Information Fusion 136 (2026) 104573



Contents lists available at ScienceDirect

Information Fusion

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



Full Length Article

# Advancing ICU decision making through multimodal data fusion in predicting cardiovascular disorders

F. Krones [a,*], T. Papastylianou [b,c], G. Qian [d], G. Parsons [a], B. Shukla [e], B. W. Papież [f], A. Mahdi [a]

[a] Oxford Internet Institute, Stephen A. Schwarzman Centre for the Humanities, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, Oxfordshire, UK
[b] School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, University of Essex, Colchester, CO4 3SQ, Essex, UK
[c] Institute of Public Health and Wellbeing, University of Essex, Wivenhoe Park, University of Essex, Colchester, CO4 3SQ, Essex, UK
[d] School of Engineering Mathematics and Technology, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol, BS8 1UB, England, UK
[e] East and North Hertfordshire NHS Trust, Lister Hospital, Coreys Mill Lane, Stevenage, SG1 4AB, Hertfordshire, UK
[f] Big Data Institute, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford, OX3 7LF, Oxfordshire, UK

ARTICLE INFO

*Keywords:*
Multimodal AI
Digital medicine
Intensive care
Medical imaging
Decision support
Model reliability

ABSTRACT

Although ICUs generate rich multimodal data, most studies examine only narrow modality combinations, limiting progress toward clinically useful decision-support systems. We propose a systematic framework for evaluating multimodal fusion strategies in the ICU, enabling structured comparisons of early, intermediate and late fusion across combinations of imaging, tabular, text and time-series data. To assess generalisability across diagnoses and patient groups, we apply the framework to six cardiovascular conditions covering a broad diagnostic spectrum and report performance broken down by age, sex and insurance. Rather than asking whether multimodal fusion helps, we ask under which conditions it helps. We find that multimodal models frequently outperform unimodal baselines, with gains of up to five percentage points in AUC, particularly when modalities provide complementary information or when prediction uncertainty is high. Across the architectures we explored, late fusion is the most robust and best-performing strategy. Using uncertainty-based ranking, we show that multimodal integration is especially effective for correcting high-uncertainty cases, substantially reducing both false-negative and false-positive rates relative to single-modality models. The most informative data are typically available within the first 24 hours of ICU admission. These findings offer guidance for prioritising modalities and deploying multimodal models in real-world clinical settings.

## 1. Introduction

*Background.* Cardiovascular diseases remain the leading cause of death worldwide, responsible for approximately 17.9 million deaths each year, accounting for 32% of global mortality [1]. Importantly, up to 80% of these deaths could be prevented through effective risk management [2]. Artificial intelligence (AI) has demonstrated considerable potential to improve healthcare by increasing diagnostic accuracy and efficiency, particularly in critical care settings. This is especially important as cardiovascular diseases account for a substantial share of ICU admissions [3].

*ICU decision support.* Decision-making in the ICU involves integrating multimodal information from laboratory results, imaging and video data, continuous monitoring signals, and clinical documentation [4]. Nevertheless, most existing AI models are limited to a single data modality [5], which does not reflect how clinicians synthesise information from multiple sources. To be incorporated in real-time hospital environments, clinical decision-support tools must be able to process multimodal data types. The limited availability of AI models with such capabilities limits their deployment to clinical settings. Hence, despite increasing interest in applying AI to intensive care and the increased interest in foundation models for medicine [6], many existing approaches remain poorly aligned with clinical workflows [7] and have not yet been widely adopted. Recent work on transformer-based foundation models trained on longitudinal electronic health records has shown that zero-shot patient trajectory models can accurately predict ICU mortality, length of stay and readmission risk directly from patient health timelines, highlighting the potential of generalist decision-support systems in critical care [8]. In practice, ICUs already rely on various scoring systems that use pre-selected variables to generate simple risk metrics. Examples include **S**equential **O**rgan **F**ailure **A**ssessment score (SOFA) [9] for organ failure, and **A**cute **P**hysiology **A**nd **C**hronic **H**ealth **E**valuation (APACHE) IV [10] for length of stay and mortality predictions. However, these scores are often applied as broad prognostic tools rather than for

* Corresponding author.
*E-mail addresses:* felix.krones@oii.ox.ac.uk (F. Krones), tasos.papastylianou@essex.ac.uk (T. Papastylianou), george.qian@bristol.ac.uk (G. Qian), guy.parsons@oii.ox.ac.uk (G. Parsons), bhavin.shukla1@nhs.net (B. Shukla), bartlomiej.papiez@bdi.ox.ac.uk (B.W. Papież), adam.mahdi@oii.ox.ac.uk (A. Mahdi).

https://doi.org/10.1016/j.inffus.2026.104573
Received 5 January 2026; Received in revised form 6 June 2026; Accepted 23 June 2026
Available online 27 June 2026
1566-2535/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

**Table 1**
**Comparison of exemplary fusion approaches.** Overview of papers on multimodal fusion in healthcare, showing that no consensus exists on a best-performing approach across all clinical applications of AI.

| Study | Application | Best approach | Modalities |
|---|---|---|---|
| Huang et al. [21] | Pulmonary embolism classification | Late | Clinical patient data, CT |
| Imrie et al. [22] | Lesion classification | Late/Ensemble | Images, tabular data |
| Guo et al. [26] | Tumor regions segmentation | Early | PET, CT, MRI |
| Holste et al. [27] | Breast cancer classification | Early/Intermediate | Clinical patient data, MRI |
| Hayat et al. [19] | Mortality & phenotyping | Intermediate | Clinical data, X-rays |

condition-specific decision support and focus on a predetermined selection of variables that have been flagged clinically important, rather than evaluating the whole spectrum of available clinical information and allowing the AI model to dynamically determine feature importance.

*Data availability.* Especially the application of multimodal prediction models to cardiovascular conditions remains under-explored [11], largely due to the lack of suitable public data. The first publicly available datasets focusing on cardio related conditions, such as NIH Chest X-ray [12], CheXpert [13] and PadChest [14], were typically limited to X-rays and basic demographic information. Thus, early multimodal studies in this area were smaller and limited in the number of modalities they could incorporate [15,16]. Considerable progress occurred in 2021 with the release of MIMIC-IV [17], a large multimodal dataset comprising over 200,000 ED patients and 65,000 ICU patients. For the first time, this enabled the linkage of different kinds of tabular data, which in the context of this paper we will be referring to as LVD (for **L**ab tests, **V**ital signs and **D**emographics), with corresponding images [18,19]. In an updated version of MIMIC-IV in 2023 [20], more data formats, such as ECGs and patients' clinical notes, were made available, allowing further linkage.

*Fusion comparison.* Given the recent availability of this multimodal dataset and the advent of foundation models for healthcare [6], the question of how to best fuse different modalities becomes critical, as effective integration determines whether models can fully exploit complementary information and generalise across clinical contexts. Data can be linked at different stages. The three main techniques correspond to when data fusion takes place and are typically referred to as *early*, *intermediate* and *late* fusion [5]. While *early* fusion uses the outputs of frozen feature extraction models as input to a final prediction model, *intermediate* fusion backpropagates the loss function through the entire network. In contrast, *late* fusion aggregates the predictions of individual models. While each of these techniques has been applied separately to classify niche applications in healthcare, few studies have systematically compared the performance of all three fusion approaches on the same dataset in a clinical context [21,22]. Consequently, there appears to be no consensus on an optimal fusion approach for multimodal data (if such an approach does exist), and, in particular, an approach relevant to data collected for clinicians in the ICU. While comparisons of different approaches have been undertaken by a few groups (Table 1), these studies typically involve either a limited number of modalities or focus on modalities that are less relevant in an ICU setting.

Complementary to the question of where modalities are fused, a separate line of work studies how to fuse them through information-theoretic objectives, which are loss functions that quantify how much information is shared between modalities through measures such as mutual information, total correlation, or maximal correlation. For example, Huang et al. [23] learn feature representations across multiple variables by maximising their total correlation, while Mei et al. [24] build on this idea for multi-modal brain tumour segmentation, where features from four MRI sequences are combined at an intermediate stage through a partial common information mask and a self-attention module.

A recent survey by Vavekanand et al. [25] provides a broader synthesis of 52 medical multimodal fusion studies, reporting intermediate fusion as the most commonly adopted strategy (60% of studies), an average AUC gain of 5–12% over unimodal baselines, and that only 12% of studies perform true external validation.

*Gaps.* The lack of a systematic evaluation of fusion techniques on ICU-relevant modalities leaves a considerable gap in our understanding of the usability of such approaches to support ICU clinicians in decision-making. The main gaps in the literature that need to be addressed include a systematic comparison of the performance of different data fusion techniques on the many data modalities relevant to patients in the ICU and, subsequently, investigating the contribution of individual data modalities to predictive performance across a range of ICU-related conditions. In addition to integrating information from multiple modalities, another major obstacle to clinical implementation in the ICU lies in accounting for the staggered availability of clinical data. Addressing those gaps is a necessary prerequisite to build robust, clinically reliable foundation models for decision-support in the ICU. However, most research on multimodal fusion assumes that all input features are available simultaneously [21,22]. Among the studies that do account for the temporal dynamics of data availability, the focus is often restricted to a single modality [28,29]. Few approaches, however, address dynamic data acquisition across multiple modalities; one such study aimed to predict mortality and other outcomes after ICU admission [30].

The core contributions of this paper are:

- **Data processing and exploration:** We created a comprehensive and efficient pipeline for MIMIC-IV processing, integrating over 100 tables into a harmonized, event-based structure with a hybrid storage approach, alongside interactive dashboards and LLM-assisted documentation to enable intuitive exploration for interdisciplinary collaboration.
- **Systematic comparison:** We conducted the first large systematic comparison of early, intermediate, and late fusion strategies on ICU-relevant clinical data modalities (imaging, time-series, tabular, and text data). In contrast to prior work that has typically evaluated a single fusion strategy on a single condition, our framework allows controlled comparisons across strategies, modalities, and conditions on the same dataset. Across the architectures we explored, late fusion was the most robust strategy, particularly when data availability was limited.
- **When does multimodal fusion help?** Beyond reporting average performance, we used uncertainty-based ranking and per-case misclassification analysis to characterise the conditions under which fusion is beneficial. We show that gains concentrate on high-uncertainty cases, while additional modalities can degrade performance when they introduce mostly noise. In a large percentage of cases, the multimodal model produces a correct prediction where at least one single-modality model would have produced a false positive.
- **Model performance over time:** We evaluated how the model's overall predictive performance and the relative importance of each feature changes over the course of a patient's ICU stay. We show that especially the data collected during the first 24 hours after ICU admission improve predictive accuracy considerably.

## 2. Methodology

This section describes in detail the analytical methods used, following the adjusted TRIPOD + AI checklist [31].

### 2.1. Data

The predictive modelling of cardiovascular conditions in the ICU requires a structured and comprehensive dataset that integrates multiple data modalities [4,32]. However, as highlighted by Kansal et al. [33], only very limited data sources provide the option to link image data, biological signal data, tabular data and free-text notes, i.e., MC-MED for ED patients [33], ADNI for dementia detection [34], and MIMIC-IV for ICU patients [20].

This is why this study used data from the MIMIC-IV (version 3.0) database [17,20]. MIMIC is a collection of freely available, de-identified medical data organised in a relational database. It contains de-identified health records of patients treated at the Beth Israel Deaconess Medical Center in Boston, USA, between 2008 and 2022. Of the 364,627 patients included in MIMIC-IV, approximately 65,000 were recorded as having been treated in the ICU. Overall, MIMIC-IV contains records of 546,028 hospitalisations for these 364,627 patients and 94,458 ICU stays for the 65,000 ICU patients.

### 2.2. Cohort

Eligible patients included those who were 18 years or older and first admitted to the emergency department, before progressing to the ICU. Only the first stay in the ICU was included in the analysis, contingent on a duration of stay of at least 24 hours. Any subsequent ICU stays were excluded. Patients who were not admitted to the ICU, not recorded as being discharged, or died during their first ICU stay were also excluded. After applying these criteria, the final cohort consisted of 28,262 patients (Table A.7).

### 2.3. Data preparation

Despite the richness of MIMIC-IV, its raw format presents challenges for direct use in predictive modelling. Data are distributed across multiple tables, often requiring complex preprocessing steps to ensure alignment across modalities. Furthermore, missing values, temporal misalignment, and variations in data quality introduce additional complexities that must be addressed to extract meaningful clinical insights. Various attempts have been made to combine the diverse types of data in MIMIC-IV, and while these efforts have contributed valuable milestones, they have proven insufficient in addressing the full range of multimodal challenges, highlighting the need for a new processing pipeline. Key contributions published through the official https://physionet.org/ platform include HAIM-MIMIC-MM [18], a software used to create a multimodal dataset by integrating MIMIC-IV v1.0 and MIMIC-CXR-JPG v2.0.0, filtering patients to those with at least one chest X-ray. However, it suffers from several limitations: it applies fixed patient inclusion criteria, reducing flexibility in cohort definition and its applicability to research questions requiring alternative eligibility; relies on outdated requirements, such as Python 3.6 and `pandas` v1, specifically for a 64-bit macOS architecture, making it difficult to extend or reproduce the pipeline on contemporary systems; and excludes ECG waveform data, limiting its use in cardiovascular research [18]. Similarly, the MIMIC-IV-Ext-MDS-ED dataset [35], focused on the first 1.5 hours after patient arrival, offers valuable data such as demographics, vital signs, lab results, and ECGs, but is constrained by a limited data collection window and the absence of imaging data, such as X-rays. Additionally, it is restricted to emergency department data, making it less applicable for ICU-specific research. The Symile-MIMIC dataset [36], while incorporating X-rays, ECGs, and blood lab tests, suffers from the lack of ICU-specific data, as its inclusion criteria are based on time after hospital admission and not necessarily on critical care, and only includes admissions containing all three modalities. The mentioned limitations have motivated the development of a more flexible and comprehensive pipeline to address the evolving demands of multimodal healthcare research.

While basic data cleaning steps have not been performed by the authors of MIMIC-IV to keep the data as close as possible to real-world datasets, patient identifiers were removed as part of de-identification and dates and times were randomly shifted. Moreover, all patients older than 89 were grouped together as being 91. We further preprocessed the data through a modular pipeline designed to integrate multimodal data sources from the MIMIC-IV database. Tabular data, both raw and processed, were merged into a single event-based table. The resultant dataset included laboratory tests, vital signs and demographic information (LVD) (among other information) for each patient, which were then linked to their X-ray images, ECG waveforms and patient notes. Temperatures were all converted to the same unit. Body temperatures below 50 °F were removed. Diastolic blood pressures above 400 mmHg were removed.

Imaging data (chest X-rays) were linked to some of the MIMIC-IV patients using the MIMIC-CXR-JPG v.2.1.0 dataset [37], which includes 377,110 images from 227,835 studies of 65,379 of the MIMIC patients. ECGs were drawn from the MIMIC-IV-ECG dataset v1.0 [38]. This dataset contains 800,035 ECGs (all 500Hz, 12-lead, and 10 seconds in duration) from 161,352 of the MIMIC patients. The data are grouped for this work into tabular data which we call LVD data (admission, administration, treatments, diagnosis codes, static patient information, low-frequency measures), text data (discharge notes, measurement and treatment comments), high-frequency time-series data/waveforms (ECGs) and images (chest X-rays). Vital signs, such as heart rate, blood pressure, etc. are grouped into low-frequency measures and counted towards tabular data, as they are only stored aggregated.

For multimodal fusion, only data containing all required modalities were used. However, the pipeline was created so that it could still operate on samples with missing modalities. To do this, an identifier vector was added to identify missing modalities, which were replaced with zeros. For missing tabular features this was done by adding a binary mask feature for all clinical features and for modalities this was done by adding a binary value for each modality before fusion. As the fusion experiments were restricted to samples with all modalities available, modality-level zero-imputation was inactive in this study and the mask was retained only for forward compatibility with future analyses on cohorts with missing modalities. For tabular LVD data, a similar approach was used. While XGBoost supports missing values by default, for the use of neural networks, missing values were replaced with column means and a mask was added to highlight any missing values. This step is particularly relevant in clinical contexts, as the absence of measurements may itself convey information: for instance, certain data are recorded only when the clinician suspects a particular condition.

### 2.4. Predictors

For LVD data, features were retained if they were present in at least one-third of all patients or had been included in prior studies [32]. Unless otherwise stated, we used the final ECGs and X-ray images taken during the considered patient's stay, as our objective was to benchmark the predictive performance achievable when all available in-stay data are used, independent of timing.

To avoid target leakage, clinical notes were first split into sections using OpenAI's gpt-4.1-mini-2025-04-14 model via the OpenAI API, chosen for its speed and cost-effectiveness at the time of the experiments. The split used a structured-output prompt with a Pydantic schema (C), so that each fragment was assigned to one of 13 predefined sections rather than rewritten. In a second stage, all text related to the *brief hospital course* or *discharge* was removed, ensuring that no post-admission outcome information was available to the prediction models. Due to resource constraints, only the 5280 notes from the fusion dataset were

**Table 2**
**Distribution of target labels across modalities.** The rows correspond to the six conditions considered in this study; the columns show how positive and negative cases were distributed across all observations within that modality, and for the respective condition. 'Fusion' denotes all cases for which all four modalities to the right were present.

| Label | Fusion | LVD | X-rays | ECGs | Text |
|---|---|---|---|---|---|
| All | 5280 | 28,262 | 110,852 | 298,270 | 202,358 |
| (1) Hypertensive diseases - 1 | 3413 (65%) | 17,455 (62%) | 69,322 (63%) | 199,148 (67%) | 106,077 (52%) |
| (1) Hypertensive diseases - 0 | 1867 (35%) | 10,797 (38%) | 41,530 (37%) | 99,122 (33%) | 96,281 (48%) |
| (2) Ischaemic heart diseases - 1 | 915 (17%) | 3271 (12%) | 12,006 (11%) | 52,972 (18%) | 9102 (4%) |
| (2) Ischaemic heart diseases - 0 | 4365 (83%) | 24,981 (88%) | 98,846 (89%) | 245,298 (82%) | 193,256 (96%) |
| (3) Chronic i. heart disease - 1 | 1474 (28%) | 6239 (22%) | 28,720 (26%) | 112,230 (38%) | 37,407 (18%) |
| (3) Chronic i. heart disease - 0 | 3806 (72%) | 22,013 (78%) | 82,132 (74%) | 186,040 (62%) | 164,951 (82%) |
| (4) Cardiomyopathies diseases - 1 | 251 (5%) | 1048 (4%) | 6000 (5%) | 18,108 (6%) | 4937 (2%) |
| (4) Cardiomyopathies diseases - 0 | 5029 (95%) | 27,204 (96%) | 104,852 (95%) | 280,162 (94%) | 197,421 (98%) |
| (5) Dysrhythmias diseases - 1 | 2007 (38%) | 8856 (31%) | 39,745 (36%) | 120,537 (40%) | 36,709 (18%) |
| (5) Dysrhythmias diseases - 0 | 3273 (62%) | 19,396 (69%) | 71,107 (64%) | 177,733 (60%) | 165,649 (82%) |
| (6) Heart failure - 1 | 1532 (29%) | 6279 (22%) | 32,429 (29%) | 95,753 (32%) | 31,515 (16%) |
| (6) Heart failure - 0 | 3748 (71%) | 21,973 (78%) | 78,423 (71%) | 202,517 (68%) | 170,843 (84%) |

cleaned. A subset of around 20 cleaned notes was additionally inspected manually against their raw counterparts; in this sample we did not find evidence of leaked outcome information. To remain conservative, we kept reporting the results including notes separately, so that none of our headline claims depend on this modality.

The following modality-specific preprocessing steps were applied. *Chest X-rays* were resized to $224 \times 224$ pixels with random cropping, rotation and horizontal flipping applied as data augmentation during training. Pixel values were normalised using ImageNet channel-wise means and standard deviations. *ECG recordings* (all 500 Hz, 12-lead, 10 seconds) were converted into mel-spectrograms using the `librosa` library, and the resulting spectrograms were treated as images and resized to $224 \times 224$ pixels, following the same augmentation pipeline as the X-ray data. *Tabular LVD data* were aggregated per patient using the methods described in Tables A.9 and A.10; for neural network models, missing values were imputed with column means and a binary mask was appended to indicate missingness, while XGBoost handled missing values natively. *Clinical text notes* were tokenised using the ClinicalBERT tokeniser with a maximum sequence length of 512 tokens for the fine-tuned model, or passed as plain text to the GPT-4.1-mini API for zero-shot classification.

### 2.5. Outcome

The objective of this study was to predict cardiovascular conditions, identified using ICD-10 diagnosis codes assigned at ICU discharge. Each was modelled as a binary classification task. The selection and grouping of these diagnoses were informed by clinical relevance, label frequency and shared risk factors, and refined after consultation with an ICU clinician, as detailed in Table A.6. The codes were grouped using the latest WHO classification[1]. To maintain analytical consistency, ICD-9 codes were mapped to ICD-10 codes [39].

For each of the above six conditions, patients were split into 'positive' vs 'negative' cases, based on whether they were given the respective diagnosis at ICU discharge or not, as obtained from their ICD-10 codes. Table 2 shows the resulting distribution of these positive and negative labels across each of the test modalities considered in this study. Note that the total number of labels under the LVD column equals the total number of patients making up the study cohort; this is because all LVD observations are compiled as a single LVD set per patient. Conversely, the total number of labels under X-rays. ECG's, and Text exceeds the cohort number, since multiple observations were possible per patient (e.g., each subject could have had multiple X-rays taken during their ICU admission, each counted individually). The Fusion column denotes cases were all 4 types of observations were present for the subjects in question.

The primary evaluation metric was the area under the receiver operating characteristic curve (AUC), which measures the model's ability to discriminate between classes across all possible thresholds. For threshold-dependent assessments, a threshold of 0.5 was used, reflecting the standard practice when no prior bias toward either class exists. We report AUC as the main metric as it is threshold independent and highlight False Negative Rates (FNR) because missed conditions can cause serious harm; equally, we recognise that False Positive Rates (FPR) are also clinically important in the ICU, where false alarms may lead to unnecessary and sometimes invasive interventions and where alarm fatigue is a serious concern [40]. Therefore, thresholds, error rates, and the choice of aggregator should be defined based on the condition and context, to balance the risks of both FNR and FPR, with guidance from clinical experts.

To improve model robustness, models were trained on balanced datasets using random over-sampling of the training data, after initial experiments showed improved results on validation data. The reported results across the paper were always reported on the original distribution to mimic real world settings. Five fold cross-validation was used for the experiments, with all splits performed at the patient level to prevent data leakage between folds; we ensured that the test data were never seen during any training step, especially when different samples were used for pre-training and fine-tuning. In each fold, 80% of patients were assigned to the training partition and the remaining 20% to the test partition. For training, the train set was further split into train and validation data using an 80/20 split, where the validation set was used for monitoring convergence and learning rate scheduling via `ReduceLROnPlateau`. This resulted in an effective split of approximately 64%/16%/20% for training, validation and testing, respectively. The same patient-level splits were used across all modalities and all experiments to ensure comparable evaluation.

### 2.6. Single-modality models

To establish a benchmark and better understand the contribution of individual modalities, single-modality models were trained first. The following models were developed for each of the four main data types available in MIMIC-IV: low-frequency tabular data (LVD, such as demographics and clinical measurements), high-frequency time-series data or waveforms (ECGs), medical imaging (chest X-rays), and clinical text notes.

[1] https://icd.who.int/browse10/2019/en

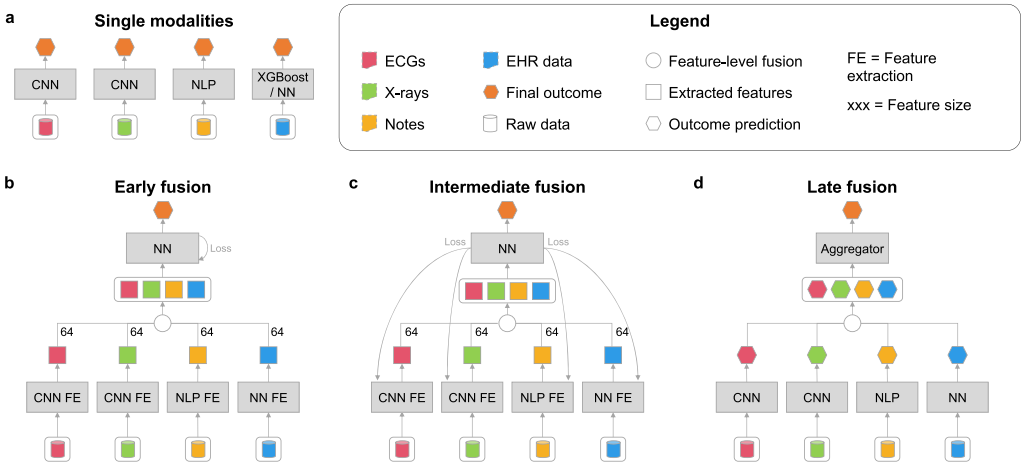



**Fig. 1. Fusion model architectures.** Overview of the four fusion model architectures developed and compared in this study. (a) Single-modality models (no fusion); (b) Early fusion, where features from each pre-trained modality are concatenated and jointly processed; (c) Intermediate fusion, with end-to-end training and loss backpropagation across all modalities; (d) Late fusion, where predictions from individual modalities are combined using different aggregation strategies (*average*, *majority voting*, and *maximum probability*).

**Image models.** For the image classification of the chest X-rays, a ResNet50 model [41] was used, following the approach described in Duvieusart et al. [32]. The final classification layer was removed to allow integration into early and intermediate multimodal fusion strategies. Given evidence that convolutional neural networks (such as ResNet50) are less sensitive to weight initialisation in medical imaging tasks than transformer-based models [42], pretrained ImageNet weights [43] were used, obtained via the `torchvision` package.

**Time-series models.** High-frequency time-series data or waveforms, in particular ECG recordings. As all ECGs were consistent in sampling frequency and duration, they were processed using spectrograms as proposed in Krones et al. [44]. The recordings were transformed into spectrograms using the `librosa` package, and subsequently modelled using the same ResNet50 architecture as applied to image data. This approach was selected as it provides a robust and generalisable alternative to more specialised analyses, achieving up to 95% of the performance of highly tailored methods [44].

**Text models.** Two models were evaluated for clinical text classification. First, ClinicalBERT [45], a BERT-based model pretrained on clinical notes, was fine-tuned using the original tokenizer and model weights from HuggingFace[2], as described in Sheetrit et al. [46]. Second, OpenAI's `gpt-4.1-mini-2025-04-14` model [47] was used in a zero-shot fashion.

**Tabular models.** Two modelling approaches were compared for the tabular data containing lab tests, vital signs and demographics: XGBoost and a feed-forward neural network similar to that in Duvieusart et al. [32]. The neural network consisted of three fully connected layers with batch normalisation between layers. The same neural network structure was used for all fusion experiments to ensure better comparability between the fusion techniques. As late fusion can also accommodate XGBoost models and due to the faster training times, we used XGBoost in several ablation studies; these instances are explicitly mentioned.

### 2.7. Data fusion approaches

Fig. 1 illustrates the three main data fusion strategies applied in this study: early, intermediate, and late fusion (implemented as per Krones et al. [11]). The schematic diagrams in Fig. 1 show how information from different modalities is integrated at each stage. In all cases, modality-specific feature extractors (a ResNet50 CNN for X-rays and ECG spectrograms, ClinicalBERT for text, and a feed-forward neural network or XGBoost for tabular data) first process the raw inputs independently. *Early fusion* concatenates the extracted features from each modality and processes them jointly through a shared classifier, without further adaptation of the feature extraction layers. Specifically, each modality-specific model outputs a 64-dimensional feature vector; these four vectors are concatenated into a single 256-dimensional vector, which is then passed through a multi-layer classifier consisting of fully connected layers with batch normalisation (Fig. 1b). *Intermediate fusion* differs by allowing error backpropagation through the entire network, which enables optimisation of feature extraction but limits the models that can be used, as the error must be able to be backpropagated. As in early fusion, the modality features are concatenated and passed through the same shared classifier, but here the gradients flow back through all modality-specific feature extractors during training, jointly optimising both the feature representations and the classifier (Fig. 1c). This requires all component models to be differentiable, which excludes XGBoost from this setting. *Late fusion* aggregates independent modality-specific predictions using strategies such as average, majority voting, or max, without any additional learned parameters (Fig. 1d). Each single-modality model is trained independently and produces a probabilistic output; these outputs are then combined at the decision level. For average aggregation, the output probabilities of the individual classifiers are averaged. In majority voting, the final prediction is determined by the most frequently predicted class after thresholding. The max strategy selects the class with the highest individual prediction probability. Thresholding was applied *after* aggregation for average and maximum. For majority voting, thresholding occurs *before* aggregation to obtain binary votes.

In this work, for early and late fusion, each modality-specific model was trained separately. In the early and intermediate fusion settings, the final classification layer of each single-modality model was removed. For early fusion, the newly added classifier was trained for 30 epochs using fixed features. In contrast, for intermediate fusion, the entire network was jointly optimised for 30 epochs. 30 epochs were chosen based on preliminary experiments that indicated stable convergence of training and validation losses.

In both early and intermediate fusion, the features were combined via a fully connected layer that connected each output node from each modality-specific model to all input nodes of the final classifier. This corresponds to a concatenation of all extracted features into a single vector. To prevent any single modality from having a disproportionate

[2] https://huggingface.co/medicalai/ClinicalBERT

influence, the final layers of all individual models were configured to contain 64 nodes. To ensure that the choice of architecture for early fusion was not suboptimal, we conducted a comprehensive sensitivity analysis. Specifically, we varied both the number of nodes in the final layers of the individual models (between 16 and 128) and the number of layers within the final layers, as all models contained a series of fully connected layers as their final layers.

**Model comparison.** The performance of the best-performing fusion approach (late fusion, as shown in Fig. 1) was compared with three recent multimodal approaches for healthcare applications: *HAIM* [18], *MedFuse* [19], and *AutoPrognosis-M* [22]. The approaches were selected based on their applicability to the MIMIC-IV dataset, public availability of code, recency (published within the past four years), and relevance to multimodal modelling. *HAIM* [18] first creates individual patient files, then generates standardised embeddings from multimodal data, and finally applies simple classifiers such as XGBoost. *MedFuse* [19] employs an LSTM-based fusion module that sequentially integrates clinical time-series and imaging data while handling missing modalities. *AutoPrognosis-M* [22] automates the construction of multimodal ensembles using AutoML to combine tabular and imaging data via early, joint, and late fusion strategies.

Official implementations as published on GitHub were used to ensure reproducibility. The same pre-processing steps as in the original publications were applied (e.g., image resizing). For each baseline, the hyperparameter configuration published in the original repository was retained, including model-specific learning rates, optimisers, schedulers, batch sizes, and any built-in early-stopping or best-checkpoint selection mechanisms. The only common change applied across baselines and our own models was a maximum training budget of 30 epochs, chosen based on convergence checks of our own fusion model. The exact configurations are visible in our public forks linked below. For models not designed to handle all modalities, only the corresponding modality data were used. Where multiple fusion approaches were proposed, the best performing configuration was selected. All fusion experiments, including training and testing under cross-validation, were performed on the same cohort of 5280 patients that had information available across all modalities (Table 2), so that the comparison between fusion strategies is not confounded by modality-level imputation.

The AutoPrognosis-M model [22] was the only repository not built on the MIMIC-IV dataset; we therefore used the same tabular features and pre-processing steps as in this work. The other models were adjusted to run with Version 3.0 of the MIMIC-IV dataset to ensure comparability. We published the adjustments as forks of the original repositories so readers can track the changes easily. For AutoPrognosis-M [22] this includes the adjustments to run with MIMIC-IV ( GitHub[3]), while we adjusted the MedFuse model [19] to incorporate data from MIMIC-IV Version 3.0 ( GitHub[4]). For HAIM [18], besides the updates to Version 3.0, several further adjustments were necessary before the code could run successfully ( GitHub[5]).

### 2.8. Ablation studies

After determining the most effective single-modality and fusion approaches, three experiments were designed to assess the contribution of individual modalities. First, the fusion model was evaluated using every modality combination to assess the contribution of each modality and determine which combinations most strongly influenced model performance across different conditions. Second, false-negative and false-positive predictions of the fusion network were analysed to determine how often one, two, or all single-modality predictions were incorrect. This analysis aimed to identify whether prediction errors primarily stemmed from weak single-modality models, the fusion strategy itself, or from erroneous single-modality predictions. Third, the behaviour of the fusion model was examined in instances where the best-performing single-modality model produced incorrect predictions. This experiment aimed to assess whether the fusion model would replicate these errors or potentially correct them.

Next, to understand which specific features contributed to the model's predictions across conditions, the influence of individual LVD features on model performance was assessed using XGBoost feature importance scores. These scores reflect the average gain across all decision trees in the model's loss function when a feature is used for splitting, thus quantifying how much each feature contributes to the model's predictive power. Moreover, Shapley values were computed using the Python package `shap` [48] to capture the contribution of each feature to individual predictions.

Given the limitations of clinical resources, a further analysis investigated whether model confidence could guide the need for collecting additional modalities. The idea was that if model confidence is informative, it could not only indicate when additional information would be beneficial but also mimic clinical reasoning, where additional tests are requested mainly when there is uncertainty. This approach was inspired by the Oracle metrics framework proposed by Tran et al. [49] and the performance gain analysis introduced by Imrie et al. [22]. In this work, the absolute distance between the probabilistic prediction value and the 0.5 decision threshold was used as the confidence measure. Predictions from single-modality models were ordered by confidence, and the corresponding increase in accuracy from applying the fusion model was quantified by calculating the F1-score difference between the single modality models and the multimodal versions.

### 2.9. Prediction improvement over time as data become available

The evolution of prediction quality and modality contributions over time was examined. Based on the typical post-admission timelines for each data type, the analyses investigated when reliable predictions become possible and how the timing of measurements affects performance. Given its strong performance in previous experiments, late fusion was selected as the representative strategy for all ablation analyses.

The availability of X-ray, ECG and LVD data over the course of patients' ICU stay was examined. To assess the impact of data availability on prediction performance, models were trained using only LVD features accessible at successive time points after admission, while keeping the diagnosis at discharge labels as target variables. This analysis focused solely on LVD features, as X-rays and ECG recordings are often already available at the point of ICU admission. The same processing steps as described in Tables A.9 and A.10 were used.

### 2.10. Ethical approval

Prior to the start of the project, ethical approval (SSH_OII_CIA_22_108) was obtained through the departmental research ethics committee using the CUREC 1A form, following the University of Oxford's Central University Research Ethics Committee (CUREC) guidelines. Although the study relied exclusively on secondary data analysis of the publicly available, de-identified MIMIC-IV database and did not require additional ethical approval, this step reflects a commitment to responsible research conduct. Access to MIMIC-IV was granted after completion of the required training on research with human participants and acceptance of PhysioNet's Data Use Agreement, which mandates safeguarding the data, refraining from any reidentification attempts, and adhering to data security standards.

### 2.11. Implementation details

For the neural network models a Binary Cross Entropy loss was used and the Adam optimiser. All images were resized to a resolution of 224 ×

[3] https://github.com/felixkrones/AutoPrognosis-Multimodal
[4] https://github.com/felixkrones/MedFuse
[5] https://github.com/felixkrones/HAIM

**Table 3**
**Comparison of proposed fusion pipeline to state-of-the-art models.** State-of-the-art models from the literature were trained and evaluated on the proposed classification task of six cardiovascular conditions, based on the ICD-10 groupings presented in Table A.6. Reported is the average test AUC over 5-fold CV and over the different conditions. If multiple fusion strategies were proposed in the literature, the best-performing was used. For a fair comparison, the same dataset of size 5280 patients was used for all models (Table 2). Table B.11 contains the per disease performance.

| Paper/Model | Modalities | Original task | Fusion strategy | AUC mean (SD) |
|---|---|---|---|---|
| Our model (baseline) | LVD, X-rays, ECGs | ICD-10 codes | Late (avg.) | 0.773 (0.030) |
| HAIM [18] | LVD, X-rays | CXR labels | Early | 0.749 (0.066) |
| MedFuse [19] | LVD, X-rays | ICD-9 codes | Intermediate | 0.772 (0.029) |
| AutoPrognosis-M [22] | LVD, X-rays | Skin lesion | Late/Ensemble | 0.763 (0.041) |

224, using random cropping, rotation and flipping during training. A learning rate of 0.0001 was set and `ReduceLROnPlateau` was used as the learning rate scheduler with a minimum rate of 0. All models were fine-tuned for 30 epochs. We used over-sampling during training to correct for imbalances. For fine-tuning either a single NVIDIA L40S or NVIDIA A100 80GB PCIe GPU was used.

### 2.12. Statistical analysis

The term significant denotes results with a p-value of less than 0.05. All statistical tests were two-sided. To assess performance differences between models evaluated on identical cross-validation folds, we used paired-samples t-tests on fold-level performance metrics, thereby accounting for the dependence induced by shared data splits. Data partitioning was performed at the patient level, so that no patient appeared in more than one fold and each patient contributed a single prediction per condition; there were therefore no repeated within-patient observations or cross-fold patient dependencies for which a mixed-effects model would be required. As the low-frequency signals were aggregated into static features rather than modelled as sequences, feature-level autocorrelation did not enter the model comparison. Given the number of pairwise comparisons across fusion strategies, conditions, and baseline models, the resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure, which ranks the raw p-values from smallest to largest and rescales each one by the total number of tests divided by its rank, controlling the expected proportion of false positives among the comparisons declared significant at a false discovery rate of 0.05. For example, with $n = 10$ tests, a raw p-value of 0.01 ranked second smallest becomes $0.01 \times 10/2 = 0.05$ after adjustment.

## 3. Results

### 3.1. Participant characteristics

Table 2 shows the prevalence of clinical conditions (as per the diagnosis at the time of discharge) among individuals in the cohort. While some conditions, such as (1) Hypertensive diseases, were present in over half the cohort, others, such as (4) Cardiomyopathies diseases, were diagnosed in only a small proportion. Most static data, including patient demographics, are available immediately after admission, as this is when most background information are recorded and basic vital signs measured. Over time, other data sources, including treatment information, low-frequency measurements, X-rays and ECG signals become available. Table A.8 shows patient demographics; the cohort comprises primarily older, white patients with a roughly equal sex distribution.

### 3.2. Model performance

#### 3.2.1. Benchmark comparison

Table 3 presents a comparison between the performance of our framework and recent state-of-the-art approaches. The average AUC scores achieved by our model are comparable to the performance of other methods, with no statistically significant pairwise differences (all BH-adjusted $p \geq 0.93$), demonstrating the validity of the proposed approach for the later ablation studies by its ability to perform classification on the target conditions with state-of-the-art performance. The table is intended as a sanity check rather than a competitive benchmark; we do not claim superiority over these baselines, and the focus of this paper remains the systematic comparison of fusion strategies under controlled conditions on the same cohort.

#### 3.2.2. Model performance across medical conditions

Table 4 shows the model performance for each medical condition. There are considerable differences in both the overall predictability of conditions and the relative performance of different models. Examining single-modality performance, the ability of each modality to predict a condition varies depending on the condition type.

#### 3.2.3. Single modality performance

As shown in Table 4, clinical notes were in all cases the best predictor. However, the results need to be interpreted with care, as the exact timing of the notes is unclear and target leakage cannot be completely ruled out. We therefore treat note-augmented rows as a separate supplementary analysis: the comparison of fusion strategies, the comparison with state-of-the-art baselines (Table 3), and the sensitivity analysis (Table 5) are all based on the no-notes rows of Table 4, so that none of our claims about fusion-strategy robustness depends on the note modality.

As shown in Table 4, ECGs were especially helpful in predicting (4) Cardiomyopathies diseases, (5) Dysrhythmias diseases and (6) Heart failure. For the other conditions either the X-rays or LVD data were most helpful. While for (3) Chronic ischaemic heart disease, (4) Cardiomyopathies, (5) Dysrhythmias, and (6) Heart failure X-rays and the LVD data had similar performances, (1) Hypertensive diseases and (2) Ischaemic heart diseases could mainly be predicted by the LVD data.

#### 3.2.4. Multi modality performance

As shown in Table 4, the best performing model was usually the late fusion one (as also confirmed by the early-fusion sensitivity analysis in Table 5). For late fusion, the three aggregation operators differ in their performance. The *average* of per-modality probabilities ($p_{\text{avg}} = \frac{1}{M}\sum_{m=1}^{M} p_m$) acts as a soft consensus: a single extreme $p_m$ shifts the fused score by at most $1/M$, so agreement across modalities drives discrimination. The *maximum* ($p_{\max} = \max_m p_m$) behaves like a permissive "OR": one confident modality can yield a high fused score, which tends to improve sensitivity (true positive rate) but often increases false positive rate (FPR). Majority voting requires first binarising each modality and then taking the mode; this ignores the magnitude of probabilities, discards information about how confident each modality is, and treats all modalities equally regardless of reliability, which together reduce its ability to distinguish between positive and negative cases.

**Table 4**
**Condition-wise AUC performance (CV).** Comparison of the different fusion approaches as shown in Fig. 1 per condition. Reported is the AUC. 5-fold CV was used and the average (SD) values are reported. For the multimodal experiments (Fig. 1), NN were used for the LVD data. Train set describes whether for training the single modality models all available data were used ('All') or only the ones available with all modalities ('Fusion'). For early and intermediate fusion, the final layers were always trained on the fusion data. For late fusion, the aggregation type used during late fusion is given: 'avg' means probabilities were averaged, 'max' means the maximum probability was used and for 'maj' a majority voting was used. In bold the best performing model with notes and without notes included. Comparing the three fusion strategies without notes (rows with Modalities = All, Notes = No, Train = All), late fusion outperformed both early and intermediate fusion at BH-adjusted $p < 0.05$ in five of the six conditions; the late vs intermediate comparison for chronic ischaemic heart disease was borderline ($p = 0.056$). Early and intermediate fusion did not differ significantly in any condition ($p \geq 0.077$).

| Fusion | Modalities | Notes included | Train set | (1) Hypertensive | (2) Ischaemic | (3) Chronic ischaemic | (4) Cardiomyop. | (5) Dysrhythmias | (6) Heart failure |
|---|---|---|---|---|---|---|---|---|---|
| None | X-ray (CNN) | N/A | All | 0.741 (0.007) | 0.691 (0.019) | 0.794 (0.006) | 0.721 (0.027) | 0.792 (0.012) | 0.815 (0.012) |
| None | X-ray (CNN) | N/A | Fusion | 0.675 (0.029) | 0.658 (0.035) | 0.740 (0.020) | 0.637 (0.045) | 0.717 (0.013) | 0.752 (0.018) |
| None | ECG (CNN) | N/A | All | 0.709 (0.010) | 0.703 (0.010) | 0.751 (0.008) | 0.761 (0.011) | 0.817 (0.011) | 0.808 (0.006) |
| None | ECG (CNN) | N/A | Fusion | 0.657 (0.021) | 0.561 (0.032) | 0.669 (0.014) | 0.595 (0.029) | 0.726 (0.013) | 0.676 (0.022) |
| None | LVD (XGB) | N/A | All | 0.798 (0.003) | 0.798 (0.007) | 0.786 (0.005) | 0.738 (0.025) | 0.827 (0.005) | 0.841 (0.007) |
| None | LVD (XGB) | N/A | Fusion | 0.785 (0.014) | 0.786 (0.028) | 0.784 (0.013) | 0.676 (0.039) | 0.809 (0.013) | 0.810 (0.012) |
| None | LVD (NN) | N/A | All | 0.791 (0.006) | 0.790 (0.008) | 0.778 (0.007) | 0.727 (0.016) | 0.800 (0.004) | 0.845 (0.009) |
| None | LVD (NN) | N/A | Fusion | 0.753 (0.011) | 0.746 (0.012) | 0.759 (0.012) | 0.688 (0.022) | 0.764 (0.021) | 0.807 (0.020) |
| None | Notes | Yes (Bert) | Fusion | 0.860 (0.007) | 0.859 (0.013) | 0.902 (0.008) | 0.783 (0.074) | 0.830 (0.017) | 0.879 (0.007) |
| None | Notes | Yes (gpt) | N/A | 0.873 (0.006) | 0.795 (0.013) | 0.740 (0.014) | 0.828 (0.032) | 0.810 (0.007) | 0.818 (0.006) |
| Early | All | No | All | 0.740 (0.021) | 0.728 (0.017) | 0.791 (0.025) | 0.742 (0.041) | 0.808 (0.015) | 0.782 (0.036) |
| Early | All | Yes (Bert) | All (fusion notes) | 0.826 (0.021) | 0.804 (0.028) | 0.889 (0.015) | 0.770 (0.078) | 0.851 (0.020) | 0.857 (0.035) |
| Interm. | All | No | All | 0.741 (0.023) | 0.714 (0.016) | 0.775 (0.039) | 0.719 (0.073) | 0.819 (0.019) | 0.813 (0.013) |
| Interm. | All | Yes (Bert) | All (fusion notes) | 0.827 (0.010) | 0.822 (0.022) | 0.897 (0.008) | 0.789 (0.060) | 0.856 (0.017) | 0.865 (0.041) |
| Late | All (max) | No | All | 0.775 (0.005) | 0.786 (0.023) | 0.814 (0.020) | 0.795 (0.037) | 0.837 (0.013) | 0.846 (0.009) |
| Late | All (maj) | No | All | 0.708 (0.006) | 0.682 (0.024) | 0.725 (0.012) | 0.668 (0.026) | 0.744 (0.020) | 0.763 (0.011) |
| Late | All (avg) | No | All | **0.799** (0.009) | **0.799** (0.020) | **0.824** (0.016) | **0.809** (0.032) | **0.845** (0.013) | **0.860** (0.007) |
| Late | All (avg) | No | Fusion | 0.757 (0.012) | 0.767 (0.016) | 0.784 (0.012) | 0.732 (0.033) | 0.792 (0.009) | 0.828 (0.007) |
| Late | All (avg) | Yes (Bert) | All (fusion notes) | 0.863 (0.008) | **0.862 (0.015)** | **0.911 (0.016)** | 0.846 (0.044) | 0.878 (0.012) | 0.898 (0.008) |
| Late | All (avg) | Yes (gpt) | All | **0.906 (0.009)** | 0.855 (0.011) | 0.871 (0.012) | **0.889 (0.025)** | **0.896 (0.009)** | **0.916 (010)** |
| Late | All (avg) | Yes (Bert) | Fusion | 0.844 (0.015) | 0.843 (0.017) | 0.896 (0.011) | 0.803 (0.041) | 0.854 (0.010) | 0.887 (0.012) |

**Table 5**
**Sensitivity analysis on early fusion architecture.** Condition-wise AUC performance for different variations of late and early fusion, all without clinical notes (see Fig. 1). Reported is the macro AUC per condition. 5-fold CV was used and the average (SD) values are reported. Average aggregation was used for late fusion.

| Fusion | Train set | Variation | (1) Hypertensive | (2) Ischae-mic | (3) Chronic ischaemic | (4) Cardiomyop. | (5) Dysrhythmias | (6) Heart failure |
|---|---|---|---|---|---|---|---|---|
| Early | All | With single modality models trained on all available data | 0.740 (0.021) | 0.728 (0.017) | 0.791 (0.025) | 0.742 (0.041) | 0.808 (0.015) | 0.782 (0.036) |
| Early | Fusion | With single modality models trained on available fusion data | 0.689 (0.022) | 0.657 (0.023) | 0.731 (0.033) | 0.636 (0.052) | 0.722 (0.031) | 0.786 (0.014) |
| Early | All | Reduced final feature layer from 64 to 16 output neurons | 0.668 (0.072) | 0.693 (0.038) | 0.746 (0.042) | 0.693 (0.075) | 0.758 (0.012) | 0.732 (0.040) |
| Early | All | Increased final feature layer from 64 to 128 output neurons | 0.712 (0.049) | 0.722 (0.026) | 0.789 (0.016) | 0.719 (0.034) | 0.820 (0.021) | 0.793 (0.042) |
| Early | All | Only one final linear layer instead multilayer network | 0.746 (0.016) | 0.760 (0.011) | 0.782 (0.023) | 0.776 (0.045) | 0.798 (0.0140) | 0.822 (0.017) |
| Early | Fusion | Only one final linear layer instead multilayer network | 0.693 (0.011) | 0.728 (0.016) | 0.733 (0.017) | 0.668 (0.029) | 0.722 (0.016) | 0.767 (0.020) |
| Late | Fusion | With single modality models trained on available fusion data | 0.757 (0.012) | 0.767 (0.016) | 0.784 (0.012) | 0.732 (0.033) | 0.792 (0.009) | 0.828 (0.007) |

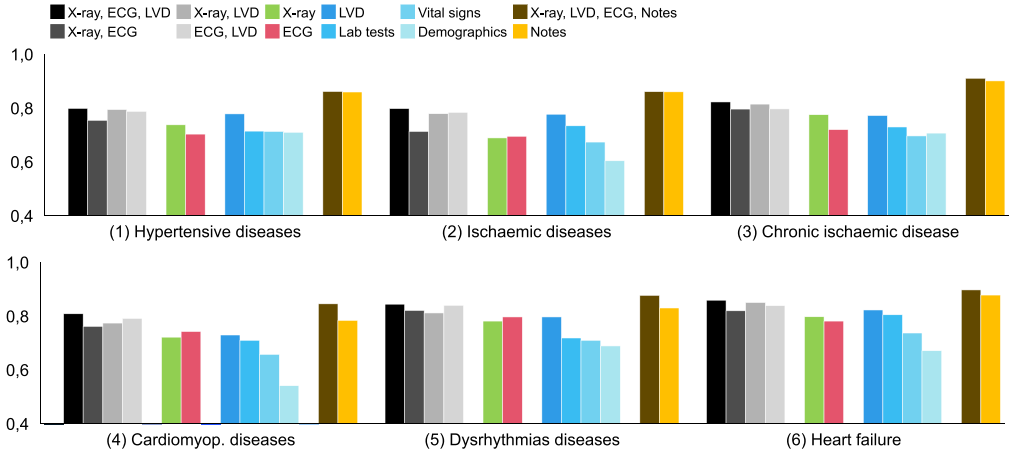



**Fig. 2. Effect of modality combinations.** The figure shows the AUC for different combinations of modalities using the late fusion approach, evaluated on the fusion test split. Each plot relates to one of the six investigated conditions (see Table 2). In blue are the different tabular LVD features. The figure shows that only the combination of the different tabular LVD features results in the often strong LVD performance.

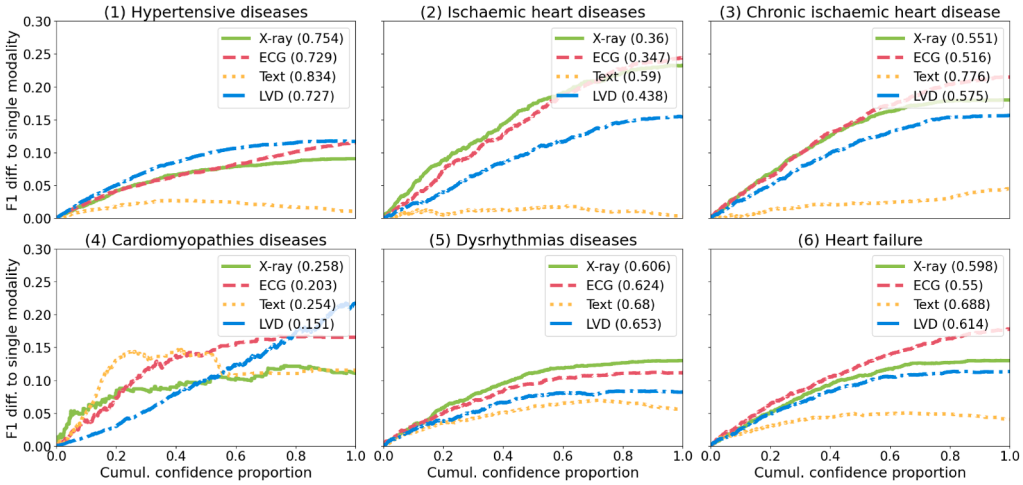



**Fig. 3. Selective acquisition of multimodal data.** The plot shows the cumulative F1-score improvement of the multimodal model over the corresponding single-modality models. The x-axis represents the cumulative proportion of patients for which the multimodal was used instead of the single modality model, ranked by the certainty of the single modality predictions, whereas the y-axis represents the difference in F1-score for that scenario, compared to the 'base' F1-score of the single modality (which is the value indicated in parentheses in the legend for each modality). For example, at 20%, the 20% most uncertain predictions of, e.g., the X-ray model were replaced with the multimodal predictions. Then the F1-score was calculated over all predictions and compared with the F1-score for just the X-rays. The graph indicates that, as the certainty of single-modality predictions increases (towards the right of the x-axis), the added value of the fusion model diminishes (the graph flattens out). This pattern suggests that the multimodal approach is most advantageous when single-modality predictions are less confident. Each plot relates to one of the six investigated conditions (see Table 2). Fig. B.12 shows the F1-score improvement per ten percent bin.

### 3.3. Improvement characteristics

#### 3.3.1. Multimodal effects

Fig. 2 shows that combining two modalities consistently outperforms the corresponding single-modality models for the six conditions of the circulatory system. However, the magnitude of this effect depends on how similar the individual performances of the two modalities are. If their performances are close, combining them leads to a substantial improvement, as seen for LVD and ECG data in (4) and (5). In contrast, if there is a large performance gap, adding the weaker modality has little effect on the stronger one, as in (1) and (2), where adding ECGs to LVD data yields only minor gains. There is a strong negative correlation ($r = -0.869$, $p < 0.001$) between the difference of the combined predictors and the resulting improvement over the stronger predictor. This means that, on average, smaller differences are associated with larger improvements. The correlation was calculated across all diseases and all possible predictor combinations, including those involving two or more predictors. The benefit of including a third modality depends on whether it compensates for a previously missing strong modality. For instance, when two strong modalities – such as X-rays and LVD data in case (1) on Hypertensive diseases – have already been combined, adding another modality (i.e., ECGs) provides little additional benefit.

Fig. B.10 shows cases where the main single-modality predictor would have been incorrect but was corrected by the fusion model. In most cases, the multimodal model fixed the false positives cases, except for (1) Hypertensive diseases, which is also the only condition with significant more positive cases in the training data. Here, the multimodal approach mainly fixed the false negative cases.

Fig. B.11 shows two main findings: First, models with a higher FNR usually have lower FPRs (compare numbers above bars in i) and iii)). However, the relationship is not linear, which means that one can tweak the model into one or the other direction depending on which error leads to the higher costs / is more severe. As long as the class imbalance is not very extreme, as, for example, for (4) Cardiomyopathies diseases (with only 5% positive cases), the class imbalance is not directly correlated to the error rates. For example, (1) Hypertensive diseases and (3) Chronic ischaemic heart diseases show similar error rates, even though (1) has almost two-third positive cases while (3) has over two-third negative cases. Second, Fig. B.11 shows that in most of the cases where the fusion model was wrong, either all or at least more that half of the single modality models were wrong. However, the figure also shows that in up to 17% of the cases, a single incorrect prediction from one modality caused the fusion model to produce an incorrect result, even though the other two single-modality models would have been correct. As *average* aggregation of positive class probabilities was used with a fixed decision threshold, a strongly confident but incorrect modality can override two correct but less confident predictors, where confidence is measured as distance to the decision threshold.

#### 3.3.2. Data efficiency

Figs. 3 and B.12 examine when adding extra modalities is most beneficial and Figs. B.13, B.14, B.15 show a deep-dive analysis into the probability distributions of the different models. As the figures show, the impact on model performance of adding additional modalities decreases as the algorithmic certainty of single-modality models increases. In clinical practice, this could suggest first using the most suitable single modality for a given condition and considering further tests primarily in cases where the model indicates high uncertainty. However, it is important to distinguish algorithmic uncertainty from clinical uncertainty, as low algorithmic confidence does not necessarily imply a clinically ambiguous case.

#### 3.3.3. Feature effects

Figs. 4 and B.9 show that for most diseases, a patient's age is among the top five most important features for classification, with higher age associated with a higher disease probability, except for cardiomyopathies, where no demographic information appears in the top ten. Blood pressure, PTT[6], and urea nitrogen[7] are among the most informative features for most diseases, where "informative" refers to variables that contribute to the model's classification decisions rather than to direct causal factors. As shown in Fig. 2, the predictive strength of the

[6] PTT (Prothrombin time) is a blood test that measures the time it takes for blood to clot, helping assess coagulation disorders and monitor heparin therapy.

[7] Urea nitrogen is a waste product formed from protein metabolism, measured in blood to evaluate kidney function and hydration status.

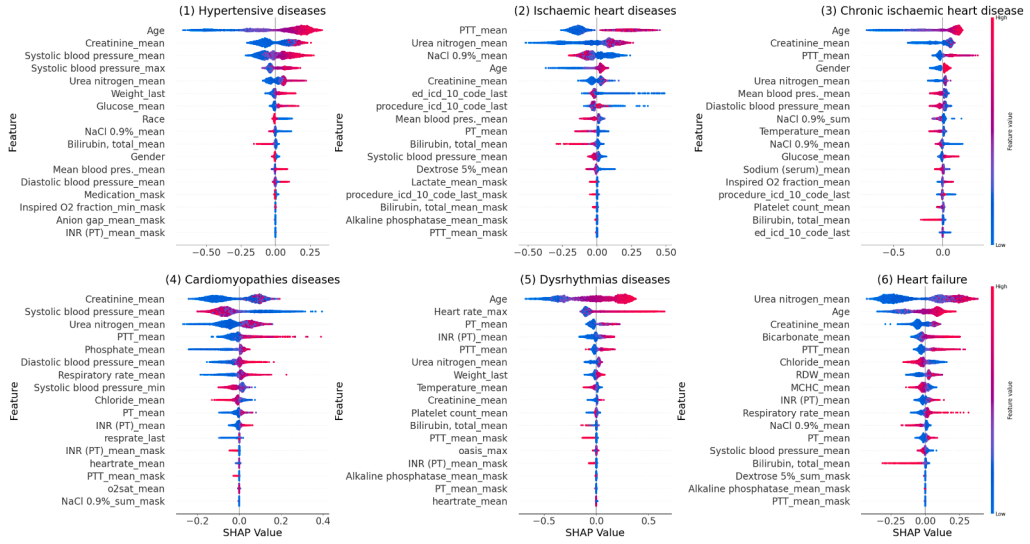



**Fig. 4. Shapley values for the top most important LVD features.** Plotted is the distribution of Shapley values from the XGBoost model across all samples, where each point represents one patient, the position on the $x$-axis indicates the impact on the model output, and the colour encodes the feature value, with higher values shown in red and lower values in blue.

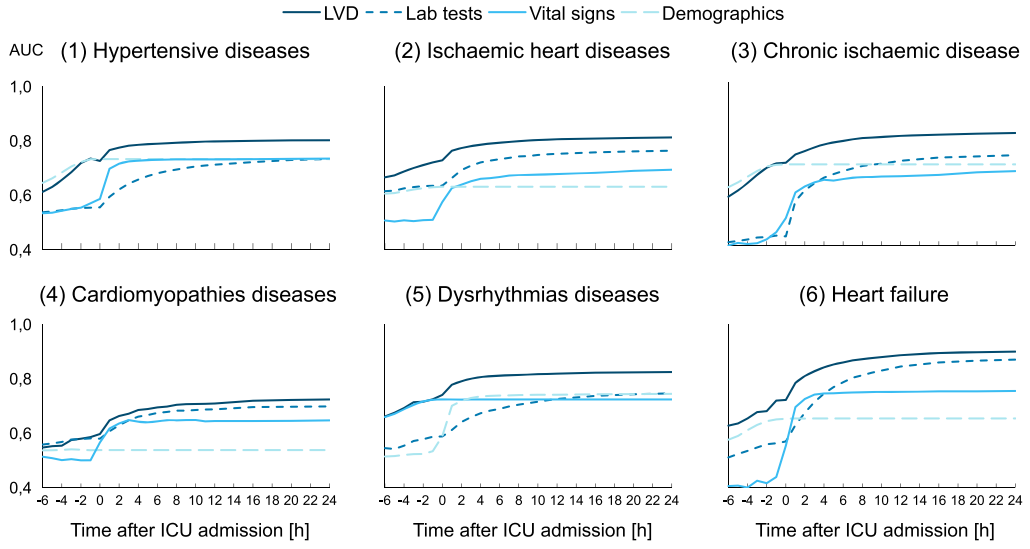



**Fig. 5. Improving model performance with increasing data availability.** Performance of the XGBoost model as a function of feature availability over time after ICU admission. Each plot corresponds to one of the six investigated conditions (see Table 2).

LVD data arises from the combination of different feature types, as the individual types perform considerably worse.

As mentioned above, we only included features that were available for at least a third of the patients. However, based on clinical guidance, we also examined where Troponin and Brain Natriuretic Peptide (BNP) would rank [50]. Using a subgroup of 822 patients with both features available (80% used for training), we re-trained the XGBoost model with these features included. As expected, Troponin ranked among the top ten most important features for (1) Hypertensive diseases, (2) Ischaemic heart diseases, (3) Chronic ischaemic heart diseases, and (6) Heart failure, while being the top predictor for (2) Ischaemic heart diseases and (3) Chronic ischaemic heart diseases. BNP ranked among the top ten most important features for (1) Hypertensive diseases, (3) Chronic ischaemic heart diseases, (5) Dysrhythmias diseases, and (6) Heart failure, while being the top predictor for (6) Heart failure.

### 3.3.4. *Subgroup performance*

To complement the AUC analysis, we report subgroup performance for the late fusion model (all modalities, average aggregation, no notes) on the same 5,280-patient fusion cohort, using chronic ischaemic heart disease as a worked example. AUCs by age, sex and insurance subgroup are shown in Table B.12. The model performed better for younger than for older patients and better for males than for females. This pattern is notable because it does not follow the relative size or positive prevalence of the subgroups: younger patients form the smaller group with fewer positive cases for chronic ischaemic heart disease, whereas males form the larger group with more positives. As Medicare primarily covers older patients, its performance mirrors the older group, showing lower performance than the other insurance types. A full fairness audit on this cohort, including per-group threshold calibration, fairness-constrained optimisation, simulation experiments characterising when per-group thresholds help, and an investigation of bias sources, is the

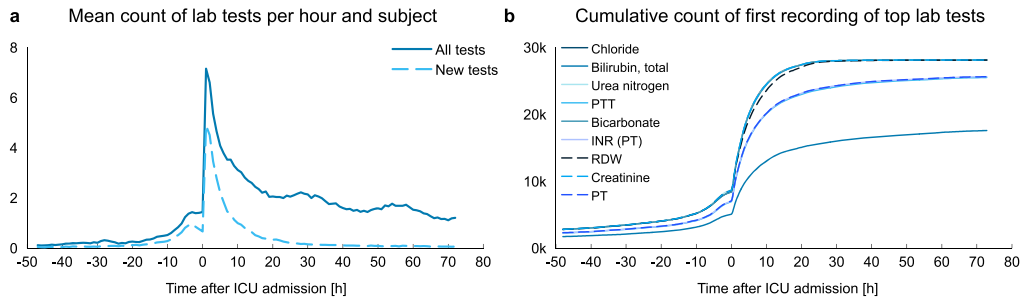



**Fig. 6. Availability of lab tests.** *Left:* Average number of lab tests over time: overall (solid blue) compared with counts including only the first occurrence of an event (light blue, dotted). *Right:* Cumulative count of the top lab tests (those ranking among the top 10 features across diseases, see Fig. B.9) at the time they were first conducted per patient. 55 subjects did not have any lab events and are excluded here.

subject of separate ongoing work that builds on the present manuscript; this paper therefore concentrates on the comparison of fusion strategies.

#### 3.3.5. Threshold-dependent performance

To complement the AUC analysis, we report accuracy, FNR and F1 for the late fusion model on chronic ischaemic heart disease, both at the default 0.5 decision threshold and at an accuracy-optimised global threshold selected on a held-out validation set, and additionally per age, sex and insurance subgroup (Table B.13). Optimising the global decision threshold raises overall accuracy compared with the default 0.5 cut-off, while overall FNR rises in parallel, reflecting the trade-off between sensitivity and specificity that comes with shifting the operating point in a class-imbalanced setting (positive class prevalence 28% in the fusion cohort, see Table 2).

### 3.4. Prediction improvement over time as data become available

Fig. 5 shows that model performance for LVD data improves, particularly during the first hours of the ICU stay, as most data become available in this period. Fig. 6 indicates that this improvement is strongly related to the timing of lab tests, especially for conditions that heavily rely on such tests (e.g., (4) Cardiomyopathies). Although most lab tests are conducted within the first few hours of an ICU stay or even earlier, the figure also reveals that some of the most predictive features are recorded only for a subset of patients during this time. While performing these tests earlier could improve predictive performance even prior to ICU admission, test ordering is typically guided by clinical suspicion and requires balancing potential diagnostic benefit against costs and risks. In this context, the framework could be extended to simulate hypothetical positive versus negative findings, allowing clinicians to assess whether the expected reduction in uncertainty justifies the test. This 'value of information' approach would help identify when a specific test offers enough differentiating power to significantly alter the prognosis or narrow the differential diagnosis, thereby guiding more targeted test ordering.

Fig. A.8 shows that much information is already available before patients are admitted to the ICU. In particular, X-ray and ECG acquisition is often decoupled from the inpatient stay and must be retrospectively mapped to the corresponding hospital stay. Only vital signs and lab test data show a broader distribution in terms of when they are first collected.

## 4. Discussion

*Main findings.* In this work we explored the conditions under which multimodal approaches are most helpful in clinical settings, by comparing and investigating different fusion approaches for integrating multimodal data in the prediction of cardiovascular conditions in ICUs. By comparing early, intermediate, and late fusion strategies, we showed that late fusion offers the most reliable performance under clinical constraints. The models combined imaging, tabular, and time-series data, allowing for a comprehensive assessment of how fusion strategies affect predictive accuracy and robustness in ICU settings.

Our primary aim in this study was to evaluate and compare multimodal fusion techniques rather than to advance clinical knowledge about individual biomarkers or to provide concrete test recommendations to clinicians. While we designed the experiments to be as clinically realistic as possible, the translation of these findings into actionable diagnostic guidance, including the evaluation of clinically sparse but highly relevant tests such as for Troponin or BNP, remains an important direction for future work.

A central finding is the robust effectiveness of late fusion across the architectures we explored. Across all 18 fusion-strategy comparisons in Table 4, late fusion outperformed both early and intermediate fusion at $p < 0.05$ in five of the six conditions, while early and intermediate fusion did not differ significantly in any condition. While early and intermediate fusion strategies theoretically enable richer feature representations by modelling cross-modal interactions [18,26,27], they often require larger datasets and more complex model architectures [5], which are frequently impractical in ICU applications. Late fusion, by contrast, achieved high predictive performance with lower data requirements and simpler model design, making it more suitable for clinical implementation. As shown in our sensitivity analysis (Table 5), early fusion performance varies with architectural choices, and we cannot rule out that a more extensive architecture search could narrow the gap. Our claim is therefore that late fusion is the most robust and best-performing strategy under the architectures and data constraints considered, rather than universally superior. These results support earlier findings by Huang et al. [5] (Table 2), who also identified late fusion as the most practical strategy and thus contribute to the understanding of optimal fusion strategies in critical care.

The results show that the predictive value of each data modality varies by condition, and that combining modalities generally improves performance, as shown in nearly all cases in Fig. 2. This supports the application of multimodal learning to clinical prediction tasks, particularly when the modalities provide complementary information. However, the findings also reveal two important caveats. First, when one modality already captures most of the relevant signal, adding further modalities yields only minor gains, which may not justify the additional complexity or cost. Second, in rare cases where an added modality primarily introduces noise, performance may decline. This indicates that additional modalities can impair performance when they fail to contribute meaningful signal. Misclassification analysis further showed that fusion models rarely corrected errors when all single-modality models failed (Fig. B.11). The largest gains occurred when combining two modalities of similar predictive strength (Fig. 2). From a clinical perspective, the findings suggests that combining data sources is generally beneficial and can improve decision support accuracy. At the same time, the findings

emphasize the need for careful selection of modalities to avoid unnecessary overhead and the risk of reduced performance in specific cases.

The study also highlighted the relevance of evaluating model confidence [49], particularly when determining whether further data collection is warranted. Providing such information when deciding whether to collect additional data can (a) reduce the risk of incorrect predictions and (b) lead to more cost-effective workflows. Efficiency curve analysis demonstrated that the benefits of multimodal fusion were most evident in cases with low prediction certainty (Fig. 3), consistent with earlier studies on uncertainty-based decision-making [22]. Fig. 3 can therefore be read as a cost-versus-benefit curve for a selective acquisition policy: the x-axis is the proportion of patients for whom the multimodal model is used, ranked from most to least uncertain by their single-modality probability, and the y-axis is the resulting cumulative F1 gain. This suggests a possible framework to help steer the decision if acquiring more data (i.e., ordering more tests) is necessary, in which additional tests are conducted only when the model's uncertainty exceeds a predefined threshold. For that, the model could be extended to simulate positive and negative inputs to evaluate the differentiability potential of the extra test with respect to the resulting predictions. Such an approach may reduce unnecessary investigations, lower costs, and lessen patient burden, while preserving diagnostic accuracy. Moreover, by investigating different measures of model uncertainty and implementing system-wide human-in-the-loop approaches where uncertain predictions are referred to a clinician at a defined cost in the decision process, represents a promising direction for future work. A formal separation of epistemic and aleatoric uncertainty, and a prospective simulation of such a clinician-in-the-loop workflow, would require prospective data or a dedicated uncertainty-modelling study and are left to future work.

The temporal analysis showed that the timing of data availability plays a central role in determining predictive performance, consistent with previous findings in other clinical contexts [28,30]. X-ray and ECG data are often available before or at the time of ICU admission, while vital signs and laboratory tests are usually collected in a staggered manner during the first hours of the ICU stay [20]. Model performance improved considerably within the first 24 hours, particularly when laboratory results became available. This suggests that if laboratory test results were available prior to admission, they could further support early assessment and inform better allocation of ICU resources. When laboratory values or vital signs are the primary drivers, reliable predictions become possible only after sufficient data are collected during the first 24 hours. In contrast, when X-rays, ECGs, or demographic features are most informative, accurate predictions could potentially be made earlier. However, as our analysis used the final available imaging and signal data, this potential for early prediction relies on the assumption that pre-admission tests hold comparable diagnostic value to those obtained during the ICU stay. Furthermore, the reliance on accumulated data highlights the value of temporal trends. While our approach implicitly captures some trajectory information by comparing mean and last values, it does not explicitly model rates of change. Given that clinicians often prioritize trends over absolute values, future work could focus on explicit trend features, though this inherently requires a sufficient observation window and recording frequency to derive meaningful gradients. In our cohort, the low-frequency laboratory and vital-sign measures are stored as aggregated values rather than dense time series, so derivative features such as lactate clearance or haemodynamic slopes, and time-aware architectures designed for densely sampled signals, cannot be applied reliably at this resolution.

*Strengths of the study in relation with other studies.* A key strength of this study is its systematic investigation of multimodal fusion strategies across a broad set of cardiovascular conditions in the ICU, addressing limitations of prior work that typically focused on individual diagnoses or a small set of modalities [11]. While previous studies have examined fusion strategies in restricted contexts – such as pulmonary embolism detection [21] or dermatological image classification [22] – this work provides a comprehensive comparison of early, intermediate, and late fusion across six clinically relevant conditions. By using the MIMIC-IV dataset [20], which integrates imaging, time-series signals, LVD data, and clinical text, this study moves beyond earlier approaches that relied on a single dominant modality with limited auxiliary features. The consistent performance of late fusion across tasks offers clarity in light of mixed findings from prior studies, where no fusion strategy showed universal superiority [21,22,26,27].

One limitation in studies addressing multiple conditions in the ICU so far was how prediction targets were defined. Many relied on radiology reports or clinical notes [18,21], which are subject to biases introduced during the diagnostic process or by the referring clinician [32]. In contrast, ICD codes summarise diagnoses across the entire hospital admission and are less susceptible to such biases. Although recent work has begun to adopt ICD codes as prediction targets [35,51], these studies were not fully aligned with the experimental setup proposed here. The limitations were both technical and clinical. Technically, the models were either limited to one fusion approach and did not compare fusion strategies [35] or attempted multi-label prediction of more than 1000 ICD codes [51], which can dilute performance, as models tend to disproportionately optimise for the most frequent codes. Clinically, they used limited observation windows (e.g., only the first 1.5 hours of admission [35]) or the models focused on a single modality [51]. To ensure clinically meaningful groupings and more balanced data distributions, the focus in here was on grouped ICD codes, selecting conditions with similar underlying pathophysiology (see Table A.6).

The presented data preparation pipeline offers several advantages over existing approaches used to process the MIMIC-IV dataset. Unlike previous solutions, such as HAIM-MIMIC-MM, MIMIC-IV-Ext-MDS-ED, and Symile-MIMIC, this pipeline integrates all major modalities available in MIMIC-IV, including ECG waveforms and chest X-rays, and combines them into an event-based format. This comprehensive approach improves temporal alignment across data sources, addressing limitations in earlier efforts, such as the exclusion of ECG data in HAIM-MIMIC-MM [18] and the restricted imaging data in MIMIC-IV-Ext-MDS-ED [35]. Another key improvement is the modular and scalable nature of the developed pipeline. While previous approaches relied on rigid inclusion criteria or outdated software requirements [18], the current framework is designed to adapt to future versions of MIMIC-IV and to operate independently of local computing environments. This flexibility enhances reproducibility and facilitates the extension of the pipeline to different settings and cohorts. Additionally, the hybrid storage strategy optimises data handling by consolidating low-intensity information while externally referencing high-intensity data. This structure enables efficient loading during model training. In summary, the developed pipeline not only addresses the fragmentation and scalability issues of MIMIC-IV but also establishes a reproducible and comprehensive framework that helps with future multimodal research. This structured approach ensures improved temporal consistency, efficient data management, and richer clinical feature representation compared to previous solutions.

*Limitations.* The models were developed and evaluated solely on the MIMIC-IV database [20]. This may restrict generalisability to other clinical environments with different patient characteristics, data availability, and recording practices. Although six cardiovascular conditions were included to broaden the scope, the analysis remains dataset-specific. This reflects an inherent challenge of multimodal research: the limited availability of comprehensive datasets limits the comparisons options across studies and settings. In Krones et al. [11] we summarised available multimodal datasets.

A further limitation concerns the scope of evaluation metrics. We focused on AUC, complemented by FNR, FPR, F1 and the threshold-dependent operating points reported in Table B.13i. Extending this with reliability diagrams, Brier scores and formal Decision Curve Analysis for the late fusion model and its single-modality components, ideally on an external cohort, would be a natural extension of this work before any clinical deployment.

The hospital from which MIMIC data were collected is located in Boston, MA, USA, and only data from patients there were collected. Hence, it is limited to the demographic structure of that region. Moreover, because it is a teaching hospital, it can be assumed that the treatment patients receive is rather extensive compared to the world's average healthcare supply. As a result, several of our quantitative findings, including the 24-hour data-availability window and the relative ranking of modalities, reflect the workflows and ordering practices of this single, well-resourced US academic centre and may differ in resource-limited or non-academic settings. External, multi-site validation is therefore a critical next step before any clinical implementation. In particular, the staggered availability of laboratory tests, vital signs and imaging in resource-limited settings would shift both the optimal time window for prediction and the relative usefulness of individual modalities, an effect that an external-cohort study with different ordering practices would be best placed to quantify. However, as it is a single recording site, the data quality can be assumed to be constant across different demographic groups.

Diagnostic labels were based on ICD codes, which are mainly used for administrative purposes. This may introduce label noise or systematic bias caused by insurance premium and reimbursement schemes [52]. An independent validation of these labels is not possible within MIMIC-IV, since the only available reference, the free-text notes, is itself one of the modelled modalities; label-noise-aware training or probabilistic label treatment is therefore a direction for future work rather than something we can verify on this dataset. Furthermore, patients who died during the hospital stay were excluded from the analysis. This concerned 4716 patients (Table A.7), a group likely to include some of the most severely ill cases, so the cohort underrepresents the highest-acuity end of the ICU population. While this exclusion could introduce a bias by omitting the most severe cases, it allowed us to ensure consistent data availability across modalities and time points, which was essential for comparing fusion techniques. We did not model mortality as a competing risk or within a joint survival-classification framework because our prediction target is the diagnostic classification of cardiovascular conditions from ICD-10 codes rather than time-to-event or survival; such frameworks would address a different question rather than correct the cohort composition.

A further limitation concerns the comparison with state-of-the-art baselines. We retained the published hyperparameter configurations of HAIM, MedFuse, and AutoPrognosis-M rather than performing an independent hyperparameter search per method on our cohort. This choice prioritises reproducibility and stays close to the configurations the original authors reported as best for their settings, but we cannot rule out that a method-specific search on our data would close part of the observed performance gap. The same applies to the choice of baselines: we included the three recent multimodal approaches whose code was publicly available and could be adapted to MIMIC-IV v3.0 within the time constraints of this work, and our published forks make any further re-tuning or extension by other researchers straightforward.

A systematic comparison of computational cost, training time, and inference latency across fusion strategies was beyond the scope of this study, which focused on predictive performance and the conditions under which fusion is beneficial; we note, however, that late fusion supports parallel execution of per-modality models and allows lightweight, modality-appropriate architectures (for example gradient boosting on tabular data), which can offset its apparent overhead at inference time. For the averaging aggregator used in this work, the aggregation step itself has negligible inference cost compared with the per-modality forward passes, so the wall-clock latency of late fusion is dominated by the slowest single-modality predictor rather than by the fusion step.

For the clinical-text modality, the cleaning pipeline relies on a proprietary API; although we use a pinned model snapshot (`gpt-4.1-mini-2025-04-14`) and release the prompts, exact long-term reproducibility outside the OpenAI API cannot be guaranteed. We also note that, given the way MIMIC-IV stores discharge notes, a temporal-control experiment that predicts the diagnosis from notes written strictly before it was made cannot be implemented cleanly within this dataset.

Finally, only a subset of modalities was included that reflects data sources typically available in real-world settings. Computed tomography, echocardiography, and wearable sensors were excluded due to limited data availability. Echocardiography in particular represents an important diagnostic tool in cardiopulmonary care; however, its integration was not feasible in this study. First, echocardiographic video analysis constitutes a separate research domain that requires dedicated pipelines and was therefore beyond the scope of this work. Second, while text-based reports would in principle provide a more direct way of incorporating echocardiography, these notes are not yet publicly released in MIMIC. In practice, extending our late-fusion pipeline to include them would be straightforward. Beyond echocardiography, future extensions should also consider modelling temporal trends over longer time windows and systematically evaluating how varying the inclusion of clinically important tests, such as Troponin or BNP, affects diagnostic differentiation and model uncertainty. A further direction is the systematic evaluation of fusion performance under missing modalities, comparing imputation schemes and quantifying how performance degrades as individual modalities become unavailable across the different conditions; the pipeline already supports masking for this purpose, but a thorough analysis is left to future work.

*Study implications.* This study showcases the value of multimodal machine learning models in addressing the diagnostic complexity of ICU settings. By systematically evaluating fusion strategies across a broad set of cardiovascular conditions, it demonstrates that combining heterogeneous data sources, such as vital signs, demographics, ECGs, laboratory tests, and imaging, can improve predictive performance and support earlier, more reliable clinical decision-making. The comparative analysis of fusion methods revealed that late fusion offers consistent benefits under realistic constraints of data availability, particularly in cases of diagnostic uncertainty. While our analysis focused on ICU data from MIMIC, the framework itself, that is, the systematic comparison of fusion strategies, the modality and timing ablations, the uncertainty-based ranking and the subgroup analysis, is database-agnostic and can be re-applied to the analysis of other multimodal databases, such as the UK Biobank[8]. The specific quantitative findings, however, should not be assumed to transfer without external validation.

Moreover, the findings provide practical guidance for prioritising data collection efforts and integrating machine learning models into clinical workflows. For clinicians, the automated integration of multimodal information has the potential to refine risk stratification and optimise intervention timing. For policymakers, the results support initiatives to improve the interoperability of electronic health record systems and to promote infrastructure that enables the effective use of multimodal data in predictive modelling. Ultimately, this work advances both methodological understanding and practical readiness for deploying uncertainty-aware, multimodal AI systems in critical care. We stress, however, that the reported gains are improvements in predictive performance and not yet validated diagnostic guidance: the framework is intended to support, not replace, clinician judgement. A path towards clinical use would require external multi-site validation, prospective evaluation in the intended setting, and regulatory assessment. We therefore position this study as a methodological benchmark rather than a deployable end-point, and view emerging clinical foundation models as a complementary direction that a systematic comparison of fusion strategies can inform.

## CRediT authorship contribution statement

**F. Krones:** Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation,

[8] https://www.ukbiobank.ac.uk/

Conceptualization; **T. Papastylianou:** Writing – review & editing, Validation, Supervision, Methodology; **G. Qian:** Writing – review & editing, Writing – original draft, Validation; **G. Parsons:** Methodology, Conceptualization; **B. Shukla:** Writing – review & editing, Validation, Methodology; **B. W. Papież:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Conceptualization; **A. Mahdi:** Supervision, Project administration, Conceptualization.

## Author contributions

FK led the study, including conceptualisation, experimental design, implementation of the models, data processing, statistical analyses and drafting of the manuscript. TP contributed to the study design, reviewed and refined the methodological approach, supported with the interpretation of the findings, and substantially revised the manuscript for structure, language and clarity. GQ contributed to the experimental design and interpretation of the findings, and substantially revised the manuscript for structure, language and clarity. GP and BS provided clinical expertise, contributed to the formulation and prioritisation of the research questions, advised on the design and interpretation of the experiments and critically revised the manuscript. BP supervised the project, contributed to the overall study conception and storyline, and reviewed the methodology, results and conclusions for coherence and scientific robustness. AM established the research group, initiated the original project idea and overarching storyline, contributed to the study design, supervised the work, and critically revised the manuscript. All authors read and approved the final manuscript.

## Code availability

We published the adjustments to all repositories. For AutoPrognosis-M [22] this includes the adjustments to run with MIMIC-IV ( GitHub[9]), while we adjusted the MedFuse model [19] to incorporate data from MIMIC-IV Version 3.0 ( GitHub[10]). For HAIM [18], besides the updates to Version 3.0, several further adjustments were necessary ( GitHub[11]).

The developed code will be published via PhysioNet, as we have also published previous work [53]. In contrast to previous work, which were shared as datasets, this submission will be released as a PhysioNet Software Project, similar to the HAIM repository [18]. Publishing the pipeline as software ensures that the underlying data remain protected under the MIMIC-IV access controls and makes it easier for others to rerun the code if the dataset is updated in the future. The software package will include code for data extraction, preprocessing, and integration across MIMIC-IV modalities; scripts to convert the data into an event-based multimodal format; visualisation tools to support expert exploration; and dummy data that reproduces the structure of the original MIMIC-IV tables for demonstration purposes. This setup enables credentialed researchers to reproduce the full preprocessing workflow using their own local copy of MIMIC-IV. To support reproducibility, the release specifies the software environment through a pinned requirements file, controls all model and preprocessing settings through a single configuration file, uses a fixed random seed throughout data splitting, resampling and training, and stores the exact patient-level cross-validation fold assignments (a 5-fold scheme with a nested 64/16/20 train, validation and test split) so that all experiments reuse identical splits.

## 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.

## Data availability

All data used in this study are derived from the publicly available MIMIC-IV database and its associated modules. Access to MIMIC-IV requires successful completion of the PhysioNet credentialing process, including human subjects research training and acceptance of a data use agreement. More information is available at https://physionet.org.

## Appendix A. Data selection

## Appendix B. Results

## Appendix C. Prompt for clinical note cleaning

The following prompt was issued to OpenAI's `gpt-4.1-mini-2025-04-14` via the OpenAI API as part of the note cleaning pipeline. The call uses structured outputs validated against a Pydantic schema, so that the model fills fixed fields rather than producing free-form text. After this prompt was applied, the *Discharge* and *Brief Hospital Course* fields were dropped, and only the remaining 11 admission-side sections were used as input to the downstream models.

```
You are a medical machine learning expert.
Your task is to split medical notes into sections based on when information gets
 ↪  collected.
Do not adjust the text, just split it into sections.
Before you reply, make sure to read the entire note, understand the context and
 ↪  that you followed the instructions below.

**SECTIONS TO SPLIT INTO:**
- Administration
- Demographics
- Allergies
- Chief complaint
- Major surgical or invasive procedure
- Medical and illness history
- Social/Family History
- Medications on admission
- Physical exams
- Imaging
- Discharge
- Brief Hospital Course:
- Other

**INSTRUCTIONS:**
1. Brief Hospital Course is everything that comes after Brief Hospital Course is
 ↪  mentioned in the note.
2. Discharge should contain all information not under Brief Hospital Course, but
 ↪  that are still discharge related, such as discharge labs, discharge exam,
 ↪  discharge medications, etc. Add all discharge related information to the
 ↪  Discharge section not just the ones I mentioned.
3. If you have exams under discharge, make sure to include them in the Discharge
 ↪  section not in the Physical exams section.
4. If you are unsure to which section a piece belongs, add it to the "Other"
 ↪  section.
5. Sections could be empty, but do not remove them.

**OUTPUT FORMAT:**
Json with two keys:
- "text": dictionary, where keys are section names and values are the text of the
 ↪  section
- "confidence": float, a confidence score between 0 and 1 indicating how
 ↪  confident you are that the text is cleaned correctly and does not contain any
 ↪  target leakage
```

[9] https://github.com/felixkrones/AutoPrognosis-Multimodal
[10] https://github.com/felixkrones/MedFuse
[11] https://github.com/felixkrones/HAIM

**Table A.6**
**Label selection for circulatory system disorders.** Refinement of WHO ICD-10 codes in order to derive prediction labels. All prediction groups are modelled as binary prediction problems.

| Label group name | ICD-10 blocks | Motive | Description |
|---|---|---|---|
| (1) Hypertensive diseases | I10 - Essential (primary) hypertension<br>I11 - Hypertensive heart disease<br>I12 - Hypertensive renal disease<br>I13 - Hypertensive heart and renal disease | All same underlying disease processes | Long-term high blood pressure that strains the heart and vessels. |
| (2) Ischaemic heart diseases | I20 - Angina pectoris<br>I21 - Acute myocardial infarction<br>I22 - Subsequent myocardial infarction<br>I24 - Other acute ischaemic heart diseases | Acute, same underlying processes | Reduced blood and oxygen supply to the heart muscle. |
| (3) Chronic ischaemic heart disease | I25 - Chronic ischaemic heart disease | Chronic, same underlying processes | Lasting heart muscle damage from reduced blood flow, often after a heart attack. |
| (4) Cardiomyopathies | I42 - Cardiomyopathy<br>I43 - Cardiomyopathy in diseases classified elsewhere | Grouping of all cardiomyopathies | Heart muscle disorders that weaken the pumping function, unrelated to high blood pressure or blocked vessels. |
| (5) Dysrhythmias | I44 - Atrioventricular and left bundle-branch block<br>I45 - Other conduction disorders<br>I47 - Paroxysmal tachycardia<br>I48 - Atrial fibrillation and flutter<br>I49 - Other cardiac arrhythmias | Grouping of all dysrhythmias | Irregular heart rhythms that are too fast, too slow, or uneven. |
| (6) Heart failure | I50 - Heart failure | Heart failure | Inability of the heart to pump enough blood to meet the body's needs. |

**Table A.7**
**Cohort filters.** Filters applied to MIMIC-IV to derive final cohort size of 28,262 patients.

| Criteria | Count |
|---|---|
| All hospitalised patients | 223,452 |
| Ambiguous ED admissions | -916 |
| No ICU stay | -157,532 |
| No ICU discharge | -11 |
| Died during first ICU stay | -4,716 |
| First ICU stay < 24 h | -12,291 |
| No ED admission | -19,724 |
| Age < 18 | -0 |
| **Cohort size** | **28,262** |

**Table A.8**
**Demographics.** Statistical summary of demographic characteristics in the cohort.

| Feature | Value |
|---|---|
| **Total** n | 28,262 patients |
| Asian | 3% |
| Black | 13% |
| Hispanic | 4% |
| Other | 17% |
| White | 63% |
| **Gender** | |
| Female | 45% |
| Male | 55% |
| **Age** mean (SD) | 65.4 (18.6) years |
| **Length of stay** mean (SD) | 272.8 (284.4) hours |

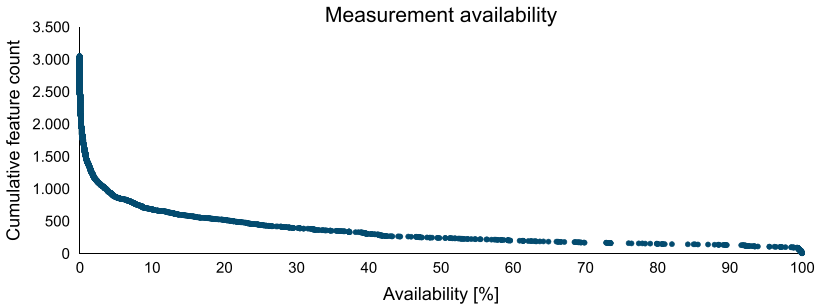



**Fig. A.7. Measurement availability.** Availability of measures in the cohort is shown, with the y-axis representing the cumulative count of measures based on their availability per patient. For example, 228 measures are available for more than 60% of the patients.

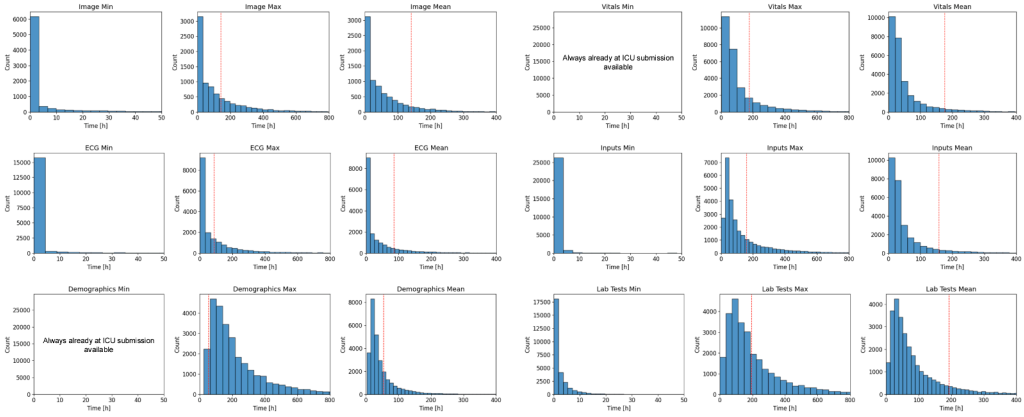



**Fig. A.8. Availability time distribution.** The figure shows for X-rays, ECGs, Demographics, Vital data, Input events and Lab Tests when the measures became available after ICU admission in hours. Per patient, the first availability (column Min), the last (column Max) and the average (column Mean) are shown. In red the overall average time the data became available. 0 means that the data were already available before ICU admission, e.g., from ED or other hospital units.

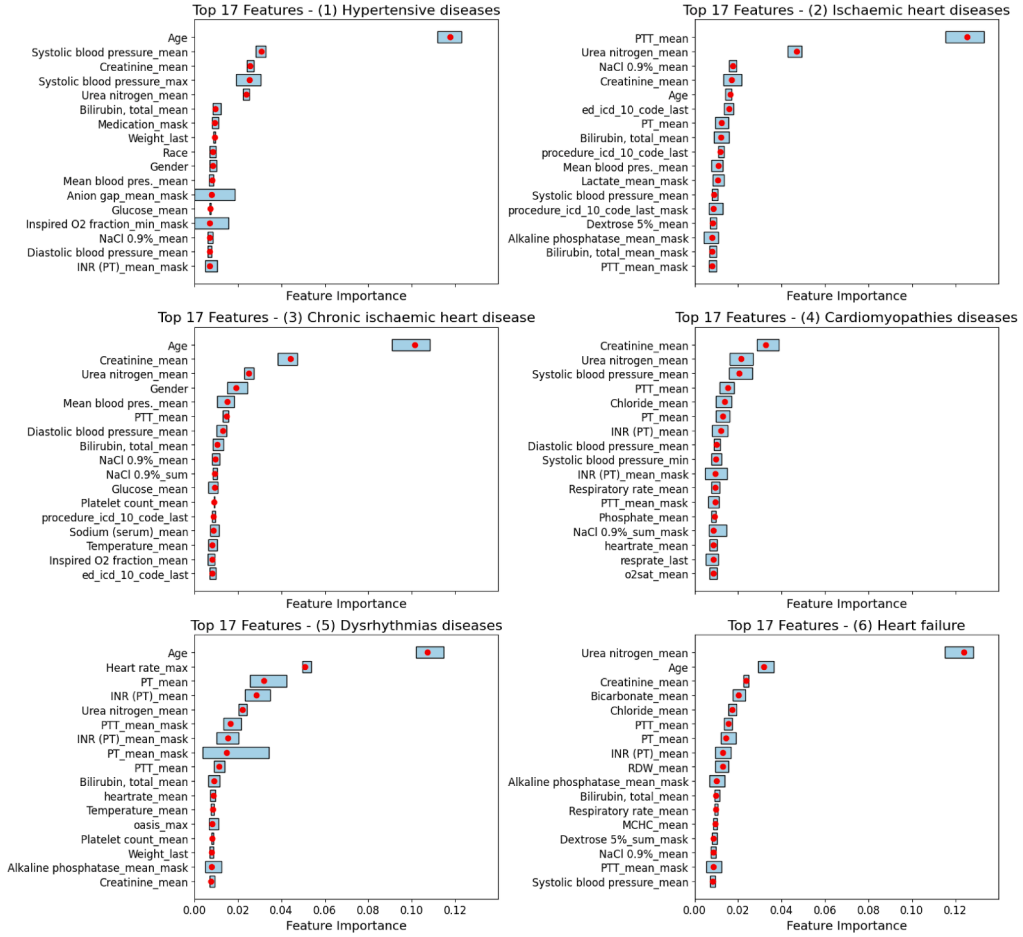



**Fig. B.9. Feature importance of LVD data.** Top features of the LVD XGB models for the various conditions, shown as their contribution to classification performance. Plotted is the range from minimum to maximum importance across folds, ordered by the mean importance (red dots). Each plot relates to one of the six investigated conditions (see Table 2). These features were found to be informative for prediction, but are not necessarily causal factors.

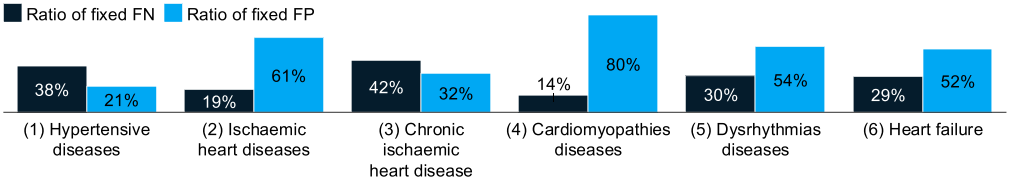



**Fig. B.10. Fusion effect on incorrect single modality predictions.** The figure shows the effect of the fusion model on incorrectly predicted cases of the main single modality model, considering X-rays, ECGs and LVD data. The figure shows how many false positives and negatives the fusion model fixed.

**Table A.9**
**Data selection for MIMIC-IV cohort.** Available features used in modelling. Aggregation refers to the method used for aggregating low-frequency features, where all aggregations were provided to the model as separate features with an option to remove highly correlated features. In bold: Modality type – Sub-type – Category – Location (data source). Availability refers to the availability in the 28k patient cohort. 1/2.

| Description | | | Usage information | |
|---|---|---|---|---|
| Feature | Item ID | Unit | Availability [%] | Aggregation |
| **Tabular - Static patient information (admissions, icustays, patients, medrecon)** | | | | |
| Age | N/A | years | 100 | at admission |
| Gender | N/A | categorical | 100 | N/A |
| Race | N/A | categorical | 80.1 | N/A |
| Prev. medication | N/A | categorical | 100 | N/A |
| **Tabular - Diagnoses (diagnosis)** | | | | |
| ED ICD codes | N/A | categorical | 100 | last |
| **Tabular - Low-frequency - Vital signs - ED (vitalsign, triage)** | | | | |
| Diastolic blood pressure | N/A | mmHg | 37.3 | first, last, mean |
| Heart rate | N/A | beats per min | 37.3 | first, last, mean |
| Order of priority | N/A | score 1–5 | 36 | first, last, mean |
| Oxygen saturation | N/A | percentage | 37.3 | first, last, mean |
| Respiratory rate | N/A | breaths per min | 37.3 | first, last, mean |
| Self-reported pain | N/A | score 0–10 | 37.3 | first, last, mean |
| Systolic blood pressure | N/A | mmHg | 37.3 | first, last, mean |
| Temperature | N/A | degrees F | 37.3 | first, last, mean |
| **Tabular - Low-frequency - Vital signs - ICU (chartevents, oasis, sofa)** | | | | |
| Albumin | 227,456 | g/dl | 42.5 | mean |
| Arterial CO2 pressure | 220,235 | mmHg | 42.3 | mean, max, min |
| Arterial O2 pressure | 220,224 | mmHg | 42.3 | mean, max, min |
| Diastolic blood pressure | 220,180 | mmHg | 99.7 | mean, max, min |
| GCS - eye opening | 220,739 | score 1–4 | 100 | last |
| GCS - motor response | 223,901 | score 1–6 | 100 | last |
| GCS - verbal response | 223,900 | score 1–5 | 100 | last |
| Heart rate | 220,045 | beats per min | 100 | mean, max, min |
| Inspired O2 fraction | 223,835 | percentage | 48.1 | mean, min |
| Mean blood pres. | 220,181 | mmHg | 99.7 | mean, max, min |
| Oxygen saturation | 220,277 | percentage | 100 | mean, min |
| Platelet count | 227,457 | k/ul | 99.4 | mean |
| Respiratory rate | 220,210 | insp/min | 99.9 | mean, max, min |
| Sodium (serum) | 220,645 | meq/l | 99.5 | mean, max, min |
| Systolic blood pressure | 220,179 | mmHg | 99.7 | mean, max, min |
| Temperature | 223,761 | degrees F | 99.8 | mean, max, min |
| Weight | 226,512 | kg | 100 | last |
| OASIS score | N/A | score | 100 | first, last, max |
| SOFA score | N/A | score | 100 | first, last, max |
| **Tabular - Low-frequency - Treatments - ICU (inputevents)** | | | | |
| Dextrose 5% | 220,949 | ml | 64.8 | sum, mean |
| Magnesium sulfate | 222,011 | grams | 45.5 | sum, mean |
| NaCl 0.9% | 225,158 | ml | 79.4 | sum, mean |
| Po intake | 226,452 | ml | 71.8 | sum, mean |
| Solution | 225,943 | ml | 48.8 | sum, mean |
| **Tabular - Low-frequency - Treatments - Hosp (procedures_icd)** | | | | |
| Procedure codes | N/A | categorical | 100 | last |
| **Images (metadata)** | | | | |
| X-rays | N/A | N/A | 24.4 | Last |
| **High-frequency measures (record_list)** | | | | |
| ECGs | N/A | N/A | 48.9 | Last |
| **Text data (triage)** | | | | |
| ED complaints | N/A | N/A | 55.7 | Last |
| Radiology reports | N/A | N/A | 75.4 | Last |

**Table A.10**
**Data selection for MIMIC-IV cohort.** Available features used in modelling. Aggregation refers to the method used for aggregating low-frequency features, where all aggregations were provided to the model as separate features with an option to remove highly correlated features. In bold: Modality type – Sub-type – Category – Location (data source). Availability refers to the availability in the 28k patient cohort. 2/2.

| Description | | | Usage information | |
|---|---|---|---|---|
| Feature | Item ID | Unit | Availability [%] | Aggregation |
| **Tabular - Low-frequency - Lab measures - Hosp (labevents)** | | | | |
| Alanine aminotransferase | 50,861 | iu/l | 69.1 | mean |
| Alkaline phosphatase | 50,863 | iu/l | 68.5 | mean |
| Anion gap | 50,868 | meq/l | 99.7 | mean |
| Bicarbonate | 50,882 | meq/l | 99.7 | mean |
| Bilirubin, total | 50,885 | mg/dl | 68.6 | mean |
| Calcium, total | 50,893 | mg/dl | 99.4 | mean |
| Chloride | 50,902 | meq/l | 99.7 | mean |
| Creatinine | 50,912 | mg/dl | 99.7 | mean |
| Glucose | 50,931 | mg/dl | 99.7 | mean |
| Hematocrit | 51,221 | percentage | 99.7 | mean |
| Hemoglobin | 51,222 | g/dl | 99.6 | mean |
| INR (PT) | 51,237 | N/A | 92.6 | mean |
| Lactate | 50,813 | mmol/l | 64.2 | mean |
| Magnesium | 50,960 | mg/dl | 99.5 | mean |
| MCH | 51,248 | pg | 99.6 | mean |
| MCHC | 51,249 | percentage | 99.6 | mean |
| MCV | 51,250 | fl | 99.7 | mean |
| Phosphate | 50,970 | mg/dl | 99.4 | mean |
| Platelet count | 51,265 | k/ul | 99.7 | mean |
| Potassium | 50,971 | meq/l | 99.7 | mean, max, min |
| PT | 51,274 | sec | 92.6 | mean |
| PTT | 51,275 | sec | 92.2 | mean |
| RDW | 51,277 | percentage | 99.6 | mean |
| Red blood cells | 51,279 | m/ul | 99.7 | mean |
| Sodium | 50,983 | meq/l | 99.7 | mean |
| Urea nitrogen | 51,006 | mg/dl | 99.7 | mean |
| WBC | 51,516 | num/hpf | 47.3 | mean, max, min |
| White blood cells | 51,301 | k/ul | 99.6 | mean |

**Table B.11**
**Comparison of proposed fusion pipeline to state-of-the-art models.** State-of-the-art models from the literature where trained and evaluated on the proposed classification task the 6 different ICD-10 code groups for the circulatory system (Table A.6). Reported is the average (SD) AUC over 5-fold CV. If multiple fusion strategies were proposed, the best reported one was used. For a fair comparison, the fusion dataset of size 5280 patients was used for all models (Table 2). In bold the best performing one.

| Paper/Model | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Baseline (Late, fusion, w/o notes) | 0.757 (0.012) | **0.767** (0.016) | **0.784** (0.012) | 0.732 (0.033) | 0.792 (0.009) | **0.828** (0.007) |
| HAIM [18] | 0.754 (0.034) | 0.700 (0.039) | 0.762 (0.027) | 0.676 (0.093) | **0.808 (0.006)** | 0.796 (0.012) |
| MedFuse [19] | **0.767** (0.015) | 0.760 (0.015) | 0.778 (0.023) | **0.775** (0.051) | 0.757 (0.021) | 0.803 (0.021) |
| AutoPrognosis-M [22] | .755 (0.012) | .729 (0.041) | .783 (0.034) | .719 (0.030) | 0.785 (0.027) | 0.808 (0.016) |

**Table B.12**
**Late fusion model: Performance by subgroup.** Mean (SD) AUC over the five cross-validation folds for the late fusion model (all modalities, average aggregation, no notes), evaluated on the held-out test data of the 5,280-patient fusion cohort. Values are reported for chronic ischaemic heart disease as a worked example. *Max gap* is the maximum AUC gap across all subgroups in the table; *Worst* is the AUC of the worst-performing subgroup.

| **Group** | **AUC** |
|---|---|
| Overall | 0.824 (0.016) |
| Max gap | 0.111 |
| Worst | 0.750 (0.023) |
| Age $< 63$ | 0.861 (0.034) |
| Age $\geq 63$ | 0.750 (0.023) |
| Sex - female | 0.789 (0.010) |
| Sex - male | 0.841 (0.023) |
| Insurance - Medicaid | 0.848 (0.053) |
| Insurance - Medicare | 0.769 (0.024) |
| Insurance - Other | 0.846 (0.123) |
| Insurance - Private | 0.865 (0.020) |

**Table B.13**
**Late fusion model: Threshold-dependent performance by subgroup.** Mean (SD) of accuracy, FNR and F1 over the five cross-validation folds for the late fusion model on chronic ischaemic heart disease, evaluated on the held-out test data. Reported are the values at the default decision threshold of 0.5 (panel *a*) and at an accuracy-optimised global threshold selected on a held-out validation set (panel *b*). *Max gap* is the largest gap across age, sex and insurance subgroups; *Worst* is the worst-performing group.

| Group | Accuracy | FNR | F1 |
|---|---|---|---|
| **a) Default threshold of 0.5** | | | |
| Overall | 0.744 (0.029) | 0.261 (0.042) | 0.713 (0.024) |
| Max gap | 0.166 | 0.111 | 0.073 |
| Worst group | 0.674 (0.030) | 0.378 (0.097) | 0.666 (0.028) |
| Age $< 63$ | 0.838 (0.030) | 0.326 (0.021) | 0.739 (0.034) |
| Age $\geq 63$ | 0.674 (0.030) | 0.288 (0.056) | 0.666 (0.028) |
| Sex - female | 0.735 (0.018) | 0.337 (0.059) | 0.674 (0.018) |
| Sex - male | 0.747 (0.039) | 0.226 (0.046) | 0.730 (0.035) |
| Insurance - Medicaid | 0.812 (0.042) | 0.378 (0.097) | 0.697 (0.052) |
| Insurance - Medicare | 0.690 (0.031) | 0.277 (0.054) | 0.679 (0.027) |
| Insurance - Other | 0.856 (0.068) | 0.306 (0.108) | 0.739 (0.087) |
| Insurance - Private | 0.802 (0.042) | 0.293 (0.063) | 0.733 (0.034) |
| **b) Accuracy-optimised global threshold** | | | |
| Overall | 0.788 (0.012) | 0.540 (0.084) | 0.703 (0.027) |
| Max gap | 0.152 | 0.134 | 0.069 |
| Worst group | 0.716 (0.017) | 0.663 (0.261) | 0.731 (0.037) |
| Age $< 63$ | 0.867 (0.012) | 0.607 (0.073) | 0.706 (0.029) |
| Age $\geq 63$ | 0.716 (0.017) | 0.564 (0.093) | 0.663 (0.034) |
| Sex - female | 0.794 (0.015) | 0.624 (0.070) | 0.662 (0.016) |
| Sex - male | 0.782 (0.020) | 0.490 (0.101) | 0.722 (0.039) |
| Insurance - Medicaid | 0.858 (0.023) | 0.623 (0.060) | 0.686 (0.035) |
| Insurance - Medicare | 0.730 (0.023) | 0.558 (0.103) | 0.672 (0.040) |
| Insurance - Other | 0.882 (0.048) | 0.663 (0.261) | 0.662 (0.144) |
| Insurance - Private | 0.851 (0.013) | 0.551 (0.096) | 0.731 (0.037) |

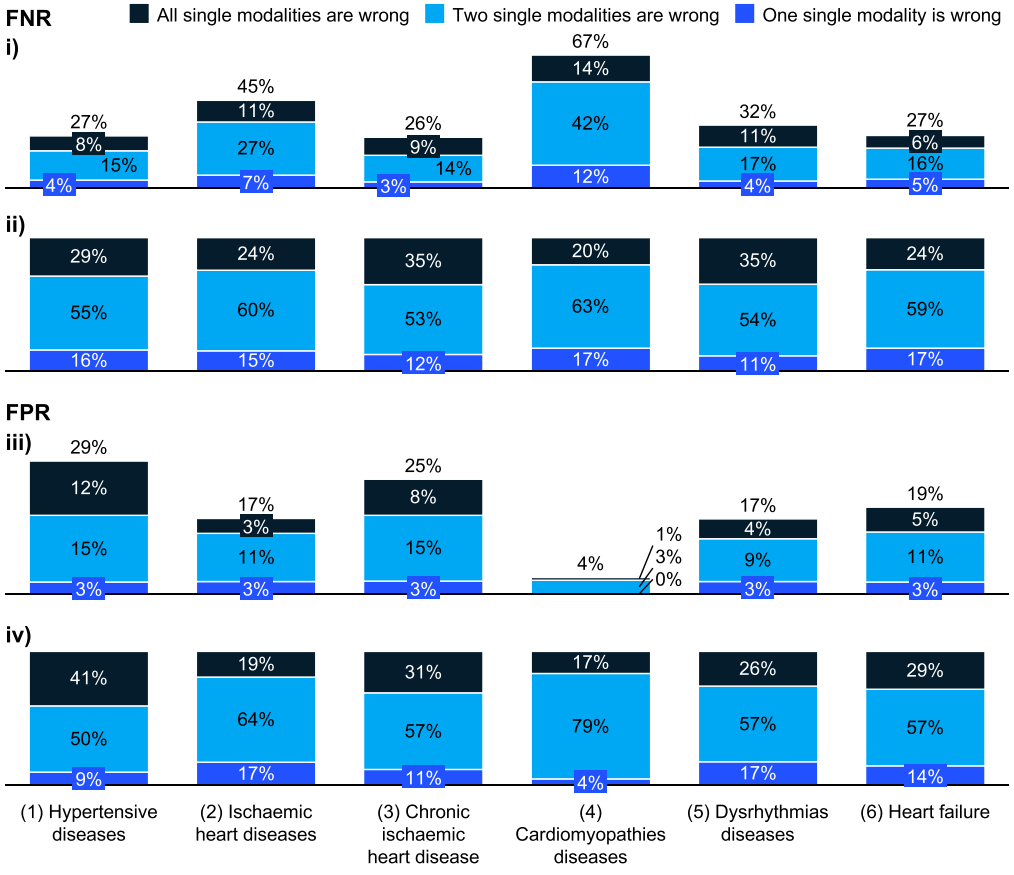



**Fig. B.11. Modality effect on false fusion predictions.** The figure shows the influence of the single modalities on false predictions of the late fusion model, considering X-rays, ECGs and LVD data. i) and ii) show the split of the FNR into whether all single modalities were wrong, only two or only one, in i) scaled to all predictions and in ii) scaled to the false negative ones. iii) and iv) do the same for the FPR. Above the bars in i) and iii) are the FNRs and FPRs shown, respectively.




**Fig. B.12. Selective acquisition of multimodal data.** The figure shows the performance gain in F1-score of the multimodal model compared with the single modality models. The x-axis is divided into ten percent buckets of patients, ranked according to the confidence of the single modality model. For each bucket, the multimodal model was applied instead of the single modality model. For example, the second bucket includes the 10% of patients whose single modality predictions had the lowest 10% to 20% confidence levels. The F1-score of the multimodal model was then calculated for each bucket and compared to the F1-score achieved using only a single modality, such as X-rays. The graph indicates that, as the certainty of single-modality predictions increases (towards the right of the x-axis), the added value of the fusion model diminishes (the graph flattens out). This pattern suggests that the multimodal approach is most advantageous when single-modality predictions are less confident. Each plot relates to one of the six investigated conditions (see Table 2).

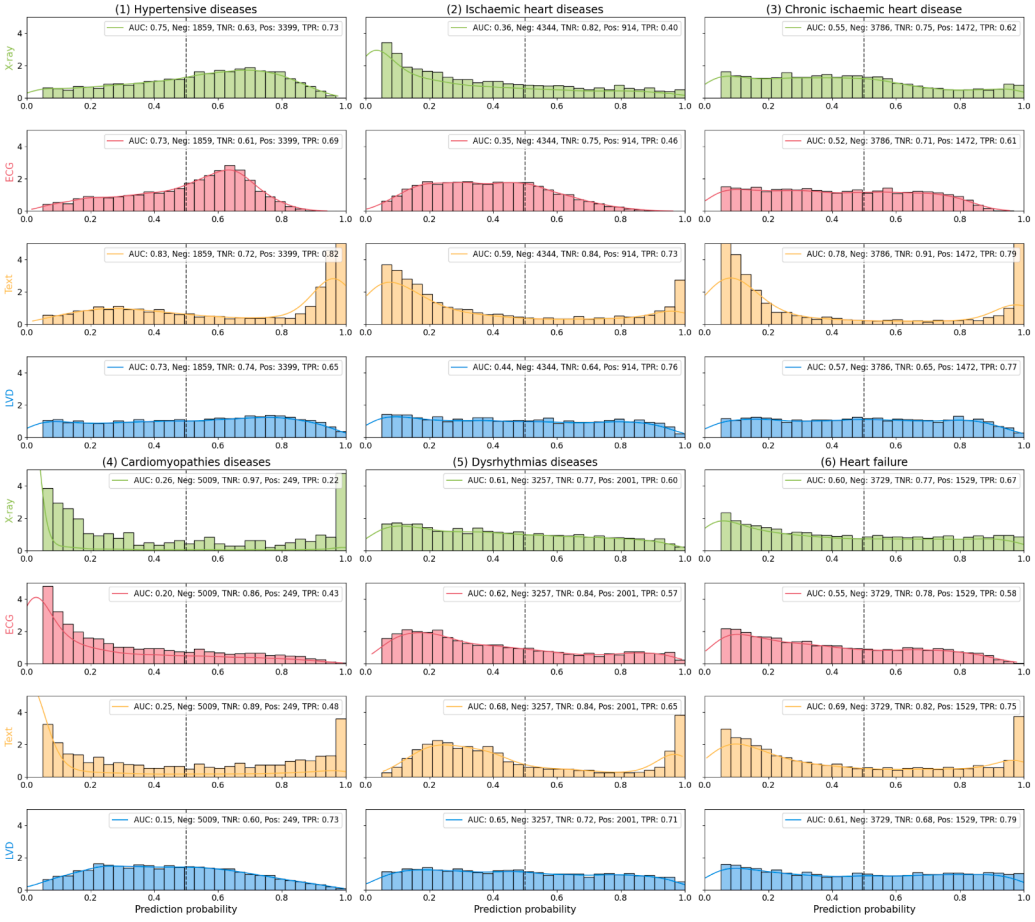



**Fig. B.13. Distribution of prediction probabilities.** Distribution of the prediction probabilities for the different diseases and modalities. In green for X-ray, red for ECG and blue for LVD data.

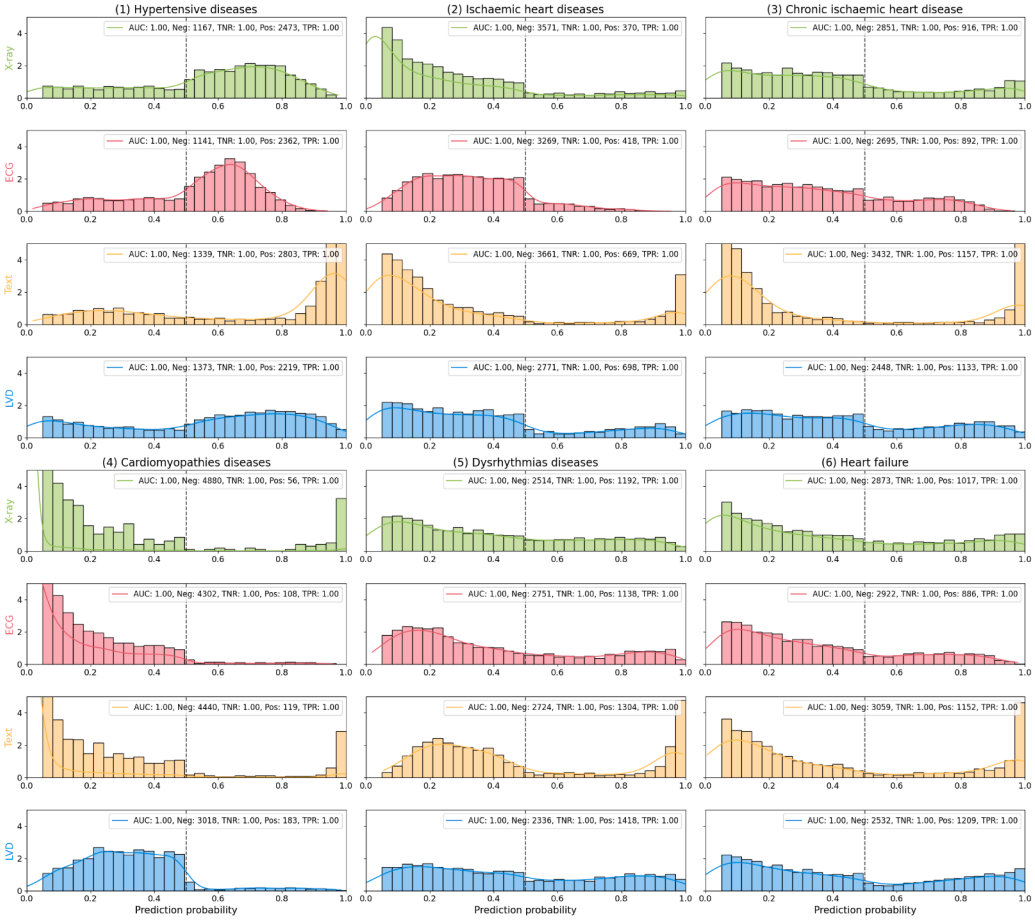



**Fig. B.14. Distribution of correct prediction probabilities.** Distribution of the prediction probabilities for the different diseases and modalities. In green for X-ray, red for ECG and blue for LVD data.

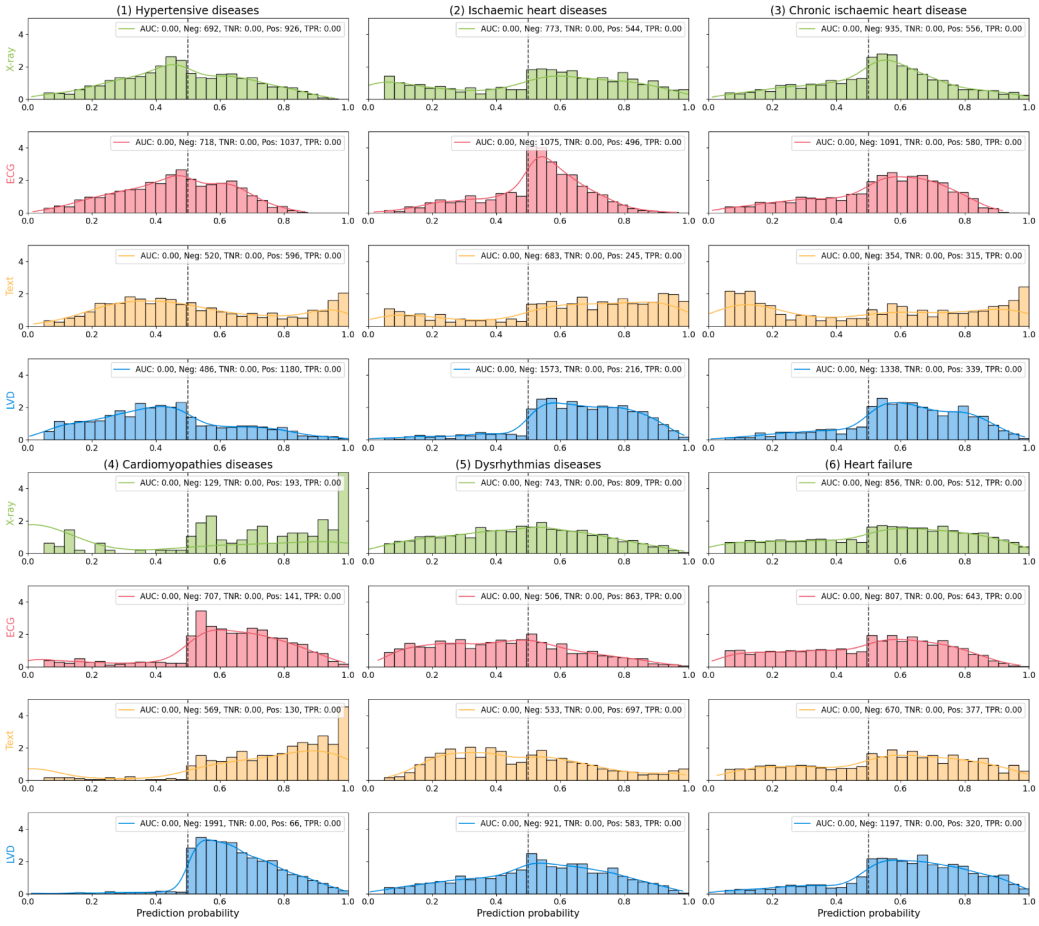



**Fig. B.15. Distribution of incorrect prediction probabilities.** Distribution of the prediction probabilities for the different diseases and modalities. In green for X-ray, red for ECG and blue for LVD data.

## Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.inffus.2026.104573.

## References


[1] World Health Organisation, Cardiovascular diseases (CVDs) — WHO.int, 2021, (https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)). Accessed in May 2023.

[2] V. Raleigh, D. Jefferies, D. Wellings, Cardiovascular disease in England, King's Fund (2022) 2022–11.

[3] D.J. Cook, S. Webb, A. Proudfoot, Assessment and management of cardiovascular disease in the intensive care unit, Heart 108 (5) (2022) 397–405.

[4] S. Amal, L. Safarnejad, J.A. Omiye, I. Ghanzouri, J.H. Cabot, E.G. Ross, Use of multimodal data and machine learning to improve cardiovascular disease care, Front. Cardiovasc. Med. 9 (2022).

[5] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ Digit. Med. 3 (1) (2020) 1–9.

[6] M. Moor, O. Banerjee, Z.S.H. Abad, H.M. Krumholz, J. Leskovec, E.J. Topol, P. Rajpurkar, Foundation models for generalist medical artificial intelligence, Nature 616 (7956) (2023) 259–265.

[7] M.R. Pinsky, A. Bedoya, A. Bihorac, L. Celi, M. Churpek, N.J. Economou-Zavlanos, P. Elbers, S. Saria, V. Liu, P.G. Lyons, et al., Use of artificial intelligence in critical care: opportunities and obstacles, Crit. Care 28 (1) (2024) 113.

[8] P. Renc, Y. Jia, A.E. Samir, J. Was, Q. Li, D.W. Bates, A. Sitek, Zero shot health trajectory prediction using transformer, NPJ Digit. Med. 7 (2024) 256.

[9] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C.K. Reinhart, P. Suter, L.G. Thijs, The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, Intensive Care Med. 22 (7) (1996) 707–710.

[10] J.E. Zimmerman, A.A. Kramer, D.S. McNair, F.M. Malila, V.L. Shaffer, Intensive care unit length of stay: benchmarking based on acute physiology and chronic health evaluation (APACHE) IV, Crit. Care Med. 34 (10) (2006) 2517–2529.

[11] F. Krones, U. Marikkar, G. Parsons, A. Szmul, A. Mahdi, Review of multimodal machine learning approaches in healthcare, Inf. Fusion 114 (2025) 102690.

[12] R.M. Summers, NIH chest X-Ray dataset of 14 common thorax disease categories, 2019,

[13] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D.A. Mong, S.S. Halabi, J.K. Sandberg, R. Jones, D.B. Larson, C.P. Langlotz, B.N. Patel, M.P. Lungren, A.Y. Ng, CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison, Proc. AAAI Conf. Artif. Intell. 33 (01) (2019) 590–597.

[14] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: a large chest X-ray image dataset with multi-label annotated reports, Med. Image Anal. 66 (2020) 101797.

[15] A. Bagheri, T.K.J. Groenhof, W.B. Veldhuis, P.A. de Jong, F.W. Asselbergs, D.L. Oberski, Multimodal learning for cardiovascular risk prediction using EHR data, arXiv:2008.11979 (2020).

[16] P. Li, Y. Hu, Z.-P. Liu, Prediction of cardiovascular diseases by integrating multimodal features with machine learning methods, Biomed. Signal Process. Control 66 (2021) 102474.

[17] A.E.W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T.J. Pollard, S. Hao, B. Moody, B. Gow, et al., MIMIC-IV, a freely accessible electronic health record dataset, Sci. Data 10 (1) (2023) 1.

[18] L.R. Soenksen, Y. Ma, C. Zeng, L.D.J. Boussioux, K.V. Carballo, L. Na, H.M. Wiberg, M.L. Li, I. Fuentes, D. Bertsimas, Integrated multimodal artificial intelligence framework for healthcare applications, arXiv:2202.12998 (2022).

[19] N. Hayat, K.J. Geras, F.E. Shamout, MedFuse: multi-modal fusion with clinical time-series data and chest X-ray images, in: Machine Learning for Healthcare Conference, PMLR, 2022, pp. 479–503.

[20] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, M. Roger, MIMIC-IV (version 3.0), 2024,

[21] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, M.P. Lungren, 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) (2020) 1–9.

[22] F. Imrie, S. Denner, L.S. Brunschwig, K. Maier-Hein, M. van der Schaar, Automated ensemble multimodal machine learning for healthcare, IEEE J. Biomed. Health Inform. (2025).

[23] S.-L. Huang, X. Xu, L. Zheng, An information-theoretic approach to unsupervised feature selection for high-dimensional data, IEEE J. Sel. Areas Inf. Theory 1 (2020) 157–166.

[24] Y. Mei, G. Venkataramani, T. Lan, Exploiting partial common information microstructure for multi-modal brain tumor segmentation, in: Workshop on Machine Learning for Multimodal Healthcare Data, Springer, 2023, pp. 64–85.

[25] R. Vavekanand, T. Kumar, S. Kumar, G. Kumar, A.A. Laghari, Multimodal machine learning approaches in predictive healthcare analytics: a comprehensive survey, Arch. Comput. Methods Eng. (2026) 1–25.

[26] Z. Guo, X. Li, H. Huang, N. Guo, Q. Li, Deep learning-based image segmentation on multimodal medical imaging, IEEE Trans. Radiat. Plasma Med. Sci. 3 (2) (2019) 162–169.

[27] G. Holste, S.C. Partridge, H. Rahbar, D. Biswas, C.I. Lee, A.M. Alessio, End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3294–3303.

[28] D. Thiele, R. Rodseth, R. Friedland, F. Berger, C. Mathew, C. Maslo, V. Moll, C. Leithner, C. Storm, A. Krannich, et al., Machine learning models for the early real-time prediction of deterioration in intensive care units–a novel approach to the early identification of high-risk patients, J. Clin. Med. 14 (2) (2025) 350.

[29] J. Kim, M. Chae, H.-J. Chang, Y.-A. Kim, E. Park, Predicting cardiac arrest and respiratory failure using feasible artificial intelligence with simple trajectories of patient data, J. Clin. Med. 8 (9) (2019) 1336.

[30] D. Placido, H.-C. Thorsen-Meyer, B.S. Kaas-Hansen, R. Reguant, S. Brunak, Development of a dynamic prediction model for unplanned ICU admission and mortality in hospitalized patients, PLOS Digit. Health 2 (6) (2023) e0000116.

[31] G.S. Collins, K.G.M. Moons, P. Dhiman, R.D. Riley, A.L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J.B. Reitsma, M. Van Smeden, et al., TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024).

[32] B. Duvieusart, F. Krones, G. Parsons, L. Tarassenko, B.W. Papież, A. Mahdi, Multimodal cardiomegaly classification with image-derived digital biomarkers, in: Annual Conference on Medical Image Understanding and Analysis, Springer, 2022, pp. 13–27.

[33] A. Kansal, E. Chen, T. Jin, P. Rajpurkar, D. Kim, Multimodal clinical monitoring in the emergency department (MC-MED), PhysioNet 1.0.0 (2025).

[34] M.W. Weiner, D.P. Veitch, P.S. Aisen, L.A. Beckett, N.J. Cairns, R.C. Green, D. Harvey, C.R. Jack, Jr, W. Jagust, J.C. Morris, et al., The Alzheimer's disease neuroimaging initiative 3: continued innovation for clinical trial improvement, Alzheimer's Dement. 13 (5) (2017) 561–571.

[35] J.M. Lopez Alcaraz, N. Strodthoff, MDS-ED: Multimodal decision support in the emergency department–a benchmark dataset for diagnoses and deterioration prediction in emergency medicine, arXiv:2407.17856 (2024).

[36] A. Saporta, A.M. Puli, M. Goldstein, R. Ranganath, Symile-MIMIC: a multimodal clinical dataset of chest X-rays, electrocardiograms, and blood labs from MIMIC-IV, 2025, (PhysioNet, version 1.0.0).

[37] A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, S. Horng, MIMIC-CXR-JPG-chest radiographs with structured labels, PhysioNet (2019).

[38] B. Gow, T. Pollard, L.A. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. Greenbaum, J.W. Waks, P. Eslami, T. Carbonati, et al., MIMIC-IV-ECG: diagnostic electrocardiogram matched subset (version 1.0), PhysioNet (2023).

[39] NBER, ICD-9-CM to and from ICD-10-CM and ICD-10-PCS crosswalk or general equivalence mappings — nber.org, 2024, (https://www.nber.org/research/data). Accessed in Oct 2024.

[40] K. Lewandowska, M. Weisbrot, A. Cieloszyk, W. Medrzycka-Dabrowska, S. Krupa, D. Ozga, Impact of alarm fatigue on the work of nurses in an intensive care environment–a systematic review, Int. J. Environ. Res. Public Health 17 (22) (2020) 8409.

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778.

[42] D. Ma, M.R. Hosseinzadeh Taher, J. Pang, N.U.I. Islam, F. Haghighi, M.B. Gotway, J. Liang, Benchmarking and boosting transformers for medical image classification, in: Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings, Springer, 2022, pp. 12–22.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[44] F. Krones, B. Walker, G. Parsons, T. Lyons, A. Mahdi, Multimodal deep learning approach to predicting neurological recovery from coma after cardiac arrest, in: 2023 Computing in Cardiology (CinC), 50, 2023, pp. 1–4.

[45] G. Wang, X. Liu, Z. Ying, G. Yang, Z. Chen, Z. Liu, M. Zhang, H. Yan, Y. Lu, Y. Gao, et al., Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial, Nat. Med. 29 (10) (2023) 2633–2642.

[46] E. Sheetrit, M. Brief, O. Elisha, Predicting unplanned readmissions in the intensive care unit: a multimodality evaluation, Sci. Rep. 13 (1) (2023) 15426.

[47] OpenAI, GPT-4o system card, 2024, (https://cdn.openai.com/gpt-4o-system-card.pdf).

[48] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, 2017,

[49] D. Tran, J. Liu, M.W. Dusenberry, D. Phan, M. Collier, J. Ren, K. Han, Z. Wang, Z. Mariet, H. Hu, et al., PLEX: Towards reliability using pretrained large model extensions, arXiv:2207.07411 (2022).

[50] M.N. Lyngbakken, P.L. Myhre, H. Røsjø, T. Omland, Novel biomarkers of cardiovascular disease: applications in clinical practice, Crit. Rev. Clin. Lab. Sci. 56 (1) (2019) 33–60.

[51] N. Strodthoff, J.M. Lopez Alcaraz, W. Haverkamp, Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: an explorative study in emergency care, Eur. Heart J.-Digit. Health 5 (4) (2024) 454–460.

[52] K.L. Dragan, S.M. Desai, J. Billings, S.A. Glied, Association of insurance mix and diagnostic coding practices in New York State hospitals, in: JAMA Health Forum, 3, American Medical Association, 2022, pp. e222919–e222919.

[53] B. Duvieusart, F. Krones, G. Parsons, L. Tarassenko, B.W. Papiez, A. Mahdi, Heart and lung segmentations for MIMIC-CXR/MIMIC-CXR-JPG and montgomery county TB databases, PhysioNet (2023).