration of data from multiple sources can be performed using three main fusion strategies: early, late, and intermediate fusion. Early fusion concatenates features from all modalities at the input level before training a single model. It is simple and computationally efficient but may struggle with heterogeneous data. Late fusion trains separate models for each modality and combines their outputs, for example via averaging or a meta-classifier. This preserves modality-specific learning and handles missing modalities, but cannot capture deep cross-modal interactions. Intermediate fusion integrates information at intermediate layers within the network, allowing the model to learn complex cross-modal relationships while retaining modality-specific features. Its performance, however, can be sensitive to data heterogeneity and noise (Guarrasi et al., 2025). Recent advances in multimodal learning have increasingly explored transformer-based architectures and attention-driven fusion mechanisms to better model interactions between heterogeneous data sources. Multimodal transformers enable

joint representation learning across modalities by using self-attention and cross-attention layers, allowing the model to capture complex dependencies between clinical variables, imaging features, and textual information (Xu et al., 2023). Cross-attention fusion, in particular, allows one modality to guide the feature extraction of another, which has shown promising results in several medical AI applications involving multimodal prediction tasks, as demonstrated in (Dhar et al., 2025; Zhang et al., 2024).

Despite their flexibility, transformer-based multimodal models typically require large-scale datasets and careful regularization to avoid overfitting, especially in clinical settings where cohorts are relatively small and modalities are highly heterogeneous (Liu et al., 2025a,b). For this reason, simpler fusion strategies such as early, late, and intermediate fusion remain widely used in medical prognosis studies, as they often provide a better trade-off between performance, robustness, and interpretability (Teoh et al., 2024; Sucre et al., 2025; Liu et al., 2025a,b). In this work, we therefore evaluate a broad spectrum of fusion strategies, including early, late, cross, and intermediate fusion schemes, to systematically assess their behavior in a realistic clinical prognosis scenario.

Across domains such as cardiology, oncology, and orthopedics, systematic reviews and methodological analyses have consistently reported that integrating multimodal data using early, late, or intermediate fusion strategies leads to improved predictive performance (Guarrasi et al., 2025; Jandoubi and Akhloufi, 2025b,a; S.-C. Huang et al., 2020; Shih-Cheng Huang et al., 2020; Sun et al., 2023; Kline et al., 2022; Mohsen et al., 2022).

In the context of PE, MDL models combining imaging and clinical data consistently outperformed both unimodal approaches and traditional clinical scoring systems across diagnostic and prognostic applications (Huang et al., 2020; Guo et al., 2024; Zhang et al., 2025; Cicek et al., 2025; Cahan et al., 2023; Zhong et al., 2024).

For binary PE detection, Huang et al. (2020) integrated imaging and EHR features using a late fusion strategy on a private dataset of 1837 CTPA studies, demonstrating that fused models outperform unimodal baselines. Similarly, using the RadFusion dataset (Zhou et al., 2021), both Guo et al. (2024) and Zhang et al. (2025) applied intermediate fusion to integrate imaging and EHR features, further confirming the benefits of multimodal approaches compared to unimodal baselines.

For PE-related mortality and severity prediction, multimodal approaches have similarly demonstrated substantial gains over traditional clinical tools. Cicek et al. (2025), using a cohort of 207 patients, showed that models integrating CTPA and clinical data significantly outperformed the PESI score for binary short-term severity prediction. In a separate study on a private cohort of 358 patients, Cahan et al. (2023) evaluated several fusion strategies combining CTPA and structured clinical data for 30-day severity prediction, reporting their best performance with intermediate fusion. More recently, Zhong et al. (2024) incorporated CTPA, clinical variables, and the PESI score into a cross-modal CoxPH framework, demonstrating on an internal dataset of 918 patients that multimodal fusion can enhance 30-day survival prediction.

A major development in PE research was the INSPECT dataset (Huang et al., 2023), the first large-scale, publicly available multimodal cohort that jointly includes CTPA images, radiology report impressions, and structured EHR data, along with diagnostic and prognostic labels. INSPECT provides a standardized benchmark for PE detection and outcome prediction, offering unimodal (CT, EHR) and multimodal (CT + EHR) baselines that primarily implement late fusion strategies. These benchmarks further reinforce the evidence observed in other studies: integrating imaging and clinical data yields clear advantages over unimodal models for both diagnostic and prognostic tasks.

2.3. Limitations and motivations

MDL approaches in PE highlight three critical limitations that directly motivate our contribution. First, the predominant focus remains on diagnostic or short-term predictive tasks rather than clinically actionable medium- and long-term prognosis. Second, most approaches rely on only two modalities, typically CT imaging and structured EHR data, thus overlooking the rich qualitative and semi-structured information contained in radiology reports. Third, reproducibility remains limited due to the use of proprietary datasets, hindering broader validation and clinical translation.

A major step toward reproducible multimodal research was the release of the INSPECT dataset (Huang et al., 2023), the first large-scale public cohort that jointly provides CTPA images, radiology report impressions, and longitudinal structured EHR data (including demographics, diagnoses, vitals, laboratory values, and medications), together with diagnostic and prognostic labels. INSPECT offers a standardized foundation for multimodal PE modeling; however, existing baselines primarily adopt late fusion and remain restricted to CT and structured EHR data, without leveraging the textual modality, leaving unexplored the potential benefits of incorporating report narratives or more advanced fusion strategies.

In this work, we leverage INSPECT to systematically evaluate unimodal and multimodal models for PE prognosis, comparing early, intermediate, and late fusion schemes and quantifying the contribution of each modality across clinically meaningful prediction horizons.

This approach not only complements prior diagnostic-focused studies but also provides a reproducible foundation for future multimodal prognostic research in PE.

3. Materials

We conducted our experiments on the INSPECT dataset (Huang et al., 2023), the only publicly available multimodal resource for PE prognosis, which includes paired CTPA scans, radiology reports, and structured EHR data. The original dataset contains in total 23,248 CTPA studies from 19,402 patients, with some patients undergoing multiple CTPA exams. Our analysis is performed at the CTPA-study level. Accordingly, the dataset provides multimodal information collected around the "index CTPA study", defined as the CT angiography performed during the clinical visit in which pulmonary embolism was suspected or diagnosed.

Each sample includes the full 3D CTPA volume acquired at this index visit (acquisition details are reported in Huang et al. (2023)) and the associated radiology report, which contains only the "Impression" section summarizing key PE-related findings. Full narrative reports are not provided.

The structured EHR data consist of the entire longitudinal clinical history available for each sample, including diagnoses (ICD codes), medications, procedures, laboratory tests, and vital signs. Importantly, INSPECT does not pre-align EHR events to the CTPA exam: each patient's EHR timeline is provided in full. In our study, we specifically selected only those events occurring prior to the index CTPA timestamp to ensure that prognostic modeling is based strictly on pre-diagnostic information, thereby avoiding any potential data leakage and reflecting clinically realistic prediction scenarios. We excluded samples for whom the timestamp of the index CTPA study was missing, since this prevents consistent extraction of pre-exam EHR events.

INSPECT provides prognostic outcomes at 1-, 6-, and 12-months, enabling systematic evaluation across short- and medium-term horizons. Outcomes are annotated as positive (death within the target window), negative (survival), or censored (insufficient follow-up). In line with the original benchmarks, we considered only positive and negative samples, discarding censored cases, as shown in Table 1. This resulted in 21,192 CTPA studies for the 1-month mortality task, 20,156 for the 6-month task, and 19,284 for the 12-month task. Across these horizons,

Table 1
Distribution of prognostic labels across 1-, 6-, and 12-month mortality prediction tasks in the INSPECT dataset. Each case includes a CTPA scan, its corresponding radiology report, and structured EHR data. Patient-level splitting ensures that all cases from the same patient are assigned to the same set (train, validation, or test). Percentages in splits indicate the proportion of positive versus negative cases.

Time	Label	All	Train	Val	Test
1 m	pos.	1143	939 (5.45%)	53 (5.23%)	151 (5.14%)
	neg.	20 049	16 305 (94.55%)	960 (94.77%)	2784 (94.86%)
	cens.	2056	–	–	–
6 m	pos.	2302	1891 (11.54%)	99 (10.21%)	312 (11.17%)
	neg.	17 854	14 501 (88.46%)	871 (89.79%)	2482 (88.83%)
	cens.	3092	–	–	–
12 m	pos.	2811	2303 (14.59%)	125 (13.51%)	383 (14.25%)
	neg.	16473	13 369 (85.31%)	800 (86.49%)	2304 (85.75%)
	cens.	3964	–	–	–

each patient contributes with slightly more than one study, with values ranging from a minimum of 1 to a maximum of 13 CTPA studies per patient across all prognostic tasks.

To ensure reproducibility and prevent information leakage, particularly because individual patients may undergo multiple CTPA exams, we adopted the official patient-level INSPECT splits, which assign CTPA studies from each patient exclusively to a single set (training, validation, or test).

3.1. Preprocessing

Each modality in INSPECT required a tailored preprocessing pipeline to ensure comparability across samples and suitability for downstream modeling.

CTPA scans. For CTPA preprocessing we follow the same pipeline described in Huang et al. (2023). Each CTPA exam was preprocessed by extracting the pixel data from the original DICOM format. Raw intensities were converted into Hounsfield Units (HU). To reduce computational cost, all slices were resized to 256 × 256 pixels and subsequently cropped around the lung region, resulting in a final resolution of 224 × 224.

To emphasize clinically relevant structures, three standard viewing windows were applied, each optimized for a specific diagnostic perspective: (i) lung parenchyma, (ii) pulmonary embolism, and (iii) mediastinum. The corresponding windowed images were stacked to form a three-channel representation of each slice (224 × 224 × 3). Finally, pixel intensities were normalized to zero mean and unit variance using ImageNet statistics (mean and standard deviation), ensuring compatibility with pretrained convolutional backbones (Deng et al., 2009).

Radiology reports. Free-text radiology reports were segmented into sentences using a custom algorithm designed to preserve clinically meaningful structures, including numbered findings (e.g., "1. PROBABLE"). The algorithm temporarily replaces periods in numbered lists (e.g., "1".) to prevent incorrect sentence splits, then applies standard sentence-ending punctuation rules by identifying spaces following ".", "?", or "!". Finally, the original numbering is restored. This procedure is summarized in Algorithm 1 and ensures that both the overall report structure and individual observations are accurately preserved. After sentence segmentation, each sentence was tokenized using the Clinical-Longformer tokenizer (Li et al., 2022), ensuring compatibility with the subsequent transformer-based encoding.

Structured EHR data. Structured EHR events were represented using a count-based featurization strategy (Rajkomar et al., 2018), producing a sparse patient-by-code matrix. Each entry in this matrix indicates the frequency with which a specific clinical code (e.g., diagnoses, medications, procedures, lab tests) occurs in a patient's record prior to the index "CTPA exam", ensuring that only information available at prediction time is used. This representation provides a compact summary of longitudinal patient history, retaining at the same time rich information about prior comorbidities, interventions, and monitoring.

Algorithm 1 Custom Sentence Splitter for Radiology Reports

```
procedure CustomSentenceSplitter(text)
    Input: raw radiology report text
    Output: list of sentences
    Replace all "\d+.\s" with "\1###"
    Split text at spaces following '.', '?', or '!'
    Restore numbering by replacing "###" with ". "
    Trim whitespace in each sentence
    Remove empty sentences
    return list of sentences
```

To incorporate EHR features into deep learning models, which typically require dense, low-dimensional representations, we investigated two complementary dimensionality reduction strategies.

The first approach relied on truncated singular value decomposition (TSVD) (Hansen, 1987), projecting each patient vector into a 128-dimensional embedding that preserved approximately 99% of the variance of the original sparse matrix. This method retains most of the salient information, substantially reducing computational complexity, thus facilitating integration with the other modalities. We refer to this representation as the EHR-TSVD embedding.

The second approach employed a feed-forward autoencoder trained to reconstruct the original sparse matrix from a compact latent representation. The encoder output defines the EHR-AE embedding, a dense representation that summarizes each patient's longitudinal clinical history and can be seamlessly combined with imaging and text embeddings within multimodal architectures (see Section 4.2 for details).

4. Methods

We implemented both unimodal and multimodal models to systematically evaluate the contribution of different data sources and fusion strategies for PE prognosis. Our methodological pipeline is organized into three main components.

First, we describe the representation extraction process, in which each modality (structured EHR, radiology reports, and CTPA imaging) is encoded into a fixed-length sample-level representation using modality-specific encoders. This section establishes the common feature space shared by all subsequent models.

Next, we present the unimodal baselines, where each modality is modeled independently using architectures tailored to its statistical and structural properties. These baselines serve two purposes: (i) they provide strong single-source predictors for PE prognosis, and (ii) they allow us to quantify the standalone predictive value of each modality before integrating them in multimodal settings.

Finally, we introduce the multimodal fusion models, which combine the representations extracted from different modalities through a set of

Table 2
Dimensionality of the intermediate representations extracted for each modality prior to aggregation.

Modality	Encoder/Method	Representation dimensionality
Structured EHR	Count-based featurization	74,303
Radiology report	Clinical-Longformer	[#Sentences, #Tokens, 768]
CTPA imaging	ResNetV2-101	[#Slices, 6145]

fusion strategies. We detail both bimodal and trimodal configurations. Our primary objective is to assess the added value of multimodal models over unimodal baselines. Among multimodal strategies, bimodal combinations can provide insights into which modality pairs contribute most to predictive performance and may serve as ablation studies related to trimodal fusion.

4.1. Representation extraction

For each data modality (structured EHR, radiology reports, and CTPA imaging) we extracted fixed-length sample-level representations using modality-specific encoders designed to capture complementary aspects of a patient's clinical profile. The dimensionality of the resulting representations is summarized in Table 2.

Structured EHR data were converted into high-dimensional count vectors summarizing each sample's clinical events up to the time of the index CTPA exam. The resulting representation is a sparse vector of size $1 \times 74{,}303$ for each patient.

In radiology reports, the text was first split into sentences and tokens (see Section 3.1) and then encoded using Clinical-Longformer (Li et al., 2022), a transformer model pretrained on large-scale clinical corpora (e.g., MIMIC-III Johnson et al., 2016) and specifically optimized for long clinical documents. The model produced a contextualized embedding $\mathbf{t}_{i,j}$ of size 768 for each token j in each sentence i. This results in a tensor of shape [#sentences, #tokens, 768], where the "#" symbol denotes "number of", thus preserving the sentence-level structure prior to hierarchical aggregation.

For CTPA imaging, after the preprocessing described in Section 3.1, each axial slice was processed using a ResNetV2-101 convolutional encoder pretrained with BigTransfer (Kolesnikov et al., 2020) and subsequently adapted to the pulmonary domain through fine-tuning on the RSPECT dataset (Colak et al., 2021), yielding slice-level representations $\mathbf{c}_i$ of size 6145 before slices-level aggregation. This results in a tensor of shape [#slices, 768], where the "#" symbol denotes "number of". To ensure consistency across samples, we retained the first 250 slices for scans exceeding this number. This cutoff was chosen to standardize input dimensionality at the same time preserving the full thoracic region relevant for prognostic assessment, thus excluding abdominal sections that do not contribute with meaningful information and reducing computational overhead.

These modality-specific representations serve as the inputs for the unimodal baselines and the multimodal fusion models described in the following sections.

4.2. Unimodal models

Unimodal baselines are illustrated in Fig. 1. Each modality was first modeled independently to establish strong single-source predictors before exploring multimodal fusion.

Structured EHR data. For the Unimodal EHR (Structured EHR) experiments, we first introduce a supervised autoencoder (EHR-AE) designed to learn compact, task-specific embeddings from structured, count-based EHR features. Let $\mathbf{x} \in \mathbb{R}^d$ denote a sample-level EHR vector, with $d = 74303$. The encoder computes a latent representation

$$\mathbf{z} = f_{\text{enc}}(\mathbf{x}), \qquad f_{\text{enc}} : \mathbb{R}^d \to \mathbb{R}^{768},$$

through a two-layer architecture with dimensions $d \to 512 \to 768$. The decoder mirrors this structure:

$$\hat{\mathbf{x}} = f_{\text{dec}}(\mathbf{z}), \qquad f_{\text{dec}} : \mathbb{R}^{768} \to \mathbb{R}^d.$$

A classification head predicts mortality:

$$\hat{y} = f_{\text{clf}}(\mathbf{z}), \qquad f_{\text{clf}} : \mathbb{R}^{768} \to [0, 1],$$

implemented as a two-layer MLP ($768 \to 128 \to 1$).

The model is trained end-to-end with a joint loss:

$$\mathcal{L} = \alpha\, \mathcal{L}_{\text{recon}}(\mathbf{x}, \hat{\mathbf{x}}) + \beta\, \mathcal{L}_{\text{BCE}}(y, \hat{y}),$$

with weights $\alpha = 0.5$ and $\beta = 1.0$, chosen to prioritize the mortality prediction task while still encouraging accurate reconstruction of the input. Here,

$$\mathcal{L}_{\text{recon}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2, \qquad \mathcal{L}_{\text{BCE}} = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right].$$

This model learns dense, task-adaptive embeddings from structured EHR data, which can be directly used for downstream multimodal fusion.

For comparison, we also reproduced the gradient boosting EHR model from the INSPECT framework (Huang et al., 2023). Let $f_{\text{GBM}}(\mathbf{x})$ denote the model prediction, defined as an ensemble of T regression trees:

$$f_{\text{GBM}}(\mathbf{x}) = \sum_{t=1}^{T} \eta\, h_t(\mathbf{x}),$$

where η is the learning rate and h_t are the leaf-wise regression trees fitted via gradient boosting on the negative gradients of the loss function. LightGBM efficiently captures non-linear interactions among sparse EHR features and serves as a validated baseline for structured data.

This setup allows us to compare the EHR-AE approach with a literature-based baseline (EHR-GBM), demonstrating the benefits of learned dense embeddings maintaining at the same time methodological continuity with prior work.

Radiology reports. For the Unimodal Radiology Report experiments, we developed a novel hierarchical text encoder that leverages token-level embeddings $\mathbf{t}_{i,j}$ produced by Clinical-Longformer. To capture both local and global semantics within each report, we designed a two-level hierarchical attention mechanism. At the first level, token embeddings $\mathbf{t}_{i,j}$ within each sentence i are aggregated according to learned attention weights, emphasizing clinically relevant terms and producing the sentence embedding $\mathbf{s}_i$:

$$\mathbf{s}_i = \sum_{j=0}^{L} w_{i,j} \mathbf{t}_{i,j}, \tag{1}$$

where $w_{i,j}$ is the weight learnt by the soft attention for the token embedding $\mathbf{t}_{i,j}$, L is the total number of tokens in the input sentence and i ranges from 1 to the total number of sentences N in the radiology report. Since each sentence may contain a variable number of tokens and each radiology report consists of a different number of sentences, both L and N are defined dynamically at runtime for each mini-batch. Specifically, L is set to the maximum number of tokens in a single sentence across all sentences in the minibatch, whereas N is defined as the maximum number of sentences across all radiology reports in the mini-batch. To ensure uniform tensor shapes in the mini-batch, we use padding with all-zero vectors of size 768, which receive a weight of zero from both the token-level and sentence-level attention layers, ensuring they do not contribute to the final representation. It is worth nothing that $\mathbf{s}_i$ has the same size 768 of the token embedding.

At the second level, sentence representations were combined through the Sentence Attention layer, that computes:

$$\mathbf{x} = \sum_{i=0}^{N} w_i \mathbf{s}_i, \tag{2}$$

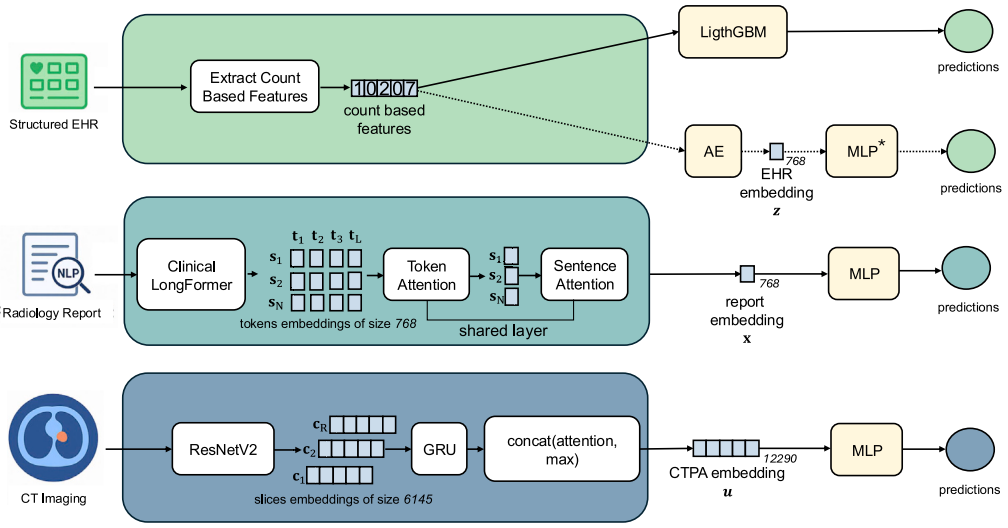


Fig. 1. Overview of unimodal prognostic models. Each modality is processed independently: (i) structured EHR data is featurized with count-based methods and modeled using gradient boosting or autoencoder, (ii) radiology reports are embedded with Clinical-Longformer and aggregated via a two-levels hierarchical attention mechanism (Token Attention and Sentence Attention), and (iii) CTPA slices are encoded with a ResNetV2-101 backbone followed by sequential modeling and pooling. Note that the MLP used for classification in the autoencoder is separate from the MLP classifiers used in the other modalities.

Table 3
Architecture of the MLP classification head.

Layer	Output size	Notes/Activation/Dropout
Input	`enc_size`	Embedding
Linear	512	Fully connected
ReLU	512	Activation
Dropout	512	$p = 0.2$
Linear	128	Fully connected
ReLU	128	Activation
Dropout	128	$p = 0.2$
Linear	1	Fully connected output layer (logit)

where w_i is the weight learnt by the soft attention for the sentence embedding $\mathbf{s}_i$, N is the total number of sentences in the radiology report. It provides a comprehensive representation of the radiology report. Importantly, this layer employs a soft attention mechanism that shares its weights with the token-level attention layer. This sharing is feasible because both token and sentence embeddings have the same dimensionality (768) making it possible to use a single soft attention module implemented as a fully connected layer of size 768. Weight sharing serves two main purposes: (i) it reduces the total number of trainable parameters, helping to prevent overfitting on limited data and lowering computational cost; and (ii) it improves feature transferability across layers, since the clinical relevance of a sentence is inherently related to the clinical relevance of its constituent entities. The final outcome is a weighted average vector, $\mathbf{x}$ of size 768. It acts as the feature vector for the MLP classification network, whose architecture is detailed in Table 3.

CTPA scans. For the Unimodal Image (CTPA) model, ResNetV2-101 slices representations $\mathbf{c}_i$ of size 6145 were processed with a bidirectional GRU to capture sequential dependencies across slices. Then, we applied a hybrid attention-and-max pooling mechanism to aggregate the sequence into a single patient-level representation $\mathbf{u}$:

$$\mathbf{u} = \left[\max_{i=0,\ldots,R} \mathbf{c}_i \;\middle\|\; \sum_{i=0}^{R} w_i \mathbf{c}_i \right], \tag{3}$$

where $\max_{i=0,\ldots,R} \mathbf{c}_i$ denotes the *max pooling* operation over the resulting GRU slice-level embeddings, $\sum_{i=0}^{R} w_i \mathbf{c}_i$ corresponds to the attention-based aggregation (R is the total number of slices in the CTPA scan, w_i is a scalar weight reflecting the contribution of the ith slice, and $\mathbf{c}_i$ represents the feature embedding of the ith slice), whereas the operator $\|$ indicates the concatenation of the two resulting vectors. The unimodal CTPA model adopts the established architecture from INSPECT (Huang et al., 2023), providing a validated baseline for structured comparison with other modalities. The result is a CTPA representation $\mathbf{u}$ of dimension 12 290, which is then fed into the MLP classification network, whose architecture is detailed in Table 3.

4.3. Multimodal fusion models

We explored a range of fusion strategies to integrate complementary information, considering all pairwise modality combinations and the joint use of all three modalities. To ensure a fair and controlled comparison across fusion strategies, modality-specific encoders are kept fixed and provide representations that are used as inputs to the fusion modules. This design allows us to isolate the contribution of the fusion mechanism itself, avoiding confounding effects due to differences in encoder architectures or training procedures, specifically because intermediate fusion models leverage structured representations, such as sentence-level embeddings for reports and slice-level features for CT scans, rather than simple flattened vectors. By processing these through cross-attention layers, the framework captures fine-grained inter-modality interactions without the prohibitive computational overhead or the architectural inconsistencies of fully end-to-end training, thereby ensuring reproducibility while maintaining the capacity for complex cross-modal modeling.

4.3.1. Early fusion

The trimodal Early Fusion strategy is illustrated in Fig. 2. In this approach, the modality-specific embeddings are directly concatenated into a single feature vector, which is then passed to a multi-layer perceptron (MLP) for classification (see Table 3 for architectural details).

Formally, let

$$\mathbf{z} \in \mathbb{R}^{d_{\text{ehr}}}, \quad \mathbf{x} \in \mathbb{R}^{d_{\text{rep}}}, \quad \mathbf{u} \in \mathbb{R}^{d_{\text{img}}},$$

denote the embeddings extracted from structured EHR data, clinical reports, and CTPA images, respectively, where d_{ehr}, d_{rep}, and d_{img} correspond to the dimensionality of the EHR, report, and image embeddings. The embedding $\mathbf{z}$ is obtained by applying either Truncated SVD or an Autoencoder to the sparse matrix of count-based EHR

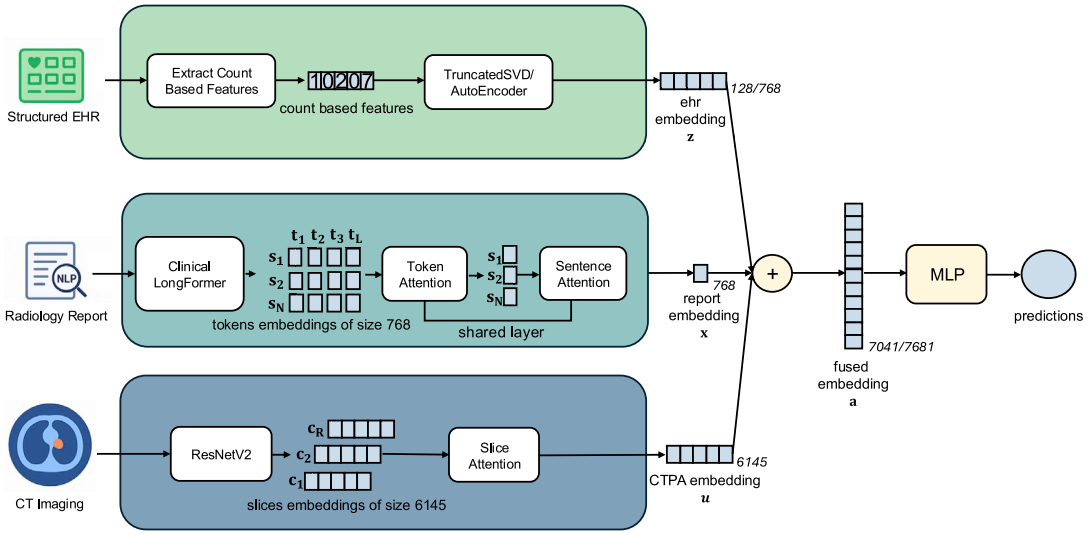


Fig. 2. Overview of the *early fusion* method, where embeddings from EHR, clinical reports, and medical images are concatenated into a joint representation, subsequently processed by an MLP classifier.

Table 4
Overview of fusion strategies used in the experiments. The first column indicates whether the model is unimodal, bimodal, or trimodal. The second column shows which modalities are included (EHR, CT, Report), and the last column reports the embedding dimension for the eraly fusion strategy.

Method	Modalities				Embedding
	EHR-TSVD	EHR-AE	CT	Report	
Unimodal	✓	–	–	–	128
Unimodal	–	✓	–	–	768
Unimodal	–	–	✓	–	12,290
Unimodal	–	–	–	✓	768
Bimodal	✓	–	✓	–	6273
Bimodal	✓	–	–	✓	896
Bimodal	–	✓	✓	–	6913
Bimodal	–	✓	–	✓	1536
Bimodal	–	–	✓	✓	6913
Trimodal	✓	–	✓	✓	7041
Trimodal	–	✓	✓	✓	7681

features, yielding a compact representation of the structured data. The embedding $\mathbf{x}$ is derived from radiology reports processed with Clinical Longformer, followed by the same two-level Hierarchical Attention Mechanism (token-level and sentence-level attention) used in the unimodal report model. Finally, the embedding $\mathbf{u}$ is computed from CT scans by first extracting slice-level representations using ResNetV2. These representations are then passed through an attention mechanism in which a fully connected layer assigns a weight to each slice. The final CT embedding is thus obtained as a weighted aggregation of all slice representations as follows:

$$\mathbf{u} = \sum_{i=1}^{R} w_i \mathbf{c}_i, \tag{4}$$

where w_i denotes the attention weight associated with slice embedding $\mathbf{c}_i$, and R is the total number of slices in the input CTPA volume. Unlike the unimodal CTPA model, we do not include the max pooling operation in the multimodal setting. This choice is motivated by two considerations: (i) the attention mechanism already captures the relative importance of each slice, making max pooling potentially redundant; and (ii) excluding max pooling reduces the dimensionality of the resulting embedding, limiting the number of trainable parameters in the fusion model.

The Early Fusion representation is then obtained by concatenation:

$$\mathbf{a} = [\,\mathbf{z} \,\|\, \mathbf{x} \,\|\, \mathbf{u}\,] \in \mathbb{R}^{d_{\text{ehr}}+d_{\text{rep}}+d_{\text{img}}},$$

where $\|$ denotes the concatenation operator. The fused vector is subsequently fed to the MLP classifier:

$$\hat{\mathbf{y}} = \sigma\left(f_{\text{MLP}}(\mathbf{a})\right),$$

with $f_{\text{MLP}}(\cdot)$ denoting the feed-forward classification network and $\sigma(\cdot)$ the final sigmoid activation function.

In our experiments, the dimensionalities of the modality-specific embeddings are reported in Table 4. Besides the trimodal configuration, we also evaluated all possible bimodal combinations using early fusion, as summarized in the same table.

4.3.2. Intermediate fusion

This strategy leverages joint training with modality-specific encoders, whose latent representations are integrated through a cross-attention mechanism. Unlike Early Fusion, where embeddings are directly concatenated, here the model explicitly learns inter-modality interactions by aligning tokens across different representation spaces.

For EHR data, we use embeddings derived from the sparse matrix of count-based features, obtained either via Truncated SVD or the Autoencoder. For Radiology reports, we adopt embeddings produced by the two-level hierarchical attention mechanism (Paolo et al., 2025), built upon representations generated by the Clinical LongFormer. Finally, for CT data, we utilize slice-level representations extracted using ResNetV2.

To ensure dimensional consistency across inputs, each modality was projected through an embedding layer that maps representations to a common size of 512. In the following subsections, we differentiate between trimodal and bimodal fusion settings.

Trimodal setting. For the trimodal setting, we focused on two intermediate fusion strategies, Joint Fusion with contrastive alignment (ARMOUR) and Cascaded Fusion (CROSS), chosen for their complementary capabilities in modeling cross-modal interactions and their suitability for multimodal clinical data.

In the first strategy, we adopted the ARMOUR framework proposed by Liu et al. (2023), which enables robust bimodal fusion for clinical prediction tasks while maintaining performance in the presence of missing modalities. In our setup, ARMOUR was applied to fuse CTPA imaging and radiology report embeddings, producing a joint image–text

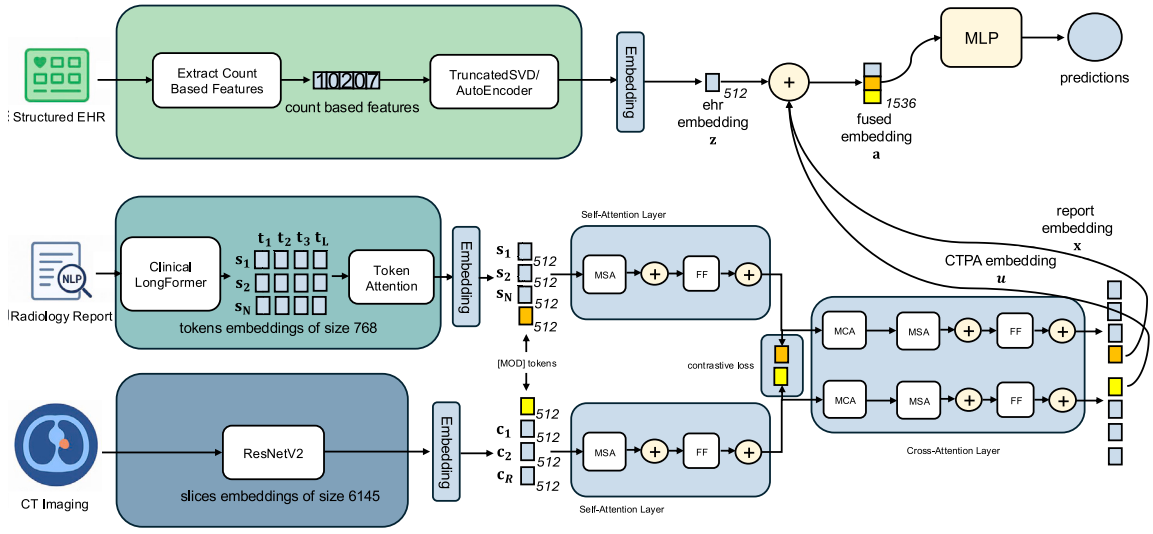


Fig. 3. Conceptual illustration of the ARMOUR workflow, showing cross-attention blocks between modality-specific encoders, the contrastive alignment mechanism, and the final fusion with the EHR embedding.

representation. This fused representation was then combined with the EHR embedding, realizing a trimodal fusion.

To enable cross-modal interaction, each modality is first projected into a shared embedding space ("Embedding" in Fig. 3) of equal dimensionality through modality-specific projection layers. Subsequently, modality-specific class tokens are introduced for each modality (e.g., imaging and reports) to act as compact global representations that summarize modality-level semantics. These class tokens are then refined via cross-attention layers, where one modality attends to the other (e.g., text attending to image features and vice versa), allowing the model to capture fine-grained interactions and contextual dependencies across modalities.

A contrastive alignment strategy, inspired by Momentum Contrast (MoCo, He et al., 2020; Chen et al., 2020), is applied to these class tokens. The contrastive loss encourages paired embeddings from the same sample to be close in the latent space while pushing apart embeddings from different samples. This promotes semantic coherence between modalities and enhances robustness under missing-modality conditions, as the network can leverage complementary information from the available modalities.

After the cross-attention layers, modality-specific class tokens are fused through concatenation. In our adaptation, the resulting fused representation from the ARMOUR cross-attention and contrastive alignment stage is then concatenated with the EHR embedding. The unified multimodal vector is finally passed to the classifier for prognostic prediction (see Fig. 3).

In the CROSS strategy, multimodal interactions are modeled through a hierarchical sequence of cross-attention operations rather than a joint or parallel fusion. Formally, the cross-attention operation is defined as:

$$\text{CrossAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{\mathbf{d}_k}}\right)\mathbf{V},$$

where $\mathbf{Q} = \mathbf{X}_Q\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}_K\mathbf{W}_K$, and $\mathbf{V} = \mathbf{X}_V\mathbf{W}_V$, with $\mathbf{X}_Q, \mathbf{X}_K, \mathbf{X}_V$ denoting the query, key, and value sequences from different modalities.

To capture richer interactions across multiple subspaces, we extend this operation to a multi-head cross-attention (MHCA) mechanism:

$$\text{MHCA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\mathbf{h}_1,\ldots,\mathbf{h}_H)\,\mathbf{W}^O,$$
$$\mathbf{h}_i = \text{CrossAttn}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V),$$

where $\mathbf{H}$ denotes the number of heads.

In this approach, the EHR embedding, represented as a single query token, acts as the central information carrier that sequentially integrates complementary information from other modalities. Specifically, in the first stage, the EHR token attends to one modality (e.g., imaging), acquiring modality-specific contextual features and generating an enriched representation. This updated token then serves as the query in a second cross-attention step, attending to another modality (e.g., textual reports), thereby incorporating additional semantic cues.

This cascaded design allows the model to progressively refine the EHR representation through modality-specific conditioning, enabling an interpretable flow of information across modalities and reducing the complexity of joint optimization. Unlike ARMOUR, which align modalities in a shared latent space simultaneously, CROSS enforces a sequential dependency that mimics a stepwise reasoning process, first grounding the representation in one modality before enriching it with another. The final token generated by the cascaded process thus encodes aggregated multimodal information and is subsequently passed to the classification layer for prognostic prediction, as illustrated in Fig. 4.

We also investigated alternative cascade configurations that vary in the ordering of modalities within the two-stage attention process. In the first configuration, the EHR representation served again as the query and sequentially attended to the report and then the image, producing a single enriched token that was directly fed to the classifier. In the second configuration, the cascade started with image tokens as the query attending to the report, followed by attention to the EHR representation. In this setup, the full token sequences were preserved rather than collapsed into a single token and the resulting sequence was averaged to obtain a mean embedding for classification.

These alternative designs allowed us to evaluate how different query–key/value hierarchies affect the propagation of cross-modal information and ultimately influence predictive performance.

Bimodal setting. For bimodal setting, experiments were carried out using only the CROSS scheme with a single Multi-head Cross-Attention (MCA) layer, where the query sequence was taken from one modality and the key–value pairs from the other. The evaluated modality pairs correspond to those considered in the early fusion setting (Table 4); however, note that the resulting embeddings in this CROSS scheme have different dimensionalities from those reported for early fusion. In the specific case of pairs with EHR data, which is represented by a single token, this token was directly used as the query. Consequently, the cross-attention operation yields an output sequence

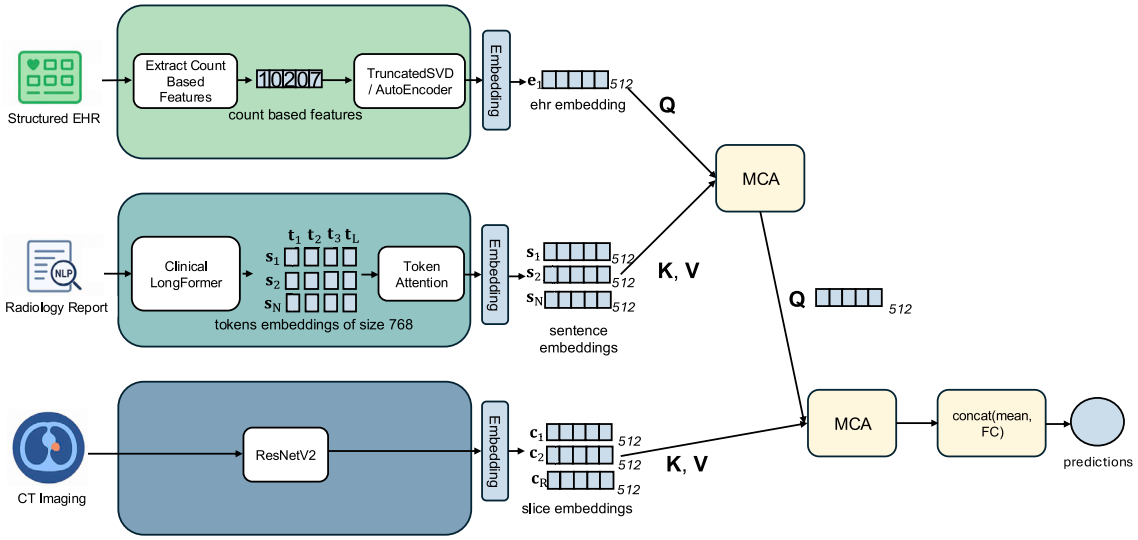


Fig. 4. Illustration of a case of trimodal *intermediate fusion* approach: A cascaded fusion mechanism is employed, where the EHR embedding sequentially attends to the other modalities through a Multi-Head Classification (MCA) layer prior to the final classification.

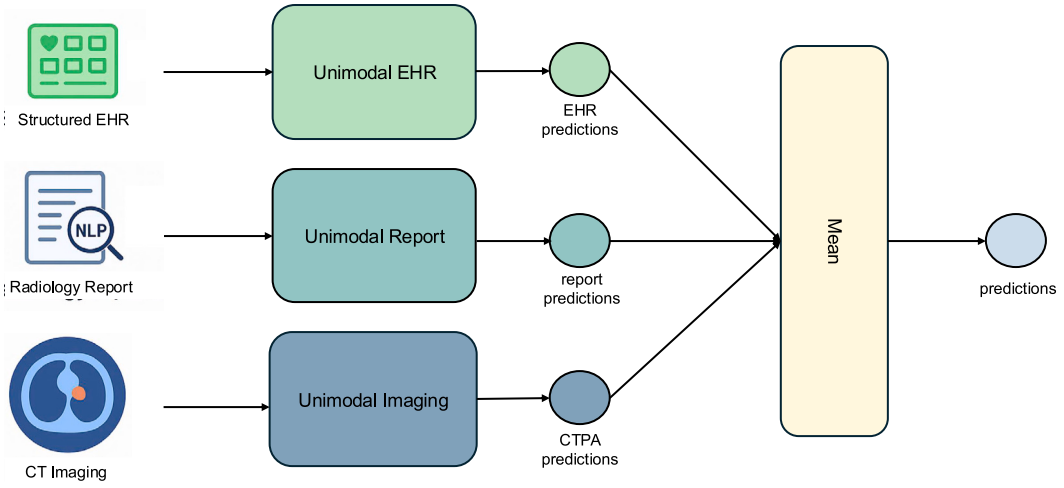


Fig. 5. Illustration of the *late fusion* method, where predictions from unimodal models are averaged to obtain the final decision.

reduced to a single token, which is subsequently passed to the classification layer. In all other cases, the output sequence produced by the cross-attention mechanism was averaged across tokens, and the resulting mean representation was fed into the classification layer.

4.3.3. Late fusion

Ensemble averaging of unimodal predictions was performed, as illustrated in Fig. 5, by combining the outputs of the three unimodal models (EHR, report, and imaging) through the averaging of their predicted probabilities. The final predicted class is then obtained by applying a threshold of 0.5 for a binary outcome. Both bimodal and trimodal combinations were considered. For the bimodal case, we used the same pairs as in the Early Fusion reported in Table 4; however, note that the resulting embeddings in this Late fusion scheme have different dimensionalities from those reported for early fusion.

4.4. Experimental settings

Unimodal and multimodal models were trained and evaluated using the patient-level hold-out splits defined in Huang et al. (2023), allocating 80% of the data to the training set, 5% to the validation set, and 15% to the test set. Our effective cohort comprises a subset of the original INSPECT dataset, totaling 23 248 CTPA studies. Details on the number of samples selected for each prognostic horizon are provided in Section 4.2. Samples lacking a valid reference time point (e.g., CT scan timestamp) were excluded to ensure that only pre-exam EHR events are used for prognostic modeling.

All experiments were implemented in `PyTorch` and executed on an NVIDIA A40 GPU with 48 GB of VRAM. Models were trained for up to 50 epochs using the AdamW optimizer with a weight decay of 0.02, a learning rate of 0.001, and a batch size of 128. The cross-entropy loss was employed as the training objective.

For the EHR-GBM model, rather than relying on the default configuration, we performed a targeted hyperparameter search over a compact grid including max depth [3, 6, –1], learning rate [0.02, 0.1, 0.5], and number of leaves [10, 25, 100].

Model performance was evaluated using the Matthews Correlation Coefficient (MCC), the Area Under the Receiver Operating Characteristic curve (AUROC), and the Area Under the Precision-Recall curve (AUPRC). The MCC is defined as

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives, respectively. MCC provides a balanced

Table 5
MCC of unimodal models across 1-, 6-, and 12-month mortality prediction tasks. Results are reported as mean ± standard deviation over five runs. Best MCC values for each time horizon are highlighted in **bold**.

Model	1-Month	6-Month	12-Month
EHR-GBM	0.258 ± 0.059	**0.367 ± 0.072**	**0.454 ± 0.023**
EHR-AE	**0.269 ± 0.013**	0.352 ± 0.016	0.395 ± 0.008
Report	0.242 ± 0.008	0.329 ± 0.030	0.354 ± 0.014
Image	0.221 ± 0.010	0.297 ± 0.016	0.304 ± 0.028

measure of prediction quality even in the presence of class imbalance, as it considers all four entries of the confusion matrix.

These metrics were chosen because the dataset is inherently imbalanced: mortality events are much less frequent than survival outcomes, especially at shorter prognostic horizons. On the one hand, accuracy can be misleading in such scenarios, as a model could achieve high accuracy by simply predicting the majority class. Similarly, F1-score does not account for true negatives, even though it captures the harmonic mean of precision and recall. On the other hand, MCC incorporates all entries of the confusion matrix, AUROC provides a threshold-independent measure of discriminative ability, and AUPRC emphasizes the model's performance on the minority positive class, which is critical for prognostic prediction.

All experiments were repeated five times with different random seeds. For each model and prediction horizon, we report the mean and standard deviation of the MCC to quantify performance variability and ensure a fair comparison across fusion strategies.

5. Results and discussion

This section presents the results of unimodal and multimodal experiments across three prognostic time horizons (1-, 6-, and 12-month mortality). Given the pronounced class imbalance in the INSPECT dataset (Table 1), with positive mortality rates of 5.45%, 11.54%, and 14.59% for 1-, 6-, and 12-month prediction tasks, respectively, the MCC was prioritized as the primary evaluation metric for model comparison and discussion, whereas additional metrics (e.g., AUROC, F1-score) are reported in the Appendix for completeness. MCC effectively balances true positives, true negatives, false positives, and false negatives, making it ideal for assessing prognostic models in imbalanced settings like PE.

We structure our findings around five key research questions, which collectively clarify the individual contribution of each modality, the behavior of different fusion strategies, and the broader implications for multimodal prognostic modeling in PE:

1. How well do unimodal models perform individually?
2. Does multimodal fusion improve predictive performance over the strongest unimodal baseline?
3. Which fusion strategy is most robust across the three prognostic horizons?
4. What are the potential pitfalls of complex fusion mechanisms such as CROSS and ARMOUR?
5. What potential clinical insights can multimodality provide?

We address these questions sequentially, beginning with unimodal performance and progressively incorporating more complex multimodal interactions, to uncover how each data source and fusion mechanism influences predictive accuracy and robustness.

(1) How well do unimodal models perform individually?

To establish baseline performance, we evaluated unimodal models trained separately on EHR, report, and imaging data across the three prognostic horizons (1-, 6-, and 12-month mortality). Table 5 summarizes the results in terms of MCC, AUC, and AUPRC.

For structured EHR data, we evaluated a Gradient Boosting Machine (GBM) and a Deep Autoencoder (AE). The EHR-AE achieved the highest mean MCC at the 1-month horizon (0.269), whereas EHR-GBM obtained the highest mean MCC at both the 6-month (0.367) and 12-month (0.454) horizons. In addition, EHR–GBM consistently showed strong AUC and AUPRC values across all tasks, highlighting the robustness of structured clinical variables.

The textual report model (Report in Table 5), based on Clinical-Longformer embeddings and hierarchical attention, achieved moderate performance, with mean MCC values of 0.242, 0.329, and 0.354 at 1-, 6-, and 12-month horizons, respectively. Meanwhile, the imaging-only model (Image in Table 5), relying on ResNetV2 features extracted from CT volumes, exhibited the lowest performance across all horizons, with mean MCC scores of 0.221, 0.297, and 0.304.

To assess whether the observed differences in MCC across the four unimodal baselines were statistically meaningful, we conducted paired one-tailed t-tests over the five independent runs for each prognostic horizon and applied Benjamini–Hochberg correction to control the false discovery rate (Benjamini and Hochberg, 1995). This analysis revealed three distinct patterns depending on the prediction window.

At the 1-month horizon, the EHR-AE clearly emerged as the strongest performer, significantly outperforming both the report-only model ($t = 3.96$, FDR-adjusted $p = 0.011$) and the imaging-only baseline ($t = 6.54$, FDR-adjusted $p = 0.0008$). Its performance, however, remained statistically indistinguishable from that of EHR–GBM ($p = 0.81$), suggesting that the two EHR-based approaches offer comparable predictive power in the short term.

At the 6-month horizon, the landscape became more balanced. No significant differences emerged between either of the EHR-based models and the report-only system. Nonetheless, the EHR–AE maintained a clear advantage over imaging ($t = 5.44$, FDR-adjusted $p = 0.004$), while the marginal superiority of EHR–GBM over imaging did not survive FDR correction ($p = 0.32$).

At the 12-month horizon EHR–GBM established itself as the dominant unimodal approach. It significantly outperformed the Autoencoder ($t = 5.42$, FDR-adjusted $p = 0.0013$), the report-only model ($t = 8.30$, FDR-adjusted $p = 0.0002$), and the imaging model ($t = 9.26$, FDR-adjusted $p = 0.0001$). Even so, the EHR-AE continued to retain a significant edge over both the report-only ($t = 5.69$, FDR-adjusted $p = 0.002$) and image-only ($t = 6.99$, FDR-adjusted $p = 0.0008$) baselines, confirming the consistent strength of structured EHR information across all horizons.

Overall, the findings delineate a clear and coherent trend: structured EHR data offer the strongest and most reliable unimodal signal for mortality prediction. The EHR-AE proves most effective in the short term, whereas the EHR–GBM progressively asserts itself as the leading baseline at longer horizons. Nevertheless, the steady, though comparatively weaker, predictive contribution of textual and imaging modalities underscores their value as complementary sources of information, supporting their integration within multimodal fusion frameworks.

(2) Does multimodal fusion improve predictive performance over the strongest unimodal baseline?

To evaluate the benefit of multimodal integration, we compared each fusion strategy against the strongest unimodal baselines, the EHR-AE at the 1-month horizon (mean MCC of 0.269) and the EHR–GBM at the 6- and 12-month horizons (mean MCCs of 0.367 and 0.454, respectively). Table B.10 reports MCC, AUC, and AUPRC for all unimodal models and the early, late, CROSS, and ARMOUR fusion schemes.

Late fusion strategies that incorporated the EHR–GBM predictions consistently achieved the highest MCC across all horizons, reaching 0.399 at 1-month, 0.479 at 6-months, and 0.497 at 12-months, compared to 0.269, 0.367, and 0.454 for the strongest unimodal baselines (Table 6). To facilitate interpretation, Fig. 6 shows a simplified visualization reporting, for each fusion category, the best-performing

Table 6
MCC of unimodal and multimodal models across 1-, 6-, and 12-month mortality prediction tasks (mean ± standard deviation over five independent runs). Best values in **bold**. In Cross Fusion models, "→" indicates the order of modality integration.

Model	1-Month	6-Month	12-Month
Unimodal			
Best unimodal	0.269 ± 0.013	0.367 ± 0.072	0.454 ± 0.023
Early fusion			
EHR-AE, report	0.296 ± 0.014	0.371 ± 0.018	0.407 ± 0.009
EHR-AE, image	0.302 ± 0.011	0.380 ± 0.016	0.405 ± 0.010
EHR-TSVD, report	0.280 ± 0.027	0.392 ± 0.006	0.419 ± 0.010
EHR-TSVD, image	0.274 ± 0.013	0.376 ± 0.011	0.390 ± 0.017
Report, image	0.244 ± 0.024	0.350 ± 0.016	0.366 ± 0.012
EHR-AE, report, image	0.302 ± 0.008	0.378 ± 0.014	0.398 ± 0.007
EHR-TSVD, report, image	0.293 ± 0.009	0.393 ± 0.008	0.426 ± 0.008
Late fusion			
Reports, EHR-AE	0.286 ± 0.015	0.388 ± 0.019	0.424 ± 0.005
Reports, EHR-GBM	**0.399 ± 0.050**	0.472 ± 0.011	0.494 ± 0.008
Image, EHR-AE	0.297 ± 0.017	0.395 ± 0.016	0.426 ± 0.004
Image, EHR-GBM	0.374 ± 0.025	0.467 ± 0.031	**0.497 ± 0.011**
Reports, image	0.275 ± 0.015	0.375 ± 0.011	0.398 ± 0.012
Reports, image, EHR-AE	0.331 ± 0.014	0.409 ± 0.014	0.444 ± 0.008
Reports, image, EHR-GBM	0.362 ± 0.016	**0.479 ± 0.011**	0.488 ± 0.002
Cross fusion			
EHR-TSVD → Image → Report	0.240 ± 0.023	0.321 ± 0.017	0.351 ± 0.028
EHR-AE → Image → Report	0.227 ± 0.011	0.327 ± 0.019	0.404 ± 0.022
EHR-TSVD → Report	0.247 ± 0.021	0.350 ± 0.028	0.377 ± 0.035
EHR-TSVD → Image	0.181 ± 0.024	0.294 ± 0.063	0.294 ± 0.119
EHR-AE → Report	0.285 ± 0.032	0.363 ± 0.019	0.391 ± 0.039
EHR-AE → Image	0.270 ± 0.036	0.370 ± 0.009	0.336 ± 0.168
Image → Report	0.236 ± 0.025	0.341 ± 0.016	0.361 ± 0.013
EHR-TSVD → Report → Image	0.084 ± 0.032	0.146 ± 0.052	0.087 ± 0.051
Image → Report → EHR-AE	0.305 ± 0.007	0.372 ± 0.023	0.397 ± 0.010
Image → Report → EHR-TSVD	0.250 ± 0.021	0.346 ± 0.017	0.388 ± 0.013
Armor fusion			
EHR-TSVD, image, report	0.239 ± 0.027	0.360 ± 0.030	0.391 ± 0.024
EHR-AE, image, report	0.284 ± 0.023	0.372 ± 0.015	0.420 ± 0.012

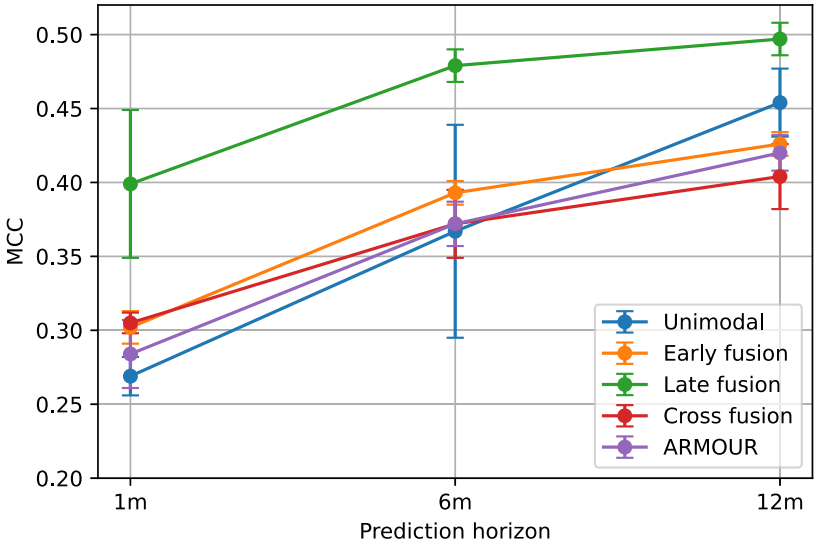


Fig. 6. Comparison of multimodal fusion strategies across prediction horizons. For each fusion type, the best-performing configuration is reported. Late fusion consistently achieves the highest MCC across 1-, 6-, and 12-month prediction.

configuration, providing a more intuitive comparison of multimodal strategies across the three prediction horizons.

Across the three prognostic horizons, different combinations of modalities emerged as the most effective. At the 1-month horizon, the best-performing configuration was the late fusion of clinical reports with EHR–GBM, reaching an MCC of 0.399 ± 0.050, which corresponds to a notable gain of +0.130 over the EHR-AE. Moving to the 6-month horizon, performance peaked with a fusion model that integrated reports, imaging, and EHR–GBM, achieving an MCC of 0.479 ± 0.011 and improving upon the EHR–GBM baseline by +0.112. At the 12-month horizon, the most effective strategy combined imaging with EHR–GBM, yielding an MCC of 0.497 ± 0.011, a +0.043 gain over the GBM alone.

Table 7
Statistical comparison of late fusion strategies incorporating EHR–GBM against the EHR-only baseline. Paired one-tailed t-tests were applied on MCC across five independent runs, with FDR correction using the Benjamini–Hochberg procedure.

Fusion strategy	1-month	6-month	12-month
Reports + EHR–GBM	$t = 4.08$, FDR-$p = 0.0020$	$t = 3.22$, FDR-$p = 0.0066$	$t = 3.67$, FDR-$p = 0.0034$
Image + EHR–GBM	$t = 4.05$, FDR-$p = 0.0021$	$t = 2.85$, FDR-$p = 0.0116$	$t = 3.77$, FDR-$p = 0.0029$
Reports + Image + EHR–GBM	$t = 3.80$, FDR-$p = 0.0029$	$t = 3.44$, FDR-$p = 0.0048$	$t = 3.29$, FDR-$p = 0.0059$

Table 8
Pairwise win comparisons across fusion strategies, aggregated over all time horizons and metrics (AUC, MCC, AUPRC). Each cell reports the proportion of wins by the row model vs. the column model, with the percentage of statistically significant wins (FDR-adjusted $p < 0.005$) in parentheses. **Abbreviations:** Uni. = Unimodal; EBi. = Early Bimodal; ETri. = Early Trimodal; LBi. = Late Bimodal; LTri. = Late Trimodal; CBi. = Cross Bimodal; CTri. = Cross Trimodal; ATri. = Armor Trimodal.

Fusion type	Uni.	EBi.	ETri.	LBi.	LTri.	CBi.	CTri.	ATri.
Uni.	–	0.3 (60%)	0.3 (100%)	0.2 (100%)	0.2 (73%)	0.6 (11%)	0.8 (66%)	0.6 (66%)
EBi.	0.7 (85%)	–	0.3 (63%)	0.1 (0%)	0.0 (0%)	0.7 (61%)	0.9 (77%)	0.8 (77%)
ETri.	0.7 (92%)	0.7 (44%)	–	0.2 (11%)	0.0 (0%)	0.8 (81%)	0.9 (88%)	0.9 (58%)
LBi.	0.8 (100%)	0.9 (88%)	0.8 (72%)	–	0.2 (78%)	1.0 (100%)	1.0 (100%)	1.0 (68%)
LTri.	0.8 (100%)	1.0 (100%)	1.0 (100%)	0.8 (79%)	–	1.0 (100%)	1.0 (100%)	1.0 (100%)
CBi.	0.4 (77%)	0.2 (49%)	0.2 (85%)	0.0 (0%)	0.0 (0%)	–	0.8 (78%)	0.5 (61%)
CTri.	0.2 (63%)	0.1 (83%)	0.1 (67%)	0.0 (0%)	0.0 (0%)	0.2 (27%)	–	0.3 (47%)
ATri.	0.4 (90%)	0.2 (71%)	0.1 (67%)	0.0 (50%)	0.0 (0%)	0.5 (70%)	0.7 (74%)	–

Notably, the three-modality late fusion model (Reports, Image, EHR–GBM) maintained strong performance across all horizons, with MCC values of 0.362, 0.479, and 0.488, reinforcing the value of integrating complementary information sources in multimodal frameworks.

Early fusion yielded more modest gains. The best early fusion variants (EHR–TSVD + Report + Image and EHR–TSVD + Report) reached MCC values up to 0.426 at 12-months but never significantly surpassed EHR–GBM on any horizon. Most CROSS-fusion and ARMOUR-fusion approaches performed poorly, with many MCC scores below or close to the unimodal report or image baselines, showing that sequential prediction and complex attention routing may introduce noise and error propagation in this setting.

To assess whether multimodal fusion genuinely contributed additional predictive value, we focused on all late fusion strategies that incorporated EHR–GBM predictions and evaluated them against the strongest EHR-only baseline. For each horizon, we applied paired one-tailed t-tests on MCC across the five independent runs, followed by FDR correction using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). This analysis, summarized in Table 7, revealed a marked contrast with respect to other integration approaches: late fusion configurations involving EHR–GBM displayed a consistent and statistically robust advantage, signaling a clear shift in performance.

In conclusion, simple decision-level late fusion of independent modality-specific predictions with the strong EHR–GBM model predictions consistently and significantly improves mortality prediction across all three prognostic horizons. The combination of structured EHR data with either narrative radiology reports, imaging features, or both yields the highest absolute performance, with gains of up to +14.1% MCC at 1-month and +11.2% MCC at 6-months. These results point out that EHR encode with radiology reports and CT imaging complementary prognostic information that is most effectively exploited through late fusion rather than joint representation learning or complex sequential architectures.

Detailed analysis of modality interactions and redundancy, which explains the slight advantage of bimodal models over trimodal configurations, is provided in Appendix C.

(3) Which fusion strategy is most robust across the three prognostic horizons?

To evaluate the robustness of each fusion strategy across the 1-, 6-, and 12-month mortality prediction tasks, we considered mean MCC performance, ranking consistency, and statistical superiority over competing strategies.

To provide a rigorous statistical synthesis of fusion strategy performance, Table 8 reports pairwise win rates across all models, metrics, and horizons. Late trimodal fusion (LTri.) dominates, achieving near-perfect win rates (0.8–1.0) with 100% significance against most categories and 86% significance vs. unimodal models. Late bimodal (LBi.) also excels, particularly against early and CROSS fusion.

Early fusion shows moderate strength against unimodal models (0.7 wins, 85%–92% significance) but is consistently outperformed by late fusion. ARMOUR trimodal wins against CROSS and unimodal models but never beats late fusion (0.0 wins). Cross fusion (CTri.) performs the weakest among the evaluated strategies, exhibiting win rates at or below 0.3 and generally lacking statistical significance in most comparisons.

These results statistically confirm that late fusion, especially trimodal, is the most robust and generalizable strategy, even when accounting for metric-specific fluctuations (e.g., occasional bimodal MCC wins at 1- and 12-month). Its superiority can be attributed to its modular design: each modality is modeled independently at its optimal capacity, particularly the strong EHR–GBM component, and only calibrated predictions are combined at the decision level. This minimizes interference between heterogeneous data types, reduces overfitting, and preserves the predictive power of structured EHR data while extracting complementary information from reports and imaging.

Across the three prediction horizons, the optimal late fusion configurations are: Radiology Reports with EHR–GBM for 1-month mortality, Radiology Reports, CT Imaging, and EHR–GBM for 6-month mortality, and CT Imaging with EHR–GBM, or alternatively all three modalities, for 12-month mortality, yielding very similar performance. These simple, interpretable ensembles consistently achieve the highest predictive performance with statistically significant improvements over other fusion strategies.

(4) What are the pitfalls of complex fusion mechanisms (CROSS and ARMOUR)?

Although advanced fusion strategies such as CROSS Fusion and ARMOUR Fusion aim to capture intricate inter-modal interactions through attention-based or bidirectional mechanisms, our results reveal significant pitfalls that undermine their effectiveness and robustness in clinical mortality prediction.

Cross Fusion exhibits severe performance degradation and instability, particularly in multi-step cascading configurations. Despite using the same modalities, performance varies dramatically with fusion order. For instance, the sequence EHR TSVD → Report → Image yields

MCC values of only 0.084 ± 0.032 (1-month), 0.146 ± 0.052 (6-month), and 0.087 ± 0.051 (12-month), drastically worse than unimodal baselines. This sensitivity to ordering indicates that the attention mechanism may amplify noise or introduce destructive interference when weaker modalities (e.g., Image) are used as queries or keys in early stages. Furthermore, high variance across runs (e.g., ±0.184 in EHR Autoencoder→ Report → Image at 12-months) suggests poor generalization and overfitting to spurious correlations in training splits.

Similarly, ARMOUR Fusion, designed for bidirectional modality refinement, fails to outperform simpler strategies. Even in its full three-modality form (EHR, Image, Report), it achieves only moderate MCCs (0.239–0.284 at 1-month, up to 0.420 at 12-months) and never surpasses Late Fusion. The incremental refinement steps appear to dilute discriminative signals from the dominant EHR-GBM predictor, adding computational overhead without proportional gain.

A key shared pitfall is the risk of over-parameterization in high-dimensional heterogeneous clinical data. Both CROSS and ARMOUR introduce numerous attention heads and fusion layers, which, although expressive, increase the risk of overfitting, especially given the limited size of medical cohorts and class imbalance in mortality tasks. In contrast, Late Fusion preserves modality-specific optimization and combines predictions at the decision level, avoiding interference.

Moreover, the behavior of the more complex fusion mechanisms reveals a lack of robustness across prediction horizons. Whereas Late Fusion improves steadily from 1 to 12 months, both CROSS and ARMOUR exhibit irregular and often unstable performance, with several configurations deteriorating sharply at longer horizons despite the availability of richer longitudinal information. This pattern suggests that these intricate fusion schemes may struggle to scale with increasing predictive difficulty.

In practical terms, these shortcomings make CROSS and ARMOUR unreliable candidates for real-world deployment. Late Fusion, by contrast, stands out as the most effective strategy: it is simple, stable across horizons, and aligns naturally with clinical workflows, confirming its role as the preferred approach. A detailed analysis supporting these findings, including noise-injection experiments, t-SNE visualizations, and Representational Similarity Analysis (RSA), is provided in Appendix D.

(5) What potential clinical insights can multimodality provide?

Multimodal models integrating EHR, radiology reports, and CT images could offer rich clinical insights beyond unimodal predictions, revealing complementary risk signals and temporal patterns in mortality prediction. These insights may support personalized clinical decision-making.

First, EHR dominates unimodal performance across all horizons (MCC: 0.269 at 1-month, 0.367 at 6-month, 0.454 at 12-months), confirming that structured clinical variables, such as vital signs, lab results, and comorbidities, carry the strongest and most stable prognostic signal. This underscores the foundational role of electronic health records in risk stratification.

However, multimodal late fusion consistently outperforms the EHR-only baseline, increasing MCC from 0.269 to 0.399 at 1-month, from 0.367 to 0.479 at 6-months, and from 0.454 to 0.497 at 12-months, indicating additive predictive value from unstructured modalities.

At the 1-month horizon, combining Reports with EHR-GBM yields a sizeable gain of +0.130 MCC (from 0.269 to 0.399). This improvement indicates that radiology reports may encode acute, clinically actionable findings, such as signs of respiratory distress or emergent complications, that are only partially reflected in structured fields.

Over longer horizons, the contribution of imaging becomes more prominent. At 6 and 12 months, the addition of CT features within the Reports + Image + EHR-GBM ensemble produces further gains (+0.112 to +0.034 MCC). This suggests that CT scans may capture slowly evolving anatomical patterns, chronic vascular changes, cardiopulmonary remodeling, and other structural markers, that carry long-term prognostic value.

Taken together, these results reveal a clear temporal structure in multimodal prediction: short-term mortality is best informed by acute descriptors encoded in narrative reports, whilst long-term risk increasingly depends on structural information visible in imaging. The complementary nature of these modalities underscores the value of multimodal fusion for a holistic, time-sensitive assessment of patient trajectory.

Furthermore, the failure of complex fusion (CROSS/ARMOUR) despite access to the same data highlights that not all inter-modal interactions are beneficial. Destructive fusion in cascading attention (e.g., EHR → Image → Report) may propagate imaging noise into textual reasoning, reducing reliability. In this context, clinical multimodal systems may benefit from modular, decision-level integration (Late Fusion) to preserve signal fidelity, though this may not generalize to all tasks or datasets.

Finally, in the best-performing models, AUPRC increases from 0.394 at 1 month to 0.581 at 12 months (see Appendix B for full results), suggesting that class separability becomes more pronounced for longer prediction horizons. Multimodal architectures seem to exploit this widening gap particularly well: they leverage early warning cues embedded in radiology reports while integrating anatomical information that becomes apparent in the CT scans at the time of the visit.

This evolution across time also suggests how such systems could be deployed in clinical practice. In the immediate, 1-month window, radiology reports combined with EHR data may offer rapid and actionable insight, supporting early risk stratification in acute scenarios. As the horizon extends to 6- and 12-months, imaging features may gain increasing predictive relevance, making them valuable for monitoring disease progression and gauging treatment response. From an implementation standpoint, late-fusion ensembles are especially practical, as they combine modality-specific predictions while remaining robust to variability in input data availability.

In essence, multimodal models reveal how different data sources contribute complementary prognostic value at different stages of the patient trajectory, turning heterogeneous clinical information into coherent, horizon-specific decision support.

Comparison with the original INSPECT study

A direct, point-by-point comparison with Huang et al. (2023) is not possible, as several methodological and practical differences separate our work from the original study. Three elements, in particular, shape this divergence.

First, our analysis is conducted on a reduced version of the INSPECT cohort. Patients lacking a reliable reference timestamp (CT acquisition time) had to be removed, since structured EHR features and prognostic labels could not be aligned consistently. This necessary filtering both reduced the available data and slightly altered class balance.

Second, although the MOTOR model (Steinberg et al., 2023), a time-to-event foundation model designed by the creators of the INSPECT dataset to extract temporal representations from structured EHR data, is officially released, its use remains out of reach in practice. The absence of operational code, documentation, and working preprocessing pipelines prevents us from reproducing its reported performance (AUC ≈ 0.924) or from incorporating it as a comparative baseline.

Third, differences in evaluation metrics complicate direct comparisons. Huang et al. (2023) reported only AUC, which can be sensitive to class imbalance and potentially misleading in rare-event settings (Saito and Rehmsmeier, 2015). In our experiments, for example, the EHR–GBM model, the same EHR-based Gradient Boosting Machine baseline used in the INSPECT study, achieves a higher AUC at 1 month (0.926) than our best late-fusion model (0.884), yet its MCC is substantially lower (0.258 vs. 0.399). This illustrates how AUC and MCC can provide complementary perspectives, and why MCC is particularly informative in the presence of class imbalance.

Despite these constraints, we were able to reproduce the main INSPECT baselines under identical preprocessing conditions. For the unimodal CT model, our AUCs reached 0.800, 0.790, and 0.781 at the

1-, 6-, and 12-month horizons, closely matching the original results (0.794, 0.755, 0.748). The unimodal EHR–GBM achieved AUCs of 0.926, 0.893, and 0.893, in line with the structured-data performance reported in INSPECT (0.848, 0.875, 0.866). Similarly, the bimodal late-fusion model combining EHR–GBM with CT imaging yielded AUCs of 0.863, 0.862, and 0.871, which remain highly comparable to the corresponding INSPECT results (0.867, 0.875, 0.866) despite the smaller cohort used in our study.

These replications indicate that our pipeline faithfully recovers the original behavior of the unimodal and bimodal models and therefore provides a solid basis for evaluating more advanced multimodal extensions.

Importantly, our Late Fusion models that incorporate radiology reports, a modality absent from the original INSPECT framework, achieve the highest MCC scores across horizons (0.399 for Reports + EHR–GBM at 1-month and 0.479 for Reports + Image + EHR–GBM at 6-month). Both configurations significantly outperform the reproduced baselines as well as the original EHR+CT benchmark ($p < 0.05$). The improvements are especially evident at the 1- and 6-month horizons because descriptive elements in the reports likely capture clinically relevant signals—such as thrombus burden or right-ventricular strain—that are only partially reflected in structured vitals or imaging-derived features.

In summary, although strict numerical comparability with Huang et al. (2023) is precluded by differences in data availability and model accessibility, our carefully reproduced baselines and the consistent gains achieved through report-augmented Late Fusion models demonstrate that incorporating narrative radiology information meaningfully extends the prognostic capacity of the INSPECT framework.

6. Conclusions

This study presents a comprehensive evaluation of unimodal and multimodal deep learning models for predicting 1-, 6-, and 12-month mortality in patients with PE, using the INSPECT dataset comprising chest CT images, radiology reports, and structured EHR data. Our findings demonstrate that multimodal integration via late fusion outperforms unimodal baselines across all prediction horizons, improving MCC from 0.269 to 0.399 at 1-month, from 0.367 to 0.479 at 6-months, and from 0.454 to 0.497 at 12-months (Table 6), establishing late fusion as the most robust fusion paradigm in this setting.

The highest MCC values were consistently achieved by late fusion models anchored on EHR-GBM, the strongest unimodal predictor, augmented with either radiology reports or imaging. Notably, bimodal fusion achieved the highest MCC at the 1-month horizon with (Reports, EHR GBM; 0.399) and at the 12-month horizon with (Image, EHR GBM; 0.497), whereas trimodal fusion (Reports, Image, EHR GBM) peaked at 6 months (0.479) and remained comparable to bimodal fusion at 12 months. This horizon-dependent synergy suggests a key insight: short-term mortality risk is primarily captured by textual descriptors in radiology reports, while longer-term prognosis increasingly benefits from structural imaging information, reflecting disease evolution not fully represented in structured EHR data.

In contrast, early fusion and ARMOUR provide only marginal gains, while CROSS fusion consistently underperforms and in some cases falls below unimodal baselines. Our additional analyses (noise-injection ablations and RSA) indicate that this behavior is driven by modality collapse, where predictions become dominated by a single modality and other sources are largely ignored. RSA further shows that fused representations tend to align with the strongest modality rather than forming a shared multimodal space.

Overall, these results suggest that increased fusion complexity does not guarantee better integration, but can instead amplify redundancy and modality dominance. In contrast, late fusion remains more robust by preserving modality-specific representations and combining them only at the decision level.

Clinically, these models integrate complementary prognostic signals from multiple sources. EHR features encoded by the GBM capture chronic comorbidity burden and overall physiological stability, whereas radiology reports provide acute, narrative-driven insights, highlighting findings such as right heart strain or clot burden. Chest CT images, in turn, reveal structural progression, including ventricular dilation or pulmonary infarction. By aligning these diverse modalities through late fusion, our approach enables horizon-aware risk stratification, supporting both short-term triage and longer-term prognosis assessment.

Future directions include improving interpretability by applying Grad-CAM to imaging, SHAP/LIME to EHR features, and attention visualization to report embeddings to better understand modality-specific contributions. Another important avenue is clinical validation through prospective studies designed to evaluate real-world impact on therapeutic decision-making and patient outcomes. Finally, adaptive fusion strategies, such as gating mechanisms that selectively activate imaging information when it is most informative, represent a promising direction for making multimodal models more efficient and clinically meaningful.

In conclusion, this work provides quantitative evidence that late fusion achieves the highest predictive performance, with MCC gains of up to +0.130 at 1-month and +0.112 at 6-months compared to unimodal EHR baselines. By prioritizing modular design over architectural complexity, we achieve superior performance, generalizability, and translational potential, paving the way for trustworthy AI augmentation of clinical prognosis in complex cardiopulmonary diseases.

CRediT authorship contribution statement

Domenico Paolo: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Paolo Soda:** Writing – review & editing, Visualization, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Matteo Tortora:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Alessandro Bria:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Rosa Sicilia:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT in order to assist with language editing and improving clarity. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Domenico Paolo is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVIII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma. This work was partially supported by PNRR MUR projects PE0000013-FAIR. This work was partially supported by PRIN 2022 MUR 20228MZFAA- AIDA (CUP C53D23003620008, CUP H53D23003480006). Resources are provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Alvis @ C3SE, partially funded by the Swedish Research Council through grant agreement no. 2022-06725 and no. 2018–05973.

Table A.9
Performance of unimodal models across 1-, 6-, and 12-month mortality prediction tasks. Results are reported as mean ± standard deviation over five independent runs. Best MCC values for each time horizon are highlighted in **bold**.

Model	1-Month			6-Month			12-Month		
	MCC	AUC	AUPRC	MCC	AUC	AUPRC	MCC	AUC	AUPRC
EHR-GBM	0.258 ± 0.059	0.926 ± 0.004	0.429 ± 0.019	**0.367 ± 0.072**	0.893 ± 0.008	0.503 ± 0.021	**0.454 ± 0.023**	0.893 ± 0.002	0.561 ± 0.007
EHR-AE	**0.269 ± 0.013**	0.790 ± 0.028	0.200 ± 0.014	0.352 ± 0.016	0.817 ± 0.014	0.354 ± 0.013	0.395 ± 0.008	0.812 ± 0.013	0.433 ± 0.012
Report	0.242 ± 0.008	0.837 ± 0.005	0.210 ± 0.007	0.329 ± 0.030	0.825 ± 0.003	0.402 ± 0.008	0.354 ± 0.014	0.821 ± 0.007	0.481 ± 0.012
Image	0.221 ± 0.010	0.800 ± 0.003	0.173 ± 0.007	0.297 ± 0.016	0.790 ± 0.008	0.323 ± 0.010	0.304 ± 0.028	0.781 ± 0.013	0.363 ± 0.013

Table B.10
Performance comparison of unimodal and multimodal models across 1-, 6-, and 12-month mortality prediction tasks. The table reports MCC, AUC, and AUPRC for each modality and fusion scheme. Results are reported as mean ± standard deviation over five independent runs. Best MCC values for each time horizon are highlighted in **bold**.

Model	1-Month			6-Month			12-Month		
	MCC	AUC	AUPRC	MCC	AUC	AUPRC	MCC	AUC	AUPRC
Unimodal									
EHR-GBM	0.258 ± 0.059	0.926 ± 0.004	0.429 ± 0.019	0.367 ± 0.072	0.893 ± 0.008	0.503 ± 0.021	0.454 ± 0.023	0.893 ± 0.002	0.561 ± 0.007
EHR-AE	0.269 ± 0.013	0.790 ± 0.028	0.200 ± 0.014	0.352 ± 0.016	0.817 ± 0.014	0.354 ± 0.013	0.395 ± 0.008	0.812 ± 0.013	0.433 ± 0.012
Report	0.242 ± 0.008	0.837 ± 0.005	0.210 ± 0.007	0.329 ± 0.030	0.825 ± 0.003	0.402 ± 0.008	0.354 ± 0.014	0.821 ± 0.007	0.481 ± 0.012
Image	0.221 ± 0.010	0.800 ± 0.003	0.173 ± 0.007	0.297 ± 0.016	0.790 ± 0.008	0.323 ± 0.010	0.304 ± 0.028	0.781 ± 0.013	0.363 ± 0.013
Early fusion									
EHR-AE, report	0.296 ± 0.014	0.845 ± 0.005	0.228 ± 0.011	0.371 ± 0.018	0.808 ± 0.002	0.365 ± 0.003	0.407 ± 0.009	0.795 ± 0.003	0.424 ± 0.007
EHR-AE, image	0.302 ± 0.011	0.842 ± 0.005	0.234 ± 0.005	0.380 ± 0.016	0.810 ± 0.004	0.369 ± 0.006	0.405 ± 0.010	0.794 ± 0.003	0.423 ± 0.003
EHR-TSVD, report	0.280 ± 0.027	0.846 ± 0.005	0.245 ± 0.013	0.392 ± 0.006	0.856 ± 0.009	0.434 ± 0.014	0.419 ± 0.010	0.858 ± 0.005	0.490 ± 0.017
EHR-TSVD, image	0.274 ± 0.013	0.842 ± 0.012	0.212 ± 0.024	0.376 ± 0.011	0.849 ± 0.011	0.402 ± 0.018	0.390 ± 0.017	0.847 ± 0.004	0.477 ± 0.018
Report, image	0.244 ± 0.024	0.841 ± 0.001	0.231 ± 0.009	0.350 ± 0.016	0.829 ± 0.001	0.397 ± 0.003	0.366 ± 0.012	0.828 ± 0.004	0.472 ± 0.007
EHR-AE, report, image	0.302 ± 0.008	0.845 ± 0.006	0.237 ± 0.014	0.378 ± 0.014	0.810 ± 0.002	0.369 ± 0.003	0.398 ± 0.007	0.790 ± 0.006	0.426 ± 0.008
EHR-TSVD, report, image	0.293 ± 0.009	0.861 ± 0.006	0.239 ± 0.006	0.393 ± 0.008	0.857 ± 0.005	0.419 ± 0.007	0.426 ± 0.008	0.853 ± 0.008	0.478 ± 0.021
Late fusion									
Reports, EHR-AE	0.286 ± 0.015	0.865 ± 0.005	0.250 ± 0.011	0.388 ± 0.019	0.861 ± 0.005	0.427 ± 0.011	0.424 ± 0.005	0.860 ± 0.005	0.519 ± 0.014
Reports, EHR-GBM	**0.399 ± 0.050**	0.884 ± 0.008	0.394 ± 0.013	0.472 ± 0.011	0.872 ± 0.008	0.473 ± 0.011	0.494 ± 0.008	0.875 ± 0.004	0.551 ± 0.007
Image, EHR-AE	0.297 ± 0.017	0.846 ± 0.011	0.268 ± 0.013	0.395 ± 0.016	0.855 ± 0.005	0.442 ± 0.009	0.426 ± 0.004	0.852 ± 0.006	0.506 ± 0.009
Image, EHR-GBM	0.374 ± 0.025	0.863 ± 0.012	0.372 ± 0.019	0.467 ± 0.031	0.862 ± 0.011	0.500 ± 0.022	**0.497 ± 0.011**	0.871 ± 0.002	0.574 ± 0.011
Reports, image	0.275 ± 0.015	0.855 ± 0.004	0.232 ± 0.005	0.375 ± 0.011	0.848 ± 0.003	0.431 ± 0.008	0.398 ± 0.012	0.841 ± 0.003	0.502 ± 0.007
Reports, image, EHR-AE	0.331 ± 0.014	0.875 ± 0.005	0.285 ± 0.012	0.409 ± 0.014	0.874 ± 0.002	0.477 ± 0.010	0.444 ± 0.008	0.871 ± 0.004	0.553 ± 0.010
Reports, image, EHR-GBM	0.362 ± 0.016	0.884 ± 0.007	0.384 ± 0.016	**0.479 ± 0.011**	0.878 ± 0.006	0.508 ± 0.015	0.488 ± 0.002	0.879 ± 0.003	0.581 ± 0.009
Cross fusion									
EHR-TSVD → Image → Report	0.240 ± 0.023	0.813 ± 0.004	0.189 ± 0.009	0.321 ± 0.017	0.811 ± 0.004	0.365 ± 0.012	0.351 ± 0.028	0.807 ± 0.004	0.447 ± 0.007
EHR-AE → Image → Report	0.227 ± 0.011	0.808 ± 0.010	0.190 ± 0.019	0.327 ± 0.019	0.807 ± 0.014	0.351 ± 0.014	0.404 ± 0.022	0.826 ± 0.015	0.468 ± 0.010
EHR-TSVD → Report	0.247 ± 0.021	0.842 ± 0.005	0.245 ± 0.027	0.350 ± 0.028	0.828 ± 0.010	0.389 ± 0.009	0.377 ± 0.035	0.831 ± 0.007	0.470 ± 0.015
EHR-TSVD → Image	0.181 ± 0.024	0.772 ± 0.031	0.151 ± 0.021	0.294 ± 0.063	0.798 ± 0.021	0.308 ± 0.048	0.294 ± 0.119	0.801 ± 0.015	0.338 ± 0.051
EHR-AE → Report	0.285 ± 0.032	0.845 ± 0.011	0.234 ± 0.018	0.363 ± 0.019	0.811 ± 0.005	0.354 ± 0.015	0.391 ± 0.039	0.801 ± 0.015	0.434 ± 0.024
EHR-AE → Image	0.270 ± 0.036	0.839 ± 0.009	0.209 ± 0.003	0.371 ± 0.009	0.822 ± 0.004	0.349 ± 0.015	0.336 ± 0.168	0.821 ± 0.005	0.402 ± 0.014
Image → Report	0.236 ± 0.025	0.823 ± 0.008	0.189 ± 0.018	0.341 ± 0.016	0.820 ± 0.006	0.369 ± 0.014	0.361 ± 0.013	0.818 ± 0.002	0.459 ± 0.006
EHR-TSVD → Report → Image	0.084 ± 0.032	0.651 ± 0.016	0.093 ± 0.007	0.146 ± 0.052	0.623 ± 0.012	0.166 ± 0.011	0.087 ± 0.051	0.630 ± 0.018	0.210 ± 0.018
EHR-AE → Report → Image	0.124 ± 0.083	0.685 ± 0.065	0.111 ± 0.031	0.204 ± 0.158	0.704 ± 0.092	0.229 ± 0.071	0.165 ± 0.184	0.669 ± 0.089	0.251 ± 0.074
Image → Report → EHR-AE	0.305 ± 0.007	0.835 ± 0.005	0.214 ± 0.011	0.372 ± 0.023	0.810 ± 0.007	0.352 ± 0.024	0.397 ± 0.010	0.794 ± 0.008	0.415 ± 0.010
Image → Report → EHR-TSVD	0.250 ± 0.021	0.813 ± 0.017	0.167 ± 0.031	0.346 ± 0.017	0.822 ± 0.005	0.333 ± 0.032	0.388 ± 0.013	0.831 ± 0.004	0.426 ± 0.009
Armor fusion									
EHR-TSVD, image, report	0.239 ± 0.027	0.801 ± 0.018	0.182 ± 0.020	0.360 ± 0.030	0.815 ± 0.013	0.341 ± 0.022	0.391 ± 0.024	0.820 ± 0.009	0.406 ± 0.023
EHR-AE, image, report	0.284 ± 0.023	0.821 ± 0.006	0.191 ± 0.007	0.372 ± 0.015	0.783 ± 0.008	0.341 ± 0.010	0.420 ± 0.012	0.782 ± 0.007	0.415 ± 0.002

Appendix A. Complete unimodal results

See Table A.9.

Appendix B. Complete multimodal results

Although advanced fusion strategies such as CROSS Fusion and ARMOUR Fusion aim to model fine-grained inter-modal dependencies through attention-based and bidirectional interaction mechanisms, our results reveal several critical limitations that undermine their effectiveness and robustness in clinical mortality prediction.

CROSS Fusion exhibits pronounced performance instability and strong sensitivity to architectural design choices. Despite operating on identical input modalities, its performance varies substantially depending on the fusion ordering. For example, the sequence EHR (TSVD) → Report → Image leads to severely degraded MCC values of 0.084 ± 0.032 (1-month), 0.146 ± 0.052 (6-month), and 0.087 ± 0.051 (12-month), which are markedly worse than several unimodal baselines.

ARMOUR Fusion, despite incorporating bidirectional cross-attention and a contrastive alignment objective, also fails to consistently outperform simpler strategies such as Late Fusion, achieving only marginal or unstable gains across horizons.

Table C.11
Pearson correlation between unimodal model predictions across 5 independent runs (mean ± std).

Task	Image - Reports	Image - EHR	Reports - EHR
1-Month mortality	**0.497 ± 0.014**	0.346 ± 0.059	0.400 ± 0.073
6-Month mortality	0.499 ± 0.015	0.432 ± 0.072	**0.545 ± 0.090**
12-Month mortality	0.501 ± 0.014	0.457 ± 0.039	**0.590 ± 0.035**

Appendix C. In-depth analysis of late fusion and modality interactions

To investigate the observed slight advantage of Late bimodal models over trimodal alternatives, we conducted an analysis focusing on information redundancy. Since Late Fusion combines independently trained unimodal predictors by averaging their output probabilities, we measured the Pearson correlation between their predictions across modalities.

The relatively small gap between bimodal and trimodal Late Fusion models can be explained by redundancy across modalities. As shown in Table C.11, unimodal predictions are moderately to highly correlated, particularly between Reports and EHR (up to 0.59 at 12 months), indicating overlapping predictive signals.

As a result, once two strongly correlated modalities are combined, the addition of a third modality provides limited incremental benefit, leading to marginal and inconsistent improvements in trimodal settings.

Appendix D. Analysis of CROSS and ARMOUR fusion strategies

To gain deeper insight into the underperformance of the intermediate fusion strategies CROSS and ARMOUR, we conducted three complementary analyses on the best-performing configuration of each model, with checkpoints selected based on the highest validation MCC. For ARMOUR we used the configuration employing the EHR-AE embedding (the best-performing across all three tasks), whilst for CROSS we adopted the ordering Image → Report → EHR-AE (denoted CROSS-A), which achieved the best results in two out of three tasks. Since the CROSS architecture is sensitive to modality ordering, we additionally evaluated the alternative ordering in which EHR and Image representations are processed first, followed by Report (denoted CROSS-B). The analyses comprise (i) targeted noise-injection experiments, (ii) t-SNE visualizations of the modalities and final fused representations, and (iii) Representational Similarity Analysis (RSA) using Representational Dissimilarity Matrices (RDMs).

For the noise-injection experiments, we corrupted one modality at a time by completely replacing its input embeddings with zero-mean Gaussian noise (`torch.randn()`, scaled to the original embedding norm with $\sigma = 0.1$), while keeping the remaining modalities unchanged. This setup ensures that the signal from the selected modality is entirely substituted by stochastic noise, enabling a controlled assessment of its contribution.

The RSA was computed by comparing the internal relational structure of modality-specific representations rather than directly comparing embedding vectors. Hence, for each modality m, we computed a Representational Dissimilarity Matrix (RDM) using the pairwise correlation distance between all samples:

$$\mathrm{RDM}_{ij}^{(m)} = 1 - \rho\left(f_i^{(m)}, f_j^{(m)}\right),$$

where $\rho(\cdot,\cdot)$ denotes the Pearson correlation coefficient. The similarity between two modalities m_1 and m_2 is then defined as the Pearson correlation between their corresponding RDMs:

$$\mathrm{RSA}(m_1, m_2) = \mathrm{corr}\left(\mathrm{vec}(\mathrm{RDM}^{(m_1)}), \mathrm{vec}(\mathrm{RDM}^{(m_2)})\right).$$

This measure captures the extent to which two modalities encode consistent relational structures among samples, rather than direct alignment in the embedding space. High RSA values indicate that two modalities organize the dataset in a similar way, whereas low values suggest divergent or independent representational geometries.

Results for CROSS-A and CROSS-B. The noise-injection results (Table D.12) show that both CROSS variants are highly sensitive to perturbations in the EHR modality. When EHR is introduced first (CROSS-B), injecting noise into the Image or Report modalities results in only minor performance variations (within ±0.04 in MCC), whereas corruption of the EHR modality leads to a pronounced degradation, with an average MCC drop of 0.23 points across horizons (−0.284 at 1-month, −0.292 at 6-month, and −0.103 at 12-month). Notably, at the 12-month horizon, CROSS-B also shows increased sensitivity to Report perturbations, with an MCC decrease of 0.276 points (0.347 → 0.071), indicating a substantially larger degradation compared to shorter time horizons. When the ordering is reversed (CROSS-A), the model becomes effectively insensitive to noise in the Image or Report modalities, while exhibiting a severe degradation under EHR corruption, with an average MCC drop of 0.38 points (−0.318 at 1-month, −0.391 at 6-month, and −0.429 at 12-month).

RSA analysis (Fig. D.7) confirms that the first fused representation aligns almost perfectly with the RDM of the dominant modality (RSA ≈ 0.93 with Report in CROSS-A; RSA = 0.92–0.98 for CROSS-B with EHR). The second fused representation remains strongly aligned with EHR in both configurations. Consequently, the final fused embedding is effectively dominated by the EHR geometry, regardless of cascade order.

Table D.12
MCC for noise-injection ablation across models. For ARMOUR, CLS_IMG and CLS_TXT denote the learned class tokens associated with the image and text modalities, respectively, obtained after bidirectional cross-attention.

Model	Noise on	1 m	6 m	12 m
CROSS-A	–	0.305	0.393	0.395
	Report	0.305	0.393	0.395
	Image	0.305	0.393	0.395
	EHR	−0.013	0.002	−0.034
CROSS-B	–	0.335	0.363	0.347
	Report	0.301	0.327	0.071
	Image	0.294	0.361	0.339
	EHR	0.051	0.071	0.244
ARMOUR	–	0.293	0.357	0.421
	Report	0.294	0.362	0.424
	Image	0.301	0.367	0.407
	EHR	0.073	0.024	−0.011
	CLS_IMG	0.296	0.371	0.329
	CLS_TXT	0.301	0.361	0.410
	(CLS_IMG, CLS_TXT)	0.296	0.375	0.407

The t-SNE projections (Figs. D.8 and D.9) illustrate that the final fused representations exhibit severe overlap between mortality classes, losing the clear separation that is visible in the unimodal EHR embeddings (especially at 6- and 12-month horizons).

Overall, rather than achieving true multimodal fusion, both CROSS variants collapse onto the strongest unimodal representation (EHR) while largely ignoring complementary information from the Image and Report modalities. This collapse not only prevents effective cross-modal integration but also degrades the original class separability provided by the EHR embedding. Hence, even though EHR alone already provides reasonable class separation, the fusion process dilutes some of its discriminative features through the need to accommodate the noisy information coming from Image and Report. The optimization therefore converges to a compromised representation that is closer to the dominant unimodal geometry but less pure in terms of class separation.

Results for ARMOUR. ARMOUR exhibits a similar fragility under EHR perturbation, with MCC drops of −0.220 at 1-month (0.293 → 0.073), −0.333 at 6-month (0.357 → 0.024), and −0.432 at 12-month (0.421 → −0.011), corresponding to an average decrease of approximately 0.33 points. In contrast, the learned class tokens following bidirectional cross-attention, i.e., the modality-specific tokens CLS_IMG and CLS_TXT for the image and text modalities, respectively, remain relatively stable under noise, with variations typically within ±0.01–0.09.

RSA values (Fig. D.7) indicate that the final fusion representation is a weighted blend of the two class tokens (RSA up to 0.74 with CLS_IMG and 0.98 with CLS_TXT), yet shows almost no geometric alignment with the original EHR RDM (≤0.07). Despite this low alignment, the model remains extremely sensitive to perturbations in EHR. The t-SNE visualizations (Fig. D.10) further reveal only partial preservation of class structure compared with the CROSS variants, and markedly poorer separation than the unimodal EHR embeddings (especially at 6- and 12-month horizons). This seemingly paradoxical behavior has a clear mechanistic explanation. Although the final fused representation no longer resembles the raw EHR geometry (low RSA), EHR plays a critical role during training: it serves as the anchor that shapes the learned class tokens through bidirectional cross-attention and the MoCo-style contrastive alignment objective. The contrastive loss enforces semantic coherence between image and text, but simultaneously forces the class tokens into a compromise subspace that maximizes cross-modal consistency at the expense of preserving the original discriminative

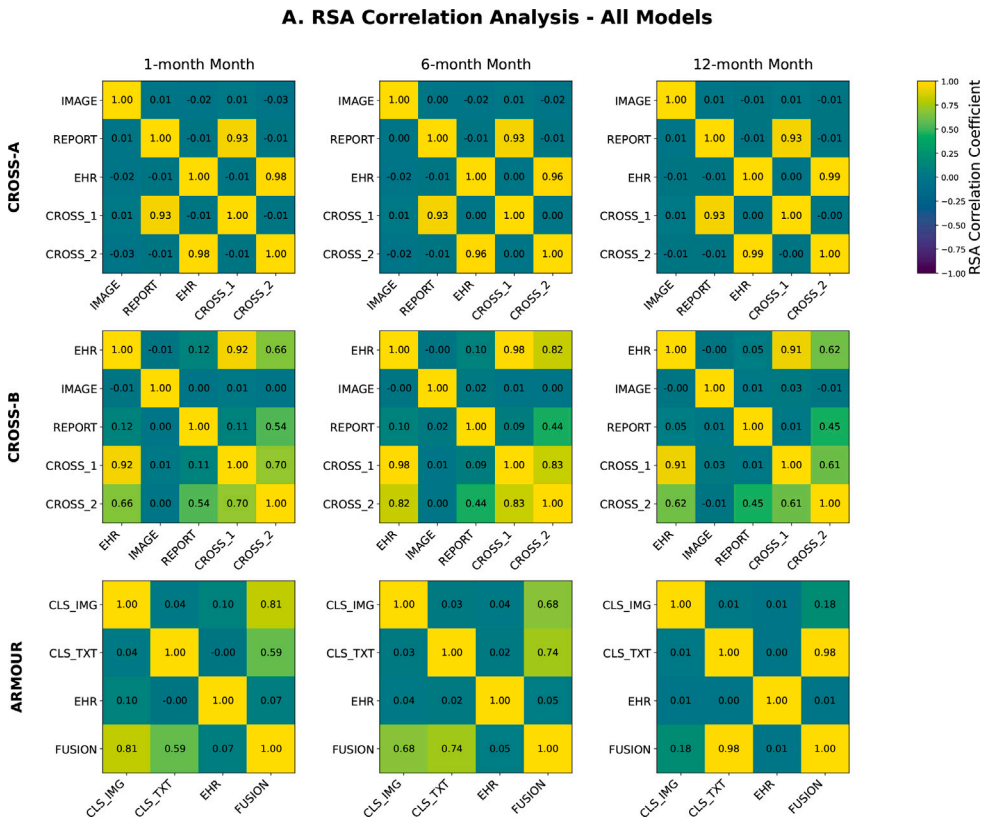


Fig. D.7. Representational Similarity Analysis (RSA) of intermediate fusion architectures. This figure illustrates the representational alignment between different model components for the ARMOUR, CROSS-A, and CROSS-B architectures across three mortality prediction horizons (1, 6, and 12 months). Each heatmap displays Pearson correlation coefficients between the Representational Dissimilarity Matrices (RDMs) of specific feature sets (e.g., EHR, Image, Report, and various Fusion layers). Values closer to 1.0 indicate that the two representational spaces perceive the similarity between patient samples in a similar manner. The analysis reveals how different modalities (EHR, Image, Text) and fusion stages (Cross-attention layers or Fusion blocks) converge or remain distinct during the learning process.

geometry of EHR. The projection of all modalities into a shared 512-dimensional space, followed by the averaging effect of cross-attention and contrastive regularization, creates a severe information bottleneck. Consequently, much of EHR's original class-separating power is diluted to satisfy the cross-modal constraints.

The learned class tokens (CLS_IMG and CLS_TXT), being highly abstracted and optimized representations, are inherently more robust to small Gaussian perturbations. In contrast, the raw EHR embedding remains the hidden teacher that the entire system was trained to trust; when it is noised at inference time, the class tokens, which were shaped under the assumption of clean EHR, lose their predictive value, leading to the observed catastrophic drop in performance.

Overall, ARMOUR does not achieve true multimodal integration. Instead, it collapses onto a compromise representation that sacrifices the strong class separability originally provided by EHR without adequately leveraging complementary information from CT imaging and radiology reports. This explains both its moderate predictive performance relative to late fusion and its extreme sensitivity to EHR noise.

Interpretation and link to modality ordering. Taken together, the RSA and noise-injection results demonstrate that intermediate fusion forces the network to over-rely on a single dominant modality, most often EHR. When this dominant modality is perturbed or when the cascade ordering is suboptimal, the model cannot recover meaningful complementary signals from the remaining modalities. As a direct consequence, class separability collapses in the fused space (t-SNE), and performance becomes catastrophically sensitive to noise. Late fusion, by simply averaging independent unimodal predictions, avoids this pitfall entirely and remains robust to perturbations in any single source.

Both CROSS and ARMOUR share a broader and more fundamental limitation: the difficulty of aligning intrinsically heterogeneous latent spaces. Their attention-based mechanisms implicitly assume that EHR, imaging, and clinical text occupy compatible representational geometries, yet the RSA analyses reveal extremely low correlations among the original modalities. This mismatch drives the models toward suboptimal solutions characterized by modality dominance rather than genuine integration.

In summary, although theoretically expressive, complex intermediate fusion strategies appear ill-suited for highly heterogeneous clinical data unless explicit alignment or regularization mechanisms are introduced. Simpler approaches such as late fusion, which preserve modality-specific representations until the final decision stage, prove substantially more stable and reliable in practice. Future work should therefore explore targeted solutions such as mutual-information-based redundancy reduction, modality-specific gating, or hierarchical fusion schemes that integrate information in a controlled and progressive manner.

Appendix E. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compmedimag.2026.102794.

Data availability

Data are publicly available via the INSPECT repository. Code supporting the analyses is available at https://github.com/som-shahlab/INSPECT_public.

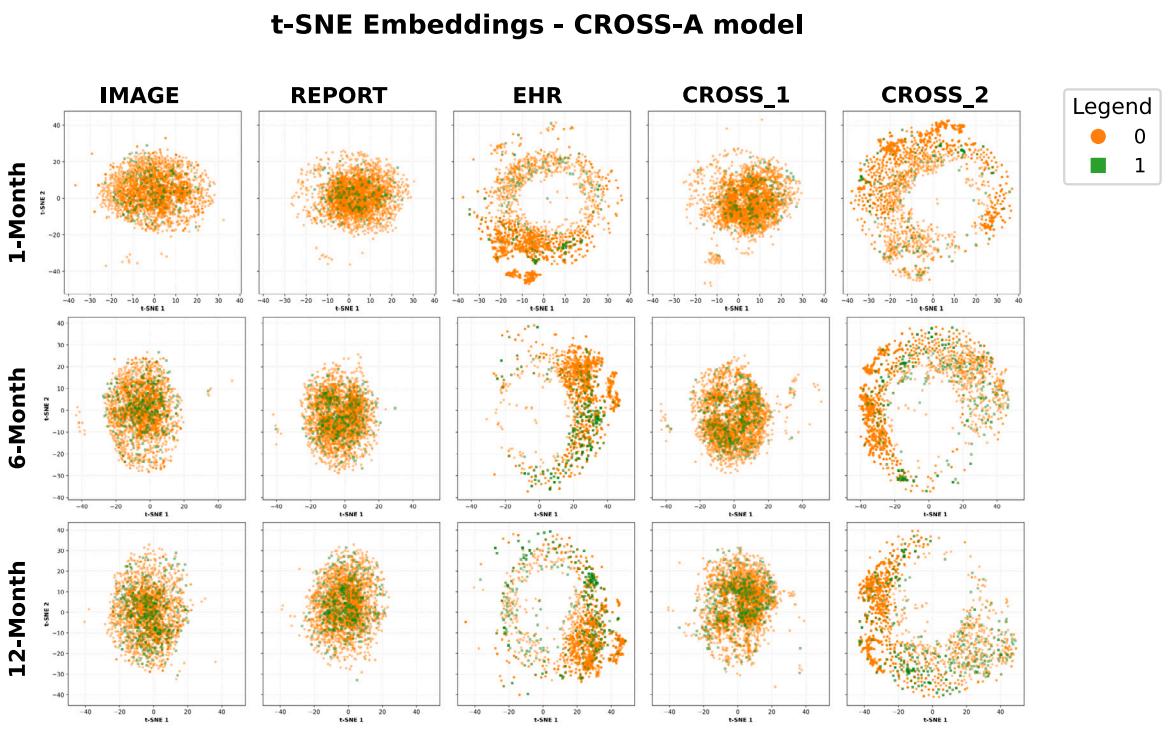


Fig. D.8. t-SNE Visualization of Patient Embeddings for the CROSS-A Architecture. This figure provides a low-dimensional visualization of high-dimensional patient representations generated by the CROSS-A architecture. The vertical layout separates the three prediction tasks: 1-month (top), 6-month (middle), and 12-month (bottom) mortality. For each task, a horizontal grid shows t-SNE projections from different stages of the model, including unimodal encoders (EHR, Image, Report) and subsequent fusion/cross-attention layers. Individual points represent patients, colored by their clinical outcome: orange points denote survivors (Class 0) and green points denote mortalities (Class 1). These plots demonstrate the model's ability to cluster patients by risk and show how discriminative information is refined as it moves through the multimodal fusion pipeline.

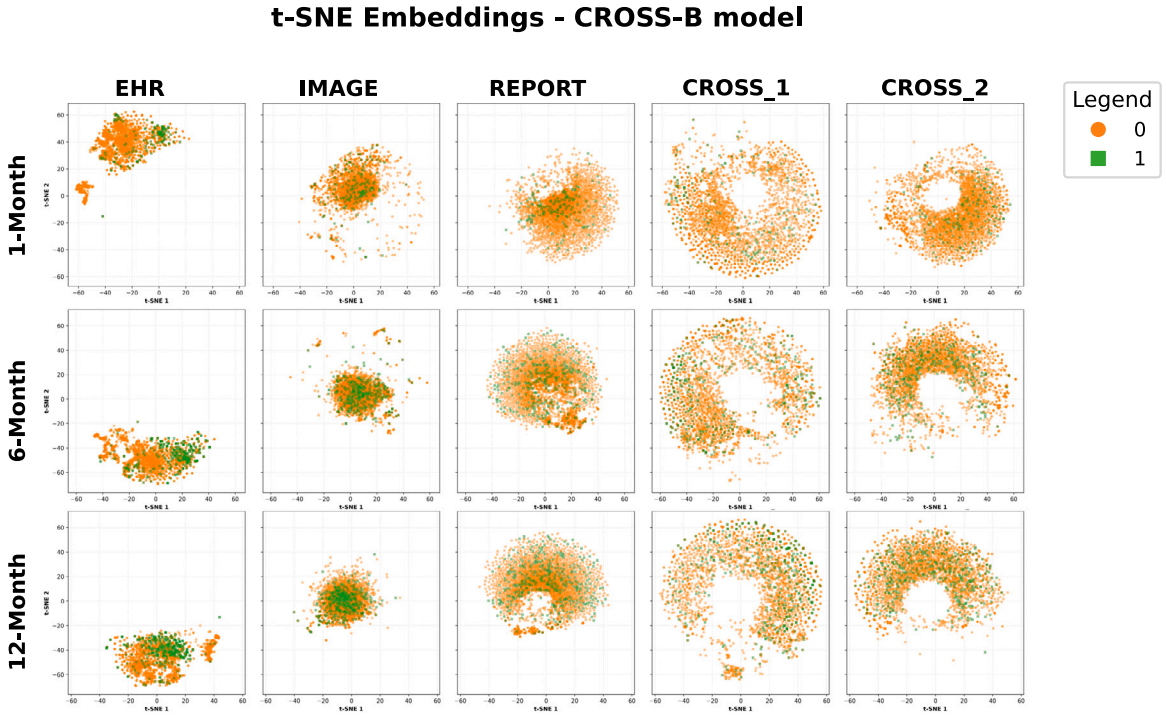


Fig. D.9. t-SNE Visualization of Patient Embeddings for the CROSS-B Architecture. This figure provides a low-dimensional visualization of high-dimensional patient representations generated by the CROSS-B architecture. The vertical layout separates the three prediction tasks: 1-month (top), 6-month (middle), and 12-month (bottom) mortality. For each task, a horizontal grid shows t-SNE projections from different stages of the model, including unimodal encoders (EHR, Image, Report) and subsequent fusion/cross-attention layers. Individual points represent patients, colored by their clinical outcome: orange points denote survivors (Class 0) and green points denote mortalities (Class 1). These plots demonstrate the model's ability to cluster patients by risk and show how discriminative information is refined as it moves through the multimodal fusion pipeline.

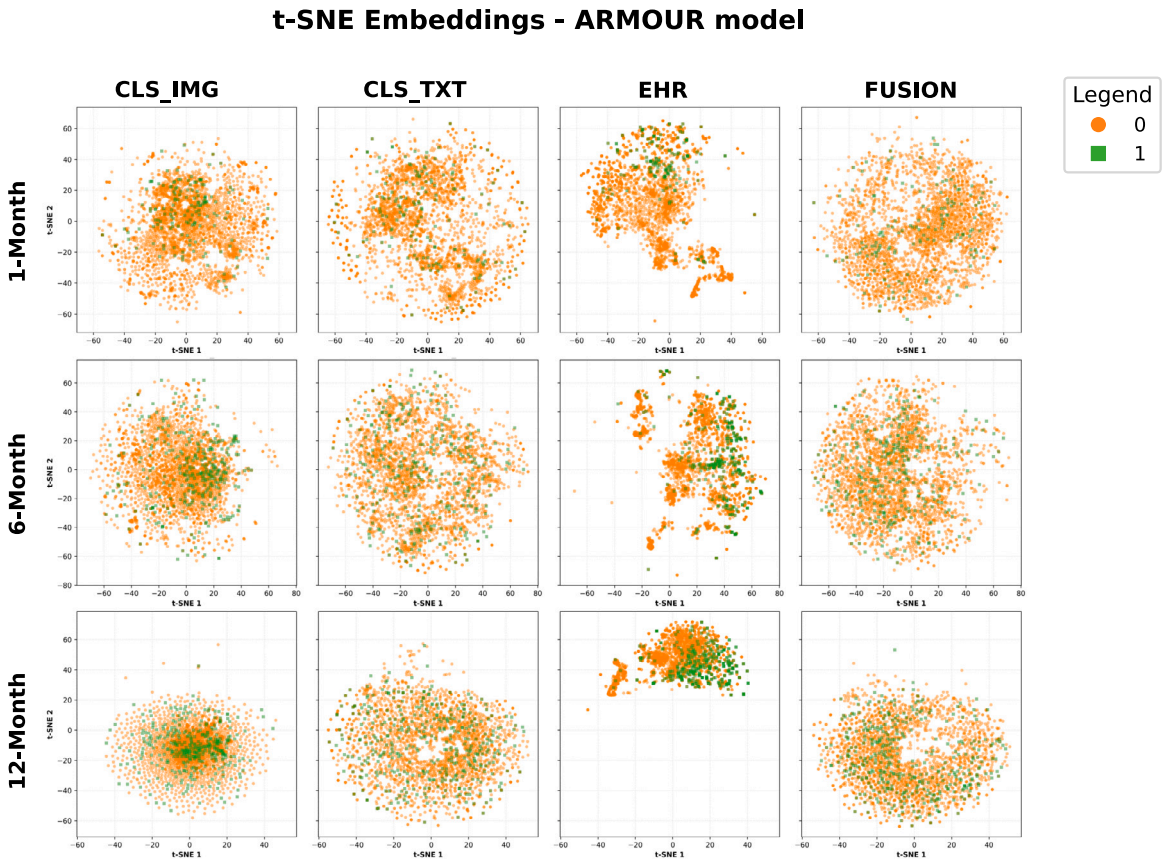


Fig. D.10. t-SNE Visualization of Patient Embeddings for the ARMOUR Architecture. This figure provides a low-dimensional visualization of high-dimensional patient representations generated by the ARMOUR architecture. The vertical layout separates the three prediction tasks: 1-month (top), 6-month (middle), and 12-month (bottom) mortality. For each task, a horizontal grid shows t-SNE projections from different stages of the model, including unimodal encoders (EHR, Image, Report) and subsequent fusion/cross-attention layers. Individual points represent patients, colored by their clinical outcome: orange points denote survivors (Class 0) and green points denote mortalities (Class 1). These plots demonstrate the model's ability to cluster patients by risk and show how discriminative information is refined as it moves through the multimodal fusion pipeline.

References

Amin, K.D., Weissler, E.H., Ratliff, W., Sullivan, A.E., Holder, T.A., Bury, C., Francis, S., Theiling, B.J., Hintze, B., Gao, M., et al., 2024. Development and validation of a natural language processing model to identify low-risk pulmonary embolism in real time to facilitate safe outpatient management. Ann. Emerg. Med. 84 (2), 118–127.

Aujesky, D., Obrosky, D.S., Stone, R.A., Auble, T.E., Perrier, A., Cornuz, J., Roy, P.-M., Fine, M.J., 2005. Derivation and validation of a prognostic model for pulmonary embolism. Am. J. Respir. Crit. Care Med. 172 (8), 1041–1046.

Barco, S., et al., 2021. Global epidemiology of pulmonary embolism and deep vein thrombosis. Lancet Respir. Med.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1), 289–300.

Cahan, Noa, Klang, Eyal, Marom, Edith M., Soffer, Shelly, Barash, Yiftach, Burshtein, Evyatar, Konen, Eli, Greenspan, Hayit, 2023. Multimodal fusion models for pulmonary embolism mortality prediction. Sci. Rep. 13 (1), 7544, Nature Publishing Group UK London.

Chen, X., Fan, H., Girshick, R., He, K., 2020. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

Cicek, V., Orhan, A.L., Saylik, F., Sharma, V., Tur, Y., Erdem, A., Babaoglu, M., Ayten, O., Taslicukur, S., Oz, A., et al., 2025. Predicting short-term mortality in patients with acute pulmonary embolism with deep learning. Circ. J. 89 (5), 602–611, Publisher: The Japanese Circulation Society.

Colak, E., Kitamura, F.C., Hobbs, S.B., Wu, C.C., Lungren, M.P., Prevedello, L.M., Kalpathy-Cramer, J., Ball, R.L., Shih, G., Stein, A., et al., 2021. The RSNA pulmonary embolism CT dataset. Radiol.: Artif. Intell. 3 (2), e200254, Radiological Society of North America.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 248–255.

Dhar, J., Zaidi, N., Haghighat, M., Roy, S., Goyal, P., Alavi, A., Kumar, V., 2025. Multimodal fusion learning with dual attention for medical imaging. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, pp. 4362–4371.

Diaz-Lorenzo, I., Alonso-Burgos, A., Reyes, A. Friera, Blanco, R.E. Pacios, de Benavides Bernaldo de Quiros, M.d.C., Madueño, G. Gallardo, 2024. Current role of CT pulmonary angiography in pulmonary embolism: A state-of-the-art review. J. Imaging 10 (12), 323.

Eini, P., Eini, P., Serpoush, H., Rezayee, M., Tremblay, J., 2025. Advancing mortality prediction in pulmonary embolism using machine learning algorithms—Systematic review and meta-analysis. Pulm. Circ. 15 (3), e70166.

Guarrasi, V., Aksu, F., Caruso, C.M., Di Feola, F., Rofena, A., Ruffini, F., Soda, P., 2025. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. Image Vis. Comput. 105509.

Guo, Z., Wang, Z., Ge, R., Yu, J., Qin, F., Tian, Y., Peng, Y., Li, Y., Wang, C., 2024. PE-MVCNET: Multi-view and cross-modal fusion network for pulmonary embolism prediction. In: 2024 IEEE International Symposium on Biomedical Imaging. ISBI, pp. 1–5, Available: https://api.semanticscholar.org/CorpusID:268032229.

Hansen, P.C., 1987. The truncated SVD as a method for regularization. BIT Numer. Math. 27 (4), 534–553.

Hassan, S.M., Mylvaganam, R., Didebulidze, T., Khalid, M., Nardelli, P., Piazza, G., Estepar, R.S.J., Cuttica, M.J., Johnson, S., Dao, N., et al., 2025. Leveraging hybrid natural language processing techniques for large-scale pulmonary embolism identification. JACC: Adv. 101845.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9729–9738.

Huang, S.-C., Huo, Z., Steinberg, E., Chiang, C.-C., Lungren, M.P., Langlotz, C.P., Yeung, S., Shah, N.H., Fries, J.A., 2023. Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. arXiv preprint arXiv:2311.10798.

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P., 2020. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit. Med. 3 (1), 136.

Huang, Shih-Cheng, Pareek, Anuj, Seyyedi, Saeed, Banerjee, Imon, Lungren, Matthew P., 2020. Fusion of medical imaging and electronic health records using deep learning:

a systematic review and implementation guidelines. NPJ Digit. Med. 3 (1), 136, Nature Publishing Group UK London.

Huang, Shih-Cheng, Pareek, Anuj, Zamanian, Roham, Banerjee, Imon, Lungren, Matthew P., 2020. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci. Rep. 10 (1), 22147, Nature Publishing Group UK London.

Jandoubi, B., Akhloufi, M.A., 2025a. Multimodal artificial intelligence in medical diagnostics. Information 16 (7), 591.

Jandoubi, B., Akhloufi, M.A., 2025b. Multimodal artificial intelligence in medical diagnostics. Information 16 (7), 591, Multidisciplinary Digital Publishing Institute.

Jiménez, David, Aujesky, Drahomir, Moores, Lisa, Gómez, Vicente, Lobo, José Luis, Uresandi, Fernando, Otero, Remedios, Monreal, Manuel, Muriel, Alfonso, Yusen, Roger D., et al., 2010. Simplification of the pulmonary embolism severity index for prognostication in patients with acute symptomatic pulmonary embolism, Archives of internal medicine. Am. Med. Assoc. 170 (15), 1383–1389.

Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G., 2016. MIMIC-III, a freely accessible critical care database. Sci. Data 3 (1), 1–9.

Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., Luo, Y., 2022. Multimodal machine learning in precision health: A scoping review. NPJ Digit. Med. 5 (1), 171.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N., 2020. Big transfer (BiT): General visual representation learning. In: European Conference on Computer Vision. Springer, pp. 491–507.

Li, Y., Wehbe, R.M., Ahmad, F.S., Wang, H., Luo, Y., 2022. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. arXiv preprint arXiv:2201.11838.

Liu, J., Capurro, D., Nguyen, A., Verspoor, K., 2023. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. J. Biomed. Inform. 145, 104466, Elsevier.

Liu, J., Cen, X., Yi, C., Wang, F.-A., Ding, J., Cheng, J., Wu, Q., Gai, B., Zhou, Y., He, R., et al., 2025a. Challenges in AI-driven biomedical multimodal data fusion and analysis. Genom. Proteom. Bioinform. 23 (1), qzaf011.

Liu, J., et al., 2025b. Challenges in AI-driven biomedical multimodal data fusion and analysis. Genom. Proteom. Bioinform. 23 (1), qzaf011.

Mohsen, F., Ali, H., El Hajj, N., Shah, Z., 2022. Artificial intelligence-based methods for fusion of electronic health records and imaging data. Sci. Rep. 12 (1), 17981.

O'Gallagher, K., Wu, J., Dobson, R.J.B., 2025. Calming the storm: Natural language processing applied to pulmonary embolism research. JACC: Adv.

Paolo, D., Greco, C., Cortellini, A., Ramella, S., Soda, P., Bria, A., Sicilia, R., 2025. Hierarchical embedding attention for overall survival prediction in lung cancer from unstructured EHRs. BMC Med. Inform. Decis. Mak. 25 (1), 169.

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., et al., 2018. Scalable and accurate deep learning with electronic health records. NPJ Digit. Med. 1 (1), 18.

Rashedi, S., Krishnathasan, D., Khairani, C., Bejjani, A., Lo, Y.C., Zarghami, M., Mahajan, S., Caraballo, C., jimenez ceja, jv, Jimenez, D., et al., 2024. Accuracy of rule-based natural language processing models for identification of pulmonary embolism. Circulation 150 (Suppl_1), A4147439.

Sadegh-Zadeh, S.-A., Sakha, H., Movahedi, S., Harandi, A.F., Ghaffari, S., Javanshir, E., Ali, S.A., Hooshanginezhad, Z., Hajizadeh, R., 2023. Advancing prognostic precision in pulmonary embolism: a clinical and laboratory-based artificial intelligence approach for enhanced early mortality risk stratification. Comput. Biol. Med. 167, 107696.

Saito, T., Rehmsmeier, M., 2015. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 10 (3), e0118432.

Simon, B.D., Ozyoruk, K.B., Gelikman, D.G., Harmon, S.A., Türkbey, B., 2025. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. Diagn. Interv. Radiol. 31 (4), 303.

Stahlschmidt, S.R., Ulfenborg, B., Synnergren, J., 2022. Multimodal deep learning for biomedical data fusion: a review. Brief. Bioinform. 23 (2), bbab569.

Steinberg, E., Fries, J., Xu, Y., Shah, N., 2023. Motor: A time-to-event foundation model for structured medical records, 2023. arXiv preprint arXiv:2301.03150. [Online]. Available: https://arxiv.org/abs/2301.03150.

Sucre, A., Sánchez, X.C., Perez-Herrera, L.V., dM Vivanco, M., García-González, M.J., López-Linares, K., Calvo, B., Garin-Muga, A., 2025. Multimodal fusion strategies for survival prediction in breast cancer: A comparative deep learning study. Comput. Struct. Biotechnol. J.

Sun, Zhaoyi, Lin, Mingquan, Zhu, Qingqing, Xie, Qianqian, Wang, Fei, Lu, Zhiyong, Peng, Yifan, 2023. A scoping review on multimodal deep learning in biomedical images and texts. J. Biomed. Inform. 146, 104482, Elsevier.

Teoh, J.R., Dong, J., Zuo, X., Lai, K.W., Hasikin, K., Wu, X., 2024. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. PeerJ Comput. Sci. 10, e2298.

Xu, P., Zhu, X., Clifton, D.A., 2023. Multimodal learning with transformers: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 45 (10), 12113–12132.

Zhang, W., Gu, Y., Ma, H., Yang, L., Zhang, B., Wang, J., Chen, M., Lu, X., Li, J., Liu, X., Yu, D., Zhao, Y., Tang, S., He, Q., 2025. A novel multimodal computer-aided diagnostic model for pulmonary embolism based on hybrid transformer-CNN and tabular transformer. Phys. Eng. Sci. Med. 48, 1107–1126, Available: https://api.semanticscholar.org/CorpusID:278884707.

Zhang, X., Wang, F., Li, Y., Lin, L., Hu, H., Chen, Y.-W., 2024. Multimodal fusion with cross-attention transformer for HCC early recurrence prediction from multi-phase CT and clinical data. In: Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence. MLMI, pp. 92–96.

Zhong, Zhusi, Zhang, Helen, Fayad, Fayez H., Lancaster, Andrew C., Sollee, John, Kulkarni, Shreyas, Lin, Cheng Ting, Li, Jie, Gao, Xinbo, Collins, Scott, et al., 2024. Pulmonary embolism mortality prediction using multimodal learning based on computed tomography angiography and clinical data. arXiv preprint arXiv:2406.01302.

Zhou, Y., Huang, S.-C., Fries, J.A., Youssef, A., Amrhein, T.J., Chang, M., Banerjee, I., Rubin, D., Xing, L., Shah, N., et al., 2021. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from CT and EHR. arXiv preprint arXiv:2111.11665.