# Evaluating the Incremental Contribution of Chest X-ray to 24-Hour ICU Deterioration Prediction

Marco Tenorio Cortés[1 [0009−0007−6309−0153]], Andreea M. Oprescu[1,2 [0000−0003−3308−4688]], and M.Carmen Romero Ternero[1,2,3 [0000−0001−6965−9485]]

[1] Departamento de Tecnología Electrónica, Universidad de Sevilla, Sevilla, Spain
[2] Instituto de Ingeniería Informática (I3US), Universidad de Sevilla, Sevilla, Spain
[3] Instituto de Biomedicina de Sevilla (IBIS), Sevilla, Spain

**Abstract.** Multimodal deterioration models increasingly combine imaging with structured clinical data. We evaluate the added contribution of CXR after conditioning on dense ICU physiological and laboratory trajectories in 57,659 time-synchronised samples from 10,810 adult ICU patients in MIMIC. The defined endpoint was new invasive mechanical ventilation, new vasopressor initiation, or death within 24 hours after radiography. We trained joint-fusion models combining vital-sign series, laboratory series, chest radiograph embeddings and metadata, and compared them with tabular, unimodal and image-free ablations. Image-free fusion of vital signs, laboratory data and metadata achieved the best composite performance (AUROC 0.831, AUPRC 0.455), exceeding the strongest tabular baseline, LightGBM (AUROC 0.783, AUPRC 0.377). Adding the radiograph did not improve performance in this setting. These findings suggest limited incremental contribution from CXR after conditioning on dense contemporaneous ICU physiological and laboratory data.

**Keywords:** Multimodal deep learning · Joint fusion · MIMIC-IV

## 1 Introduction

In the intensive care unit (ICU), clinical deterioration can occur over short time scales, and decisions to escalate support, such as initiating invasive mechanical ventilation or vasopressors, are informed by multiple sources of information. Vital signs, laboratory results, chest radiography (CXR) and clinical context all contribute to bedside assessment of severity, while established severity and early-warning tools combine selected physiological and clinical variables into summary scores, such as the National Early Warning Score 2 (NEWS2) developed by the Royal College of Physicians [18] and the SOFA score proposed by Vincent et al. [21]. This multimodal nature of ICU assessment motivates models that can evaluate whether different data sources provide complementary prognostic information. Existing deterioration models have shown that both structured physiological data and imaging can be informative. However, for near-term ICU deterioration, it remains unclear CXR adds measurable predictive value once dense

physiological and laboratory trajectories are already available. This question is particularly relevant in the ICU, where patients are continuously monitored and escalation decisions are closely tied to evolving physiology. We address this question by developing a multimodal joint-fusion model for 24-hour deterioration prediction after CXR in MIMIC-IV/MIMIC-CXR. The endpoint is a clinically actionable composite of new invasive mechanical ventilation, new vasopressor initiation or death within 24 hours. We compare image-inclusive fusion against image-free ablations, unimodal models and aggregated tabular baselines to isolate the incremental contribution of the radiograph.

## 2 Related Work

Rule-based early-warning scores such as the NEWS2 remain widely used for standardized assessment of acute illness severity and are based on bedside physiological observations measured at a given time point. Machine-learning early-warning systems using electronic health records (EHR) data have shown strong performance in clinical deterioration prediction. Edelson et al. [2] further reported that, in some head-to-head evaluations, these systems outperform aggregate rule-based scores. Hyland et al. [6] further showed that high-resolution physiological monitoring data can predict circulatory failure in the ICU with an AUROC of 0.94.

CheXNet [17] and CheXpert [7] established DenseNet121 [5] as the de facto backbone for CXR analysis, initially for pathology detection. More recent work targets clinical deterioration rather than diagnosis. Rahman et al. [16] compare several architectures to predict ICU transfer or death within 24 h from radiographs alone, reaching an AUROC of 0.734 on a ward population.

Several studies have combined CXR with structured clinical data, including MIMIC-based multimodal benchmarks. Hayat et al. [4] uses MIMIC-IV clinical time series linked to MIMIC-CXR, while Soenksen et al. [19] integrates tabular data, time series, text and images from MIMIC-IV and MIMIC-CXR-JPG. More closely related to clinical deterioration, Mu et al. [14] evaluated whether CXR could augment existing EHR-based deterioration indices.

## 3 Methodology

We use data from MIMIC-IV v3.1 [8] and MIMIC-CXR-JPG v2.1.0 [9], two large databases collected at the Beth Israel Deaconess Medical Center Hospital in Boston since 2008 and distributed by PhysioNet [3]. MIMIC-IV provides structured EHR containing demographics, admissions, ICU stays, charted vital signs and laboratory results, while MIMIC-CXR-JPG provides 377,110 chest radiographs with acquisition timestamps and view position. The two databases share common patient identifiers, which makes paired multimodal study possible: of the 65,366 adult ICU patients in MIMIC-IV, 30.8 % also have at least one image in MIMIC-CXR-JPG.

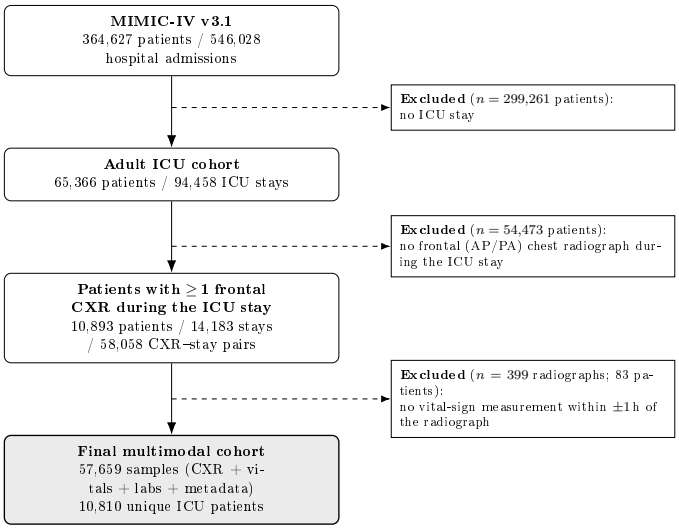



**Fig. 1.** Inclusion diagram. Successive filters from adult ICU stays to the final synchronised multimodal cohort, with the prevalence of the endpoint reported at each step.

### 3.1 Cohort definition

The cohort is built with the following successive filters (see Fig. 1). Starting with all adult ICU stays (age $\geq 18$), we select patients with at least one frontal chest radiograph (posteroanterior or anteroposterior view) acquired during the ICU stay. The resulting set is dominated by portable AP films, as expected in critically ill patients. Lastly, we require at least one vital-sign measurement within a strict $\pm 1$ h window of the radiograph, which guarantees that the physiological snapshot fed to the model is synchronous with the image and not a later state contaminated by the very deterioration we aim to predict. The final cohort comprises 57,659 multimodal samples from 10,810 adult ICU patients, with 52 % of the vital-sign readings falling within $\pm 30$ min of the image.

### 3.2 Composite 24-hour endpoint

We aim to predict a clinically actionable composite deterioration endpoint over the window $(t_i, t_i + 24\text{h}]$ following each radiograph acquired at time $t_i$:

$$y_i = \mathbb{I}[\,\text{newIMV}_i \;\vee\; \text{newVaso}_i \;\vee\; \text{death24h}_i\,] \tag{1}$$

where the three components are a new start of invasive mechanical ventilation (newIMV), a new vasopressor initiation (newVaso) among the five main agents and death within 24 h (death24h). The "new" condition excludes patients already receiving the corresponding support at $t_i$, so that the label captures an escalation of care rather than its continuation. The component prevalences are 4.1 % (new

**Table 1.** Clinical range limits used to flag physiologically impossible values (mapped to missing before winsorisation).

| Variable | Range | Variable | Range |
|---|---|---|---|
| Heart rate | $[10, 300]$ bpm | Haemoglobin | $[1, 30]$ g/dL |
| Respiratory rate | $[2, 70]$ min$^{-1}$ | Haematocrit | $[5, 80]$ % |
| $SpO_2$ | $[30, 100]$ % | White blood cells | $[0.05, 200] \times 10^3/\mu$L |
| Systolic BP | $[30, 350]$ mmHg | Platelets | $[1, 2000] \times 10^3/\mu$L |
| Diastolic BP | $[10, 250]$ mmHg | Creatinine | $[0.05, 50]$ mg/dL |
| Mean BP | $[20, 300]$ mmHg | Lactate | $[0.05, 40]$ mmol/L |
| Temperature | $[28, 45]$ °C | Urea | $[0.5, 300]$ mg/dL |
| | | Bicarbonate | $[2, 60]$ mEq/L |

IMV), 8.5 % (new vasopressors) and 1.8 % (death within 24 h); since the events partially overlap, the composite prevalence is 12.0 %.

### 3.3 Features and preprocessing

We used the vital signs available in the dataset, namely, heart rate, respiratory rate, peripheral oxygen saturation, temperature, and blood pressure measurements. We restricted blood pressure to the non-invasive measurements to maximise availability across patients, resulting in seven vital-sign variables: heart rate, respiratory rate, $SpO_2$, systolic, diastolic and mean non-invasive blood pressure, and temperature. We restricted the set to parameters recorded continuously and quantitatively, excluding intermittent or qualitative observations (e.g. level of consciousness) and invasive measurements available only in a minority of patients with an arterial or central line (e.g. invasive arterial pressure). Laboratory variables were selected as a compact set of routinely measured ICU analytes covering domains relevant to acute deterioration: haematological status and oxygen-carrying capacity (haemoglobin, haematocrit, white blood cells and platelets), renal function (creatinine and urea), acid–base/metabolic status (bicarbonate), and tissue hypoperfusion or shock physiology (lactate). These domains overlap with established ICU severity and organ-dysfunction frameworks: APACHE II includes haematocrit, white blood cell count and creatinine; SAPS II includes white blood cell count, urea and bicarbonate; and SOFA uses platelet count and creatinine for coagulation and renal dysfunction [11,12,21]. Lactate was included separately because of its established association with tissue hypoperfusion and mortality in critically ill patients [15].

Physiologically impossible values are mapped to missing using clinical range limits (Table 1), remaining extreme values are winsorised at $Q_1 - 3\,\mathrm{IQR}$ / $Q_3 + 3\,\mathrm{IQR}$ (2,128 values affected), and all variables are $z$-normalised with statistics fitted on the training split only to prevent data leakage. We derive two input representations from the same raw measurements. For the tabular baselines, each vital sign is summarised within $\pm 1$ h by five statistics (mean, standard deviation, minimum, maximum and temporal slope) and each laboratory variable by its

**Table 2.** Patient-grouped split for the 24-hour composite endpoint.

| Subset | Samples | Patients | Prevalence |
|---|---|---|---|
| Train | 39,611 | 7,566 | 12.0 % |
| Validation | 9,085 | 1,622 | 12.3 % |
| Test | 8,963 | 1,622 | 11.6 % |
| **Total** | **57,659** | **10,810** | **12.0 %** |

nearest value within ±6 h, together with binary missingness indicators; residual missing values are imputed with the training-set median, yielding 56 features. For the sequential encoders, each modality is discretised into 24 temporal bins—±6 h (30-min bins) for the vital signs and ±12 h (1-h bins) for the laboratory variables—imputed by forward-fill (with an initial back-fill for the slowly-varying laboratory series, and zero-filling for channels never measured in the window) and augmented with an explicit binary presence mask per channel, so the model can tell measured values from imputed ones. To avoid subject-level leakage from a patient contributing several radiographs, all samples of a given patient are assigned to the same dataset split. The endpoint prevalence stays comparable across the three subsets (see Table 2).

### 3.4 Multimodal fusion

We propose a joint-fusion architecture (see Fig. 2). Every modality is processed by a specialised encoder that produces a fixed-length embedding, the embeddings are concatenated into a shared representation space, and a common multi-task head produces four outputs, one per component of the endpoint and one for the composite. Joint fusion lets the modalities interact during training, unlike late fusion, while still allowing each encoder to exploit modality-specific inductive biases and pretrained weights, unlike early fusion.

The radiograph is encoded with DenseNet121 [5] pretrained on ImageNet [7,17]. Images are preprocessed with Medical Open Network for AI [1]: converted to grayscale, resized to $384 \times 384$, intensity-scaled to $[0, 1]$ and normalised with ImageNet statistics so that the pretrained weights remain usable. Training-time augmentations include random rotation ($\pm 10°$), affine jitter, contrast and histogram shifts, Gaussian noise and coarse dropout; we exclude horizontal flipping, which would simulate situs inversus (the heart on the right) and teach the model a pathological anatomy as a normal variant. To avoid the severe over-fitting observed when training end-to-end from the start, the backbone is fine-tuned in two phases: a warm-up in which only the new classification head is trained with the backbone frozen, followed by full fine-tuning with discriminative learning rates (a smaller rate for the pretrained backbone than for the new head). The penultimate layer yields a 1024-d image embedding.

Non-sequential metadata (age, sex, ICU length of stay and two binary baseline indicators of whether the patient is already on invasive mechanical ventilation and on vasopressors, respectively, at $t_i$) are encoded by a small two-layer

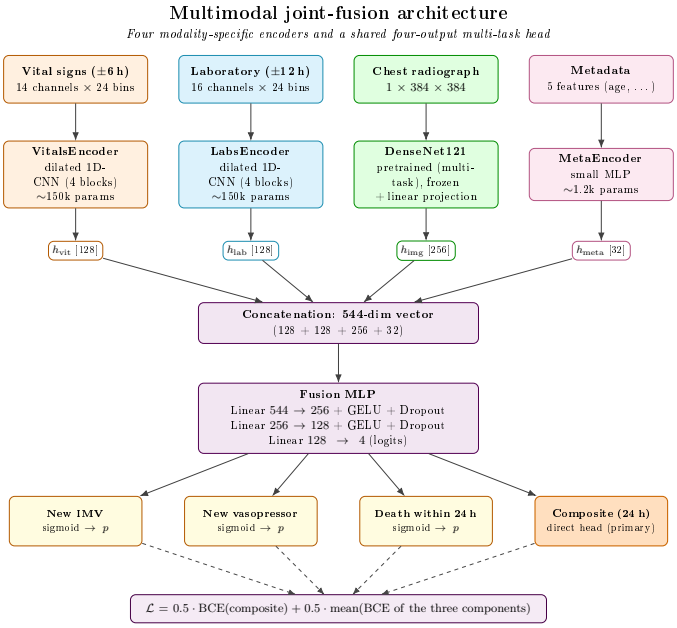



**Fig. 2.** Joint-fusion architecture. A DenseNet121 image encoder, two dilated 1D-CNN encoders for the vital-sign and laboratory time series, and a small metadata MLP produce embeddings that are concatenated and passed to a shared four-head classifier (three components and the composite).

MLP ($5 \to 32$). The baseline-support indicators are clinically important because they encode the patient's starting point on the escalation pathway.

The four embeddings—vitals (128-d), labs (128-d), image (1024-d, linearly projected to 256-d) and metadata (32-d)—are concatenated into a 544-d vector and passed to a fusion MLP ($544 \to 256 \to 128 \to 4$, with GELU and dropout 0.4) ending in four sigmoid outputs. We observed that training only the three component heads and reconstructing the composite a posteriori (by probabilistic OR, maximum, mean or a learned logistic combination) is sub-optimal: the reconstructed composite falls below the per-component scores, because the composite carries structural information about which patients are close to any severe deterioration that is not recovered by combining atomic tasks. We therefore add the composite as an explicit fourth task, and train with a balanced loss

$$\mathcal{L} = 0.5\,\mathcal{L}_{\text{comp}} + 0.5 \cdot \frac{1}{3} \sum_{i \in \{\text{imv,vaso,death}\}} \mathcal{L}_i, \tag{2}$$

where each $\mathcal{L}_i$ is a binary cross-entropy with a per-task positive weight pos_weight $= \sqrt{n_{\text{neg}}/n_{\text{pos}}}$ to counter class imbalance.

## 4 Experiments and Results

Two tabular baselines are trained on the 56 aggregated features: Logistic Regression and LightGBM [10]. Unimodal deep baselines are the DenseNet121 image model and each sequential encoder with its own four-head classifier. All deep models are optimised with AdamW [13], weight decay and early stopping on the validation AUROC; in the fusion model the pretrained sequential encoders are fine-tuned with a smaller learning rate than the freshly initialised image projection, metadata encoder and fusion head (discriminative learning rates). To keep fusion training efficient, image embeddings are pre-computed once with the frozen DenseNet and cached, reducing the per-epoch cost from minutes to seconds.

The validation split is used only for model development (monitoring early stopping, selecting the best checkpoint, tuning hyperparameters and choosing the operating-point threshold (Youden's index)), while the test split is held out and used for the final evaluation. We report AUROC and AUPRC on the composite endpoint (see Table 3), the per-component AUROC of the key models (see Table 4) and the corresponding ROC and Precision–Recall curves (see Fig. 3). AUPRC is reported alongside AUROC because, under the 12 % prevalence of the endpoint, it is more informative about performance on the minority class.

**Table 3.** Test-set performance on the 24-hour composite endpoint. Modalities: T = aggregated tabular, V = vital-sign series, L = laboratory series, I = image, M = metadata. Higher AUROC performance in **bold**.

| Model | Modalities | AUROC | AUPRC |
|---|---|---|---|
| Logistic Regression (tabular) | T | 0.723 | 0.304 |
| LightGBM (tabular) | T | 0.783 | 0.377 |
| DenseNet121 (composite) | I | 0.610 | 0.166 |
| DenseNet121 (multi-task) | I | 0.597 | 0.164 |
| VitalsEncoder (1D-CNN) | V | 0.708 | 0.274 |
| LabsEncoder (1D-CNN) | L | 0.689 | 0.261 |
| Full fusion (frozen image emb.) | V+L+I+M | 0.812 | 0.408 |
| **Fusion without image (ablation)** | **V+L+M** | **0.831** | **0.455** |
| Fusion end-to-end (trainable DenseNet) | V+L+I+M | 0.808 | 0.412 |
| Fusion with Transformer encoders | V+L+M | 0.828 | 0.456 |

The aggregated-feature (T) baselines performed strongly relative to the image-only model, with AUROCs of 0.723 for Logistic Regression and 0.783 for LightGBM, compared with 0.610 for DenseNet121. The individual sequential encoders achieved intermediate performance, with AUROCs of 0.708 for the VitalsEncoder and 0.689 for the LabsEncoder. Fusion of structured ICU modalities improved over these non-fusion baselines, increasing the composite AUROC from 0.783 for LightGBM to 0.831 for the image-free fusion model. The corresponding AUPRC

**Table 4.** Per-component test AUROC for selected models. The image-free fusion matches or exceeds the image fusion on every target.

| Model | IMV | Vaso | Death | Composite |
|---|---|---|---|---|
| DenseNet121 (multi-task) | 0.645 | 0.608 | 0.683 | 0.597 |
| VitalsEncoder (1D-CNN) | 0.769 | 0.726 | 0.775 | 0.708 |
| LabsEncoder (1D-CNN) | 0.724 | 0.689 | 0.820 | 0.689 |
| Full fusion | 0.921 | 0.869 | 0.863 | 0.812 |
| **Fusion without image** | **0.936** | **0.881** | **0.894** | **0.831** |
| Fusion end-to-end | 0.924 | 0.870 | 0.824 | 0.808 |

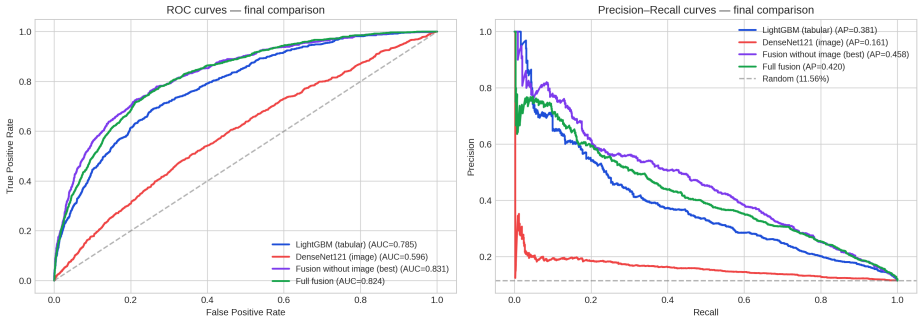



**Fig. 3.** ROC (left) and Precision–Recall (right) on the composite endpoint for LightGBM, DenseNet121, the full fusion and the image-free fusion. The image-free fusion achieves the strongest summary performance among the displayed models.

increased from 0.377 to 0.455, with the Transformer variant achieving a similar AUPRC of 0.456.

The highest composite AUROC was achieved by the image-free fusion model using vital-sign series, laboratory series and metadata (0.831). Models incorporating the radiograph performed below this image-free fusion model, with composite AUROCs of 0.812 for frozen image embeddings and 0.808 for end-to-end DenseNet fine-tuning. Thus, in this experimental setting, adding the radiograph did not improve prediction of the 24-hour composite endpoint after conditioning on structured ICU data. Across five random seeds this ordering was consistent (image-free $0.8305 \pm 0.0027$ vs. image $0.8120 \pm 0.0035$ composite AUROC; image-free higher in 5/5 runs), indicating that the gap is stable rather than an artefact of a single training run.

Table 4 shows that this pattern was also observed across the individual endpoint components. The image-free fusion model achieved the highest AUROC for new invasive mechanical ventilation (0.936), new vasopressor initiation (0.881), death within 24 h (0.894), and the composite endpoint (0.831). The image-only multi-task DenseNet performed below the structured-data models across all targets, whereas the vital-sign and laboratory encoders provided complementary

**Table 5.** Test-set operating point for the final image-free fusion model at the validation-selected threshold of 0.23. The threshold was chosen by Youden's index and applied unchanged to the test set ($n = 8{,}963$; 1,036 positives, 11.6 %).

| | Confusion matrix | | | Metrics | |
|---|---|---|---|---|---|
| Actual class | Pred. det. | Pred. stable | Total | Metric | Value |
| Deterioration | 778 | 258 | 1,036 | Sensitivity | 0.751 |
| Stable | 1,889 | 6,038 | 7,927 | Specificity | 0.762 |
| Total | 2,667 | 6,296 | 8,963 | Precision | 0.292 |
| | | | | F1-score | 0.420 |

signal. In particular, the laboratory encoder achieved the strongest unimodal AUROC for death within 24 h (0.820).

Replacing the dilated 1D-CNN encoders with Transformer encoders [20] leaves the composite AUROC essentially unchanged (0.828 vs. 0.831), so the result does not depend on the sequential-encoder architecture. It is also stable across the image-training scheme (frozen embeddings vs. end-to-end fine-tuning, both below the image-free model) and across the prediction horizon: shortening it to 6 and 12 h raises the LightGBM AUROC to 0.827 and 0.805 respectively (versus 0.790 at 24 h, as the event moves closer in time), yet the physiological series remain more predictive than the image at every horizon.

At the selected threshold, the model achieved a sensitivity of 0.751 and specificity of 0.762. Precision was 0.292, reflecting the low event prevalence and the expected false-positive burden of a sensitive early-warning threshold. In an early-warning context this operating point deliberately favours sensitivity: a false positive entails re-checking a patient who is already under continuous monitoring, whereas a false negative is a missed deterioration, so the higher false-positive rate is an accepted trade-off for patient safety. The threshold is tunable, and moving it towards higher precision (e.g. a max-F1 criterion) reduces false alarms at the cost of some sensitivity, letting the operating point be matched to the clinical use case.

## 5 Discussion

Our main finding is that, for predicting 24-hour ICU deterioration after CXR, dense structured clinical data provided the strongest predictive signal. Fusion of vital-sign, laboratory and metadata outperformed all unimodal and tabular baselines, indicating that temporal physiological and laboratory context is important for this near-term endpoint. However, in our experimental setting and for our defined endpoint, adding the chest radiograph did not improve performance beyond this image-free fusion model, either with frozen image embeddings or with end-to-end image fine-tuning. This suggests that, in this cohort and prediction window, the incremental contribution of CXR was limited after conditioning on contemporaneous ICU physiology and laboratory data. This

finding should not be interpreted as evidence that chest radiographs are clinically uninformative. Rather, the endpoint studied here (new invasive mechanical ventilation, new vasopressor initiation or death within 24 h) is closely tied to physiological instability. In a highly monitored ICU setting, this instability may already be reflected enough in vital-sign and laboratory trajectories around the time of imaging. We examined several methodological explanations for this finding. The pattern was not confined to a single image-model configuration: models using frozen DenseNet embeddings, an end-to-end fine-tuned DenseNet, and a pretrained multi-task image backbone all performed below the image-free fusion model. It was also not specific to the sequential encoder architecture, as replacing the dilated 1D-CNN with Transformer encoders yielded similar composite performance. Finally, the finding was not restricted to the 24-hour horizon, with structured physiological models remaining stronger than image-only models at 6, 12 and 24 hours. These analyses do not imply that CXR are intrinsically non-prognostic; rather, they suggest that their incremental value may be attenuated when dense, contemporaneous ICU physiological and laboratory context is already available. This result should also be interpreted in relation to prior work by Mu et al. [14], who found that CXR could augment EHR-based deterioration indices, particularly for low-risk identification. The difference likely reflects that the two studies do not share the same research question: Mu et al. assess whether CXR improves existing EHR-based deterioration indices, whereas we evaluate whether the radiograph provides additional signal once the model already has access to dense vital-sign and laboratory trajectories around the time of imaging in ICU patients. Our composite AUROC (0.831) is within the range of multimodal MIMIC benchmarks, such as MedFuse (AUROC $\sim$0.86 for in-hospital mortality [4]).

This study presents the following limitations. The data comes from a single-centre, retrospective database (MIMIC, 2008–2022), so the conclusion should be confirmed by external, multi-centre validation. We use a single radiograph per sample and therefore do not exploit the dynamics of serial imaging, which may carry signal that a static image does not. The multimodal cohort covers only patients who received a radiograph and is thus biased toward more severe cases. Finally, we report point estimates of AUROC and AUPRC without confidence intervals; future work should quantify statistical uncertainty, for example using bootstrap resampling or paired tests across model predictions.

## 6 Conclusion

We developed a complete multimodal joint-fusion system for 24-hour ICU deterioration prediction on a radiograph-centred cohort with a vital-sign measurement within $\pm 1$ h and used systematic ablations to assess the incremental contribution of CXR. In this setting, models incorporating the image did not improve over fusion models based on structured physiological data alone. The model with higher performance is an image-free fusion (AUROC 0.831 on the composite endpoint, 0.936 on new intubation), and the result is robust to the sequential-encoder ar-

chitecture, the image-training scheme and the prediction horizon. Future work includes longer prediction horizons (48–72 h) and less densely monitored settings such as general wards or the emergency department, more recent vision foundation models or radiology-specific encoders, modeling the dynamics of serial radiographs, analysing clinically homogeneous subgroups (e.g. ARDS, where deterioration is predominantly respiratory and imaging may still help), and external multi-centre validation.

## Acknowledgements

This work was partially funded by grant PID2022-141045OB-C41,C42,C43 funded by MCIN/AEI/10.13039/501100011033/ and Feder a way of making Europe in artifacts project: Generation of Reliable Synthetic Health Data for Federated Learning in Secure Data Spaces.

The authors used AI assistants (Anthropic Claude and Perplexity) for coding support and language editing and take full responsibility for the final content.

## References

1. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., et al.: MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
2. Edelson, D.P., Churpek, M.M., Carey, K.A., Lin, Z., Huang, C., Siner, J.M., Johnson, J., Krumholz, H.M., Rhodes, D.J.: Early Warning Scores With and Without Artificial Intelligence. JAMA Network Open **7**(10), e2438986 (2024). https://doi.org/10.1001/jamanetworkopen.2024.38986
3. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000). https://doi.org/10.1161/01.CIR.101.23.e215
4. Hayat, N., Geras, K.J., Shamout, F.E.: MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In: Proceedings of the 7th Machine Learning for Healthcare Conference (MLHC). PMLR, vol. 182, pp. 479–503 (2022)
5. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4700–4708 (2017). https://doi.org/10.1109/CVPR.2017.243
6. Hyland, S.L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B., Zimmermann, M., Bodenham, D., Borgwardt, K., Rätsch, G., Merz, T.M.: Early prediction of circulatory failure in the intensive care unit using machine learning. Nature Medicine **26**(3), 364–373 (2020). https://doi.org/10.1038/s41591-020-0789-4
7. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019). https://doi.org/10.1609/aaai.v33i01.3301590

8. Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., Lehman, L.w.H., Celi, L.A., Mark, R.G.: MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data **10**(1), 1 (2023). https://doi.org/10.1038/s41597-022-01899-x
9. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019). https://doi.org/10.1038/s41597-019-0322-0
10. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30, pp. 3146–3154 (2017)
11. Knaus, W.A., Draper, E.A., Wagner, D.P., Zimmerman, J.E.: APACHE II: A severity of disease classification system. Critical Care Medicine **13**(10), 818–829 (1985)
12. Le Gall, J.R., Lemeshow, S., Saulnier, F.: A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA **270**(24), 2957–2963 (1993)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
14. Mu, E., Jabbour, S., Dalca, A.V., Guttag, J., Wiens, J., Sjoding, M.W.: Augmenting existing deterioration indices with chest radiographs to predict clinical deterioration. PLoS One **17**(2), e0263922 (2022). https://doi.org/10.1371/journal.pone.0263922
15. Nichol, A.D., Egi, M., Pettilä, V., Bellomo, R., French, C., Hart, G., Davies, A., Stachowski, E., Reade, M.C., Bailey, M., Cooper, D.J.: Relative hyperlactatemia and hospital mortality in critically ill patients: a retrospective multi-centre study. Critical Care **14**(1), R25 (2010)
16. Rahman, M., Gao, J., Carey, K.A., Edelson, D.P., Afshar, A., Garrett, J.W., Chen, G., Afshar, M., Churpek, M.M.: Comparison of deep learning approaches using chest radiographs for predicting clinical deterioration: retrospective observational study. JMIR AI **4**, e67144 (2025). https://doi.org/10.2196/67144
17. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C.P., Shpanskaya, K., Lungren, M.P., Ng, A.Y.: CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
18. Royal College of Physicians: National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. Updated report of a working party. Tech. rep., Royal College of Physicians, London (Dec 2017), https://www.rcp.ac.uk/media/a4ibkkbf/news2-final-report_0_0.pdf
19. Soenksen, L.R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H.M., Li, M.L., Fuentes, I., Bertsimas, D.: Integrated multimodal artificial intelligence framework for healthcare applications. npj Digital Medicine **5**(1), 149 (2022). https://doi.org/10.1038/s41746-022-00689-4
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30, pp. 5998–6008 (2017)
21. Vincent, J.L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C.K., Suter, P.M., Thijs, L.G.: The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Medicine **22**(7), 707–710 (1996). https://doi.org/10.1007/BF01709751