RespireTransFuse: A Bidirectional Cross-Attention Fusion for Respiratory Deterioration Prediction

by

Reza Arab Neidi

A dissertation submitted to the
Department of Computer Science
in conformity with the requirements for
the degree of Master of Science

Bishop's University
Canada
November 2026

Copyright © Reza Arab Neidi, 2026
released under a CC BY-SA 4.0 License

Abstract

Respiratory deterioration in critically ill patients may progress to acute respiratory failure and require invasive or non-invasive ventilatory support, making early identification important for monitoring and clinical preparation [8, 30]. Prediction is challenging because electronic health records (EHRs) describe evolving physiological conditions, while chest radiographs (CXRs) provide complementary spatial evidence of pulmonary abnormalities. Effectively integrating these sources may improve the prediction of future respiratory-support requirements. This thesis develops RespireTransFuse, a multimodal framework that combines a Transformer-based EHR encoder with an EfficientNet-B0 image encoder through reciprocal cross-attention. Image tokens attend to temporal clinical representations, and EHR tokens attend to spatial radiographic features; their cross-conditioned outputs are combined with the original modality summaries. Using paired MIMIC-IV and MIMIC-CXR data [35, 36], each prediction uses an index ICU radiograph and 24 hours of preceding EHR measurements to predict recorded respiratory-support procedures within the following 48 hours. The final cohort contains 9,509 samples from 4,668 patients. Seven image-only, EHR-only, and multimodal models, including adapted MedFuse baselines [22], were evaluated across three random seeds using patient-exclusive splits. On the held-out test set of 1,429 samples (17.42% positive), RespireTransFuse achieved the highest mean area under the receiver operating characteristic curve (AUROC, 0.7134) and area under the precision–recall curve (AUPRC, 0.4071). Compared with Early Fusion, these represent improvements of 0.0165 AUROC and 0.0502 AUPRC. The full architecture also obtained the highest observed discrimination in the ablation study. After validation-based recalibration, its mean Brier score was 0.1282 and expected calibration error was 0.0220. The findings indicate that preceding EHR measurements provide the stronger standalone predictive signal, while reciprocal spatial–temporal fusion achieves the best overall discrimination among the evaluated architectures.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Madjid Allili, for his invaluable guidance, insightful feedback, and unwavering support throughout the course of my research. His expertise and encouragement have been essential to both the development of this thesis and my growth as a researcher. I am also sincerely thankful to Professor Mohammed Ayoub Alaoui Mhamdi for his valuable advice and constructive feedback, which have greatly enriched my academic journey.

My heartfelt appreciation goes to my family for their constant encouragement, love, and understanding. In particular, I wish to thank my brother Jalal, whose unwavering belief in me has been a source of strength and motivation during challenging moments. Their support has been the foundation that allowed me to pursue my studies with determination.

Lastly, I extend my thanks to Bishop's University, especially the Department of Computer Science, for providing an inspiring academic environment and the opportunity to explore a field I am truly passionate about.

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Problem Statement . . . 2
- 1.3 Objectives . . . 2
- 1.4 Contributions . . . 3
- 1.5 Thesis Organization . . . 4

2 Literature Review **5**
- 2.1 Respiratory Deterioration as a Prediction Problem . . . 5
- 2.2 Data Resources and Modality Characteristics . . . 6
- 2.3 Longitudinal EHR Modeling . . . 7
- 2.4 Deep Learning for Chest Radiographs . . . 8
- 2.5 Multimodal Learning and Fusion . . . 9
- 2.6 Transformers and Cross-Modal Attention . . . 10
- 2.7 Related Work in Multimodal Clinical Prediction . . . 12
- 2.8 Synthesis and Research Gap . . . 13

3 Methodology **15**
- 3.1 Study Design and Experimental Overview . . . 15
- 3.2 Data Sources, Identifiers, and Temporal Linkage . . . 16
- 3.3 Cohort Construction and Outcome Definition . . . 18
- 3.4 Multimodal Preprocessing Pipeline . . . 20
- 3.5 Notation and Shared Input Interface . . . 25
- 3.6 Adopted Baseline and Proposed Models . . . 25
- 3.7 Objective Functions . . . 43
- 3.8 Optimization and Training Protocol . . . 44
- 3.9 Evaluation Measures . . . 45
- 3.10 Ablation Analysis . . . 47
- 3.11 Implementation and Reproducibility Controls . . . 48
- 3.12 Methodological Scope and Limitations . . . 48

4 Experimental Results **49**
4.1 Cohort Distribution 49
4.2 Overall Performance and Uncertainty 50
4.3 Model-Specific Training Profiles 52
4.4 Comparative Interpretation 61

5 Conclusion and Future Work **69**
5.1 Conclusion 69
5.2 Future Work 70

Bibliography **71**

List of Tables

2.1 Representative image–EHR studies and the remaining task-specific gap. 14

3.1 26-variable EHR registry and fixed tensor order. 23
3.2 Models evaluated under the common cohort, label, and partitions. . 26
3.3 Stage-level architecture of the `tf_efficientnet_b0.ns_jft_in1k` feature extractor used in the image-only model. 29
3.4 Principal training settings for the seven evaluated models. 45

4.1 Patient-exclusive cohort and positive-event prevalence. 49
4.2 Three-seed discrimination on the held-out test set. 52
4.3 Differences in three-seed mean AUROC and AUPRC between Respire-TransFuse and the six comparator models. 53
4.4 Total and trainable parameter counts for the seven evaluated models using the 26-variable EHR configuration. 63
4.5 Optimization settings for the seven evaluated models. 65
4.6 Matched fusion ablations (A0–A4). 66
4.7 Paired ablation differences from the full A4 model. 66

List of Figures

3.1 Representative indexed chest radiographs from the paired cohort. . . 17
3.2 Temporal definition of model inputs and outcome labels. 18
3.3 Overview of cohort construction and EHR preprocessing. 21
3.4 Illustrative 24×26 EHR tensor for a positive test sample over the 24 hours preceding the index chest radiograph. 22
3.5 MedFuse architecture used as the basis of the three external baseline configurations. 28
3.6 Overview of the image-only architecture. 31
3.7 Overview of the EHR-only architecture. 33
3.8 Reference Transformer encoder–decoder architecture. 34
3.9 Detailed Transformer architecture inside the EHR-only model. 37
3.10 Early Fusion architecture. 39
3.11 Overview of the current RespireTransFuse architecture. 41

4.1 Patient-level training, validation, and test distributions in the final cohort. 50
4.2 Class composition and positive-event timing in the final cohort. . . . 51
4.3 Image-Only CNN: three-seed mean training and validation loss, AUROC, and AUPRC across 20 epochs, with standard-deviation bands. . 54
4.4 Adapted MedFuse Uni-CXR: three-seed mean training and validation loss, AUROC, and AUPRC across 20 epochs, with standard-deviation bands. 55
4.5 EHR-Only Transformer: three-seed mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands. 56
4.6 Adapted MedFuse Uni-EHR: three-seed mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands. 57
4.7 Early Fusion: three-seed mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands, alongside the three-seed mean calibration curve and its variability. 58

4.8 Adapted MedFuse Multimodal LSTM: three-seed mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands, alongside the three-seed mean calibration curve and its variability. 59
4.9 RespireTransFuse: three-run mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands, alongside the mean calibration curve and its variability. 60
4.10 Precision–recall curves for the seven models on the held-out test set (seed 40). The dashed line indicates the positive-event prevalence (17.42%). 61
4.11 Receiver operating characteristic curves for the seven models on the held-out test set (seed 40). 62
4.12 Representative held-out test predictions (seed 40): TP, TN, FP, and FN examples. 67

Chapter 1

Introduction

1.1 Context

Respiratory deterioration is a major concern in intensive care units (ICUs), where worsening pulmonary function may lead to acute respiratory failure and require invasive or non-invasive ventilatory support [3, 8]. Early prediction may provide additional time for clinical reassessment, monitoring, and preparation. However, respiratory deterioration is difficult to anticipate because it involves complex physiological changes that may not be adequately characterized by individual clinical indicators [30, 64].

EHRs contain longitudinal measurements, including vital signs, oxygen requirements, blood gases, and laboratory results, which provide information about evolving physiological conditions. Nevertheless, these measurements are frequently irregular and incomplete, requiring models capable of representing temporal dependencies and missing observations [1, 11]. CXRs, in contrast, provide spatial information about pulmonary abnormalities that may not be fully reflected in structured clinical measurements. These modalities therefore offer potentially complementary perspectives on respiratory condition.

Recent multimodal learning approaches have explored the integration of EHR and CXR information for clinical prediction [22, 33, 38]. However, the effectiveness of multimodal learning depends on how information is represented and combined. Conventional concatenation integrates global modality summaries but does not explicitly model relationships between individual image regions and temporal observations. Cross-attention offers a mechanism for learning such interactions, although its benefits must be established experimentally.

This thesis introduces *RespireTransFuse*, a multimodal framework that integrates an EfficientNet-based image encoder with a temporal EHR Transformer through reciprocal cross-attention. Using paired MIMIC-IV and MIMIC-CXR data [35, 36], the study investigates whether spatial–temporal interactions improve prediction of future respiratory-support procedures compared with unimodal and alternative

multimodal approaches.

1.2 Problem Statement

Although multimodal clinical prediction has advanced considerably, several challenges remain. First, many existing EHR–CXR approaches emphasize diagnosis, mortality, or disease classification rather than temporally defined respiratory-support prediction [22, 33]. Developing a prospective prediction task from retrospective records requires careful temporal alignment, outcome definition, and prevention of information leakage.

Second, EHR measurements and radiographs have fundamentally different structures. EHRs represent physiological changes over time, whereas CXRs contain spatial pulmonary information. Combining only their global representations may obscure relationships between clinical observations and radiographic regions. More expressive fusion mechanisms may capture these interactions but introduce additional complexity.

Third, the presence of multiple modalities does not guarantee improved prediction. A stronger modality may dominate the outcome, and architectural improvements must be distinguished from the predictive strength of the underlying encoders. Consequently, meaningful evaluation requires substantive unimodal baselines, alternative fusion approaches, and controlled ablation experiments.

Finally, respiratory-support events are relatively infrequent, making accuracy alone insufficient for evaluating predictive performance. Discrimination, probability calibration, training variability, and statistical uncertainty must be considered separately [61, 69].

1.3 Objectives

The primary objective of this thesis is to develop and systematically evaluate RespireTransFuse for predicting recorded respiratory-support procedures using paired EHR sequences and chest radiographs.

The specific objectives are:

1. **Construct a temporally aligned prediction framework:** Establish a paired MIMIC-IV and MIMIC-CXR cohort using a 24-hour EHR observation window, an index radiograph, and a 48-hour respiratory-support prediction horizon, with explicit eligibility rules and patient-exclusive data partitions.

2. **Develop and compare predictive architectures:** Implement substantive image-only and EHR-only models, conventional fusion approaches, and RespireTransFuse to evaluate the relative contributions of individual modalities and multimodal integration.

3. **Investigate fusion mechanisms:** Examine whether reciprocal cross-attention provides benefits beyond summary-level concatenation and single-direction attention, including the contribution of auxiliary branch supervision.

4. **Evaluate performance and reliability:** Assess discrimination, calibration, statistical uncertainty, and training variability through multiple random seeds, patient-level analyses, controlled ablations, and qualitative examination of model predictions.

1.4 Contributions

The principal contributions of this thesis are summarized as follows:

1. **A temporally controlled multimodal prediction task.**

   A paired cohort of 9,509 observations from 4,668 patients was constructed using MIMIC-IV and MIMIC-CXR. Each prediction combines an index radiograph with 24 hours of preceding EHR information to predict recorded respiratory-support procedures within 48 hours. The cohort incorporates explicit event eligibility, a 72-hour observation requirement for negative examples, and patient-level partitioning to reduce information leakage.

2. **A compact reciprocal cross-attention framework.**

   RespireTransFuse integrates an EfficientNet-B0 image encoder and a missingness-aware temporal EHR Transformer. Separately parameterized cross-attention modules allow image tokens to attend to EHR observations and EHR tokens to attend to image regions. The resulting cross-conditioned representations are combined with the original modality summaries, preserving independently learned information alongside multimodal context. Auxiliary losses additionally supervise the unimodal branches.

3. **A comprehensive comparison of prediction architectures.**

   Seven models were evaluated using common data partitions and outcome definitions: Image-Only CNN, EHR-Only Transformer, Early Fusion, adapted MedFuse Uni-CXR, adapted MedFuse Uni-EHR, adapted MedFuse Multimodal LSTM, and RespireTransFuse. This comparison investigates unimodal predictive strength and the relative effectiveness of concatenation-based, recurrent, and attention-based fusion.

4. **Controlled architectural ablation experiments.**

   Five configurations were investigated to isolate the contributions of the fusion mechanism: concatenation without cross-attention, image-to-EHR attention, EHR-to-image attention, bidirectional attention without auxiliary supervision, and the complete model. These experiments examine attention

directionality and branch supervision rather than attributing performance differences solely to architectural complexity.

5. **An expanded evaluation of predictive reliability.**

   Beyond AUROC and AUPRC, the experimental analysis considers multiple random seeds, patient-cluster bootstrap confidence intervals, paired statistical comparisons, probability calibration, and threshold-dependent performance. Qualitative examinations of radiographic attribution maps, EHR trajectories, and representative prediction errors provide additional insight into model behavior and limitations.

Across the seven evaluated models, RespireTransFuse achieved the highest observed mean AUROC and AUPRC in the internal test evaluation. These findings support the potential of reciprocal spatial–temporal fusion while emphasizing that numerical advantages, probability calibration, and statistical evidence must be interpreted separately. External and prospective validation remain necessary before clinical application.

1.5 Thesis Organization

The remainder of this thesis is organized into four chapters. Chapter 2 reviews the literature on respiratory deterioration prediction, temporal EHR modeling, chest-radiograph analysis, multimodal learning, and attention-based fusion, identifying the research gaps addressed in this study. Building on this foundation, Chapter 3 presents the methodology, covering cohort construction, temporal alignment, data preprocessing, model architectures, training procedures, evaluation metrics, and ablation configurations. Chapter 4 presents and discusses the experimental results, including comparisons among the seven evaluated models, discrimination and calibration performance, statistical uncertainty, architectural ablations, and representative predictions. Finally, Chapter 5 summarizes the main findings and contributions, discusses methodological limitations, and outlines directions for future research.

Chapter 2

Literature Review

This chapter synthesizes the clinical and technical literature supporting multimodal prediction of respiratory deterioration from EHR time series and chest radiographs. It reviews the prediction problem, the two modalities, fusion strategies, attention-based interaction, and the closest EHR–CXR studies. The emphasis is on what prior evidence establishes, where findings conflict, and which question remains unresolved.

2.1 Respiratory Deterioration as a Prediction Problem

Respiratory deterioration is a trajectory rather than a single diagnosis. It may progress through worsening gas exchange, increasing oxygen requirements, ventilatory support, and acute respiratory failure. The Berlin definition formalizes acute respiratory distress syndrome using timing, bilateral opacities, exclusion of hydrostatic oedema as the principal cause, and hypoxaemia severity [3]. Bellani et al. [8] showed that this severe syndrome remains common, under-recognized, and associated with substantial mortality. These findings motivate early warning but also require each prediction study to define its future event, index time, and horizon precisely.

Temporal context is essential because one abnormal measurement may be transient, while apparently acceptable oxygen saturation can coexist with worsening respiratory rate, inspired oxygen, blood gases, or monitoring intensity. Hüser et al. [30] reported that multivariable time-series information could anticipate respiratory failure earlier than a conventional oxygenation baseline. Siu et al. [64] likewise found predictive information for intubation in routinely collected ICU variables. These studies establish structured physiology as a meaningful unimodal reference.

Task definition remains a major boundary between studies. Jabbour et al. [33] combined chest radiographs and EHR variables to distinguish causes of acute respiratory failure; their target was diagnostic. Lu and Nemati [47] instead examined prospective invasive-mechanical-ventilation prediction and reported benefit from

adding CXR representations to EHR data. Similar modalities therefore do not make tasks equivalent when endpoints, index times, and clinical states differ. Multimodal respiratory prediction is plausible because some mechanisms are expressed spatially and others physiologically, but improvement can be limited by stale imaging, sparse EHR measurements, redundancy, or weak temporal alignment [1, 78].

Respiratory deterioration is etiologically heterogeneous; Pneumonia, pulmonary oedema, aspiration, atelectasis, pleural disorders, systemic infection, and impaired neurological ventilatory control can produce comparable escalation events through distinct pathophysiological mechanisms [40]. Chest radiographs primarily depict structural and spatial thoracic abnormalities and therefore provide limited direct information about recent changes in oxygen requirement, ventilation, haemodynamic status, or laboratory measurements. Conversely, longitudinal EHR data can characterize evolving physiological instability without necessarily localizing its thoracic cause [29, 48]. This informational asymmetry provides a rationale for multimodal fusion, and studies integrating chest radiographs with clinical variables have reported improvements over unimodal models [38, 48].

2.2 Data Resources and Modality Characteristics

MIMIC-IV contains de-identified hospital and ICU data linked through patient, admission, and ICU-stay identifiers, providing temporally ordered laboratory, procedure, and bedside records for longitudinal prediction [35]. Harutyunyan et al. [21] established reproducible MIMIC-III benchmarks for mortality, decompensation, length of stay, and phenotyping, helping formalize patient-level partitioning and temporal task construction. However, these records arise from routine care; irregular observations, selective test ordering, and documentation frequency can encode clinical behaviour as well as physiological state [2, 63].

MIMIC-CXR contains 377,110 radiographs from 227,835 imaging studies with associated reports [36] and provides processed images and report-derived labels [37]. Shared patient identifiers and timestamps permit linkage with MIMIC-IV, but requiring an eligible radiograph reduces the available ICU cohort and introduces selection through modality availability [22].

Each radiograph captures both thoracic appearance and aspects of its acquisition process. Interpretation can be affected by clinical indication, acquisition time, projection, patient positioning, support devices, and image quality [10]. Temporal asynchrony creates an additional challenge; the latest CXR may no longer represent the physiological state recorded near the prediction index [76]. Moreover, acquisition and institutional characteristics can provide shortcut signals that impair generalization [18, 58, 60, 79]. Linked EHR–CXR studies must therefore distinguish information contained in image pixels from information introduced by clinical workflow, acquisition timing, eligibility criteria, and pairing decisions.

Hayat et al. [22] emphasized that requiring complete EHR–CXR pairing can

sharply reduce the cohort. Yao et al. [77] treated missing and inconsistent modalities as central properties of clinical fusion, while Yin et al. [78] found that multimodal gains depend strongly on modality completeness. The linked MIMIC ecosystem is therefore valuable for controlled comparison, but its retrospective, single-centre acquisition process limits claims of transportability [1, 35].

2.3 Longitudinal EHR Modeling

2.3.1 Recurrent, Convolutional, and Missingness-Aware Models

Conventional statistical and tree-based models remain essential baselines because carefully engineered temporal summaries such as the most recent value, minimum, maximum, mean, variability, and slope can retain substantial predictive signal with limited complexity. In critical-care applications, flexible regression, random forests, and gradient-boosted trees have achieved strong deterioration prediction and may match or outperform recurrent models [15, 31]. More broadly, a systematic review found no consistent performance advantage of machine-learning methods over logistic regression in clinical prediction studies [14]. The principal limitation of summary-based representations is therefore not an inability to predict, but the loss of temporal structure: patients with identical aggregate values may differ in the timing, volatility, and direction of physiological change. Temporal neural models should consequently be evaluated against competent summarized baselines rather than presumed superior because of architectural complexity [56].

Recurrent neural networks address part of this limitation by updating a hidden state over ordered observations and became standard comparators in ICU time-series benchmarks [21, 56]. RETAIN introduced reverse-time attention, allowing visits and variables to contribute unequally to a prediction [13]. Nevertheless, recurrent computation remains sequential, and reducing a long clinical history to a terminal hidden representation can compress information from clinically distinct periods.

Temporal convolution learns local patterns in parallel. Rocheteau et al. [59] demonstrated efficient temporal pointwise convolution for ICU prediction, supporting the combination of local operations with a broader sequence model. Missingness must also be represented explicitly. Che et al. [11] introduced GRU-D, which combines values, observation masks, and elapsed time and learns decay as measurements become older. Baytas et al. [7] modified LSTM memory for irregular intervals. Together, these approaches show that values, measurement availability, and time gaps encode different aspects of clinical state.

Informative observation creates a methodological caution. Frequent measurement or a newly ordered radiograph may indicate that clinicians already suspect instability. Such signals can improve retrospective prediction while weakening

transportability or reducing the apparent lead over clinical judgement. Transparent reporting should therefore distinguish physiologic information from acquisition and documentation processes where possible [1, 16].

2.3.2 Transformers for Clinical Sequences

Vaswani et al. [70] replaced recurrent state updates with self-attention, allowing every token to interact directly with other positions. Li et al. [42] adapted this principle to longitudinal EHR histories in BEHRT using clinical tokens with age and positional information. ICU numerical time series differ from coded visit histories, yet the core requirements remain: positional information preserves order, masks exclude unavailable observations, and attention represents dependencies beyond adjacent time steps.

Transformers are not universally superior. They add capacity, require regularization, and can be more sensitive than simpler models to cohort size, preprocessing, and missing-data handling. The literature consequently supports attention as one strong sequence mechanism rather than a replacement for recurrent and convolutional approaches [11, 21, 42, 59]. Its particular value for multimodal work is that the same query–key–value [70] formalism extends from relations within the EHR to relations between EHR and image tokens.

2.4 Deep Learning for Chest Radiographs

Large CXR datasets enabled general radiographic representation learning. Wang et al. [72] introduced ChestX-ray8 for multi-label classification and localization, and Irvin et al. [32] developed CheXpert with explicit treatment of uncertain report-derived labels. MIMIC-CXR adds scale and linkage to the surrounding hospital record [36]. These datasets reduce dependence on small disease-specific collections but inherit label uncertainty and acquisition bias.

Convolutional networks remain common because local receptive fields suit spatial patterns. Huang et al. [28] developed DenseNet to encourage feature reuse, and Rajpurkar et al. [57] helped establish DenseNet transfer learning as a CXR baseline. Tan and Le [66] introduced compound scaling in EfficientNet, providing parameter-efficient visual encoders. Transformer-based vision models instead treat images as tokens, although Shamshad et al. [62] noted their dependence on scale and pretraining in medical imaging.

A radiograph supplies spatial evidence at acquisition time but not the recent trajectory of oxygen support, blood gases, haemodynamics, or treatment. Conversely, it may reveal pulmonary abnormalities absent from structured variables. This complementarity supports fusion [33, 38], yet an abnormal image is not equivalent to future deterioration. Zech et al. [79] showed that CXR classifiers can exploit institution-specific prevalence and acquisition patterns and may generalize poorly.

Transfer learning therefore reduces label requirements without making image representations clinically neutral; multimodal claims require strong image-only and EHR-only comparisons.

2.5 Multimodal Learning and Fusion

2.5.1 Representation and Fusion Taxonomy

Baltrušaitis et al. [6] organize multimodal learning around representation, translation, alignment, fusion, and co-learning. Biomedical reviews add concerns that are especially important in healthcare such as heterogeneous inputs, limited paired observations, modality absence, and optimization dominated by the strongest branch [1, 17, 65]. Image–EHR models therefore require modality-appropriate encoders and an explicit decision about where interaction occurs.

Early fusion combines raw or low-level inputs, but direct raw fusion is rarely natural for a two-dimensional image and an irregular clinical sequence. Intermediate fusion combines learned features and ranges from summary concatenation to fine-grained interaction. Late fusion combines branch predictions and is modular but cannot learn feature-level correspondence. Ngiam et al. [49] showed that multimodal representation learning can exploit complementary and redundant evidence, while Polsterl et al. [55] demonstrated that tabular variables can modulate medical image features beyond simple concatenation.

Fusion location determines both the representational capacity of a multimodal model and the comparison being performed [1, 6, 65]. Summary-level feature fusion retains modality-specific encoders and combines their pooled representations, thereby testing whether global image and EHR information is jointly predictive [17]. Feature-modulation methods introduce a stronger interaction by allowing one modality to rescale or shift the intermediate features of another [55]. Token-level fusion delays global pooling and enables spatial image regions and temporal EHR periods to interact directly, for example through cross-attention [38, 70]. In contrast, decision-level fusion combines independently generated branch predictions. This preserves clearer branch boundaries and can be configured to operate when a modality is unavailable, although robustness to missing inputs must be explicitly designed rather than assumed [22, 77].

No fusion location is universally optimal. Fine-grained interaction offers greater capacity to learn localized cross-modal relationships, but it also requires reliable alignment and may learn spurious correspondence when paired data are limited or noisy. Later fusion restricts such interactions but generally provides lower complexity, clearer modality-specific ablations, and greater flexibility when modalities are incomplete [6, 65, 71].

Concatenation is an important baseline because it tests whether globally encoded modalities add information without a large interaction module. More expressive fusion is justified only if it improves upon both unimodal branches and simple combination. Wang et al. [71] showed that multimodal optimization can overfit at different rates across modalities, so an apparently sophisticated model may simply rely on the strongest encoder. Reviews similarly caution that parameter count, pairing policy, and baseline quality can be mistaken for fusion benefit [1, 17].

2.5.2 Alignment, Missing Modalities, and Temporal Relevance

Clinical modalities are asynchronous. EHR variables may update frequently, whereas the latest radiograph may remain unchanged for hours. Availability and relevance are therefore distinct: an image can exist but be too old to represent the current state. Hayat et al. [22] addressed paired and partially paired learning, while Yao et al. [76] used intervening EHR observations to estimate an updated latent CXR representation. DrFuse [77] separates shared and modality-specific information to manage absent or conflicting evidence.

Complete-case and missing-modality studies answer different questions. A paired cohort offers a controlled comparison because every fusion system receives both inputs, but it narrows the population to encounters with appropriately timed imaging. Training on partially paired data may increase sample size and deployment realism, yet performance can then reflect the method used to represent absence. These design choices should be treated as part of the scientific estimand rather than interchangeable preprocessing details.

Disagreement need not be noise. Stable physiology can coexist with a persistent radiographic abnormality, and physiological decline can precede visible change. Yin et al. [78] reported that EHR–CXR benefit varies by target, modality completeness, and unimodal strength. The defensible experimental strategy is therefore to hold cohort, task, and eligible observations constant while comparing model families, rather than infer complementarity from one multimodal score.

2.6 Transformers and Cross-Modal Attention

Attention-based interaction predates the Transformer. Bahdanau et al. [5] introduced encoder–decoder attention to remove the requirement that an input sequence be represented by a single fixed-length vector. Vaswani et al. [70] subsequently formulated scaled dot-product attention as

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

where queries determine the relevance of the keys and the corresponding values form the contextualized representation. Self-attention obtains queries, keys, and

values from the same sequence, whereas cross-attention obtains queries from one representation and keys and values from another. A Transformer operating only within an image or EHR encoder therefore remains unimodal unless its attention layers receive information from both modalities.

Lu et al. [46] developed hierarchical co-attention to contextualize image regions and words jointly. Tsai et al. [68] subsequently used directional cross-modal attention for unaligned multimodal sequences, while Lu et al. [45] introduced a two-stream Transformer in which separate visual and linguistic streams interact through co-attentional layers. In biomedical prediction, Chen et al. [12] used co-attention to relate histopathology regions to genomic representations, and Khader et al. [38] demonstrated Transformer-based integration of chest radiographs and clinical parameters. These studies establish cross-attention as a mechanism for conditioning one modality on another, but they do not establish that reciprocal interaction is always preferable to summary-level fusion.

Attention direction has a concrete representational meaning. In the proposed image–EHR setting, image-to-EHR attention asks which clinical periods are relevant to each spatial image token, whereas EHR-to-image attention asks which image regions contextualize each temporal clinical state. Although both directions use the same attention equation, exchanging the source of the queries, keys, and values changes the resulting conditional representation [45, 68]. Bidirectional cross-attention can therefore represent reciprocal contextualization without permanently assigning either modality the dominant role.

Cross-attention differs from summary concatenation because cross-modal interaction occurs before global compression. It can associate spatial image regions with periods of clinical instability, but a learned association under outcome supervision is not automatically causal or clinically explanatory. Hessel and Lee [25] showed that apparently multimodal models may rely primarily on additive unimodal evidence rather than substantive cross-modal interaction. Jain and Wallace [34] found that attention weights can disagree with other measures of feature importance and that substantially different attention distributions may produce similar predictions. Wiegreffe and Pinter [73] subsequently argued that the explanatory usefulness of attention depends on the definition of explanation and the evaluation procedure.

Token-level interaction also introduces computational and representational tradeoffs. For n_q query tokens and n_k key tokens, the cross-attention score matrix contains $n_q n_k$ elements, giving computational cost proportional to $O(n_q n_k d_k)$ and attention-map memory proportional to $O(n_q n_k)$ per head [67, 70]. Retaining a large spatial feature map can therefore make fine-grained image–EHR interaction unnecessarily expensive. Pooling or projecting each modality to a compact token set improves tractability but removes resolution before interaction. The token counts, attention directions, and retention of unimodal summary representations thus determine which information remains accessible to the fusion layer.

2.7 Related Work in Multimodal Clinical Prediction

2.7.1 Image and Clinical-Context Fusion

Medical image models frequently benefit from non-image context, but the form of context varies. Polsterl et al. [55] modulated image features using tabular variables. Khader et al. [38] combined CXRs and clinical parameters with a Transformer and improved mean diagnostic AUC over unimodal models on MIMIC and internal data. Hsieh et al. [27] combined MIMIC-IV-ED variables with CXR features for abnormality localization. These studies support contextual imaging but address diagnosis or localization rather than a temporally ordered early-warning endpoint.

Jabbour et al. [33] provide a respiratory diagnostic precedent. Their combined CXR–EHR model improved differentiation among pneumonia, heart failure, and chronic obstructive pulmonary disease, with performance varying by condition and external cohort. Lee et al. [41] similarly used CXR–EHR information to screen dyspnoeic emergency-department patients for abnormal NT-proBNP associated with acute heart failure. Both demonstrate target-dependent complementarity, but neither directly evaluates future respiratory deterioration.

2.7.2 EHR–CXR Fusion in the MIMIC Ecosystem

MedFuse is a central EHR–CXR comparator. Hayat et al. [22] used modality-specific encoders and recurrent fusion for mortality and phenotyping on MIMIC-IV and MIMIC-CXR, reporting gains on paired data and robustness to missing images.

MedFuse is especially relevant as a baseline because it retains modality-specific processing while combining asynchronous clinical and imaging information through a recurrent mechanism. It therefore tests a different interaction family from direct feature concatenation and cross-attention. At the same time, its mortality and phenotyping outcomes are broader than respiratory escalation, so its reported performance cannot be transferred directly to the present task. Reimplementation on a common cohort is necessary to distinguish architectural differences from differences in labels and eligible observations.

DrFuse addresses absent and inconsistent evidence through shared and modality-specific representations and disease-wise attention [77]. Lin et al. [43] combined physiological measurements, report representations, thoracic findings, and CXR features for ICU mortality; the inclusion of text means that its information setting differs from image plus structured EHR alone. DDL-CXR instead focuses on temporal staleness by updating a latent image representation from later clinical observations [76]. Yin et al. [78] benchmarked several EHR–CXR settings and concluded that improvement is conditional on the target, completeness, and strength of the individual modalities.

These studies are complementary rather than directly rankable. MedFuse emphasizes asynchronous availability, DrFuse separates shared and distinct information, DDL-CXR addresses stale images, and the benchmark work compares conditional gains. Their outcomes also differ, so a superior metric in one study cannot identify the best architecture for prospective respiratory risk. The transferable lesson is to align the fusion mechanism with the task and to retain strong unimodal comparisons.

Respiratory-specific evidence remains comparatively limited. EHR-only studies predict respiratory failure and intubation but cannot directly use lung appearance [30, 64]. Jabbour et al. [33] study diagnosis after respiratory failure is present, whereas Lu and Nemati [47] more directly study prospective ventilation risk using EHR and CXR representations. The latter is the closest task precedent but remains a recent preprint with a different cohort and fusion formulation.

The distinction between diagnosis and prognosis affects the useful evidence in each modality. Diagnostic models can exploit radiographic abnormalities that characterize an existing cause, whereas prospective models must determine whether the current combination of anatomy and trajectory precedes a later intervention. A finding can therefore be highly diagnostic yet weakly predictive of near-term escalation, or predictive only when interpreted together with recent physiology. This difference explains why respiratory diagnostic success supports multimodal feasibility without resolving the early-warning question.

Table 2.1 preserves the central comparison and its remaining boundaries. Performance values are not ranked because the studies use different populations, endpoints, and evaluation settings.

2.8 Synthesis and Research Gap

The literature supports four conclusions. First, temporal EHR data contain substantial respiratory and general ICU predictive signal [21, 30, 64]. Second, chest radiographs provide spatial pulmonary evidence unavailable from structured measurements alone [33, 36]. Third, multimodal models can improve diagnosis, phenotyping, mortality prediction, and localization [22, 27, 38, 77]. Fourth, gains are conditional on task definition, pairing, temporal relevance, missingness, optimization balance, and unimodal strength [71, 78].

The unresolved gap lies at their intersection. Much EHR–CXR research targets mortality, broad phenotypes, diagnosis, or image findings rather than forward-looking respiratory deterioration. Respiratory EHR studies lack radiographic input, while multimodal diagnostic studies do not establish whether imaging improves future risk beyond a substantial temporal EHR encoder. Existing prospective evidence is promising but does not determine whether fine-grained spatial–temporal interaction improves upon summary concatenation and recurrent fusion.

A convincing comparison must distinguish the benefit of adding a second

Table 2.1: Representative image–EHR studies and the remaining task-specific gap.

Study	Modalities and data	Task and fusion emphasis	Relevance and boundary
Jabbour et al. [33]	CXR and EHR; internal and external cohorts	Diagnosis of causes of acute respiratory failure	Respiratory and multimodal, but diagnostic rather than prospective
Hayat et al. [22]	MIMIC-IV time series and MIMIC-CXR	Mortality and phenotyping; recurrent fusion for asynchronous and partially paired data	Direct EHR–CXR baseline with broader outcomes
Khader et al. [38]	CXR and clinical parameters; MIMIC and internal data	Transformer-based diagnosis of multiple conditions	Attention-based precedent without temporal early warning
Yao et al. [77]	MIMIC-IV EHR and MIMIC-CXR	Phenotyping with shared/distinct representations and disease-wise attention	Addresses absent and conflicting modalities; multi-label target
Yin et al. [78]	MIMIC-IV EHR and MIMIC-CXR	Benchmark of several fusion settings	Tests conditional multimodal benefit; recent preprint
Lu and Nemati [47]	ICU EHR and CXR representations	Prospective invasive-mechanical-ventilation prediction	Closest prospective respiratory study; different fusion formulation

modality, the improvement achieved through joint prediction over the strongest unimodal model, and the contribution of the selected interaction mechanism. Keeping the cohort, outcome definition, data partitions, and unimodal branches constant reduces the likelihood that an observed improvement results from additional data, different encoders, or favourable sample selection rather than the fusion method itself.

Cross-attention tests whether relationships between image regions and EHR periods are lost when both modalities are globally pooled before interaction. Reciprocal cross-attention allows each token set to contextualize the other, although the underlying operation is already established in the literature [38, 46, 68]. The remaining research need is therefore evaluative and requires a controlled comparison of capable unimodal models, lower-complexity fusion, recurrent EHR–CXR fusion, and bidirectional token-level interaction using one paired cohort, one outcome definition, and patient-exclusive partitions. Chapter 3 presents the methodology used to conduct this comparison.

Chapter 3

Methodology

This chapter describes the complete experimental procedure used to construct and evaluate RespireTransFuse. It presents the study-specific decisions concerning data linkage, temporal eligibility, outcome construction, preprocessing, model specification, optimization, and evaluation. The central design principle was controlled comparison. All primary models were trained and evaluated using the same patient-exclusive partitions, paired radiograph–EHR observations, and binary outcome definition. Performance differences could therefore be compared without differences in cohort composition or prediction target, while interpretation still accounted for differences in architecture, trainability, and training objective.

3.1 Study Design and Experimental Overview

The study was a retrospective prediction-model development and internal evaluation study using de-identified data from MIMIC-IV and MIMIC-CXR. MIMIC-IV supplied ICU-stay identifiers, charted physiological observations, laboratory measurements, and timestamped respiratory-support procedures [35]. MIMIC-CXR supplied radiograph metadata and the corresponding locally available image files [36, 37]. Both resources are distributed through PhysioNet under credentialed access procedures [19, 53, 54]. The unit of analysis was an indexed frontal chest radiograph, identified at the DICOM-image level, acquired during an ICU stay. Its study acquisition timestamp defined the prediction index.

For each index radiograph, structured clinical observations were collected over the preceding 24 hours and binned into 24 one-hour intervals. The prediction target was the occurrence of a predefined respiratory-support intervention after the index. An event within 48 hours produced a positive label. A sample was negative only when no event occurred through 72 hours and the patient remained under hospital observation throughout that period; samples whose first future event occurred after 48 hours but no later than 72 hours were excluded. An earlier qualifying procedure

in the index ICU stay or a respiratory-support interval active at the index also excluded the candidate. The first future event was sought across ICU stays in the same hospital admission. This stable-negative construction created temporal separation between the positive and negative groups and reduced label ambiguity near the 48-hour decision boundary.

Seven models were evaluated: Image-Only CNN, EHR-Only Transformer, Early Fusion, RespireTransFuse, adapted MedFuse Uni-EHR, adapted MedFuse Uni-CXR, and adapted MedFuse Multimodal LSTM [22]. Together, they compared single-modality information, summary concatenation, reciprocal token interaction, and recurrent external baselines under one cohort and label.

The procedure constructed the paired cohort and labels, assigned patients to exclusive partitions, created training-derived inputs, and evaluated the selected checkpoints on the held-out test set across three training seeds (40, 41, and 42). The resulting cohort contained 9,509 paired samples from 4,668 patients, including 1,667 positive samples.

3.2 Data Sources, Identifiers, and Temporal Linkage

3.2.1 MIMIC-CXR radiographs

The radiograph source consisted of MIMIC-CXR metadata and the locally indexed image archive. Study date and time were parsed into a study timestamp. The metadata identifiers were matched to local image files using the available subject, study, and DICOM identifiers. Only frontal anteroposterior or posteroanterior views were retained. Lateral views were removed because the image encoder and the primary experimental comparison were defined for one frontal input per sample. A successful metadata-to-file match was required before a row could enter the cohort.

When more than one local representation was available, a deterministic image path was retained. The cohort stored database identifiers and a stable sample identifier, allowing every prediction to be traced to its image, ICU stay, index time, label, and EHR tensor position.

Images were decoded as RGB tensors and resized to 224×224 pixels, the resolution specified by the image configurations. Pixel scaling followed the normalization expected by the pretrained EfficientNet implementation [66]. The training transform was deliberately conservative; no horizontal flipping was used, random affine transformation was applied with probability 0.15, and mild colour jitter was applied with probability 0.03. Validation and test transforms were deterministic. The same image preprocessing was used in the four thesis-developed models so that changes in performance were not caused by different augmentation policies. Figure 3.1 shows three representative radiographs from the final paired cohort.

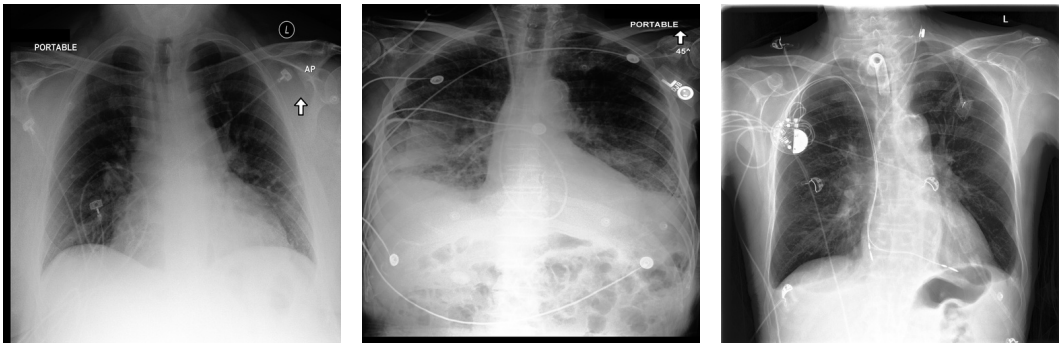


Figure 3.1: Representative indexed chest radiographs from the paired cohort.

3.2.2 MIMIC-IV ICU and event tables

MIMIC-IV ICU-stay records supplied `subject_id`, `hadm_id`, `stay_id`, admission time, and discharge time. Timestamped chart and laboratory tables supplied the candidate EHR variables. A radiograph was eligible only if its acquisition time fell within the corresponding ICU interval. This constraint prevented an image from a different encounter or a period outside critical care from being paired with the selected EHR window.

Respiratory-support events were assembled from the predefined procedure map (intubation: item 224385; invasive ventilation: 225792; non-invasive ventilation: 225794) and ordered by time. Earlier qualifying procedures within the index ICU stay, or a respiratory-support procedure interval active at the index, excluded a candidate. The first later qualifying event was then sought across ICU stays within the same hospital admission, subject to discharge and death boundaries. This procedure-based target is not an independently adjudicated diagnosis or proof of newly initiated support.

3.2.3 Index time and information boundary

Let $t_i^{(0)}$ denote the acquisition timestamp of the index CXR for sample i. The EHR observation interval was

$$\mathcal{W}_i = [t_i^{(0)} - 24\ \text{h},\ t_i^{(0)}], \tag{3.1}$$

and all outcome evidence occurred after $t_i^{(0)}$. The final tensor builder included observations timestamped exactly at the index boundary, whereas the earlier strict-candidate audit used a strictly pre-index boundary. This distinction is recorded because it affects exact reproducibility. The production tensor therefore corresponds to Equation 3.1 with an inclusive right endpoint. No measurement after the CXR index was admitted to the model input.

Figure 3.2 shows the temporal definition used throughout the study. The EHR

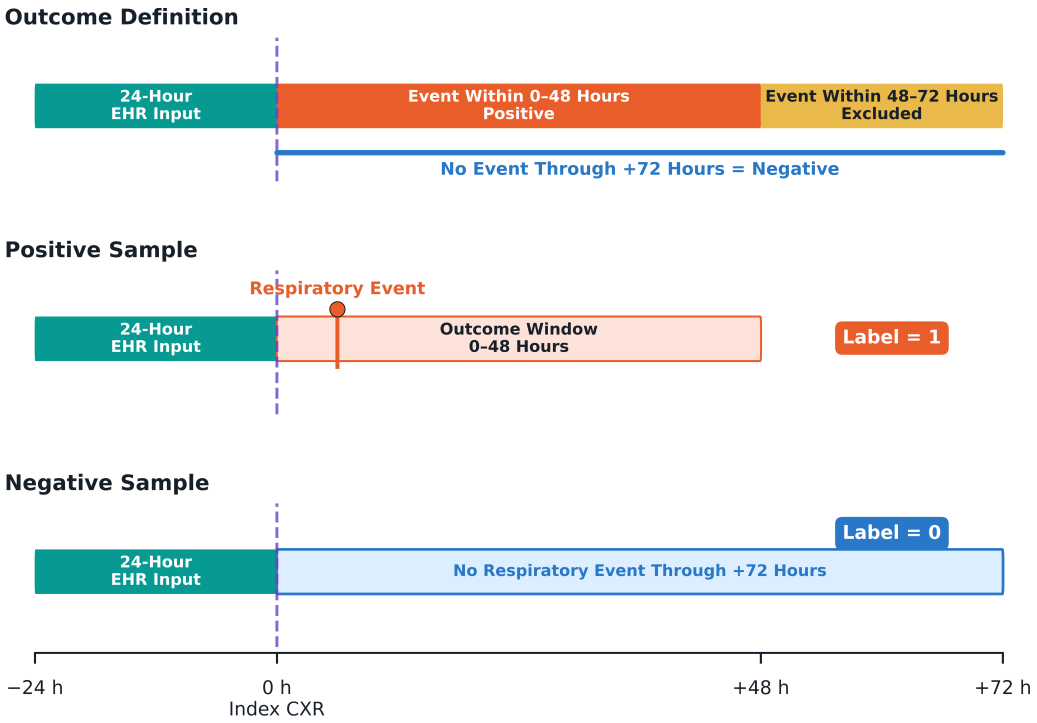


Figure 3.2: Temporal definition of model inputs and outcome labels. A positive sample has a first future respiratory event within 48 hours after the CXR index. A negative sample remains event-free through 72 hours. Samples with a first event between 48 and 72 hours are excluded.

window always precedes or terminates at the CXR index, and the label is assigned from future respiratory-support timing.

3.3 Cohort Construction and Outcome Definition

3.3.1 Eligibility procedure

The cohort was built before any feature-selection statistic or model fitting. An eligible row satisfied all of the following conditions:

- the radiograph was a locally available frontal AP or PA image with a valid study timestamp;
- the image could be associated with a MIMIC-IV subject, hospital admission, and ICU stay;
- the CXR timestamp occurred within the ICU-stay interval;

- respiratory-support timing could be determined from the predefined procedure table;
- no qualifying intervention was recorded earlier in the index ICU stay and no qualifying support interval was active at the CXR index;
- future events were checked across ICU stays within the same hospital admission;
- the row was positive within 48 hours or was event-free and under hospital observation for at least 72 hours; otherwise it was excluded.

Excluding interventions recorded earlier in the index stay, including support active at the index, reduced the risk of predicting an already documented intervention. However, earlier ICU stays in the same admission were not screened for all past procedure exposure, so newly initiated support cannot always be verified. Excluding first events in the 48–72-hour interval avoided labeling near-future procedures as stable negative.

Let Δ_i be the elapsed time in hours from the index to the first subsequent qualifying procedure, with $\Delta_i = \infty$ if none was recorded. Let $O_i(72) = 1$ indicate uninterrupted hospital observation through 72 hours without earlier discharge or death. After the index-stay prior-support exclusions, the label was

$$y_i = \begin{cases} 1, & 0 < \Delta_i \leq 48, \\ 0, & \Delta_i > 72 \text{ and } O_i(72) = 1, \\ \text{excluded}, & \text{otherwise.} \end{cases}$$

The event map and label are clinical-operational definitions for this retrospective study; they are not presented as a universal definition of respiratory failure.

3.3.2 Unit of analysis and repeated patients

The unit of analysis was the eligible indexed radiograph. A patient could contribute more than one indexed study when each study independently satisfied the cohort eligibility and temporal-definition criteria. Consequently, the number of analytic samples exceeded the number of unique patients.

To prevent information leakage across repeated studies from the same individual, all samples associated with a given `subject_id` were assigned to a single partition. A patient therefore could not contribute samples to more than one of the training, validation, or test sets. Patient-level grouping was also preserved during uncertainty estimation, for which patients rather than individual rows were resampled. The stable `sample_id` provided the common identifier used to align cohort records, EHR tensors, radiographs, model predictions, and resampling outputs throughout the experimental pipeline.

3.3.3 Patient-level partitioning

The final analytic cohort was divided into approximately 70% training, 15% validation, and 15% test partitions using a seeded patient-level split search. Whole-patient assignment was preserved throughout the procedure, ensuring that no `subject_id` appeared in more than one partition. The split search jointly targeted the desired sample proportions and closely matched positive-event prevalence while retaining all samples from the same patient within one partition.

The validation partition was used for checkpoint selection, early stopping, hyperparameter comparison, and selection of the final operating threshold. Where post-hoc calibration was reported, its parameters were also determined without consulting the test labels. The test partition was not used for feature selection, normalization parameter estimation, hyperparameter selection, checkpoint selection, or threshold determination. This separation prevented preprocessing or model-selection decisions from becoming indirect forms of test-set tuning.

3.4 Multimodal Preprocessing Pipeline

The ordered preprocessing scripts produced the fixed paired-cohort CSV, verified chest-radiograph paths, and train-ready EHR archive used by all evaluated models. Figure 3.3 summarizes the complete workflow. EHR feature-selection evidence and normalization parameters were estimated exclusively from the training partition and were subsequently applied unchanged to the validation and test partitions. The preprocessing and model-training scripts are publicly available in the RespireTransFuse GitHub repository at https://github.com/arabneidi/RespireTransFuse.

3.4.1 Hourly EHR tensor construction

Each sample was represented by 24 ordered hourly bins covering the 24-hour observation window before the index radiograph. For a recorded event at time t, the corresponding hour index was computed relative to $t_i^{(0)} - 24$ hours and clipped to the valid range. When a variable was measured more than once within the same hour, the arithmetic mean of those measurements was used. This aggregation reduced duplicate charting while preserving a fixed temporal resolution suitable for batching. The raw value tensor and observation mask were

$$X_i^{\text{raw}} \in \mathbb{R}^{24\times F}, \qquad M_i \in \{0,1\}^{24\times F},$$

where $M_{i,t,f} = 1$ indicates that feature f had at least one recorded value in hour t. Missing values remained `NaN` during feature selection and normalization estimation so that absence of measurement was not conflated with a true numerical zero.

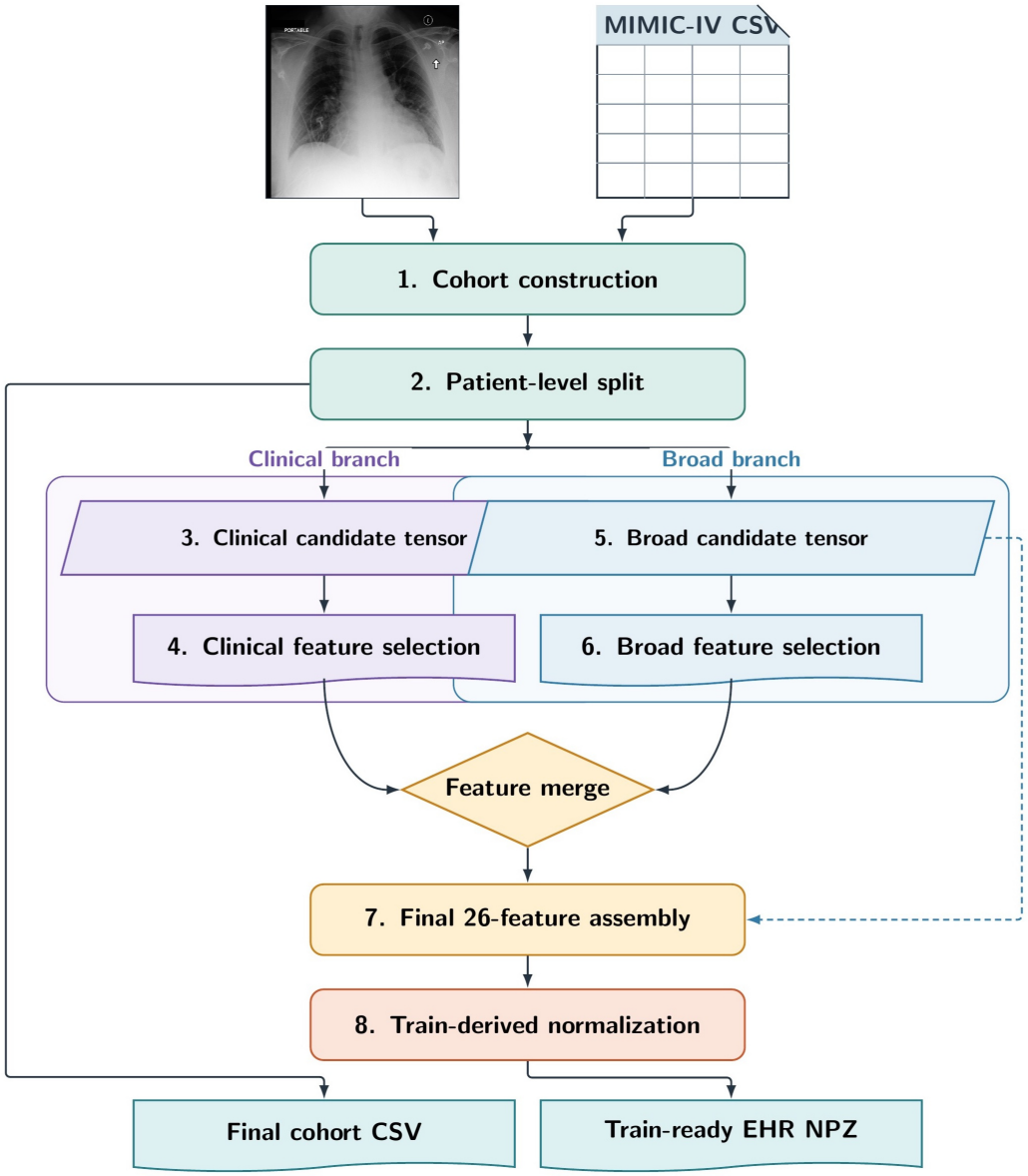


Figure 3.3: Overview of cohort construction and EHR preprocessing. The paired cohort is divided by patient, candidate features are selected and merged into the final 26-variable registry, and training-derived normalization produces the train-ready EHR arrays.

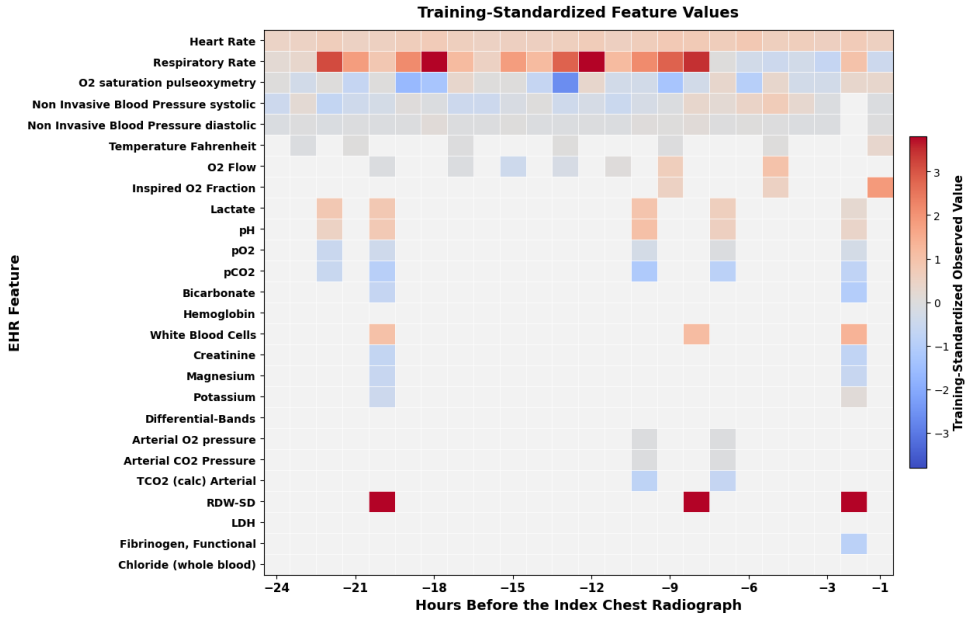


Figure 3.4: Illustrative 24 × 26 EHR tensor for a positive test sample over the 24 hours preceding the index chest radiograph. Coloured cells represent training-standardized observed values, whereas light-grey cells indicate unobserved feature–hour positions.

This hourly representation preserved approximate measurement timing without implying that every variable was sampled once per hour. Charted variables were included only if their chart time was no later than the index radiograph, and laboratory variables were included only if their store time was no later than the index. This avoided introducing test results that would not yet have been available at prediction time. Figure 3.4 shows an illustrative final 24 × 26 EHR tensor from a positive test case and highlights the temporal sparsity of the structured input.

3.4.2 Strict clinically constrained candidate branch

The clinically constrained branch restricted the initial candidate space to conventional numeric bedside physiology and common blood, serum, and blood-gas measurements. Administrative fields, clinical scores, alarms, ventilator-device settings, procedures, and most treatment-related variables were excluded. Pre-index oxygen-flow and inspired-oxygen-fraction measurements were retained as clinically relevant support-intensity variables. At most 80 chart and 80 laboratory candidates were retained. This branch ensured that established measurements of vital status, oxygenation, acid–base balance, metabolism, haematology, renal function, electrolytes, and coagulation remained represented.

Measurements from the preceding 24 hours were aggregated into hourly tensors. Variables with no training observations were removed, and eight temporal summaries were evaluated using training-only point-biserial association, AUROC, AUPRC lift, mutual information, coverage, correlation clustering, mRMR, and repeated Elastic Net stability fits [52, 80].

3.4.3 Broad statistically screened candidate branch

The broad branch searched a substantially larger set of numeric chart and laboratory variables. After excluding likely leakage, administrative, treatment, procedure, and device-status fields, the same statistical evidence procedure was applied, with Elastic Net analysis restricted to the 120 strongest eligible candidates. This branch allowed less clinically obvious but potentially complementary inflammatory, coagulation, metabolic, and haematological measurements to be considered.

The branches produced ranked evidence rather than an automatically selected final feature list. The final 26-variable registry combined a clinically constrained core with additional measurements supported by the broader statistical screening. This hybrid design preserved clinical interpretability while reducing the possibility that manual clinical screening alone would omit empirically informative measurements.

3.4.4 Final 26-variable registry

Table 3.1 gives the fixed tensor order. Joins required feature name, source table, and item identifier, preventing similarly named measurements from being merged.

Table 3.1: 26-variable EHR registry and fixed tensor order.

No.	Variable	Source	Item ID
1	Heart Rate	`chartevents`	220045
2	Respiratory Rate	`chartevents`	220210
3	O2 saturation pulseoxymetry	`chartevents`	220277
4	Non-Invasive Blood Pressure systolic	`chartevents`	220179
5	Non-Invasive Blood Pressure diastolic	`chartevents`	220180
6	Temperature Fahrenheit	`chartevents`	223761
7	O2 Flow	`chartevents`	223834
8	Inspired O2 Fraction	`chartevents`	223835
9	Lactate	`labevents`	50813
10	pH	`labevents`	50820

Continued on next page

Table 3.1 – continued

No.	Variable	Source	Item ID
11	pO2	labevents	50821
12	pCO2	labevents	50818
13	Bicarbonate	labevents	50882
14	Hemoglobin	labevents	50811
15	White Blood Cells	labevents	51301
16	Creatinine	labevents	50912
17	Magnesium	labevents	50960
18	Potassium	labevents	50971
19	Differential–Bands	chartevents	225638
20	Arterial O2 pressure	chartevents	220224
21	Arterial CO2 pressure	chartevents	220235
22	Arterial total CO2 (calc.)	chartevents	225698
23	RDW-SD	labevents	52172
24	LDH	chartevents	220632
25	Fibrinogen, Functional	labevents	51214
26	Chloride (whole blood)	chartevents	226536

3.4.5 Training-derived normalization and missingness representation

For feature f, the mean μ_f and standard deviation σ_f were estimated from finite observed cells in the training partition only. Validation and test values did not influence these statistics. When the estimated standard deviation was nonfinite or smaller than 10^{-6}, its replacement value was 1 (rather than 10^{-6}). The observed values were standardized as

$$X_{i,t,f} = \frac{X^{\text{raw}}_{i,t,f} - \mu_f}{\widetilde{\sigma}_f}, \qquad \widetilde{\sigma}_f = \begin{cases} 1, & \sigma_f < 10^{-6} \text{ or } \sigma_f \text{ is nonfinite,} \\ \sigma_f, & \text{otherwise.} \end{cases}$$

After standardization, unobserved cells were assigned zero in X while remaining zero in M. Consequently, an input zero could mean either an observed value equal to the training mean or a missing value, but the mask distinguished these cases. The final archive contained normalized X, binary M, unnormalized X^{raw}, labels, split assignments, sample identifiers, feature metadata, and the training statistics.

The EHR encoder consumed values and masks jointly. A time-position validity

indicator was also derived:

$$m_{i,t} = \mathbb{I}\left(\sum_{f=1}^{26} M_{i,t,f} > 0\right). \tag{3.2}$$

An hour with $m_{i,t} = 0$ was excluded when it served as an attention key or contributed to masked temporal means. This represented observed-data availability without label-dependent imputation and retained the informative sampling pattern emphasized in clinical time-series research [11].

3.5 Notation and Shared Input Interface

For a mini-batch of size B, the image input was $I \in \mathbb{R}^{B\times3\times224\times224}$, normalized EHR values were $X \in \mathbb{R}^{B\times24\times26}$, the observation mask was $M \in \{0,1\}^{B\times24\times26}$, and labels were $y \in \{0,1\}^{B}$. Each model produced a scalar logit z_i and probability $\widehat{p}_i = \sigma(z_i)$. Multimodal models additionally returned image-branch and EHR-branch logits for auxiliary training and error analysis.

Model interfaces were consistent even when an architecture used only one modality. Image-only ignored X and M; EHR-only ignored I; multimodal models required all three. The final cohort contained complete paired inputs, so the primary experiment did not evaluate missing-modality inference. Table 3.2 summarizes the role of each system.

3.6 Adopted Baseline and Proposed Models

3.6.1 Adapted MedFuse Baseline Models

MedFuse was included as an external EHR–CXR comparator because Hayat et al. [22] developed the architecture for multimodal learning from clinical time series and chest radiographs. The original code structure was adapted to read the final cohort CSV and train-ready EHR archive, return a single binary respiratory-deterioration output, and preserve the patient-level training, validation, and test assignments. The adaptation did not claim to reproduce the original study endpoint or its reported numerical results; instead, it transferred the MedFuse architecture to the prediction task defined in this thesis. Three adapted configurations were evaluated: Uni-EHR, Uni-CXR, and multimodal LSTM fusion. The recurrent and convolutional baselines provide an important test of whether temporal self-attention and cross-modal attention are necessary for this particular endpoint, rather than assuming that a Transformer is universally preferable.

Table 3.2: Models evaluated under the common cohort, label, and partitions.

Model	Inputs	Methodological role
Image-only	CXR	Shared EfficientNet branch; radiographic information only.
EHR-only	X, M	Shared mask-aware Transformer branch; 24-hour temporal information only.
Early Fusion	CXR, X, M	Independently normalized modality summaries followed by concatenation.
MedFuse Uni-CXR	CXR	External radiology-pretrained ResNet-34 comparator with a frozen image backbone and newly trained binary classification head.
MedFuse Uni-EHR	EHR sequence	External recurrent EHR comparator.
MedFuse LSTM	CXR, EHR sequence	External multimodal recurrent-fusion comparator.
RespireTransFuse	CXR, X, M	Reciprocal cross-attention between compact image and EHR token sequences, followed by direct joint classification.

MedFuse Uni-EHR

The `uni_ehr` mode was selected to test whether gated sequential recurrence could summarize the short, irregular 24-hour clinical history without a Transformer. LSTM gates preserve or discard prior state as new observations arrive, offering a meaningful alternative when relevant information is predominantly local or sequential. Omitting the recurrent encoder would remove the baseline's temporal memory and reduce it to a non-recurrent input summary. The `uni_ehr` mode provided the unimodal EHR MedFuse baseline. It used the framework's LSTM clinical encoder and disabled contribution from the CXR branch. The recurrent encoder used a hidden dimension of 256, two stacked recurrent layers, and dropout of 0.20 in the adapted configuration. LSTM recurrence follows the cell-state and input, forget, and output-gate formulation introduced by Hochreiter and Schmidhuber [26]. The final recurrent representation was passed to a single binary classification output for respiratory-deterioration prediction.

This baseline differs architecturally from the thesis-developed EHR-only model. It uses recurrent rather than Transformer-based temporal modeling and retains the MedFuse data and training structure. It was not designed to match the EHR-only Transformer parameter-for-parameter; its purpose was to compare the proposed temporal representation with an established recurrent clinical encoder.

MedFuse Uni-CXR

This configuration tests whether a radiology-pretrained convolutional representation alone is informative for the future support endpoint. Freezing the pretrained ResNet-34 isolates its transferable radiographic features, but also restricts adaptation to task-specific image patterns; removing the classification head would leave no trained mapping to the binary outcome. The `uni_cxr` mode provided the unimodal image MedFuse baseline. Its visual branch used the ResNet-34 architecture introduced by He et al. [23]. The branch was initialized from the available MedFuse chest-radiograph checkpoint, which had originally been trained with a 14-label radiology classifier. Because that output layer was incompatible with the binary respiratory-deterioration endpoint, the original classifier was excluded and replaced with a single-output binary classification layer.

Compatible ResNet-34 tensors were transferred from the radiology checkpoint, while the task-specific classifier was initialized for the prevalence of the respiratory-deterioration outcome. In the final Uni-CXR configuration, the pretrained visual backbone remained frozen and only the new binary classifier was optimized. The resulting model therefore functioned as a linear evaluation of the MedFuse radiographic representation rather than a fully fine-tuned image network.

This baseline differs from the thesis-developed Image-Only CNN in both backbone and adaptation strategy. The MedFuse Uni-CXR model used a frozen ResNet-34 encoder with a 513-parameter binary output layer, whereas the Image-Only CNN used the project-specific EfficientNet-B0 training configuration.

MedFuse LSTM Multimodal Model

The recurrent multimodal baseline was retained to test whether a sequential fusion mechanism, which does not explicitly align individual image regions with individual EHR hours, can nonetheless combine clinically useful representations. Its performance provides a counterpoint to the added complexity of token-level reciprocal attention; without the CXR branch, it reduces to a recurrent EHR comparison and cannot test radiographic complementarity. The multimodal MedFuse baseline used the `lstm` fusion mode. Its clinical branch employed the same recurrent encoder as the Uni-EHR model, while its visual branch used the ResNet-34 image encoder [23]. The modality representations were combined by the recurrent fusion mechanism implemented in MedFuse, producing a joint binary output while retaining separate clinical and radiographic encoders. In the reported multimodal runs, the matching-seed Uni-EHR checkpoint and radiology-pretrained CXR checkpoint supplied the frozen unimodal branches, allowing the recurrent fusion stage to combine established features.

Unlike the unimodal configurations, this model used information from both the EHR sequence and the chest radiograph. Its purpose was to provide an external

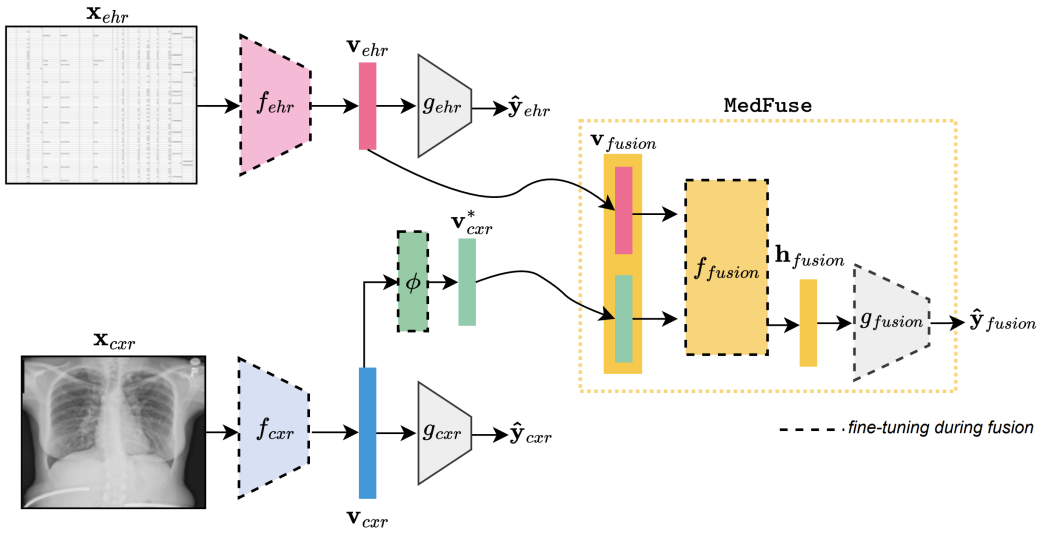


Figure 3.5: MedFuse architecture used as the basis of the three external baseline configurations. The Uni-EHR experiment uses the clinical branch, the Uni-CXR experiment uses the radiographic branch with a task-specific binary classifier, and the multimodal LSTM experiment combines the clinical and CXR representations through the MedFuse fusion module. Adapted from Hayat et al. [22].

multimodal comparison against Early Fusion and RespireTransFuse while preserving the principal recurrent fusion structure supplied by the MedFuse framework. Figure 3.5 summarizes the baseline architecture and its three evaluated configurations.

3.6.2 Image-Only Model

The image-only model measures how much of the prospective respiratory-support signal can be recovered from a single radiograph without access to preceding physiological trends. EfficientNet-B0 was chosen as a comparatively compact pretrained convolutional encoder; freezing it limits trainable complexity on the paired cohort. Global pooling captures distributed lung and thoracic features, whereas learned spatial-attention pooling can emphasize localized abnormalities; their learned interpolation avoids assuming either view is always preferable. Removing focused pooling would leave only a spatially uniform summary, potentially missing localized evidence, while removing global pooling could make predictions overly dependent on a few attended regions. Neither pooling contribution was isolated in the reported matched fusion ablations, so these are architectural expectations rather than independently demonstrated effects.

Table 3.3: Stage-level architecture of the `tf_efficientnet_b0.ns_jft_in1k` feature extractor used in the image-only model.

Stage	Operation	Expansion	Kernel/stride	Repeats	Output tensor
Input	Chest-radiograph tensor	–	–	–	$3 \times 224 \times 224$
Stem	Convolution, batch normalization, SiLU	–	$3 \times 3/2$	1	$32 \times 112 \times 112$
1	MBConv1 with squeeze-and-excitation	1	$3 \times 3/1$	1	$16 \times 112 \times 112$
2	MBConv6 with squeeze-and-excitation	6	$3 \times 3/2$	2	$24 \times 56 \times 56$
3	MBConv6 with squeeze-and-excitation	6	$5 \times 5/2$	2	$40 \times 28 \times 28$
4	MBConv6 with squeeze-and-excitation	6	$3 \times 3/2$	3	$80 \times 14 \times 14$
5	MBConv6 with squeeze-and-excitation	6	$5 \times 5/1$	3	$112 \times 14 \times 14$
6	MBConv6 with squeeze-and-excitation	6	$5 \times 5/2$	4	$192 \times 7 \times 7$
7	MBConv6 with squeeze-and-excitation	6	$3 \times 3/1$	1	$320 \times 7 \times 7$
Head	Pointwise convolution, batch normalization, SiLU	–	$1 \times 1/1$	1	$1280 \times 7 \times 7$
Excluded head	Original global pooling and 1,000-class classifier	–	–	–	Not used

EfficientNet feature extractor

The image-only system used `tf_efficientnet_b0.ns_jft_in1k` from the PyTorch Image Models library [74]. EfficientNet was introduced by Tan and Le [66] as a family of convolutional networks that jointly scales network depth, width, and input resolution. The B0 configuration forms the smallest baseline member of this family and processes images at a nominal resolution of 224×224 pixels.

The selected `ns_jft_in1k` weights were obtained using the Noisy Student training procedure described by Xie et al. [75]. This pretraining combines labelled ImageNet data with pseudo-labelled examples from JFT-300M before ImageNet-1k adaptation. The pretrained representation was used only as a generic visual feature extractor; no ImageNet class predictions were used in the respiratory deterioration task.

The backbone was instantiated without its original 1,000-class linear classifier. During the forward pass, the implementation called `forward_features`, thereby bypassing the original global pooling and classification path and retaining the final convolutional feature map. For an input chest radiograph

$$I_i \in \mathbb{R}^{3\times224\times224},$$

the extracted representation was

$$A_i = \phi_{\mathrm{EffB0}}(I_i) \in \mathbb{R}^{1280\times7\times7}.$$

Table 3.3 summarizes the EfficientNet-B0 feature-extraction stages used in this thesis. One row is reported for each architectural stage, following the stage-level description used in the original EfficientNet specification. The repeat column gives the number of individual blocks contained within each stage, and the reported tensor dimensions correspond to the output of that stage for a 224×224 input.

Each MBConv6 block first expands the channel dimension using a 1×1 pointwise convolution, applies a depthwise convolution with the stage-specific kernel, recalibrates channels using a squeeze-and-excitation module, and projects the result using a second 1×1 convolution. MBConv1 follows the same general depthwise-separable structure but does not perform the sixfold channel expansion. Residual connections are applied when the spatial resolution and channel dimensions permit identity addition.

Spatial resolution was progressively reduced from 224×224 to 7×7, while the channel dimension increased from 3 input channels to 1,280 high-level feature channels. The final 1×1 convolution expanded the 320-channel output of the last MBConv stage to the 1,280-channel map retained by the proposed image-processing layers.

All EfficientNet parameters were frozen by setting `requires_grad=False`. The backbone was also maintained in evaluation mode during training, preventing updates to both its learned weights and batch-normalization running statistics. Only the newly introduced spatial-pooling and classification components were optimized. This design preserved the pretrained visual representation while limiting the number of trainable parameters and reducing the risk of overfitting.

Figure 3.6 summarizes the complete image-only path. The frozen EfficientNet-B0 backbone produces the $1280 \times 7 \times 7$ spatial feature map, which is subsequently processed by global and focused spatial pooling, learned interpolation, and the task-specific prediction head.

Global and focused spatial pooling

Two summaries were derived from A_i. Global average pooling produced

$$g_i = \frac{1}{49} \sum_{p=1}^{49} A_{i,p} \in \mathbb{R}^{1280}.$$

In parallel, a compact scorer applied a 1×1 convolution from 1280 channels to four channels, GELU activation, and a second 1×1 convolution to one score per spatial position. A softmax over the 49 positions produced normalized weights $a_{i,p}$, from which the focused summary was

$$q_i = \sum_{p=1}^{49} a_{i,p} A_{i,p}, \qquad \sum_{p=1}^{49} a_{i,p} = 1.$$

The final visual feature interpolated between the broad and focused summaries:

$$v_i^{(I)} = g_i + \alpha(q_i - g_i), \qquad \alpha = \sigma(a_{\text{mix}}). \tag{3.3}$$

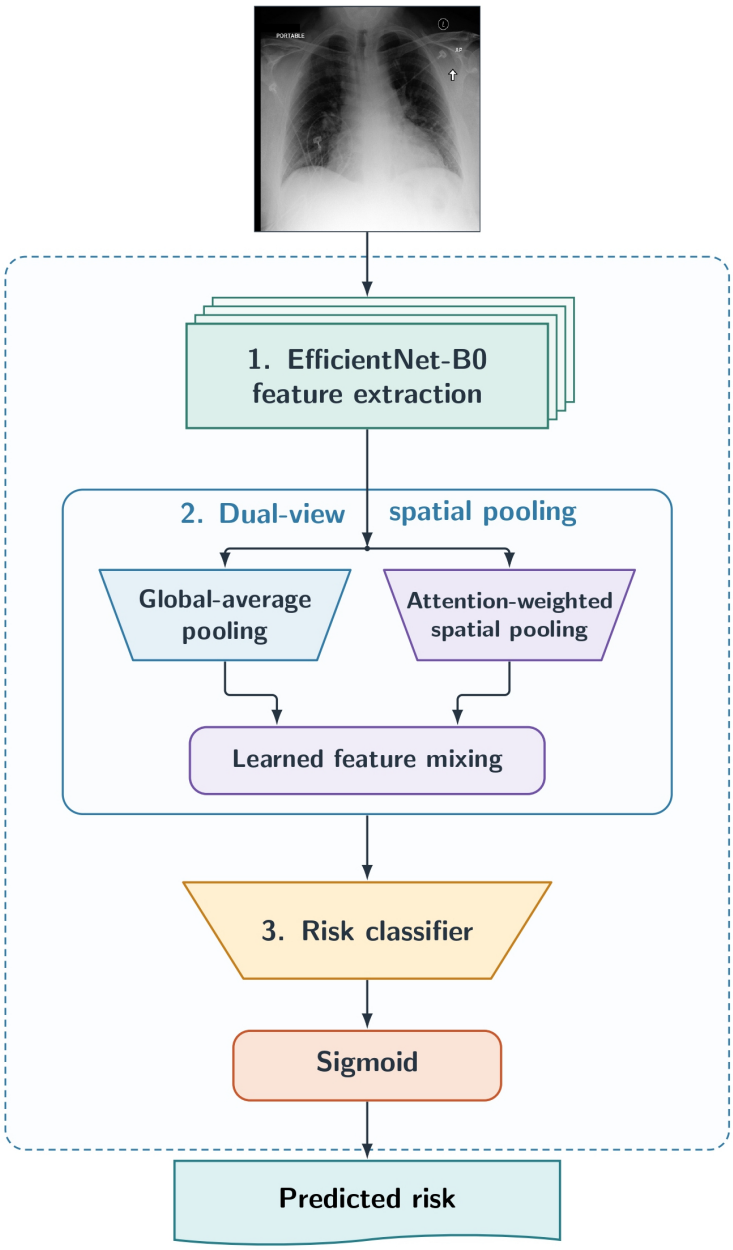


Figure 3.6: Overview of the image-only architecture. A frozen EfficientNet-B0 extracts spatial features, which are summarized through global-average and attention-weighted pooling, combined by learned feature mixing, and mapped to the predicted respiratory-deterioration risk.

The scalar a_{mix} was initialized to -2 ($\alpha \approx 0.119$), initially favouring global pooling. The final scorer convolution was initialized at zero, producing approximately uniform spatial weights.

The image classifier applied layer normalization to $v_i^{(I)}$, dropout 0.35, a $1280 \to 128$ linear layer, GELU, another dropout layer, and a $128 \to 1$ output layer. Its logit was

$$z_i^{(I)} = h_I(v_i^{(I)}), \qquad \widehat{p}_i^{(I)} = \sigma(z_i^{(I)}).$$

The scorer, interpolation scalar, and classifier were trainable. Multimodal models reused the same $v_i^{(I)}$ and A_i.

3.6.3 EHR-Only Model

The EHR-only system combined mask-aware projection, local convolution, sinusoidal positions, Transformer self-attention, and multi-view aggregation. It also formed the clinical branch of both thesis-developed fusion models. The design responds to two characteristics of this task: respiratory-support decisions may reflect changes over several separated hours, while EHR variables are sampled unevenly. The input mask distinguishes unavailable values from physiologically average observations; short-kernel convolutions emphasize nearby fluctuations, and self-attention allows measurements at nonadjacent times to influence each other directly. Multiple summary views preserve overall context and the latest observed state. Removing the mask would conflate missingness with standardized zero, removing temporal convolutions could weaken local-pattern extraction, and removing attention would eliminate learned global pairwise temporal interaction unless another sequence operator replaced it. Attention, layer normalization, and GELU followed established formulations [4, 24, 70]. Figure 3.7 gives the complete EHR-only path from the values and mask tensors to the binary logit.

Reference Transformer encoder–decoder structure

Figure 3.8 records the standard encoder–decoder Transformer structure that motivates the attention operations used in the implemented models [70]. The generic decoder contains masked self-attention and encoder–decoder attention for autoregressive generation. This study predicts one binary outcome rather than generating a sequence, so the EHR branch uses only the encoder path. RespireTransFuse then introduces two explicit cross-attention directions between the encoded EHR and image tokens; it does not instantiate the autoregressive decoder stack.

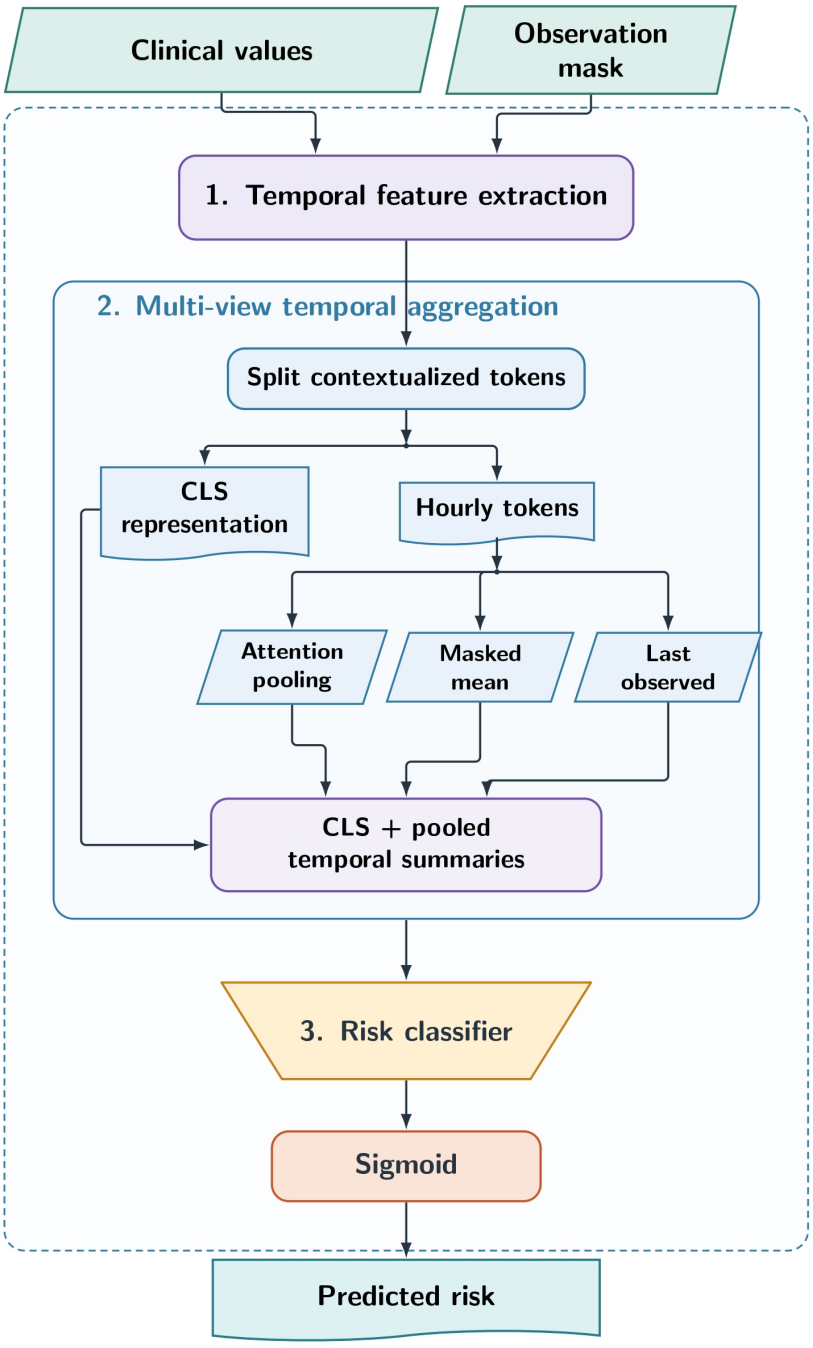


Figure 3.7: Overview of the EHR-only architecture. The value and observation-mask channels are jointly projected, locally processed, contextually encoded, summarized through complementary pooling views, and classified.

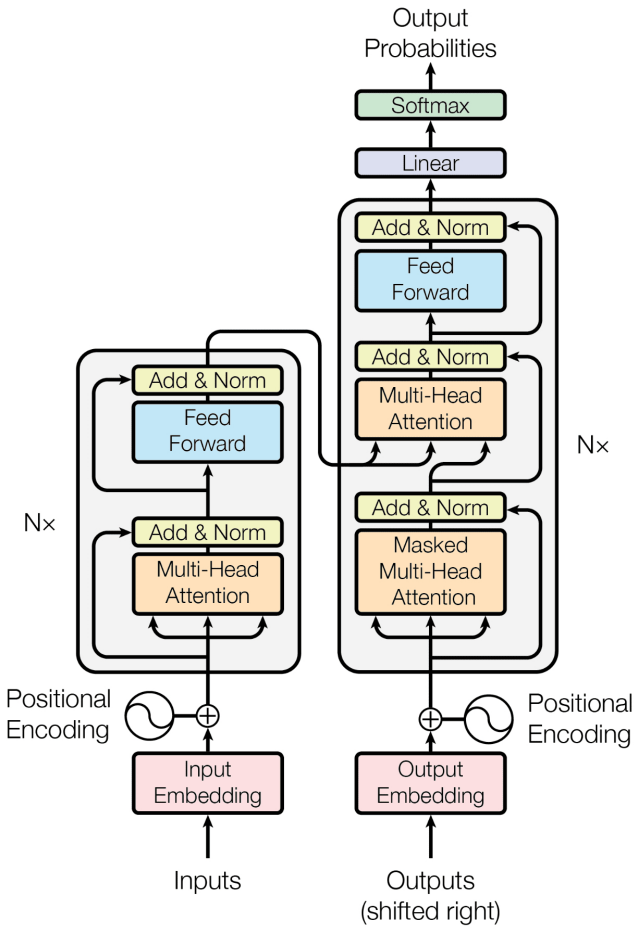


Figure 3.8: Reference Transformer encoder–decoder architecture. The EHR-only branch uses the encoder pathway; the autoregressive decoder shown on the right is included to distinguish the standard sequence-generation architecture from the binary prediction architecture implemented in this study. Adapted from Vaswani et al. [70].

Values–mask projection

At each hour, the 26 normalized values and 26 binary observation indicators were concatenated:

$$u_{i,t} = [X_{i,t,:} \| M_{i,t,:}] \in \mathbb{R}^{52}.$$

A linear layer, layer normalization, GELU, and dropout 0.35 mapped the vector to $d = 128$:

$$H_{i,t}^{(0)} = \text{Dropout}(\text{GELU}(\text{LN}(W_0 u_{i,t} + b_0))) .$$

The mask distinguished an observed mean-valued cell from a missing cell; Equation 3.2 identified empty hours.

Multi-scale local temporal block

Before global self-attention, a local block processed each feature channel with depthwise one-dimensional convolutions of width three and five. With $\widetilde{H} = \text{LN}(H^{(0)})$, the update was

$$\begin{aligned} R_3 &= \text{DWConv}_3(\widetilde{H}), \\ R_5 &= \text{DWConv}_5(\widetilde{H}), \\ R &= \text{Dropout}\left(\text{GELU}\left(\text{Conv}_1\left([R_3 \| R_5]\right)\right)\right), \\ H^{(1)} &= H^{(0)} + \beta R, \qquad \beta = \sigma(b_{\text{local}}). \end{aligned}$$

The depthwise paths retained 128 channels, concatenated to 256, and returned to 128 through the pointwise convolution. Empty-hour updates were zeroed. The scale logit was initialized to -1.1, and kernels three and five represented short local patterns.

Positional encoding and Transformer encoder

A Transformer encoder was preferred to an autoregressive encoder–decoder because this task produces one future-event probability from a fixed observation window, not a sequence of predicted values. Self-attention is potentially useful when a recent oxygenation change must be interpreted together with earlier respiratory rate, oxygen delivery, or laboratory measurements, especially if their records are separated by missing hours. Sinusoidal positions retain hourly order, while key-padding masks prevent entirely unobserved hours from being used as attention evidence. Without positions the same multiset of hourly values would carry less temporal-order information; without the key-padding mask, unobserved hours could dilute attention. However, the 24-hour window is short, some cases have sparse observations, and a gated recurrent encoder may already capture sufficient temporal history. Transformers therefore do not inherently guarantee higher

AUROC or AUPRC, and these alternatives are compared empirically rather than attributing all differences to self-attention alone.

Fixed sinusoidal positional encodings were added to the 24 local representations to preserve temporal order. A learned classification token was prepended, producing 25 positions. The classification position was always valid; the remaining positions used the hour mask. Two pre-normalized Transformer encoder layers processed the sequence. Each layer used four attention heads, model width 128, feed-forward width 320, GELU, and dropout 0.35.

For one self-attention head, the contextual update is represented by

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \mathcal{P}_i\right) V,$$

where $\mathcal{P}_i$ masks invalid keys. The output was a classification vector $c_i \in \mathbb{R}^{128}$ and contextual tokens $T_i \in \mathbb{R}^{24\times128}$.

Figure 3.9 expands the EHR Transformer itself. It shows the pre-normalized attention and feed-forward residual sublayers repeated twice, as implemented by the two-layer encoder configuration. The observation-derived hour mask enters self-attention as a key-padding mask; it is not an outcome-dependent mask.

Multi-view temporal aggregation

Three summaries were derived from T_i: a learned attention-weighted sum, a masked arithmetic mean, and the last observed hourly token. The learned attention summary used a layer-normalized $128 \to 64 \to 1$ scoring network with tanh and dropout. The three summaries were stacked and scored by a separate layer-normalized linear function. A softmax produced sample-specific weights $\omega_{i,k}$:

$$r_i = \sum_{k=1}^{3} \omega_{i,k} s_{i,k}, \qquad \omega_{i,k} \geq 0, \quad \sum_{k=1}^{3} \omega_{i,k} = 1.$$

Zero score initialization gave equal initial contribution, after which the model could emphasize attention pooling, the mean, or the latest observation per sample.

The final EHR representation concatenated the classification token and mixed temporal summary:

$$v_i^{(E)} = [c_i \| r_i] \in \mathbb{R}^{256}.$$

Its classifier used layer normalization, dropout 0.35, $256 \to 96$, GELU, dropout, and $96 \to 1$. RespireTransFuse-specific adapters projected tokens $128 \to 48$ and the summary $256 \to 48$; they were disabled in EHR-only.

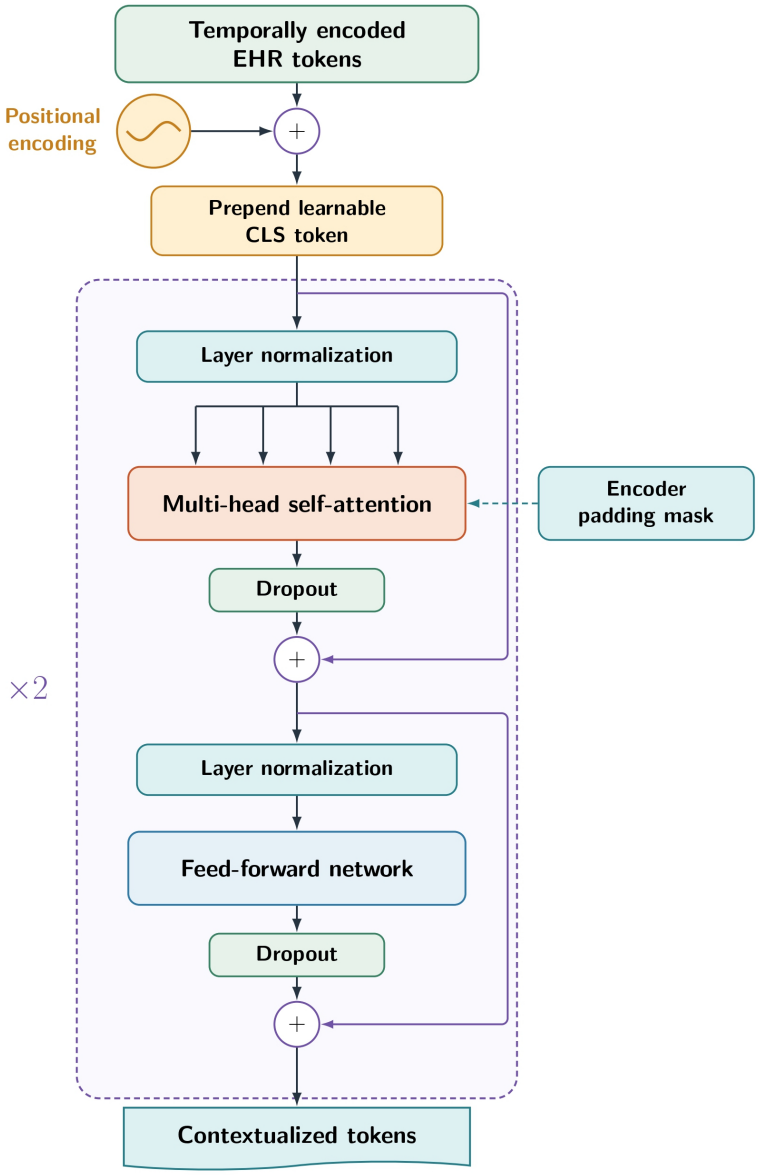


Figure 3.9: Detailed Transformer architecture inside the EHR-only model. After mask-aware input projection and multi-scale local temporal processing, positional information and a learned classification token enter two pre-normalized Transformer encoder layers.

3.6.4 Early-Fusion Model

Early Fusion was included as a low-complexity reference that tests whether two individually useful modality summaries are sufficient without learning region-to-hour correspondences. Separate projections and L_2 normalization put the unequal-dimensional image and EHR summaries on comparable scales before classification. If normalization were removed, relative feature magnitudes could dominate fusion; if token interaction is beneficial, its absence may limit this baseline. Conversely, when the modalities are redundant or the paired cohort is too small to support extra attention parameters, this simpler design may generalize as well as or better than attention-based fusion. The matched A0 concatenation ablation provides a more direct test of cross-attention removal than the separately optimized Early-Fusion baseline.

The Early Fusion model combines learned modality summaries after independent image and EHR encoding. It is therefore a feature-level concatenation baseline rather than a method that fuses raw inputs or exchanges spatial and temporal tokens. The experiment name is retained for consistency with the model configurations and reported results. The image and EHR encoders follow the architectures described in Sections 3.6.2 and 3.6.3, with model-specific training settings. Figure 3.10 illustrates the summary-level fusion path and the auxiliary outputs of the two branches.

The image encoder provides a pooled visual representation $v_i^{(I)} \in \mathbb{R}^{1280}$, obtained from its global and attention-weighted spatial pooling. The temporal EHR encoder produces $v_i^{(E)} \in \mathbb{R}^{256}$ by concatenating its classification-token representation and aggregated temporal summary. The two representations are independently projected to a common width of 96 using layer normalization, a linear transformation, and GELU activation:

$$p_i^{(I)} = \mathrm{GELU}\Big(W_I\,\mathrm{LN}(v_i^{(I)}) + b_I\Big),$$
$$p_i^{(E)} = \mathrm{GELU}\Big(W_E\,\mathrm{LN}(v_i^{(E)}) + b_E\Big),$$

where $W_I \in \mathbb{R}^{96\times1280}$ and $W_E \in \mathbb{R}^{96\times256}$. Both projected features are then normalized to unit Euclidean length, with $\varepsilon = 10^{-6}$ to avoid division by zero:

$$\bar{p}_i^{(m)} = \frac{p_i^{(m)}}{\max\Big(\|p_i^{(m)}\|_2, 10^{-6}\Big)}, \quad m \in \{I, E\}.$$

The normalized projections are concatenated to form $f_i^{(\mathrm{early})} = [\bar{p}_i^{(I)} \| \bar{p}_i^{(E)}] \in \mathbb{R}^{192}$. This vector passes through layer normalization, dropout (0.20), and a $192 \to 1$ linear classifier to produce the fusion logit. The equal projection widths and L2 normalization reduce differences in feature dimensionality and magnitude before classification. Unlike RespireTransFuse, this model does not condition image features on individual EHR time steps or EHR features on spatial image tokens.

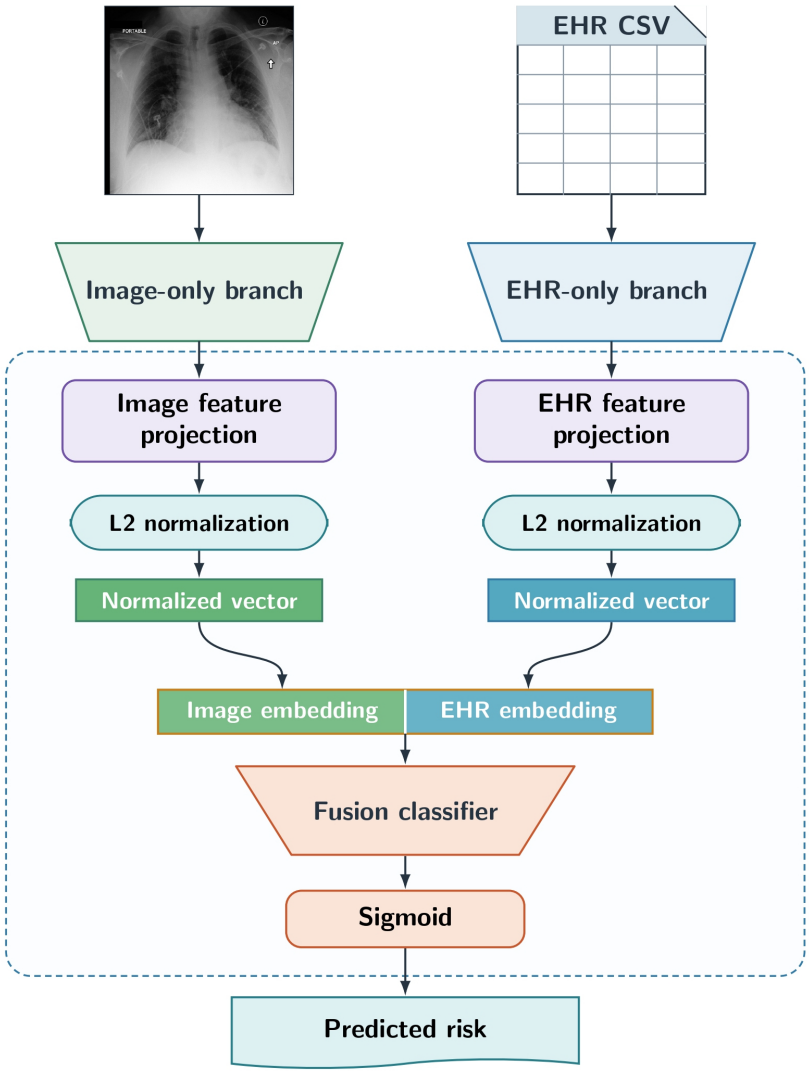


Figure 3.10: Early Fusion architecture. Independently encoded image and EHR summaries are projected to 96 dimensions, L2-normalized, concatenated, and classified. The two modality branches also produce auxiliary predictions; no token-level cross-attention is applied.

The image and EHR branches also produce separate prediction logits, which are used for auxiliary supervision and branch-level analysis rather than as inputs to the fusion classifier. The pretrained EfficientNet backbone remains frozen, while its spatial attention-pooling module and classification head are trainable. The EHR encoder and its prediction head, both 96-dimensional fusion projections, and the final fusion classifier are trainable; EHR adapters intended for token-level fusion remain frozen in this baseline.

With the 26-variable EHR input and its 26 observation-mask channels, the model contains 4,723,628 total parameters, including 696,784 trainable parameters. Of these, 151,297 trainable parameters belong to the two fusion projections and the final fusion classifier. The 96-dimensional projection setting therefore gives a 192-dimensional concatenated representation while preserving a comparatively compact fusion head.

3.6.5 RespireTransFuse Model

RespireTransFuse was designed to test whether a radiographic region becomes more informative when conditioned on the patient's recent clinical trajectory, and whether a particular EHR hour becomes more informative in the context of the contemporaneous radiograph. A compact 2×2 image-token grid and 48-dimensional common token width control attention cost; the two independently parameterized directions allow distinct clinical-to-visual and visual-to-clinical associations. Residual updates preserve existing features, while concatenating original summaries with cross-conditioned summaries retains direct unimodal information even if cross-attention is unhelpful for a sample. Removing cross-attention would reduce fusion to summary-level combination; removing a direction would eliminate one possible conditioning pathway; removing original summaries would force the classifier to rely entirely on cross-conditioned tokens. The first two changes are directly examined by the matched ablations; preservation of original summaries was not separately ablated.

RespireTransFuse preserved both complete branches, generated compact token sequences, applied one bidirectional cross-attention layer, and directly classified original plus cross-conditioned summaries. Branch logits were used only for auxiliary supervision and analysis. Figure 3.11 illustrates the proposed architecture and the two cross-attention directions.

Image spatial tokens and summary

The final EfficientNet map $A_i \in \mathbb{R}^{1280\times7\times7}$ was adaptively average-pooled to a 2×2 grid and rearranged into four 1280-dimensional spatial tokens. Each token was transformed by layer normalization, a $1280 \rightarrow 48$ linear layer, GELU, and dropout

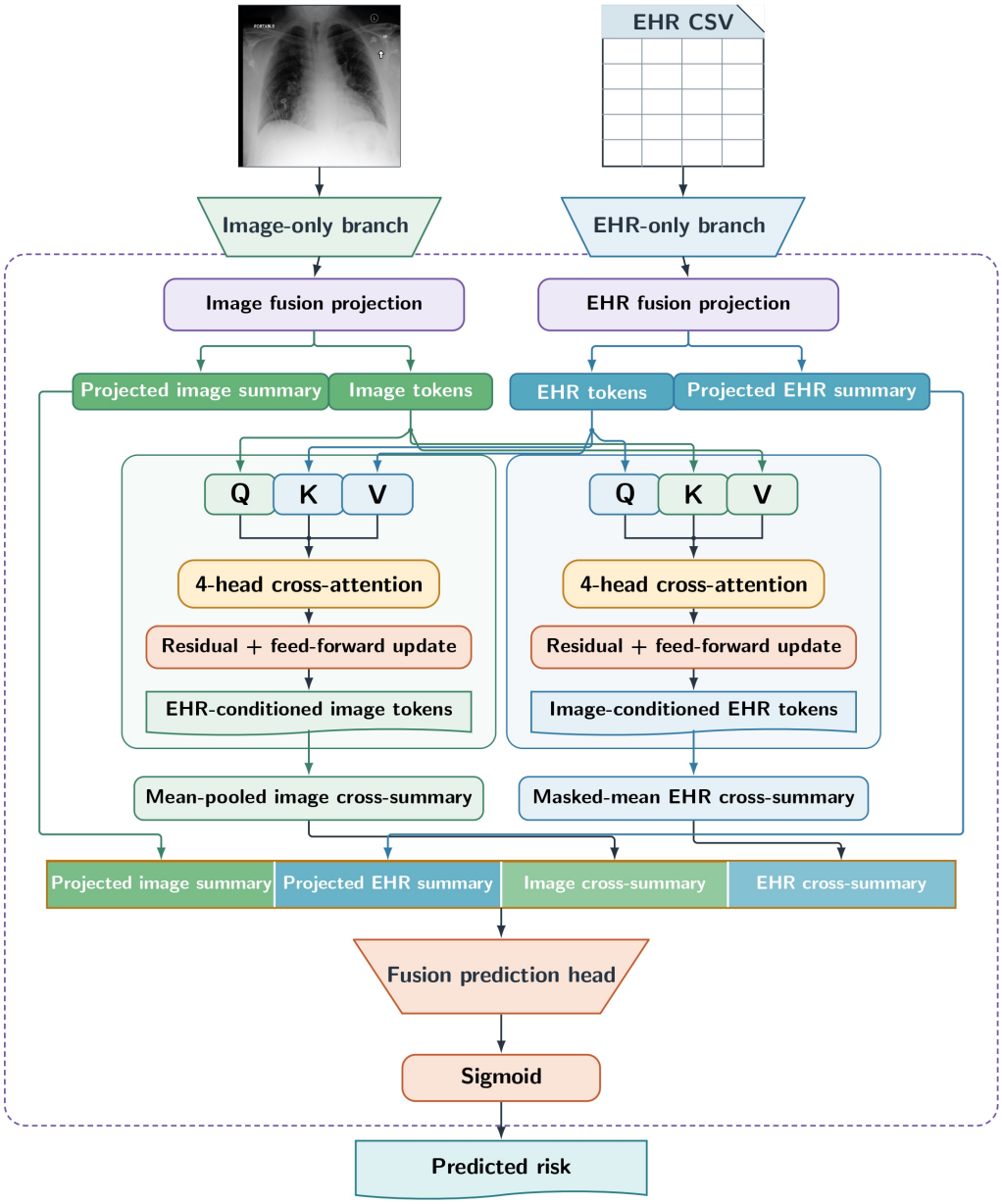


Figure 3.11: Overview of the current RespireTransFuse architecture. The image and EHR branches create original summaries and compact token sequences. Two separately parameterized attention directions condition image tokens on the EHR and EHR tokens on the image. The four 48-dimensional summaries form a 192-dimensional vector that is classified directly.

0.20:

$$S_i^{(I)} = \text{Dropout}\left(\text{GELU}\left(W_{IT}\,\text{LN}\left[\text{reshape}\left(\text{AAP}_{2\times2}(A_i)\right)\right] + b_{IT}\right)\right) \in \mathbb{R}^{4\times48}.$$

Adaptive pooling fixed four broad spatial tokens. In parallel, $v_i^{(I)}$ from Equation 3.3 was layer-normalized and projected $1280 \to 48$ to form $s_i^{(I)}$.

EHR temporal tokens and summary

The 24 contextual tokens from the EHR Transformer were passed through layer normalization, a $128 \to 48$ projection, GELU, and dropout to produce

$$S_i^{(E)} \in \mathbb{R}^{24\times48}.$$

The hour mask identified valid tokens. The mature EHR representation $v_i^{(E)}$ was independently normalized and projected $256 \to 48$ to form $s_i^{(E)}$.

Bidirectional cross-attention

The fusion module contained one reciprocal layer with two independently parameterized attention blocks. Both used width $d_f = 48$, four heads, and head dimension $d_k = 12$. In the first direction, the four image tokens were queries and the 24 EHR tokens were keys and values. For head h,

$$O_{i,h}^{I\leftarrow E} = \text{softmax}\left(\frac{Q_{i,h}^{I}(K_{i,h}^{E})^{\top}}{\sqrt{12}} + \mathcal{P}_i^{E}\right)V_{i,h}^{E},$$

where $\mathcal{P}_i^{E}$ masked completely unobserved EHR hours. In the reciprocal direction, the 24 EHR tokens queried the four image tokens:

$$O_i^{E\leftarrow I} = \text{MHA}\left(\text{LN}(S_i^{(E)}), \text{LN}(S_i^{(I)}), \text{LN}(S_i^{(I)})\right).$$

All four image tokens were valid. Separate parameters allowed the two query directions to learn different relations.

Each direction used a pre-normalized attention update and a pre-normalized feed-forward update. Both were added using a fixed scale $\rho = 0.40$:

$$\widehat{S} = S + \rho\,\text{Dropout}(O),$$
$$S' = \widehat{S} + \rho\,\text{Dropout}\left(W_2\,\text{Dropout}\left[\text{GELU}\left(W_1\,\text{LN}(\widehat{S}) + b_1\right)\right] + b_2\right).$$

The feed-forward network expanded $48 \to 96 \to 48$ with dropout 0.20. The residual scale moderated initial cross-modal updates.

Cross-conditioned pooling and direct classifier

After reciprocal interaction, the cross-conditioned image summary was the mean of four tokens:

$$c_i^{(I)} = \frac{1}{4}\sum_{p=1}^{4} C_{i,p}^{(I)}.$$

The cross-conditioned EHR summary was a masked temporal mean:

$$c_i^{(E)} = \frac{\sum_{t=1}^{24} m_{i,t} C_{i,t}^{(E)}}{\max\left(1, \sum_{t=1}^{24} m_{i,t}\right)}.$$

The final vector concatenated four 48-dimensional components in a fixed order:

$$f_i = [s_i^{(I)} \| s_i^{(E)} \| c_i^{(I)} \| c_i^{(E)}] \in \mathbb{R}^{192}.$$

The multimodal classifier applied layer normalization, dropout 0.20, a $192 \to 48$ linear layer, GELU, dropout, layer normalization at width 48, a $48 \to 24$ linear layer, GELU, dropout, and a final $24 \to 1$ layer:

$$z_i^{(F)} = h_F(f_i), \qquad \widehat{p}_i^{(F)} = \sigma(z_i^{(F)}).$$

Optional outputs included tokens, attention matrices, pooling weights, and branch logits for inspection.

3.7 Objective Functions

3.7.1 Binary cross-entropy

All primary thesis-developed predictions used binary cross-entropy with logits. For a mini-batch of size B, the unweighted loss was

$$\mathcal{L}_{\mathrm{BCE}}(z, y) = -\frac{1}{B}\sum_{i=1}^{B}\left[y_i \log \sigma(z_i) + (1 - y_i)\log(1 - \sigma(z_i))\right]. \tag{3.4}$$

The implementation used the numerically stable logit form rather than explicitly computing the logarithms in Equation 3.4. Although the shared training code supported positive-class weighting and additional ranking or approximation terms, the principal configurations set the effective positive weight to one and all additional terms to zero. Mini-batches were not class-balanced. Thus, the active objective remained ordinary BCE on the natural training distribution.

For Image-only and EHR-only,

$$\mathcal{L}_I = \mathcal{L}_{\mathrm{BCE}}(z^{(I)}, y), \qquad \mathcal{L}_E = \mathcal{L}_{\mathrm{BCE}}(z^{(E)}, y).$$

The natural class distribution was retained because AUPRC and precision depend on prevalence and because oversampling changes the optimization distribution. Class imbalance was addressed primarily by reporting prevalence-aware metrics and threshold-free discrimination rather than by altering sample frequency.

3.7.2 Multimodal auxiliary supervision

Early Fusion and RespireTransFuse returned a multimodal logit together with both branch logits. Their objective was

$$\mathcal{L}_{\text{multi}} = \lambda_F \mathcal{L}_{\text{BCE}}(z^{(F)}, y) + \lambda_E \mathcal{L}_{\text{BCE}}(z^{(E)}, y) + \lambda_I \mathcal{L}_{\text{BCE}}(z^{(I)}, y).$$

The direct multimodal output was the primary term with $\lambda_F = 1.00$. In RespireTransFuse, $\lambda_E = 0.05$ and $\lambda_I = 0.02$; Early Fusion used $\lambda_E = \lambda_I = 0.05$. The auxiliary terms encouraged both encoders to retain individually predictive representations during joint training without making either branch the mathematical basis of the final logit. At inference, only $z^{(F)}$ determined the multimodal probability. The MedFuse trainer used its original probability-space binary cross-entropy. I

3.8 Optimization and Training Protocol

3.8.1 Initialization and trainability

The models were implemented in PyTorch [50]. EfficientNet used generic pretrained weights; all task-specific heads, the EHR encoder, projections, and fusion modules began from their module initializations. The thesis-developed multimodal runs did not load completed unimodal checkpoints; the adapted MedFuse multimodal LSTM loaded its matching-seed Uni-EHR checkpoint and the available radiology-pretrained ResNet-34 checkpoint. EfficientNet remained frozen, while image pooling and classification, the EHR core, and the applicable projection and fusion layers were trainable.

3.8.2 AdamW and learning-rate schedules

The four thesis-developed models used AdamW, which decouples weight decay from the adaptive gradient update [44]. The three adapted MedFuse configurations used Adam [39]. The patient partition was fixed across the experiments; model training and comparison used seeds 40, 41, and 42. Table 3.4 lists the principal settings.

3.8.3 Epoch execution and data isolation

Training batches were shuffled only within the training partition, and augmentation was training-only. Validation and test loaders were deterministic and retained final

Table 3.4: Principal training settings for the seven evaluated models.

Model	Epochs	Batch	Optimizer
Image-only	20	32	AdamW
EHR-only	20	32	AdamW
Early Fusion	20	32	AdamW
RespireTransFuse	20	32	AdamW
MedFuse Uni-CXR	20	32	Adam
MedFuse Uni-EHR	20	32	Adam
MedFuse Multimodal LSTM	20	32	Adam

incomplete batches. Validation used evaluation mode and no gradients. Epoch records contained loss components, AUROC, AUPRC, probability scores and the checkpoint-selection decision. All reported models used a maximum of 20 epochs and early-stopping patience 10, with batch size 32. They retained model-specific learning rates, weight decay, dropout, and optimizers rather than imposing identical parameterizations on different architectures. The EHR dropout was 0.35; the Image-Only CNN used 0.35, Early Fusion used image/EHR/fusion dropouts of 0.30/0.30/0.20, and RespireTransFuse used 0.40/0.35/0.20. MedFuse Uni-CXR, Uni-EHR, and multimodal LSTM used dropouts of 0.30, 0.20, and 0.40, respectively.

3.9 Evaluation Measures

3.9.1 Discrimination and model selection

AUROC and AUPRC were the primary measures of discrimination. AUROC evaluates the ability of a model to assign higher scores to positive cases than to negative cases across all possible classification thresholds. AUPRC summarizes the relationship between precision and sensitivity and is particularly informative for imbalanced binary outcomes [61]. Because the no-skill AUPRC baseline is equal to the positive-class prevalence, AUPRC was interpreted relative to the 17.42% event prevalence (249/1,429) in the held-out test set.

Training and validation loss, AUROC, and AUPRC were recorded at every epoch for all seven models. Validation AUPRC was used as the principal checkpoint-selection measure, and the checkpoint with the strongest validation performance was retained for final evaluation. Validation labels were also used to select F1 operating thresholds and fit logistic recalibration where reported; all such choices were applied unchanged to test predictions. The test partition was used only after model development and checkpoint selection had been completed.

Final test-set discrimination was assessed using both receiver operating characteristic and precision–recall curves. The ROC curve plots the true-positive rate

against the false-positive rate, whereas the precision–recall curve plots precision against sensitivity. A random-ranking diagonal was included in the ROC analysis, and the test-set prevalence was included as the no-skill reference in the precision–recall analysis.

Bootstrap Uncertainty Estimation

Uncertainty in three-seed mean discrimination was estimated using 2,000 paired patient-cluster bootstrap resamples. Because patients could contribute more than one indexed radiograph, patients rather than individual encounters were sampled with replacement, retaining all of their test encounters together. The same sampled clusters were applied to each model and seed within a resample. AUROC and AUPRC were computed for each seed and averaged, and two-sided 95% percentile confidence intervals were obtained from the 2.5th and 97.5th percentiles of the resulting mean-score distribution.

For paired model comparisons, the AUROC and AUPRC of each comparator were subtracted from the corresponding RespireTransFuse metric within every bootstrap iteration. A paired difference interval excluding zero was interpreted as evidence that the observed difference was stable under patient-cluster resampling. These intervals were not adjusted for multiple comparisons.

Fixed-threshold operating-point uncertainty was estimated with patient-cluster resampling. The validation-selected threshold for each model and seed remained fixed throughout test resampling; threshold selection was not repeated on bootstrap test samples. Two-sided 95% percentile confidence intervals were obtained from the 2.5th and 97.5th percentiles. These were absolute intervals for each model rather than paired intervals for between-model differences.

In addition to paired bootstrap intervals, between-model discrimination differences were evaluated with 20,000 two-sided patient-level permutations and Holm correction across the six comparators and two primary metrics. Three-seed model means and sample standard deviations were reported separately from their patient-cluster bootstrap intervals. A numerical advantage or a confidence interval for one model alone was not treated as evidence of a reliable paired improvement.

3.9.2 Probability Quality

The Brier score measured the mean squared difference between the predicted probability and the observed binary outcome [9]:

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{p}_i - y_i \right)^2 .$$

Lower Brier-score values indicate better overall probability accuracy. The Brier score complements AUROC and AUPRC because a model can rank cases correctly

while still producing poorly scaled risk estimates.

3.9.3 Calibration assessment

Post-hoc logistic recalibration $z_i^{\text{cal}} = a + bz_i$ was fitted using validation logits and labels separately for each model and seed, then applied unchanged to test logits. Calibration was evaluated for all seven models using reliability curves constructed from ten probability bins. Within each bin, the mean predicted probability was compared with the observed deterioration rate. Perfect calibration corresponds to equality between these two quantities and was represented by the 45-degree reference line.

Expected calibration error (ECE) summarized the weighted absolute difference between predicted and observed risk across the bins [20]:

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \left| \widehat{p}_b - \widehat{o}_b \right|,$$

where $B = 10$, n_b is the number of samples in bin b, $\widehat{p}_b$ is the mean predicted probability, and $\widehat{o}_b$ is the observed event rate in that bin.

Importantly, AUROC and AUPRC measure ranking and do not directly measure the accuracy of predicted probabilities. Logistic recalibration with positive slope preserves ranking while changing the Brier score and ECE. Improved discrimination from additional Transformer or cross-attention layers therefore does not guarantee improved calibration; these properties were evaluated independently.

3.10 Ablation Analysis

The seven-model comparison evaluates the predictive value of each modality and contrasts distinct encoder and fusion families. The Image-Only CNN and EHR-Only Transformer establish unimodal references; Early Fusion combines summaries without spatial–temporal token interaction; and the three adapted MedFuse systems supply external convolutional, recurrent, and recurrent-fusion comparisons. Since these independently optimized models differ in several architectural and training choices, their comparison alone cannot isolate the effect of a particular attention component.

To address this limitation, the latest experiments also evaluated five matched RespireTransFuse configurations using the same paired inputs, patient partitions, and primary prediction objective: A0, summary concatenation without cross-attention; A1, image tokens querying EHR tokens only; A2, EHR tokens querying image tokens only; A3, reciprocal attention without auxiliary branch losses; and A4, the complete reciprocal-attention model with auxiliary supervision. Other architectural and training settings followed the full model. These comparisons probe the

removal of both attention directions, the importance of each direction individually, and the influence of preserving branch-level predictive supervision.

3.11 Implementation and Reproducibility Controls

The preprocessing, model-development, training, and evaluation pipelines were implemented in Python using PyTorch and scikit-learn [50, 51]. The project source code is maintained in a public GitHub repository, which provides the version-controlled implementation of the RespireTransFuse framework, its comparison models, configuration files, and supporting training and evaluation scripts.

Raw MIMIC-IV and MIMIC-CXR data were not redistributed through the GitHub repository because both datasets are restricted-access resources governed by PhysioNet credentialing, training, and data-use requirements [53, 54]. Reproduction therefore requires authorized access to the source datasets, after which the published preprocessing and training scripts can be used to reconstruct the analytic inputs and experimental workflow.

3.12 Methodological Scope and Limitations

The paired cohort does not represent every ICU encounter, both sources originate from one health system, and the procedure-derived outcome is not a prospectively adjudicated syndrome. Hourly bins simplify irregular events; a frozen visual backbone may under-adapt and reciprocal interaction may overfit. Patient-exclusive splitting, strict timing, training-only preprocessing, strong comparators, multi-seed evaluation, paired intervals, and matched ablations reduce but do not remove these limitations. The model comparisons provide indirect evidence of cross-modal complementarity, but the study does not directly quantify redundancy or sample-level disagreement between the EHR and radiographic information. EHR Transformer and recurrent baselines are not parameter-matched, and numerical fusion ablation advantages were not statistically established. External validation and inference with a completely absent modality were not studied. The study supports internal comparison under one retrospective protocol, not transportability, workflow benefit, or causality.

Chapter 4

Experimental Results

This chapter reports the final experiments on the paired MIMIC-IV–MIMIC-CXR cohort. The analysis covers cohort distribution, overall performance, individual training profiles, and comparative interpretation. It also presents three-seed evaluation, controlled architectural ablations, and representative radiograph/EHR trajectories from the final study.

4.1 Cohort Distribution

The final analytic cohort contained 9,509 paired observations from 4,668 patients, including 1,667 positive instances and 7,842 stable negatives. The patient-exclusive training, validation, and test partitions contained 6,691, 1,389, and 1,429 observations, respectively, with 1,175, 243, and 249 positives. Positive prevalence was 17.56%, 17.49%, and 17.42% across these partitions, and 17.53% overall. The held-out test observations came from 703 patients. Patient-level separation prevented observations from the same individual appearing in more than one partition. Table 4.1 summarizes the split counts and event prevalence, while Figure 4.1 illustrates the distributions.

Across the complete cohort, positive cases accounted for 17.53% of samples,

Table 4.1: Patient-exclusive cohort and positive-event prevalence.

Partition	**Total**	**Positive**	**Negative**	**Positive (%)**
Training	6,691	1,175	5,516	17.56
Validation	1,389	243	1,146	17.49
Test	1,429	249	1,180	17.42
Overall	9,509	1,667	7,842	17.53

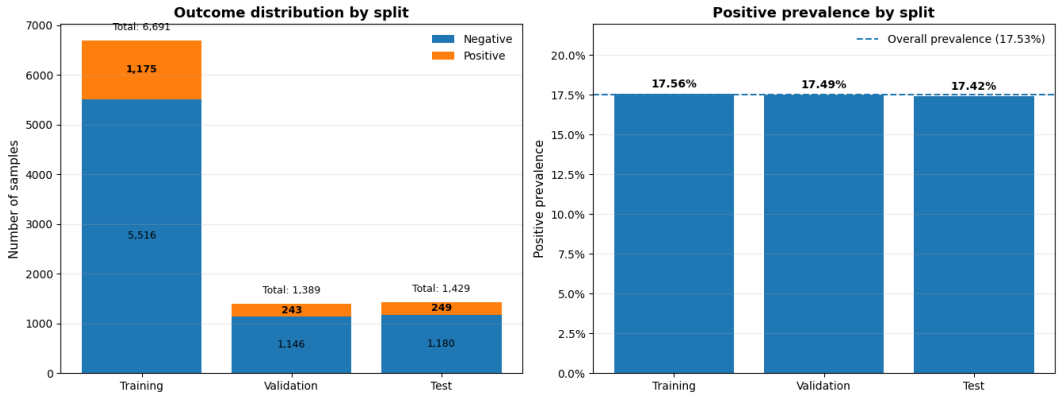


Figure 4.1: Patient-level training, validation, and test distributions in the final cohort. The left panel displays positive and negative sample counts; the right panel shows within-partition event prevalence. Patients do not overlap between partitions.

leaving approximately 4.70 negative observations per positive observation. Because respiratory-support events form a minority class, the study emphasizes AUROC and AUPRC rather than accuracy [61]. The no-skill precision baseline on the held-out test set is the observed event prevalence:

$$\text{AUPRC}_{\text{no-skill}} \approx \frac{249}{1429} = 0.1742.$$

AUPRC above this reference indicates better positive-case retrieval than a random ranking. The outcome is defined from the first qualifying respiratory-support procedure within 48 hours of an indexed CXR, whereas the negative class requires no event and complete follow-up through 72 hours.

Figure 4.2 complements the patient partition analysis by examining overall class composition and the time from the index CXR to the first qualifying intervention. Positive events are grouped into 0–6, 6–12, 12–24, and 24–48-hour intervals.

4.2 Overall Performance and Uncertainty

All seven models were trained and evaluated using the same patient-exclusive partitions, with independently seeded runs 40, 41, and 42. Table 4.2 presents mean AUROC and AUPRC, their across-seed sample standard deviations, and 95% patient-cluster bootstrap confidence intervals from 2,000 resamples. The standard deviation describes variation between fitted seeds, whereas the patient bootstrap addresses sampling uncertainty while retaining within-patient dependence. The model comparisons were additionally analyzed with 20,000 paired patient-level permutations and Holm adjustment across 12 tests (six comparators and two metrics).

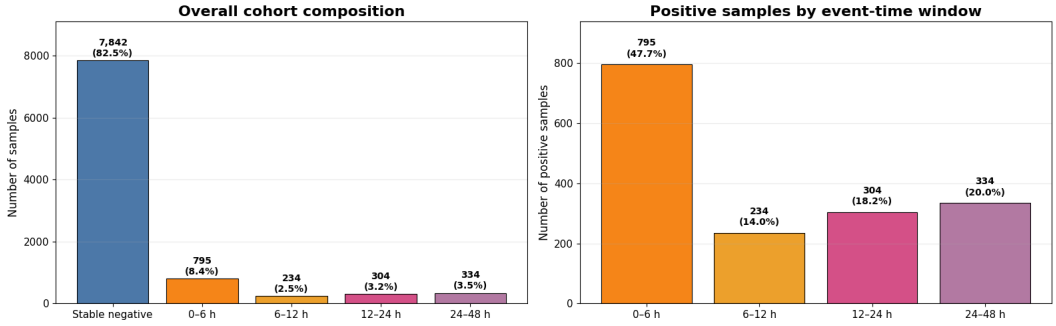


Figure 4.2: Class composition and positive-event timing in the final cohort. The left panel distinguishes 7,842 stable negatives from 1,667 positives; the right panel divides positive events into 0–6, 6–12, 12–24, and 24–48-hour intervals after the index CXR. The time-to-event groups describe the 48-hour outcome window.

The adjusted comparisons should not be inferred from the marginal confidence intervals shown in the main table.

RespireTransFuse achieved the highest observed mean AUROC (0.7134) and AUPRC (0.4071). The AUPRC was approximately 2.34 times the random-ranking reference of 0.1742. Across models, the image-only systems had the lowest discrimination, while EHR-containing models were generally stronger. Nonetheless, the EHR-only Transformer did not outperform the recurrent EHR comparator on every metric, and adding an image through recurrent MedFuse did not improve on its EHR-only version. These results caution against equating additional model complexity with uniform performance gains.

The results presented in Table 4.3 show that RespireTransFuse achieved higher three-seed mean AUROC and AUPRC values than all six comparator models. The largest improvements were observed relative to the image-only approaches, with AUROC and AUPRC differences of 0.1331 and 0.1563, respectively, compared with the Image-Only CNN. Smaller improvements were observed relative to the EHR-based and multimodal models, indicating that these approaches already captured substantial predictive information. Nevertheless, RespireTransFuse consistently achieved higher mean discrimination performance, suggesting that its multimodal fusion architecture may offer additional predictive benefits. The differences in this table are point estimates; the paired bootstrap intervals and adjusted permutation tests provide the corresponding inferential assessment.

The paired patient-level permutation analysis used 20,000 permutations and Holm adjustment across 12 comparisons. RespireTransFuse exceeded Image-Only CNN on AUROC and AUPRC (Holm-adjusted $p = 0.0006$ for each metric) and adapted MedFuse Uni-CXR on AUROC ($p = 0.0020$) and AUPRC ($p = 0.0027$). Its other four comparator comparisons were numerically favorable in both measures,

Table 4.2: Three-seed discrimination on the held-out test set. Each model entry reports mean ± sample standard deviation and the 95% patient-cluster bootstrap confidence interval in brackets. The models are ordered by mean AUROC.

Model	**AUROC ↑**	**AUPRC ↑**
Image-Only CNN	0.5803 ± 0.0032	0.2508 ± 0.0043
	[0.5366, 0.6238]	[0.2044, 0.3131]
Adapted MedFuse Uni-CXR	0.6085 ± 0.0006	0.2592 ± 0.0012
	[0.5683, 0.6496]	[0.2108, 0.3213]
Adapted MedFuse Multimodal LSTM	0.6733 ± 0.0057	0.3484 ± 0.0175
	[0.6253, 0.7192]	[0.2915, 0.4128]
Adapted MedFuse Uni-EHR	0.6820 ± 0.0063	0.3557 ± 0.0020
	[0.6348, 0.7272]	[0.2984, 0.4212]
EHR-Only Transformer	0.6922 ± 0.0031	0.3308 ± 0.0039
	[0.6420, 0.7364]	[0.2751, 0.3965]
Early Fusion	0.6969 ± 0.0085	0.3569 ± 0.0042
	[0.6552, 0.7384]	[0.2969, 0.4328]
RespireTransFuse	**0.7134 ± 0.0113**	**0.4071 ± 0.0164**
	[0.6715, 0.7547]	**[0.3400, 0.4830]**

but their Holm-adjusted p-values exceeded 0.05. The confidence intervals in Table 4.2 describe sampling uncertainty for individual model means and should not be interpreted as paired comparison tests.

4.3 Model-Specific Training Profiles

Training and validation loss, AUROC, and AUPRC provide complementary evidence about the convergence of the seven model families. Across-seed summaries use the three independently initialized runs, with standard-deviation bands calculated at each observed epoch. The reported test metrics below are three-seed summaries, not the measurements from any one curve. The calibration panels present the three-seed mean relationship between predicted probabilities and observed event rates, with shaded regions representing ±1 standard deviation across runs. The dashed diagonal indicates perfect calibration, while the Brier score and expected calibration error (ECE) provide quantitative measures of prediction reliability.

Table 4.3: Differences in three-seed mean AUROC and AUPRC between RespireTransFuse and the six comparator models. Positive values indicate higher performance for RespireTransFuse.

Comparator	ΔAUROC	ΔAUPRC
Image-Only CNN	+0.1331	+0.1563
Adapted MedFuse Uni-CXR	+0.1049	+0.1479
Adapted MedFuse Multimodal LSTM	+0.0401	+0.0587
Adapted MedFuse Uni-EHR	+0.0314	+0.0514
EHR-Only Transformer	+0.0212	+0.0763
Early Fusion	+0.0165	+0.0502

4.3.1 Image-Only CNN

The Image-Only CNN achieved a three-seed mean test AUROC of 0.5803 ± 0.0032 and AUPRC of 0.2508 ± 0.0043. Its AUPRC was approximately 1.44 times the no-skill baseline. The EfficientNet-B0 feature extractor was frozen, so training focused on the pooling and classification head. The radiograph-only model provided measurable discrimination, although its held-out performance was below that of the EHR-based and multimodal models.

Figure 4.3 shows rapid convergence during the initial epochs. Training loss decreased from approximately 0.60 to 0.47, while validation loss declined from approximately 0.52 to 0.46. Both losses stabilized after roughly eight epochs. Training and validation AUROC converged near 0.58, and their AUPRC curves converged near 0.23. The close alignment of the training and validation curves during the remaining epochs indicates stable convergence, without a sustained training–validation gap suggesting overfitting.

The calibration curve lies close to the perfect-calibration diagonal over the observed probability range, with some deviations in individual bins. The displayed Brier score of 0.142 and ECE of 0.008 summarize the model's probability prediction accuracy and average calibration discrepancy, respectively. These calibration results are distinct from the held-out AUROC and AUPRC summaries reported in Table 4.2.

4.3.2 Adapted MedFuse Uni-CXR

The adapted MedFuse Uni-CXR achieved a three-seed mean test AUROC of 0.6085 ± 0.0006 and AUPRC of 0.2592 ± 0.0012, exceeding the Image-Only CNN by 0.0282 AUROC and 0.0084 AUPRC. This model used a frozen, radiology-pretrained ResNet-34 feature extractor and optimized the binary classification head, so its learning curves reflect adaptation of the head rather than fine-tuning of the encoder.

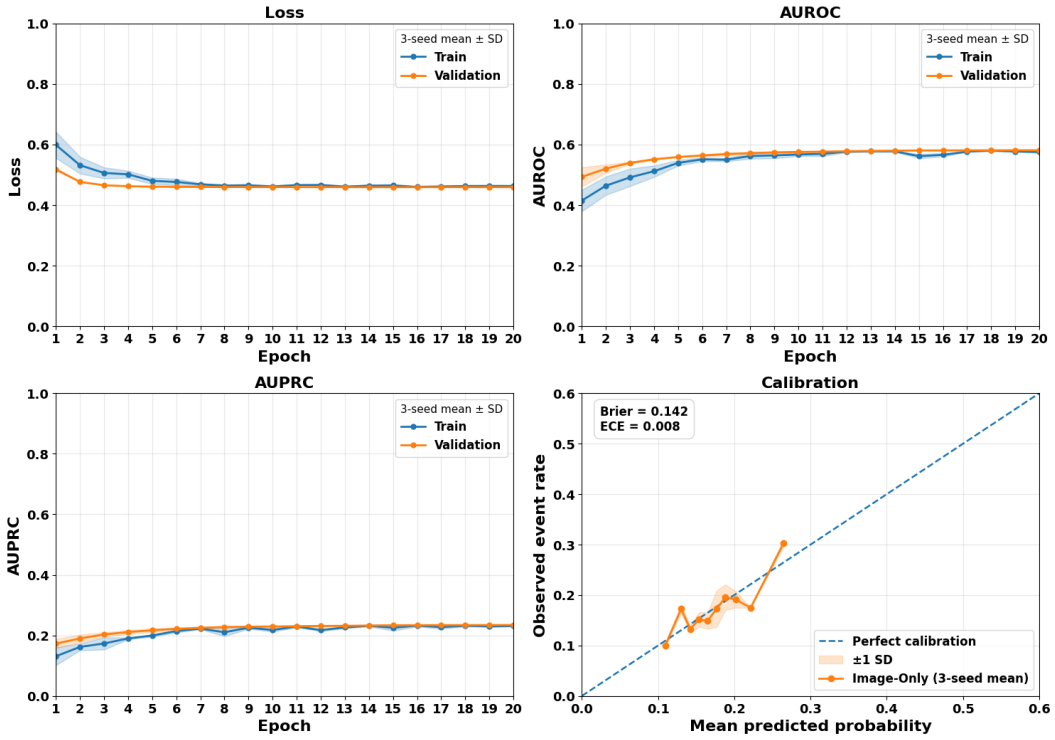


Figure 4.3: Image-Only CNN: three-seed mean training and validation loss, AUROC, and AUPRC across 20 epochs, with standard-deviation bands. The calibration panel shows the mean reliability curve, its across-seed variability, and the reported Brier score and ECE.

Figure 4.4 shows that training loss decreased from approximately 0.55 toward 0.46, converging with validation loss, which remained near 0.46. Training AUROC rose from approximately 0.50 to 0.61, approaching the validation AUROC near 0.61. Similarly, training and validation AUPRC converged near 0.25. The discrimination curves reached a stable plateau around epochs 10–12, with no sustained separation between training and validation performance. This pattern indicates convergence rather than progressive overfitting.

The calibration panel generally follows the perfect-calibration diagonal across the model's predicted probability range, although deviations remain in individual bins. Its Brier score was 0.140 and ECE was 0.010. Relative to the Image-Only CNN, the Brier score was slightly lower (0.140 versus 0.142), whereas ECE was slightly higher (0.010 versus 0.008); the calibration metrics therefore do not show a uniform improvement. The three-seed held-out discrimination results are summarized separately in Table 4.2.

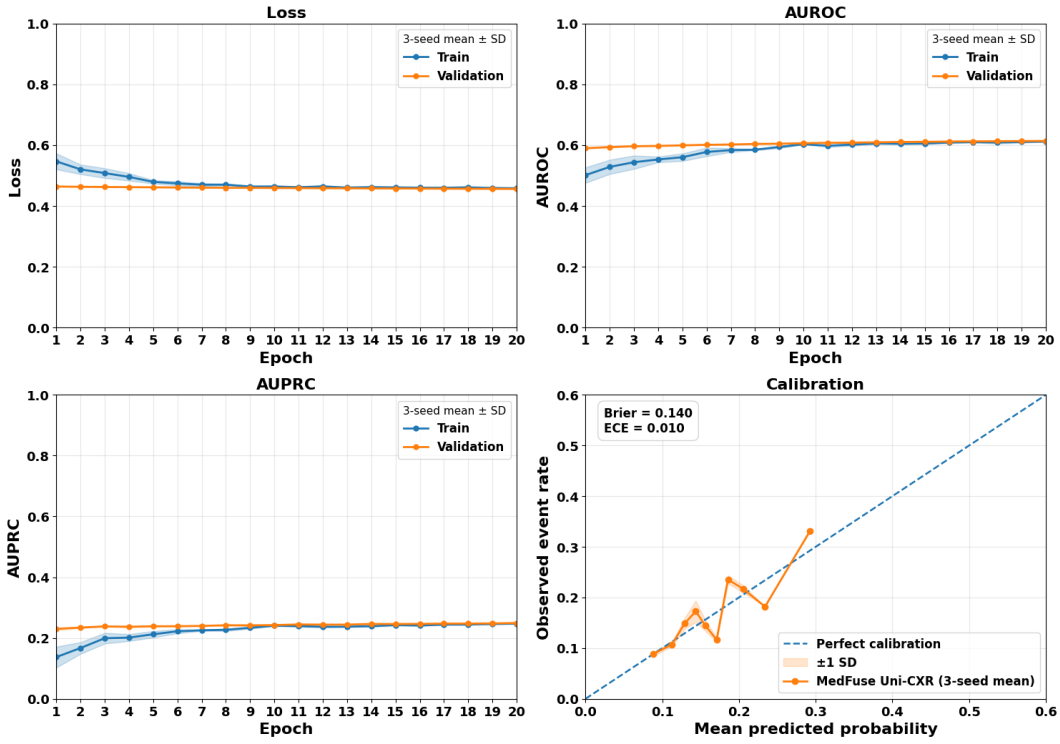


Figure 4.4: Adapted MedFuse Uni-CXR: three-seed mean training and validation loss, AUROC, and AUPRC across 20 epochs, with standard-deviation bands. The calibration panel shows the mean reliability curve, its across-seed variability, and the reported Brier score and ECE.

4.3.3 EHR-Only Transformer

The EHR-Only Transformer achieved a three-seed mean test AUROC of 0.6922 ± 0.0031 and AUPRC of 0.3308 ± 0.0039. The model processed 24 hours of mask-aware physiological measurements using multi-scale temporal convolutions and Transformer self-attention. Relative to the adapted MedFuse Uni-EHR model, it achieved a higher mean AUROC (0.6922 versus 0.6820), but a lower mean AUPRC (0.3308 versus 0.3557). This indicates stronger overall discrimination as measured by AUROC, whereas the recurrent baseline attained better precision–recall performance on the held-out test set.

Figure 4.5 shows that both discrimination metrics largely stabilized by epochs 7–10. The narrowing gaps and stable validation performance indicate convergence without a sustained divergence between training and validation curves.

The calibration curve follows the perfect-calibration diagonal reasonably closely over the model's observed prediction range, with deviations in some probability

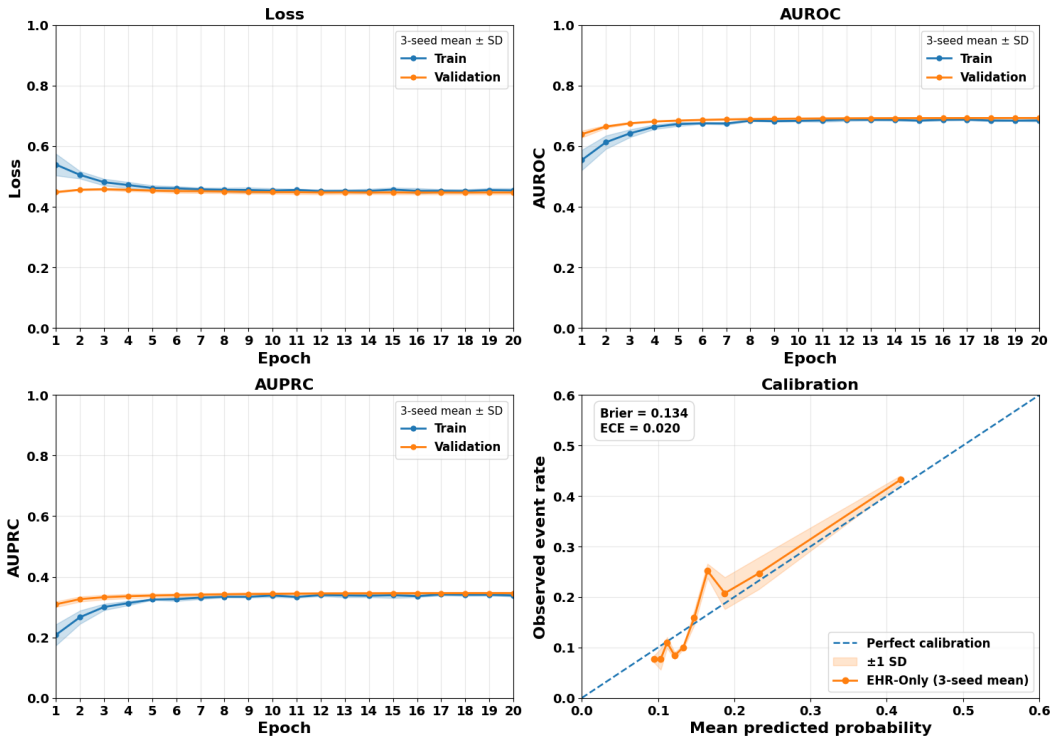


Figure 4.5: EHR-Only Transformer: three-seed mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands. The calibration panel presents the three-seed mean reliability curve, across-seed variability, Brier score, and expected calibration error (ECE).

bins. The figure reports a Brier score of 0.134 and an ECE of 0.020. These values describe probability prediction quality and calibration, respectively, and complement the three-seed test discrimination results reported in Table 4.2.

4.3.4 Adapted MedFuse Uni-EHR

The adapted MedFuse Uni-EHR model achieved a three-seed mean test AUROC of 0.6820 ± 0.0063 and AUPRC of 0.3557 ± 0.0020. Compared with the EHR-Only Transformer, its mean AUROC was lower by 0.0102, whereas its mean AUPRC was higher by 0.0249. Thus, the recurrent model produced better precision–recall performance, despite its lower overall AUROC. Because these models differ in architecture and training configuration, the comparison describes their observed performance rather than isolating the effect of recurrence versus self-attention.

Figure 4.6 shows a substantial reduction in loss during the first three epochs, followed by gradual convergence. Training AUROC increased from approximately

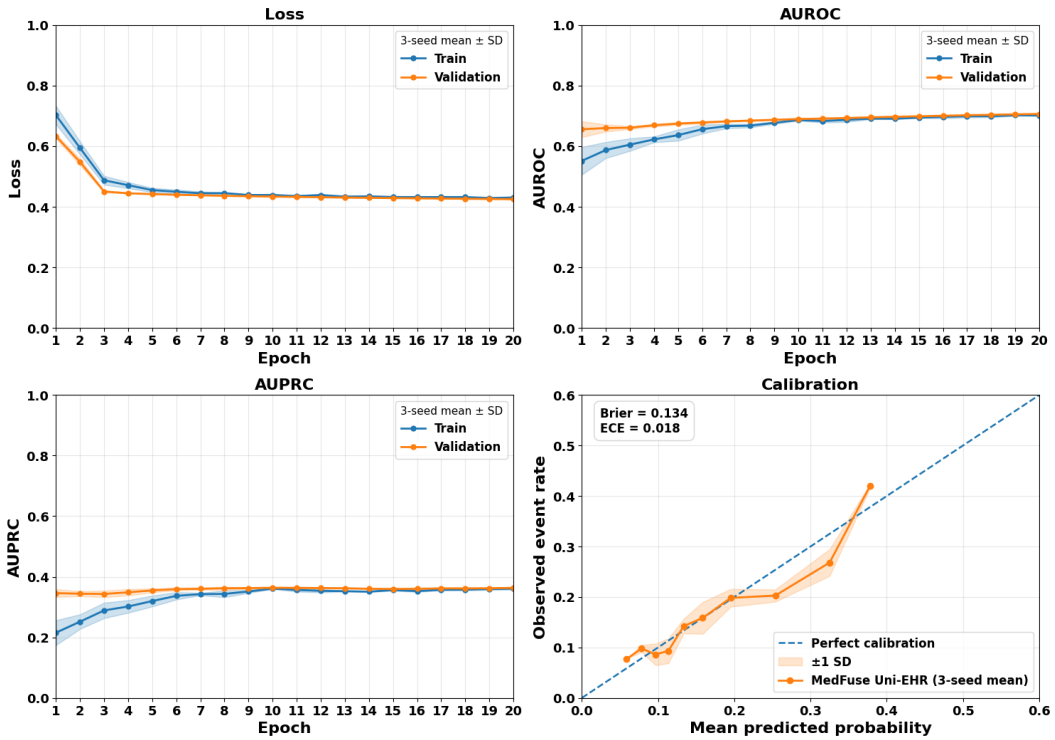


Figure 4.6: Adapted MedFuse Uni-EHR: three-seed mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands. The calibration panel presents the three-seed mean reliability curve, across-seed variability, Brier score, and ECE.

0.55 to 0.70, approaching validation AUROC near 0.71. Training AUPRC rose from approximately 0.22 to 0.36, converging toward validation AUPRC near 0.37. By approximately epochs 10–12, the training and validation curves had reached stable levels and remained closely aligned. The figure therefore shows stable convergence without a sustained deterioration in validation performance.

The calibration curve generally tracks the perfect-calibration diagonal, with visible departures in individual probability bins. Its reported Brier score was 0.134 and ECE was 0.018. The Brier score was identical to that of the EHR-Only Transformer at the displayed precision, while ECE was slightly lower (0.018 versus 0.020). These small descriptive differences do not, by themselves, establish a statistically meaningful calibration advantage. The held-out discrimination results are reported separately in Table 4.2.

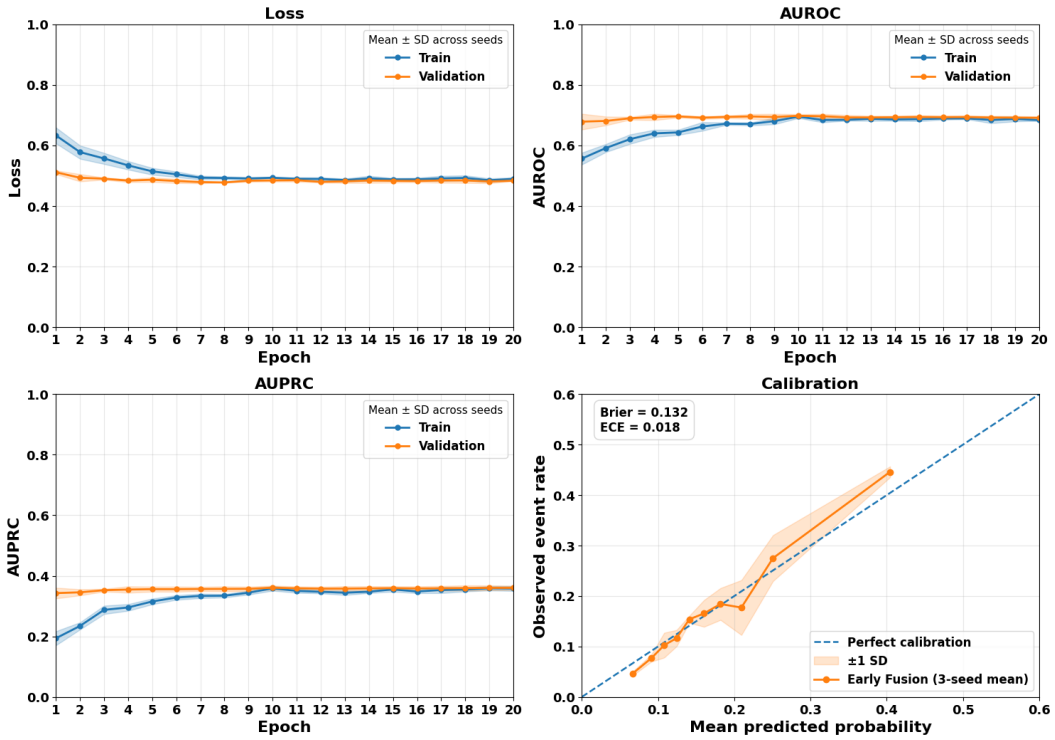


Figure 4.7: Early Fusion: three-seed mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands, alongside the three-seed mean calibration curve and its variability.

4.3.5 Early Fusion

Early Fusion achieved a three-seed mean test AUROC of 0.6969 ± 0.0085 and AUPRC of 0.3569 ± 0.0042, exceeding the corresponding EHR-Only Transformer means. It combines global image and EHR representations without token-level cross-modal attention, providing a competitive baseline for evaluating more structured fusion mechanisms.

Figure 4.7 shows that training loss declines during the initial epochs and approaches the validation loss. Training AUROC and AUPRC rise steadily toward their validation counterparts before both stabilize. The small gaps between the curves in later epochs indicate consistent convergence, with no sustained validation deterioration. The calibration curve generally follows the ideal diagonal, with localized deviations; the displayed Brier score and ECE are 0.132 and 0.018, respectively.

The stable training profile and competitive held-out performance establish Early

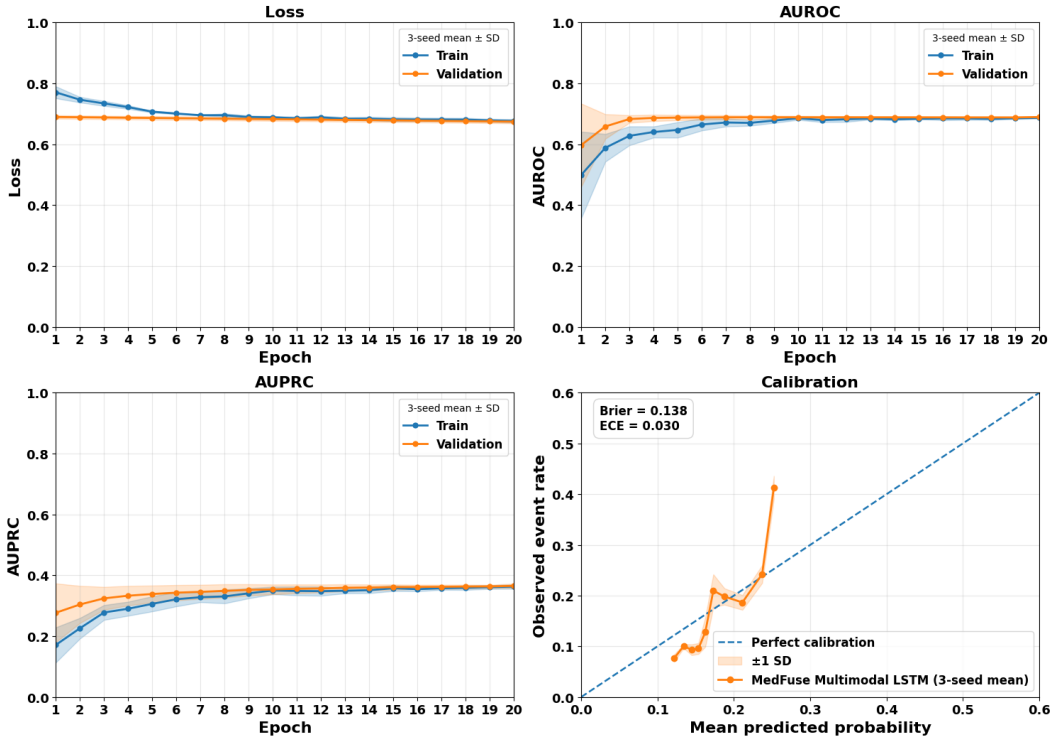


Figure 4.8: Adapted MedFuse Multimodal LSTM: three-seed mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands, alongside the three-seed mean calibration curve and its variability.

Fusion as a meaningful multimodal comparator. Its results demonstrate that combining image and EHR information can be effective even without explicit cross-attention.

4.3.6 Adapted MedFuse Multimodal LSTM

The adapted MedFuse Multimodal LSTM achieved a three-seed mean test AUROC of 0.6733 ± 0.0057 and AUPRC of 0.3484 ± 0.0175. Both were below the adapted MedFuse Uni-EHR means, indicating that the additional image information did not improve discrimination under this fusion configuration. Its comparatively larger AUPRC variation also indicates greater sensitivity to the training seed.

Figure 4.8 shows that the training loss progressively decreases and approaches validation loss, with both stabilizing during later epochs. Training and validation AUROC converge near 0.69, while AUPRC reaches approximately 0.36, indicating stable convergence without evident overfitting. The calibration curve remains relatively close to the ideal diagonal across lower and intermediate probabilities but

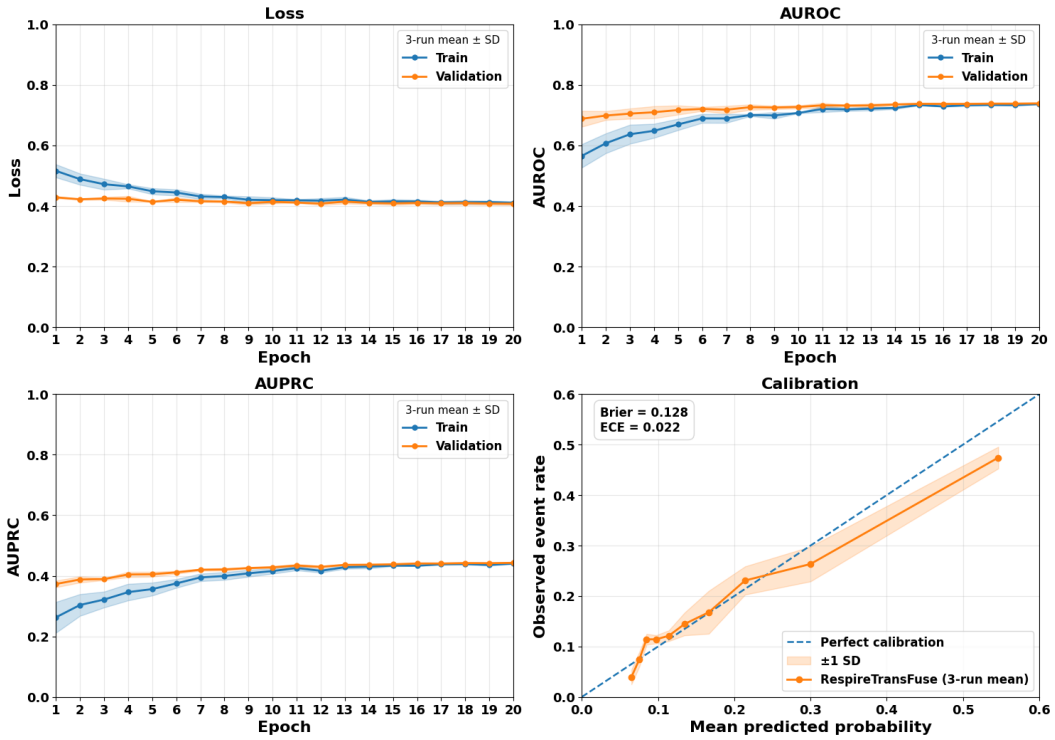


Figure 4.9: RespireTransFuse: three-run mean training and validation loss, AUROC, and AUPRC over 20 epochs, with standard-deviation bands, alongside the mean calibration curve and its variability.

shows a notable deviation in the highest-probability bin, where the model underestimates deterioration risk. The Brier score of 0.138 and ECE of 0.030 summarize its probability prediction accuracy and calibration discrepancy.

4.3.7 RespireTransFuse

RespireTransFuse achieved the highest three-seed mean test discrimination among the seven evaluated models, with AUROC 0.7134 ± 0.0113 and AUPRC 0.4071 ± 0.0164. Its bidirectional cross-attention mechanism integrates image and EHR tokens while retaining modality-specific summaries, enabling learned interactions between the two information sources.

Figure 4.9 shows consistent convergence across the three runs. Training loss decreases and aligns closely with validation loss, while both AUROC and AUPRC improve before stabilizing in the later epochs. There is no sustained divergence between training and validation discrimination in the displayed mean curves. The

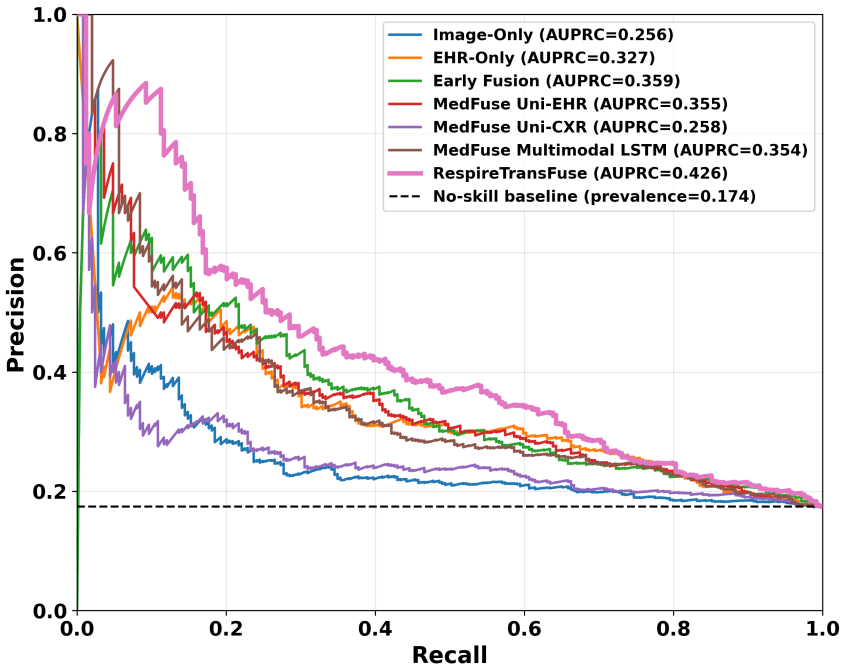


Figure 4.10: Precision–recall curves for the seven models on the held-out test set (seed 40). The dashed line indicates the positive-event prevalence (17.42%). Three-seed mean AUPRC values are reported in Table 4.2.

calibration panel shows a broadly consistent relationship between predicted probabilities and observed event rates, although some bins depart from the ideal diagonal. The displayed Brier score is 0.128, with ECE 0.022.

4.4 Comparative Interpretation

4.4.1 Test-Set Discrimination

Figures 4.10 and 4.11 compare the seven models on the held-out test set. The curves show predictions from seed 40, whereas Table 4.2 summarizes performance across seeds 40, 41, and 42. The distinction is important because the area displayed in each figure corresponds to one trained model, not the three-seed mean. In the precision–recall plot, the dashed horizontal line indicates the positive-event prevalence of 17.42%; in the ROC plot, the diagonal indicates random ranking.

The precision–recall curves highlight the difficulty of identifying the minority positive class. Both image-only approaches remain relatively close to the prevalence baseline over much of the recall range, whereas models incorporating EHR information generally achieve higher precision. RespireTransFuse maintains the

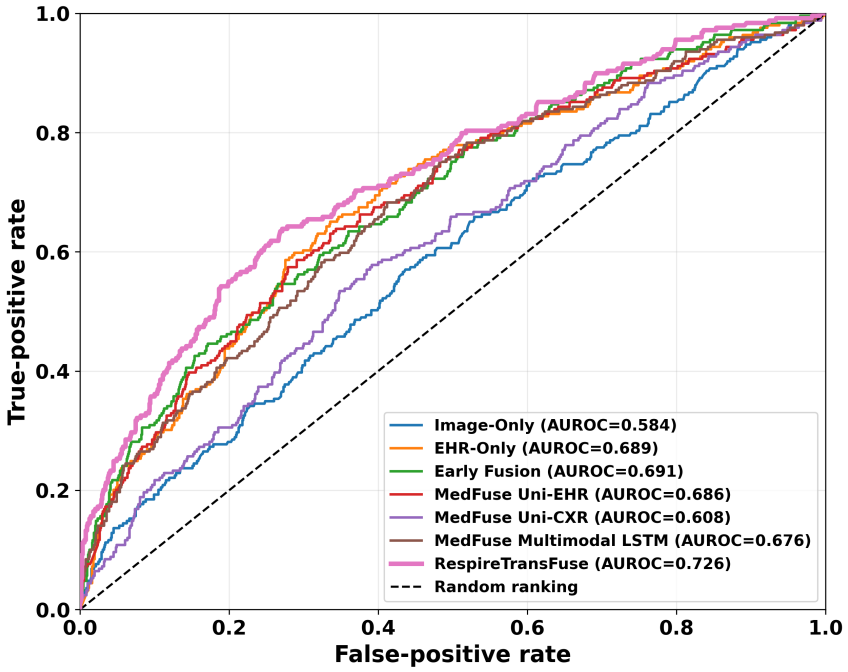


Figure 4.11: Receiver operating characteristic curves for the seven models on the held-out test set (seed 40). The dashed diagonal represents random ranking. Three-seed mean AUROC values are reported in Table 4.2.

strongest overall precision–recall performance, with a seed-40 AUPRC of 0.426. Its advantage is especially visible over the low-to-intermediate recall range, where identifying patients with high predicted risk is most relevant to prioritizing clinical review. This curve-level result is consistent with its leading three-seed mean AUPRC of 0.4071.

The ROC curves show a similar ordering. RespireTransFuse achieves the highest seed-40 AUROC (0.726) and generally provides a better trade-off between true-positive and false-positive rates than the comparison models. Across the three seeds, its mean AUROC is 0.7134, compared with 0.6969 for Early Fusion, the next-highest mean. RespireTransFuse also improves mean AUPRC from 0.3569 to 0.4071. Thus, relative to Early Fusion, the proposed model gains 0.0165 in AUROC and 0.0502 in AUPRC. The larger AUPRC difference is particularly relevant because positive outcomes represent only 17.42% of the test samples.

The comparison also shows that simply adding a second modality does not guarantee better discrimination. Early Fusion slightly exceeds the EHR-Only Transformer in both mean AUROC and AUPRC, while the adapted MedFuse Multimodal

Table 4.4: Total and trainable parameter counts for the seven evaluated models using the 26-variable EHR configuration. Adapted MedFuse modes share a registered model container; its total count does not represent the number of parameters used in every forward pass.

Model	Total Parameters	Trainable Parameters
Image-Only CNN	4,179,335	171,787
Adapted MedFuse Uni-CXR	23,838,532	513
Adapted MedFuse Uni-EHR	23,838,532	2,553,347
Early Fusion	4,723,628	696,784
EHR-Only Transformer	392,996	373,700
Adapted MedFuse Multimodal LSTM	23,838,532	1,709,314
RespireTransFuse	4,749,484	741,936

LSTM performs below its Uni-EHR counterpart on both metrics. The stronger results of RespireTransFuse therefore support investigating how the modalities interact, rather than treating the presence of two inputs as sufficient for an improvement.

4.4.2 Parameter Comparison

Model size provides useful context for interpreting the performance differences. Table 4.4 reports total registered and trainable parameter counts for the evaluated configurations, using 26 EHR variables and 26 corresponding observation-mask channels. Thus, the input projection receives 52 channels at each time step. Total parameters include weights registered within a model, while trainable parameters include only those updated during optimization. These counts differ substantially when pretrained components are frozen or when a model container includes components that are inactive in a particular evaluation mode.

The Image-Only CNN contains more than four million registered parameters, but only 171,787 are trainable because its pretrained visual backbone is frozen. The EHR-Only Transformer is considerably smaller, with 392,996 parameters in total. The adapted MedFuse variants share a larger registered parameter count, although their trainable counts differ according to the active branch and freezing configuration. These differences illustrate why the total number of parameters should not be interpreted as the computational cost of every model's forward pass.

Early Fusion contains 4,723,628 total parameters, of which 696,784 are trainable. RespireTransFuse contains 4,749,484 total and 741,936 trainable parameters, corresponding to increases of 25,856 and 45,152 parameters, respectively. The additional trainable capacity supports the token projections and reciprocal attention mechanism while keeping the overall model size close to that of Early Fusion. The

improvement in predictive performance is therefore obtained with a relatively modest increase in model size. Nevertheless, parameter count alone does not describe computational requirements or isolate the effect of an architectural component.

4.4.3 Architectural Comparisons

The unimodal comparisons establish the value of temporal clinical information for this prediction task. The EHR-Only Transformer achieves a mean AUROC of 0.6922 and AUPRC of 0.3308, exceeding the Image-Only CNN (0.5803 and 0.2508) on both measures. A similar difference appears within the adapted MedFuse models: Uni-EHR obtains an AUROC of 0.6820 and AUPRC of 0.3557, compared with 0.6085 and 0.2592 for Uni-CXR. In this cohort, physiological measurements collected before the radiograph are therefore more informative on their own than the radiographic image.

The choice of temporal encoder produces a more nuanced comparison. The EHR-Only Transformer has a higher AUROC than adapted MedFuse Uni-EHR (0.6922 versus 0.6820), while the recurrent EHR model has a higher AUPRC (0.3557 versus 0.3308). These results indicate that the two temporal representations have different discrimination characteristics; the comparison also includes differences in their training configurations and is not a controlled replacement of an LSTM by a Transformer.

Among multimodal approaches, Early Fusion improves on the EHR-Only Transformer, reaching an AUROC of 0.6969 and AUPRC of 0.3569. In contrast, adapted MedFuse Multimodal LSTM records 0.6733 and 0.3484, slightly below the adapted MedFuse Uni-EHR model. These outcomes demonstrate that successful fusion depends on the architecture and the information retained from each modality. A second input may add useful evidence, but it may also introduce redundant or difficult-to-optimize features.

RespireTransFuse provides the best mean results across both discrimination metrics. Unlike summary-level concatenation, it allows compact image-region tokens and hourly clinical tokens to condition one another before final pooling. It also retains the original image and EHR summaries, allowing the classifier to use modality-specific evidence alongside cross-conditioned information. The ablation experiments examine these design choices more directly by comparing concatenation, individual attention directions, reciprocal attention, and auxiliary supervision while retaining the same underlying modality branches.

4.4.4 Optimizer and Fusion Learning-Rate Comparison

The models were trained with architecture-specific optimization settings, summarized in Table 4.5. All experiments used a maximum of 20 epochs, batch size 32, and early-stopping patience of 10 epochs. AdamW was used for the Image-Only CNN, EHR-Only Transformer, Early Fusion, and RespireTransFuse; the adapted MedFuse

Table 4.5: Optimization settings for the seven evaluated models. All configurations used a maximum of 20 epochs, batch size 32, and early-stopping patience 10. Slash-separated entries refer to image-head, EHR, and fusion parameter groups.

Model	Optimizer	Learning Rate	Weight Decay	Dropout
Image-Only CNN	AdamW	2×10^{-5}	10^{-3}	0.35
EHR-Only Transformer	AdamW	4×10^{-4}	10^{-3}	0.35
Early Fusion	AdamW	$2 \times 10^{-4}/4 \times 10^{-5}/10^{-4}$	$10^{-4}/3 \times 10^{-4}/3 \times 10^{-4}$	0.30/0.30/0.20
MedFuse Uni-CXR	Adam	3×10^{-5}	3×10^{-4}	0.30
MedFuse Uni-EHR	Adam	3×10^{-4}	3×10^{-4}	0.20
MedFuse Multimodal LSTM	Adam	3×10^{-4}	3×10^{-4}	0.40
RespireTransFuse	AdamW	$10^{-4}/6 \times 10^{-4}/6 \times 10^{-4}$	$10^{-4}/3.5 \times 10^{-4}/2.5 \times 10^{-4}$	0.40/0.35/0.20

models used Adam. The multimodal architectures used separate learning rates and weight decay values for their image, EHR, and fusion parameter groups to accommodate differences in representation learning and fusion training.

The use of parameter-group learning rates permits the image, EHR, and fusion components to be updated at different rates, rather than imposing one step size on every trainable weight. The table documents the configurations used for the comparative experiments; it does not imply that a common optimizer or a single learning-rate setting was optimal for every architecture. Keeping these distinctions explicit helps interpret the performance results alongside the different model capacities and training strategies.

4.4.5 Matched Cross-Attention and Auxiliary-Supervision Ablations

To test the contribution of cross-modal interaction more directly, five configurations were evaluated on the same cohort and patient partitions with the full model's principal training choices. A0 used summary concatenation without cross-attention; A1 allowed image queries to attend to EHR tokens only; A2 allowed EHR queries to attend to image tokens only; A3 used both directions without auxiliary branch supervision; and A4 was the full reciprocal-attention model with auxiliary supervision. Table 4.6 reports their observed discrimination and patient-bootstrap intervals.

The complete configuration had the highest point estimates in both metrics. However, A1 did not exceed A0 in AUROC, showing that adding a single attention direction need not improve performance. A2 exceeded A1 on both measures, suggesting that the direction of cross-conditioning mattered in these fitted models. A3 achieved higher AUPRC than A2 but slightly lower AUROC, so reciprocal attention without auxiliary losses was not uniformly advantageous. Relative to A3, adding auxiliary supervision in A4 increased AUROC by 0.0106 and AUPRC by 0.0058 in the point estimates.

Table 4.6: Matched fusion ablations (A0–A4). Intervals are 95% patient-cluster bootstrap confidence intervals. "Aux." indicates the presence of branch-specific auxiliary losses.

Variant	Cross-attention	Aux.	AUROC	AUPRC
A0	None (concatenation)	Yes	0.6966	0.3679
			[0.6547, 0.7379]	[0.3007, 0.4437]
A1	Image → EHR	Yes	0.6900	0.3693
			[0.6481, 0.7313]	[0.3021, 0.4451]
A2	EHR → Image	Yes	0.7038	0.3815
			[0.6619, 0.7451]	[0.3144, 0.4574]
A3	Both directions	No	0.7028	0.4013
			[0.6579, 0.7409]	[0.3324, 0.4751]
A4 (Full)	Both directions	Yes	**0.7134**	**0.4071**
			[0.6715, 0.7547]	**[0.3400, 0.4830]**

Table 4.7: Paired ablation differences from the full A4 model. Each interval is a 95% paired patient-cluster bootstrap confidence interval. An interval including zero does not establish a statistically reliable improvement.

Comparison	ΔAUROC [95% CI]	ΔAUPRC [95% CI]
A4 – A0	+0.0168 [−0.0072, 0.0425]	+0.0392 [−0.0087, 0.0887]
A4 – A1	+0.0234 [−0.0027, 0.0510]	+0.0378 [−0.0076, 0.0749]
A4 – A2	+0.0096 [−0.0152, 0.0359]	+0.0256 [−0.0219, 0.0717]
A4 – A3	+0.0106 [−0.0152, 0.0375]	+0.0058 [−0.0380, 0.0488]

All four point-estimate comparisons favored A4, but each paired 95% interval in Table 4.7 includes zero. The ablations are therefore consistent with an advantage for the full configuration but do not establish that any individual attention direction or auxiliary-loss component yields a statistically reliable gain. This distinction limits claims about the necessity of each component while clarifying the behavior of the investigated alternatives.

4.4.6 Qualitative Prediction Examples and EHR Trajectories

Figure 4.12 presents four representative held-out seed-40 observations: a true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Cases were selected from distinct patients, with attention to differences between fusion and

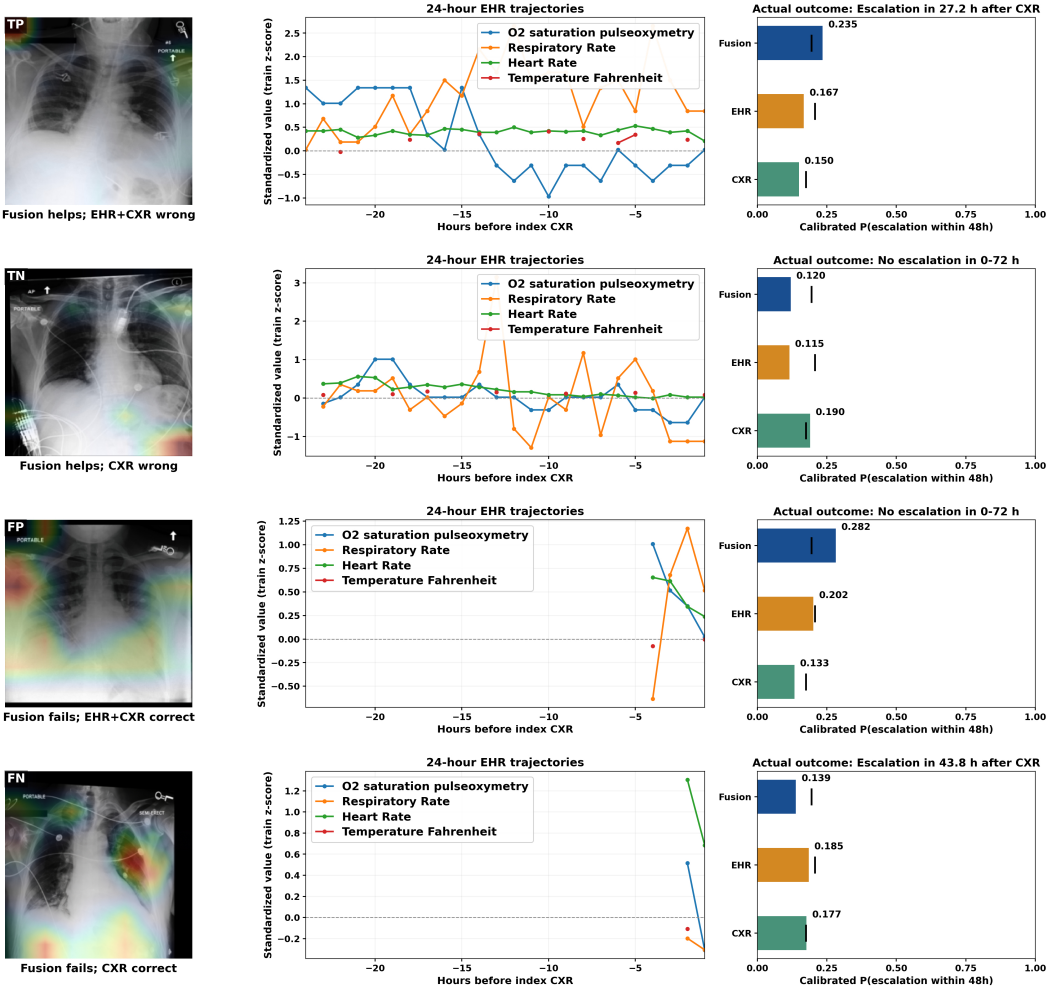


Figure 4.12: Representative held-out test predictions (seed 40): TP, TN, FP, and FN examples. Each row combines the indexed chest radiograph with fusion-output attribution, selected preceding EHR trajectories, and calibrated risk scores for fusion, EHR-only, and CXR-only models. Dashed or sparse clinical traces reflect available observations; the panels illustrate model agreement and disagreement rather than causal explanations.

unimodal correctness when feasible. The image panels show fusion-output Grad-CAM attribution conditional on both CXR and EHR inputs, the middle panels show selected standardized EHR trajectories over the preceding 24 hours, and the right panels compare calibrated model probabilities. Values in the image heatmaps were normalized independently; they are qualitative attribution illustrations, not quantitative proof that a particular region caused a prediction.

The TP and TN examples illustrate situations in which fusion can make a correct classification despite an error in one or both unimodal predictions. The FP and FN examples show that adding a second modality can also introduce or fail to correct mistakes. Uneven temporal measurement density is visible in the EHR trajectories, particularly for laboratory-type observations. This analysis complements the aggregate discrimination results, but a small purposively selected set of cases cannot establish general interpretability or subgroup performance.

4.4.7 Limitations of the Internal Comparison

Several factors constrain the interpretation of the observed gains. First, the endpoint is a recorded procedure and may not reliably distinguish newly initiated from continuing respiratory support. An audit identified 661 included samples from 301 patients with a qualifying procedure in an earlier ICU stay of the same hospitalization; prior-procedure screening for inclusion was limited to the index ICU stay. Second, selection into the paired cohort depends on obtaining a frontal radiograph and having sufficient preceding clinical observations; those clinical recording processes are not random. Third, a negative label requires observation through 72 hours, excluding some patients discharged or dying earlier without a documented positive procedure. Finally, the models were evaluated internally on data from the MIMIC environment and required both modalities at inference; generalization to other institutions and missing-modality scenarios was not tested.

Taken together, the final experiments support a numerical discrimination advantage for RespireTransFuse within the evaluated seven-model set and show that the choice and direction of cross-modal interaction affect point estimates. They do not establish clinical deployability, causal interpretation of attention weights, or statistically reliable gains from any single fusion component. External and temporally separated validation, additional subgroup analysis, and prospective evaluation remain necessary.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis developed and internally evaluated RespireTransFuse for prediction of recorded respiratory-support procedures using an index chest radiograph and the preceding 24 hours of electronic health record measurements. The temporally defined cohort comprised 9,509 paired samples from 4,668 patients. Seven models were evaluated on patient-exclusive partitions across seeds 40, 41, and 42: Image-Only CNN, EHR-Only Transformer, Early Fusion, adapted MedFuse Uni-CXR, adapted MedFuse Uni-EHR, adapted MedFuse Multimodal LSTM, and RespireTransFuse.

Temporal EHR information provided a stronger standalone predictive signal than chest radiography. The EHR-Only Transformer obtained mean AUROC 0.6922 and AUPRC 0.3308, compared with 0.5803 and 0.2508 for the Image-Only CNN. The adapted MedFuse Uni-EHR model likewise outperformed its Uni-CXR counterpart in both metrics. These findings indicate that recent physiological measurements are particularly informative for the study-defined procedure outcome, although radiographic information can add predictive value when integrated effectively.

RespireTransFuse achieved the highest observed mean test AUROC (0.7134) and AUPRC (0.4071) among the seven evaluated systems. Relative to Early Fusion, its mean improvements were 0.0165 AUROC and 0.0502 AUPRC. The adapted MedFuse multimodal LSTM, by contrast, remained below its Uni-EHR comparator on both measures, showing that access to two modalities does not automatically improve discrimination. Paired patient-level permutation tests with Holm correction supported differences between RespireTransFuse and the Image-Only CNN and adapted MedFuse Uni-CXR, while the differences from the stronger EHR and multimodal comparators did not meet the multiplicity-adjusted threshold.

The architectural ablations also favored the full reciprocal cross-attention configuration in observed AUROC and AUPRC. Relative to summary concatenation (A0), the complete model improved AUROC by 0.0168 and AUPRC by 0.0392; however,

the paired ablation confidence intervals contained zero. These experiments support the complete design's numerical advantage while leaving uncertainty about the separate effects of individual attention directions and auxiliary supervision.

Validation-based logistic recalibration produced a mean test Brier score of 0.1282 and ten-bin expected calibration error of 0.0220 for RespireTransFuse. These results complement the discrimination measures by characterizing overall probability error and calibration discrepancy. The endpoint remains based on documented respiratory-support procedures rather than independently confirmed initiation of respiratory failure, and the analysis is restricted to a paired retrospective cohort. External, temporal, and prospective evaluation, together with a more specific characterization of respiratory-support state, remains necessary before clinical application.

5.2 Future Work

Several extensions are required before RespireTransFuse can be considered for clinical evaluation beyond the present retrospective study. The most important next step is external validation using paired chest-radiograph and temporal EHR data from hospitals that were not represented during model development. Such evaluation would determine whether the observed performance transfers across differences in patient populations, imaging devices, clinical documentation, measurement frequency, treatment practices, and respiratory-support protocols. Temporal validation using a later patient cohort from the same institution would provide an additional assessment of robustness to changes in clinical practice and data collection over time.

Probability calibration should be investigated more systematically. Although RespireTransFuse achieved the strongest discrimination, its raw outputs should not automatically be interpreted as absolute clinical risks.

Finally, prospective evaluation would be required before deployment. Such work should examine inference latency, data availability, integration with hospital information systems, alert frequency, model drift, and clinician response. A silent prospective study could first generate predictions without influencing care, allowing assessment of real-time data quality and operational reliability. Only after successful external, temporal, calibration, subgroup, and prospective validation should the effect of RespireTransFuse on clinical decisions and patient outcomes be studied.

Bibliography

[1] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, *Multimodal Biomedical AI*, Nature Medicine, 28 (2022), pp. 1773–1784.

[2] D. Agniel, I. S. Kohane, and G. M. Weber, *Biases in Electronic Health Record Data Due to Processes Within the Healthcare System: Retrospective Observational Study*, BMJ, 361 (2018), p. k1479.

[3] ARDS Definition Task Force, V. M. Ranieri, G. D. Rubenfeld, B. T. Thompson, N. D. Ferguson, E. Caldwell, E. Fan, L. Camporota, and A. S. Slutsky, *Acute Respiratory Distress Syndrome: The Berlin Definition*, JAMA, 307 (2012), pp. 2526–2533.

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton, *Layer Normalization*, 2016. Preprint.

[5] D. Bahdanau, K. Cho, and Y. Bengio, *Neural Machine Translation by Jointly Learning to Align and Translate*, in Proceedings of the 3rd International Conference on Learning Representations, 2015.

[6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, *Multimodal Machine Learning: A Survey and Taxonomy*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41 (2019), pp. 423–443.

[7] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, *Patient Subtyping via Time-Aware LSTM Networks*, in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 65–74.

[8] G. Bellani, J. G. Laffey, T. Pham, E. Fan, L. Brochard, A. Esteban, L. Gattinoni, F. van Haren, A. Larsson, D. F. McAuley, M. Ranieri, G. Rubenfeld, T. Thompson, H. Wrigge, A. S. Slutsky, and A. Pesenti, *Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries*, JAMA, 315 (2016), pp. 788–800.

[9] G. W. Brier, *Verification of Forecasts Expressed in Terms of Probability*, Monthly Weather Review, 78 (1950), pp. 1–3.

[10] E. Çalli, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, *Deep Learning for Chest X-Ray Analysis: A Survey*, Medical Image Analysis, 72 (2021), p. 102125.

[11] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, *Recurrent Neural Networks for Multivariate Time Series with Missing Values*, Scientific Reports, 8 (2018), p. 6085.

[12] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. K. Williamson, T. Manz, M. Shady, and F. Mahmood, *Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3995–4005.

[13] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. F. Stewart, *RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism*, in Advances in Neural Information Processing Systems, vol. 29, 2016, pp. 3504–3512.

[14] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, *A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models*, Journal of Clinical Epidemiology, 110 (2019), pp. 12–22.

[15] M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson, *Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards*, Critical Care Medicine, 44 (2016), pp. 368–374.

[16] 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, A.-L. Boulesteix, J. C. Camaradou, L. A. Celi, S. Denaxas, A. K. Denniston, B. Glocker, R. M. Golub, H. Harvey, G. Heinze, M. M. Hoffman, A. P. Kengne, E. Lam, N. Lee, V. López-Miranda, L. Maier-Hein, B. A. Mateen, M. D. McCradden, L. Oakden-Rayner, J. Ordish, Z. Parnianpour, S. Rose, K. Singh, L. Wynants, and P. Logullo, *TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods*, BMJ, 385 (2024), p. e078378.

[17] C. Cui, H. Yang, Y. Wang, S. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson, B. A. Landman, and Y. Huo, *Deep Multimodal Fusion of Image and Non-Image Data in Disease Diagnosis and Prognosis: A Review*, Progress in Biomedical Engineering, 5 (2023), p. 022001.

[18] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, *AI for Radiographic COVID-19 Detection Selects Shortcuts over Signal*, Nature Machine Intelligence, 3 (2021), pp. 610–619.

[19] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, *PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals*, Circulation, 101 (2000), pp. e215–e220.

[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, *On Calibration of Modern Neural Networks*, in Proceedings of the 34th International Conference on Machine Learning, vol. 70 of Proceedings of Machine Learning Research, 2017, pp. 1321–1330.

[21] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, *Multitask Learning and Benchmarking with Clinical Time Series Data*, Scientific Data, 6 (2019), p. 96.

[22] N. Hayat, K. J. Geras, and F. E. Shamout, *MedFuse: Multi-Modal Fusion with Clinical Time-Series Data and Chest X-Ray Images*, in Proceedings of the 7th Machine Learning for Healthcare Conference, vol. 182 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 479–503.

[23] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[24] D. Hendrycks and K. Gimpel, *Gaussian Error Linear Units (GELUs)*, 2016. Preprint.

[25] J. Hessel and L. Lee, *Does My Multimodal Model Learn Cross-Modal Interactions? It's Harder to Tell Than You Might Think!*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 861–877.

[26] S. Hochreiter and J. Schmidhuber, *Long Short-Term Memory*, Neural Computation, 9 (1997), pp. 1735–1780.

[27] C. Hsieh, I. B. Nobre, S. C. Sousa, C. Ouyang, M. Brereton, J. C. Nascimento, J. Jorge, and C. Moreira, *MDF-Net for Abnormality Detection by Fusing X-Rays with Clinical Data*, Scientific Reports, 13 (2023), p. 15873.

[28] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely Connected Convolutional Networks*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[29] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, *Fusion of Medical Imaging and Electronic Health Records Using Deep Learning: A Systematic Review and Implementation Guidelines*, npj Digital Medicine, 3 (2020), p. 136.

[30] M. Hüser, M. Faltys, X. Lyu, C. Barber, S. L. Hyland, T. M. Merz, and G. Rätsch, *Early Prediction of Respiratory Failure in the Intensive Care Unit*, 2021. Preprint.

[31] S. L. Hyland, M. Faltys, M. Hüser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn, M. Moor, B. Rieck, M. Zimmermann, D. Bodenham, K. M. Borgwardt, G. Rätsch, and T. M. Merz, *Early Prediction of Circulatory Failure in the Intensive Care Unit Using Machine Learning*, Nature Medicine, 26 (2020), pp. 364–373.

[32] 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, and A. Y. Ng, *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 590–597.

[33] S. Jabbour, D. Fouhey, E. A. Kazerooni, J. Wiens, and M. W. Sjoding, *Combining Chest X-Rays and Electronic Health Record Data Using Machine Learning to Diagnose Acute Respiratory Failure*, Journal of the American Medical Informatics Association, 29 (2022), pp. 1060–1068.

[34] S. Jain and B. C. Wallace, *Attention Is Not Explanation*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 3543–3556.

[35] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, *MIMIC-IV, a Freely Accessible Electronic Health Record Dataset*, Scientific Data, 10 (2023), p. 1.

[36] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, *MIMIC-CXR, a De-identified Publicly Available Database of Chest Radiographs with Free-Text Reports*, Scientific Data, 6 (2019), p. 317.

[37] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, *MIMIC-CXR-JPG, a Large Publicly Available Database of Labeled Chest Radiographs*, 2019. Preprint.

[38] F. Khader, G. Müller-Franzes, T. Wang, T. Han, S. Tayebi Arasteh, C. Haarburger, J. Stegmaier, K. Bressem, C. Kuhl, S. Nebelung, J. N. Kather, and D. Truhn, *Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters: A Case for Transformers*, Radiology, 309 (2023), p. e230806.

[39] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, in International Conference on Learning Representations, 2015.

[40] M. Lagina and T. S. Valley, *Diagnosis and Management of Acute Respiratory Failure*, Critical Care Clinics, 40 (2024), pp. 235–253.

[41] C.-K. Lee, T.-L. Chen, J.-E. Wu, M.-T. Liao, C. Wang, W. Wang, and C.-Y. Chou, *Multimodal Deep Learning Models Utilizing Chest X-Ray and Electronic Health Record Data for Predictive Screening of Acute Heart Failure in Emergency Department*, Computer Methods and Programs in Biomedicine, 255 (2024), p. 108357.

[42] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi, *BEHRT: Transformer for Electronic Health Records*, Scientific Reports, 10 (2020), p. 7155.

[43] M. Lin, S. Wang, Y. Ding, L. Zhao, F. Wang, and Y. Peng, *An Empirical Study of Using Radiology Reports and Images to Improve Intensive Care Unit Mortality Prediction*, JAMIA Open, 8 (2025), p. ooae137.

[44] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, in International Conference on Learning Representations, 2019.

[45] J. Lu, D. Batra, D. Parikh, and S. Lee, *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*, in Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 13–23.

[46] J. Lu, J. Yang, D. Batra, and D. Parikh, *Hierarchical Question-Image Co-Attention for Visual Question Answering*, in Advances in Neural Information Processing Systems, vol. 29, 2016, pp. 289–297.

[47] X. Lu and S. Nemati, *Prospective Evaluation of Multimodal Respiratory Failure Prediction: Do Chest X-Rays Improve Performance Beyond EHR Signals?*, 2026. Preprint.

[48] E. Mu, S. Jabbour, A. V. Dalca, J. Guttag, J. Wiens, and M. W. Sjoding, *Augmenting Existing Deterioration Indices With Chest Radiographs to Predict Clinical Deterioration*, PLOS ONE, 17 (2022), p. e0263922.

[49] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, *Multimodal Deep Learning*, in Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 689–696.

[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, in

Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 8024–8035.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[52] H. Peng, F. Long, and C. Ding, *Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (2005), pp. 1226–1238.

[53] PhysioNet, *PhysioNet Credentialed Health Data License for MIMIC-CXR*, 2026. Accessed: July 20, 2026.

[54] ——, *PhysioNet Credentialed Health Data Use Agreement for MIMIC-IV*, 2026. Accessed: July 20, 2026.

[55] S. Pölsterl, T. N. Wolf, and C. Wachinger, *Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform*, in Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, vol. 12905 of Lecture Notes in Computer Science, Springer, 2021, pp. 688–698.

[56] S. Purushotham, C. Meng, Z. Che, and Y. Liu, *Benchmarking Deep Learning Models on Large Healthcare Datasets*, Journal of Biomedical Informatics, 83 (2018), pp. 112–134.

[57] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*, 2017. Preprint.

[58] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, et al., *Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans*, Nature Machine Intelligence, 3 (2021), pp. 199–217.

[59] E. Rocheteau, P. Liò, and S. Hyland, *Temporal Pointwise Convolutional Networks for Length of Stay Prediction in the Intensive Care Unit*, in Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 58–68.

[60] J. Rueckel, L. Trappmann, B. Schachtner, P. Wesp, B. F. Hoppe, N. Fink, J. Ricke, J. Dinkel, M. Ingrisch, and B. O. Sabel, *Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs*, Investigative Radiology, 55 (2020), pp. 792–798.

[61] T. Saito and M. Rehmsmeier, *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*, PLOS ONE, 10 (2015), p. e0118432.

[62] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, *Transformers in Medical Imaging: A Survey*, Medical Image Analysis, 88 (2023), p. 102802.

[63] R. Sisk, L. Lin, M. Sperrin, J. K. Barrett, B. D. M. Tom, K. Diaz-Ordaz, N. Peek, and G. P. Martin, *Informative Presence and Observation in Routine Health Data: A Review of Methodology for Clinical Risk Prediction*, Journal of the American Medical Informatics Association, 28 (2021), pp. 155–166.

[64] B. M. K. Siu, G. H. Kwak, L. Ling, and P. Hui, *Predicting the Need for Intubation in the First 24 Hours After Critical Care Admission Using Machine Learning Approaches*, Scientific Reports, 10 (2020), p. 20931.

[65] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, *Multimodal Deep Learning for Biomedical Data Fusion: A Review*, Briefings in Bioinformatics, 23 (2022), p. bbab569.

[66] M. Tan and Q. V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, in Proceedings of the 36th International Conference on Machine Learning, vol. 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 6105–6114.

[67] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, *Efficient Transformers: A Survey*, ACM Computing Surveys, 55 (2022), pp. 1–28.

[68] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, *Multimodal Transformer for Unaligned Multimodal Language Sequences*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 6558–6569.

[69] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, *Calibration: The Achilles Heel of Predictive Analytics*, BMC Medicine, 17 (2019), p. 230.

[70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention Is All You Need*, in Advances in Neural Information Processing Systems, vol. 30, 2017.

[71] W. Wang, D. Tran, and M. Feiszli, *What Makes Training Multi-Modal Classification Networks Hard?*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.

[72] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, *ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.

[73] S. Wiegreffe and Y. Pinter, *Attention Is Not Not Explanation*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 11–20.

[74] R. Wightman, *rwightman/pytorch-image-models: v0.4.12*, 2021.

[75] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, *Self-Training with Noisy Student Improves ImageNet Classification*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10687–10698.

[76] W. Yao, C. Liu, K. Yin, W. K. Cheung, and J. Qin, *Addressing Asynchronicity in Clinical Multimodal Fusion via Individualized Chest X-Ray Generation*, in Advances in Neural Information Processing Systems, vol. 37, 2024.

[77] W. Yao, K. Yin, W. K. Cheung, J. Liu, and J. Qin, *DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 16416–16424.

[78] K. Yin, H. Xu, W. Yao, C. Liu, Z. Chen, Y. H. Cheung, W. K. Cheung, and J. Qin, *When Does Multimodal Learning Help in Healthcare? A Benchmark on EHR and Chest X-Ray Fusion*, 2026. Preprint.

[79] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, *Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study*, PLOS Medicine, 15 (2018), p. e1002683.

[80] H. Zou and T. Hastie, *Regularization and Variable Selection via the Elastic Net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 301–320.