ACTIONWAVE: A SYSTEMATIC BENCHMARK FOR WAVEFORM-CENTRIC CLINICAL PREDICTION ACROSS CARE SCENARIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous bedside waveforms record rapid physiological changes that can inform clinical decisions. Existing benchmarks examine selected prediction conditions, but it remains unclear how model advantages depend on signal duration, warning time, and available inputs across care settings. We introduce **Actionwave**, a benchmark for waveform-centric clinical prediction across intensive-care, operating-room, and emergency-department settings. **Actionwave** uses a common decision-time prediction interface and groups clinically grounded tasks by what is inferred and when. It varies conditions relevant to clinical use, including the warning required before an intervention. We compare diverse architectures and pretrained biosignal encoders under specified model-size and compute budgets. The protocol requires patient-level splits, strict information cutoffs, and comparisons that isolate individual modeling factors. Across 20 tasks and 1.4M task-labelled windows from source datasets, the experimental grids account for over 87,000 model fits. These comparisons identify task-specific choices and show why matched cohorts, unimodal controls, and explicit timing are necessary when interpreting waveform model performance.

1 INTRODUCTION

Continuous bedside monitoring captures physiological dynamics missed by intermittent clinical measurements (Sundrani et al., 2023). Electrocardiography (ECG), arterial blood pressure (ABP), photoplethysmography (PPG), and respiration (RESP) capture changes over seconds to minutes for clinical prediction (Moody et al., 2020; Lee et al., 2022; Kansal et al., 2025). Their predictive value depends on prediction conditions (Saeb et al., 2017; Lauritsen et al., 2021). Intensive-care units (ICU), operating rooms (OR), and emergency departments (ED) differ in available sensors, monitoring duration, and the warning time required before an intervention. Studying these conditions reveals which modeling choices work, where benefits persist, and what needs improvement.

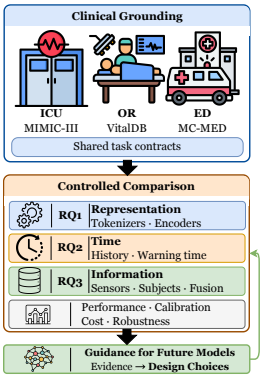


Figure 1: **Actionwave**: clinically grounded tasks and controlled comparisons provide evidence to guide future model development. The comparison axes are summarized in Table 4.

Waveform encoders and biosignal foundation models offer diverse approaches to clinical prediction. YAIB, MC-BEC, and SignalMC-MED support reproducible evaluation across clinical definitions, modalities, and signal durations (van de Water et al., 2024; Chen et al., 2023; Gustafsson et al., 2026). Two issues obscure which raw-waveform modeling choices offer consistent gains across care settings and prediction conditions. First, encoder comparisons often also change tokenization, sampling, input channels, or computational resources, obscuring architecture's contribution. Second, changing observation length or forecast gap can alter episode eligibility, follow-up availability, and outcome prevalence. Performance differences can therefore reflect population shifts or task difficulty alongside modeling choices. These comparisons require control over model inputs and prediction-task construction. Table 1 compares our raw-waveform task coverage across sources and experimental axes with complementary benchmarks. We ask three questions that inform clinical waveform model development:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with clinical time-series and waveform benchmarks, with experimental axes grouped by our three research questions. Scope columns denote raw-waveform availability, independent source cohorts, and clinical task families. Systolic and diastolic blood pressure count as one family. RQ columns indicate reported comparisons (✓) or none (–). Input panels include feature groups, and gap/horizon coverage includes multiple target horizons. The **Actionwave** row summarizes the specified benchmark design.

	Scope			RQ1	RQ2: Time		RQ3: Information		
Benchmark	Raw waveform	Multiple sources	Multiple tasks	Tokenization	Observation length	Gap or horizon	Input panels	Missingness	Data budget
MIMIC-III (Harutyunyan et al., 2019)	–	–	✓	–	–	–	–	–	–
YAIB (van de Water et al., 2024)	–	✓	✓	–	–	–	✓	–	✓†
PulseDB (Wang et al., 2023)	✓	✓	–	–	–	–	–	–	–
MC-BEC (Chen et al., 2023)	✓	–	✓	–	–	✓	✓	✓	–
MDS-ED (Lopez Alcaraz et al., 2025)	✓	–	✓	–	–	✓§	✓	✓‡	–
SignalMC-MED (Gustafsson et al., 2026)	✓	–	✓	–	✓	–	✓	–	✓
Actionwave (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

† YAIB varies target-cohort training/fine-tuning sample size. ‡ MDS-ED compares including versus omitting missingness masks. § MDS-ED evaluates multiple mortality and ICU-admission target horizons.

(RQ1) Representation and architecture. How do waveform tokenization and encoder choices affect predictive performance and model rankings under controlled comparisons?

(RQ2) Temporal modeling. How do observation history and warning time affect predictability, and how do episode eligibility and outcome prevalence change across these comparisons?

(RQ3) Information availability. When do additional sensors, training subjects, and clinical modalities improve prediction, and how robust are these gains to missing inputs?

We introduce **Actionwave** (Figure 1), a waveform-centric benchmark covering 20 tasks across MIMIC-III intensive-care records, VitalDB surgical records, and MC-MED emergency-department records (Moody et al., 2020; Lee et al., 2022; Kansal et al., 2025). **Actionwave** represents each prediction as a time-indexed episode with a clinical outcome, observation window, available inputs, and a warning interval where applicable. This shared interface makes observation history, forecast gap, sensor access, and clinical context explicit experimental factors, providing the following contributions: ❶ We provide a common framework for task definitions and evaluation across care settings. ❷ We compare representations, temporal context, and information availability in grids accounting for over 87,000 model fits, including hyperparameter searches and repetitions (Appendix A.3). ❸ We identify task-specific trade-offs in discrimination, calibration, efficiency, and robustness to guide model choice under clinical data and warning-time constraints. Together, these findings establish a foundation to enable future model development for clinical waveform prediction.

2 Related Work

Clinical waveform benchmarks. Clinical time-series and waveform benchmarks standardize cohorts, tasks, and input protocols for reproducible comparisons across acute-care settings (Harutyunyan et al., 2019; van de Water et al., 2024; Chen et al., 2023; Wang et al., 2023; Lopez Alcaraz et al., 2025; Gustafsson et al., 2026). The MIMIC-III clinical time-series benchmark defined mortality, decompensation, length-of-stay, and phenotyping tasks using structured ICU data rather than raw waveforms (Harutyunyan et al., 2019). The MIMIC-III Waveform Database Matched Subset linked multichannel bedside signals to clinical records without prescribing matched comparisons across temporal or sensing conditions (Moody et al., 2020). MC-BEC extended multimodal benchmarking to emergency care with decompensation, disposition, and revisit prediction, including missing-modality tests (Chen et al., 2023). MDS-ED evaluated raw and engineered ECG alongside routine clinical data for diagnosis and deterioration prediction in a common emergency-department cohort, including multiple mortality and ICU-admission horizons (Lopez Alcaraz et al., 2025). SignalMC-MED compared biosignal foundation models on synchronized ECG and PPG while varying modality, fusion, signal duration, and model scale (Gustafsson et al., 2026). **Actionwave** treats tokenization, imputation, temporal geometry, sensing, data budget, architecture, fusion, and source domain as explicit experimental factors across ICU, operating-room, and emergency-department settings. Its protocol calls for varying one factor at a time under matched episodes, splits, and budgets (Table 1).

Waveform representation learning. Time-series models depend on both the representation of samples, channels, and irregular observations and the modeling of temporal and cross-channel depen-

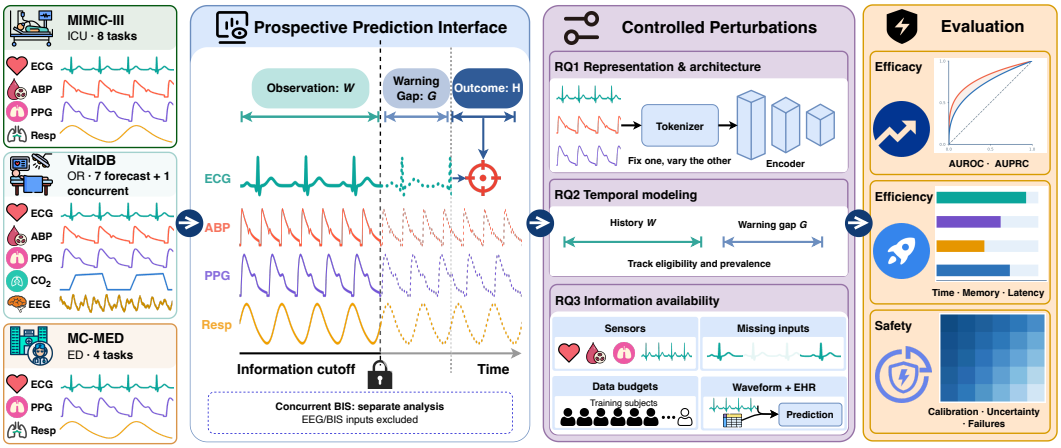


Figure 2: **Actionwave** links three data sources to a shared decision-time prediction interface and varies representation, time, and information availability to assess efficacy, efficiency, and safety. Task names and cohort support are listed in Tables 3 and 6.

dencies (Nie et al., 2023; Zeng et al., 2023; Zhang & Yan, 2023; Liu et al., 2024; Horn et al., 2020; Shukla & Marlin, 2021). PatchTST segments each channel into fixed-length patches and applies a channel-independent Transformer (Vaswani et al., 2017), showing that patching and channel independence can affect long-horizon forecasting accuracy and efficiency (Nie et al., 2023). LTSF-Linear (Zeng et al., 2023) shows that simple linear models can outperform several Transformer-based methods on standard long-term forecasting datasets, underscoring the importance of baseline choice and experimental design. Whether these tokenizer and encoder findings hold for clinical waveforms across tasks, sensing policies, missingness, and source domains remains an open question. **Actionwave** examines their applicability through comparisons organized around shared task definitions and resource constraints.

3 Benchmark and Shared Experimental Protocol

Actionwave covers clinical prediction tasks across intensive-care, operating-room, and emergency-department settings (Sec. 3.1). A shared episode interface defines outcomes, observation windows, applicable warning intervals, and information cutoffs (Sec. 3.2). Each comparison specifies the factor varied, the quantities held fixed, and the model-selection procedure (Sec. 3.3). Figure 2 illustrates the workflow and schematic comparisons of efficacy, efficiency, and safety (Sec. 3.4).

3.1 Clinical Sources and Task Coverage

Clinical Sources. **Actionwave** spans three care settings: the intensive care unit (MIMIC-III), operating room (VitalDB), and emergency department (MC-MED) (Moody et al., 2020; Lee et al., 2022; Kansal et al., 2025). Waveform inputs include ICU monitoring channels, twelve native intraoperative streams, and ED ECG-II and respiration, respectively. Table 2 summarizes unique patients and clinical episodes after task-specific eligibility filtering. *Clinical episodes* are ICU stays, surgical cases, or ED visits. *Windows* are task-labelled prediction examples. One episode can supply several windows across tasks or time points.

Table 2: Eligible cohort unions. Patients and clinical episodes are deduplicated across tasks. Windows count task-labelled rows across training, validation, and test.

Dataset	Patients	Episodes	Windows
MIMIC-III	8,535	9,772	979,920
VitalDB	5,346	5,575	385,667
MC-MED	15,852	18,536	34,379
In total	29,733	33,883	1,399,966

Task Coverage. Full task definitions appear in Appendix B. *MIMIC-III* includes 8 tasks: acute deterioration, dynamic mortality, respiratory escalation, circulatory deterioration, fluid response, KDIGO stage 2/3 AKI, ventilator liberation, and Sepsis-3. *VitalDB* includes 8 tasks: intraoperative hypotension, accelerating MAP decline, oxygen desaturation, core hypothermia, PCO_2–$ETCO_2$ discordance, abrupt $ETCO_2$ fall, coupled hemodynamic response, and concurrent light BIS state. We define 4 tasks in *MC-MED*: cardiorespiratory decompensation (60 min), pending hyperkalemia, lactate non-clearance, and pending severe anemia.

These sources provide around 1.4M task-labelled windows across care settings and endpoints. Concurrent BIS is analyzed separately from forecasting. Table 6 gives per-task cohort counts.

3.2 Episode construction and information boundaries

Episode Construction. Using the interface in Figure 2, we construct windows within eligible clinical episodes and represent each example as

$$\begin{aligned} \mathcal{E} &= (d, s, i, T, W, C, G, H, Y, M), \\ \mathcal{O} &= [T - W, T), \qquad \mathcal{F} = [T + G, T + G + H), \end{aligned} \tag{1}$$

where d, s, and i identify the dataset, care setting, and subject. T is the prediction anchor, W is the observation duration, and C denotes permitted channels. G and H specify the forecast gap and target horizon. Y is the outcome, and M records input validity and outcome ascertainment separately. Forecasting tasks use observations within $\mathcal{O}$ to predict an endpoint ascertained within $\mathcal{F}$. The concurrent VitalDB probe targets light BIS state within the observation interval and is analyzed separately. Candidate anchors follow prespecified surveillance, event, or daily schedules. Incident-event eligibility excludes prevalent events, handles events in the forecast gap according to the task contract, and requires sufficient follow-up to assign a negative label. Incompletely ascertained outcomes remain masked. Waveform views end at the common prediction anchor: from the extracted MIMIC-III history $[T - 6\,\mathrm{h}, T)$, a five-minute input uses $[T - 5\,\mathrm{min}, T)$ and a one-hour input uses $[T - 1\,\mathrm{h}, T)$. Raw waveform samples from the selected interval enter the model for prediction.

Information Boundaries. Waveforms and EHR values must precede the applicable decision-time cutoff. Future test results, label and censoring fields, and direct or derived target features are excluded. The concurrent VitalDB probe withholds EEG, BIS, and other label-defining inputs. Only input-availability masks enter the model. Outcome-ascertainment masks select labels for training and evaluation. Subject-level partitions keep repeated visits, surgical cases, and prediction windows within one split or fold. Dataset-specific preprocessing preserves physical time and input-presence masks. Normalization and learned preprocessing use training subjects only. Windows with unavailable channels remain eligible. Appendix B.1 details these controls.

3.3 Controlled comparisons and model selection

Each RQ1–RQ3 comparison in Figure 2 specifies one experimental change and quantities held fixed. Competing configurations should share eligible windows, labels, patient partitions, permitted information, and evaluation procedures within each task, except for the factor varied. We report changes in eligibility or available training subjects.

❶ **Representation Choices (RQ1).** Encoder comparisons require a common tokenizer[1] to isolate encoder effects. Tokenizer comparisons require a common encoder. Selected crossed comparisons test whether tokenizer changes alter encoder rankings. The reported architecture comparisons use approximately 500k trainable parameters. When tokenizers are selected for each encoder, we identify the comparison as one of encoder–tokenizer combinations. Parameter matching does not equate token counts, training cost, or hyperparameter-search effort. These require separate resource specifications. ❷ **Physiological Context and Warning Time (RQ2).** The observation-length comparisons use nested intervals $[T-W, T)$ on the same anchors, holding the prediction cutoff, labels, and patient partitions fixed. For warning-time comparisons, patient assignments to training, validation, and test partitions remain fixed across gaps, as do the observation view and target horizon. Each gap defines its own outcome window. Eligibility, follow-up support, and prevalence can therefore vary within these fixed patient partitions. These controls compare additional history at the same decision time and longer intervention lead times without reassigning patients between partitions. ❸ **Information Availability (RQ3).** Sensor comparisons preserve the evaluation cohort when selected inputs are unavailable, and channel selectors or imputers use training subjects only. Inference-time missing-input stresses use paired masks without changing training or model selection. Data-budget comparisons fix validation and test cohorts. The reported MIMIC-III comparison varies training-window quotas, while VitalDB and MC-MED vary unique training subjects and retain their windows. Fusion com-

[1] A tokenizer maps raw waveform samples to a sequence of feature vectors, each called a token. The encoder processes this token sequence.

parisons align windows and pre-cutoff information across waveform-only, EHR-only, and fusion models. Gains are assessed against both unimodal references, including the stronger one.

Across five seeds, validation selects checkpoint epochs and hyperparameters, including learning rates and tokenizers where applicable. Selected checkpoints are evaluated on the held-out test set. Appendices B.7 and A.2 detail the searches, configuration selection, and test-score aggregation.

3.4 Evaluation: efficacy, efficiency, and safety

The main comparisons and companion figures report predictive performance, training cost, and reliability. ❶ **Efficacy.** AUROC is the main discrimination metric. AUPRC characterizes precision–recall performance. Higher values are better for both. Task-level scores accompany equal-task macro summaries where recorded. VitalDB forecast summaries weight gaps equally within each task, then average the seven tasks. Concurrent BIS remains separate. Prevalence, outcome support, and the scoring estimator are specified in the corresponding study details. ❷ **Efficiency.** We report training time per optimizer step for observation-length comparisons and total training wall time for other experiments, together with accelerator memory and time per training sample. MC-MED temporal panels measure shared-device memory in MiB. Wall time is measured per run on MIMIC-III and MC-MED and per validation training cell on VitalDB. The figures label $1000/\text{training throughput}$ as latency, but it measures training cost, not inference or alert-delivery latency. Hardware, batch size, precision, and schedule constrain comparisons across studies. ❸ **Safety.** Recorded reliability metrics include expected calibration error (ECE, 15 bins), area under the risk–coverage curve (AURC), and failure rates from nonfinite-event, training-attempt, or prediction-file audits, as specified per study. Lower ECE indicates better calibration, lower AURC better selective prediction, and lower failure rates greater numerical stability. Separate VitalDB analyses also report Brier score, log loss, and paired sensor/modality-dropout effects. Seed variation describes training variability. Whole-subject bootstrap intervals estimate uncertainty conditional on fitted models. Metric definitions, aggregation, and selection qualifications are in Appendix A.2. Shaded placeholder panels are excluded from the evidence.

4 Representation Choices and Model Rankings

Representation performance depends on task and care setting. Figure 3 compares nine encoders (a), seven tokenizers on MIMIC-III and VitalDB and eight on MC-MED with a fixed Transformer (b), and tokenizer changes within each encoder (c). Encoders are Transformer (Vaswani et al., 2017), BiLSTM (Graves & Schmidhuber, 2005), Mamba-3 (Lahoti et al., 2026), InceptionTime (Ismail Fawaz et al., 2020), TCN (Bai et al., 2018), ResNet1D (Wang et al., 2017), S4 (Gu et al., 2022), FITS (Xu et al., 2024), and DLinear (Zeng et al., 2023). Conv1D projects fixed-length waveform patches following PatchTST (Nie et al., 2023). Other tokenizers are SeFT (Horn et al., 2020), Crossformer (Zhang & Yan, 2023), STraTS (Tipirneni & Reddy, 2022), iTransformer (Liu et al., 2024), FFT spectral projection (Yi et al., 2023), and mTAND (Shukla & Marlin, 2021). We adapt tokenizers to the shared encoder budget without reproducing complete systems. MC-MED additionally compares the local Conv2D tokenizer.

Outlined markers in panels (a,b) identify each task's highest displayed mean. Panel (c) shows $100 \times \Delta\mathrm{AUROC}$ from Conv1D for nine encoders (dots), their mean and SD (diamonds and bars), and counts agreeing with the unrounded mean's sign. Macros average eight MIMIC-III, four MC-MED, and seven VitalDB forecast tasks. Appendices B.2, B.7, and C give task names, protocols, and full scores, respectively. Restricted waveform inputs and required warning intervals make task AUROCs lower than the 0.8–0.9 reported (Appendix G).

Tokenization. Conv1D alone achieves top-three macro AUROC across all datasets, leading VitalDB (0.631) and MC-MED (0.574) and supporting local patches as a shared baseline (Figure 3(b)). SeFT leads on MIMIC-III (0.615 versus 0.611), driven by Sepsis-3 (0.669 versus 0.630). The other seven tasks collectively favor Conv1D, including fluid response (-0.012). SeFT handles irregular, unaligned timestamped observations (Horn et al., 2020), fitting MIMIC-III's sparse channel coverage. Aggregation outperforms local patches for sepsis, a syndrome of infection-related organ dysfunction (Singer et al., 2016), but offers less for intervention-linked blood-pressure changes. See Appendix C.1 for comparisons.

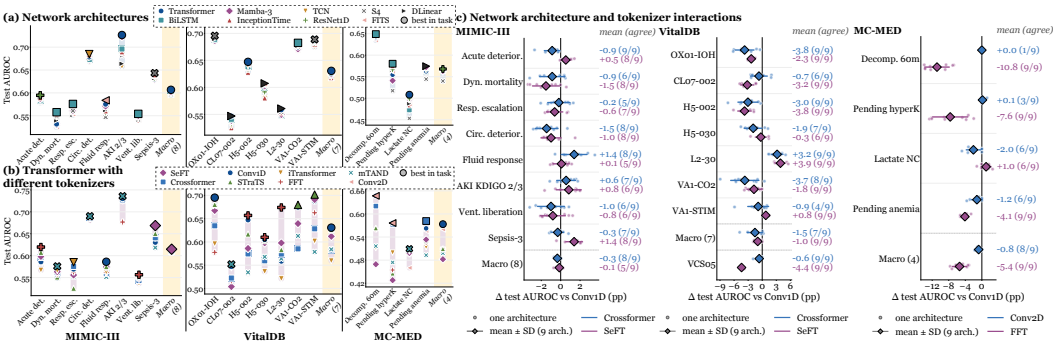


Figure 3: **Representation choices and model rankings.** (a) Encoders with selected tokenizers. (b) Tokenizers with a fixed Transformer. (c) Tokenizer–architecture interactions relative to Conv1D, including Conv2D and FFT on MC-MED. Numerical results are in Tables 10 and 11.

Architecture. Transformer leads MIMIC-III (0.607) and VitalDB (0.631). On MC-MED, ResNet1D (0.568), DLinear (0.567), and Transformer (0.567) have closely ranked macros. On MC-MED, BiLSTM leads decompensation (0.649) and hyperkalemia (0.580), DLinear leads anemia (0.574), and Transformer leads lactate nonclearance (0.508). DLinear's linear temporal projections (Zeng et al., 2023) remain competitive for anemia, while the recurrent encoder leads decompensation and hyperkalemia. Lactate AUROC is 0.453–0.508 across all nine encoders despite decompensation reaching 0.649, consistent with lactate reflecting production, clearance, and treatment over hours beyond the current cardiorespiratory state (Vincent et al., 2016). Table 10 reports the selected encoder–tokenizer combinations and fixed-Transformer tokenizer comparisons.

Tokenizer–architecture interactions. On MC-MED, FFT lowers mean macro AUROC by 0.054 across nine encoders, with losses of 0.108 for decompensation and 0.076 for hyperkalemia in all nine backbones. Its lactate gain is only 0.010 and occurs in six. FFT exposes within-patch frequency structure (Yi et al., 2023), but local patches are stronger for these ED tasks. Despite ECGs carrying potassium-related information (Galloway et al., 2019), repeated FFT losses show that this signal alone does not favor spectral tokenization. VitalDB shows a reversal: with a fixed Transformer, FFT gains 0.115 on gas discordance but loses 0.118 on hypotension, yielding a macro change of -0.006. Gas exchange reflects ventilation–perfusion physiology, while hypotension prediction draws directly on arterial-waveform changes (Nassar & Schmidt, 2016; Hatib et al., 2018). These results favor spectral projection for gas discordance and local patches for pressure transitions. Full contrasts and task-level MIMIC-III interaction curves appear in Appendix C.3.

5 Physiological Context and Advance Warning

Figure 4 compares observation lengths and blank forecast gaps across the three datasets (RQ2). All panels report test AUROC. VitalDB macros exclude the concurrent BIS probe. Protocols and separate validation results appear in Appendices D and F.2.

Observation length. Extending MIMIC-III history from ten to 60 minutes changes macro AUROC from 0.606 to 0.609: respiratory escalation improves from 0.575 to 0.593, while mortality remains near 0.517 (Figure 4(a)). On VitalDB, extending five minutes to 60 improves hypotension (0.689 to 0.727) and gas discordance (0.560 to 0.594), but lowers oxygen-desaturation AUROC (0.635 to 0.625). The pressure-specific gain aligns with predictive pre-event arterial-waveform patterns (Hatib et al., 2018). Longer history does not benefit all transitions equally. MC-MED's macro peaks at ten minutes (0.578), then falls at 15 minutes (0.567): lactate nonclearance improves from 0.502 to 0.577, while the other three tasks decline. The lactate-specific gain accords with its serial-measurement endpoint and evolving production, clearance, and treatment history (Vincent et al., 2016), whereas pending hyperkalemia and anemia concern tests ordered before prediction. These matched-anchor comparisons favor endpoint-specific history lengths (Appendix D).

Warning time[2]**.** Training, validation, and test patient partitions remain fixed across warning gaps. For MIMIC-III, mortality AUROC decreases from 0.545 at three hours to 0.537 at 24 hours, whereas AKI peaks at six hours (0.697) and Sepsis-3 recovers after a four-hour dip. AKI's intermediate peak

[2]Also called forecast gap.

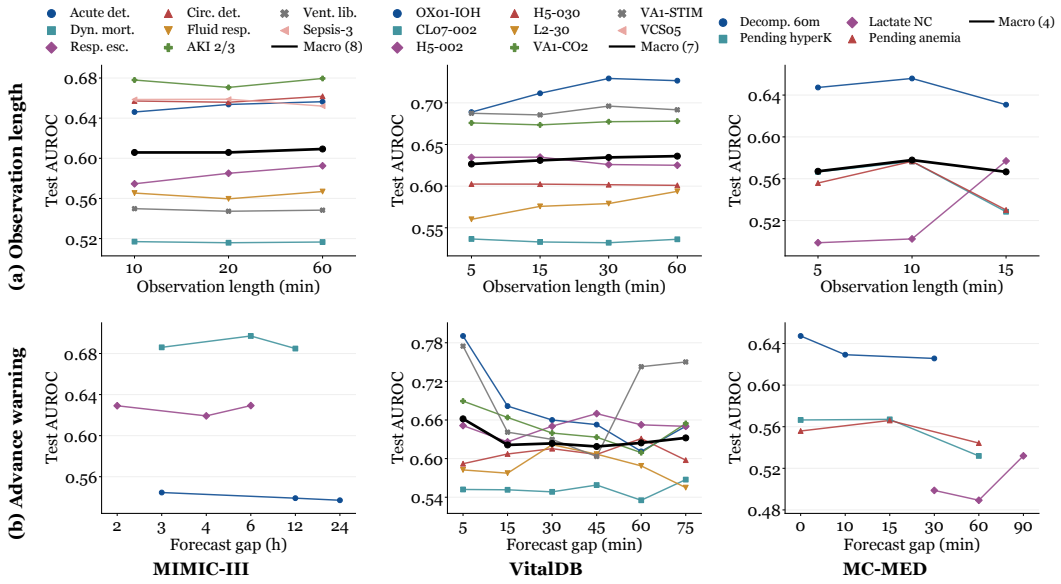


Figure 4: **Physiological context and advance warning.** (a) Nested observation intervals ending at matched anchors. (b) Task-specific forecast gaps with fixed patient partitions. In the MIMIC-III warning panel, blue circles denote mortality, teal squares AKI, and purple diamonds Sepsis-3. Task-level values are in Tables 39–44.

accords with renal criteria assessed over time (KDIGO Acute Kidney Injury Work Group, 2012): shifting outcome windows changes diagnostic timing alongside warning distance. For VitalDB, macro AUROC falls from 0.662 at five minutes to 0.621 at 15 minutes, then ranges from 0.619 to 0.632. This range conceals coupled-response recovery from 0.603 at 45 minutes to 0.750 at 75 minutes, while gas discordance declines from 0.622 at 30 minutes to 0.555 at 75 minutes. We interpret the coupled-response rebound and 60-minute hypothermia peak as reflecting changing surgical stimulation and heat balance (Ledowski, 2019; Sessler, 2016). For MC-MED, decompensation falls from 0.647 at zero gap to 0.626 at 30 minutes. Hyperkalemia and anemia retain stronger discrimination at zero or 15 minutes than at 60 minutes. Useful warning ranges differ between imminent vital-sign transitions and laboratory-defined outcomes. Comparisons must preserve required intervention intervals and account for shifted outcome windows.

6 Information Availability and Multimodal Value

Figure 5 compares channel selection, missing-input handling, waveform–EHR fusion, training subsets, and MIMIC-IV transfer (Johnson et al., 2023) (RQ3). AUROC references are panel-specific (Appendix E). Channel selectors include mutual information (MI) (Peng et al., 2005), Lasso (Tibshirani, 1996), and signal quality index (SQI) (Clifford et al., 2012), plus random, clinical-priority, and temporal-MI variants. Missing-input methods include zero-mask plus time-gap encoding (Che et al., 2018), GRU-D (Che et al., 2018), GRIN (Cini et al., 2022), Raindrop (Zhang et al., 2022), positional kriging adapted from IGNNK (Wu et al., 2021), lead-agnostic SSL (Oh et al., 2022), and adaptations of Rectified Flow (Liu et al., 2023), TrigFlow (Lu & Song, 2025), and Drifting (Deng et al., 2026). Subset selectors use diversity coverage (Sener & Savarese, 2018), diversity clustering (McInnes et al., 2018; Sener & Savarese, 2018), gradient matching (Mirzasoleiman et al., 2020; Killamsetty et al., 2021), an NLDR-inspired policy (Vazifeh & Fleischer, 2026), and training dynamics (Swayamdipta et al., 2020; Paul et al., 2021), plus benchmark-specific noise filtering and task-anchor heuristics. Appendix B.8 details benchmark adaptations.

❶ **Sensors and missingness.** Clinical/SQI selection preserves MIMIC-III macro AUROC with ten channels. All VitalDB selectors lower it, with temporal MI closest to the all-channel reference (Figure 5(a,b)). For MIMIC-III missing inputs, lead-agnostic SSL and GRU-D lead the macro, but all eight alternatives to zero-mask plus Δ lower AKI and fluid-response AUROC. On VitalDB, Raindrop approximately matches the native-handling macro, eight alternatives lower it, and all nine reduce desaturation AUROC. GRIN's cross-sensor reconstruction (Cini et al., 2022) improves gas

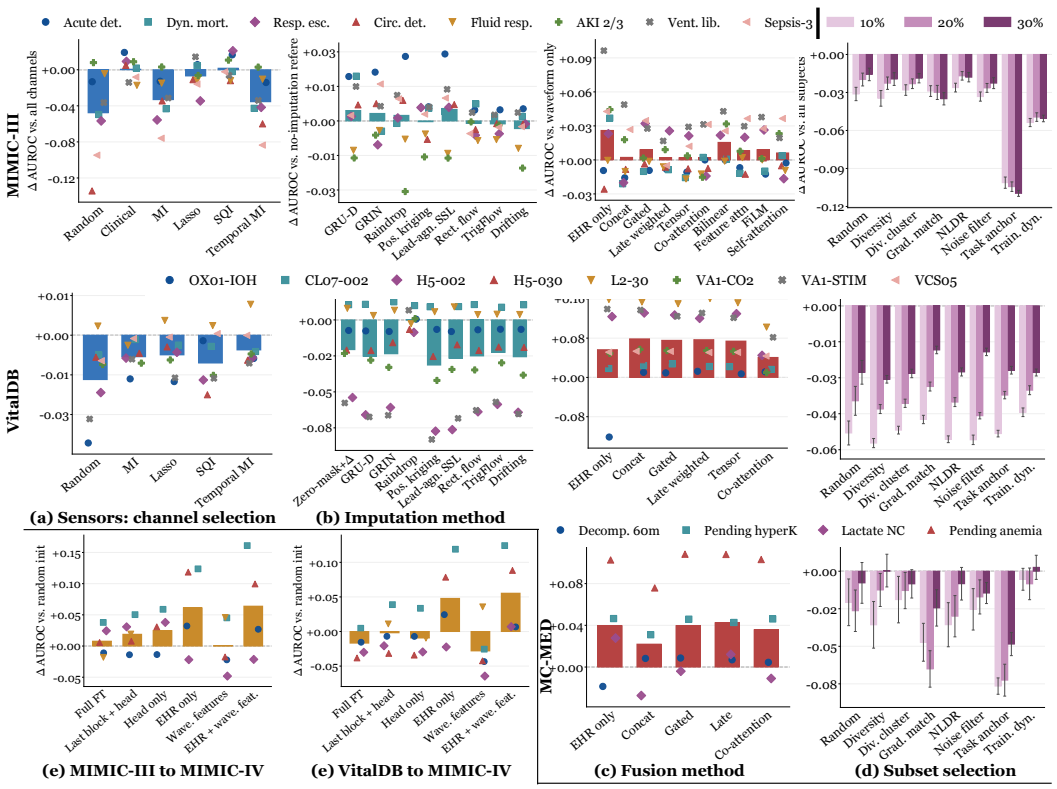


Figure 5: **Information availability and multimodal value.** AUROC changes for (a) sensor selection, (b) missing inputs, (c) waveform–EHR fusion, (d) training-data budgets, and (e) transfer to MIMIC-IV in the two lower-left panels. Light-to-dark shades order window budgets for MIMIC-III and subject budgets for VitalDB and MC-MED (Appendix E.3). Bars summarize task averages. Numerical results are in Tables 66–75 and 78.

discordance by 0.007 but lowers desaturation and coupled-response AUROC by 0.061 and 0.066. Gains for the ventilation–perfusion relation (Nassar & Schmidt, 2016) do not extend to these transitions. MC-MED omits these comparisons with only two waveform channels. In a separate ensemble test, 50% sensor dropout lowers AUROC from 0.619 to 0.583: handling natural missingness does not ensure resilience to new sensor loss (Appendix E.1).

❷ **Multimodal fusion.** Bilinear fusion raises MIMIC-III's waveform-only macro most (0.606 to 0.622), below EHR-only (0.632). VitalDB concatenation leads both unimodal models (0.712 versus waveform-only 0.634 and EHR-only 0.690, Figure 5(c)). All VitalDB fusion methods improve hypotension, supporting concatenation or prediction weighting without cross-modal attention. The waveform branch retains value for MIMIC-III circulatory deterioration and VitalDB hypotension, consistent with hemodynamic changes preceding circulatory events (Hyland et al., 2020; Hatib et al., 2018). On MC-MED, late fusion leads the macro (0.609), but its lactate AUROC remains below EHR-only (0.520 versus 0.536), consistent with the metabolic and treatment history governing lactate kinetics (Vincent et al., 2016). Practical fusion is bilinear fusion on MIMIC-III, concatenation or weighted late fusion on VitalDB, and late fusion on MC-MED. EHR-only controls assess the waveform contribution beyond clinical context (Appendix E.2).

❸ **Training-data budgets.** Full training data give the highest MIMIC-III macro and VitalDB forecast-macro AUROC. At the largest tested subsets, random selection leads MIMIC-III (30% of windows), while gradient matching and noise filtering lead VitalDB (20% of subjects). MIMIC-III task-anchor selection worsens with larger quotas: prioritizing transitions and hard negatives remains weaker than random sampling (Figure 5(d)). On MC-MED, training dynamics and diversity at 50% of subjects slightly exceed the full-cohort mean (0.5695 and 0.5673 versus 0.5669). In a separate fixed-recipe VitalDB test, 20% gradient-matched subjects preserve concurrent BIS AUROC (0.884 versus 0.883) but reduce forecast AUROC from 0.630 to 0.578. Approximating full-data gradients through selection (Killamsetty et al., 2021) better preserves current anesthetic-state ranking than pre-event warning information. BIS AUPRC nevertheless falls from 0.606 to 0.592 (Appendix E.3).

❹ **Transfer to MIMIC-IV.** For supervised Transformer/Conv1D transfer to five MIMIC-IV targets, head-only adaptation gives the highest tuning macro with MIMIC-III initialization. VitalDB initialization offers no consistent macro benefit (Figure 5(e)). Learning a target decision boundary from fixed source features gives the largest aggregate gain without encoder updates. On circulatory deterioration, head-only adaptation beats random initialization, while logistic regression on the same frozen features falls below it. Waveform adaptation outperforms EHR alone for circulatory deterioration, whereas EHR accounts for most combined-feature gains on AKI and delirium. Hemodynamic monitoring captures circulatory instability, while AKI is defined by renal measurements (Hyland et al., 2020; KDIGO Acute Kidney Injury Work Group, 2012) (Appendix E.5).

Efficacy, efficiency, and safety. MIMIC-III clinical channel selection retains AUPRC (0.096 versus 0.098) and reduces peak memory (2.353% versus 3.395%), but increases ECE (0.228 versus 0.178). On VitalDB, weighted late and gated fusion have similar AUPRC (0.1255 and 0.1256) but different ECE (0.0251 and 0.0306). Selecting configurations by AUROC rewards correct ordering without directly optimizing calibration. Sensor and fusion choices can therefore preserve discrimination while changing probability accuracy. For risk-based bedside alerts, ranking patients correctly and assigning reliable probabilities for clinical review are distinct requirements (Van Calster et al., 2019). Companion results appear in Appendices C.4, D.1, and E.6. Definitions are in Appendix A.2.

(RQ1) Representation and architecture: Tokenizer and encoder choices change model rankings by endpoint. Conv1D leads the VitalDB and MC-MED tokenizer macros, while Transformer leads the MIMIC-III and VitalDB architecture macros. MC-MED favors BiLSTM for decompensation and hyperkalemia and DLinear for anemia. **(RQ2) Temporal modeling:** Longer histories offer task-dependent gains that often plateau. Longer warning gaps reduce discrimination for some imminent events but produce nonmonotonic effects elsewhere. Patient partitions stay fixed while shifted outcome windows can change eligibility and prevalence. **(RQ3) Information availability:** Full training sets lead MIMIC-III macro and VitalDB forecast-macro AUROC. Clinical-priority and SQI selection preserve MIMIC-III performance with fewer sensors, whereas VitalDB favors all channels. Missing-input handling remains task-dependent: Raindrop matches VitalDB's native-handling macro, but all added methods lower desaturation AUROC. Detailed comparisons appear in Appendices C–E. Appendix H gives task-specific recommendations.

MoE with self-supervised pretraining. To evaluate our benchmark as a foundation for model development, we combine mixture-of-experts (MoE) routing (Fedus et al., 2022) with self-supervised learning (SSL) (Oh et al., 2022) in **MoE+SSL**. SSL learns representations from unlabeled waveforms, while an input-dependent router combines expert outputs for clinical prediction. Pretraining uses training patients only, retaining the benchmark's splits, information cutoffs, and validation-based selection. We evaluate five-minute waveform inputs on six surveillance-style MIMIC-III tasks: acute and circulatory deterioration, respiratory escalation, sepsis onset, ventilator liberation, and dynamic mortality. Figure 6 shows test AUROC gains on all six tasks over the parameter-matched baseline, ranging from $+0.006$ to $+0.045$, with a macro gain of $+0.033$.

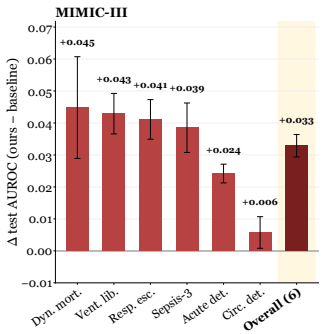


Figure 6: AUROC gains over the parameter-matched baseline: six MIMIC-III tasks and macro, five-seed means. Bars show SE of the difference.

7 Conclusion

Actionwave benchmarks representation, temporal context, and information availability across 20 clinical waveform classification tasks. Endpoint differences persist across model families: sepsis drives SeFT's MIMIC-III lead, while FFT lowers MC-MED decompensation and hyperkalemia AUROC across all nine encoders. History and warning time produce opposing task trends. Fusion and transfer reveal complementary EHR information for hypotension and predominantly EHR-driven gains for AKI and delirium. These task-level comparisons align waveform representations and training configurations with each endpoint's clinical information and intervention time. The clinically defined events and states support bedside screening, warning decisions, and comparisons across care settings. Future work will extend to regression tasks that estimate physiological trajectories and quantify the magnitude of clinical change.

AI USE STATEMENT

Generative AI tools were used only to correct typographical errors in the manuscript. The authors take responsibility for the final text, scientific claims, and reported results.

ETHICS STATEMENT

Actionwave is a retrospective benchmark using existing de-identified clinical datasets: MIMIC-III, VitalDB, and MC-MED, with additional transfer experiments on MIMIC-IV (Moody et al., 2020; Lee et al., 2022; Kansal et al., 2025; Johnson et al., 2023). The study analyzes previously collected records and involves no new patient recruitment or clinical intervention. Access to and reuse of the source data remain subject to each provider's access conditions and data-use terms. Reproducing the benchmark requires obtaining the relevant data from those providers.

REPRODUCIBILITY STATEMENT

Section 3 specifies the prediction interface, information cutoffs, patient-level partitioning, and rules for controlled comparisons. Appendix B details task definitions, eligibility, preprocessing, and cohort support. Appendix B.7 documents training settings, search grids, random seeds, and validation-based model selection, with method adaptations in Appendix B.8. Appendix A.2 defines metrics, aggregation, uncertainty estimates, and resource measurements. Appendices D and E describe the temporal and information-availability comparisons, including the MIMIC-IV transfer protocol in Appendix E.5. Appendix I provides the numerical results underlying the reported comparisons.

REFERENCES

Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361:k1479, 2018. doi: 10.1136/bmj.k1479.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv preprint arXiv:1803.01271*, 2018. URL https://arxiv.org/abs/1803.01271.

Brett K Beaulieu-Jones, William Yuan, Gabriel A Brat, Andrew L Beam, Griffin Weber, Marshall Ruffin, and Isaac S Kohane. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ digital medicine*, 4(1):62, 2021. doi: 10.1038/s41746-021-00426-3.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Rachel Reisler, David A. Kim, and Pranav Rajpurkar. Multimodal clinical benchmark for emergency care (MC-BEC): A comprehensive benchmark for evaluating foundation models in emergency medicine. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45794–45811, 2023.

Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the Gaps: Multivariate Time Series Imputation by Graph Neural Networks. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2108.00298.

G. D. Clifford, J. Behar, Q. Li, and I. Rezek. Signal Quality Indices and Data Fusion for Determining Clinical Acceptability of Electrocardiograms. *Physiological Measurement*, 33(9):1419–1433, 2012. doi: 10.1088/0967-3334/33/9/1419.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative Modeling via Drifting. *arXiv preprint arXiv:2602.04770*, 2026. URL https://arxiv.org/abs/2602.04770.

Johannes Enevoldsen and Simon T Vistisen. Performance of the Hypotension Prediction Index May Be Overestimated Due to Selection Bias. *Anesthesiology*, 137(3):283–289, 2022. doi: 10.1097/ALN.0000000000004320.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://www.jmlr.org/papers/v23/21-0998.html.

Conner D. Galloway, Alexander V. Valys, Jacqueline B. Shreibati, Daniel L. Treiman, Frank L. Petterson, Vivek P. Gundotra, David E. Albert, Zachi I. Attia, Rickey E. Carter, Samuel J. Asirvatham, Michael J. Ackerman, Peter A. Noseworthy, John J. Dillon, and Paul A. Friedman. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiology*, 4(5):428–436, 2019. doi: 10.1001/jamacardio.2019.0640.

Alex Graves and Jürgen Schmidhuber. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks*, 18(5–6):602–610, 2005. doi: 10.1016/j.neunet.2005.06.042. URL https://www.cs.toronto.edu/~graves/nn_2005.pdf.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.00396.

Fredrik K. Gustafsson, Xiao Gu, Mattia Carletti, Patitapaban Palo, David W. Eyre, and David A. Clifton. SignalMC-MED: A multimodal benchmark for evaluating biosignal foundation models on single-lead ECG and PPG. *arXiv preprint arXiv:2603.09940*, 2026.

J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:96, 2019. doi: 10.1038/s41597-019-0103-9.

Feras Hatib, Zhongping Jian, Sai Buddi, Christine Lee, Jos Settels, Karen Sibert, Joseph Rinehart, and Maxime Cannesson. Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. *Anesthesiology*, 129(4):663–674, 2018. ISSN 1528-1175. doi: 10.1097/ALN.0000000000002300.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set Functions for Time Series. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4353–4363. PMLR, 2020. URL https://proceedings.mlr.press/v119/horn20a.html.

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020. doi: 10.1038/s41591-020-0789-4.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for Time Series Classification. *Data Mining and Knowledge Discovery*, 2020. URL https://arxiv.org/abs/1909.04939.

Alistair E W Johnson, Jerome Aboab, Jesse D Raffa, Tom J Pollard, Rodrigo O Deliberato, Leo A Celi, and David J Stone. A Comparative Analysis of Sepsis Identification Methods in an Electronic Database. *Critical care medicine*, 46(4):494–499, 2018. doi: 10.1097/CCM.0000000000002965.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A. Kim. MC-MED, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12:1094, 2025. doi: 10.1038/s41597-025-05419-5.

KDIGO Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney International Supplements*, 2(1):1–138, 2012. URL https://kdigo.org/wp-content/uploads/2016/10/KDIGO-2012-AKI-Guideline-English.pdf.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient Matching based Data Subset Selection for Efficient Deep Model Training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5464–5474, 2021. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Joon-Myoung Kwon, Younghoon Cho, Ki-Hyun Jeon, Soohyun Cho, Kyung-Hee Kim, Seung Don Baek, Soomin Jeung, Jinsik Park, and Byung-Hee Oh. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. *The Lancet Digital Health*, 2(7):e358–e367, 2020. doi: 10.1016/S2589-7500(20)30108-4.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved Sequence Modeling using State Space Principles. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.15569.

Simon Meyer Lauritsen, Bo Thiesson, Marianne Johansson Jørgensen, Anders Hammerich Riis, Ulrick Skipper Espelund, Jesper Bo Weile, and Jeppe Lange. The Framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. *npj Digital Medicine*, 4(1):158, 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00529-x.

Thomas Ledowski. Objective monitoring of nociception: a review of current commercial solutions. *British journal of anaesthesia*, 123(2):e312–e321, 2019. doi: 10.1016/j.bja.2019.03.024.

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9:279, 2022. doi: 10.1038/s41597-022-01411-5.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

Juan Miguel Lopez Alcaraz, Hjalmar Bouma, and Nils Strodthoff. Enhancing clinical decision support with physiological waveforms—a multimodal benchmark in emergency care. *Computers in Biology and Medicine*, 192:110196, 2025. doi: 10.1016/j.compbiomed.2025.110196.

Cheng Lu and Yang Song. Simplifying, Stabilizing and Scaling Continuous-Time Consistency Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7e9c2053258b1bdd32ff2654802cd594-Abstract-Conference.html.

Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749–760, 2018. doi: 10.1038/s41551-018-0304-0.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*, 2018. URL https://arxiv.org/abs/1802.03426.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for Data-efficient Training of Machine Learning Models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6950–6960, 2020. URL https://proceedings.mlr.press/v119/mirzasoleiman20a.html.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III waveform database matched subset. *PhysioNet*, 2020. doi: 10.13026/c2294b. Version 1.0.

Boulos S Nassar and Gregory A Schmidt. Capnography During Critical Illness. *Chest*, 149(2): 576–585, 2016. doi: 10.1378/chest.15-1369.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic Self-supervised Learning for Local and Global Representations of Electrocardiogram. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174, pp. 338–353, 2022. URL https://proceedings.mlr.press/v174/oh22a.html.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep Learning on a Data Diet: Finding Important Examples Early in Training. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2107.07075.

Hanchuan Peng, Fuhui Long, and Chris 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(8):1226–1238, 2005. doi: 10.1109/TPAMI.2005.159. URL https://home.penglab.com/papersall/docpdf/2005_TPAMI_FeaSel.pdf.

Charalampos Pierrakos, Dimitrios Velissaris, Sabino Scolletta, Sarah Heenen, Daniel De Backer, and Jean-Louis Vincent. Can changes in arterial pressure be used to detect changes in cardiac index during fluid challenge in patients with septic shock? *Intensive care medicine*, 38(3):422–428, 2012. doi: 10.1007/s00134-011-2457-0.

Richard D Riley, Joie Ensor, Kym I E Snell, Frank E Harrell, Jr., Glen P Martin, Johannes B Reitsma, Karel G M Moons, Gary Collins, and Maarten van Smeden. Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368:m441, 2020. doi: 10.1136/bmj.m441.

Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, and Konrad P Kording. The need to approximate the use-case in clinical machine learning. *GigaScience*, 6(5):1–9, 2017. doi: 10.1093/gigascience/gix019.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Ozan Sener and Silvio Savarese. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1708.00489.

Daniel I Sessler. Perioperative thermoregulation and heat balance. *The Lancet*, 387(10038):2655–2664, 2016. doi: 10.1016/S0140-6736(15)00981-2.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-Time Attention Networks for Irregularly Sampled Time Series. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.10318.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016. doi: 10.1001/jama.2016.0287.

Sameer Sundrani, Julie Chen, Boyang Tom Jin, Zahra Shakeri Hossein Abad, Pranav Rajpurkar, and David Kim. Predicting patient decompensation from continuous physiologic monitoring in the emergency department. *NPJ digital medicine*, 6(1):60, 2023. doi: 10.1038/s41746-023-00803-0.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://arxiv.org/abs/2009.10795.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Sindhu Tipirneni and Chandan K. Reddy. Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series. *ACM Transactions on Knowledge Discovery from Data*, 16(6), 2022. doi: 10.1145/3516367. URL https://arxiv.org/abs/2107.14293.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían O Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R Baker, Kelly Peterson, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Joseph R Ledsam, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019. doi: 10.1038/s41586-019-1390-1.

Flavia Torrini, Ségolène Gendreau, Johanna Morel, Guillaume Carteaux, Arnaud W Thille, Massimo Antonelli, and Armand Mekontso Dessap. Prediction of extubation outcome in critically ill patients: a systematic review and meta-analysis. *Critical care (London, England)*, 25(1):391, 2021. doi: 10.1186/s13054-021-03802-3.

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: The achilles heel of predictive analytics. *BMC Medicine*, 17:230, 2019. doi: 10.1186/s12916-019-1466-7.

Robin van de Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another ICU benchmark: A flexible multi-center framework for clinical ML. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c26a89073d972f2e6643617b0f3a9e8a-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762.

Amir Reza Vazifeh and Jason W. Fleischer. Manifold Learning for Personalized and Label-Free Detection of Cardiac Arrhythmias. *Informatics in Medicine Unlocked*, 64:101770, 2026. doi: 10.1016/j.imu.2026.101770. URL https://arxiv.org/abs/2506.16494.

Jean-Louis Vincent, Amanda Quintairos E Silva, Lúcio Couto, Jr., and Fabio S Taccone. The value of blood lactate kinetics in critically ill patients: a systematic review. *Critical care (London, England)*, 20(1):257, 2016. doi: 10.1186/s13054-016-1403-5.

Weinan Wang, Pedram Mohseni, Kevin L. Kilgore, and Laleh Najafizadeh. PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023. doi: 10.3389/fdgth.2022.1090854.

Zhiguang Wang, Weizhong Yan, and Tim Oates. Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline. In *International Joint Conference on Neural Networks*, pp. 1578–1585, 2017. doi: 10.1109/IJCNN.2017.7966039. URL https://arxiv.org/abs/1611.06455.

Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. Inductive Graph Neural Networks for Spatiotemporal Kriging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4478–4485, 2021. doi: 10.1609/aaai.v35i5.16575. URL https://ojs.aaai.org/index.php/AAAI/article/view/16575.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling Time Series with 10k Parameters. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/701251e1db4a2e4dd2ef23f5265d5936-Abstract-Conference.html.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Defu Lian, Ning An, Longbing Cao, and Zhendong Niu. Frequency-domain MLPs are more effective learners in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f1d16af76939f476b5f040fd1398c0a3-Abstract-Conference.html.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, volume 37 of *AAAI'23/IAAI'23/EAAI'23*, pp. 11121–11128. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i9.26317.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-Guided Network for Irregularly Sampled Multivariate Time Series. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.05357.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vSVLM2j9eie.

A Benchmark Definition Details

A.1 Clinical-Inference and Temporal-Geometry Taxonomies

Clinical-inference taxonomy. We group tasks by their primary clinical inference target, rather than by dataset, input modality, or anatomical system alone. Acute homeostatic failure and safety-boundary transitions concern impending changes from relative stability to clinically actionable instability. Pathophysiologic state and organ dysfunction concern the onset, persistence, or presence of a defined biological or organ-level abnormality. Functional reserve and intervention response concern responses to treatment, physiological challenge, or withdrawal of organ support. Patient trajectory and care-disposition outcomes cover downstream survival and healthcare utilization. Organ systems provide a secondary grouping; dataset, prediction horizon, and task formulation describe other aspects of each task. We assign each task by its primary endpoint construct. The dataset-specific registries below define each endpoint, including those for the four selected MC-MED tasks.

Temporal-geometry taxonomy. Temporal geometry describes when observations and targets occur relative to the prediction anchor T. ❶ *Surveillance* issues predictions at repeated eligible anchors as new observations accumulate. ❷ *Gapped forecasts* use a nonzero G to make the minimum warning time explicit. ❸ *Event- and daily-anchored predictions* align T to a recorded action, measurement, or recurring care time rather than a uniform schedule. ❹ *Concurrent held-out-state probes* withhold a target state or channel within the observation interval and measure shared information instead of advance warning; they are not pooled with future-event forecasts. The main comparison uses one primary endpoint and geometry per task. Alternative heads, gaps, and horizons are evaluated in sensitivity analyses.

A.2 Evaluation and Aggregation

Metrics and uncertainty. The numeric supplement reports AUROC, AUPRC, Brier score, log loss, and expected calibration error (ECE) for the completed VitalDB analyses. Forecast summaries give equal weight to gaps within each task and then to the seven forecasting tasks. Concurrent BIS is reported separately.

For validation, we pool predictions from disjoint held folds within each seed, compute a metric for each seed, and average the five values. Fixed-fold test summaries average metrics from the five frozen fold models. The separately labeled ensemble-test stress analysis instead averages model probabilities before computing metrics. The original subject-selection panel averages saved seed–fold validation metrics for both the selected subsets and the full-data comparator. The later gradient-matching test uses individual model metrics and is reported separately in Appendix E.3.

Gap and fusion intervals use 2,000 shared whole-subject bootstrap draws, keeping repeated episodes from each subject together. These pointwise intervals are conditional on the fitted models: they do not account for retraining or selection uncertainty and are not adjusted for multiplicity. The 25 seed–fold fits are not 25 independent patient samples. To compare fusion with the stronger unimodal envelope, we select the better unimodal metric separately for each head and draw before aggregation. This envelope is a comparison rule, not a single deployable model.

Runtime and reliability. We report training wall time, peak accelerator memory, and time per training sample for the precision, schedule, and workload. Failure summaries use the recorded nonfinite-loss, gradient, and prediction counters; their measurement scope is specified below.

Companion efficacy, efficiency, and safety metrics. The companion figures in Appendices C.4, D.1, and E.6 report AUPRC, training cost, calibration, selective risk, and numerical failures. AUPRC summarizes precision–recall performance and depends on outcome prevalence. Observation-length comparisons report mean training time per optimizer step in milliseconds, computed for each fit as $1000\, t_{\mathrm{train}}/N_{\mathrm{updates}}$ before task and seed averaging. The MC-MED records provide both quantities for all 60 fits at batch size 128; other observation-length timing entries remain unreported until their update counts are verified. Other experiments retain training wall time per run on MIMIC-III and MC-MED, or per validation training cell on VitalDB. MIMIC-III memory is peak accelerator memory as a percentage of device capacity; the refreshed MC-MED temporal figures instead report sampled shared-device memory in MiB, which does not isolate a model's memory use. The plotted latency is $1000/$training throughput in milliseconds per training sample, not inference or

alert-delivery latency. ECE uses 15 bins, and AURC integrates selective-prediction risk over coverage; lower values are preferable for both. MIMIC-III and MC-MED failure rates use the available nonfinite-loss, gradient, or prediction records. For the refreshed MC-MED temporal figures, the failure rate is failed training attempts divided by started attempts, using the campaign audit (zero of 10,112 attempts); the same audit records zero nonfinite events across 5,320,037 training steps. For VitalDB, the audited rate of nonfinite or out-of-range prediction files among completed selected models is zero of 78,700 test files; all-attempt training-failure counters were not recorded.

Bars average available task means when conditions cover the same tasks; VitalDB forecast macros exclude concurrent VCS05. Colored markers show individual tasks; grouped bars distinguish subject budgets from light (smallest) to dark (largest). Hollow marks identify validation metrics, and filled cost bars show training telemetry. Task markers replace macro bars when task coverage differs across conditions. Panels show only recorded metrics and omit missing measurements. The VitalDB sensor-selection figure pools the available two-, four-, and eight-channel results under each policy. The source tables propagate task seed SDs as $\sqrt{\sum_t s_t^2}/K$ for K tasks; this summary omits cross-task covariance and is not a patient-level confidence interval. The per-step timing table instead averages the four task values within each seed and reports the mean and sample SD of the five seed-level macros.

Selection and measurement scope. Checkpoint selection and search grids follow Appendix B.7. VitalDB ECE and AURC come from the September 15 safety supplement, except fusion ECE, which retains the original fusion export used in the main text. MC-MED representation and fusion safety metrics are recomputed from saved test predictions; earlier-study calibration estimates remain separately identified in Appendix F. VitalDB training telemetry comes from earlier grids, whereas its predictive metrics come from the updated test release; cost comparisons are within device and workload, not across datasets. The MC-MED representation companion retains the earlier campaign, including its additional FFT sweep; Table 10 and the AUROC dot plots use the September 17 update. Its AUPRC and reliability results therefore describe that earlier campaign, rather than the updated AUROC configurations. The MIMIC-III companion and updated main AUROC table also use different Conv1D search summaries and some different tokenizer seed statistics. The temporal companions now use the September 17 collection corresponding to Figure 4: duration comparisons average three MIMIC-III and five VitalDB encoder–tokenizer pairs, and MC-MED uses the refreshed ResNet1D/Conv2D campaign. MC-MED temporal ECE and AURC cover all four endpoints; duration figures report their macro, and gap figures show task-specific values. Information-availability companions likewise retain earlier exports: their MIMIC-III missingness reference is architecture-only and their subset budgets are reported by subject, whereas the refreshed AUROC study uses zero-mask plus Δ and training-window quotas, respectively. MC-MED fusion AUPRC, training time, and failure records cover all four endpoints; its ECE and AURC panels report decompensation. MC-MED subject-budget results and the full-cohort AUPRC reference use five-seed validation means.

A.3 Experiment Count

We count one model fit for each task/head, experimental condition, hyperparameter setting, training seed, and subject fold. The reported grids imply a conservative total of 87,825 fits (Table 4): 5,033 task/head-specific conditions, including learning-rate choices, before repetitions across seeds and folds. MIMIC-III contributes eight task heads and five seeds, MC-MED four heads and five seeds, and the VitalDB core grid 15 heads and five seeds across five subject folds. Thus, a condition evaluated across a dataset's core heads contributes 40, 20, or 375 fits, respectively. We include the specified learning-rate searches and count overlapping search conditions and shared reference models only once. Where the search grid is unspecified, we count one learning-rate setting. We exclude additional unreported tuning trials, restarts, bootstrap draws, and inference-only controls.

MIMIC-III and VitalDB each contain $9 \times 3 + 4 = 31$ encoder–tokenizer configurations, including the four Transformer tokenizers outside their architecture grids. MC-MED contains $9 \times 3 + 6 + 8 = 41$ configurations: 27 native architecture-grid pairs, six additional Transformer configurations, and eight additional non-Transformer FFT pairs. Separately trained configurations within the same tokenizer family count as separate fits. For MIMIC-III, the architecture search contains $9 \times 3 \times 3 = 81$ pair–learning-rate conditions, and the tokenizer search contains $6 \times 5 + 3 = 33$. The two searches share eight conditions: two Conv1D learning rates and three each for Crossformer and SeFT with a

Transformer. Their union gives $(81 + 33 - 8) \times 8 \times 5 = 4,240$ fits. VitalDB searches three rates for each of its 31 pairs, giving $31 \times 3 \times 15 \times 5 \times 5 = 34,875$ fits. MC-MED also searches three learning rates for each of its 41 configurations, giving $41 \times 3 \times 4 \times 5 = 2,460$ representation fits.

The history count retains the original inventory of MIMIC-III fits at 10, 20, and 60 minutes, two additional MC-MED durations, and three additional durations for each of five VitalDB pairs (Appendices D and F.1). The refreshed MIMIC-III curve excludes the earlier five-minute point, which came from a separate campaign. Only the 700 explicitly new VitalDB longer-gap fits are added; the 350 reused fits and the MIMIC-III and MC-MED gap extensions are excluded from this lower bound (Appendix F.2).

The documented MIMIC-III and MC-MED grids contribute six channel-selection alternatives and nine missing-input methods per task. The workload tally retains these original MC-MED grids, although their result panels are omitted because the current input configuration has only two available waveform channels. The fuller VitalDB grids contribute $48 - 3 = 45$ channel conditions and $30 - 3 = 27$ missingness conditions after removing the three fixed-pair references (Appendix F.1); their subsets in the main figure are not counted again. Subject selection contributes seven methods at three fractions per dataset, excluding the full-data reference. The refreshed eight-selector grids and added MIMIC-III EHR-only control are not added to this conservative fit inventory. The counted fusion grids contain nine MIMIC-III methods and five fusion methods plus EHR only on VitalDB and MC-MED; waveform-only references are not added again. This inventory is broader than the current MC-MED fusion display, which contains four fusion methods plus EHR only (Table 9); transfer-to-MIMIC-IV fits are not added to this lower bound.

B Dataset and Task Details

B.1 Episode Construction and Data-Control Details

Anchors and temporal windows. For each task, we generate candidate decision anchors T, then construct the observation interval $[T - W, T)$, blank forecast gap G, and outcome horizon H. MIMIC-III combines sliding-surveillance, daily, and event anchors; VitalDB separates future-event anchors from the concurrent BIS probe; and MC-MED applies the task-specific eligibility rules and target windows described in Appendix B.5. Observation intervals are fixed independently of future labels to prevent post-cutoff information from entering the inputs. For MIMIC-III, we extract the six-hour history $[T - 6\,\mathrm{h}, T)$ and construct each input as its suffix $[T - W, T)$: five minutes uses $[T - 5\,\mathrm{min}, T)$, and one hour uses $[T - 1\,\mathrm{h}, T)$. Observation-length comparisons retain the same anchors and labels across these nested inputs.

Eligibility and ascertainment. Eligibility rules determine whether an anchor can be assigned an interpretable label. For incident-event tasks, a prevalent event before T excludes the episode, an event during G is excluded or masked according to the task contract, and only a qualifying event during H receives a positive label. A negative label requires sufficient follow-up and target availability to establish that the event did not occur; missing, stale, or incompletely ascertained outcomes remain masked. Response tasks instead anchor on a documented action and apply prespecified baseline, response, and non-overlap rules. Repeated or overlapping episodes remain valid longitudinal observations but are not treated as independent patients in aggregation or uncertainty estimation.

Splits and missingness. We assign subjects to partitions before constructing model inputs. Repeated surgical cases from the same VitalDB subject remain in the same partition, and no subject contributes waveform fragments or prediction anchors to more than one split. Patient split and fold assignments remain fixed across observation lengths and warning gaps. Raw waveform samples enter the model tokenizer with channel-presence and sample-validity masks; missingness is recorded explicitly and is not converted into a negative observation. We fit normalization statistics, vocabularies, and learned quality filters on the training partition only.

Leakage controls and auditability. Temporal and feature-level controls limit inputs to information available at the decision time. Every EHR value must be available before the applicable input cutoff, and waveform samples at or after that cutoff are inaccessible to the model. Outcome-derived columns, post-anchor results and interventions, final disposition and diagnoses, realized length of stay, label and censoring fields, and direct or derived target channels are excluded. Documented

pre-anchor treatment or support is retained when it would have been known at the decision time, although an already active intervention can make an episode ineligible for an incident-event task. Each released registry records the dataset version, subject split, anchor, input and label bounds, channel mask, and preprocessing hash so that the temporal audit can be reproduced without opening held-out outcomes.

B.2 Short Task Labels and Full Clinical Names

The main result tables and figures use the short labels listed in Table 3. Aliases identify the same endpoint; the dataset-specific definitions in Appendices B.3, B.4, and B.5 specify its eligibility, timing, and outcome rules. Macro (8), Macro (7), and Macro (4) denote dataset summaries, not additional tasks; VitalDB's forecast macro excludes VCS05. The separate five-target MIMIC-IV transfer labels and full names appear in Table 7; their temporal contracts are specific to the adaptation study in Appendix E.5.

Table 3 lists the full names.

B.3 MIMIC-III Details

The MIMIC-III Waveform Database Matched Subset contains 10,282 adult ICU patients under the frozen 80/10/10 patient-level split. Native waveforms are sampled at 125 Hz and may include ECG, arterial blood pressure, plethysmography, and respiration. Availability varies across channels and time; channels need not be simultaneous, and the benchmark retains presence and validity masks. Each ID below counts as one benchmark task. For tasks with multiple heads, we declare one primary head before evaluation for the main comparison and report all heads in the supplement.

❶**Acute deterioration / failure to rescue** predicts near-term deterioration from the preceding six hours. The primary surveillance task uses a zero-gap six-hour horizon, with a three-hour-gap sensitivity analysis for rescue preparation. ❷**Dynamic in-hospital mortality** predicts death within 24 hours after a six-hour forecast gap from a six-hour waveform history, serving as a severity and representation benchmark rather than a stand-alone paging rule. ❸**Respiratory deterioration and escalation** predicts impending airway or oxygen-support escalation over six hours, with two- and three-hour sensitivities testing whether useful warning persists earlier in the course. ❹**Circulatory deterioration and shock** uses the primary `y_new_pressor` head to predict a new vasopressor start within six hours, with zero gap, among patients off vasopressors at the anchor. Escalation and 24-hour recovery are auxiliary heads from the broader task registry and are excluded from the reported primary-head comparisons. ❺**Rapid documented-fluid response** uses the six hours before a fluid event to predict a documented MAP improvement 30–60 minutes afterward. We interpret this endpoint as a hemodynamic-response proxy, not true fluid responsiveness. ❻**Incident KDIGO stage 2/3 AKI** predicts new moderate-to-severe kidney injury over the subsequent 23 hours after a one-hour gap, supporting earlier renal-risk monitoring among patients not already meeting the endpoint. ❼**Ventilator course and liberation** uses daily and event-anchored predictions to characterize readiness for weaning or extubation and the subsequent need for prolonged ventilation or tracheostomy. ❽**Sepsis-3 onset and shock trajectory** predicts sepsis-related organ dysfunction over the next five hours after a one-hour gap and tracks progression toward shock under an explicit suspected-infection rule.

B.4 VitalDB Details

The waveform-only VitalDB benchmark uses every official waveform stream present in a case and permitted by the task's input rules, drawn from the twelve-stream registry in Table 5. In waveform-only A–D, missingness, and gap experiments, numeric and laboratory tracks define eligibility and outcomes but are not model inputs. Fusion experiments also use EHR predictors, subject to task-specific pre-cutoff availability and target-source exclusions. Source sampling rates range from 62.5 to 500 Hz. For each episode, the hour ending at the task-specific legal input cutoff is divided into twelve consecutive five-minute fragments. The model tokenizer receives raw waveform samples from the selected input interval at the sampling rates recorded in the source headers. The twelve native streams are placed into a fixed 66-slot canonical tensor layout; the remaining slots are structural placeholders rather than additional observed channels. Unrecorded streams and invalid samples re-

main NaN and are accompanied by explicit channel-presence and sample-validity masks. We do not require ECG-II, interpolate across absent streams, or apply complete-case filtering. Channel-wise median and interquartile-range normalization is fitted independently within the four outer-training folds and clipped to $[-5, 5]$. For models that cannot consume NaNs directly, missing normalized values are replaced by neutral zero only after the observation mask has been constructed, and that mask is provided to the model as a separate input.

VitalDB contributes seven future-event tasks and one concurrent held-out-state probe. The core future-event panel uses waveform histories ending strictly before the prediction anchor, gaps of 5 or 15 minutes, and task-specific target horizons of 15 or 60 minutes. The separate 30-minute-input gap extension adds 30, 45, 60, and 75 minutes (Appendix F.2). Direct target channels, target-derived features, and post-cutoff samples are excluded. The concurrent task is analyzed separately and is not pooled with the future-event forecasts. Missing, stale, or incompletely ascertained outcomes are masked rather than treated as negative.

❶**Sustained intraoperative hypotension** predicts a new episode in which invasive MAP remains below 65 mmHg for at least two continuous minutes after five stable pre-anchor minutes at or above 65 mmHg. ❷**Accelerating MAP decline** predicts an accelerating ten-minute decrease in invasive MAP after a stable pre-anchor period. It captures trajectory shape before a conventional hypotension threshold is necessarily crossed. ❸**Incident oxygen desaturation** predicts a sustained $\mathrm{SpO}_2 < 92\%$ episode after five stable minutes at or above 94%, with oxygen-delivery, target-quality, and future-oximetry information excluded from the predictors. ❹**Incident core hypothermia** predicts a newly observed core-temperature episode below $35.5^{\circ}\mathrm{C}$ lasting at least five minutes, subject to explicit temperature freshness and coverage requirements. ❺**Abrupt** ETCO_2 **fall** predicts a decrease of at least 10 mmHg from the recent ETCO_2 baseline that develops within one minute and persists long enough to distinguish it from an isolated sample. ❻PCO_2**–**ETCO_2 **discordance** predicts a newly measured difference of at least 10 mmHg between blood-gas PCO_2 and temporally matched ETCO_2 over a 60-minute target horizon; the label is masked when no qualifying future blood-gas measurement is available. ❼**Coupled hemodynamic response** predicts a joint increase of at least 20 beats/min in heart rate and 20 mmHg in invasive MAP, with the two responses beginning within three minutes and overlapping for at least one minute. This endpoint represents a coupled hemodynamic pattern and is not interpreted as proof of pain, awareness, or inadequate anesthesia. ❽**Concurrent light BIS state** evaluates whether non-BIS physiology identifies a simultaneous, quality-valid BIS value above 60 lasting at least two minutes. The five-minute label interval is concurrent with the most recent five minutes of the waveform input; longer input views add only earlier physiological context. EEG1, EEG2, numeric BIS, BIS SQI, and other label-defining BIS inputs are withheld from every model and selector view. Accordingly, this task is a held-out-state representation probe rather than a forecast and is analyzed separately from the seven future-event tasks.

B.5 MC-MED Details

MC-MED contains 70,545 adult ED patients across 118,385 visits and uses the supplied 78/11/11 chronological, patient-disjoint split; completed representation-comparison test results are reported in Figure 3. Waveform sampling rates are channel-dependent and recorded in each WFDB header. At least one waveform is available in 83,623 visits (70.6%), while monitoring interruptions produce segment gaps. The source provides ECG lead II, plethysmography, and respiration; the native packs used for the reported MC-MED comparisons contain ECG-II and respiration, with channel presence and validity represented explicitly.

The selected MC-MED panel contains four binary prediction tasks, defined relative to prediction time T:

- **Physiological decompensation** (`decompensation_60m`): predict whether an initially stable patient crosses a heart-rate (HR), mean-arterial-pressure (MAP), or oxygen-saturation (SpO_2) deterioration threshold within 60 minutes after T.
- **Pending hyperkalemia** (`pending_hyperkalemia`): predict whether the first result of a potassium test ordered before T is ≥ 5.5 mmol/L during the next 240 minutes.
- **Lactate nonclearance** (`lactate_nonclearance`): after an initial lactate ≥ 2 mmol/L, predict whether the minimum repeat lactate measured 30–240 minutes after T remains ≥ 2 mmol/L and $> 80\%$ of baseline.

- **Pending severe anemia** (`pending_severe_anemia`): predict whether the first result of a hemoglobin test ordered before T is < 8 g/dL during the next 240 minutes.

Decompensation targets an acute safety-boundary transition, while the laboratory endpoints target the presence or persistence of a pathophysiologic abnormality. The pending-test tasks condition on an order placed before the prediction time; their future laboratory results define outcomes and are excluded from model inputs. Lactate nonclearance requires an eligible baseline and repeat measurement; missing target measurements remain unascertained rather than negative. Table 6 reports task-specific patients, ED visits, windows, and positive windows across training, validation, and test.

B.6 Task-level Cohort Support

Table 6 reports the supplied cohort counts for all 20 tasks and all reported development and test partitions. Clinical episodes are ICU stays for MIMIC-III, surgical cases for VitalDB, and ED visits for MC-MED. Within each partition, the union rows count each patient and clinical episode once across tasks. Window counts refer to task-labelled prediction examples. One stay or visit can contribute windows to several tasks; MC-MED has one window per visit *per task*, rather than one across the whole task panel. Positive-window totals also count task-labelled outcomes, not unique patients with positive outcomes.

Across its patient-disjoint train, validation, and test partitions, MIMIC-III contains 8,535 unique patients, 9,772 ICU stays, and 979,920 task windows. VitalDB contains 5,346 unique patients, 5,575 surgical cases, and 385,667 task windows across development and test; its five-fold development cohort is counted once. MC-MED contains 15,852 unique patients, 18,536 ED visits, and 34,379 task windows across all partitions. Together, the benchmark includes 29,733 patients, 33,883 clinical episodes, and 1,399,966 task-labelled windows. Patients and episodes are deduplicated across tasks, while windows count each task and forecast-gap head separately without multiplication by model, fold, seed, or observation-length setting.

B.7 Architecture–Tokenizer Evaluation Protocol

Common protocol. The representation comparisons use waveform-only inputs on all three datasets. Unavailable channels are encoded as NaN with a binary mask; we do not apply complete-case filtering. Each cell of Table 10 gives the mean and sample SD of test AUROC over five seeds (0–4). For macro rows, we first average tasks within each seed, then summarize the five seed-level macros.

For each candidate configuration, validation performance selects the checkpoint epoch; it also selects hyperparameters, including learning rate and tokenizer where applicable. The selected checkpoints are then evaluated on the held-out test set to obtain the reported performance. The dataset-specific search grids are listed below. Bold and underlining indicate the highest and next-highest unrounded means, respectively; distinct values can round to the same displayed score.

Controlled comparisons and input views. Within each comparison, models share the task-specific episodes, patient partitions, permitted channels, and missingness representation. The MIMIC-III representation input is the five-minute suffix $[T - 5\,\mathrm{min}, T)$ of the extracted six-hour history; VitalDB uses five-minute inputs ending at its legal cutoff. The selected tokenizer converts raw waveform samples into the token sequence processed by the encoder.

Part (a): architectures. Nine waveform encoders are compared at a matched budget of approximately 500k trainable parameters ($499{,}915 - -500{,}091$): Transformer, BiLSTM, Mamba-3, InceptionTime, TCN, ResNet1D, S4, FITS and DLinear. MIMIC-III and VitalDB compare Conv1D, Crossformer, and SeFT under their dataset-specific aggregation rules. MC-MED uses Conv1D across its nine displayed encoders. **Part (b): tokenizers.** The tokenizer study fixes a Transformer backbone and varies the input tokenizer. Conv1D projects local waveform patches into tokens. Macro means and SDs are recomputed from the selected per-seed task scores. SeFT embeds each observed (time, channel, value) triple as a set element. Crossformer uses dimension-segment-wise embedding with cross-channel attention, and STraTS embeds observation triplets with continuous value embeddings. iTransformer treats each channel as one token; FFT tokenizes each patch's frequency spectrum; mTAND applies multi-time attention over irregular sampling. MC-MED additionally reports a Conv2D tokenizer.

MIMIC-III. Eight ICU tasks are derived from the MIMIC-III clinical records and paired with the matched-subset bedside waveforms. Each task extracts a six-hour history ending at the anchor T: acute deterioration (major rescue event within $[T, T{+}6\,\mathrm{h})$), dynamic in-hospital mortality (six-hour gap, then death within the following 24 h), respiratory escalation (new HFNC/NIV/IMV or an upward respiratory-support transition within 6 h), circulatory deterioration (new vasopressor start within 6 h; primary head `y_new_pressor`), fluid response (hemodynamic improvement 30–60 min after a documented rapid crystalloid course; event-anchored), AKI KDIGO stage 2/3 (incident injury within $[T{+}1\,\mathrm{h}, T{+}24\,\mathrm{h})$), ventilator liberation (readiness and durable liberation over 24 h at daily ventilator checkpoints) and Sepsis-3 (infection-associated organ dysfunction within $[T{+}1\,\mathrm{h}, T{+}6\,\mathrm{h})$). The representation baseline uses raw samples in $[T - 5\,\mathrm{min}, T)$ on a 62-channel canonical layout at 125 Hz, giving 37,500 samples per channel before tokenization. Longer observation views use the corresponding suffix $[T - W, T)$ of the same extracted history. Splits are patient-level with a single held-out test set (854–81,359 test windows per task). Models are trained for 10 epochs with batch size 128, bf16 mixed precision, dropout 0.1 and weight decay 10^{-4}. The learning-rate grid is $\{10^{-3}, 10^{-4}, 10^{-5}\}$ in part (a) and $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 5{\times}10^{-6}\}$ in part (b) (the Conv1D baseline searched $\{10^{-2}, 10^{-3}, 10^{-4}\}$). A learning-rate group is eligible only if all five seeds finished without non-finite gradient steps.

VitalDB. The VitalDB Core-8 study uses intra-operative waveforms. All 12 available native waveform channels are kept in a canonical 66-slot geometry, and robust normalisation is fitted on the training folds only. Seven forecast tasks (OX01-IOH, CL07-002, H5-002, H5-030, L2-30, VA1-CO2-DROP, VA1-STIM-RESPONSE) are each evaluated at forecast gaps of 5 and 15 min; the two gap heads are averaged with equal weight into one value per task. The concurrent-state task VCS05 (one head, EEG slots excluded) is reported separately and is not part of the forecast macro. Training uses five seeds $\times$ five outer whole-subject folds (4,596 validation subjects). In part (a), the selected tokenizer is Conv1D for all nine encoders. The average test AUROC is got from the five fitted-fold models within each seed on the fixed cohort of 750 test subjects, and the mean and sample SD are reported across five seeds. The learning rate is selected per architecture/tokenizer condition from $\{10^{-4}, 3{\times}10^{-4}, 10^{-3}\}$. In part (b), Conv1D, SeFT, and Crossformer use the Transformer conditions of the architecture grid; the reported seven-tokenizer comparison also includes STraTS, iTransformer, FFT, and mTAND (Figure 3 and Table 10).

MC-MED. MC-MED uses the four selected tasks in Appendix B.5: physiological decompensation, pending hyperkalemia, lactate nonclearance, and pending severe anemia. It compares nine architectures at the 500k-parameter budget, eight displayed Transformer tokenizers, five seeds, and an equal-task macro over the four endpoints. Table 10 records the updated per-method means and seed SDs. Architecture and tokenizer panels select configurations separately, so their Transformer and Conv1D entries can differ. The rounded interaction annotations from Figure 3(c) are recorded in Table 11.

B.8 Method Attribution and Benchmark Adaptations

The citations at first mention in Sections 4 and 6 identify the methodological sources of the displayed configurations. The benchmark adapts these methods to its waveform inputs, task heads, and parameter budgets. A shared method name does not imply use of the original data, pretraining, ensembling, or optimization schedules.

Encoders and tokenizers. The nine encoder families retain their cited architectural basis while using the benchmark's classification heads and matched parameter target. The Conv1D projection is a local shared baseline inspired by PatchTST's patching design; this attribution does not assign the origin of generic Conv1D tokenization to PatchTST or imply reproduction of its channel-independent architecture. SeFT, Crossformer, STraTS, iTransformer, and mTAND supply representation mechanisms used with the common Transformer, so their tokenizer results do not evaluate the complete original systems. The FFT tokenizer applies a Fourier transform and learned spectral projection. FreTS (Yi et al., 2023) provides a reference for neural learning after frequency-domain conversion; our local tokenizer feeds the shared encoder and does not reproduce the complete FreTS model. FITS is a separate encoder family.

Missing-input handling. Zero-mask plus Δ supplies absence and elapsed-time features without reconstructing channels, using the missingness representation formalized in GRU-D (Che et al., 2018).

GRIN and Raindrop motivate graph-based sharing across observed channels, while the positional-kriging variant applies the IGNNK idea to a fixed sensor layout (Wu et al., 2021). Lead-agnostic SSL uses the random-lead-masking principle of Oh et al. (2022). Rectified Flow, TrigFlow, and Drifting identify generative objectives adapted for conditional missing-input handling; the cited generative papers do not themselves establish clinical waveform-imputation performance.

Training-subset policies. Diversity uses greedy coverage in a patient-representation space; diversity clustering combines manifold projection with proportional coverage within clusters (Sener & Savarese, 2018; McInnes et al., 2018). Gradient matching adapts coreset and matching-pursuit ideas to patient-level proxy gradients (Mirzasoleiman et al., 2020; Killamsetty et al., 2021). Training-dynamics selection combines confidence and variability with an early error score (Swayamdipta et al., 2020; Paul et al., 2021). The NLDR policy draws inspiration from label-free arrhythmia structure in nonlinear embeddings (Vazifeh & Fleischer, 2026); allocation across rare patient clusters is a benchmark adaptation, not a subset-selection algorithm evaluated in that paper. Noise filtering and task anchors are local controls that prioritize recorded signal quality and physiological transitions/hard negatives, respectively. The current MIMIC-III comparison applies training-window quotas, while VitalDB and MC-MED use subject quotas (Appendix E.3).

Channel-selection policies. MI and Lasso use published feature-selection principles (Peng et al., 2005; Tibshirani, 1996); temporal MI is a benchmark variation of the information-based score. SQI applies quality-based channel ranking, motivated by work on ECG signal acceptability (Clifford et al., 2012), without claiming to reproduce that paper's clinical-quality classifier across every waveform type. Random selection and clinical-priority channel lists are benchmark controls.

C Detailed Representation Results

C.1 Task-level and Macro Test AUROC

Table 10 reports the detailed MIMIC-III, VitalDB, and MC-MED architecture and tokenizer results summarized in Figure 3(a,b). The MIMIC-III macro averages eight tasks; VitalDB first weights the two forecast gaps equally within each task and then averages seven forecasting tasks, with VCS05 reported separately. MC-MED averages its four endpoints with equal weight, retaining the supplied macro means and seed SDs. Selection and scoring rules are specified in Appendix B.7.

Tokenizer comparisons. The tokenizers cover local waveform projections, observation sets or triplets, channel-wise segments, channel tokens, spectral tokens, and multi-time attention; their definitions and search grids appear in Appendix B.7. Conv1D is the only tokenizer in the macro-AUROC top three on all datasets: it rounds to the same MIMIC-III macro as Crossformer (0.611) and leads VitalDB (0.631) and MC-MED (0.574). SeFT's MIMIC-III lead (0.615) depends on its Sepsis-3 gain of 0.038 over Conv1D, computed before rounding; the other seven task differences sum to -0.010, and their means favor Conv1D (0.6085 versus 0.6070). SeFT accommodates irregular observations through timestamped-set aggregation (Horn et al., 2020), but lowers fluid-response AUROC from 0.586 to 0.574. On VitalDB, STraTS and FFT approach Conv1D at 0.627 and 0.625, whereas Crossformer falls to 0.583. On MC-MED, Conv1D and Crossformer lead the macro (0.574 and 0.570), whereas FFT falls to 0.509; Conv1D and Conv2D are nearly tied for decompensation (0.647) and hyperkalemia (0.576), but Crossformer leads anemia (0.582). Conv1D is a strong shared starting tokenizer, with SeFT retained for MIMIC-III sepsis and Crossformer for MC-MED anemia.

Architecture comparisons. Table 10(a) compares selected encoder–tokenizer configurations at the matched parameter budget under the selection rules in Appendix B.7. Transformer has the highest macro AUROC on MIMIC-III (0.607) and VitalDB (0.631); on MC-MED, ResNet1D (0.568) is closely followed by DLinear, Transformer, and InceptionTime (each 0.567 at the displayed precision). BiLSTM improves MIMIC-III mortality over Transformer (0.558 versus 0.532) but scores lower on AKI (0.696 versus 0.727). VitalDB also has endpoint-specific alternatives: DLinear exceeds Transformer on hypothermia (0.608 versus 0.605), and BiLSTM leads abrupt $ETCO_2$ fall (0.683 versus 0.675). DLinear leads MC-MED anemia (0.574), extending evidence that linear temporal projections can compete with attention models (Zeng et al., 2023) to this endpoint; BiLSTM instead leads decompensation (0.649) and hyperkalemia (0.580). Transformer leads lactate non-clearance (0.508 versus ResNet1D's 0.499), so ResNet1D's small macro lead comes from its gains on hyperkalemia and anemia. Lactate AUROCs remain near chance across architectures (0.453–

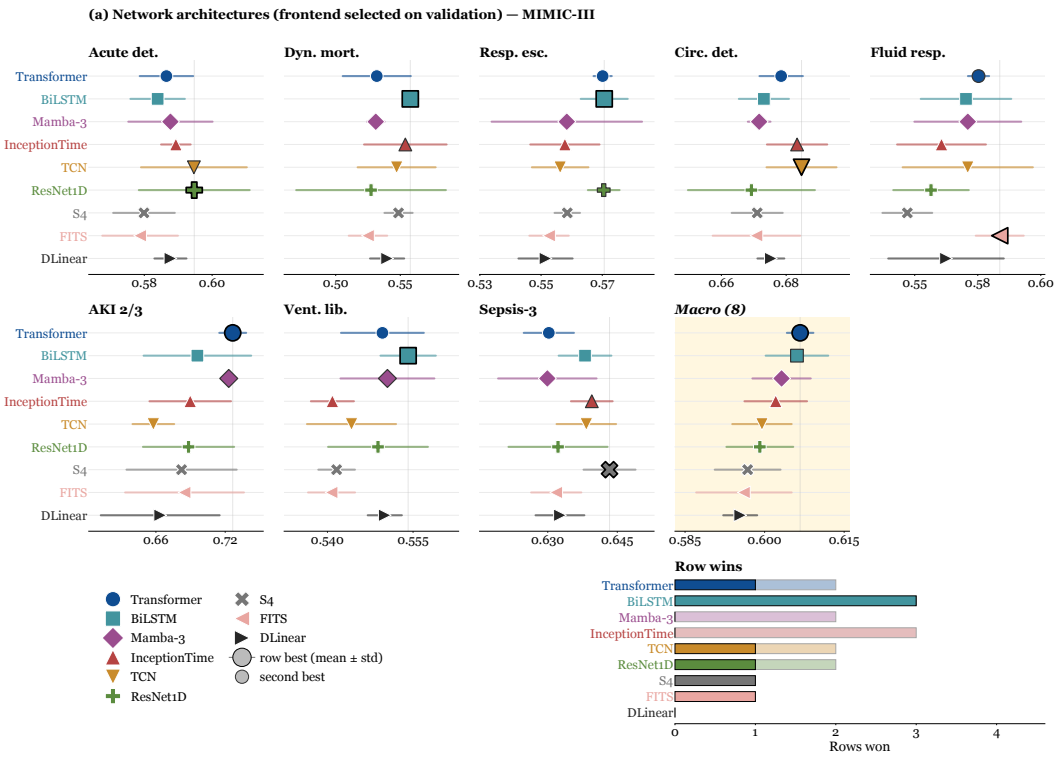


Figure 7: **Architecture AUROC by task — MIMIC-III.** Nine approximately 500k-parameter encoders with selected tokenizers across eight ICU tasks and their macro. Marker and SD-bar definitions follow Appendix C.2. Numerical values are in Table 10.

0.508), consistent with a five-minute waveform capturing only part of the production, clearance, and treatment history that shapes lactate kinetics over hours (Vincent et al., 2016). Transformer is a useful initial ICU and operating-room encoder; MC-MED favors a task-specific shortlist including BiLSTM, DLinear, Transformer, and ResNet1D.

C.2 Task-level AUROC Dot Plots

Figures 7, 8, and 9 show architecture results; Figures 10, 11, and 12 show Transformer tokenizer results. They expand Table 10 into individual task panels: markers are five-seed means, horizontal bars are ± 1 seed SD, and enlarged outlined markers identify the highest displayed means. Axis ranges differ by task; shaded panels are macros, and row-win summaries count first and second places over non-macro tasks, including VitalDB's concurrent probe.

C.3 Tokenizer–Encoder Test Contrasts

Table 11 reports the AUROC differences shown in the rounded annotations of Figure 3(c) for all three datasets. For each encoder and task, we subtract the Conv1D test AUROC from the indicated tokenizer's test AUROC. We report these differences on the 0–1 AUROC scale, dividing the figure annotations by 100. Each mean averages the nine encoder-specific contrasts. Agreement counts how many contrasts have the same sign as the unrounded mean. MIMIC-III macros average eight tasks; VitalDB forecast macros average seven tasks after equal weighting of the two gaps, with concurrent VCS05 reported separately; MC-MED macros average four tasks. MC-MED compares Conv2D and FFT with Conv1D; changes are rounded to three decimals, and agreement follows the sign of the unrounded mean. Figure 13 shows the absolute MIMIC-III scores behind the crossed-grid comparison, making encoder-rank changes visible within each task. This crossed grid is separate from the fixed-Transformer tokenizer search in Table 10(b), whose reported cell means differ.

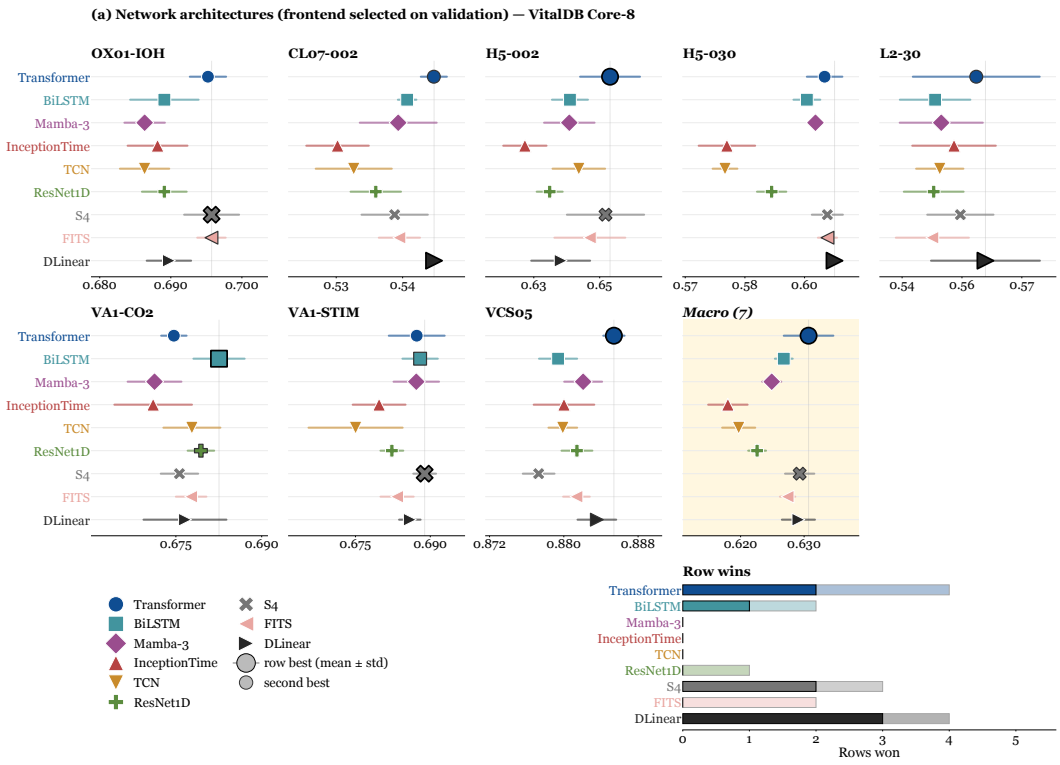


Figure 8: **Architecture AUROC by task — VitalDB Core-8.** Updated test results for nine encoders with selected tokenizers. Each forecast task gives equal weight to its two gaps; the seven-task macro excludes the concurrent VCS05 probe. Scores average fitted-fold model metrics within each seed, without probability ensembling. Numerical values are in Table 10.

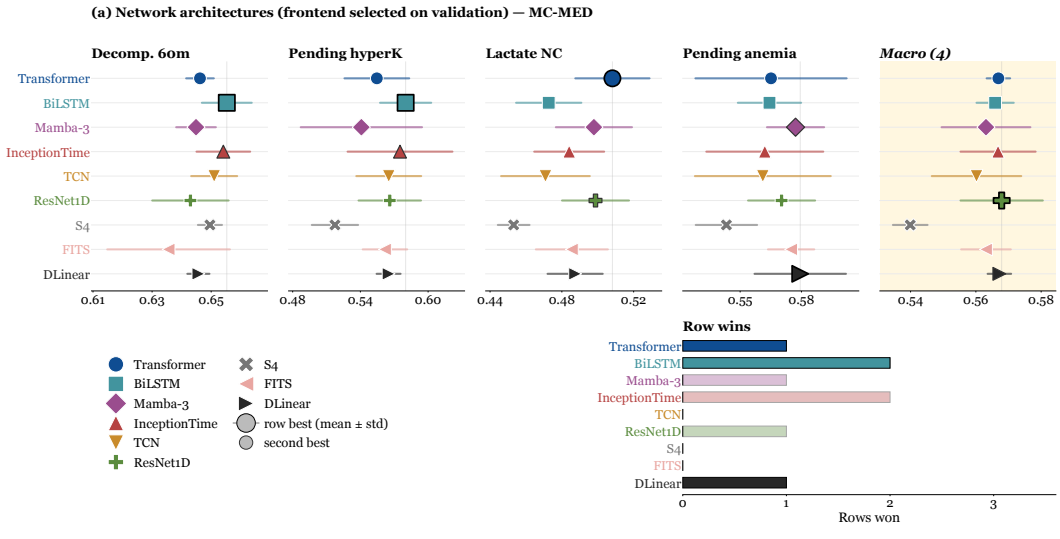


Figure 9: **Architecture AUROC by task — MC-MED.** Nine encoders with selected tokenizers on four ED endpoints and their macro, using the input and search grids in Appendix B.7. Numerical values are in Table 10.

MIMIC-III interaction patterns. The mean macro changes relative to Conv1D are -0.003 AUROC for Crossformer and -0.001 for SeFT; eight and five encoders, respectively, share the mean's negative sign. For fluid response, Crossformer improves AUROC by 0.014 on average, with eight

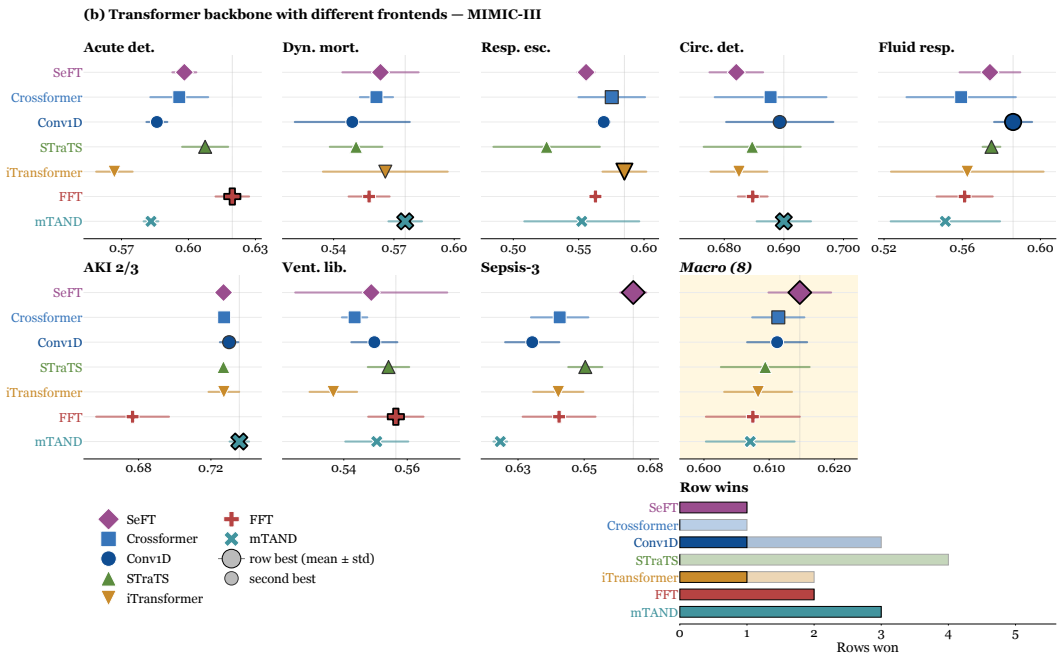


Figure 10: **Transformer tokenizer AUROC by task — MIMIC-III.** Seven tokenizers across eight ICU tasks; SeFT has the highest macro AUROC. Numerical values are in Table 10.

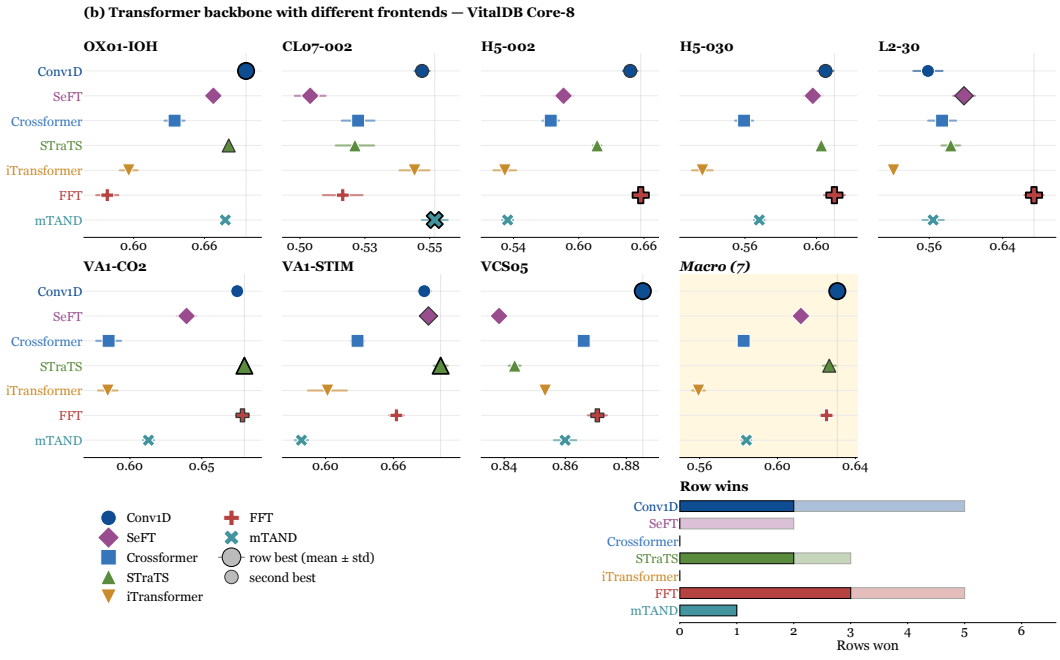


Figure 11: **Transformer tokenizer AUROC by task — VitalDB Core-8.** Seven tokenizers with the gap weighting and scoring of Figure 8. FFT's gas-discordance gain and hypotension loss largely offset in the macro; concurrent VCS05 is separate. Numerical values are in Table 10.

of nine encoders improving, while SeFT improves it by 0.001, with five encoders improving. For Sepsis-3, Crossformer's mean change is -0.003 AUROC, whereas SeFT's is $+0.014$ AUROC, with eight of nine encoders improving under SeFT. The fixed-Transformer FFT comparison gives another reversal: acute-deterioration AUROC rises from 0.586 to 0.620, while AKI falls from 0.730 to

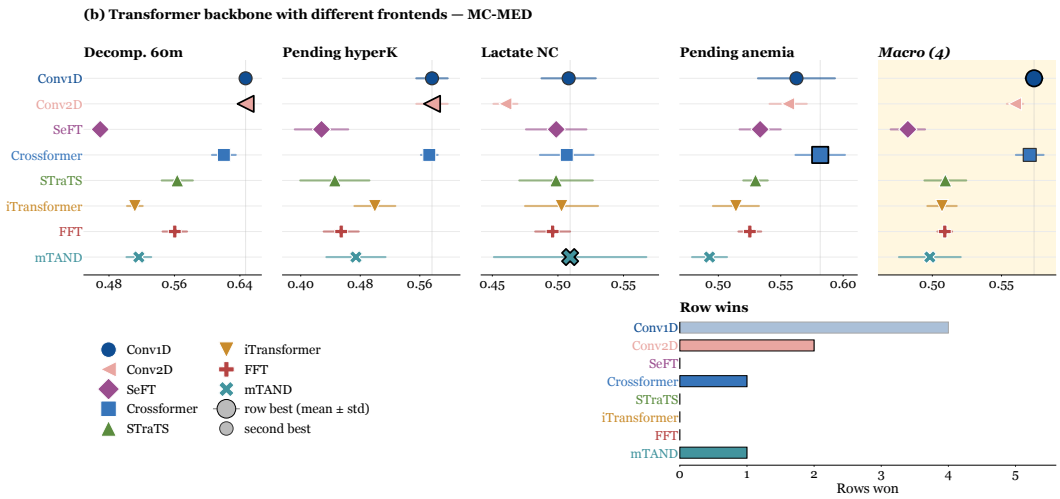


Figure 12: **Transformer tokenizer AUROC by task — MC-MED.** Eight tokenizers, reported as five-seed means and SDs. Conv1D leads the macro; Conv1D and Conv2D are nearly tied on decompensation and hyperkalemia, while Crossformer leads anemia. Numerical values are in Table 10.

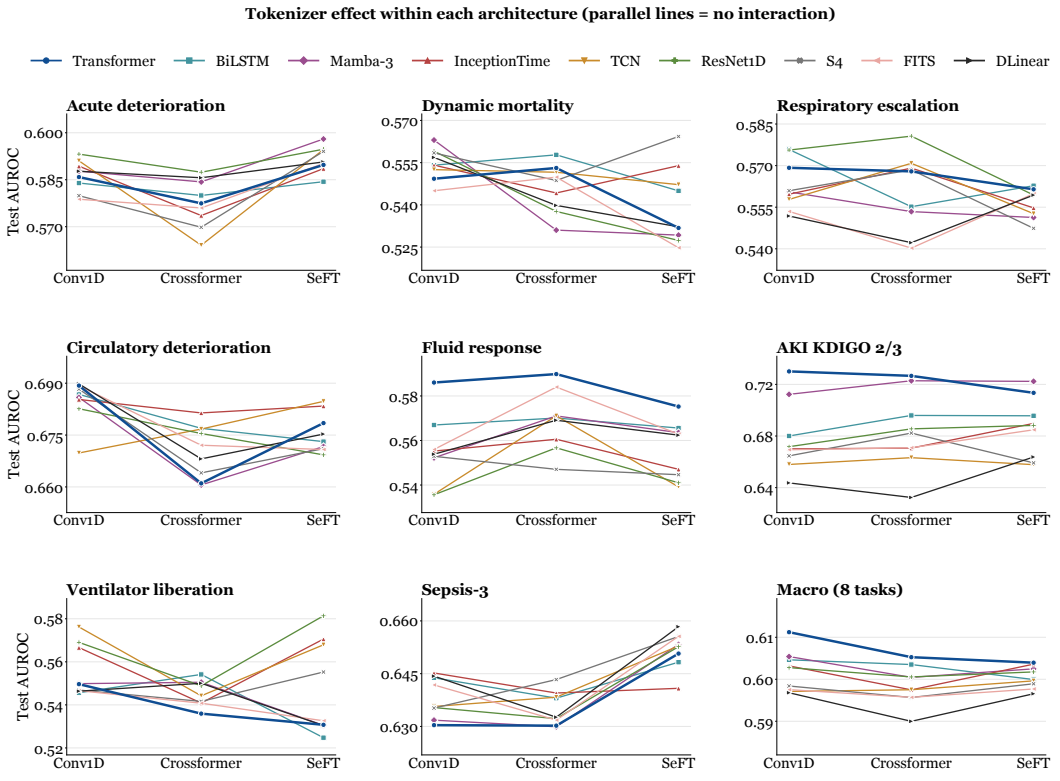


Figure 13: **Task-level tokenizer–architecture interactions on MIMIC-III.** Test AUROC for nine encoders with Conv1D, Crossformer, and SeFT across eight tasks and their macro; each line follows one encoder, with nonparallel trends indicating encoder-dependent tokenizer effects and crossings indicating changes in encoder ranking. Mean task-level contrasts are listed in Table 11.

0.677. The results favor a shortlist of Conv1D and SeFT for sepsis, and Conv1D and Crossformer for fluid response, where the alternative tokenizer's gain recurs across most encoders.

VitalDB interaction patterns. The mean forecast-macro changes are -0.015 AUROC for Crossformer and -0.010 for SeFT, with negative changes in seven and nine encoders, respectively. Both tokenizers lower hypotension AUROC across all nine encoders (mean changes of -0.038 and -0.023), yet improve blood-gas/$ETCO_2$-discordance AUROC by 0.032 and 0.039, with all nine encoders improving under each tokenizer. For coupled hemodynamic response (VA1-STIM), SeFT improves AUROC by 0.008 in all nine encoders, whereas Crossformer has a mean change of -0.009 with only four encoders sharing the negative sign. With a fixed Transformer, FFT similarly improves gas-discordance AUROC (0.674 versus 0.559) but lowers hypotension AUROC (0.578 versus 0.695). Conv1D's repeated advantage for hypotension is consistent with the predictive value of local arterial-waveform changes reported by Hatib et al. (2018), whereas the gas-discordance target estimates an arterial-to-expired gas relation with different ventilation–perfusion determinants (Nassar & Schmidt, 2016). The concurrent VCS05 probe also favors Conv1D over Crossformer and SeFT across all nine encoders, with mean losses of 0.006 and 0.044, respectively. Conv1D is better suited to the pressure and concurrent-state endpoints in this comparison; FFT, SeFT, and Crossformer are stronger candidates for gas discordance.

MC-MED interaction patterns. FFT lowers mean AUROC for decompensation by 0.108, hyperkalemia by 0.076, and anemia by 0.041, with losses in all nine encoders for each endpoint. Its lactate gain of 0.010 occurs in six encoders and does not offset these losses: macro AUROC falls by 0.054 in the nine-encoder average, with all nine declining. Conv2D changes decompensation and hyperkalemia means by approximately 0.000 and 0.001, but lowers lactate nonclearance by 0.020 and anemia by 0.012, giving a macro change of -0.008 with eight encoders declining. The FFT losses on hyperkalemia distinguish the clinical information in ECGs (Galloway et al., 2019) from the representation that best exposes it to the tested models: frequency projection is less effective than local patches for this endpoint. Conv1D is the stronger shared tokenizer for this ED battery; Conv2D remains a close task-specific comparator for decompensation and hyperkalemia.

Comparison scope. The alternative tokenizers differ across datasets: Crossformer and SeFT in the MIMIC-III and VitalDB interaction panels, and Conv2D and FFT on MC-MED. The updated VitalDB representation table averages fitted-fold model metrics within each seed; the validation and probability-ensemble studies in Appendix F.1 use separate estimators.

C.4 Efficacy, Efficiency, and Safety of Representation Choices

Figures 14–16 compare architectures, and Figures 17–19 compare Transformer tokenizers. Appendix A.2 specifies metric definitions, model selection, and export differences.

Discrimination, calibration, and selective risk. On MIMIC-III, TCN has higher architecture macro AUPRC than Transformer (0.098 versus 0.093) and lower ECE (0.138 versus 0.206), but higher AURC (0.319 versus 0.308). On VitalDB, Transformer leads architecture AUPRC at 0.078, while BiLSTM has the lowest ECE at 0.036 versus Transformer's 0.038. In the earlier MC-MED companion campaign, ResNet1D has the highest AUPRC (0.169), S4 the lowest ECE (0.245), and InceptionTime the lowest AURC (0.326). The different AUPRC and ECE leaders agree with the distinction between discrimination and probability calibration emphasized by Van Calster et al. (2019). For AUPRC, TCN on MIMIC-III, Transformer on VitalDB, and ResNet1D on MC-MED are strong initial baselines, with the ECE and AURC leaders retained when calibrated risk or selective prediction is the primary objective.

Spectral tokenization across endpoints and metrics. For MIMIC-III Transformer tokenizers, FFT improves AUPRC from 0.098 to 0.100, ECE from 0.178 to 0.128, and AURC from 0.330 to 0.271 relative to Conv1D. In that earlier MC-MED campaign, FFT improves AUPRC (0.166 versus 0.148), but increases ECE (0.282 versus 0.255) and AURC (0.450 versus 0.326); these are separate configurations from the updated AUROC comparison in Table 10. VitalDB's gas-discordance AUPRC rises from 0.101 with Conv1D to 0.138 with FFT, alongside the AUROC increase from 0.559 to 0.674. On hypotension, however, AUROC falls from 0.695 to 0.578, leaving FFT below Conv1D in both macro AUROC and AUPRC (0.076 versus 0.078), despite its lower macro ECE (0.035 versus 0.038).

Tokenizer cost and choice. MIMIC-III tokenizer training time ranges from 108.6 seconds for Conv1D to 1,084 seconds for mTAND, while peak memory increases from 3.395% to 17.2% of device capacity and AUPRC decreases from 0.098 to 0.095. FFT's AUPRC, ECE, and AURC im-

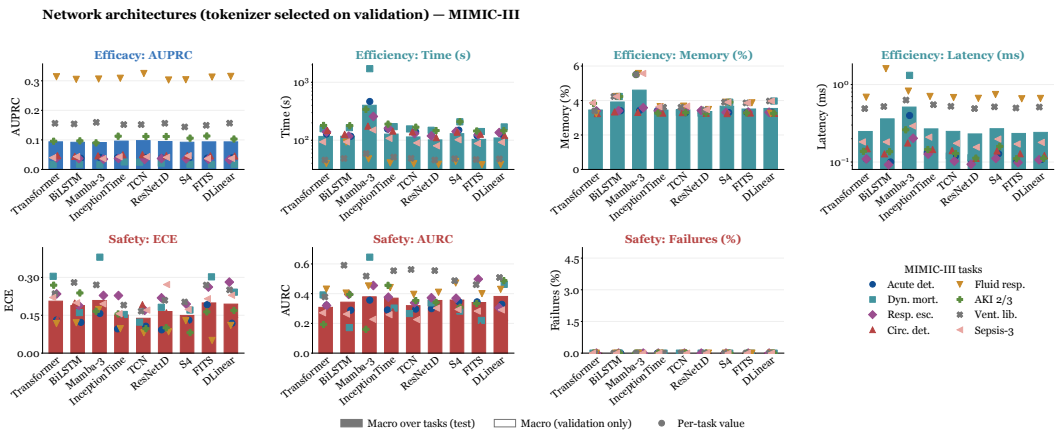


Figure 14: **Network architectures — MIMIC-III.** Approximately 500k-parameter encoders with selected tokenizers. Panels show AUPRC, training time, memory, time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 80 and 81.

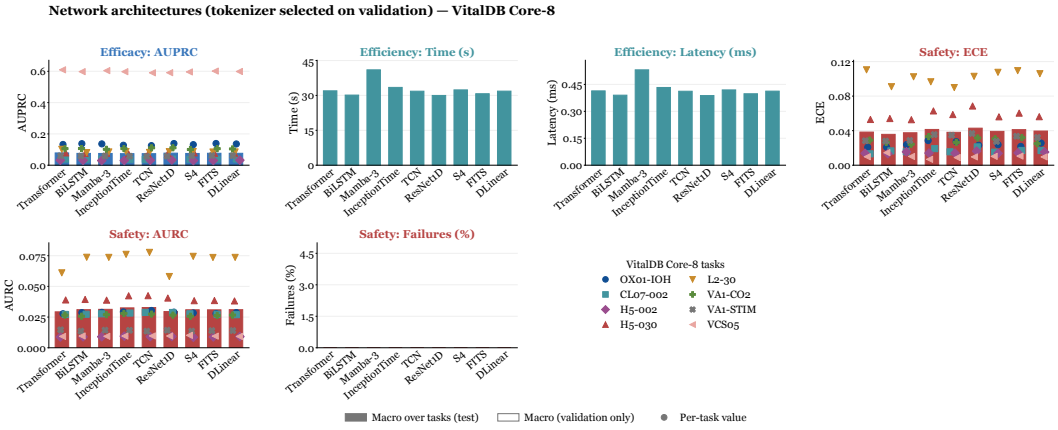


Figure 15: **Network architectures — VitalDB Core-8.** Approximately 500k-parameter encoders with selected tokenizers. Forecast macros exclude concurrent VCS05. Cost panels use earlier-grid telemetry per validation cell. Panels show AUPRC, training time, time per training sample, ECE, AURC, and audited prediction-file failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 82 and 83.

provements in this comparison require 142.6 seconds and 3.404% memory, substantially less than mTAND. VitalDB tokenizer cost records cover Conv1D, SeFT, and Crossformer. Conv1D and FFT offer a useful cost comparison on MIMIC-III; the updated VitalDB and MC-MED AUROC comparisons favor Conv1D for macro discrimination.

D Temporal Comparison Details

Comparison grids and evaluation status. Figure 4 compares two timing choices: how much waveform history is available before prediction, and how long the interval is between the input cutoff and the target window. Table 8 records the displayed observation lengths and forecast-gap axis values. All six panels in Figure 4 are labeled as test AUROC. The VitalDB warning curve is distinct from the validation curve and the fixed-30-minute Transformer/Patch test results in Appendices F.2 and F.2. The scores, uncertainty intervals, and probability-quality metrics reported there apply only to those studies. The MIMIC-III warning panel contains only dynamic mortality, AKI, and Sepsis-3; MC-MED includes all four endpoints with task-specific gaps. The refreshed MIMIC-III duration com-

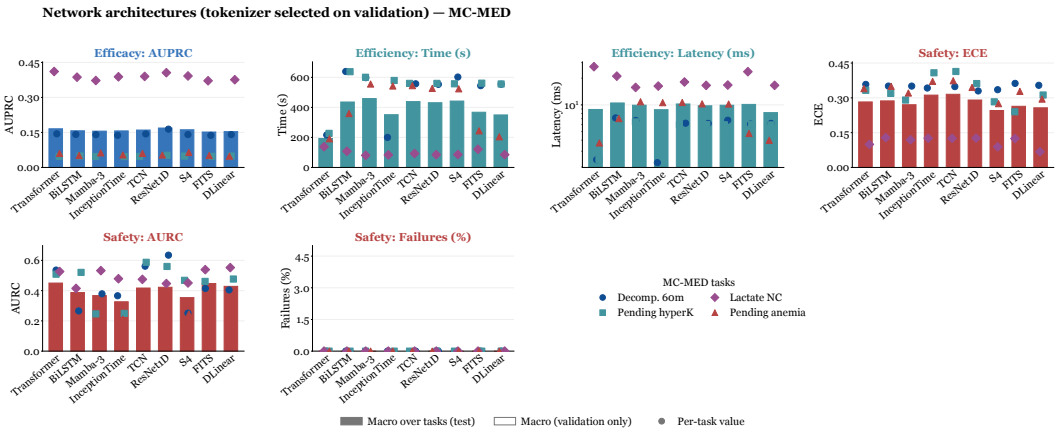


Figure 16: **Network architectures — MC-MED, earlier companion campaign.** Approximately 500k-parameter encoders with selected tokenizers, including FFT candidates. Panels show AUPRC, training time, time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 84 and 85.

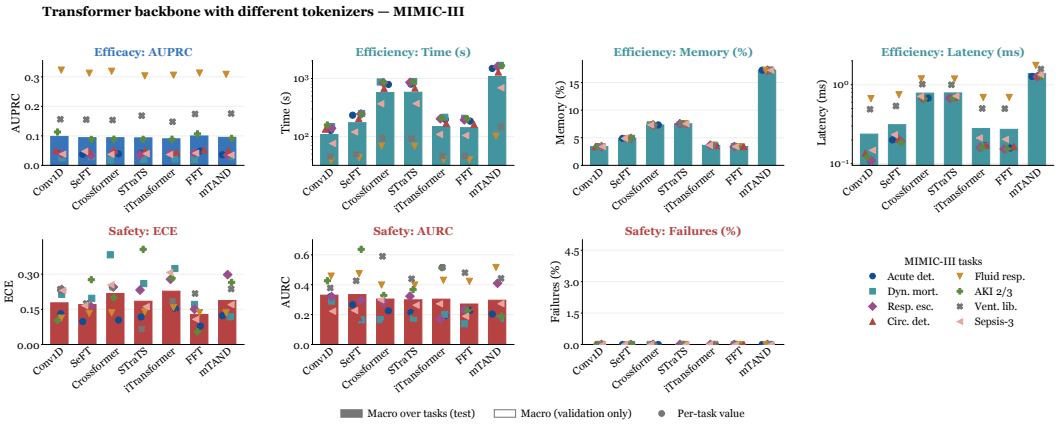


Figure 17: **Transformer tokenizers — MIMIC-III.** Seven tokenizers with a fixed Transformer backbone. Panels show AUPRC, training time, memory, time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 86 and 87.

parison contains 10-, 20-, and 60-minute histories; the earlier five-minute result belongs to a separate campaign. The duration curves average three encoder–tokenizer configurations on MIMIC-III and five on VitalDB; MC-MED uses ResNet1D with Conv2D. The VitalDB gap study uses Transformer with patch embedding.

Observation length. Black lines in panel (a) summarize eight MIMIC-III tasks, seven VitalDB forecast tasks, and four MC-MED tasks. VitalDB weights its two core gaps equally within each forecast task; concurrent VCS05 is excluded from the forecast macro. Numerical observations use the task-level exports underlying the updated figure. MIMIC-III macro AUROC is 0.606 at both ten and 20 minutes and 0.609 at 60 minutes; respiratory escalation improves from 0.575 to 0.593 and acute deterioration from 0.646 to 0.656 across ten to 60 minutes. Mortality stays near 0.517, while Sepsis-3 decreases from 0.659 to 0.652; the modest macro change therefore combines opposing endpoint trends. VitalDB macro AUROC increases from 0.627 at five minutes to 0.636 at 60 minutes, with hypotension peaking at 30 minutes (0.729) and gas discordance at 60 minutes (0.594). The hypotension gain aligns with prior evidence of predictive pre-event arterial-waveform patterns (Hatib et al., 2018), whereas oxygen desaturation decreases from 0.635 to 0.625 and hypothermia stays near 0.602. MC-MED macro AUROC peaks at ten minutes (0.578 versus 0.567 at both five and 15

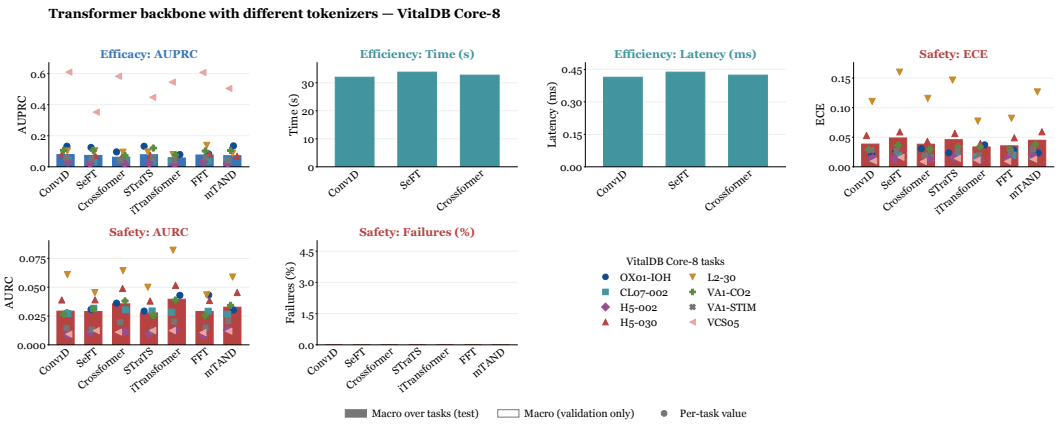


Figure 18: **Transformer tokenizers — VitalDB Core-8.** Seven tokenizers with a fixed Transformer backbone. Forecast macros exclude concurrent VCS05. Cost panels use earlier-grid telemetry per validation cell. Panels show AUPRC, training time, time per training sample, ECE, AURC, and audited prediction-file failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 88 and 89.

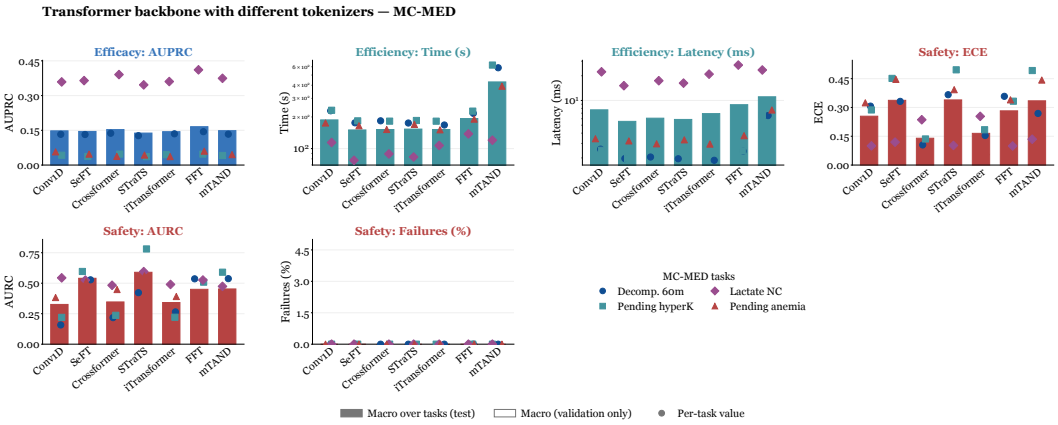


Figure 19: **Transformer tokenizers — MC-MED, earlier companion campaign.** Seven tokenizers with a fixed Transformer backbone. Panels show AUPRC, training time, time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 90 and 91.

minutes). From ten to 15 minutes, decompensation falls from 0.656 to 0.631, hyperkalemia from 0.576 to 0.528, and anemia from 0.577 to 0.530; lactate nonclearance instead rises from 0.502 to 0.577. This lactate-specific gain is consistent with a serial-measurement endpoint shaped by evolving metabolism and treatment (Vincent et al., 2016). For AUROC, ten-minute histories are useful starting points for the MIMIC-III battery and the three non-lactate MC-MED tasks, while longer windows suit respiratory escalation, intraoperative hypotension and gas discordance, and MC-MED lactate nonclearance in these sweeps.

Forecast gap. Panel (b) shifts the target window beyond the same input cutoff, changing both the warning requirement and which future events qualify. For MIMIC-III, mortality decreases from 0.545 at three hours to 0.539 at twelve hours and 0.537 at 24 hours; AKI instead peaks at six hours (0.697 versus 0.686 at three hours and 0.685 at twelve hours). Sepsis-3 returns to 0.629 at six hours after falling to 0.619 at four hours. AKI's intermediate peak is consistent with creatinine and urine-output criteria being assessed over time (KDIGO Acute Kidney Injury Work Group, 2012), linking its warning curve to outcome ascertainment as well as physiology. For VitalDB, macro AUROC falls from 0.662 at five minutes to 0.621 at 15 minutes, then ranges from 0.619 to 0.632. Coupled response rebounds from 0.603 at 45 minutes to 0.750 at 75 minutes, while gas discordance declines

from 0.622 at 30 minutes to 0.555 at 75 minutes. Hypotension falls from 0.791 at five minutes to 0.611 at 60 minutes before recovering to 0.650 at 75 minutes; its strongest discrimination at short gaps aligns with prior arterial-waveform forecasting results (Hatib et al., 2018). In our clinical interpretation, the 60-minute hypothermia peak reflects evolving redistribution and warming (Sessler, 2016), and the coupled-response rebound reflects a later operative phase with changing surgical stimulation and anesthetic context (Ledowski, 2019). For MC-MED, decompensation declines from 0.647 at zero gap to 0.629 at ten minutes and 0.626 at 30 minutes. Hyperkalemia is almost unchanged from zero to 15 minutes (0.566 and 0.567), while anemia improves from 0.556 to 0.566; both decline at 60 minutes (0.532 and 0.544). Lactate nonclearance falls from 0.499 at 30 minutes to 0.489 at 60 minutes, then rises to 0.532 at 90 minutes. The laboratory endpoints therefore have different preferred intervals, with the lactate rebound remaining close to chance. These curves use fixed patient partitions but shifted outcome windows and gap-specific eligibility, so they describe prediction of different future intervals within the same partitioning scheme. Model comparisons should fix the gap to the intended clinical intervention window and interpret later-gap gains alongside outcome timing and support.

Temporal controls and interpretation. The observation-length comparisons in Figure 4(a) use the same anchors and nested raw-waveform intervals $[T - W, T)$, holding the task, gap, horizon, channels, labels, and patient partitions fixed. For MIMIC-III, the extracted history is $[T - 6\,\mathrm{h}, T)$; the five-minute and one-hour conditions take $[T - 5\,\mathrm{min}, T)$ and $[T - 1\,\mathrm{h}, T)$, respectively. For warning-time comparisons, each patient retains the same training, validation, or test assignment across gaps, including the same fold assignment where folds are used. The observation view and target horizon remain fixed, while the outcome interval shifts to $[T + G, T + G + H)$. This fixes patient allocation across gaps; eligible anchors, outcome labels, and prevalence can still change with follow-up and task-specific exclusion rules. The separate VitalDB validation extension reports this gap-specific support and prevalence with subject-bootstrap intervals conditional on fitted models (Appendix F.2).

D.1 Efficacy, Efficiency, and Safety Across Temporal Settings

Figures 20–22 compare observation lengths, and Figures 23–25 compare forecast gaps. These figures use the September 17 temporal collection corresponding to the AUROC studies in Figure 4; observation histories and task-specific gap grids follow Appendix D. Predictive and confidence metrics report test results; VitalDB training costs retain earlier-grid telemetry, and MC-MED memory records sampled shared-device usage in MiB.

Additional history and measured cost. MIMIC-III macro AUPRC changes from 0.1192 at ten minutes to 0.1177 at 20 and 0.1227 at 60, while 20 minutes gives the lowest macro ECE (0.1538) and AURC (0.3190). Respiratory-escalation AUPRC rises from 0.0549 to 0.0725 across ten to 60 minutes, but its ECE increases from 0.2264 to 0.2787 and AURC from 0.3363 to 0.4530: extra waveform history improves discrimination without improving the confidence estimates used to prioritize review. For the ICU battery, 20- and 60-minute views offer different starting points for confidence quality and precision–recall performance, respectively.

VitalDB macro AUPRC rises from 0.0762 at five minutes to 0.0810 at 30 and 0.0812 at 60; macro ECE decreases from 0.0411 to 0.0351 and AURC from 0.0300 to 0.0277 across five to 60 minutes. Hypotension AUPRC increases from 0.1321 to 0.1597, consistent with predictive information in pre-event arterial-pressure patterns (Hatib et al., 2018), whereas oxygen-desaturation AUPRC decreases from 0.0295 to 0.0237. Thirty- and 60-minute histories are useful hypotension candidates, while desaturation favors retaining the five-minute view.

MC-MED macro AUPRC increases from 0.1809 at five minutes to 0.1977 at ten and 0.2208 at 15, even though macro AUROC peaks at ten minutes. Decompensation and lactate drive the ten-to-15-minute AUPRC gain (0.2763 to 0.3166 and 0.3966 to 0.4689); hyperkalemia and anemia both decline. The lowest macro ECE occurs at ten minutes (0.0873), whereas five minutes has the lowest AURC (0.2385 versus 0.2661 at 15 minutes). At batch size 128, mean training time per optimizer step increases from 14.20 ms at five minutes to 18.80 ms at ten and 25.27 ms at 15, separating update cost from differences in the number of training windows and selected epochs. Sampled shared-device memory is nonmonotonic and does not measure a model's isolated memory requirement. Ten

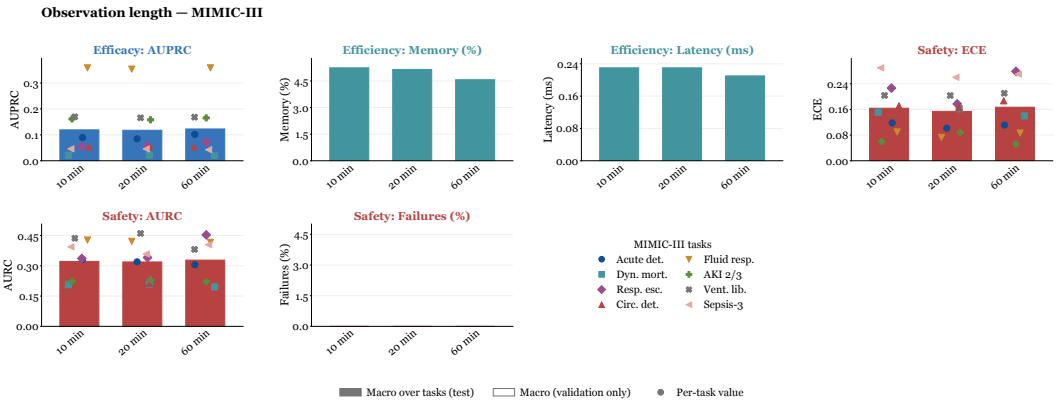


Figure 20: **Observation length — MIMIC-III.** Ten-, 20-, and 60-minute histories across eight tasks. Panels show AUPRC, peak accelerator memory (%), time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 92 and 93.

minutes is the stronger AUROC/calibration starting point, with 15 minutes as the AUPRC alternative for decompensation and lactate and five minutes retained for selective-risk comparisons.

Earlier warning and endpoint differences. MIMIC-III AKI AUPRC decreases from 0.1509 at three hours to 0.1404 at six and 0.1202 at twelve, despite AUROC peaking at six hours. Sepsis-3 AUPRC also falls from 0.0362 at two hours to 0.0301 at six, while ECE improves from 0.3108 to 0.2197. Thus, neither the AKI AUROC peak nor the improved sepsis calibration identifies the gap with the strongest precision–recall performance.

VitalDB gap macro AUPRC decreases from 0.100 at five minutes to 0.071 at 15 minutes, then varies nonmonotonically and reaches 0.082 at 75 minutes. Coupled hemodynamic response recovers from 0.047 at 45 minutes to 0.113 at 75 minutes, while oxygen-desaturation AUPRC ranges from 0.022 to 0.027 across 15–75 minutes. We interpret the coupled-response rebound as reflecting a later operative period entering the outcome window, since stimulation and anesthetic context affect hemodynamic responses (Ledowski, 2019). Across the six gaps, macro ECE ranges from 0.030 to 0.035 and AURC from 0.027 to 0.030. The 45-minute gap has the lowest ECE (0.030) but also the lowest AUPRC (0.068), consistent with calibration and discrimination measuring different properties (Van Calster et al., 2019). MC-MED decompensation AUPRC decreases from 0.2636 at zero gap to 0.1566 at 30 minutes, agreeing with its AUROC decline. Anemia instead gains AUPRC from 0.0470 at zero gap to 0.1999 at 60 minutes, while AUROC falls from 0.556 to 0.544 and ECE increases from 0.0348 to 0.0776. This later-window AUPRC gain accompanies weaker pairwise ranking and calibration; shifted outcome windows and prevalence must be considered when interpreting it. MIMIC-III and MC-MED cover different task subsets at different gaps, so their figures show task markers without macro bars. For waveform warning models, fix the intervention gap before comparing configurations, then assess precision–recall performance and confidence quality within that task and outcome window.

E Information Availability Comparison Details

References and evidence scope. Figure 5 reports AUROC differences relative to a separate reference for each experimental axis (Table 9). The top two rows represent MIMIC-III and VitalDB; the lower-right cells show MC-MED fusion and subject selection, and the two lower-left cells show transfer to MIMIC-IV. Colored symbols identify tasks within each dataset, and bars summarize aggregate differences; the transfer cells use the distinct five-target key in Appendix E.5. The task legends contain eight ICU tasks, seven VitalDB forecasting tasks plus the concurrent VCS05 probe, and four MC-MED tasks. The concurrent probe remains separate from the forecast aggregate under the documented VitalDB protocol. Comparisons in Section 6 describe the plotted means; uncertainty definitions follow the corresponding detailed studies below. The composite and task-level

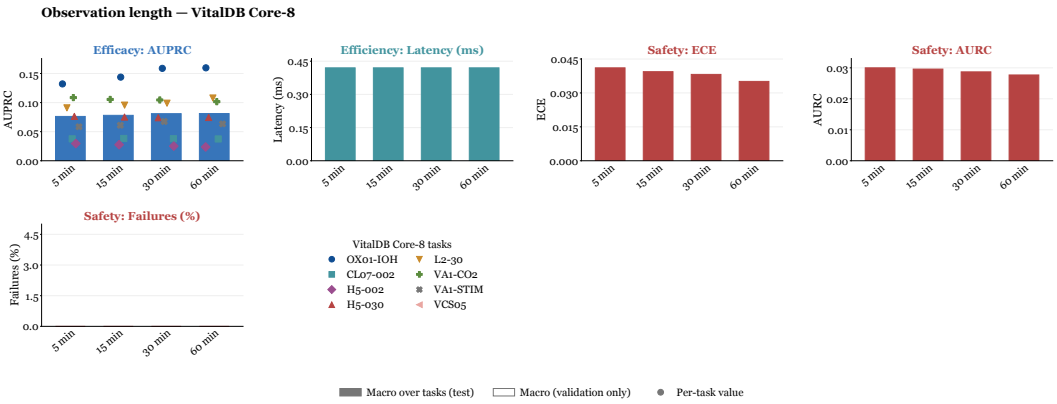


Figure 21: **Observation length — VitalDB Core-8.** Five-, 15-, 30-, and 60-minute histories; forecast macros exclude concurrent VCS05. Per-sample cost uses earlier-grid telemetry. Panels show AUPRC, time per training sample, ECE, AURC, and audited prediction-file failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 94 and 95.

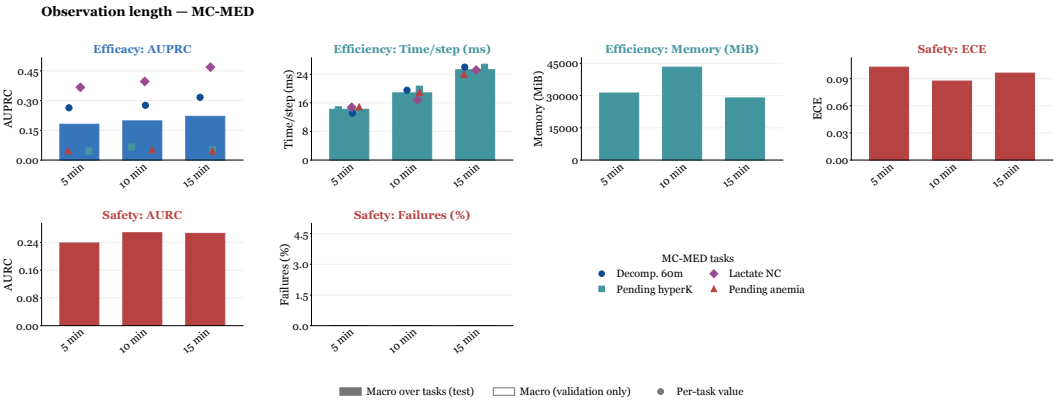


Figure 22: **Observation length — MC-MED.** Five-, ten-, and 15-minute histories across four ED tasks. Panels show AUPRC, training time per optimizer step (ms; batch size 128), sampled shared-device memory (MiB), ECE, AURC, and training-attempt failures. Step times divide each fit's training seconds by its recorded update count before averaging. Memory is a device-level measurement rather than an isolated model peak. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 96 and 97.

figures use the September 17 collection; MC-MED subject selection and fusion share the September 16 campaign with ECG-II and respiration inputs. MIMIC-III sensor, missingness, and subset results average Transformer/Patch, Transformer/Crossformer, and selective-SSM/Patch; VitalDB averages Transformer/Patch, Transformer/Crossformer, and Mamba-3/Patch. Fusion uses Transformer/Patch on MIMIC-III and MC-MED, and Mamba-3/Patch with 60-minute history on VitalDB. The revised VitalDB plotted means are separate from the earlier validation summaries and completed fixed-recipe tests in Appendices E.3 and F.3. Transfer uses held-out outer-role test predictions under the retrospective development protocol in Appendix E.5.

E.1 Sensor Access and Missing Inputs

Channel selection. Figure 5(a) compares channel-selection policies against each dataset's all-channel reference; Figures 26 and 27 give task-level AUROCs. MC-MED channel selection and channel imputation are omitted because its input configuration has only two available waveform channels. At ten channels, MIMIC-III clinical-priority and SQI selection reach macro AUROCs of 0.6062 and 0.6073 versus the all-channel 0.6056, while random selection falls to 0.5579. These poli-

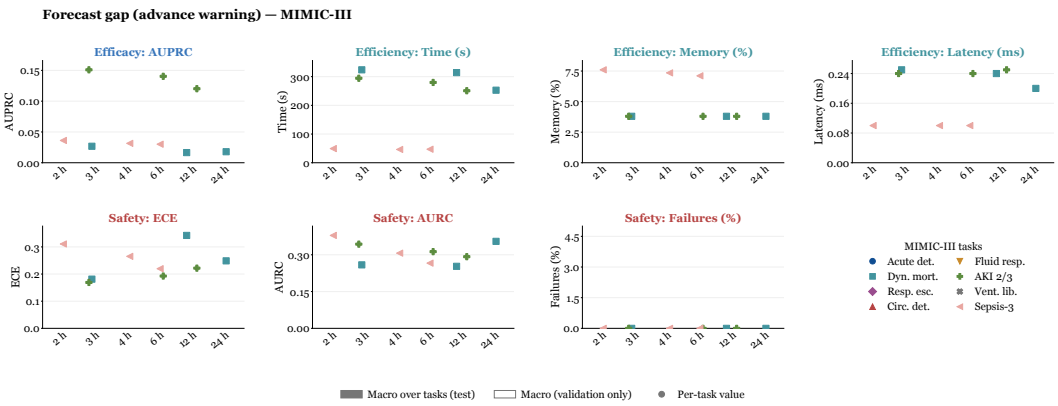


Figure 23: **Forecast gap — MIMIC-III.** Task-specific intervals for mortality (3/12/24 h), AKI (3/6/12 h), and Sepsis-3 (2/4/6 h). Task coverage varies by gap, so macro bars are omitted. Panels show AUPRC, training time, peak accelerator memory (%), time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 98 and 99.

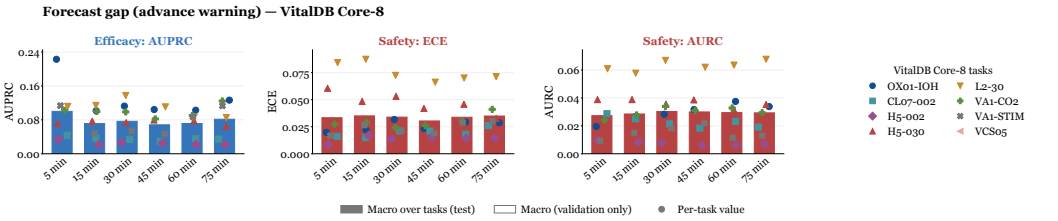


Figure 24: **Forecast gap — VitalDB Core-8.** Five- to 75-minute advance-warning intervals across seven forecasts, excluding concurrent VCS05. Panels show AUPRC, ECE, and AURC. Metric and mark definitions follow Appendix A.2. Numerical values are in Table 100.

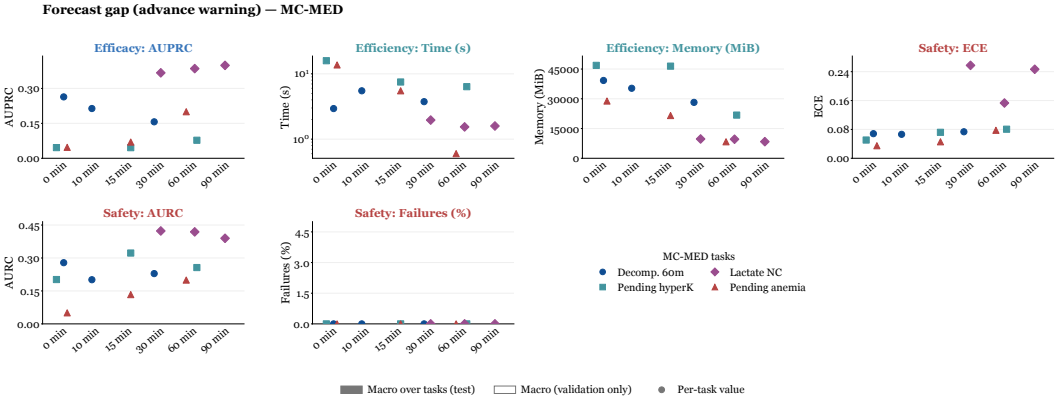


Figure 25: **Forecast gap — MC-MED.** Task-specific advance-warning intervals for four ED endpoints. Task coverage varies by gap, so macro bars are omitted. Panels show AUPRC, training time, sampled shared-device memory (MiB), ECE, AURC, and training-attempt failures. Memory measures device usage rather than an isolated model peak. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 101 and 102.

cies prioritize clinical relevance or signal reliability, whereas MI-based policies prioritize statistical association with the training label. Every VitalDB selector lowers macro AUROC, with temporal MI giving the smallest loss and random selection the largest, particularly for hypotension. At the displayed eight-channel budget, temporal MI improves gas-discordance AUROC from 0.5559 to

0.5676, the largest gain among selectors; SQI also improves it, to 0.5596. Clinical-priority and SQI are the stronger MIMIC-III choices under a sensor budget; VitalDB favors all available channels, with temporal MI as its closest reduced-channel comparator.

Missing-input handling. Panel (b) compares eight methods with zero-mask plus Δ on MIMIC-III and nine methods with architecture-only handling on VitalDB; Figures 28 and 29 give task-level AUROCs. Lead-agnostic SSL and GRU-D lead the MIMIC-III macro at 0.6086 and 0.6082 versus 0.6037 for the reference, but all eight alternatives improve acute deterioration and ventilator liberation while reducing AKI and fluid-response AUROC. GRU-D models observation and missingness histories (Che et al., 2018); its mortality gain of 0.0195 contrasts with a fluid-response loss of 0.0128, favoring different handling for prognosis and the short response to an intervention. On VitalDB, Raindrop approximately matches the architecture-only macro (both 0.6128); the other eight methods lower it by 0.0207–0.0314. Gas discordance improves with eight methods and accelerating MAP decline with all nine, while oxygen desaturation declines with every method. Raindrop is the exception for coupled hemodynamic response, improving AUROC by 0.0067 while the other eight methods lower it. The separate Transformer/Patch companion favors native handling in macro AUPRC (0.078) over GRIN (0.076) and lead-agnostic SSL (0.077). In the separate VitalDB ensemble test, native handling loses AUROC from 0.619 to 0.583 under 50% channel dropout, identifying additional sensor loss as a failure mode despite its favorable natural-missingness ranking (Appendix F.1). Lead-agnostic SSL and GRU-D are useful MIMIC-III macro candidates, retaining zero-mask plus Δ for AKI and fluid response; VitalDB favors native handling or Raindrop for macro AUROC, with other methods reserved for gas discordance and MAP decline.

E.2 Fusion Gains and Their Interpretation

Figure 5(c) compares nine fusion methods on MIMIC-III, five on VitalDB, and four on MC-MED with waveform-only and EHR-only controls. On MIMIC-III, all nine fusion methods improve macro AUROC over waveform-only (0.6062), but the strongest, bilinear fusion (0.6217), remains below EHR-only (0.6322). EHR-only also exceeds every fusion method for AKI, mortality, sepsis, and ventilator liberation; its mortality AUROC is 0.5898 versus at most 0.5567 for fusion, consistent with prognostic information in clinical records and clinician actions (Beaulieu-Jones et al., 2021). Circulatory deterioration instead favors waveform-based prediction: bilinear fusion reaches 0.6901 versus waveform-only 0.6878 and EHR-only 0.6620, aligning with the direct hemodynamic information used for circulatory warning (Hyland et al., 2020). VitalDB concatenation leads the macro at 0.7123, followed by weighted late (0.7106) and gated fusion (0.7093), all above EHR-only (0.6902). Every VitalDB fusion method improves hypotension AUROC despite EHR-only underperforming waveforms, showing that a weak standalone modality can still add useful conditional information. MC-MED late fusion leads macro AUROC (0.6093 versus waveform-only 0.5669 and EHR-only 0.6065), while gated and late fusion are closely ranked for anemia (0.6708 and 0.6706; Figure 32). EHR-only leads lactate nonclearance at 0.5360; late fusion is the only fusion method to improve this task over waveform-only, reaching 0.5203 versus 0.5082. Lactate's preference for EHR information is consistent with the role of metabolic and treatment trajectories in lactate kinetics (Vincent et al., 2016). The separate matched VitalDB validation, test, and missing-branch controls are detailed in Appendix F.3. Bilinear fusion is the leading MIMIC-III candidate for augmenting waveform models, concatenation, weighted late, and gated fusion form the VitalDB shortlist, and late fusion leads the MC-MED macro; retain EHR-only controls to determine which gains include an additional waveform contribution.

E.3 Training-Data Budgets

The comparison holds evaluation cohorts fixed and applies eight selectors, including random selection, under restricted training-data budgets. MIMIC-III uses quotas of 10/20/30% of training windows; VitalDB and MC-MED select whole subjects at 5/10/20% and 10/20/50%, respectively. These dataset-specific fractions run from light to dark in Figures 33–35; the composite's numeric shade legend illustrates MIMIC-III's fractions. The earlier MC-MED efficacy/efficiency companion remains a separate validation sweep (Appendix E.6). Every reduced-data condition lowers MIMIC-III macro AUROC and VitalDB forecast-macro AUROC relative to the full training set. At 30% of MIMIC-III windows, random selection leads at 0.5898 versus the full-data 0.6056; NLDR reaches 0.5887 at 20%, with no further improvement at 30%. Task-anchor selection worsens from 0.5044 to

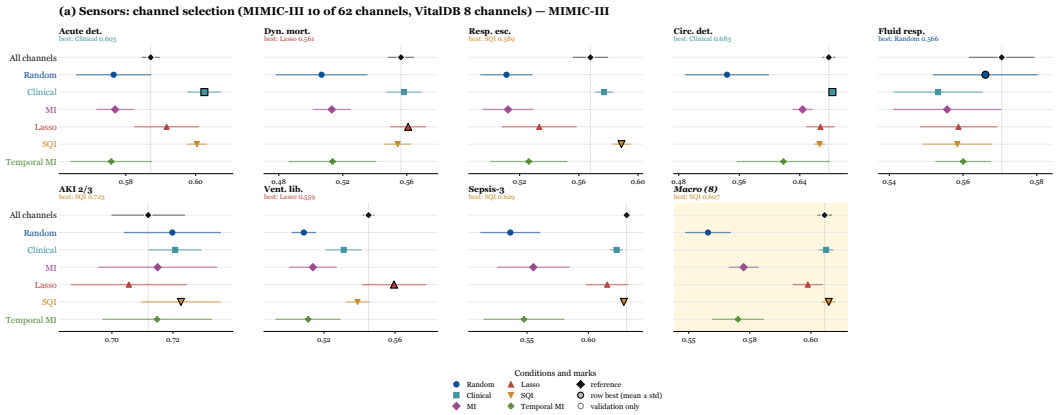


Figure 26: **Channel selection: MIMIC-III.** Comparison of six policies selecting ten of 62 canonical channels with all channels across eight tasks and their macro. Mark and uncertainty definitions follow Appendix E.4. Numerical values are in Table 66.

0.4960 as its window quota increases from 10% to 30%, so additional transition-focused examples do not recover the performance of random sampling. VitalDB deficits shrink as subsets grow: at 20%, gradient matching and noise filtering lead at 0.5941 and 0.5933 versus the full-cohort 0.6125. On MC-MED, most subsets lower test macro AUROC; at 50%, diversity and training-dynamics selection reach 0.5673 and 0.5695 versus the full-cohort 0.5669. Gradient matching and task-anchor selection remain below the MC-MED reference at every displayed budget. Under data constraints, random selection at 30% or NLDR at 20% are MIMIC-III starting points, gradient matching and noise filtering at 20% form the VitalDB shortlist, and training dynamics and diversity at 50% are the MC-MED candidates.

Completed VitalDB test: training-subject budgets. The later test release evaluates the frozen five-minute Transformer/Patch recipe with gradient matching at 5%, 10%, and 20% of training subjects and a matched full-data comparator. It averages 25 individual model metrics, giving equal weight to folds within each seed and then to the five seeds, without probability ensembling. All subject budgets retain 600 training updates, and intervals use 2,000 shared whole-subject draws conditional on fitted models. At 10% and 20%, concurrent BIS AUROC is 0.882 and 0.884 versus 0.883 with all subjects, whereas seven-task forecast AUROC is 0.544 and 0.578 versus 0.630. The 20%-minus-full forecast AUROC difference is -0.0520 (95% subject-bootstrap CI -0.0692 to -0.0340). The preserved concurrent-state AUROC and lower forecast AUROC indicate that a subset sufficient for estimating current anesthetic state is insufficient for learning future-event discrimination under the same recipe. BIS AUPRC remains lower at 10% and 20% (0.584 and 0.592 versus 0.606), and both paired AUPRC intervals exclude zero in the negative direction. The full cohort is the stronger forecasting choice, whereas 10–20% gradient-matched subsets are viable for concurrent-state AUROC, where the observed loss is small.

E.4 Task-Level Information Availability Results

Figures 26–35 show absolute test AUROC for each task and its dataset macro under the four information-availability comparisons. Markers show five-seed means and horizontal intervals show one seed SD; the macro interval propagates task SDs as $\sqrt{\sum_t s_t^2}/K$ and omits cross-task covariance. Black diamonds and vertical dotted lines identify the reference; black outlines mark the highest-mean tested alternative, excluding the reference. VitalDB macros average seven forecasts, with concurrent VCS05 displayed separately where available. These figures and the main composite use the same September 17 collection; MC-MED inputs are ECG-II and respiration (Appendix E).

E.5 Transfer to MIMIC-IV

Targets and patient roles. The two transfer cells in Figure 5(e) evaluate five waveform-eligible MIMIC-IV targets (Table 7). Figures 36 and 37 provide the detailed target-level results. Each

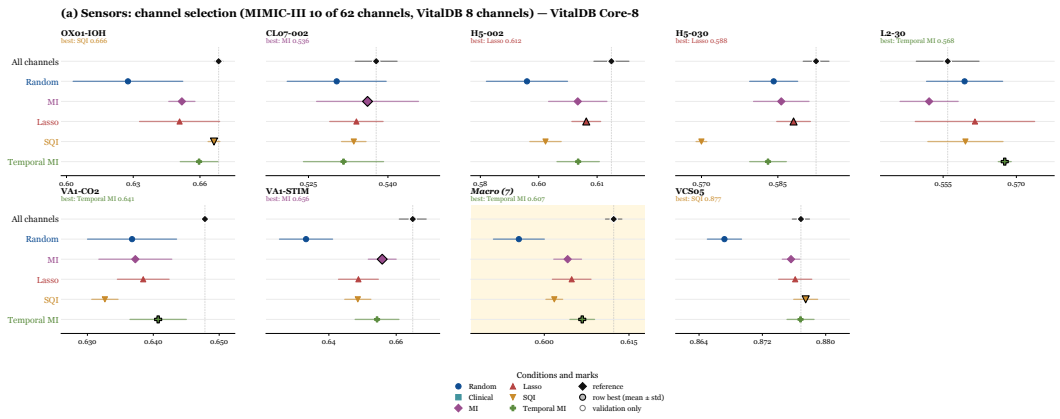


Figure 27: **Channel selection: VitalDB.** Five eight-channel policies against all permitted channels. The macro excludes concurrent VCS05. This fixed-budget AUROC comparison differs from the pooled-budget efficacy/efficiency companion. Mark definitions follow Appendix E.4. Numerical values are in Table 67.

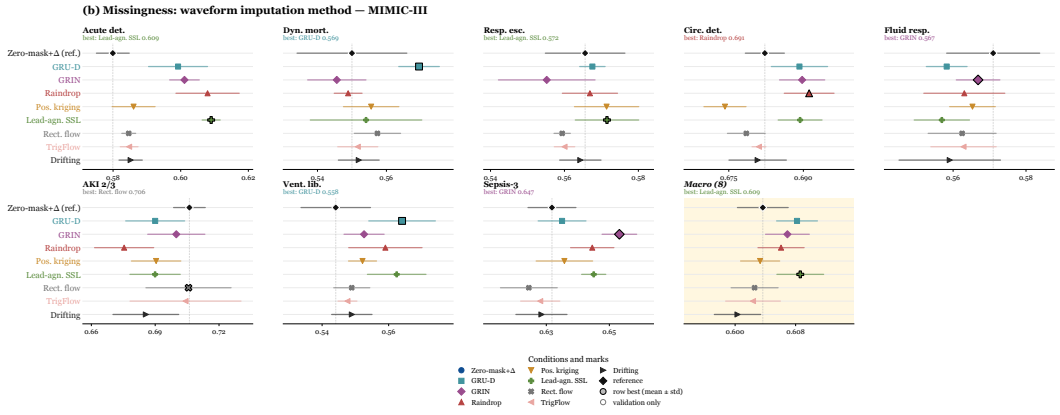


Figure 28: **Missing-input handling: MIMIC-III.** Eight alternatives compared with zero-mask plus Δ across eight tasks and their macro. The zero-mask reference supplies missingness features without reconstructing channels. Mark and uncertainty definitions follow Appendix E.4. Numerical values are in Table 68.

target is an operational forecasting label with its original temporal contract; these target cohorts form an additional adaptation study and are not added to the core 20-task battery or its window count. Patients overlap across targets, so the table cannot be summed to obtain a unique-patient total. Mortality has 157 waveform-eligible patients and eight positive patients but is not estimable under this protocol and is omitted. The target cohorts were inspected in earlier AutoAdapt work, and MIMIC-III/MIMIC-IV patient overlap remains unverified; these transfer results are retrospective development comparisons.

Patients rotate through five outer roles, with approximately 20% used for adaptation and 80% for testing in each role. Each patient enters adaptation once and testing four times; each adaptation role has four inner folds, except AKI role 4, which has three. Both patient predictions and patient labels are the maxima over that patient's windows. This patient-level endpoint and its repeated outer-role evaluation differ from the core battery's window-level scores.

Source models and target inputs. Both source conditions use our supervised Transformer with the Conv1D tokenizer, approximately 500k parameters in total: 275k in the encoder/tokenizer and 225k in the task head. For MIMIC-III, the acute-deterioration source model initializes the acute-deterioration, delirium/neurological, and future-SOFA targets; AKI and circulatory-deterioration

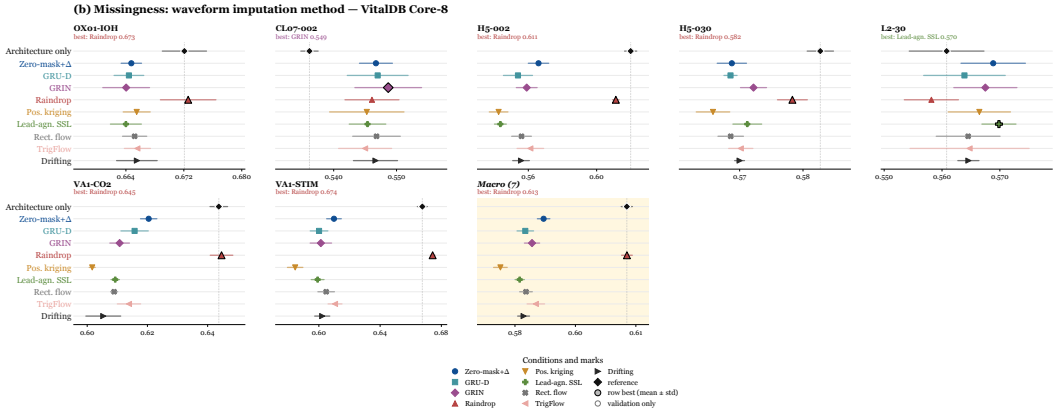


Figure 29: **Missing-input handling: VitalDB.** Nine methods compared with architecture-only handling under natural channel availability, for seven forecasting tasks and their macro. Mark and uncertainty definitions follow Appendix E.4. Numerical values are in Table 69.

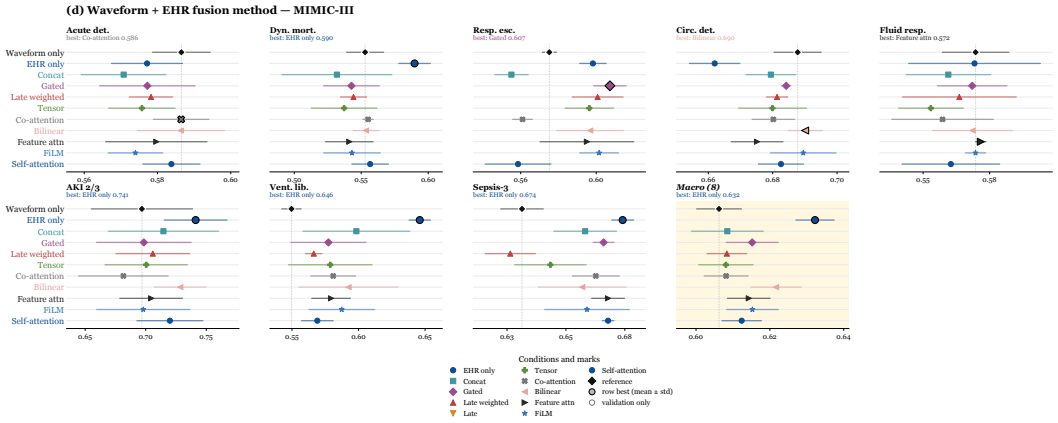


Figure 30: **Waveform–EHR fusion: MIMIC-III.** Nine fusion methods and EHR-only prediction against the Transformer/Patch waveform-only reference across eight tasks and their macro. Mark and uncertainty definitions follow Appendix E.4. Numerical values are in Table 70.

source models initialize their corresponding targets. The VitalDB source is the OX01-IOH model with a five-minute forecast gap. Five source seeds are matched to five target-initialization seeds; the VitalDB checkpoints are also fold matched. The source head and optimizer are discarded, and each target receives a fresh head.

MIMIC-IV inputs use raw waveform samples from the selected window, passed directly to the model tokenizer; NaN marks unobserved channels. The MIMIC-III-source condition uses raw physical units on the canonical 62-channel layout. The VitalDB-source condition pads this layout with four unobserved slots and uses median/IQR normalization fitted on adaptation patients only. Each condition has a randomly initialized waveform reference with matching inputs and preprocessing. Cross-source differences therefore combine initialization and preprocessing effects.

Adaptation and model selection. The four waveform arms update the full model, the last Transformer block plus head, the head only, or a matched randomly initialized full model. Each fit uses batch size 128, BF16, at most 600 updates, and learning rates $\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$. Checkpoint updates and learning rates are selected using inner-validation patient-max AUROC. The revised protocol used in Figure 5(e) refits once on all adaptation patients in each outer role, using the selected learning rate and an update budget equal to the lower median of the selected inner update counts. That refit is evaluated on the role's test patients; it is not an ensemble of the inner-fold models. The detailed figures also retain the original protocol's selected-inner-fit results for comparison.

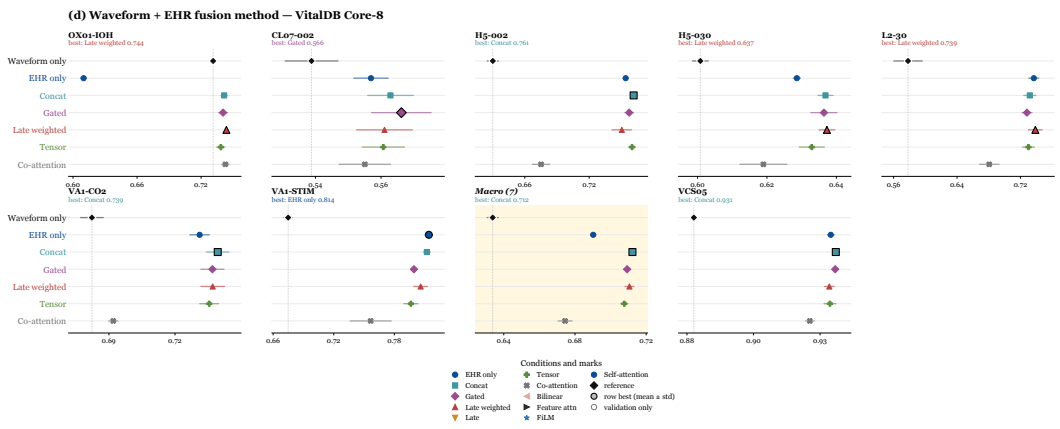


Figure 31: **Waveform–EHR fusion: VitalDB.** Five fusion methods and EHR-only prediction against the Mamba-3/Patch waveform-only reference with 60-minute history. The forecast macro excludes concurrent VCS05. Mark and uncertainty definitions follow Appendix E.4. Numerical values are in Table 71.

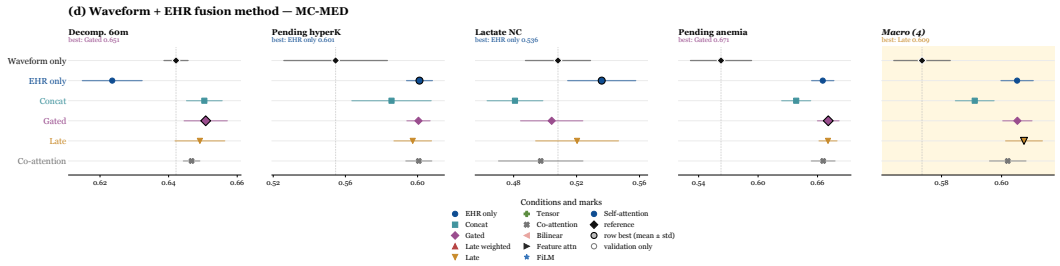


Figure 32: **Waveform–EHR fusion: MC-MED.** Four fusion methods and EHR-only prediction against the Transformer/Patch waveform-only reference with ECG-II and respiration across four tasks. Five-seed test results match the main composite. Mark definitions follow Appendix E.4. Numerical values are in Table 72.

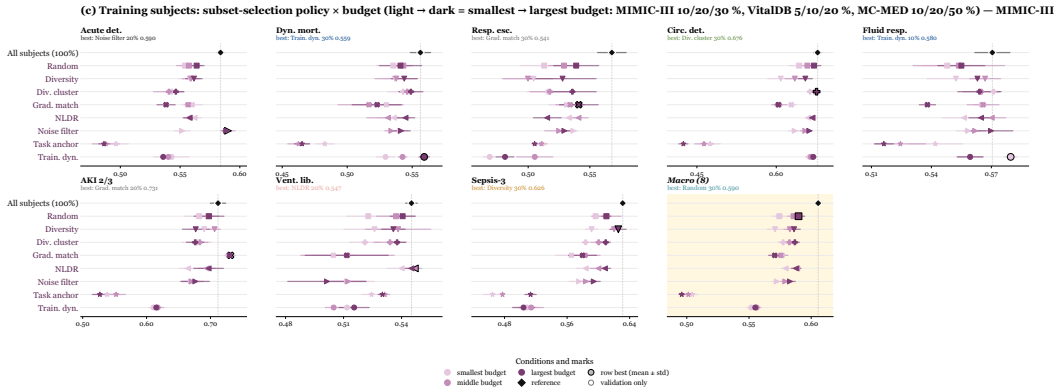


Figure 33: **Training-window selection: MIMIC-III.** Eight policies, including Random, at 10%, 20%, and 30% of training windows, ordered from light to dark, are compared with the full training set. Mark and uncertainty definitions follow Appendix E.4. Numerical values are in Table 73.

Feature-based controls. EHR only uses regularized logistic regression on 183 numeric features and one categorical feature summarizing the six hours before the index time, with observed-value indicators. Waveform features use logistic regression on the source encoder's frozen 128-dimensional embedding, and EHR plus waveform features combines both inputs. All controls use the same

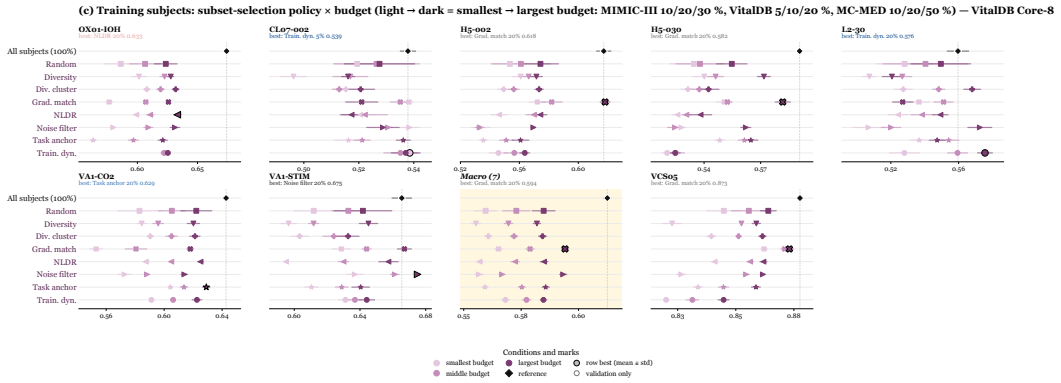


Figure 34: **Training-subject selection: VitalDB.** Eight policies, including Random, at 5%, 10%, and 20% of training subjects, ordered from light to dark, against the full cohort. VCS05 is separate from the forecast macro. Mark definitions follow Appendix E.4. Numerical values are in Table 74.

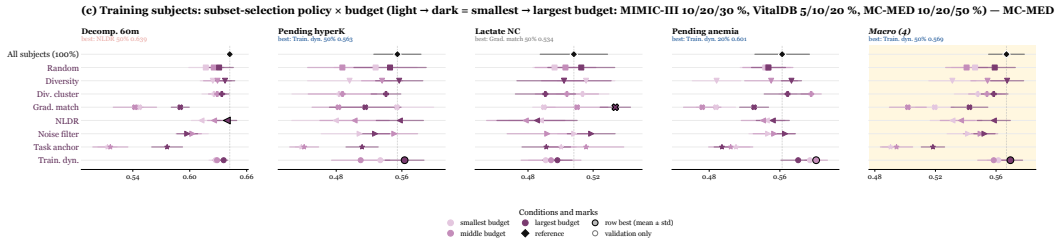


Figure 35: **Training-subject selection: MC-MED.** Eight selectors, including Random, using 10%, 20%, and 50% of training subjects, ordered from light to dark, compared with the full cohort. Results are five-seed test means from the September 16 campaign. Mark definitions follow Appendix E.4. Numerical values are in Table 75.

waveform-eligible patients as the waveform arms. Regularization uses $C \in \{0.01, 0.1, 1, 10\}$; imputation and scaling are fitted on adaptation patients only, and each patient receives equal total weight. The selected classifier is refitted once on all adaptation patients and scored once on the role's test patients.

Result encoding. In Figure 5(e), bars average AUROC changes across the five targets relative to the revised random-initialization refit for the corresponding source condition; markers show individual target changes. Blue circles denote acute deterioration, teal squares AKI, purple diamonds circulatory deterioration, red upward triangles delirium/neurological outcomes, and gold downward triangles future SOFA. Test values average outer-role results over five matched seeds; EHR-only logistic regression is deterministic. Figures 36 and 37 show absolute scores for each target, with MIMIC-III sources in row (a) and VitalDB sources in row (b). Light and dark blue bars distinguish original selected inner fits from revised refits on all adaptation patients; the feature-based controls are shown separately. Hollow diamonds report the original inner out-of-fold validation scores and should be distinguished from the test bars. The AUPRC panel evaluates the same AUROC-selected configurations.

Adaptation depth and clinical information. MIMIC-III head-only adaptation has the highest macro AUROC among the three tuning strategies, making preservation of the source encoder competitive with updating it. Circulatory deterioration and AKI improve over random initialization under all three MIMIC-III adaptation strategies. VitalDB initialization offers no consistent macro benefit, with partial adaptation helping AKI but circulatory deterioration and delirium remaining below random initialization. For MIMIC-III circulatory deterioration, head-only adaptation improves AUROC while logistic regression on the same frozen encoder features lowers it, locating this difference in the readout and fitting procedure rather than encoder updates. EHR-only and combined-feature controls give the largest macro gains, about 0.05 AUROC in the VitalDB-source condition, driven mainly by

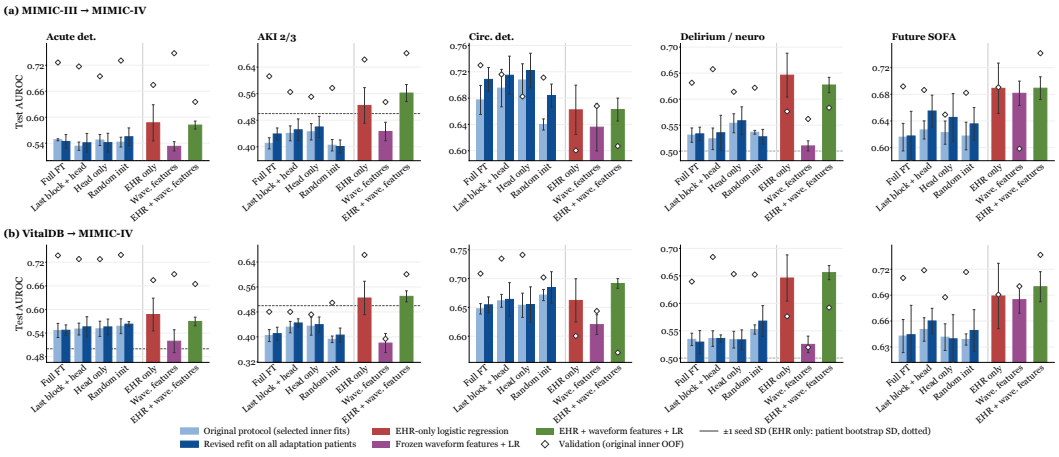


Figure 36: **Detailed transfer to MIMIC-IV: AUROC.** Absolute patient-level test AUROC for five targets, using (a) MIMIC-III and (b) VitalDB source models. Light/dark blue bars show the original selected-inner-fit protocol and revised full-adaptation refits; red, purple, and green bars show EHR-only, frozen waveform-feature, and combined logistic-regression controls. Hollow diamonds are original inner-OOF validation scores. Error bars denote ± 1 seed SD, except EHR-only, which uses patient-bootstrap SD (dotted); these are not confidence intervals for paired gains. Inputs, patient roles, and selection follow this subsection. Numerical values are in Table 78.

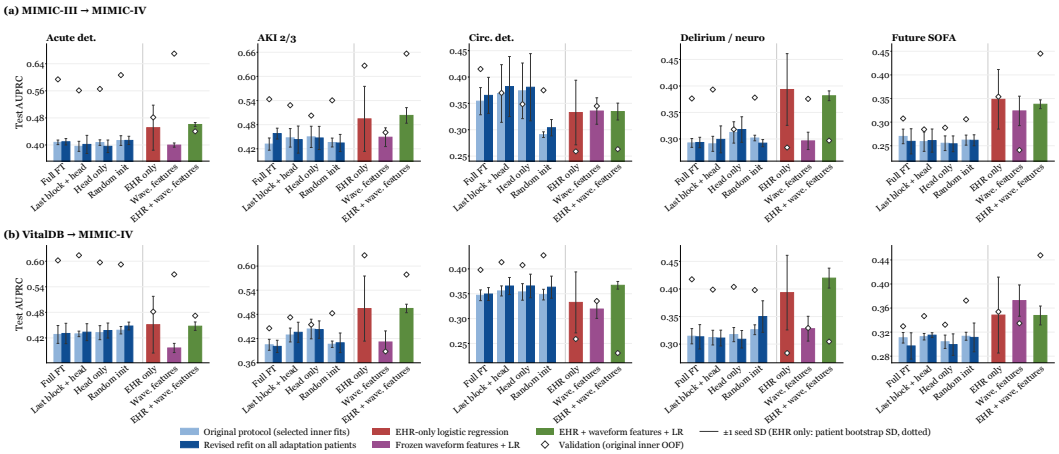


Figure 37: **Detailed transfer to MIMIC-IV: AUPRC.** Patient-level test AUPRC for the same AUROC-selected configurations and targets as Figure 36, with (a) MIMIC-III and (b) VitalDB sources. Colors, validation diamonds, and SD bars use the same encoding; EHR-only bars use patient-bootstrap SD rather than seed variability. Compare methods within each target because prevalence differs across targets. Numerical values are in Table 79.

AKI and delirium. Waveform adaptation's advantage over EHR-only for circulatory deterioration aligns with the relevance of hemodynamic measurements to circulatory failure (Hyland et al., 2020). Conversely, EHR-based gains on AKI align with its dependence on renal measurements in the diagnostic criteria (KDIGO Acute Kidney Injury Work Group, 2012). For MIMIC-III-source AKI, partial adaptation has higher AUROC than full fine-tuning, while full fine-tuning has the highest AUPRC among the adaptation arms. MIMIC-III head-only adaptation is a useful waveform-transfer starting point, with partial and full fine-tuning added to the AKI shortlist according to whether AUROC or AUPRC is the development objective; EHR-only remains a control for the value added by waveform adaptation.

E.6 Efficacy, Efficiency, and Safety Under Information Constraints

The companion figures report precision–recall performance, training cost, calibration, and selective risk under changes in available information. They retain earlier experimental exports; the refreshed AUROC comparison uses different model averages and a zero-mask plus Δ reference for MIMIC-III missingness (Appendix E). Appendix A.2 specifies the definitions and measurement coverage.

Sensor selection. Figures 38 and 39 compare the predictive and resource costs of channel reduction. MIMIC-III clinical channel selection retains macro AUPRC 0.096 versus 0.098 with all channels, while peak memory decreases from 3.395% to 2.353% and training time from 108.6 to 100.1 seconds. Its ECE increases from 0.178 to 0.228 and AURC from 0.330 to 0.382, making the memory saving less attractive when confidence quality is a priority. Clinical selection is useful when MIMIC-III memory use is constrained; all channels provide the stronger reference for calibration and selective risk.

Missing-input handling. Figures 40 and 41 assess whether added missing-input modules improve predictive quality enough to justify their cost. MIMIC-III GRU-D improves AUPRC from 0.098 to 0.100 and ECE from 0.183 to 0.136, consistent with prior evidence for modeling informative missingness (Che et al., 2018). Its AURC increases from 0.340 to 0.347 and training time from 109.0 to 369.3 seconds, so these benefits require more computation without improving selective risk. On VitalDB, all nine missing-input methods have higher macro ECE and AURC than architecture-only handling; their AUPRCs span 0.073–0.078 versus the reference 0.078. Native handling is a practical initial configuration; GRU-D offers better MIMIC-III calibration when its additional training cost is acceptable.

Subject budgets and optimization schedules. Figures 42–44 compare reduced training cohorts. At a 20% budget, MIMIC-III diversity-cluster selection has AUPRC 0.095 and training time 35.0 seconds, versus 0.098 and 108.6 seconds for the full cohort; ECE rises from 0.178 to 0.213. VitalDB reduced-cohort AUPRCs remain below the full-cohort value of 0.080. At 20%, noise filtering and diversity-cluster selection both round to AUPRC 0.072, but noise filtering has lower ECE (0.022 versus 0.044) and AURC (0.029 versus 0.033). VitalDB's earlier training cells remain between 31.8 and 32.2 seconds under fixed update budgets; the smaller cohorts receive similar scheduled optimization time. MC-MED subject-budget results and the full-cohort reference both summarize five seeds. At a 20% budget, diversity-cluster selection reduces MIMIC-III training time, while noise filtering is preferable to diversity clustering on VitalDB when confidence quality matters.

Fusion across tasks and metrics. Figures 45–47 compare waveform–EHR fusion. MIMIC-III fusion AUPRCs remain between 0.096 and 0.098, while co-attention reduces ECE from the waveform-only 0.183 to 0.136 and AURC from 0.340 to 0.310. On VitalDB, gated and weighted late fusion have similar AUPRCs (0.1256 and 0.1255), but weighted late fusion has lower ECE (0.0251 versus 0.0306). For hypotension in this companion study, weighted late fusion improves AUPRC over EHR alone (0.151 versus 0.093), yet remains below waveform-only (0.161). MC-MED fusion now covers all four endpoints for AUPRC: late fusion records macro 0.210, EHR-only 0.209, co-attention 0.193, and waveform-only 0.179. For pending hyperkalemia and anemia, EHR-only has higher AUPRC than every fusion method, whereas late fusion leads decompensation and gated fusion leads lactate nonclearance. MC-MED fusion ECE and AURC cover decompensation alone: late fusion improves its AUPRC from 0.210 to 0.260 but increases AURC from 0.279 to 0.327, while co-attention has the lowest AURC at 0.178. Weighted late and gated fusion are the leading VitalDB AUPRC candidates, with lower ECE favoring weighted late fusion; late fusion is the strongest MC-MED macro-AUPRC option for waveform augmentation, with EHR-only retained as a control.

F Additional VitalDB Results

F.1 Architecture, Data, and Missingness

Scope. The waveform-only grids comprise 27 architecture/tokenizer pairs, 20 duration conditions, eight subject selectors, 48 channel conditions, and 30 missingness conditions. Across five seeds and five subject folds, normalization uses training data only; missing streams and source exclusions are preserved. Forecast summaries give equal weight to the 5/15-minute gaps within each task, then to the seven tasks. Concurrent BIS is reported separately, with its EEG and BIS sources withheld.

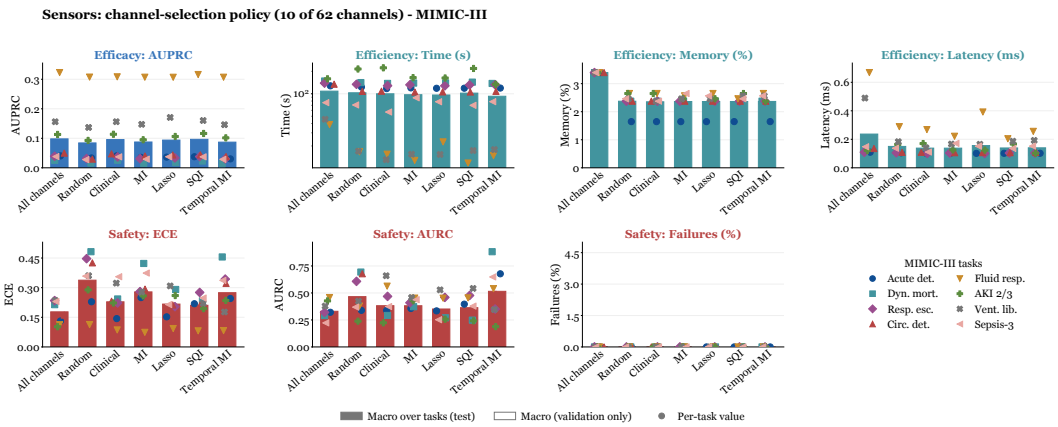


Figure 38: **Sensor selection — MIMIC-III.** Ten selected channels out of 62 canonical slots against the all-channel reference. Panels show AUPRC, training time, memory, time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 103 and 104.

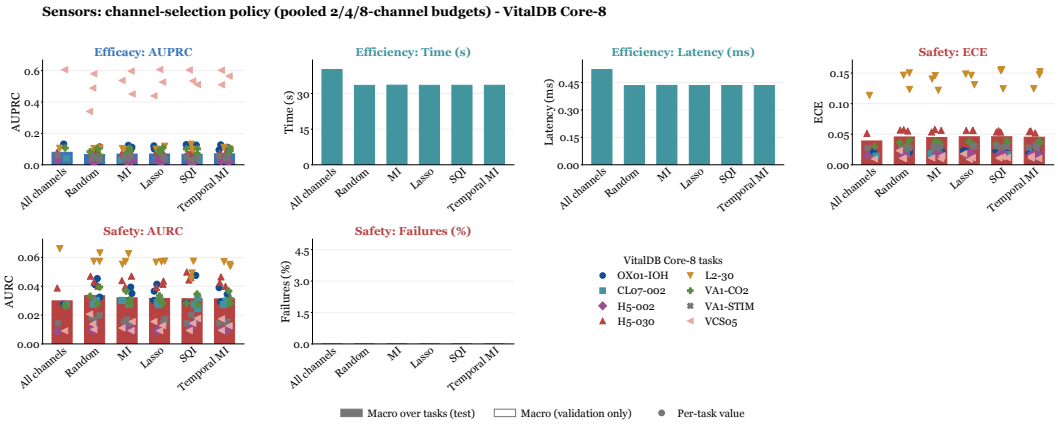


Figure 39: **Sensor selection — VitalDB Core-8.** Policy means pool two-, four-, and eight-channel budgets against the all-channel reference. Forecast macros exclude concurrent VCS05. Cost panels use earlier-grid telemetry per validation cell. Panels show AUPRC, training time, time per training sample, ECE, AURC, and audited prediction-file failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 105 and 106.

Figure 48 retains the earlier crossed grid and ensemble-dropout controls, which are not included in the current test-metric figures in Appendix E.6.

Representation and history. Patch gives the highest forecast-validation AUROC for every architecture, with Transformer and FITS tied at 0.635. The encoder ordering nevertheless changes with the tokenizer: Transformer exceeds S4 with Patch (0.635 versus 0.632), while S4 exceeds Transformer with Crossformer (0.621 versus 0.573). Extending S4/SeFT history from five to 60 minutes lowers concurrent BIS AUROC from 0.812 to 0.607, making this longer-history configuration unfavorable for current-state estimation. Transformer/Patch and FITS/Patch provide a competitive initial pair from this grid, with the five-minute view sufficient for the concurrent-state comparison.

Data and sensor constraints. Gradient matching at 10–20% of subjects approaches full-data concurrent AUROC but still falls short in forecasting. Eight temporally selected channels approach full-panel performance, while native masking gives the highest natural-missingness validation mean. The native-masking ensemble's test AUROC falls from 0.619 to 0.583 under 50% dropout. Forecast development can start from native masking and the full cohort, then compare the eight-channel temporal selector under a sensor budget and dropout under additional sensor loss.

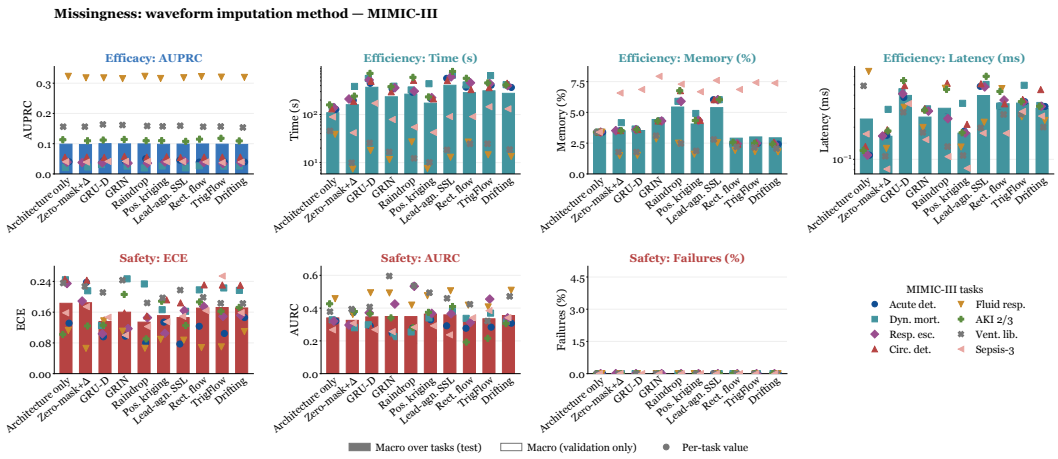


Figure 40: **Missing-input handling — MIMIC-III.** Missing-input methods compared with the architecture-only reference. Panels show AUPRC, training time, memory, time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 107 and 108.

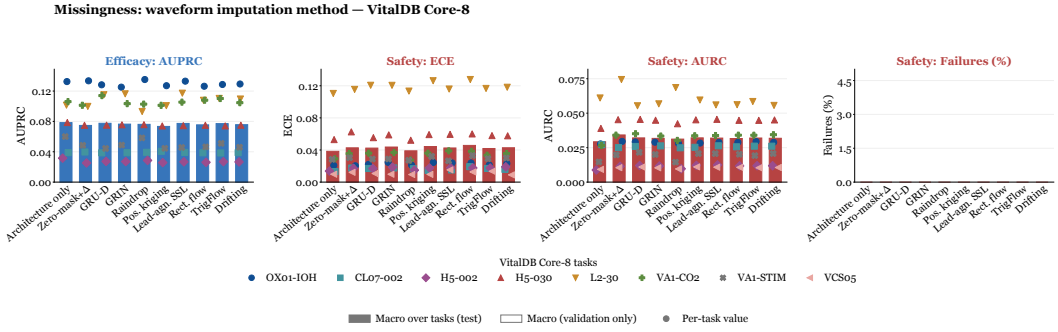


Figure 41: **Missing-input handling — VitalDB Core-8.** Missing-input methods against the architecture-only reference. Forecast macros exclude concurrent VCS05. Panels show AUPRC, ECE, AURC, and audited prediction-file failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 109 and 110.

F.2 Forecast Gaps from Five to 75 Minutes

Design. A fixed 30-minute Transformer/Patch recipe is evaluated at six gaps on seven forecast tasks: 1,050 seed–fold fits, including 350 reused 5/15-minute references and 700 new fits. Figure 49 reports this validation study, separately from the main test curves in Figure 4. Metrics pool held-fold predictions within each seed, average five seeds, then weight tasks equally. The concurrent probe is excluded.

Results and probability quality. From five to 75 minutes, macro AUROC declines from 0.643 (95% CI 0.635–0.651) to 0.561 (0.546–0.575), and AUPRC from 0.078 to 0.052. Brier score increases from 0.0522 to 0.0608, log loss from 0.214 to 0.288, and ECE from 0.042 to 0.057. Task responses differ: core hypothermia is nonmonotonic, while MAP-decline AUROC at 75 minutes is 0.491 (0.446–0.535). This configuration is therefore unsuitable as a starting point for long-lead MAP-decline warning. The overall lead-time loss agrees with earlier arterial-waveform hypotension forecasting (Hatib et al., 2018), while the nonmonotonic temperature curve shows that this pattern is endpoint dependent.

Support and interpretation. At 75 minutes, tasks retain 710–2,251 subjects, including 94–359 with positive windows. All 42 heads pass support checks; the smallest held fold contains 130 subjects and 16 with positive windows. Intervals use 2,000 shared whole-subject Poisson draws over the

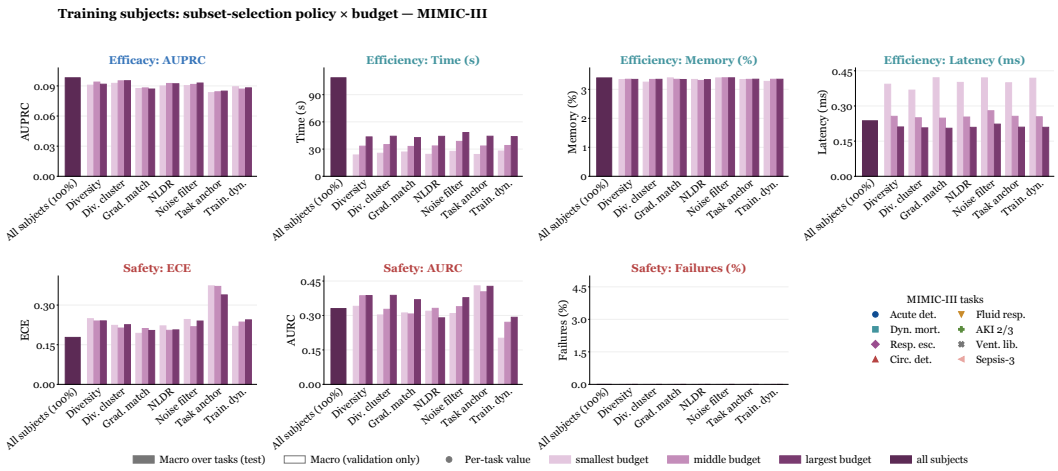


Figure 42: **Training-subject budgets — MIMIC-III.** Grouped bars compare 10%, 20%, and 30% subject budgets with the full cohort. Panels show AUPRC, training time, memory, time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 111 and 112.

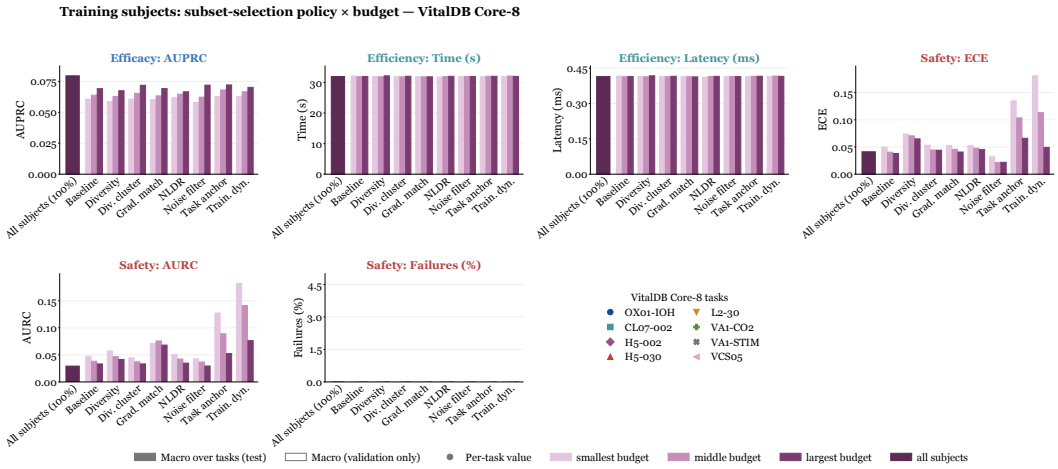


Figure 43: **Training-subject budgets — VitalDB Core-8.** Grouped bars compare 5%, 10%, and 20% subject budgets with the full cohort. Forecast macros exclude concurrent VCS05. Cost panels use earlier-grid telemetry per validation cell. Panels show AUPRC, training time, time per training sample, ECE, AURC, and audited prediction-file failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 113 and 114.

4,311-subject union with gap-specific eligibility masks. They condition on fitted models and omit retraining, selection, and multiplicity uncertainty. Patient fold assignments remain fixed across gaps, while follow-up support, event exclusions, and prevalence vary among the eligible anchors. No temporal-order permutation control was run.

Completed test: warning time. A later test readout of the fixed 30-minute Transformer/Patch recipe averages 25 individual fold/seed model metrics and seven tasks equally at each gap, separately from the validation estimator above. Macro AUROC decreases from 0.659 (95% CI 0.638–0.679) at five minutes to 0.558 (0.530–0.585) at 75 minutes; AUPRC decreases from 0.094 to 0.060 and ECE increases from 0.041 to 0.058. The test readout uses fixed patient partitions, gap-specific eligible anchors, and 2,000 whole-subject draws conditional on fitted models; paired short–long-gap contrast intervals are not supplied. Its numerical trajectory differs from the main-figure test curve, so these exact scores and intervals remain specific to this fixed-recipe readout. This Transformer/Patch

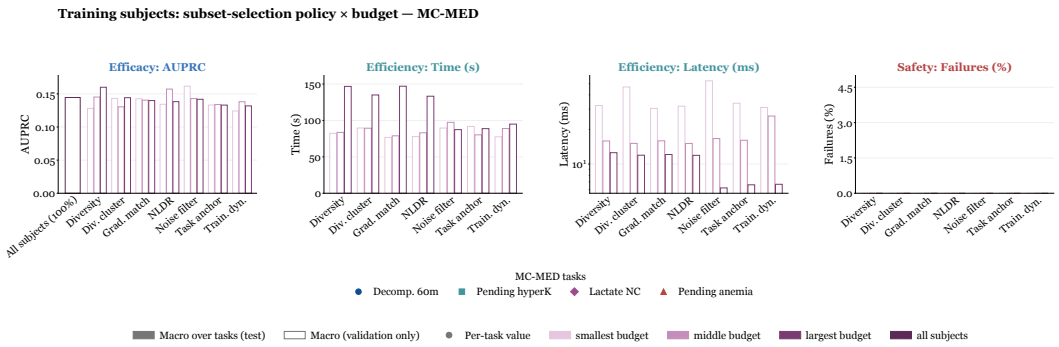


Figure 44: **Training-subject budgets — MC-MED.** Grouped bars compare 10%, 20%, and 50% subject budgets with the full cohort. Hollow bars show validation results averaged over five seeds, including the full-cohort reference. Panels show AUPRC, training time, time per training sample, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 115 and 116.

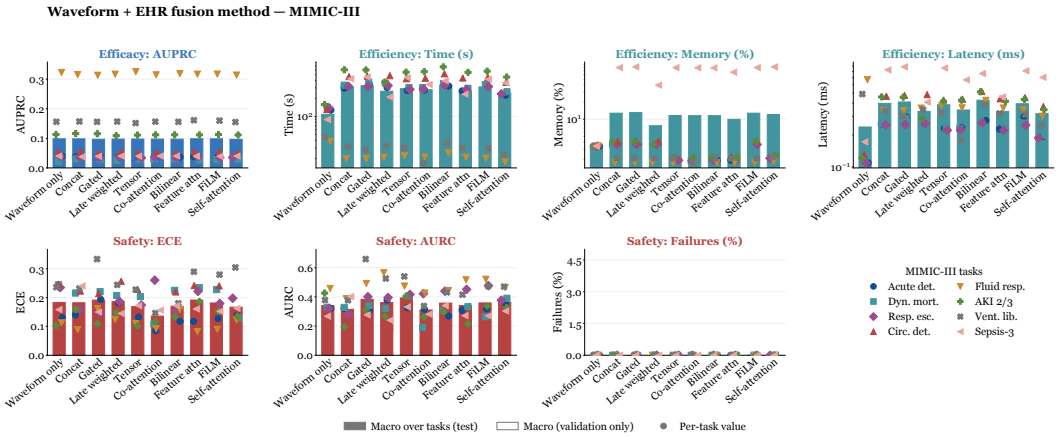


Figure 45: **Waveform–EHR fusion — MIMIC-III.** Fusion methods against waveform-only prediction. Panels show AUPRC, training time, memory, time per training sample, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 117 and 118.

recipe provides a baseline for five- to 15-minute warning, with additional clinical context evaluated for longer-gap prediction where waveform discrimination is weaker.

F.3 Waveform–EHR Fusion and Missing-Input Controls

Design. All seven conditions use identical episodes and task-specific exclusions. Waveform arms use 60-minute Mamba-3/Patch inputs; EHR arms use pre-cutoff information permitted by the task input rules. The conditions comprise two unimodal references and five fusion methods. The validation battery has 15 heads (seven tasks at 5/15-minute gaps and a separate concurrent probe), five folds, and five seeds. New fits use 600 updates at batch size 128. Forecast aggregation follows Appendix F.1. Figure 50 shows the original validation discrimination and inference-time controls; the current test AUPRC/ECE comparison appears in Figure 46.

Validation gains and probability quality. The stronger-unimodal envelope selects the better unimodal metric per head and bootstrap draw before aggregation. Weighted late fusion reaches AUROC 0.6716 versus 0.6483 for EHR and 0.6146 for waveforms. Its paired AUROC gain over the envelope is 0.0087 (95% CI 0.0043–0.0127), whereas its AUPRC gain is 0.0016 (−0.0012–0.0043). Brier score, log loss, and ECE worsen by 0.00116, 0.00447, and 0.00156, respectively, aligning with the established distinction between discrimination and probability quality (Van Calster et al., 2019).

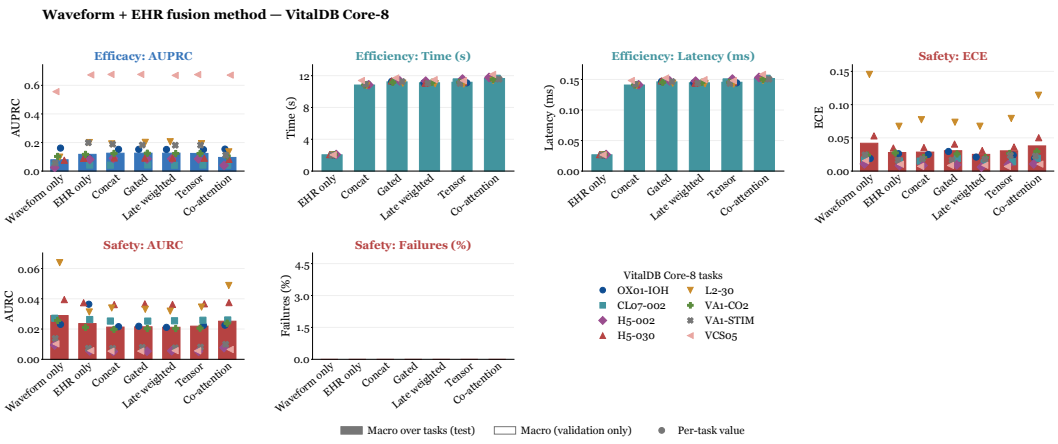


Figure 46: **Waveform–EHR fusion — VitalDB Core-8.** Fusion methods against waveform-only and EHR-only references. Forecast macros exclude concurrent VCS05. Cost panels use earlier-grid telemetry per validation cell. Panels show AUPRC, training time, time per training sample, ECE, AURC, and audited prediction-file failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 119 and 120.

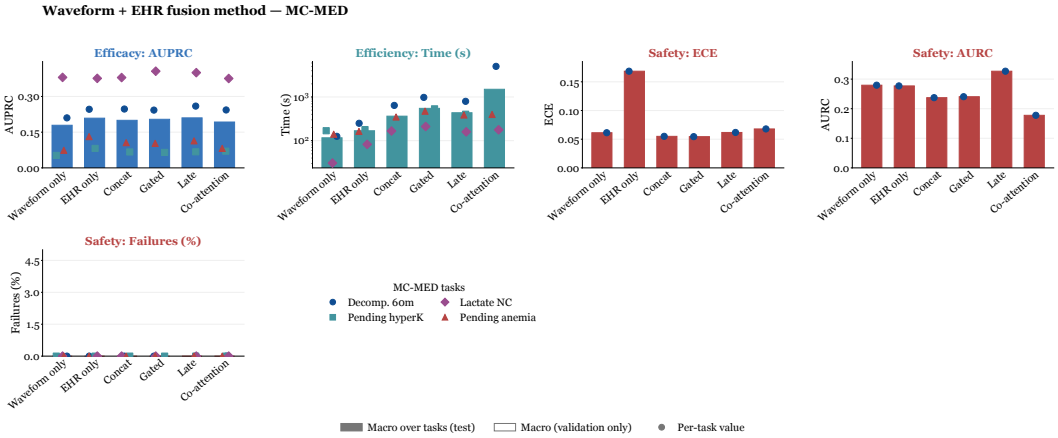


Figure 47: **Waveform–EHR fusion — MC-MED.** Fusion methods against waveform-only and EHR-only references. AUPRC, time, and failures cover four tasks; ECE and AURC cover decompensation only. Panels show AUPRC, training time, ECE, AURC, and recorded failures. Metric and mark definitions follow Appendix A.2. Numerical values are in Tables 121 and 122.

Concurrent late-fusion AUROC is 0.9069 versus EHR's 0.9060 (paired difference CI −0.0026–0.0046), leaving little observed benefit from fusion for this state-estimation task. Weighted late fusion is the stronger candidate for augmenting waveform forecasts; the small concurrent-state gain over EHR-only does not establish a comparable benefit from the waveform branch for that endpoint.

Robustness and uncertainty. All 1,875 frozen fusion checkpoints reproduce their natural-input predictions bitwise before we apply the 9,375 controls. Every method/control pair reduces forecast AUROC and AUPRC (all 50 unadjusted intervals exclude zero), with task-level exceptions. Late-fusion EHR zeroing lowers Brier to 0.04478 despite worse discrimination, whereas EHR branch absence raises it to 0.17294. Intervals use 2,000 shared subject draws over 4,596 subjects, with the same conditional uncertainty limits as the gap analysis. Robustness evaluations should distinguish EHR-value disruption from complete branch absence, because retaining observation masks and removing the branch produce substantially different probability errors.

Completed matched test comparison. The subsequent test release verifies identical rows, subjects, and labels across waveform and EHR conditions. It reports the mean of 25 individual model met-

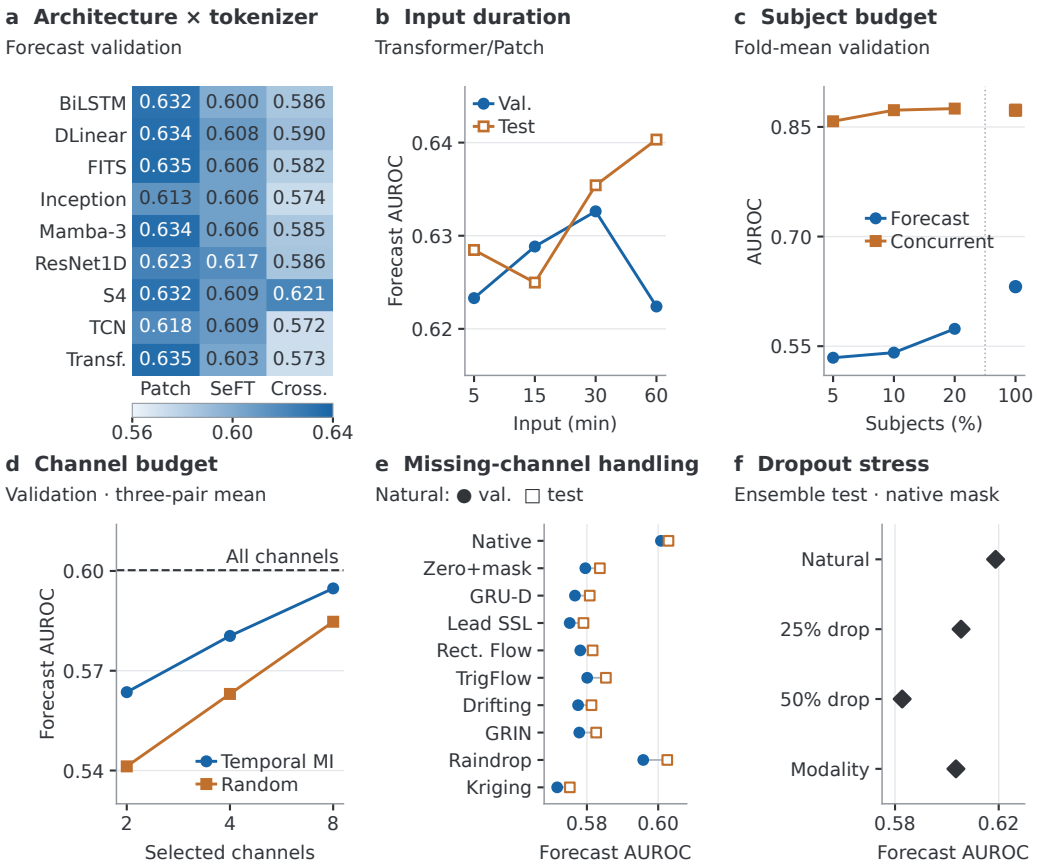


Figure 48: **VitalDB preprocessing and missingness.** All 27 architecture/tokenizer pairs and ten natural-missingness methods; selected duration, subject, and channel contrasts. Validation pools held-fold predictions within seed; test averages fixed-fold model metrics. Subject selection instead uses fold-mean validation, including its full-data comparator; the later reduced-budget test is reported separately in Appendix E.3. The separate stress panel scores five-fold probability ensembles on test. These estimators are not interchangeable. Channel and missingness panels average three fixed pairs. Architecture validation follows recipe selection, not nested evaluation. Numerical values are in Table 123.

rics, with equal gap weights within each forecast task and equal weights across the seven tasks; these scores are distinct from both pooled validation and probability-ensemble test estimates. Weighted late fusion reaches macro AUROC 0.6937 and AUPRC 0.1166, versus 0.6723 and 0.1074 for EHR-only and 0.6247 and 0.0729 for waveform-only. Relative to EHR-only, its paired AUROC gain is 0.0214 (95% CI 0.0124–0.0307) and AUPRC gain is 0.0092 (0.0041–0.0155). Relative to the head-wise stronger-unimodal envelope, however, neither gain is conclusive: AUROC $+0.0053$ (-0.0052 to 0.0120) and AUPRC $+0.0012$ (-0.0060 to 0.0071). ECE is approximately 0.0302 for late fusion versus 0.0306 for EHR-only, with an inconclusive paired difference; relative to the metric-specific envelope, ECE worsens by 0.00149 (0.00010–0.00417). For blood-gas/$ETCO_2$ discordance, gap-averaged AUROC favors EHR-only over late fusion (0.730 versus 0.697). The 2,000 shared whole-subject bootstrap draws condition on fitted models and are not adjusted for multiplicity. Weighted late fusion is a useful waveform-augmentation candidate for this forecast battery, with calibration assessed alongside discrimination and EHR-only retained as a control, especially for gas discordance.

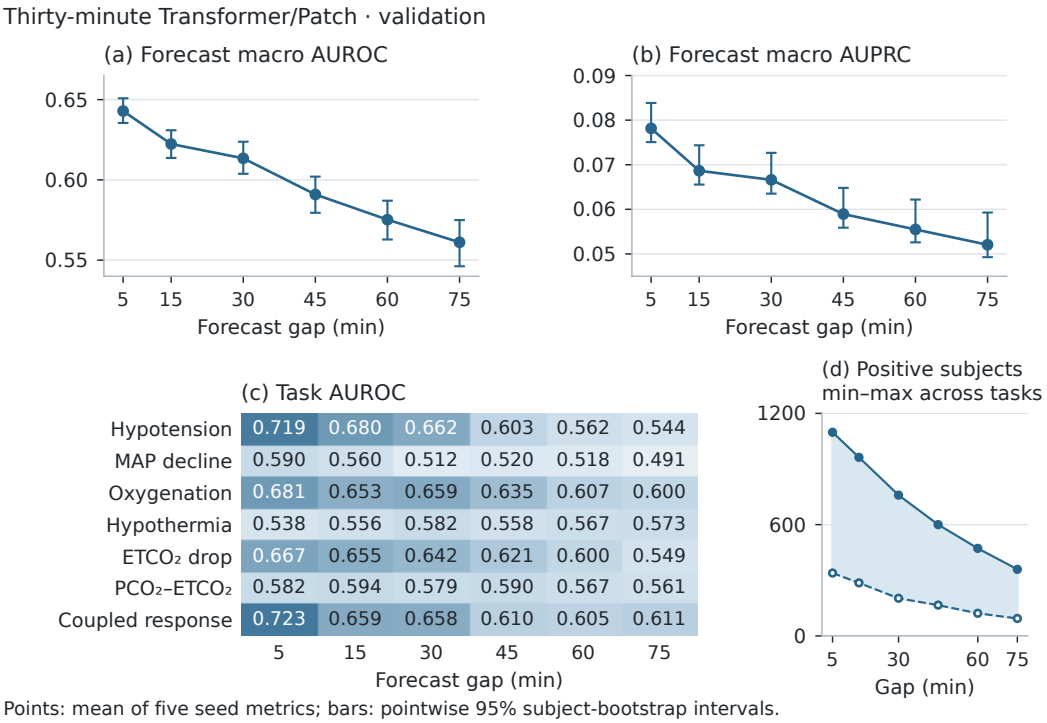


Figure 49: **Longer-gap forecasting.** Macro discrimination with pointwise 95% subject-bootstrap intervals, task AUROC, and the range of positive-window subject counts across tasks. Support ranges are neither pooled counts nor matched-cohort sizes. A positive-window subject has at least one positive episode and may also contribute negative episodes. Eligibility and incidence definitions differ across gaps; the plotted trend is descriptive. Numerical values and subject support are in Table 124.

G Interpreting Absolute Waveform-Only Performance

G.1 What the Reported AUROCs Establish

To interpret why many tasks fall below AUROC 0.8–0.9, we examine the available clinical information, effective warning time, and outcome ascertainment. Table 12 anchors the discussion in the numerical comparisons already reported in Table 10. For example, the MIMIC-III Transformer architecture entry ranges from 0.532 for dynamic mortality to 0.727 for AKI stage 2/3; the corresponding VitalDB forecast entries range from 0.547 for accelerating MAP decline to 0.695 for hypotension. The best MIMIC-III tokenizer macro, 0.615, comes from a different learning-rate search from the architecture panel, whose highest macro is 0.607 (Appendix B.7).

The concurrent BIS result demonstrates strong current-state discrimination with EEG and BIS-derived inputs excluded, whereas the forecasting tasks require anticipating a different endpoint after the input cutoff. Within-task evidence comes from the separate VitalDB gap study, where macro validation AUROC falls from 0.643 at five minutes to 0.561 at 75 minutes as warning time and gap-specific eligibility change (Appendix F.2).

G.2 Information, Timing, and Evaluation Constraints

Restricted information relative to clinical comparators. The waveform-only arm omits laboratory values, medication history, and other charted context that may directly inform organ dysfunction and treatment decisions. For example, the circulatory-failure system of Hyland et al. (2020) reports AUROC 0.94 using a broader clinical information set, while Tomašev et al. (2019) predicts creatinine-defined AKI from longitudinal EHR data. Within our VitalDB fusion validation study,

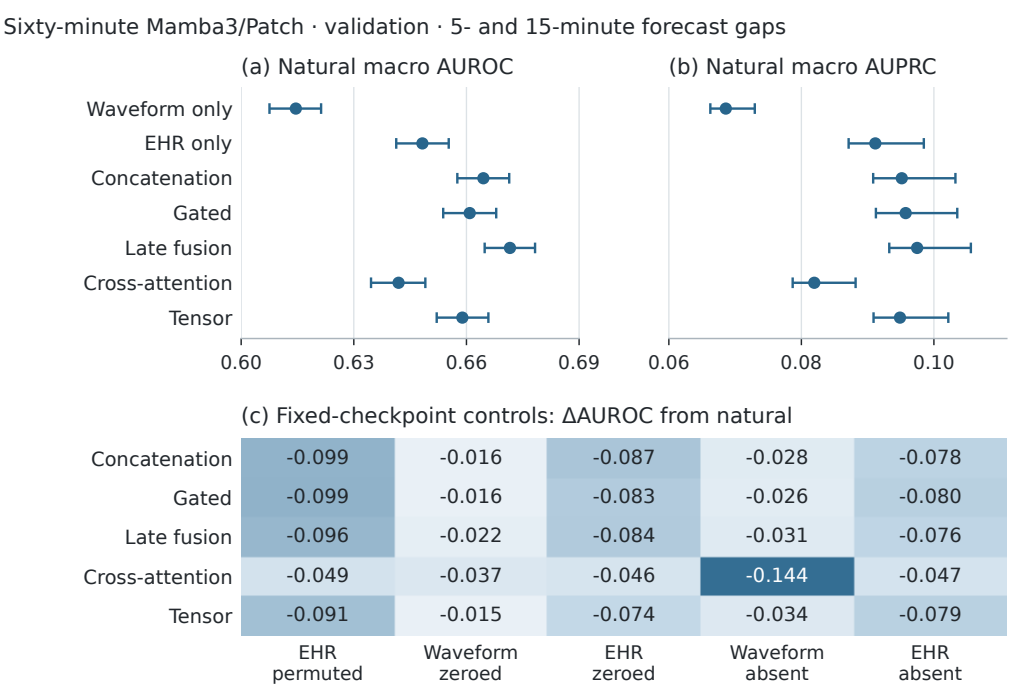


Figure 50: **Fusion discrimination and sensitivity.** Natural-input forecast means with 95% subject-bootstrap intervals, and each control's AUROC difference from its own natural-input model. EHR permutation shuffles observed feature values within the held fold, preserving recipient masks. Value zeroing retains masks; branch absence uses each architecture's frozen missing-branch behavior. Controls are inference interventions, not retrained unimodal baselines or causal feature importance. Numerical values are in Tables 125 and 126.

EHR-only macro AUROC is 0.6483 versus 0.6146 for waveforms, using matched episodes; these values belong to that study's 60-minute waveform protocol (Appendix F.3). Clinical actions themselves can carry predictive information: Beaulieu-Jones et al. (2021) obtained mortality AUROC 0.89 from administrative and clinician-initiated data. High waveform-only AUROCs are also possible for particular endpoints and input resolutions, as illustrated by arterial-waveform hypotension prediction (Hatib et al., 2018).

Warning requirements and input recency. Future-event tasks require information to remain useful beyond the waveform cutoff; concurrent BIS and pending laboratory-result tasks have distinct temporal interpretations. The MIMIC-III headline baseline uses $[T - 5\,\mathrm{min}, T)$ from the extracted six-hour history, so its input ends at the prediction anchor. The target interval $[T + G, T + G + H)$ begins G after this cutoff; for dynamic mortality, the six-hour gap and 24-hour horizon place it 6–30 hours after the cutoff. Observation-length comparisons extend the input backward from the same T, preserving the anchor and outcome window.

Sparse and uneven sensor availability. Canonical channel counts describe tensor positions; observed sensor coverage varies by patient. The MIMIC-III waveform audit covers 2,679 matched records from 2,315 subjects and reports ECG-II in 97.4% of records, plethysmography in 66.9%, arterial pressure in 49.6%, and central venous pressure in 17.8%. These are record-level statistics for the audited subset, not per-window coverage rates for the full benchmark. Consequently, some examples lack the pressure or respiratory signals most closely related to their endpoint, and the available channels can change over time. Explicit masks preserve information about monitoring decisions and illness severity carried by missingness (Che et al., 2018). The channel and imputation comparisons show that recovery of predictive value is task-dependent (Appendix E.1).

Input representation and optimization budget. The models receive raw waveform samples: a five-minute MIMIC-III channel at 125 Hz contains 37,500 samples before the selected tokenizer constructs its representation. The approximately 500k-parameter models and schedules of ten epochs on MIMIC-III or 600 updates on VitalDB define a bounded comparison budget. The 87,825 training runs distribute this budget across tasks, conditions, hyperparameters, seeds, and folds.

The evaluated population and negative examples. Evaluation retains the anchors meeting each task's eligibility and outcome-ascertainment rules, with patient-disjoint partitions and no additional selection of only easy negative examples. Task-specific stability criteria restrict evaluation to patients meeting the stated pre-anchor conditions. Selecting positive and negative segments using different blood-pressure ranges can exaggerate apparent discrimination; Enevoldsen & Vistisen (2022) identified this concern in the development and validation design of the Hypotension Prediction Index. Likewise, record-wise validation can overestimate generalization to new patients (Saeb et al., 2017).

Outcome support and uncertainty near chance. Low prevalence alone does not mathematically reduce population AUROC when the class-conditional score distributions are unchanged: AUROC measures the probability of correctly ordering a positive and a negative example (Hanley & McNeil, 1982). A small number of independent positive patients can nevertheless limit learning, widen uncertainty, and make learning-rate or checkpoint selection unstable. In MC-MED, validation contains only 23 positive lactate windows, 33 positive anemia windows, and 48 positive hyperkalemia windows (Table 6); repeated windows elsewhere in the benchmark do not create independent patient-level evidence. Sample-size adequacy depends on the outcome frequency, expected predictive signal, and model complexity, rather than a universal events-per-parameter rule (Riley et al., 2020). AUPRC and threshold-specific sensitivity and positive predictive value help interpret prediction under rare outcomes (Saito & Rehmsmeier, 2015). Calibration separately assesses whether predicted risks agree with observed event frequencies (Van Calster et al., 2019).

Recorded endpoints and healthcare processes. Laboratory ascertainment and interventions such as ventilation or vasopressor initiation reflect both patient physiology and clinical decisions. Their recorded times can differ from the onset of the underlying physiological change, while treatment after the input cutoff can alter the eventual outcome. The influence of ordering and recording processes on EHR-derived prediction targets is well documented (Agniel et al., 2018). Outcome-coverage requirements therefore determine which anchors can receive an observed negative label.

G.3 Why Difficulty Differs Across Clinical Tasks

MIMIC-III: downstream outcomes and intervention responses. Dynamic mortality concerns a delayed trajectory influenced by organ dysfunction, treatment, and comorbidity beyond a short waveform fragment; the composite acute-deterioration endpoint combines several such pathways. Respiratory escalation and ventilator liberation additionally depend on clinical assessments that are incompletely represented by monitor waveforms. For extubation, a meta-analysis identifies respiratory, circulatory, neurological, and secretion-management factors, supporting a broader assessment than a single physiological index (Torrini et al., 2021). The fluid-response task predicts a documented MAP improvement, which is a limited surrogate for a change in cardiac output: changes in MAP did not reliably track cardiac-index responses in the septic-shock study of Pierrakos et al. (2012). Circulatory deterioration is more directly related to available hemodynamic signals, consistent with its relatively stronger discrimination, while AKI classification uses serum creatinine or urine-output criteria (KDIGO Acute Kidney Injury Work Group, 2012), whose trajectories are excluded from the waveform-only input. Sepsis identification also depends on the operational definition of infection and organ dysfunction; different retrospective definitions produce different cohorts and outcome incidences (Johnson et al., 2018).

VitalDB: observable states, future transitions, and external influences. Hypotension is directly related to arterial-pressure dynamics, whereas predicting accelerating MAP decline requires anticipating a trajectory change rather than identifying an already low pressure. Oxygen-desaturation prediction can draw on preoperative patient characteristics and intraoperative measurements, including ventilation parameters (Lundberg et al., 2018). Core hypothermia is influenced by anesthetic effects, redistribution of body heat, and warming (Sessler, 2016). A short waveform input may not fully represent these influences. The blood-gas–end-tidal CO_2 discordance endpoint also requires

a qualifying future blood-gas measurement; capnography is informative about ventilation and perfusion but has important limits as a surrogate for arterial CO_2 (Nassar & Schmidt, 2016). Abrupt $ETCO_2$ falls may combine cardiopulmonary changes with equipment or ventilation changes, whose antecedents need not be present in the waveform history. Coupled hemodynamic responses can depend on subsequent stimulation or treatment, and their HR/MAP definition does not identify nociception specifically (Ledowski, 2019). The BIS probe measures a concurrent state and is reported separately from future-event forecasting.

MC-MED: restricted cohorts and laboratory information. Decompensation is evaluated among initially stable patients and combines several future vital-sign abnormalities into one endpoint. The related ED study of Sundrani et al. (2023) predicts the abnormalities separately over 90 minutes and combines monitoring features with triage information, so its AUROCs are not directly comparable with our 60-minute waveform-only composite. The pending hyperkalemia and anemia tasks estimate the values of tests ordered before T; a result arriving after T need not represent a newly developing physiological abnormality. Large ECG screening studies also differ in scale and acquisition: Galloway et al. (2019) trained on approximately 1.6 million ECGs and evaluated two- and four-lead inputs, while Kwon et al. (2020) studied anemia screening using standard ECG records. Our corresponding training sets contain 360 and 180 positive windows, respectively (Table 6). Lactate nonclearance depends on metabolic and treatment trajectories over hours (Vincent et al., 2016); its training cohort contains only 518 patients and its validation set 85 windows.

H Recommendations for Waveform Model Development

Architecture shortlists refer to the selected encoder–tokenizer configurations in Table 10(a); tokenizer shortlists refer to its fixed-Transformer comparisons in Table 10(b). History and input choices come from separate sweeps; their preferred settings have not been jointly evaluated. Recommendations concern waveform models and their augmentation; EHR-only models serve as controls for the additional value of waveform information.

H.1 ICU Tasks: MIMIC-III

Sepsis-3 and acute deterioration. With a fixed Transformer, SeFT is the first alternative to Conv1D for Sepsis-3: it raises AUROC from 0.630 to 0.669, and its gain recurs in eight of nine encoders in the interaction study. For acute deterioration, the corresponding shortlist is FFT and STraTS, with AUROCs of 0.620 and 0.607 versus Conv1D's 0.586. If one tokenizer must serve the ICU battery, Conv1D is a practical shared baseline (Appendix C.3).

Circulatory deterioration and AKI stage 2/3. For waveform-only encoder searches, TCN and InceptionTime are strong circulatory-deterioration candidates, while Transformer and Mamba-3 lead the AKI comparison. With a fixed Transformer, Conv1D and mTAND form a useful shortlist for both endpoints; FFT is less attractive for AKI, reducing AUROC from 0.730 to 0.677. Under a training-time or memory constraint, Conv1D is the better initial choice: the MIMIC-III tokenizer sweep records approximately tenfold higher training time and fivefold higher peak memory for mTAND, with only a small task-AUROC advantage on these two endpoints (Appendix C.4). The AKI gap sweep peaks at six hours for AUROC but at three hours for AUPRC; use the required clinical warning interval to fix the comparison, because outcome timing and eligibility also change (Appendices D and D.1). With EHR access, bilinear fusion is a useful waveform-augmentation candidate for both tasks, but only circulatory deterioration exceeds the EHR-only control in this comparison (Appendix E.2).

Dynamic mortality. With waveform-only inputs, BiLSTM and InceptionTime are suitable architecture candidates, while mTAND and iTransformer lead the fixed-Transformer tokenizer comparison. Ten minutes is the preferred initial history among the tested views: extending to 20 or 60 minutes does not improve mortality AUROC, which remains near 0.517. For missing-input handling, GRU-D improves mortality AUROC by 0.0195 over zero-mask plus Δ; adding EHR through the tested fusion methods yields at most 0.557, below the EHR-only control of 0.590 (Appendix E.1).

Respiratory escalation and ventilator liberation. For respiratory escalation, the selected BiLSTM, ResNet1D, and Transformer configurations form a close architecture shortlist; when the Transformer is fixed, iTransformer and Crossformer are the leading tokenizer alternatives. For venti-

lator liberation, BiLSTM, Mamba-3, and Transformer are reasonable architecture candidates, while FFT and STraTS lead the tokenizer comparison by small absolute margins. For respiratory escalation, the 60-minute view improves AUROC and AUPRC, while the 20-minute view has lower ECE and AURC and is worth retaining when confidence quality guides clinical review (Appendices D and D.1).

Documented fluid response. Under a fixed Transformer, Conv1D is the preferred tokenizer, while FITS is a useful candidate when the architecture can also change. For a broader encoder–tokenizer search, include Crossformer: it improves fluid-response AUROC in eight of nine encoders, despite lowering the fixed-Transformer score. This interaction makes fluid response a case for joint encoder–tokenizer comparison, rather than selecting the tokenizer from its performance on one backbone (Appendix C.3).

H.2 Intraoperative Tasks: VitalDB

Hypotension. With waveform-only inputs, Conv1D is the strongest fixed-Transformer tokenizer; Transformer, S4, and FITS provide a closely ranked architecture shortlist. When more pre-cutoff history is available, 30- and 60-minute inputs are useful candidates because longer history improves both hypotension AUROC and AUPRC, while increasing the forecast gap weakens discrimination (Appendices D and D.1). If EHR inputs are available, compare concatenation, weighted late, and gated fusion with both unimodal references: these methods improve AUROC, but weighted late fusion has lower hypotension AUPRC than waveform-only prediction. For a battery-level AUPRC and calibration objective, weighted late fusion is preferable to gated fusion when their AUPRCs are similar, because its ECE is lower (0.0251 versus 0.0306; Appendix E.6).

Gas discordance and accelerating MAP decline. With a fixed Transformer, FFT leads the gas-discordance tokenizer comparison, with SeFT as the next candidate; MAP decline favors mTAND and Conv1D. When inputs are incomplete, these endpoints justify targeted comparisons of missing-input methods: eight of nine methods improve gas discordance and all nine improve MAP decline, without establishing one common best imputer. For gas discordance with EHR access, weighted late fusion is a candidate for augmenting the waveform model, evaluated against EHR-only to measure the waveform branch's additional value (Appendices E.1 and F.3).

Oxygen desaturation and coupled hemodynamic response. For fixed-Transformer development, FFT and Conv1D are the leading desaturation tokenizers, whereas STraTS and SeFT lead coupled response. Under natural channel missingness, native handling leads desaturation; for coupled response, Raindrop is the only added method that improves AUROC, by 0.0067 (Appendix E.1). If desaturation AUPRC is the objective, keep short histories in the comparison, because its AUPRC falls as history increases even while the battery macro improves (Appendix D.1). For longer warning intervals, reevaluate coupled response at the required gap, since its rebound accompanies a shifted outcome window (Appendix D).

Hypothermia and abrupt $ETCO_2$ fall. Under an architecture search, DLinear is a strong hypothermia candidate and BiLSTM leads abrupt $ETCO_2$ fall. With the Transformer fixed, compare FFT and Conv1D for hypothermia, and STraTS, FFT, and Conv1D for $ETCO_2$ fall. For hypothermia, retain the five-minute history because longer views do not improve AUROC in the reported sweep (Appendix D).

Concurrent light-BIS state. For waveform-only state estimation, Transformer/Conv1D with the five-minute view is a strong starting configuration. Under a restricted subject budget and an AUROC objective, 10–20% gradient-matched subsets nearly preserve the full-data score; if AUPRC is the objective, the full cohort remains preferable because both subsets lose precision–recall performance (Appendix E.3). The separate fusion validation study finds little concurrent-state AUROC gain over EHR-only, so its results do not favor adding fusion complexity to improve the waveform branch's contribution (Appendix F.3).

H.3 Emergency-Department Tasks: MC-MED

Physiological decompensation. For waveform-only development, BiLSTM and InceptionTime are the leading architecture candidates (0.649 and 0.648), while Conv1D and Conv2D are nearly tied with a fixed Transformer (both 0.647 at the displayed precision). FFT is unfavorable for this end-

point, lowering AUROC in all nine encoders in the interaction study. For an AUROC objective, ten minutes is preferable to 15 minutes for decompensation (0.656 versus 0.631) and also gives the highest four-task macro. For an AUPRC objective, retain the 15-minute waveform view, which instead improves this task from 0.2763 to 0.3166 (Appendix D.1). If EHR is available and AUPRC is the objective, start with late fusion because it leads decompensation. Its worse AURC makes co-attention worth retaining when selective prediction is the objective (Appendix E.6).

Pending hyperkalemia and severe anemia. For waveform-only hyperkalemia, BiLSTM and InceptionTime form the architecture shortlist; Conv1D, Conv2D, and Crossformer lead the fixed-Transformer comparison, while FFT lowers AUROC across all nine encoders. For anemia, compare DLinear, Mamba-3, and FITS when choosing the encoder, and Crossformer with Conv1D when the Transformer is fixed; Crossformer reaches 0.582 versus Conv1D's 0.563. If longer recordings are available, ten minutes is the stronger history for both hyperkalemia and anemia; extending to 15 minutes reduces their AUROCs by 0.048 and 0.046, respectively, and also lowers both AUPRCs (Appendices D and D.1). For waveform augmentation, gated and late fusion are the leading anemia-AUROC candidates; the companion study's higher EHR-only AUPRC for both laboratory tasks makes that control necessary when assessing the added waveform contribution (Appendices E.2 and E.6).

Lactate nonclearance. Transformer and ResNet1D lead the architecture comparison (0.508 and 0.499); with the Transformer fixed, mTAND, Conv1D, and Crossformer score 0.509, 0.508, and 0.507. These configurations form a near-chance shortlist; FFT's small mean gain across encoders does not translate into the best fixed-Transformer result. The 15-minute view is a useful candidate when history can be extended, because both AUROC and AUPRC improve for this endpoint in the duration sweep (Appendix D.1). With EHR access, late fusion is the only tested fusion method to improve lactate AUROC over waveform-only, while gated fusion leads task AUPRC in the companion study; retain EHR-only as a control because the AUROC improvement does not exceed it (Appendices E.2 and E.6).

H.4 Shared Resource Constraints and Transfer

Reduced sensors and training data. For MIMIC-III under a sensor budget, clinical-priority and SQI selection are stronger initial choices than random or generic statistical selection; clinical selection also reduces memory, at the cost of worse calibration and selective risk. For VitalDB, retain all available channels when possible, and use temporal MI as the first reduced-channel comparator. Under a MIMIC-III window budget, random selection at 30% and NLDR at 20% are useful starting points; for subject reduction, gradient matching or noise filtering at 20% of VitalDB subjects and training dynamics or diversity at 50% of MC-MED subjects form the shortlists. Full training sets remain stronger for MIMIC-III macro and VitalDB forecast-macro AUROC (Appendix E.3). If the operating condition includes additional sensor loss, evaluate that condition separately: natural-missingness performance does not predict the observed AUROC decline under 50% sensor dropout (Appendix E.1).

Adapting to MIMIC-IV. In the reported comparison of circulatory deterioration with MIMIC-III source weights, head-only adaptation outperforms both random initialization and EHR-only, making it a strong initial choice. For AKI, include partial adaptation when prioritizing AUROC and full fine-tuning when prioritizing AUPRC; with EHR access, retain EHR-only and combined-feature controls for both AKI and delirium, where they account for much of the gain. For acute-deterioration and future-SOFA targets, compare head-only waveform adaptation with matched random initialization, using EHR-only to assess whether the waveform models add information beyond clinical context. Under VitalDB initialization, retain the matched randomly initialized model as a primary candidate because source pretraining gives no consistent macro advantage in this adaptation study (Appendix E.5).

I Numerical Tables and Figure Data

This section collects the cohort, protocol, and numerical result tables. Updated summaries retain the precision supplied in the latest numerical exports; task-level tables retain the data used to draw the corresponding figures. Earlier companion experiments are identified separately, so values from

Table 3: Short labels and full names of the 20 clinical tasks.

Short label in results	Full clinical task name
MIMIC-III	
Acute det. / Acute deterior.	Acute deterioration / failure to rescue
Dyn. mort. / Dyn. mortality	Dynamic in-hospital mortality
Resp. esc. / Resp. escalation	Respiratory deterioration and escalation
Circ. det. / Circ. deterior.	Circulatory deterioration and shock
Fluid resp. / Fluid response	Rapid documented-fluid response (MAP-response proxy)
AKI 2/3 / AKI KDIGO 2/3	Incident Kidney Disease: Improving Global Outcomes (KDIGO) stage 2/3 acute kidney injury
Vent. lib. / Vent. liberation	Ventilator course and liberation
Sepsis-3	Sepsis-3 onset and shock trajectory
VitalDB Core-8	
OX01-IOH	Sustained intraoperative hypotension
CL07-002	Accelerating mean arterial pressure (MAP) decline
H5-002	Incident oxygen desaturation
H5-030	Incident core hypothermia
L2-30	Blood-gas PCO_2–end-tidal CO_2 discordance
VA1-CO2 / VA1-CO2-DROP	Abrupt end-tidal CO_2 ($ETCO_2$) fall
VA1-STIM / VA1-STIM-RESPONSE	Coupled hemodynamic response
VCS05	Concurrent light bispectral-index (BIS) state
MC-MED	
Decomp. 60m	Physiological decompensation within 60 minutes
Pending hyperK	Pending hyperkalemia
Lactate NC	Lactate nonclearance
Pending anemia	Pending severe anemia

different model searches or scoring estimators are not combined. A dash denotes an unavailable measurement or an unevaluated combination, not zero. Metric definitions and uncertainty conventions follow Appendix A.2.

These tables give the task markers and shared telemetry from each EES figure. Forecast macros exclude concurrent BIS; a condition-level telemetry value is reported once as shared telemetry. Earlier representation and information-availability companions retain their original measurements, whereas the temporal tables use the refreshed collection.

Table 4: Conservative model-fit counts implied by the reported grids, including specified learning-rate searches, seeds, and folds, with shared conditions counted once.

Comparison	MIMIC-III	VitalDB	MC-MED	Total
Representation	4,240	34,875	2,460	41,575
Additional histories	120	5,625	40	5,785
Additional forecast gaps	–	700	–	700
Channel selection	240	16,875	120	17,235
Missing-input handling	360	10,125	180	10,665
Subject selection	840	7,875	420	9,135
Fusion and EHR-only models	360	2,250	120	2,730
Total	6,160	78,325	3,340	87,825

Table 5: Official VitalDB waveform streams used by the waveform-only benchmark. Canonical IDs denote positions in the fixed 66-slot representation; they are not a count of observed channels. Native rates are nominal, while preprocessing uses the rate recorded in each source file.

ID	Canonical name	VitalDB source track	Nominal Hz
0	ECG-II	`SNUADC/ECG_II`	500
59	ECG-V5	`SNUADC/ECG_V5`	500
2	PLETH	`SNUADC/PLETH`	500
13	ART	`SNUADC/ART`	500
8	CVP	`SNUADC/CVP`	500
42	Femoral AP	`SNUADC/FEM`	500
28	CO2	`Primus/CO2`	62.5
62	Airway pressure	`Primus/AWP`	62.5
63	EEG1	`BIS/EEG1_WAV`	128
64	EEG2	`BIS/EEG2_WAV`	128
4	CardioQ ABP	`CardioQ/ABP`	180
65	CardioQ flow	`CardioQ/FLOW`	180

Table 6: Task-level cohort support across development and test partitions. MIMIC-III and MC-MED cells give unique patients / clinical episodes / task-labelled windows; episodes are ICU stays and ED visits, respectively. VitalDB cells give patients / windows because surgical-case counts were not supplied. Parentheses give positive-window counts where reported.

MIMIC-III: patient-disjoint partitions

Task	Train	Validation	Test
Acute deterioration	5,151 / 5,761 / 116,929	633 / 703 / 14,399	644 / 739 / 15,257
Dynamic mortality	5,327 / 5,953 / 130,504	651 / 722 / 15,215	894 / 1,052 / 81,359
Respiratory escalation	4,393 / 4,863 / 127,005	556 / 598 / 15,509	547 / 622 / 16,966
Circulatory deterioration	4,788 / 5,320 / 99,417	595 / 649 / 12,031	877 / 1,020 / 64,619
Fluid response	1,740 / 1,839 / 5,787	209 / 217 / 696	247 / 256 / 854
AKI KDIGO 2/3	3,844 / 4,126 / 130,944	466 / 490 / 16,117	485 / 514 / 15,288
Ventilator liberation	2,064 / 2,188 / 9,302	239 / 252 / 1,120	246 / 262 / 1,239
Sepsis-3	3,896 / 4,190 / 51,482	480 / 510 / 6,635	729 / 816 / 31,246
Union / task totals	6,787 / 7,753 / 671,370	851 / 964 / 81,722	897 / 1,055 / 226,828

VitalDB Core-8: development uses five outer whole-subject folds

Task	Development	Test
Intraoperative hypotension	2,443 / 27,855 (1,252)	355 / 3,442 (189)
Accelerating MAP decline	2,447 / 28,903 (918)	365 / 3,927 (123)
Oxygen desaturation	4,175 / 61,293 (849)	608 / 8,395 (121)
Core hypothermia	3,437 / 40,665 (2,164)	514 / 5,736 (314)
PCO_2–$ETCO_2$ discordance	1,297 / 9,338 (782)	169 / 1,078 (65)
Abrupt $ETCO_2$ fall	3,580 / 32,602 (1,629)	500 / 4,409 (204)
Coupled hemodynamic response	2,238 / 23,265 (634)	333 / 3,240 (84)
Concurrent light BIS state	4,221 / 112,425 (5,231)	705 / 19,094 (890)
Union / task totals	4,596 / 336,346 (13,459)	750 / 49,321 (1,990)

MC-MED: episodes are ED visits; one window per visit and task

Task	Train	Validation	Test
Cardiorespiratory decompensation (60 min)	7,875 / 9,016 / 9,016 (1,106)	1,321 / 1,390 / 1,390 (167)	2,100 / 2,288 / 2,288 (289)
Pending hyperkalemia	8,171 / 9,422 / 9,422 (360)	1,440 / 1,529 / 1,529 (48)	2,139 / 2,364 / 2,364 (90)
Lactate non-clearance	518 / 540 / 540 (145)	84 / 85 / 85 (23)	187 / 196 / 196 (70)
Pending severe anemia	4,715 / 5,130 / 5,130 (180)	896 / 929 / 929 (33)	1,388 / 1,490 / 1,490 (54)
Union / task totals	11,051 / 13,216 / 24,108 (1,791)	1,891 / 2,044 / 3,933 (271)	– / – / 6,338 (503)

Union patients and clinical episodes are deduplicated across tasks within each partition; windows and positive windows are summed over task-labelled rows. The same stay or visit can therefore contribute to several tasks. VitalDB development is counted once, not once as training and again as validation. Its forecast-window counts include the reported 5- and 15-minute gap heads; concurrent BIS is a separate task. – denotes an unreported count, not zero. In particular, MC-MED test union patients and visits cannot be recovered by summing the overlapping task cohorts.

Table 7: MIMIC-IV transfer targets, full task names, and support in the frozen manifests. Positive counts refer to patients; windows precede patient-level aggregation.

Short label	Full target name	Patients	Positive	Windows
Acute det.	Acute deterioration	87	30	2,506
AKI 2/3	Acute kidney injury	46	19	932
Circ. det.	Circulatory deterioration	139	26	5,332
Delirium / neuro	Delirium / neurological outcome	91	22	1,778
Future SOFA	Future Sequential Organ Failure Assessment outcome	147	25	6,823

Table 8: Temporal grids displayed in Figure 4. All panels report test AUROC. The warning grids are listed by endpoint; separate VitalDB validation results appear in Appendix F.2.

Dataset	Observation length (min)	Task-specific forecast gaps
MIMIC-III	10, 20, 60	Mortality: 3, 12, 24 h; AKI: 3, 6, 12 h; Sepsis-3: 2, 4, 6 h.
VitalDB	5, 15, 30, 60	All seven forecasts: 5, 15, 30, 45, 60, 75 min.
MC-MED	5, 10, 15	Decompensation: 0, 10, 30 min; hyperkalemia and anemia: 0, 15, 60 min; lactate: 30, 60, 90 min.

Table 9: References and comparison grids in Figure 5. Positive differences favor the displayed condition over its panel-specific reference.

Panel	Reference	Displayed comparison
(a) Sensors	All permitted channels	MIMIC-III: 10 of 62 channels; VitalDB: eight channels. Policies are Random, Clinical, MI, Lasso, SQI, and temporal MI; Clinical is not displayed for VitalDB.
(b) Missing inputs	MIMIC-III: zero-mask plus Δ; VitalDB: architecture only	Eight MIMIC-III alternatives: GRU-D, GRIN, Raindrop, positional kriging, lead-agnostic SSL, Rectified Flow, TrigFlow, and Drifting. VitalDB also tests zero-mask plus Δ, giving nine alternatives.
(c) Fusion	Waveform only	Nine fusion methods on MIMIC-III, five on VitalDB, and four on MC-MED (concatenation, gated, late, and co-attention); all three include EHR-only controls.
(d) Training data	Full training set	Eight selectors: random, diversity, diversity clustering, gradient matching, NLDR, noise filtering, task anchors, and training dynamics. Budgets are 10/20/30% of windows (MIMIC-III), 5/10/20% of subjects (VitalDB), and 10/20/50% of subjects (MC-MED).
(e) Transfer	Source-matched random initialization	MIMIC-III or VitalDB Transformer/Conv1D to MIMIC-IV: full fine-tuning, last block plus head, head only, and three logistic-regression controls (EHR, frozen waveform features, or both).

Table 10: Updated test AUROC for Figure 3(a,b), as mean ± sample SD over five seeds: (a) nine approximately 500k-parameter encoder–frontend configurations; (b) a Transformer with seven frontends on MIMIC-III and VitalDB and eight on MC-MED. Bold and underlining mark the highest and next-highest unrounded means. Dashes indicate unevaluated combinations. Search rules are in Appendix B.7.

(a) Network architectures (selected frontends)		Transformer	BiLSTM	Mamba-3	InceptionTime	TCN	ResNet1D	S4	FITS	DLinear
	Acute det.	0.587 ± 0.008	0.584 ± 0.008	0.588 ± 0.012	0.589 ± 0.004	0.595 ± 0.016	0.595 ± 0.016	0.580 ± 0.009	0.579 ± 0.011	0.588 ± 0.005
	Dyn. mort.	0.532 ± 0.026	0.558 ± 0.005	0.531 ± 0.007	0.554 ± 0.032	0.547 ± 0.030	0.527 ± 0.058	0.549 ± 0.011	0.525 ± 0.015	0.540 ± 0.013
	Resp. esc.	0.575 ± 0.004	0.576 ± 0.010	0.561 ± 0.031	0.560 ± 0.014	0.558 ± 0.012	0.576 ± 0.006	0.561 ± 0.005	0.553 ± 0.008	0.552 ± 0.011
	Circ. det.	0.678 ± 0.007	0.673 ± 0.008	0.672 ± 0.003	0.683 ± 0.009	0.685 ± 0.011	0.669 ± 0.020	0.671 ± 0.008	0.671 ± 0.014	0.675 ± 0.004
MIMIC-III	Fluid resp.	0.575 ± 0.004	0.570 ± 0.018	0.571 ± 0.021	0.561 ± 0.018	0.571 ± 0.026	0.556 ± 0.015	0.547 ± 0.010	0.584 ± 0.010	0.562 ± 0.023
	AKI 2/3	0.727 ± 0.012	0.696 ± 0.047	0.723 ± 0.007	0.690 ± 0.035	0.658 ± 0.018	0.688 ± 0.039	0.682 ± 0.048	0.685 ± 0.051	0.664 ± 0.051
	Vent. lib.	0.550 ± 0.007	0.554 ± 0.005	0.551 ± 0.008	0.541 ± 0.004	0.544 ± 0.008	0.549 ± 0.009	0.542 ± 0.003	0.541 ± 0.004	0.550 ± 0.003
	Sepsis-3	0.630 ± 0.005	0.638 ± 0.006	0.630 ± 0.011	0.640 ± 0.004	0.638 ± 0.006	0.632 ± 0.011	0.643 ± 0.006	0.632 ± 0.005	0.633 ± 0.005
	Macro (8)	0.607 ± 0.003	0.606 ± 0.006	0.603 ± 0.005	0.602 ± 0.006	0.600 ± 0.006	0.599 ± 0.006	0.597 ± 0.006	0.596 ± 0.009	0.595 ± 0.003
	OX01-IOH	0.695 ± 0.003	0.689 ± 0.005	0.686 ± 0.003	0.688 ± 0.004	0.686 ± 0.003	0.689 ± 0.003	0.696 ± 0.004	0.696 ± 0.002	0.690 ± 0.003
	CL07-002	0.547 ± 0.003	0.541 ± 0.002	0.539 ± 0.009	0.525 ± 0.007	0.529 ± 0.009	0.534 ± 0.006	0.538 ± 0.007	0.539 ± 0.005	0.547 ± 0.001
	H5-002	0.647 ± 0.007	0.638 ± 0.004	0.638 ± 0.006	0.628 ± 0.005	0.640 ± 0.006	0.634 ± 0.003	0.646 ± 0.009	0.643 ± 0.008	0.636 ± 0.007
	H5-030	0.605 ± 0.004	0.601 ± 0.003	0.603 ± 0.001	0.580 ± 0.007	0.580 ± 0.003	0.592 ± 0.004	0.606 ± 0.004	0.606 ± 0.002	0.608 ± 0.001
VitalDB	L2-30	0.559 ± 0.016	0.548 ± 0.009	0.550 ± 0.010	0.553 ± 0.010	0.549 ± 0.006	0.548 ± 0.007	0.555 ± 0.008	0.548 ± 0.009	0.561 ± 0.014
	VA1-CO2	0.675 ± 0.002	0.683 ± 0.004	0.671 ± 0.005	0.671 ± 0.007	0.678 ± 0.005	0.679 ± 0.002	0.676 ± 0.003	0.678 ± 0.003	0.677 ± 0.007
	VA1-STIM	0.687 ± 0.006	0.688 ± 0.004	0.687 ± 0.005	0.680 ± 0.005	0.675 ± 0.009	0.682 ± 0.002	0.689 ± 0.002	0.683 ± 0.003	0.686 ± 0.002
	VCS05	0.885 ± 0.001	0.879 ± 0.002	0.882 ± 0.002	0.880 ± 0.003	0.880 ± 0.002	0.881 ± 0.002	0.877 ± 0.002	0.881 ± 0.001	0.884 ± 0.002
	Macro (7)	0.631 ± 0.004	0.627 ± 0.001	0.625 ± 0.002	0.618 ± 0.003	0.620 ± 0.003	0.623 ± 0.001	0.629 ± 0.002	0.627 ± 0.001	0.629 ± 0.003
	Decomp. 60m	0.642 ± 0.004	0.649 ± 0.006	0.641 ± 0.005	0.648 ± 0.007	0.646 ± 0.006	0.640 ± 0.010	0.645 ± 0.003	0.634 ± 0.016	0.642 ± 0.003
	Pending hyperK	0.555 ± 0.029	0.580 ± 0.023	0.541 ± 0.054	0.575 ± 0.046	0.565 ± 0.029	0.566 ± 0.028	0.517 ± 0.020	0.562 ± 0.020	0.565 ± 0.011
MC-MED	Lactate NC	0.508 ± 0.021	0.473 ± 0.018	0.498 ± 0.021	0.484 ± 0.019	0.471 ± 0.025	0.499 ± 0.019	0.453 ± 0.009	0.486 ± 0.020	0.487 ± 0.015
	Pending anemia	0.563 ± 0.031	0.562 ± 0.013	0.573 ± 0.012	0.560 ± 0.024	0.559 ± 0.028	0.567 ± 0.014	0.544 ± 0.013	0.571 ± 0.009	0.574 ± 0.019
	Macro (4)	0.567 ± 0.004	0.566 ± 0.006	0.563 ± 0.014	0.567 ± 0.011	0.560 ± 0.014	0.568 ± 0.013	0.540 ± 0.005	0.563 ± 0.008	0.567 ± 0.004

(b) Transformer backbone with different frontends		Conv1D	Conv2D	SeFT	Crossformer	STraTS	iTransformer	FFT	mTAND
	Acute det.	0.586 ± 0.005	–	0.598 ± 0.005	0.596 ± 0.013	0.607 ± 0.010	0.567 ± 0.008	0.620 ± 0.008	0.583 ± 0.003
	Dyn. mort.	0.549 ± 0.029	–	0.563 ± 0.019	0.561 ± 0.008	0.551 ± 0.013	0.566 ± 0.031	0.558 ± 0.010	0.576 ± 0.008
	Resp. esc.	0.569 ± 0.005	–	0.556 ± 0.007	0.575 ± 0.025	0.525 ± 0.041	0.585 ± 0.017	0.563 ± 0.003	0.552 ± 0.044
	Circ. det.	0.689 ± 0.009	–	0.682 ± 0.004	0.688 ± 0.009	0.685 ± 0.008	0.682 ± 0.005	0.685 ± 0.002	0.690 ± 0.005
MIMIC-III	Fluid resp.	0.586 ± 0.010	–	0.574 ± 0.016	0.559 ± 0.028	0.575 ± 0.005	0.563 ± 0.039	0.561 ± 0.014	0.551 ± 0.028
	AKI 2/3	0.730 ± 0.005	–	0.727 ± 0.000	0.727 ± 0.001	0.727 ± 0.001	0.727 ± 0.008	0.677 ± 0.020	0.736 ± 0.005
	Vent. lib.	0.550 ± 0.007	–	0.549 ± 0.024	0.543 ± 0.004	0.554 ± 0.007	0.537 ± 0.007	0.556 ± 0.009	0.550 ± 0.010
	Sepsis-3	0.630 ± 0.010	–	0.669 ± 0.005	0.641 ± 0.011	0.650 ± 0.006	0.640 ± 0.009	0.641 ± 0.014	0.618 ± 0.003
	Macro (8)	0.611 ± 0.004	–	0.615 ± 0.005	0.611 ± 0.004	0.609 ± 0.007	0.608 ± 0.005	0.608 ± 0.007	0.607 ± 0.007
	OX01-IOH	0.695 ± 0.003	–	0.668 ± 0.005	0.635 ± 0.009	0.681 ± 0.002	0.596 ± 0.008	0.578 ± 0.009	0.678 ± 0.003
	CL07-002	0.547 ± 0.003	–	0.504 ± 0.006	0.522 ± 0.006	0.521 ± 0.007	0.544 ± 0.006	0.516 ± 0.008	0.552 ± 0.005
	H5-002	0.647 ± 0.007	–	0.586 ± 0.008	0.574 ± 0.008	0.617 ± 0.004	0.532 ± 0.010	0.657 ± 0.005	0.535 ± 0.005
	H5-030	0.605 ± 0.004	–	0.598 ± 0.005	0.560 ± 0.005	0.603 ± 0.002	0.536 ± 0.006	0.610 ± 0.006	0.568 ± 0.003
VitalDB	L2-30	0.559 ± 0.016	–	0.598 ± 0.012	0.574 ± 0.015	0.583 ± 0.010	0.521 ± 0.003	0.674 ± 0.010	0.564 ± 0.012
	VA1-CO2	0.675 ± 0.002	–	0.639 ± 0.006	0.585 ± 0.009	0.680 ± 0.001	0.585 ± 0.007	0.678 ± 0.005	0.613 ± 0.004
	VA1-STIM	0.687 ± 0.006	–	0.691 ± 0.004	0.628 ± 0.006	0.702 ± 0.006	0.602 ± 0.017	0.663 ± 0.006	0.579 ± 0.006
	VCS05	0.885 ± 0.001	–	0.838 ± 0.002	0.866 ± 0.001	0.843 ± 0.002	0.853 ± 0.001	0.871 ± 0.003	0.860 ± 0.004
	Macro (7)	0.631 ± 0.004	–	0.612 ± 0.003	0.583 ± 0.002	0.627 ± 0.003	0.559 ± 0.003	0.625 ± 0.003	0.584 ± 0.003
	Decomp. 60m	0.647 ± 0.004	0.647 ± 0.004	0.469 ± 0.005	0.620 ± 0.015	0.563 ± 0.019	0.512 ± 0.010	0.560 ± 0.015	0.516 ± 0.015
	Pending hyperK	0.576 ± 0.021	0.576 ± 0.021	0.428 ± 0.036	0.573 ± 0.011	0.446 ± 0.046	0.500 ± 0.027	0.454 ± 0.024	0.474 ± 0.040
MC-MED	Lactate NC	0.508 ± 0.021	0.460 ± 0.009	0.499 ± 0.023	0.507 ± 0.020	0.499 ± 0.028	0.503 ± 0.028	0.496 ± 0.013	0.509 ± 0.058
	Pending anemia	0.563 ± 0.031	0.556 ± 0.015	0.533 ± 0.016	0.582 ± 0.020	0.530 ± 0.010	0.514 ± 0.018	0.525 ± 0.009	0.493 ± 0.014
	Macro (4)	0.574 ± 0.006	0.560 ± 0.006	0.482 ± 0.012	0.570 ± 0.010	0.509 ± 0.015	0.507 ± 0.011	0.509 ± 0.005	0.498 ± 0.022

Table 11: Tokenizer–encoder contrasts transcribed from Figure 3(c): Crossformer and SeFT on MIMIC-III and VitalDB, and Conv2D and FFT on MC-MED. Entries give mean AUROC differences relative to Conv1D; agreement is out of nine encoders and follows the unrounded mean. Concurrent VCS05 is excluded from the VitalDB forecast macro.

MIMIC-III				
	Crossformer		SeFT	
Task	ΔAUROC	Agreement	ΔAUROC	Agreement
Acute deterioration	−0.009	9/9	+0.005	8/9
Dynamic mortality	−0.009	6/9	−0.015	8/9
Respiratory escalation	−0.002	5/9	−0.006	7/9
Circulatory deterioration	−0.015	8/9	−0.010	8/9
Fluid response	+0.014	8/9	+0.001	5/9
AKI KDIGO stage 2/3	+0.006	7/9	+0.008	6/9
Ventilator liberation	−0.010	6/9	−0.008	6/9
Sepsis-3	−0.003	7/9	+0.014	8/9
Macro (8)	−0.003	8/9	−0.001	5/9
VitalDB				
OX01-IOH	−0.038	9/9	−0.023	9/9
CL07-002	−0.007	6/9	−0.032	9/9
H5-002	−0.030	9/9	−0.038	9/9
H5-030	−0.019	7/9	−0.003	6/9
L2-30	+0.032	9/9	+0.039	9/9
VA1-CO2	−0.037	8/9	−0.018	9/9
VA1-STIM	−0.009	4/9	+0.008	9/9
Macro (7)	−0.015	7/9	−0.010	9/9
VCS05 (concurrent)	−0.006	9/9	−0.044	9/9

MC-MED: contrasts versus Conv1D				
	Conv2D		FFT	
Task	ΔAUROC	Agreement	ΔAUROC	Agreement
Decompensation (60 min)	+0.000	1/9	−0.108	9/9
Pending hyperkalemia	+0.001	3/9	−0.076	9/9
Lactate nonclearance	−0.020	6/9	+0.010	6/9
Pending severe anemia	−0.012	6/9	−0.041	9/9
Macro (4)	−0.008	8/9	−0.054	9/9

Table 12: Selected test results show the scale of absolute performance. Values are means over five seeds; the VitalDB forecast macro first averages gaps within task. These are different prediction problems, not a controlled ablation of forecast gap.

Setting	Reported comparison	AUROC
MIMIC-III	Eight-task macro; Transformer with SeFT in the tokenizer comparison	0.615
VitalDB forecasts	Seven-task macro; Transformer in the architecture comparison	0.631
VitalDB concurrent BIS	Held-out-state probe; Transformer with Conv1D	0.885

Table 13: Architecture summaries — MIMIC-III. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
BiLSTM	0.656 ± 0.005	0.606 ± 0.006	0.104 ± 0.003	0.189	0.343
DLinear	0.647 ± 0.006	0.595 ± 0.003	0.101 ± 0.002	0.194	0.381
FITS	0.645 ± 0.006	0.596 ± 0.009	0.104 ± 0.004	0.199	0.341
InceptionTime	0.655 ± 0.002	0.602 ± 0.006	0.105 ± 0.002	0.151	0.370
Mamba-3	0.657 ± 0.005	0.603 ± 0.005	0.104 ± 0.001	0.209	0.379
ResNet1D	0.654 ± 0.006	0.599 ± 0.006	0.103 ± 0.004	0.165	0.354
S4	0.647 ± 0.005	0.597 ± 0.006	0.101 ± 0.004	0.149	0.357
TCN	0.652 ± 0.006	0.600 ± 0.006	0.103 ± 0.003	0.138	0.319
Transformer	0.649 ± 0.004	0.607 ± 0.003	0.106 ± 0.001	0.206	0.308

Table 14: Architecture summaries — MIMIC-III. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	ms/sample	Failures (%)
BiLSTM	117.3	3.9 %	0.36	0.0
DLinear	109.5	3.5 %	0.24	0.0
FITS	103.7	3.5 %	0.23	0.0
InceptionTime	127.7	3.4 %	0.27	0.0
Mamba-3	401.4	4.6 %	0.51	0.0
ResNet1D	101.4	3.4 %	0.23	0.0
S4	129.8	3.7 %	0.27	0.0
TCN	114.0	3.5 %	0.25	0.0
Transformer	116.4	3.5 %	0.25	0.0

Table 15: Architecture summaries — VitalDB. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
BiLSTM	0.636 ± 0.004	0.627 ± 0.001	0.075 ± 0.001	0.036 ± 0.001	0.031 ± 0.001
DLinear	0.636 ± 0.002	0.629 ± 0.003	0.076 ± 0.002	0.040 ± 0.002	0.031 ± 0.001
FITS	0.635 ± 0.002	0.627 ± 0.001	0.076 ± 0.001	0.041 ± 0.002	0.031 ± 0.001
InceptionTime	0.636 ± 0.003	0.618 ± 0.003	0.074 ± 0.002	0.041 ± 0.003	0.033 ± 0.002
Mamba-3	0.641 ± 0.002	0.625 ± 0.002	0.075 ± 0.002	0.038 ± 0.003	0.031 ± 0.001
ResNet1D	0.642 ± 0.000	0.623 ± 0.001	0.077 ± 0.001	0.043 ± 0.001	0.029 ± 0.000
S4	0.641 ± 0.003	0.629 ± 0.002	0.075 ± 0.001	0.039 ± 0.003	0.031 ± 0.001
TCN	0.638 ± 0.002	0.620 ± 0.003	0.074 ± 0.002	0.038 ± 0.001	0.033 ± 0.002
Transformer	0.640 ± 0.002	0.631 ± 0.004	0.078 ± 0.002	0.038 ± 0.002	0.029 ± 0.001

Table 16: Architecture summaries — VitalDB. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	ms/sample	Failures (%)
BiLSTM	30.1	0.39	0.0
DLinear	31.9	0.41	0.0
FITS	30.7	0.40	0.0
InceptionTime	33.5	0.43	0.0
Mamba-3	41.0	0.53	0.0
ResNet1D	30.0	0.39	0.0
S4	32.4	0.42	0.0
TCN	31.8	0.41	0.0
Transformer	32.0	0.41	0.0

Table 17: Architecture summaries — MC-MED. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
BiLSTM	0.579 ± 0.012	0.566 ± 0.006	0.179 ± 0.004	0.094 ± 0.019	0.350 ± 0.063
DLinear	0.572 ± 0.006	0.567 ± 0.004	0.179 ± 0.004	0.123 ± 0.051	0.278 ± 0.059
FITS	0.571 ± 0.010	0.563 ± 0.008	0.177 ± 0.003	0.124 ± 0.039	0.299 ± 0.087
InceptionTime	0.601 ± 0.014	0.567 ± 0.011	0.177 ± 0.008	0.171 ± 0.052	0.333 ± 0.106
Mamba-3	0.573 ± 0.009	0.563 ± 0.014	0.181 ± 0.006	0.098 ± 0.027	0.287 ± 0.084
ResNet1D	0.590 ± 0.013	0.568 ± 0.013	0.180 ± 0.004	0.080 ± 0.016	0.241 ± 0.058
S4	0.574 ± 0.005	0.540 ± 0.005	0.165 ± 0.003	0.070 ± 0.012	0.330 ± 0.041
TCN	0.602 ± 0.010	0.560 ± 0.014	0.175 ± 0.004	0.128 ± 0.029	0.265 ± 0.028
Transformer	0.579 ± 0.014	0.567 ± 0.004	0.182 ± 0.004	0.079 ± 0.016	0.296 ± 0.057

Table 18: Architecture summaries — MC-MED. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	Failures (%)
BiLSTM	17.7	33996 MiB	0.0
DLinear	21.4	33658 MiB	0.0
FITS	16.3	33300 MiB	0.0
InceptionTime	14.5	32260 MiB	0.0
Mamba-3	60.8	39008 MiB	0.0
ResNet1D	8.2	31609 MiB	0.0
S4	21.5	32066 MiB	0.0
TCN	11.9	31664 MiB	0.0
Transformer	20.7	32063 MiB	0.0

Table 19: Tokenizer summaries — MIMIC-III. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Conv1D	0.648 ± 0.003	0.611 ± 0.004	0.108 ± 0.001	–	–
Crossformer	0.639 ± 0.005	0.611 ± 0.004	0.106 ± 0.002	0.217	0.304
FFT	0.653 ± 0.001	0.608 ± 0.007	0.108 ± 0.005	0.128	0.271
iTransformer	0.634 ± 0.004	0.608 ± 0.005	0.103 ± 0.002	0.227	0.303
mTAND	0.641 ± 0.003	0.607 ± 0.007	0.107 ± 0.005	0.188	0.296
SeFT	0.649 ± 0.003	0.615 ± 0.005	0.107 ± 0.004	0.170	0.334
STraTS	0.651 ± 0.004	0.609 ± 0.007	0.107 ± 0.001	0.185	0.299

Table 20: Tokenizer summaries — MIMIC-III. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	ms/sample	Failures (%)
Conv1D	–	–	–	–
Crossformer	572.6	7.3 %	0.78	0.0
FFT	142.6	3.4 %	0.27	0.0
iTransformer	148.8	3.6 %	0.28	0.0
mTAND	1084.1	17.2 %	1.39	0.0
SeFT	173.7	4.9 %	0.31	0.0
STraTS	580.9	7.5 %	0.79	0.0

Table 21: Tokenizer summaries — VitalDB. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Crossformer	0.573 ± 0.005	0.583 ± 0.002	0.062 ± 0.002	0.038 ± 0.002	0.036 ± 0.001
FFT	0.631 ± 0.004	0.625 ± 0.003	0.076 ± 0.003	0.035 ± 0.001	0.029 ± 0.000
iTransformer	0.558 ± 0.002	0.559 ± 0.003	0.057 ± 0.001	0.034 ± 0.002	0.040 ± 0.001
mTAND	0.592 ± 0.004	0.584 ± 0.003	0.071 ± 0.001	0.045 ± 0.003	0.033 ± 0.000
Conv1D	0.640 ± 0.002	0.631 ± 0.004	0.078 ± 0.002	0.038 ± 0.002	0.029 ± 0.001
SeFT	0.602 ± 0.004	0.612 ± 0.003	0.073 ± 0.002	0.049 ± 0.002	0.029 ± 0.000
STraTS	0.627 ± 0.003	0.627 ± 0.003	0.078 ± 0.001	0.046 ± 0.003	0.028 ± 0.000

Table 22: Tokenizer summaries — VitalDB. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	ms/sample	Failures (%)
Crossformer	32.7	0.42	0.0
FFT	–	–	0.0
iTransformer	–	–	0.0
mTAND	–	–	0.0
Conv1D	32.0	0.41	0.0
SeFT	33.8	0.44	0.0
STraTS	–	–	0.0

Table 23: Tokenizer summaries — MC-MED. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Conv2D	0.572 ± 0.010	0.560 ± 0.006	0.175 ± 0.004	0.077 ± 0.025	0.327 ± 0.053
Crossformer	0.560 ± 0.009	0.570 ± 0.010	0.180 ± 0.007	0.088 ± 0.033	0.272 ± 0.031
FFT	0.532 ± 0.004	0.509 ± 0.005	0.162 ± 0.004	0.222 ± 0.059	0.300 ± 0.120
iTransformer	0.510 ± 0.007	0.507 ± 0.011	0.151 ± 0.006	0.144 ± 0.008	0.259 ± 0.049
mTAND	0.502 ± 0.014	0.498 ± 0.022	0.146 ± 0.011	0.362 ± 0.025	0.544 ± 0.200
SeFT	0.510 ± 0.012	0.482 ± 0.012	0.143 ± 0.004	0.332 ± 0.056	0.506 ± 0.223
STraTS	0.551 ± 0.013	0.509 ± 0.015	0.156 ± 0.010	0.277 ± 0.055	0.461 ± 0.134
Conv1D	0.582 ± 0.011	0.574 ± 0.006	0.183	0.083	0.301

Table 24: Tokenizer summaries — MC-MED. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	Failures (%)
Conv2D	26.9	33142 MiB	0.0
Crossformer	21.8	32390 MiB	0.0
FFT	22.9	33180 MiB	0.0
iTransformer	3.9	30141 MiB	0.0
mTAND	140.9	32674 MiB	0.0
SeFT	4.2	30526 MiB	0.0
STraTS	32.9	34699 MiB	0.0
Conv1D	–	–	–

Table 25: Crossed encoder–tokenizer grid — VitalDB. Updated export; values retain the supplied mean and SD precision. Only configurations present in the numerical export are listed.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
BiLSTM / Crossformer	0.603 ± 0.005	0.615 ± 0.003	0.071 ± 0.001	0.041 ± 0.003	0.030 ± 0.001
BiLSTM / Conv1D	0.636 ± 0.004	0.627 ± 0.001	0.075 ± 0.001	0.036 ± 0.001	0.031 ± 0.001
BiLSTM / SeFT	0.603 ± 0.002	0.610 ± 0.001	0.072 ± 0.001	0.047 ± 0.001	0.029 ± 0.000
DLinear / Crossformer	0.605 ± 0.003	0.611 ± 0.002	0.068 ± 0.001	0.045 ± 0.002	0.032 ± 0.002
DLinear / Conv1D	0.636 ± 0.002	0.629 ± 0.003	0.076 ± 0.002	0.040 ± 0.002	0.031 ± 0.001
DLinear / SeFT	0.609 ± 0.003	0.616 ± 0.001	0.076 ± 0.001	0.047 ± 0.002	0.029 ± 0.001
FITS / Crossformer	0.597 ± 0.002	0.606 ± 0.001	0.068 ± 0.001	0.040 ± 0.002	0.031 ± 0.001
FITS / Conv1D	0.635 ± 0.002	0.627 ± 0.001	0.076 ± 0.001	0.041 ± 0.002	0.031 ± 0.001
FITS / SeFT	0.610 ± 0.004	0.616 ± 0.001	0.075 ± 0.003	0.049 ± 0.001	0.029 ± 0.000
InceptionTime / Crossformer	0.614 ± 0.003	0.621 ± 0.003	0.072 ± 0.002	0.047 ± 0.003	0.032 ± 0.003
InceptionTime / Conv1D	0.636 ± 0.003	0.618 ± 0.003	0.074 ± 0.002	0.041 ± 0.003	0.033 ± 0.002
InceptionTime / SeFT	0.610 ± 0.003	0.617 ± 0.004	0.077 ± 0.001	0.049 ± 0.003	0.030 ± 0.001
Mamba-3 / sel. SSM / Crossformer	0.587 ± 0.001	0.594 ± 0.005	0.064 ± 0.003	0.040 ± 0.002	0.034 ± 0.001
Mamba-3 / sel. SSM / Conv1D	0.641 ± 0.002	0.625 ± 0.002	0.075 ± 0.002	0.038 ± 0.003	0.031 ± 0.001
Mamba-3 / sel. SSM / SeFT	0.610 ± 0.004	0.617 ± 0.001	0.075 ± 0.001	0.048 ± 0.002	0.029 ± 0.000
ResNet1D / Crossformer	0.612 ± 0.003	0.614 ± 0.002	0.070 ± 0.000	0.040 ± 0.002	0.030 ± 0.000
ResNet1D / Conv1D	0.642 ± 0.000	0.623 ± 0.001	0.077 ± 0.001	0.043 ± 0.001	0.029 ± 0.000
ResNet1D / SeFT	0.618 ± 0.004	0.619 ± 0.003	0.076 ± 0.002	0.043 ± 0.002	0.029 ± 0.001
S4 / Crossformer	0.629 ± 0.002	0.627 ± 0.003	0.079 ± 0.001	0.041 ± 0.003	0.029 ± 0.001
S4 / Conv1D	0.641 ± 0.003	0.629 ± 0.002	0.075 ± 0.001	0.039 ± 0.003	0.031 ± 0.001
S4 / SeFT	0.614 ± 0.004	0.617 ± 0.002	0.075 ± 0.001	0.047 ± 0.002	0.029 ± 0.000
TCN / Crossformer	0.607 ± 0.004	0.620 ± 0.002	0.073 ± 0.002	0.035 ± 0.002	0.029 ± 0.001
TCN / Conv1D	0.638 ± 0.002	0.620 ± 0.003	0.074 ± 0.002	0.038 ± 0.001	0.033 ± 0.002
TCN / SeFT	0.616 ± 0.002	0.620 ± 0.002	0.076 ± 0.002	0.043 ± 0.003	0.029 ± 0.000
Transformer / Crossformer	0.573 ± 0.005	0.583 ± 0.002	0.062 ± 0.002	0.038 ± 0.002	0.036 ± 0.001
Transformer / FFT	0.631 ± 0.004	0.625 ± 0.003	0.076 ± 0.003	0.035 ± 0.001	0.029 ± 0.000
Transformer / iTransformer	0.558 ± 0.002	0.559 ± 0.003	0.057 ± 0.001	0.034 ± 0.002	0.040 ± 0.001
Transformer / mTAND	0.592 ± 0.004	0.584 ± 0.003	0.071 ± 0.001	0.045 ± 0.003	0.033 ± 0.000
Transformer / Conv1D	0.640 ± 0.002	0.631 ± 0.004	0.078 ± 0.002	0.038 ± 0.002	0.029 ± 0.001
Transformer / SeFT	0.602 ± 0.004	0.612 ± 0.003	0.073 ± 0.002	0.049 ± 0.002	0.029 ± 0.000
Transformer / STraTS	0.627 ± 0.003	0.627 ± 0.003	0.078 ± 0.001	0.046 ± 0.003	0.028 ± 0.000

Table 26: Crossed encoder–tokenizer grid — MC-MED. Updated export; values retain the supplied mean and SD precision. Only configurations present in the numerical export are listed.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
BiLSTM / Conv1D	0.579 ± 0.012	0.566 ± 0.006	0.179 ± 0.004	0.094 ± 0.019	0.350 ± 0.063
BiLSTM / Conv2D	0.579 ± 0.012	0.566 ± 0.006	0.179 ± 0.004	0.094 ± 0.019	0.350 ± 0.063
BiLSTM / FFT	0.522 ± 0.003	0.495 ± 0.011	0.153 ± 0.003	0.226 ± 0.052	0.373 ± 0.151
DLinear / Conv1D	0.569 ± 0.008	0.557 ± 0.005	0.178 ± 0.004	0.087 ± 0.016	0.317 ± 0.051
DLinear / Conv2D	0.569 ± 0.008	0.557 ± 0.005	0.178 ± 0.004	0.087 ± 0.016	0.317 ± 0.051
DLinear / FFT	0.519 ± 0.008	0.512 ± 0.009	0.157 ± 0.003	0.216 ± 0.045	0.395 ± 0.097
FITS / Conv1D	0.567 ± 0.008	0.559 ± 0.003	0.175 ± 0.005	0.087 ± 0.018	0.335 ± 0.049
FITS / Conv2D	0.567 ± 0.008	0.559 ± 0.003	0.175 ± 0.005	0.087 ± 0.018	0.336 ± 0.050
FITS / FFT	0.520 ± 0.005	0.499 ± 0.006	0.155 ± 0.002	0.194 ± 0.044	0.365 ± 0.140
InceptionTime / Conv1D	0.601 ± 0.014	0.567 ± 0.011	0.177 ± 0.008	0.171 ± 0.052	0.333 ± 0.106
InceptionTime / Conv2D	0.601 ± 0.014	0.567 ± 0.011	0.177 ± 0.008	0.171 ± 0.052	0.333 ± 0.106
InceptionTime / FFT	0.524 ± 0.011	0.532 ± 0.008	0.159 ± 0.004	0.140 ± 0.025	0.307 ± 0.070
Mamba-3 / sel. SSM / Conv1D	0.571 ± 0.004	0.565 ± 0.007	0.174 ± 0.002	0.066 ± 0.021	0.329 ± 0.023
Mamba-3 / sel. SSM / Conv2D	0.571 ± 0.004	0.565 ± 0.007	0.174 ± 0.002	0.066 ± 0.021	0.329 ± 0.023
Mamba-3 / sel. SSM / FFT	0.513 ± 0.006	0.513 ± 0.020	0.155 ± 0.003	0.216 ± 0.026	0.372 ± 0.112
ResNet1D / Conv1D	0.590 ± 0.010	0.567 ± 0.011	0.181 ± 0.004	0.103 ± 0.020	0.238 ± 0.060
ResNet1D / Conv2D	0.590 ± 0.010	0.567 ± 0.011	0.181 ± 0.004	0.103 ± 0.020	0.238 ± 0.060
ResNet1D / FFT	0.528 ± 0.009	0.536 ± 0.013	0.159 ± 0.004	0.176 ± 0.035	0.332 ± 0.149
S4 / Conv1D	0.574 ± 0.005	0.540 ± 0.005	0.165 ± 0.003	0.070 ± 0.012	0.330 ± 0.041
S4 / Conv2D	0.574 ± 0.005	0.540 ± 0.005	0.165 ± 0.003	0.070 ± 0.012	0.330 ± 0.041
S4 / FFT	0.513 ± 0.008	0.500 ± 0.014	0.152 ± 0.005	0.222 ± 0.024	0.386 ± 0.096
TCN / Conv1D	0.602 ± 0.010	0.560 ± 0.014	0.175 ± 0.004	0.128 ± 0.029	0.265 ± 0.028
TCN / Conv2D	0.602 ± 0.010	0.560 ± 0.014	0.175 ± 0.004	0.128 ± 0.029	0.265 ± 0.028
TCN / FFT	0.529 ± 0.007	0.529 ± 0.011	0.156 ± 0.011	0.119 ± 0.025	0.354 ± 0.045
Transformer / Conv2D	0.572 ± 0.010	0.560 ± 0.006	0.175 ± 0.004	0.077 ± 0.025	0.327 ± 0.053
Transformer / Crossformer	0.560 ± 0.009	0.570 ± 0.010	0.180 ± 0.007	0.088 ± 0.033	0.272 ± 0.031
Transformer / FFT	0.532 ± 0.004	0.509 ± 0.005	0.162 ± 0.004	0.222 ± 0.059	0.300 ± 0.120
Transformer / iTransformer	0.510 ± 0.007	0.507 ± 0.011	0.151 ± 0.006	0.144 ± 0.008	0.259 ± 0.049
Transformer / mTAND	0.502 ± 0.014	0.498 ± 0.022	0.146 ± 0.011	0.362 ± 0.025	0.544 ± 0.200
Transformer / SeFT	0.510 ± 0.012	0.482 ± 0.012	0.143 ± 0.004	0.332 ± 0.056	0.506 ± 0.223
Transformer / STraTS	0.551 ± 0.013	0.509 ± 0.015	0.156 ± 0.010	0.277 ± 0.055	0.461 ± 0.134
Transformer / Conv1D	0.582 ± 0.011	0.574 ± 0.006	0.183	0.083	0.301

Table 27: Observation-history summaries — MIMIC-III. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
10 min	0.634 ± 0.006	0.606 ± 0.004	0.119 ± 0.004	0.164 ± 0.012	0.322 ± 0.016
20 min	0.634 ± 0.002	0.606 ± 0.002	0.118 ± 0.001	0.154 ± 0.008	0.319 ± 0.015
60 min	0.636 ± 0.002	0.609 ± 0.005	0.123 ± 0.004	0.167 ± 0.017	0.328 ± 0.017

Table 28: Observation-history summaries — MIMIC-III. Training time per optimizer step is computed for each fit before task and seed averaging; mean ± seed SD is shown where recorded update counts are available. Other units are retained per cell.

Configuration	Time/step (ms)	Memory	ms/sample	Failures (%)
10 min	–	5.2 %	0.23	0.0
20 min	–	5.2 %	0.23	0.0
60 min	–	4.6 %	0.21	0.0

Table 29: Observation-history summaries — VitalDB. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
5 min	0.632 ± 0.001	0.627 ± 0.001	0.076 ± 0.001	0.041 ± 0.001	0.030 ± 0.001
15 min	0.636 ± 0.001	0.631 ± 0.001	0.078 ± 0.000	0.039 ± 0.001	0.030 ± 0.000
30 min	0.637 ± 0.001	0.634 ± 0.001	0.081 ± 0.001	0.038 ± 0.001	0.029 ± 0.000
60 min	0.638 ± 0.001	0.636 ± 0.001	0.081 ± 0.001	0.035 ± 0.001	0.028 ± 0.000

Table 30: Observation-history summaries — VitalDB. Training time per optimizer step is computed for each fit before task and seed averaging; mean ± seed SD is shown where recorded update counts are available. Other units are retained per cell.

Configuration	Time/step (ms)	ms/sample	Failures (%)
5 min	–	0.42	0.0
15 min	–	0.42	0.0
30 min	–	0.42	0.0
60 min	–	0.42	0.0

Table 31: Observation-history summaries — MC-MED. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
5 min	0.590 ± 0.010	0.567 ± 0.011	0.181 ± 0.004	0.103 ± 0.020	0.238 ± 0.060
10 min	0.598 ± 0.016	0.578 ± 0.006	0.198 ± 0.005	0.087 ± 0.026	0.268 ± 0.051
15 min	0.571 ± 0.014	0.567 ± 0.007	0.221 ± 0.008	0.096 ± 0.021	0.266 ± 0.038

Table 32: Observation-history summaries — MC-MED. Training time per optimizer step is computed for each fit before task and seed averaging; mean ± seed SD is shown where recorded update counts are available. Other units are retained per cell.

Configuration	Time/step (ms)	Memory	Failures (%)
5 min	14.20 ± 4.92	31176 MiB	0.0
10 min	18.80 ± 6.96	43212 MiB	0.0
15 min	25.27 ± 9.48	28902 MiB	0.0

Table 33: Forecast-gap results — MIMIC-III. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Sepsis-3; 2 h	0.714 ± 0.004	0.629 ± 0.007	0.036 ± 0.001	0.311 ± 0.071	0.379 ± 0.049
AKI 2/3; 3 h	0.680 ± 0.011	0.686 ± 0.017	0.151 ± 0.010	0.169 ± 0.026	0.344 ± 0.072
Dyn. mort.; 3 h	0.595 ± 0.017	0.545 ± 0.011	0.027 ± 0.002	0.181 ± 0.033	0.259 ± 0.080
Sepsis-3; 4 h	0.710 ± 0.010	0.619 ± 0.010	0.032 ± 0.001	0.265 ± 0.043	0.307 ± 0.084
AKI 2/3; 6 h	0.685 ± 0.022	0.697 ± 0.028	0.140 ± 0.012	0.193 ± 0.039	0.313 ± 0.165
Sepsis-3; 6 h	0.686 ± 0.006	0.629 ± 0.004	0.030 ± 0.002	0.220 ± 0.035	0.266 ± 0.047
AKI 2/3; 12 h	0.666 ± 0.011	0.685 ± 0.009	0.120 ± 0.001	0.222 ± 0.066	0.292 ± 0.061
Dyn. mort.; 12 h	0.553 ± 0.019	0.539 ± 0.011	0.017 ± 0.001	0.342 ± 0.036	0.253 ± 0.172
Dyn. mort.; 24 h	0.523 ± 0.022	0.537 ± 0.009	0.018 ± 0.001	0.249 ± 0.045	0.355 ± 0.051

Table 34: Forecast-gap results — MIMIC-III. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	ms/sample	Failures (%)
Sepsis-3; 2 h	49.3	7.6 %	0.10	0.0
AKI 2/3; 3 h	294.5	3.8 %	0.24	0.0
Dyn. mort.; 3 h	323.8	3.8 %	0.25	0.0
Sepsis-3; 4 h	46.4	7.3 %	0.10	0.0
AKI 2/3; 6 h	279.8	3.8 %	0.24	0.0
Sepsis-3; 6 h	47.1	7.1 %	0.10	0.0
AKI 2/3; 12 h	250.9	3.8 %	0.25	0.0
Dyn. mort.; 12 h	314.0	3.8 %	0.24	0.0
Dyn. mort.; 24 h	252.8	3.8 %	0.20	0.0

Table 35: Forecast-gap results — VitalDB. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Macro; 5 min	0.658 ± 0.007	0.662 ± 0.004	0.100 ± 0.001	0.033 ± 0.002	0.027 ± 0.001
CL07-002; 5 min	0.589 ± 0.020	0.552 ± 0.012	0.044 ± 0.002	0.016 ± 0.005	0.029 ± 0.002
H5-002; 5 min	0.680 ± 0.009	0.651 ± 0.002	0.034 ± 0.002	0.008 ± 0.003	0.010 ± 0.000
H5-030; 5 min	0.555 ± 0.009	0.592 ± 0.007	0.072 ± 0.003	0.061 ± 0.004	0.039 ± 0.001
L2-30; 5 min	0.631 ± 0.006	0.582 ± 0.010	0.112 ± 0.009	0.084 ± 0.008	0.061 ± 0.003
OX01-IOH; 5 min	0.739 ± 0.005	0.790 ± 0.001	0.223 ± 0.004	0.020 ± 0.009	0.020 ± 0.000
VA1-CO2; 5 min	0.681 ± 0.002	0.689 ± 0.005	0.102 ± 0.004	0.027 ± 0.005	0.024 ± 0.001
VA1-STIM; 5 min	0.730 ± 0.015	0.775 ± 0.006	0.114 ± 0.004	0.017 ± 0.003	0.009 ± 0.000
Macro; 15 min	0.639 ± 0.001	0.621 ± 0.004	0.071 ± 0.001	0.035 ± 0.004	0.029 ± 0.001
CL07-002; 15 min	0.570 ± 0.012	0.551 ± 0.016	0.036 ± 0.002	0.015 ± 0.007	0.025 ± 0.001
H5-002; 15 min	0.659 ± 0.010	0.626 ± 0.012	0.022 ± 0.002	0.017 ± 0.002	0.008 ± 0.000
H5-030; 15 min	0.566 ± 0.007	0.607 ± 0.005	0.077 ± 0.002	0.048 ± 0.006	0.039 ± 0.001
L2-30; 15 min	0.634 ± 0.006	0.577 ± 0.009	0.114 ± 0.005	0.087 ± 0.012	0.058 ± 0.004
OX01-IOH; 15 min	0.695 ± 0.007	0.682 ± 0.003	0.101 ± 0.005	0.022 ± 0.003	0.027 ± 0.001
VA1-CO2; 15 min	0.666 ± 0.006	0.664 ± 0.007	0.100 ± 0.004	0.029 ± 0.007	0.028 ± 0.001
VA1-STIM; 15 min	0.686 ± 0.005	0.641 ± 0.013	0.048 ± 0.002	0.026 ± 0.006	0.015 ± 0.001
Macro; 30 min	0.640 ± 0.002	0.624 ± 0.002	0.077 ± 0.002	0.034 ± 0.003	0.030 ± 0.001
CL07-002; 30 min	0.546 ± 0.008	0.548 ± 0.009	0.034 ± 0.001	0.021 ± 0.005	0.021 ± 0.001
H5-002; 30 min	0.666 ± 0.005	0.650 ± 0.009	0.027 ± 0.005	0.013 ± 0.002	0.008 ± 0.000
H5-030; 30 min	0.597 ± 0.012	0.615 ± 0.006	0.073 ± 0.003	0.053 ± 0.008	0.035 ± 0.001
L2-30; 30 min	0.629 ± 0.010	0.622 ± 0.002	0.137 ± 0.007	0.072 ± 0.017	0.067 ± 0.004
OX01-IOH; 30 min	0.685 ± 0.007	0.660 ± 0.008	0.113 ± 0.006	0.032 ± 0.006	0.028 ± 0.001
VA1-CO2; 30 min	0.669 ± 0.006	0.640 ± 0.011	0.099 ± 0.007	0.024 ± 0.005	0.034 ± 0.002
VA1-STIM; 30 min	0.690 ± 0.009	0.630 ± 0.008	0.053 ± 0.002	0.021 ± 0.002	0.018 ± 0.001
Macro; 45 min	0.639 ± 0.006	0.619 ± 0.004	0.068 ± 0.000	0.030 ± 0.002	0.030 ± 0.000
CL07-002; 45 min	0.569 ± 0.013	0.559 ± 0.016	0.029 ± 0.002	0.021 ± 0.003	0.018 ± 0.001
H5-002; 45 min	0.650 ± 0.015	0.670 ± 0.010	0.025 ± 0.001	0.015 ± 0.005	0.006 ± 0.000
H5-030; 45 min	0.587 ± 0.006	0.607 ± 0.010	0.080 ± 0.002	0.042 ± 0.006	0.039 ± 0.001
L2-30; 45 min	0.632 ± 0.011	0.607 ± 0.007	0.111 ± 0.005	0.066 ± 0.009	0.062 ± 0.003
OX01-IOH; 45 min	0.657 ± 0.003	0.653 ± 0.006	0.104 ± 0.005	0.023 ± 0.004	0.032 ± 0.001
VA1-CO2; 45 min	0.676 ± 0.005	0.633 ± 0.002	0.082 ± 0.002	0.026 ± 0.003	0.031 ± 0.001
VA1-STIM; 45 min	0.703 ± 0.006	0.603 ± 0.009	0.047 ± 0.003	0.019 ± 0.006	0.022 ± 0.001
Macro; 60 min	0.620 ± 0.005	0.624 ± 0.004	0.071 ± 0.002	0.034 ± 0.003	0.029 ± 0.001
CL07-002; 60 min	0.555 ± 0.014	0.535 ± 0.006	0.036 ± 0.005	0.018 ± 0.002	0.023 ± 0.002
H5-002; 60 min	0.622 ± 0.018	0.652 ± 0.010	0.022 ± 0.001	0.014 ± 0.004	0.006 ± 0.000
H5-030; 60 min	0.592 ± 0.004	0.631 ± 0.009	0.078 ± 0.002	0.046 ± 0.014	0.032 ± 0.001
L2-30; 60 min	0.644 ± 0.008	0.589 ± 0.006	0.086 ± 0.002	0.070 ± 0.006	0.064 ± 0.004
OX01-IOH; 60 min	0.637 ± 0.007	0.611 ± 0.007	0.103 ± 0.006	0.029 ± 0.007	0.037 ± 0.001
VA1-CO2; 60 min	0.655 ± 0.008	0.609 ± 0.014	0.090 ± 0.011	0.030 ± 0.006	0.033 ± 0.002
VA1-STIM; 60 min	0.634 ± 0.018	0.743 ± 0.014	0.087 ± 0.008	0.029 ± 0.006	0.011 ± 0.001

Table 35: Forecast-gap results — VitalDB. Updated export; values retain the supplied mean and SD precision. (Continued.)

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Macro; 75 min	0.619 ± 0.006	0.632 ± 0.004	0.082 ± 0.004	0.035 ± 0.004	0.029 ± 0.001
CL07-002; 75 min	0.560 ± 0.014	0.567 ± 0.014	0.035 ± 0.003	0.026 ± 0.007	0.019 ± 0.001
H5-002; 75 min	0.629 ± 0.019	0.650 ± 0.012	0.022 ± 0.004	0.015 ± 0.002	0.007 ± 0.001
H5-030; 75 min	0.593 ± 0.008	0.598 ± 0.007	0.065 ± 0.002	0.033 ± 0.007	0.035 ± 0.002
L2-30; 75 min	0.649 ± 0.005	0.555 ± 0.012	0.085 ± 0.011	0.071 ± 0.014	0.068 ± 0.006
OX01-IOH; 75 min	0.619 ± 0.007	0.650 ± 0.010	0.127 ± 0.008	0.029 ± 0.004	0.034 ± 0.002
VA1-CO2; 75 min	0.643 ± 0.014	0.655 ± 0.018	0.124 ± 0.005	0.041 ± 0.007	0.029 ± 0.004
VA1-STIM; 75 min	0.636 ± 0.008	0.750 ± 0.028	0.113 ± 0.018	0.030 ± 0.007	0.013 ± 0.002

Table 36: Forecast-gap results — MC-MED. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Decomp. 60m; 0 min	0.675 ± 0.005	0.647 ± 0.003	0.264 ± 0.002	0.068 ± 0.031	0.279 ± 0.065
Pending hyperK; 0 min	0.578 ± 0.028	0.567 ± 0.035	0.046 ± 0.004	0.051 ± 0.027	0.202 ± 0.143
Pending anemia; 0 min	0.538 ± 0.032	0.556 ± 0.014	0.047 ± 0.002	0.035 ± 0.015	0.050 ± 0.020
Decomp. 60m; 10 min	0.636 ± 0.014	0.629 ± 0.016	0.214 ± 0.009	0.066 ± 0.035	0.201 ± 0.066
Pending hyperK; 15 min	0.573 ± 0.015	0.567 ± 0.030	0.046 ± 0.009	0.072 ± 0.039	0.323 ± 0.126
Pending anemia; 15 min	0.573 ± 0.047	0.566 ± 0.016	0.070 ± 0.003	0.046 ± 0.022	0.133 ± 0.063
Decomp. 60m; 30 min	0.664 ± 0.019	0.626 ± 0.017	0.157 ± 0.010	0.073 ± 0.031	0.229 ± 0.088
Lactate NC; 30 min	0.567 ± 0.022	0.499 ± 0.020	0.367 ± 0.015	0.258 ± 0.028	0.423 ± 0.067
Lactate NC; 60 min	0.534 ± 0.023	0.489 ± 0.016	0.385 ± 0.011	0.153 ± 0.034	0.419 ± 0.046
Pending hyperK; 60 min	0.555 ± 0.023	0.532 ± 0.017	0.077 ± 0.013	0.081 ± 0.068	0.257 ± 0.173
Pending anemia; 60 min	0.544 ± 0.053	0.544 ± 0.057	0.200 ± 0.031	0.078 ± 0.018	0.200 ± 0.086
Lactate NC; 90 min	0.577 ± 0.039	0.532 ± 0.014	0.399 ± 0.017	0.247 ± 0.037	0.390 ± 0.047

Table 37: Forecast-gap results — MC-MED. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	Failures (%)
Decomp. 60m; 0 min	2.9	39238 MiB	0.0
Pending hyperK; 0 min	15.9	46850 MiB	0.0
Pending anemia; 0 min	13.7	28876 MiB	0.0
Decomp. 60m; 10 min	5.5	35301 MiB	0.0
Pending hyperK; 15 min	7.5	46461 MiB	0.0
Pending anemia; 15 min	5.5	21562 MiB	0.0
Decomp. 60m; 30 min	3.8	28188 MiB	0.0
Lactate NC; 30 min	2.0	9741 MiB	0.0
Lactate NC; 60 min	1.5	9606 MiB	0.0
Pending hyperK; 60 min	6.4	21782 MiB	0.0
Pending anemia; 60 min	0.6	8324 MiB	0.0
Lactate NC; 90 min	1.6	8381 MiB	0.0

Table 38: Temporal-order control — MC-MED. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC
10-s block shuffle	0.576 ± 0.011	0.563 ± 0.009	0.179 ± 0.004

Table 39: Observation length — MIMIC-III. Task-level test AUROC (mean $\pm$ seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

Acute det. / AKI 2/3 / Circ. det. / Fluid resp.

Condition / budget	Acute det.	AKI 2/3	Circ. det.	Fluid resp.
10 min	0.646 ± 0.009	0.678 ± 0.009	0.657 ± 0.007	0.565 ± 0.010
20 min	0.654 ± 0.015	0.671 ± 0.007	0.656 ± 0.004	0.560 ± 0.008
60 min	0.656 ± 0.009	0.680 ± 0.018	0.662 ± 0.004	0.567 ± 0.004

Dyn. mort. / Resp. esc. / Sepsis-3 / Vent. lib.

Condition / budget	Dyn. mort.	Resp. esc.	Sepsis-3	Vent. lib.
10 min	0.517 ± 0.006	0.575 ± 0.034	0.659 ± 0.010	0.550 ± 0.006
20 min	0.516 ± 0.006	0.585 ± 0.022	0.659 ± 0.004	0.547 ± 0.007
60 min	0.517 ± 0.012	0.593 ± 0.019	0.652 ± 0.005	0.548 ± 0.004

Table 40: Observation length — VitalDB. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

CL07-002 / H5-002 / H5-030 / L2-30				
Condition / budget	CL07-002	H5-002	H5-030	L2-30
5 min	0.537 ± 0.001	0.635 ± 0.003	0.602 ± 0.001	0.560 ± 0.004
15 min	0.533 ± 0.002	0.635 ± 0.002	0.602 ± 0.002	0.576 ± 0.004
30 min	0.532 ± 0.002	0.626 ± 0.002	0.602 ± 0.001	0.579 ± 0.002
60 min	0.536 ± 0.003	0.625 ± 0.002	0.601 ± 0.002	0.594 ± 0.002
OX01-IOH / VA1-CO2 / VA1-STIM				
Condition / budget	OX01-IOH	VA1-CO2	VA1-STIM	
5 min	0.689 ± 0.003	0.676 ± 0.002	0.687 ± 0.002	
15 min	0.712 ± 0.002	0.674 ± 0.003	0.686 ± 0.003	
30 min	0.729 ± 0.001	0.677 ± 0.001	0.696 ± 0.002	
60 min	0.727 ± 0.004	0.678 ± 0.001	0.692 ± 0.003	

Table 41: Observation length — MC-MED. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

Decomp. 60m / Lactate NC / Pending hyperK / Pending anemia				
Condition / budget	Decomp. 60m	Lactate NC	Pending hyperK	Pending anemia
5 min	0.647 ± 0.003	0.499 ± 0.020	0.566 ± 0.035	0.556 ± 0.014
10 min	0.656 ± 0.009	0.502 ± 0.013	0.576 ± 0.027	0.577 ± 0.022
15 min	0.631 ± 0.014	0.577 ± 0.022	0.528 ± 0.057	0.530 ± 0.050

Table 42: Forecast gap — MIMIC-III. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

AKI 2/3 / Dyn. mort. / Sepsis-3			
Condition / budget	AKI 2/3	Dyn. mort.	Sepsis-3
2 h	–	–	0.629 ± 0.007
3 h	0.686 ± 0.017	0.545 ± 0.011	–
4 h	–	–	0.619 ± 0.010
6 h	0.697 ± 0.028	–	0.629 ± 0.004
12 h	0.685 ± 0.009	0.539 ± 0.011	–
24 h	–	0.537 ± 0.009	–

Table 43: Forecast gap — VitalDB. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

CL07-002 / H5-002 / H5-030 / L2-30				
Condition / budget	CL07-002	H5-002	H5-030	L2-30
5 min	0.552 ± 0.012	0.651 ± 0.002	0.592 ± 0.007	0.582 ± 0.010
15 min	0.552 ± 0.016	0.626 ± 0.012	0.607 ± 0.005	0.577 ± 0.009
30 min	0.548 ± 0.009	0.650 ± 0.010	0.615 ± 0.006	0.622 ± 0.002
45 min	0.559 ± 0.016	0.670 ± 0.010	0.606 ± 0.010	0.607 ± 0.007
60 min	0.535 ± 0.006	0.652 ± 0.010	0.631 ± 0.009	0.589 ± 0.006
75 min	0.567 ± 0.014	0.650 ± 0.012	0.598 ± 0.007	0.555 ± 0.012
OX01-IOH / VA1-CO2 / VA1-STIM				
Condition / budget	OX01-IOH	VA1-CO2	VA1-STIM	
5 min	0.791 ± 0.001	0.689 ± 0.005	0.775 ± 0.006	
15 min	0.682 ± 0.003	0.664 ± 0.007	0.641 ± 0.013	
30 min	0.660 ± 0.008	0.640 ± 0.011	0.630 ± 0.008	
45 min	0.653 ± 0.006	0.633 ± 0.002	0.603 ± 0.009	
60 min	0.611 ± 0.008	0.609 ± 0.014	0.743 ± 0.014	
75 min	0.650 ± 0.010	0.655 ± 0.018	0.750 ± 0.028	

Table 44: Forecast gap — MC-MED. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

Decomp. 60m / Lactate NC / Pending hyperK / Pending anemia				
Condition / budget	Decomp. 60m	Lactate NC	Pending hyperK	Pending anemia
0 min	0.647 ± 0.003	–	0.566 ± 0.035	0.556 ± 0.014
10 min	0.629 ± 0.016	–	–	–
15 min	–	–	0.567 ± 0.031	0.566 ± 0.016
30 min	0.626 ± 0.017	0.499 ± 0.020	–	–
60 min	–	0.489 ± 0.016	0.532 ± 0.017	0.544 ± 0.057
90 min	–	0.532 ± 0.014	–	–

Table 45: Training-budget summaries — MIMIC-III. Updated export; values retain the supplied mean and SD precision. Budgets refer to training windows on MIMIC-III and subjects on VitalDB and MC-MED.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Diversity (10 %)	0.596 ± 0.004	0.571 ± 0.008	0.098 ± 0.003	0.240 ± 0.014	0.338 ± 0.057
Diversity (20 %)	0.611 ± 0.011	0.583 ± 0.002	0.102 ± 0.002	0.236 ± 0.012	0.368 ± 0.030
Diversity (30 %)	0.609 ± 0.008	0.586 ± 0.007	0.101 ± 0.002	0.237 ± 0.012	0.381 ± 0.044
Diversity cluster (10 %)	0.607 ± 0.002	0.578 ± 0.003	0.101 ± 0.002	0.238 ± 0.018	0.338 ± 0.056
Diversity cluster (20 %)	0.611 ± 0.004	0.583 ± 0.004	0.101 ± 0.002	0.236 ± 0.022	0.347 ± 0.051
Diversity cluster (30 %)	0.620 ± 0.002	0.587 ± 0.003	0.101 ± 0.003	0.226 ± 0.009	0.363 ± 0.028
Full cohort (100 %)	0.635 ± 0.001	0.606 ± 0.002	0.107 ± 0.001	0.189 ± 0.019	0.344 ± 0.019
Gradient match (10 %)	0.599 ± 0.006	0.577 ± 0.003	0.098 ± 0.001	0.202 ± 0.014	0.341 ± 0.024
Gradient match (20 %)	0.604 ± 0.007	0.576 ± 0.001	0.099 ± 0.001	0.219 ± 0.009	0.348 ± 0.026
Gradient match (30 %)	0.602 ± 0.004	0.571 ± 0.005	0.097 ± 0.002	0.216 ± 0.015	0.366 ± 0.049
NLDR (10 %)	0.606 ± 0.003	0.580 ± 0.002	0.098 ± 0.001	0.234 ± 0.011	0.330 ± 0.029
NLDR (20 %)	0.620 ± 0.002	0.589 ± 0.004	0.101 ± 0.002	0.207 ± 0.021	0.322 ± 0.016
NLDR (30 %)	0.623 ± 0.003	0.588 ± 0.004	0.101 ± 0.002	0.216 ± 0.016	0.334 ± 0.025
Noise filter (10 %)	0.594 ± 0.004	0.572 ± 0.003	0.099 ± 0.003	0.250 ± 0.029	0.333 ± 0.050
Noise filter (20 %)	0.607 ± 0.006	0.579 ± 0.004	0.098 ± 0.001	0.224 ± 0.013	0.347 ± 0.040
Noise filter (30 %)	0.608 ± 0.004	0.583 ± 0.003	0.100 ± 0.003	0.226 ± 0.013	0.348 ± 0.055
Random (10 %)	0.607 ± 0.004	0.574 ± 0.004	0.098 ± 0.003	0.246 ± 0.011	0.354 ± 0.007
Random (20 %)	0.618 ± 0.003	0.586 ± 0.004	0.102 ± 0.003	0.218 ± 0.026	0.319 ± 0.030
Random (30 %)	0.625 ± 0.006	0.590 ± 0.003	0.102 ± 0.001	0.241 ± 0.015	0.346 ± 0.016
Task anchor (10 %)	0.502 ± 0.005	0.504 ± 0.008	0.090 ± 0.002	0.340 ± 0.009	0.422 ± 0.022
Task anchor (20 %)	0.506 ± 0.005	0.501 ± 0.003	0.090 ± 0.002	0.331 ± 0.012	0.421 ± 0.036
Task anchor (30 %)	0.510 ± 0.001	0.496 ± 0.002	0.087 ± 0.001	0.309 ± 0.012	0.435 ± 0.017
Training dynamics (10 %)	0.589 ± 0.005	0.552 ± 0.002	0.094 ± 0.001	0.219 ± 0.017	0.242 ± 0.034
Training dynamics (20 %)	0.592 ± 0.002	0.556 ± 0.003	0.093 ± 0.002	0.271 ± 0.014	0.287 ± 0.037
Training dynamics (30 %)	0.571 ± 0.004	0.555 ± 0.001	0.094 ± 0.001	0.241 ± 0.019	0.309 ± 0.045

Table 46: Training-budget summaries — MIMIC-III. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	ms/sample	Failures (%)
Diversity (10 %)	23.5	3.3 %	0.39	0.0
Diversity (20 %)	33.2	3.4 %	0.26	0.0
Diversity (30 %)	43.4	3.3 %	0.21	0.0
Diversity cluster (10 %)	25.4	3.2 %	0.37	0.0
Diversity cluster (20 %)	35.0	3.3 %	0.25	0.0
Diversity cluster (30 %)	44.2	3.4 %	0.21	0.0
Full cohort (100 %)	108.6	3.4 %	0.24	0.0
Gradient match (10 %)	26.8	3.4 %	0.42	0.0
Gradient match (20 %)	33.0	3.4 %	0.25	0.0
Gradient match (30 %)	42.6	3.3 %	0.21	0.0
NLDR (10 %)	24.2	3.3 %	0.40	0.0
NLDR (20 %)	33.5	3.3 %	0.25	0.0
NLDR (30 %)	44.1	3.3 %	0.21	0.0
Noise filter (10 %)	27.5	3.4 %	0.42	0.0
Noise filter (20 %)	38.5	3.4 %	0.28	0.0
Noise filter (30 %)	48.2	3.4 %	0.22	0.0
Random (10 %)	–	–	–	–
Random (20 %)	–	–	–	–
Random (30 %)	–	–	–	–
Task anchor (10 %)	24.1	3.3 %	0.40	0.0
Task anchor (20 %)	33.4	3.4 %	0.26	0.0
Task anchor (30 %)	44.2	3.4 %	0.21	0.0
Training dynamics (10 %)	27.9	3.3 %	0.42	0.0
Training dynamics (20 %)	34.0	3.4 %	0.25	0.0
Training dynamics (30 %)	43.8	3.4 %	0.21	0.0

Table 47: Training-budget summaries — VitalDB. Updated export; values retain the supplied mean and SD precision. Budgets refer to training windows on MIMIC-III and subjects on VitalDB and MC-MED.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Diversity (5 %)	0.556 ± 0.002	0.555 ± 0.002	0.056 ± 0.001	0.075 ± 0.002	0.057 ± 0.001
Diversity (10 %)	0.572 ± 0.001	0.570 ± 0.002	0.059 ± 0.001	0.068 ± 0.001	0.049 ± 0.001
Diversity (20 %)	0.590 ± 0.002	0.582 ± 0.002	0.063 ± 0.001	0.067 ± 0.001	0.043 ± 0.001
Diversity cluster (5 %)	0.559 ± 0.003	0.561 ± 0.002	0.058 ± 0.001	0.053 ± 0.001	0.046 ± 0.001
Diversity cluster (10 %)	0.576 ± 0.002	0.572 ± 0.001	0.061 ± 0.001	0.048 ± 0.002	0.040 ± 0.001
Diversity cluster (20 %)	0.590 ± 0.002	0.584 ± 0.001	0.066 ± 0.001	0.046 ± 0.001	0.036 ± 0.001
Full cohort (100 %)	0.615 ± 0.002	0.613 ± 0.001	0.072 ± 0.001	0.039 ± 0.002	0.032 ± 0.001
Gradient match (5 %)	0.561 ± 0.001	0.565 ± 0.002	0.057 ± 0.000	0.051 ± 0.001	0.065 ± 0.002
Gradient match (10 %)	0.578 ± 0.002	0.579 ± 0.003	0.060 ± 0.001	0.044 ± 0.001	0.069 ± 0.002
Gradient match (20 %)	0.594 ± 0.002	0.594 ± 0.002	0.065 ± 0.001	0.042 ± 0.002	0.065 ± 0.002
NLDR (5 %)	0.559 ± 0.002	0.557 ± 0.001	0.058 ± 0.000	0.055 ± 0.002	0.049 ± 0.001
NLDR (10 %)	0.573 ± 0.001	0.572 ± 0.002	0.061 ± 0.001	0.049 ± 0.001	0.045 ± 0.002
NLDR (20 %)	0.589 ± 0.001	0.585 ± 0.001	0.063 ± 0.001	0.046 ± 0.002	0.038 ± 0.002
Noise filter (5 %)	0.552 ± 0.002	0.557 ± 0.003	0.056 ± 0.001	0.036 ± 0.003	0.042 ± 0.002
Noise filter (10 %)	0.569 ± 0.002	0.567 ± 0.001	0.059 ± 0.000	0.027 ± 0.001	0.039 ± 0.002
Noise filter (20 %)	0.589 ± 0.001	0.593 ± 0.001	0.066 ± 0.001	0.025 ± 0.001	0.033 ± 0.000
Random (5 %)	0.565 ± 0.004	0.559 ± 0.003	0.057 ± 0.002	0.053 ± 0.004	0.049 ± 0.003
Random (10 %)	0.581 ± 0.004	0.573 ± 0.005	0.060 ± 0.001	0.043 ± 0.002	0.041 ± 0.001
Random (20 %)	0.594 ± 0.003	0.585 ± 0.004	0.064 ± 0.002	0.041 ± 0.002	0.036 ± 0.001
Task anchor (5 %)	0.569 ± 0.002	0.559 ± 0.001	0.059 ± 0.001	0.129 ± 0.001	0.117 ± 0.003
Task anchor (10 %)	0.583 ± 0.001	0.575 ± 0.002	0.063 ± 0.001	0.100 ± 0.003	0.086 ± 0.002
Task anchor (20 %)	0.593 ± 0.001	0.586 ± 0.003	0.066 ± 0.001	0.071 ± 0.001	0.054 ± 0.001
Training dynamics (5 %)	0.569 ± 0.001	0.568 ± 0.002	0.059 ± 0.001	0.191 ± 0.001	0.178 ± 0.002
Training dynamics (10 %)	0.580 ± 0.002	0.577 ± 0.001	0.061 ± 0.001	0.117 ± 0.002	0.136 ± 0.002
Training dynamics (20 %)	0.591 ± 0.001	0.585 ± 0.001	0.065 ± 0.001	0.055 ± 0.003	0.071 ± 0.001

Table 48: Training-budget summaries — VitalDB. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	ms/sample	Failures (%)
Diversity (5 %)	32.0	0.41	0.0
Diversity (10 %)	31.9	0.41	0.0
Diversity (20 %)	32.2	0.42	0.0
Diversity cluster (5 %)	32.0	0.42	0.0
Diversity cluster (10 %)	31.9	0.41	0.0
Diversity cluster (20 %)	32.1	0.42	0.0
Full cohort (100 %)	32.0	0.41	0.0
Gradient match (5 %)	31.9	0.41	0.0
Gradient match (10 %)	31.9	0.41	0.0
Gradient match (20 %)	31.9	0.41	0.0
NLDR (5 %)	31.8	0.41	0.0
NLDR (10 %)	31.9	0.41	0.0
NLDR (20 %)	32.0	0.41	0.0
Noise filter (5 %)	31.9	0.41	0.0
Noise filter (10 %)	32.0	0.41	0.0
Noise filter (20 %)	32.0	0.41	0.0
Random (5 %)	32.1	0.42	0.0
Random (10 %)	31.9	0.41	0.0
Random (20 %)	32.0	0.41	0.0
Task anchor (5 %)	31.9	0.41	0.0
Task anchor (10 %)	32.0	0.42	0.0
Task anchor (20 %)	32.0	0.42	0.0
Training dynamics (5 %)	32.0	0.41	0.0
Training dynamics (10 %)	32.1	0.42	0.0
Training dynamics (20 %)	32.0	0.41	0.0

Table 49: Training-budget summaries — MC-MED. Updated export; values retain the supplied mean and SD precision. Budgets refer to training windows on MIMIC-III and subjects on VitalDB and MC-MED.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Diversity (10 %)	0.552 ± 0.015	0.531 ± 0.019	0.168 ± 0.002	0.229 ± 0.064	0.465 ± 0.106
Diversity (20 %)	0.555 ± 0.014	0.554 ± 0.010	0.171 ± 0.004	0.133 ± 0.016	0.356 ± 0.092
Diversity (50 %)	0.566 ± 0.004	0.567 ± 0.012	0.179 ± 0.007	0.106 ± 0.032	0.354 ± 0.091
Diversity cluster (10 %)	0.568 ± 0.014	0.548 ± 0.013	0.172 ± 0.002	0.146 ± 0.051	0.469 ± 0.137
Diversity cluster (20 %)	0.560 ± 0.010	0.554 ± 0.015	0.173 ± 0.004	0.128 ± 0.050	0.356 ± 0.133
Diversity cluster (50 %)	0.568 ± 0.013	0.558 ± 0.008	0.175 ± 0.004	0.101 ± 0.021	0.305 ± 0.086
Full cohort (100 %)	0.579 ± 0.014	0.567 ± 0.004	0.182 ± 0.004	0.079 ± 0.016	0.296 ± 0.057
Gradient match (10 %)	0.497 ± 0.005	0.520 ± 0.017	0.152 ± 0.003	0.319 ± 0.064	0.440 ± 0.159
Gradient match (20 %)	0.517 ± 0.013	0.502 ± 0.016	0.151 ± 0.005	0.335 ± 0.063	0.532 ± 0.204
Gradient match (50 %)	0.541 ± 0.008	0.542 ± 0.014	0.172 ± 0.004	0.126 ± 0.030	0.323 ± 0.063
NLDR (10 %)	0.560 ± 0.017	0.531 ± 0.015	0.162 ± 0.006	0.152 ± 0.041	0.398 ± 0.053
NLDR (20 %)	0.567 ± 0.007	0.537 ± 0.014	0.166 ± 0.005	0.127 ± 0.022	0.274 ± 0.096
NLDR (50 %)	0.581 ± 0.020	0.559 ± 0.019	0.172 ± 0.005	0.132 ± 0.029	0.350 ± 0.099
Noise filter (10 %)	0.524 ± 0.019	0.541 ± 0.019	0.163 ± 0.008	0.197 ± 0.039	0.374 ± 0.160
Noise filter (20 %)	0.547 ± 0.018	0.550 ± 0.013	0.164 ± 0.007	0.126 ± 0.013	0.364 ± 0.061
Noise filter (50 %)	0.566 ± 0.008	0.552 ± 0.004	0.168 ± 0.003	0.157 ± 0.022	0.472 ± 0.125
Random (10 %)	0.550 ± 0.020	0.546 ± 0.019	0.166 ± 0.005	0.170 ± 0.057	0.284 ± 0.126
Random (20 %)	0.561 ± 0.010	0.541 ± 0.021	0.171 ± 0.008	0.160 ± 0.022	0.366 ± 0.047
Random (50 %)	0.572 ± 0.011	0.559 ± 0.013	0.176 ± 0.005	0.127 ± 0.048	0.296 ± 0.080
Task anchor (10 %)	0.538 ± 0.015	0.490 ± 0.004	0.149 ± 0.002	0.182 ± 0.021	0.413 ± 0.112
Task anchor (20 %)	0.516 ± 0.010	0.494 ± 0.012	0.149 ± 0.006	0.224 ± 0.060	0.467 ± 0.095
Task anchor (50 %)	0.512 ± 0.007	0.518 ± 0.004	0.161 ± 0.003	0.187 ± 0.022	0.460 ± 0.069
Training dynamics (10 %)	0.561 ± 0.014	0.561 ± 0.004	0.170 ± 0.004	0.167 ± 0.054	0.380 ± 0.109
Training dynamics (20 %)	0.567 ± 0.010	0.558 ± 0.011	0.171 ± 0.005	0.140 ± 0.018	0.363 ± 0.064
Training dynamics (50 %)	0.574 ± 0.016	0.570 ± 0.006	0.175 ± 0.002	0.076 ± 0.023	0.291 ± 0.056

Table 50: Training-budget summaries — MC-MED. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	Failures (%)
Diversity (10 %)	2.2	32554 MiB	0.0
Diversity (20 %)	5.0	32338 MiB	0.0
Diversity (50 %)	11.8	32630 MiB	0.0
Diversity cluster (10 %)	2.5	32554 MiB	0.0
Diversity cluster (20 %)	4.9	32338 MiB	0.0
Diversity cluster (50 %)	12.9	31479 MiB	0.0
Full cohort (100 %)	20.7	32063 MiB	0.0
Gradient match (10 %)	1.2	32494 MiB	0.0
Gradient match (20 %)	1.6	32414 MiB	0.0
Gradient match (50 %)	12.4	32338 MiB	0.0
NLDR (10 %)	1.5	32913 MiB	0.0
NLDR (20 %)	4.8	32270 MiB	0.0
NLDR (50 %)	9.9	32646 MiB	0.0
Noise filter (10 %)	2.4	31806 MiB	0.0
Noise filter (20 %)	4.7	32600 MiB	0.0
Noise filter (50 %)	8.8	32166 MiB	0.0
Random (10 %)	2.4	32968 MiB	0.0
Random (20 %)	4.8	32610 MiB	0.0
Random (50 %)	9.2	32968 MiB	0.0
Task anchor (10 %)	1.6	32070 MiB	0.0
Task anchor (20 %)	3.1	32070 MiB	0.0
Task anchor (50 %)	8.1	32353 MiB	0.0
Training dynamics (10 %)	2.2	32913 MiB	0.0
Training dynamics (20 %)	4.5	32346 MiB	0.0
Training dynamics (50 %)	10.9	32697 MiB	0.0

Table 51: Channel-budget summaries — MIMIC-III. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
All channels (62)	0.635 ± 0.001	0.606 ± 0.002	0.107 ± 0.001	–	–
Lasso (10)	0.625 ± 0.006	0.599 ± 0.004	0.105 ± 0.001	0.204 ± 0.029	0.347 ± 0.032
Mutual information (10)	0.598 ± 0.007	0.573 ± 0.005	0.098 ± 0.002	0.283 ± 0.032	0.357 ± 0.077
Random (10)	0.579 ± 0.008	0.558 ± 0.008	0.097 ± 0.001	0.334 ± 0.033	0.451 ± 0.037
Clinical (10)	0.633 ± 0.002	0.606 ± 0.004	0.104 ± 0.002	0.202 ± 0.021	0.343 ± 0.031
Waveform SQI (10)	0.635 ± 0.004	0.607 ± 0.003	0.106 ± 0.002	0.197 ± 0.009	0.359 ± 0.031
Temporal MI (10)	0.595 ± 0.010	0.570 ± 0.008	0.099 ± 0.003	0.285 ± 0.033	0.444 ± 0.059

Table 52: Channel-budget summaries — MIMIC-III. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	ms/sample	Failures (%)
All channels (62)	108.6	3.4 %	0.24	0.0
Lasso (10)	96.4	2.4 %	0.16	0.0
Mutual information (10)	97.2	2.4 %	0.14	0.0
Random (10)	103.2	2.4 %	0.15	0.0
Clinical (10)	100.1	2.4 %	0.14	0.0
Waveform SQI (10)	102.1	2.4 %	0.14	0.0
Temporal MI (10)	92.8	2.4 %	0.14	0.0

Table 53: Channel-budget summaries — VitalDB. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
All channels (12)	0.617 ± 0.001	0.612 ± 0.001	0.072 ± 0.001	0.039 ± 0.001	0.032 ± 0.001
Lasso (2)	0.571 ± 0.005	0.565 ± 0.006	0.058 ± 0.001	0.044 ± 0.002	0.034 ± 0.000
Lasso (4)	0.591 ± 0.002	0.587 ± 0.005	0.066 ± 0.001	0.045 ± 0.001	0.033 ± 0.001
Lasso (8)	0.610 ± 0.002	0.605 ± 0.003	0.071 ± 0.001	0.042 ± 0.001	0.032 ± 0.001
Mutual information (2)	0.569 ± 0.006	0.563 ± 0.006	0.058 ± 0.002	0.044 ± 0.001	0.035 ± 0.001
Mutual information (4)	0.591 ± 0.004	0.584 ± 0.003	0.066 ± 0.002	0.044 ± 0.001	0.033 ± 0.001
Mutual information (8)	0.611 ± 0.001	0.604 ± 0.002	0.070 ± 0.001	0.041 ± 0.003	0.032 ± 0.001
Random (2)	0.552 ± 0.005	0.546 ± 0.005	0.056 ± 0.001	0.045 ± 0.001	0.036 ± 0.001
Random (4)	0.576 ± 0.004	0.568 ± 0.005	0.061 ± 0.002	0.045 ± 0.001	0.035 ± 0.001
Random (8)	0.601 ± 0.005	0.596 ± 0.005	0.069 ± 0.002	0.042 ± 0.001	0.033 ± 0.001
Waveform SQI (2)	0.542 ± 0.001	0.546 ± 0.003	0.054 ± 0.001	0.047 ± 0.001	0.036 ± 0.001
Waveform SQI (4)	0.590 ± 0.001	0.594 ± 0.001	0.071 ± 0.001	0.045 ± 0.001	0.031 ± 0.000
Waveform SQI (8)	0.606 ± 0.002	0.602 ± 0.002	0.068 ± 0.001	0.041 ± 0.001	0.033 ± 0.001
Temporal MI (2)	0.575 ± 0.003	0.572 ± 0.003	0.060 ± 0.001	0.044 ± 0.002	0.034 ± 0.001
Temporal MI (4)	0.594 ± 0.002	0.589 ± 0.003	0.066 ± 0.002	0.044 ± 0.002	0.033 ± 0.001
Temporal MI (8)	0.612 ± 0.002	0.607 ± 0.002	0.071 ± 0.001	0.041 ± 0.002	0.032 ± 0.001

Table 54: Channel-budget summaries — VitalDB. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	ms/sample	Failures (%)
All channels (12)	–	–	–
Lasso (2)	33.3	0.43	0.0
Lasso (4)	33.4	0.43	0.0
Lasso (8)	33.6	0.44	0.0
Mutual information (2)	33.5	0.43	0.0
Mutual information (4)	33.6	0.44	0.0
Mutual information (8)	33.5	0.43	0.0
Random (2)	33.3	0.43	0.0
Random (4)	33.8	0.44	0.0
Random (8)	33.2	0.43	0.0
Waveform SQI (2)	33.4	0.43	0.0
Waveform SQI (4)	33.7	0.44	0.0
Waveform SQI (8)	33.4	0.43	0.0
Temporal MI (2)	33.6	0.44	0.0
Temporal MI (4)	33.5	0.43	0.0
Temporal MI (8)	33.3	0.43	0.0

Table 55: Missing-input summaries — MIMIC-III. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Drifting model	0.631 ± 0.001	0.600 ± 0.004	0.103 ± 0.001	0.182 ± 0.019	0.361 ± 0.023
GRIN	0.639 ± 0.003	0.607 ± 0.003	0.106 ± 0.001	0.188 ± 0.014	0.349 ± 0.039
GRU-D	0.638 ± 0.004	0.608 ± 0.004	0.107 ± 0.001	0.174 ± 0.012	0.361 ± 0.033
Lead-agnostic SSL	0.642 ± 0.003	0.609 ± 0.003	0.106 ± 0.002	0.154 ± 0.007	0.347 ± 0.018
Positional kriging	0.637 ± 0.002	0.603 ± 0.003	0.104 ± 0.002	0.189 ± 0.007	0.358 ± 0.034
Raindrop	0.638 ± 0.003	0.606 ± 0.003	0.105 ± 0.001	0.162 ± 0.014	0.329 ± 0.013
Rectified flow	0.635 ± 0.003	0.603 ± 0.004	0.106 ± 0.002	0.202 ± 0.010	0.357 ± 0.023
TrigFlow	0.635 ± 0.003	0.602 ± 0.003	0.105 ± 0.001	0.198 ± 0.015	0.349 ± 0.023
Zero-mask + Δ	0.637 ± 0.003	0.604 ± 0.003	0.105 ± 0.001	0.206 ± 0.022	0.360 ± 0.050

Table 56: Missing-input summaries — MIMIC-III. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	ms/sample	Failures (%)
Drifting model	276.0	2.9 %	0.30	0.0
GRIN	234.2	4.4 %	0.25	0.0
GRU-D	369.3	3.5 %	0.40	0.0
Lead-agnostic SSL	406.2	5.4 %	0.40	0.0
Positional kriging	172.7	4.1 %	0.18	0.0
Raindrop	265.5	5.4 %	0.30	0.0
Rectified flow	285.0	2.9 %	0.33	0.0
TrigFlow	311.6	3.0 %	0.34	0.0
Zero-mask + Δ	160.0	3.5 %	0.16	0.0

Table 57: Missing-input summaries — VitalDB. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Native masks	0.618 ± 0.001	0.613 ± 0.002	0.072 ± 0.001	0.038 ± 0.001	0.032 ± 0.000
Drifting model	0.592 ± 0.002	0.587 ± 0.001	0.069 ± 0.001	0.042 ± 0.001	0.033 ± 0.001
GRIN	0.593 ± 0.001	0.589 ± 0.002	0.070 ± 0.001	0.042 ± 0.001	0.033 ± 0.001
GRU-D	0.590 ± 0.001	0.588 ± 0.002	0.071 ± 0.001	0.041 ± 0.002	0.034 ± 0.000
Lead-agnostic SSL	0.591 ± 0.002	0.586 ± 0.001	0.071 ± 0.000	0.043 ± 0.001	0.033 ± 0.000
Positional kriging	0.587 ± 0.002	0.581 ± 0.001	0.067 ± 0.001	0.042 ± 0.002	0.034 ± 0.000
Raindrop	0.610 ± 0.002	0.613 ± 0.001	0.072 ± 0.001	0.039 ± 0.001	0.033 ± 0.000
Rectified flow	0.591 ± 0.001	0.588 ± 0.001	0.069 ± 0.001	0.042 ± 0.001	0.033 ± 0.001
TrigFlow	0.593 ± 0.001	0.590 ± 0.002	0.070 ± 0.001	0.041 ± 0.002	0.034 ± 0.000
Zero-mask + Δ	0.590 ± 0.002	0.592 ± 0.001	0.070 ± 0.001	0.041 ± 0.001	0.034 ± 0.001

Table 58: Missing-input summaries — VitalDB. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Failures (%)
Native masks	0.0
Drifting model	0.0
GRIN	0.0
GRU-D	0.0
Lead-agnostic SSL	0.0
Positional kriging	0.0
Raindrop	0.0
Rectified flow	0.0
TrigFlow	0.0
Zero-mask + Δ	0.0

Table 59: Fusion summaries — MIMIC-III. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
EHR only	0.663 ± 0.003	0.636 ± 0.003	0.118 ± 0.002	0.148 ± 0.023	0.372 ± 0.064
Late concatenation	0.645 ± 0.003	0.619 ± 0.006	0.112 ± 0.005	0.150 ± 0.021	0.335 ± 0.054
Tensor fusion	0.640 ± 0.007	0.610 ± 0.005	0.105 ± 0.002	0.181 ± 0.009	0.336 ± 0.047
Waveform only	0.638 ± 0.008	0.604 ± 0.007	0.106 ± 0.003	0.196 ± 0.013	0.366 ± 0.062

Table 60: Fusion summaries — VitalDB. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
EHR only	0.665 ± 0.003	0.690 ± 0.001	0.117 ± 0.001	0.024 ± 0.002	0.024 ± 0.000
EHR-query attention	0.663 ± 0.003	0.674 ± 0.004	0.096 ± 0.001	0.029 ± 0.002	0.025 ± 0.000
Gated	0.681 ± 0.004	0.709 ± 0.003	0.126 ± 0.003	0.026 ± 0.002	0.021 ± 0.001
Weighted late fusion	0.692 ± 0.003	0.711 ± 0.001	0.126 ± 0.002	0.020 ± 0.001	0.021 ± 0.000
Concatenation	0.689 ± 0.001	0.712 ± 0.002	0.125 ± 0.002	0.023 ± 0.002	0.021 ± 0.000
Tensor fusion	0.684 ± 0.004	0.708 ± 0.002	0.124 ± 0.002	0.024 ± 0.001	0.022 ± 0.000
Waveform only	0.639 ± 0.001	0.634 ± 0.003	0.079 ± 0.002	0.034 ± 0.002	0.029 ± 0.001

Table 61: Fusion summaries — VitalDB. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	ms/sample	Failures (%)
EHR only	2.0	0.03	0.0
EHR-query attention	11.7	0.15	0.0
Gated	11.3	0.15	0.0
Weighted late fusion	11.2	0.15	0.0
Concatenation	10.9	0.14	0.0
Tensor fusion	11.2	0.15	0.0
Waveform only	–	–	0.0

Table 62: Fusion summaries — MC-MED. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Co-attention	0.628 ± 0.006	0.603 ± 0.010	0.186 ± 0.008	0.106 ± 0.026	0.286 ± 0.059
Co-attention | resnet1d	0.611 ± 0.016	0.605 ± 0.011	0.179 ± 0.007	0.207 ± 0.055	0.398 ± 0.167
Early concatenation	0.604 ± 0.013	0.589 ± 0.007	0.186 ± 0.007	0.096 ± 0.029	0.320 ± 0.050
Concatenation (early) | resnet1d	0.607 ± 0.010	0.585 ± 0.009	0.184 ± 0.003	0.102 ± 0.015	0.248 ± 0.032
EHR only	0.614 ± 0.010	0.607 ± 0.006	0.182 ± 0.005	0.099 ± 0.022	0.311 ± 0.047
EHR only | resnet1d	0.600 ± 0.004	0.595 ± 0.013	0.178 ± 0.009	0.125 ± 0.021	0.303 ± 0.060
Gated	0.628 ± 0.005	0.607 ± 0.004	0.189 ± 0.007	0.095 ± 0.022	0.276 ± 0.046
Gated | resnet1d	0.603 ± 0.009	0.599 ± 0.004	0.183 ± 0.006	0.157 ± 0.063	0.267 ± 0.131
Late fusion	0.628 ± 0.004	0.609 ± 0.004	0.192 ± 0.005	0.115 ± 0.028	0.285 ± 0.056
Late fusion | resnet1d	0.614 ± 0.009	0.601 ± 0.004	0.180 ± 0.006	0.152 ± 0.033	0.286 ± 0.096
Waveform only	0.579 ± 0.014	0.567 ± 0.004	0.182 ± 0.004	0.079 ± 0.016	0.296 ± 0.057

Table 63: Fusion summaries — MC-MED. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	Failures (%)
Co-attention	24.4	31944 MiB	0.0
Co-attention | resnet1d	6.9	31645 MiB	0.0
Early concatenation	21.7	32791 MiB	0.0
Concatenation (early) | resnet1d	8.5	31588 MiB	0.0
EHR only	6.5	3818 MiB	0.0
EHR only | resnet1d	2.3	3917 MiB	0.0
Gated	20.6	32838 MiB	0.0
Gated | resnet1d	8.4	31992 MiB	0.0
Late fusion	19.2	32383 MiB	0.0
Late fusion | resnet1d	8.5	31694 MiB	0.0
Waveform only	20.7	32063 MiB	0.0

Table 64: Expanded MIMIC-III fusion grid — MIMIC-III. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Bilinear | Mamba-3/Conv1D	0.639 ± 0.008	0.608 ± 0.007	0.106 ± 0.005	0.175 ± 0.024	0.387 ± 0.056
Bilinear | Transformer/Crossformer	0.638 ± 0.011	0.599 ± 0.008	0.103 ± 0.003	0.171 ± 0.022	0.379 ± 0.102
Bilinear | Transformer/Conv1D	0.646 ± 0.007	0.622 ± 0.010	0.111 ± 0.007	0.175 ± 0.034	0.320 ± 0.039
Concat + product | Mamba-3/Conv1D	0.639 ± 0.003	0.608 ± 0.007	0.105 ± 0.004	0.186 ± 0.020	0.346 ± 0.055
Concat + product | Transformer/Crossformer	0.639 ± 0.006	0.599 ± 0.003	0.102 ± 0.002	0.177 ± 0.027	0.332 ± 0.043
Concat + product | Transformer/Conv1D	0.645 ± 0.009	0.609 ± 0.011	0.110 ± 0.008	0.180 ± 0.038	0.351 ± 0.039
EHR only | Mamba-3/Conv1D	0.664 ± 0.005	0.635 ± 0.004	0.116 ± 0.003	0.149 ± 0.021	0.301 ± 0.052
EHR only | Transformer/Crossformer	0.662 ± 0.004	0.636 ± 0.003	0.120 ± 0.003	0.140 ± 0.031	0.302 ± 0.024
EHR only | Transformer/Conv1D	0.660 ± 0.004	0.632 ± 0.004	0.118 ± 0.004	0.169 ± 0.024	0.308 ± 0.092
Feature attention | Transformer/Crossformer	0.644 ± 0.006	0.602 ± 0.009	0.102 ± 0.004	0.199 ± 0.039	0.411 ± 0.051
Feature attention | Transformer/Conv1D	0.643 ± 0.005	0.614 ± 0.003	0.109 ± 0.003	0.165 ± 0.017	0.323 ± 0.043
FiLM | Mamba-3/Conv1D	0.639 ± 0.009	0.608 ± 0.006	0.106 ± 0.005	0.171 ± 0.021	0.362 ± 0.044
FiLM | Transformer/Crossformer	0.640 ± 0.013	0.601 ± 0.009	0.102 ± 0.004	0.197 ± 0.034	0.389 ± 0.021
FiLM | Transformer/Conv1D	0.643 ± 0.010	0.615 ± 0.012	0.108 ± 0.005	0.174 ± 0.017	0.353 ± 0.048
Gated | Mamba-3/Conv1D	0.640 ± 0.015	0.609 ± 0.007	0.106 ± 0.002	0.155 ± 0.027	0.295 ± 0.043
Gated | Transformer/Crossformer	0.641 ± 0.009	0.599 ± 0.005	0.103 ± 0.004	0.209 ± 0.020	0.348 ± 0.054
Gated | Transformer/Conv1D	0.642 ± 0.008	0.615 ± 0.007	0.109 ± 0.005	0.149 ± 0.014	0.322 ± 0.051
Weighted mixture | Mamba-3/Conv1D	0.640 ± 0.009	0.605 ± 0.007	0.105 ± 0.005	0.193 ± 0.032	0.352 ± 0.045
Weighted mixture | Transformer/Crossformer	0.643 ± 0.003	0.606 ± 0.004	0.105 ± 0.002	0.194 ± 0.010	0.328 ± 0.021
Weighted mixture | Transformer/Conv1D	0.644 ± 0.006	0.608 ± 0.006	0.106 ± 0.005	0.169 ± 0.019	0.342 ± 0.040
Tensor fusion | Mamba-3/Conv1D	0.638 ± 0.007	0.607 ± 0.006	0.105 ± 0.002	0.217 ± 0.018	0.363 ± 0.068
Tensor fusion | Transformer/Crossformer	0.640 ± 0.014	0.600 ± 0.014	0.104 ± 0.005	0.173 ± 0.042	0.369 ± 0.036
Tensor fusion | Transformer/Conv1D	0.639 ± 0.006	0.608 ± 0.010	0.106 ± 0.005	0.171 ± 0.016	0.359 ± 0.025
Co-attention | Transformer/Crossformer	0.634 ± 0.005	0.601 ± 0.004	0.103 ± 0.002	0.208 ± 0.033	0.386 ± 0.059

Table 64: Expanded MIMIC-III fusion grid — MIMIC-III. Updated export; values retain the supplied mean and SD precision. (Continued.)

Configuration	Val. AUROC	Test AUROC	Test AUPRC	ECE	AURC
Co-attention | Transformer/Conv1D	0.641 ± 0.005	0.608 ± 0.007	0.106 ± 0.005	0.157 ± 0.027	0.350 ± 0.026
Self-attention | Transformer/Crossformer	0.635 ± 0.008	0.606 ± 0.008	0.105 ± 0.003	0.217 ± 0.030	0.317 ± 0.035
Self-attention | Transformer/Conv1D	0.646 ± 0.005	0.612 ± 0.007	0.107 ± 0.003	0.181 ± 0.023	0.308 ± 0.035
Waveform only | Mamba-3/Conv1D	0.636 ± 0.005	0.605 ± 0.007	0.107 ± 0.004	0.192 ± 0.030	0.370 ± 0.051
Waveform only | Transformer/Crossformer	0.633 ± 0.004	0.605 ± 0.003	0.105 ± 0.002	0.196 ± 0.037	0.331 ± 0.049
Waveform only | Transformer/Conv1D	0.635 ± 0.005	0.606 ± 0.004	0.108 ± 0.003	0.178 ± 0.016	0.330 ± 0.040

Table 65: Expanded MIMIC-III fusion grid — MIMIC-III. Training-cost and failure measurements from the updated export; units are retained per cell.

Configuration	Time (s)	Memory	Failures (%)
Bilinear | Mamba-3/Conv1D	541.2	5050 MiB	0.0
Bilinear | Transformer/Crossformer	439.8	7251 MiB	0.0
Bilinear | Transformer/Conv1D	405.6	4462 MiB	0.0
Concat + product | Mamba-3/Conv1D	461.3	5022 MiB	0.0
Concat + product | Transformer/Crossformer	386.6	7250 MiB	0.0
Concat + product | Transformer/Conv1D	321.7	4458 MiB	0.0
EHR only | Mamba-3/Conv1D	–	–	–
EHR only | Transformer/Crossformer	–	–	–
EHR only | Transformer/Conv1D	–	–	–
Feature attention | Transformer/Crossformer	399.2	7238 MiB	0.0
Feature attention | Transformer/Conv1D	304.7	4457 MiB	0.0
FiLM | Mamba-3/Conv1D	473.1	5017 MiB	0.0
FiLM | Transformer/Crossformer	393.1	7236 MiB	0.0
FiLM | Transformer/Conv1D	305.0	4452 MiB	0.0
Gated | Mamba-3/Conv1D	471.3	5020 MiB	0.0
Gated | Transformer/Crossformer	395.4	7250 MiB	0.0
Gated | Transformer/Conv1D	315.8	4453 MiB	0.0
Weighted mixture | Mamba-3/Conv1D	486.4	5017 MiB	0.0
Weighted mixture | Transformer/Crossformer	389.3	7236 MiB	0.0
Weighted mixture | Transformer/Conv1D	350.3	4488 MiB	0.0
Tensor fusion | Mamba-3/Conv1D	484.6	5032 MiB	0.0
Tensor fusion | Transformer/Crossformer	396.7	7238 MiB	0.0
Tensor fusion | Transformer/Conv1D	329.4	4460 MiB	0.0
Co-attention | Transformer/Crossformer	400.4	6382 MiB	0.0

Table 65: Expanded MIMIC-III fusion grid — MIMIC-III. Training-cost and failure measurements from the updated export; units are retained per cell. (Continued.)

Configuration	Time (s)	Memory	Failures (%)
Co-attention | Transformer/Conv1D	315.3	4369 MiB	0.0
Self-attention | Transformer/Crossformer	360.4	8849 MiB	0.0
Self-attention | Transformer/Conv1D	257.2	4984 MiB	0.0
Waveform only | Mamba-3/Conv1D	–	–	–
Waveform only | Transformer/Crossformer	–	–	–
Waveform only | Transformer/Conv1D	–	–	–

Table 66: Channel selection — MIMIC-III. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

Acute det. / AKI 2/3 / Circ. det. / Fluid resp.

Condition / budget	Acute det.	AKI 2/3	Circ. det.	Fluid resp.
All channels (62)	0.584 ± 0.003	0.712 ± 0.012	0.678 ± 0.008	0.570 ± 0.009
Lasso (10)	0.590 ± 0.011	0.706 ± 0.019	0.667 ± 0.018	0.559 ± 0.010
MI (10)	0.571 ± 0.007	0.715 ± 0.019	0.644 ± 0.013	0.556 ± 0.015
Random (10)	0.571 ± 0.013	0.720 ± 0.016	0.544 ± 0.055	0.566 ± 0.014
Clinical (10)	0.603 ± 0.006	0.721 ± 0.009	0.683 ± 0.004	0.553 ± 0.012
SQI (10)	0.600 ± 0.003	0.723 ± 0.013	0.666 ± 0.006	0.558 ± 0.009
Temporal MI (10)	0.570 ± 0.014	0.715 ± 0.018	0.618 ± 0.061	0.560 ± 0.007

Dyn. mort. / Resp. esc. / Sepsis-3 / Vent. lib.

Condition / budget	Dyn. mort.	Resp. esc.	Sepsis-3	Vent. lib.
All channels (62)	0.556 ± 0.008	0.568 ± 0.012	0.631 ± 0.003	0.545 ± 0.003
Lasso (10)	0.561 ± 0.011	0.533 ± 0.025	0.615 ± 0.017	0.559 ± 0.018
MI (10)	0.513 ± 0.012	0.512 ± 0.017	0.555 ± 0.029	0.514 ± 0.013
Random (10)	0.507 ± 0.028	0.511 ± 0.017	0.536 ± 0.024	0.509 ± 0.007
Clinical (10)	0.558 ± 0.011	0.577 ± 0.006	0.623 ± 0.005	0.531 ± 0.010
SQI (10)	0.554 ± 0.008	0.589 ± 0.006	0.629 ± 0.002	0.539 ± 0.006
Temporal MI (10)	0.514 ± 0.027	0.526 ± 0.026	0.548 ± 0.033	0.511 ± 0.018

Table 67: Channel selection — VitalDB. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

VCS05 / CL07-002 / H5-002 / H5-030				
Condition / budget	VCS05	CL07-002	H5-002	H5-030
All channels (10)	0.877 ± 0.001	–	–	–
All channels (12)	–	0.538 ± 0.004	0.619 ± 0.004	0.593 ± 0.003
Lasso (2)	0.819 ± 0.008	0.507 ± 0.013	0.553 ± 0.012	0.562 ± 0.005
Lasso (4)	0.846 ± 0.008	0.524 ± 0.008	0.581 ± 0.011	0.573 ± 0.005
Lasso (8)	0.876 ± 0.002	0.534 ± 0.005	0.612 ± 0.004	0.588 ± 0.003
MI (2)	0.826 ± 0.010	0.511 ± 0.006	0.554 ± 0.009	0.545 ± 0.007
MI (4)	0.858 ± 0.006	0.517 ± 0.008	0.586 ± 0.008	0.563 ± 0.008
MI (8)	0.876 ± 0.001	0.536 ± 0.010	0.610 ± 0.007	0.586 ± 0.005
Random (2)	0.755 ± 0.035	0.509 ± 0.009	0.525 ± 0.007	0.548 ± 0.006
Random (4)	0.832 ± 0.011	0.519 ± 0.011	0.565 ± 0.012	0.568 ± 0.005
Random (8)	0.867 ± 0.002	0.530 ± 0.009	0.597 ± 0.010	0.584 ± 0.005
SQI (2)	0.809 ± 0.001	0.488 ± 0.006	0.501 ± 0.002	0.555 ± 0.002
SQI (4)	0.835 ± 0.002	0.552 ± 0.003	0.549 ± 0.008	0.542 ± 0.003
SQI (8)	0.877 ± 0.002	0.534 ± 0.002	0.602 ± 0.004	0.570 ± 0.001
Temporal MI (2)	0.817 ± 0.006	0.516 ± 0.008	0.569 ± 0.012	0.550 ± 0.004
Temporal MI (4)	0.851 ± 0.007	0.525 ± 0.011	0.591 ± 0.007	0.569 ± 0.007
Temporal MI (8)	0.877 ± 0.002	0.532 ± 0.008	0.610 ± 0.005	0.583 ± 0.004

Table 67: Channel selection — VitalDB. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid. (Continued.)

L2-30 / OX01-IOH / VA1-CO2 / VA1-STIM				
Condition / budget	L2-30	OX01-IOH	VA1-CO2	VA1-STIM
All channels (10)	–	–	–	–
All channels (12)	0.556 ± 0.006	0.668 ± 0.002	0.648 ± 0.001	0.665 ± 0.004
Lasso (2)	0.551 ± 0.006	0.577 ± 0.018	0.605 ± 0.012	0.597 ± 0.010
Lasso (4)	0.570 ± 0.007	0.608 ± 0.018	0.624 ± 0.006	0.625 ± 0.012
Lasso (8)	0.562 ± 0.012	0.651 ± 0.018	0.638 ± 0.004	0.649 ± 0.006
MI (2)	0.552 ± 0.007	0.588 ± 0.009	0.577 ± 0.012	0.615 ± 0.018
MI (4)	0.559 ± 0.010	0.615 ± 0.020	0.615 ± 0.005	0.635 ± 0.003
MI (8)	0.552 ± 0.006	0.652 ± 0.006	0.637 ± 0.006	0.656 ± 0.004
Random (2)	0.562 ± 0.026	0.551 ± 0.026	0.554 ± 0.007	0.573 ± 0.020
Random (4)	0.560 ± 0.015	0.574 ± 0.024	0.599 ± 0.010	0.593 ± 0.011
Random (8)	0.559 ± 0.008	0.628 ± 0.025	0.637 ± 0.007	0.633 ± 0.008
SQI (2)	0.566 ± 0.006	0.518 ± 0.001	0.610 ± 0.001	0.587 ± 0.006
SQI (4)	0.615 ± 0.005	0.664 ± 0.002	0.639 ± 0.002	0.594 ± 0.007
SQI (8)	0.560 ± 0.008	0.666 ± 0.003	0.633 ± 0.002	0.649 ± 0.004
Temporal MI (2)	0.560 ± 0.013	0.593 ± 0.028	0.579 ± 0.022	0.635 ± 0.013
Temporal MI (4)	0.561 ± 0.010	0.619 ± 0.015	0.616 ± 0.019	0.643 ± 0.003
Temporal MI (8)	0.568 ± 0.001	0.660 ± 0.008	0.641 ± 0.004	0.654 ± 0.006

Table 68: Missingness — MIMIC-III. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

Acute det. / AKI 2/3 / Circ. det. / Fluid resp.

Condition / budget	Acute det.	AKI 2/3	Circ. det.	Fluid resp.
Drifting	0.585 ± 0.003	0.686 ± 0.015	0.681 ± 0.006	0.559 ± 0.014
GRIN	0.601 ± 0.004	0.700 ± 0.013	0.690 ± 0.005	0.567 ± 0.006
GRU-D	0.599 ± 0.009	0.690 ± 0.014	0.689 ± 0.006	0.558 ± 0.006
Lead-agn. SSL	0.609 ± 0.003	0.690 ± 0.012	0.689 ± 0.004	0.557 ± 0.008
Pos. kriging	0.586 ± 0.006	0.690 ± 0.012	0.674 ± 0.004	0.565 ± 0.006
Raindrop	0.608 ± 0.009	0.675 ± 0.014	0.691 ± 0.005	0.563 ± 0.011
Rect. flow	0.585 ± 0.002	0.706 ± 0.020	0.679 ± 0.004	0.562 ± 0.009
TrigFlow	0.585 ± 0.003	0.704 ± 0.026	0.681 ± 0.001	0.563 ± 0.009
Zero-mask+Δ (ref.)	0.580 ± 0.005	0.706 ± 0.007	0.682 ± 0.004	0.571 ± 0.013

Dyn. mort. / Resp. esc. / Sepsis-3 / Vent. lib.

Condition / budget	Dyn. mort.	Resp. esc.	Sepsis-3	Vent. lib.
Drifting	0.552 ± 0.006	0.564 ± 0.006	0.629 ± 0.006	0.547 ± 0.005
GRIN	0.546 ± 0.008	0.555 ± 0.013	0.647 ± 0.004	0.549 ± 0.005
GRU-D	0.569 ± 0.006	0.568 ± 0.003	0.634 ± 0.006	0.558 ± 0.008
Lead-agn. SSL	0.554 ± 0.016	0.572 ± 0.009	0.641 ± 0.003	0.557 ± 0.007
Pos. kriging	0.556 ± 0.008	0.571 ± 0.009	0.634 ± 0.007	0.549 ± 0.003
Raindrop	0.549 ± 0.004	0.567 ± 0.007	0.641 ± 0.005	0.554 ± 0.008
Rect. flow	0.557 ± 0.007	0.559 ± 0.002	0.626 ± 0.007	0.547 ± 0.004
TrigFlow	0.552 ± 0.006	0.560 ± 0.003	0.629 ± 0.005	0.546 ± 0.002
Zero-mask+Δ (ref.)	0.550 ± 0.016	0.566 ± 0.011	0.631 ± 0.006	0.543 ± 0.008

Table 69: Missingness — VitalDB. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

CL07-002 / H5-002 / H5-030 / L2-30

Condition / budget	CL07-002	H5-002	H5-030	L2-30
Architecture only	0.536 ± 0.001	0.620 ± 0.004	0.589 ± 0.003	0.561 ± 0.006
Drifting	0.547 ± 0.004	0.556 ± 0.005	0.570 ± 0.001	0.565 ± 0.002
GRIN	0.549 ± 0.005	0.559 ± 0.006	0.573 ± 0.003	0.567 ± 0.005
GRU-D	0.547 ± 0.005	0.554 ± 0.009	0.568 ± 0.002	0.564 ± 0.007
Lead-agn. SSL	0.545 ± 0.003	0.544 ± 0.003	0.572 ± 0.003	0.570 ± 0.003
Pos. kriging	0.545 ± 0.006	0.543 ± 0.005	0.564 ± 0.004	0.566 ± 0.005
Raindrop	0.546 ± 0.004	0.611 ± 0.002	0.582 ± 0.004	0.558 ± 0.005
Rect. flow	0.547 ± 0.004	0.556 ± 0.006	0.568 ± 0.003	0.565 ± 0.006
TrigFlow	0.545 ± 0.004	0.561 ± 0.008	0.570 ± 0.003	0.565 ± 0.010
Zero-mask+Δ	0.547 ± 0.003	0.566 ± 0.006	0.568 ± 0.003	0.569 ± 0.006

OX01-IOH / VA1-CO2 / VA1-STIM

Condition / budget	OX01-IOH	VA1-CO2	VA1-STIM
Architecture only	0.672 ± 0.003	0.644 ± 0.003	0.668 ± 0.003
Drifting	0.666 ± 0.003	0.605 ± 0.006	0.602 ± 0.005
GRIN	0.664 ± 0.003	0.611 ± 0.003	0.601 ± 0.007
GRU-D	0.664 ± 0.002	0.616 ± 0.005	0.600 ± 0.006
Lead-agn. SSL	0.664 ± 0.002	0.609 ± 0.001	0.599 ± 0.004
Pos. kriging	0.665 ± 0.002	0.602 ± 0.001	0.585 ± 0.005
Raindrop	0.673 ± 0.004	0.645 ± 0.004	0.674 ± 0.001
Rect. flow	0.665 ± 0.002	0.609 ± 0.001	0.605 ± 0.005
TrigFlow	0.666 ± 0.002	0.614 ± 0.004	0.611 ± 0.005
Zero-mask+Δ	0.665 ± 0.001	0.620 ± 0.003	0.610 ± 0.005

Table 70: Ehr fusion — MIMIC-III. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

Acute det. / AKI 2/3 / Circ. det. / Fluid resp.

Condition / budget	Acute det.	AKI 2/3	Circ. det.	Fluid resp.
Bilinear	0.586 ± 0.012	0.729 ± 0.021	0.690 ± 0.005	0.569 ± 0.015
Concat	0.571 ± 0.012	0.715 ± 0.045	0.679 ± 0.008	0.559 ± 0.016
EHR only	0.577 ± 0.010	0.741 ± 0.026	0.662 ± 0.008	0.569 ± 0.025
Feature attn	0.580 ± 0.014	0.704 ± 0.026	0.675 ± 0.008	0.572 ± 0.002
FiLM	0.574 ± 0.007	0.698 ± 0.038	0.690 ± 0.010	0.570 ± 0.004
Gated	0.577 ± 0.013	0.698 ± 0.039	0.684 ± 0.001	0.568 ± 0.013
Late weighted	0.578 ± 0.006	0.706 ± 0.030	0.681 ± 0.003	0.564 ± 0.021
Tensor	0.576 ± 0.009	0.700 ± 0.034	0.680 ± 0.011	0.553 ± 0.012
Co-attention	0.586 ± 0.008	0.682 ± 0.037	0.680 ± 0.007	0.557 ± 0.019
Self-attention	0.584 ± 0.008	0.720 ± 0.027	0.683 ± 0.007	0.560 ± 0.018
Waveform only	0.587 ± 0.008	0.697 ± 0.042	0.688 ± 0.007	0.570 ± 0.013

Dyn. mort. / Resp. esc. / Sepsis-3 / Vent. lib.

Condition / budget	Dyn. mort.	Resp. esc.	Sepsis-3	Vent. lib.
Bilinear	0.554 ± 0.010	0.597 ± 0.018	0.657 ± 0.019	0.592 ± 0.037
Concat	0.532 ± 0.041	0.555 ± 0.009	0.658 ± 0.013	0.598 ± 0.040
EHR only	0.590 ± 0.012	0.598 ± 0.007	0.674 ± 0.005	0.646 ± 0.008
Feature attn	0.541 ± 0.018	0.595 ± 0.025	0.668 ± 0.007	0.579 ± 0.015
FiLM	0.543 ± 0.021	0.602 ± 0.010	0.659 ± 0.018	0.587 ± 0.025
Gated	0.543 ± 0.021	0.607 ± 0.009	0.666 ± 0.004	0.577 ± 0.028
Late weighted	0.544 ± 0.010	0.601 ± 0.013	0.626 ± 0.011	0.566 ± 0.006
Tensor	0.537 ± 0.025	0.596 ± 0.013	0.643 ± 0.015	0.579 ± 0.031
Co-attention	0.555 ± 0.004	0.561 ± 0.005	0.663 ± 0.010	0.581 ± 0.017
Self-attention	0.557 ± 0.014	0.559 ± 0.017	0.668 ± 0.002	0.569 ± 0.012
Waveform only	0.553 ± 0.014	0.575 ± 0.004	0.631 ± 0.009	0.550 ± 0.007

Table 71: Ehr fusion — VitalDB. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

CL07-002 / H5-002 / H5-030 / L2-30

Condition / budget	CL07-002	H5-002	H5-030	L2-30
EHR only	0.557 ± 0.005	0.754 ± 0.002	0.629 ± 0.001	0.737 ± 0.006
Co-attention	0.555 ± 0.008	0.675 ± 0.008	0.619 ± 0.007	0.681 ± 0.012
Gated	0.566 ± 0.009	0.757 ± 0.004	0.636 ± 0.004	0.728 ± 0.006
Late weighted	0.561 ± 0.009	0.750 ± 0.009	0.637 ± 0.002	0.739 ± 0.009
Concat	0.563 ± 0.007	0.761 ± 0.002	0.637 ± 0.002	0.732 ± 0.008
Tensor	0.561 ± 0.006	0.760 ± 0.003	0.633 ± 0.004	0.730 ± 0.008
Waveform only	0.539 ± 0.008	0.630 ± 0.005	0.601 ± 0.002	0.578 ± 0.018

OX01-IOH / VA1-CO2 / VA1-STIM / VCS05

Condition / budget	OX01-IOH	VA1-CO2	VA1-STIM	VCS05
EHR only	0.610 ± 0.003	0.731 ± 0.004	0.814 ± 0.004	0.929 ± 0.001
Co-attention	0.743 ± 0.003	0.692 ± 0.002	0.757 ± 0.020	0.921 ± 0.002
Gated	0.741 ± 0.004	0.737 ± 0.005	0.799 ± 0.003	0.931 ± 0.001
Late weighted	0.744 ± 0.001	0.737 ± 0.005	0.806 ± 0.007	0.928 ± 0.002
Concat	0.742 ± 0.003	0.739 ± 0.005	0.812 ± 0.004	0.931 ± 0.001
Tensor	0.739 ± 0.004	0.736 ± 0.004	0.796 ± 0.007	0.929 ± 0.002
Waveform only	0.731 ± 0.003	0.682 ± 0.005	0.675 ± 0.004	0.878 ± 0.001

Table 72: Ehr fusion — MC-MED. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid.

Decomp. 60m / Lactate NC / Pending hyperK / Pending anemia				
Condition / budget	Decomp. 60m	Lactate NC	Pending hyperK	Pending anemia
Co-attention	0.647 ± 0.002	0.497 ± 0.027	0.601 ± 0.007	0.666 ± 0.012
Concat	0.650 ± 0.005	0.481 ± 0.018	0.586 ± 0.022	0.638 ± 0.015
EHR only	0.624 ± 0.009	0.536 ± 0.022	0.601 ± 0.007	0.665 ± 0.011
Gated	0.651 ± 0.006	0.504 ± 0.020	0.600 ± 0.006	0.671 ± 0.011
Late	0.649 ± 0.007	0.520 ± 0.026	0.597 ± 0.010	0.671 ± 0.009
Waveform only	0.642 ± 0.004	0.508 ± 0.021	0.555 ± 0.029	0.563 ± 0.031

Table 73: Training subjects — MIMIC-III. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid. Budget values follow the plotted export; MIMIC-III uses training-window quotas.

Acute det. / AKI 2/3 / Circ. det. / Fluid resp.				
Condition / budget	Acute det.	AKI 2/3	Circ. det.	Fluid resp.
Diversity (10%)	0.558 ± 0.009	0.690 ± 0.022	0.609 ± 0.024	0.552 ± 0.014
Diversity (20%)	0.559 ± 0.007	0.706 ± 0.009	0.635 ± 0.018	0.567 ± 0.008
Diversity (30%)	0.562 ± 0.007	0.677 ± 0.021	0.655 ± 0.013	0.563 ± 0.001
Div. cluster (10%)	0.543 ± 0.007	0.679 ± 0.023	0.664 ± 0.003	0.571 ± 0.003
Div. cluster (20%)	0.541 ± 0.013	0.683 ± 0.013	0.673 ± 0.006	0.565 ± 0.009
Div. cluster (30%)	0.546 ± 0.006	0.676 ± 0.014	0.676 ± 0.002	0.564 ± 0.011
Full cohort (100%)	0.584 ± 0.003	0.712 ± 0.012	0.678 ± 0.008	0.570 ± 0.009
Grad. match (10%)	0.560 ± 0.009	0.728 ± 0.007	0.629 ± 0.010	0.565 ± 0.004
Grad. match (20%)	0.557 ± 0.003	0.731 ± 0.007	0.605 ± 0.014	0.566 ± 0.008
Grad. match (30%)	0.538 ± 0.008	0.730 ± 0.001	0.604 ± 0.007	0.538 ± 0.004
NLDR (10%)	0.562 ± 0.006	0.665 ± 0.013	0.660 ± 0.004	0.556 ± 0.011
NLDR (20%)	0.557 ± 0.005	0.694 ± 0.013	0.670 ± 0.005	0.570 ± 0.007
NLDR (30%)	0.558 ± 0.005	0.697 ± 0.023	0.668 ± 0.006	0.565 ± 0.006
Noise filter (10%)	0.552 ± 0.006	0.669 ± 0.014	0.635 ± 0.012	0.558 ± 0.006
Noise filter (20%)	0.590 ± 0.006	0.671 ± 0.018	0.655 ± 0.009	0.562 ± 0.005
Noise filter (30%)	0.589 ± 0.003	0.676 ± 0.022	0.663 ± 0.005	0.570 ± 0.011
Random (10%)	0.554 ± 0.004	0.682 ± 0.023	0.643 ± 0.019	0.548 ± 0.013
Random (20%)	0.557 ± 0.010	0.697 ± 0.016	0.659 ± 0.010	0.554 ± 0.023
Random (30%)	0.564 ± 0.006	0.698 ± 0.023	0.671 ± 0.013	0.555 ± 0.012
Task anchor (10%)	0.496 ± 0.010	0.538 ± 0.020	0.475 ± 0.019	0.542 ± 0.014
Task anchor (20%)	0.488 ± 0.011	0.552 ± 0.015	0.463 ± 0.018	0.524 ± 0.013
Task anchor (30%)	0.486 ± 0.004	0.527 ± 0.012	0.424 ± 0.009	0.516 ± 0.005
Train. dyn. (10%)	0.543 ± 0.015	0.611 ± 0.008	0.663 ± 0.004	0.580 ± 0.003
Train. dyn. (20%)	0.540 ± 0.003	0.617 ± 0.009	0.666 ± 0.009	0.559 ± 0.007
Train. dyn. (30%)	0.536 ± 0.000	0.615 ± 0.006	0.670 ± 0.005	0.559 ± 0.006

Table 73: Training subjects — MIMIC-III. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid. Budget values follow the plotted export; MIMIC-III uses training-window quotas. (Continued.)

Dyn. mort. / Resp. esc. / Sepsis-3 / Vent. lib.

Condition / budget	Dyn. mort.	Resp. esc.	Sepsis-3	Vent. lib.
Diversity (10%)	0.537 ± 0.009	0.504 ± 0.026	0.591 ± 0.008	0.526 ± 0.016
Diversity (20%)	0.538 ± 0.016	0.500 ± 0.020	0.620 ± 0.006	0.538 ± 0.017
Diversity (30%)	0.544 ± 0.010	0.528 ± 0.027	0.626 ± 0.010	0.536 ± 0.006
Div. cluster (10%)	0.543 ± 0.012	0.516 ± 0.017	0.584 ± 0.005	0.521 ± 0.008
Div. cluster (20%)	0.546 ± 0.006	0.518 ± 0.016	0.600 ± 0.010	0.534 ± 0.008
Div. cluster (30%)	0.549 ± 0.009	0.536 ± 0.019	0.609 ± 0.006	0.538 ± 0.004
Full cohort (100%)	0.556 ± 0.008	0.568 ± 0.012	0.631 ± 0.003	0.545 ± 0.003
Grad. match (10%)	0.530 ± 0.014	0.531 ± 0.014	0.566 ± 0.020	0.505 ± 0.011
Grad. match (20%)	0.517 ± 0.024	0.534 ± 0.012	0.583 ± 0.019	0.512 ± 0.024
Grad. match (30%)	0.523 ± 0.019	0.541 ± 0.016	0.580 ± 0.018	0.512 ± 0.021
NLDR (10%)	0.536 ± 0.013	0.533 ± 0.008	0.584 ± 0.012	0.540 ± 0.008
NLDR (20%)	0.531 ± 0.016	0.541 ± 0.008	0.600 ± 0.010	0.547 ± 0.003
NLDR (30%)	0.545 ± 0.007	0.515 ± 0.011	0.608 ± 0.007	0.545 ± 0.003
Noise filter (10%)	0.541 ± 0.006	0.537 ± 0.003	0.575 ± 0.013	0.512 ± 0.014
Noise filter (20%)	0.534 ± 0.005	0.525 ± 0.012	0.584 ± 0.008	0.512 ± 0.009
Noise filter (30%)	0.541 ± 0.007	0.529 ± 0.004	0.594 ± 0.008	0.501 ± 0.020
Random (10%)	0.535 ± 0.013	0.513 ± 0.019	0.597 ± 0.005	0.523 ± 0.011
Random (20%)	0.543 ± 0.007	0.529 ± 0.013	0.610 ± 0.019	0.537 ± 0.010
Random (30%)	0.541 ± 0.016	0.539 ± 0.018	0.610 ± 0.006	0.541 ± 0.006
Task anchor (10%)	0.483 ± 0.007	0.512 ± 0.003	0.465 ± 0.012	0.524 ± 0.005
Task anchor (20%)	0.463 ± 0.008	0.511 ± 0.005	0.477 ± 0.001	0.531 ± 0.003
Task anchor (30%)	0.465 ± 0.006	0.505 ± 0.002	0.513 ± 0.007	0.530 ± 0.002
Train. dyn. (10%)	0.529 ± 0.007	0.469 ± 0.005	0.509 ± 0.016	0.512 ± 0.004
Train. dyn. (20%)	0.543 ± 0.003	0.505 ± 0.015	0.514 ± 0.016	0.505 ± 0.004
Train. dyn. (30%)	0.559 ± 0.004	0.481 ± 0.007	0.504 ± 0.014	0.515 ± 0.008

Table 74: Training subjects — VitalDB. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid. Budget values follow the plotted export; MIMIC-III uses training-window quotas.

CL07-002 / H5-002 / H5-030 / L2-30

Condition / budget	CL07-002	H5-002	H5-030	L2-30
Diversity (5%)	0.496 ± 0.005	0.561 ± 0.006	0.540 ± 0.006	0.507 ± 0.004
Diversity (10%)	0.517 ± 0.006	0.566 ± 0.007	0.546 ± 0.004	0.527 ± 0.006
Diversity (20%)	0.516 ± 0.002	0.572 ± 0.004	0.572 ± 0.003	0.521 ± 0.006
Div. cluster (5%)	0.515 ± 0.005	0.549 ± 0.004	0.531 ± 0.004	0.527 ± 0.005
Div. cluster (10%)	0.513 ± 0.002	0.556 ± 0.004	0.537 ± 0.006	0.548 ± 0.004
Div. cluster (20%)	0.521 ± 0.005	0.573 ± 0.003	0.542 ± 0.006	0.568 ± 0.005
Full cohort (100%)	0.538 ± 0.003	0.617 ± 0.005	0.591 ± 0.001	0.560 ± 0.006
Grad. match (5%)	0.538 ± 0.001	0.572 ± 0.005	0.551 ± 0.004	0.537 ± 0.006
Grad. match (10%)	0.535 ± 0.004	0.582 ± 0.007	0.552 ± 0.002	0.551 ± 0.006
Grad. match (20%)	0.521 ± 0.006	0.618 ± 0.003	0.582 ± 0.004	0.527 ± 0.006
NLDR (5%)	0.521 ± 0.003	0.546 ± 0.005	0.526 ± 0.003	0.524 ± 0.006
NLDR (10%)	0.522 ± 0.008	0.570 ± 0.005	0.531 ± 0.002	0.538 ± 0.007
NLDR (20%)	0.517 ± 0.004	0.574 ± 0.004	0.538 ± 0.006	0.549 ± 0.004
Noise filter (5%)	0.538 ± 0.004	0.534 ± 0.006	0.528 ± 0.002	0.507 ± 0.009
Noise filter (10%)	0.531 ± 0.002	0.533 ± 0.003	0.525 ± 0.004	0.520 ± 0.006
Noise filter (20%)	0.529 ± 0.006	0.569 ± 0.002	0.562 ± 0.003	0.573 ± 0.007
Random (5%)	0.519 ± 0.009	0.554 ± 0.012	0.535 ± 0.013	0.528 ± 0.017
Random (10%)	0.526 ± 0.016	0.561 ± 0.020	0.538 ± 0.013	0.540 ± 0.025
Random (20%)	0.528 ± 0.013	0.574 ± 0.012	0.555 ± 0.008	0.550 ± 0.018
Task anchor (5%)	0.516 ± 0.005	0.536 ± 0.003	0.548 ± 0.003	0.537 ± 0.005
Task anchor (10%)	0.521 ± 0.003	0.551 ± 0.005	0.561 ± 0.003	0.554 ± 0.006
Task anchor (20%)	0.536 ± 0.003	0.561 ± 0.006	0.565 ± 0.004	0.547 ± 0.005
Train. dyn. (5%)	0.539 ± 0.003	0.546 ± 0.006	0.520 ± 0.003	0.528 ± 0.009
Train. dyn. (10%)	0.535 ± 0.006	0.557 ± 0.007	0.525 ± 0.005	0.559 ± 0.006
Train. dyn. (20%)	0.537 ± 0.005	0.564 ± 0.003	0.525 ± 0.003	0.576 ± 0.004

Table 74: Training subjects — VitalDB. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid. Budget values follow the plotted export; MIMIC-III uses training-window quotas. (Continued.)

OX01-IOH / VA1-CO2 / VA1-STIM / VCS05

Condition / budget	OX01-IOH	VA1-CO2	VA1-STIM	VCS05
Diversity (5%)	0.601 ± 0.006	0.585 ± 0.003	0.597 ± 0.005	0.823 ± 0.004
Diversity (10%)	0.623 ± 0.004	0.596 ± 0.004	0.612 ± 0.003	0.852 ± 0.000
Diversity (20%)	0.628 ± 0.002	0.620 ± 0.004	0.645 ± 0.006	0.859 ± 0.002
Div. cluster (5%)	0.608 ± 0.003	0.590 ± 0.002	0.603 ± 0.005	0.839 ± 0.002
Div. cluster (10%)	0.619 ± 0.003	0.605 ± 0.004	0.624 ± 0.007	0.851 ± 0.001
Div. cluster (20%)	0.632 ± 0.003	0.621 ± 0.003	0.633 ± 0.007	0.861 ± 0.002
Full cohort (100%)	0.674 ± 0.002	0.643 ± 0.002	0.665 ± 0.006	0.878 ± 0.001
Grad. match (5%)	0.577 ± 0.003	0.553 ± 0.004	0.629 ± 0.006	0.862 ± 0.001
Grad. match (10%)	0.607 ± 0.002	0.581 ± 0.007	0.644 ± 0.003	0.871 ± 0.002
Grad. match (20%)	0.626 ± 0.001	0.618 ± 0.002	0.667 ± 0.004	0.873 ± 0.002
NLDR (5%)	0.599 ± 0.003	0.587 ± 0.002	0.595 ± 0.001	0.841 ± 0.002
NLDR (10%)	0.611 ± 0.003	0.605 ± 0.002	0.630 ± 0.003	0.856 ± 0.002
NLDR (20%)	0.633 ± 0.003	0.625 ± 0.002	0.657 ± 0.006	0.862 ± 0.002
Noise filter (5%)	0.580 ± 0.002	0.573 ± 0.005	0.637 ± 0.005	0.827 ± 0.002
Noise filter (10%)	0.609 ± 0.004	0.588 ± 0.002	0.661 ± 0.003	0.854 ± 0.001
Noise filter (20%)	0.631 ± 0.004	0.614 ± 0.002	0.675 ± 0.002	0.862 ± 0.001
Random (5%)	0.586 ± 0.014	0.583 ± 0.015	0.612 ± 0.012	0.845 ± 0.007
Random (10%)	0.606 ± 0.013	0.605 ± 0.009	0.633 ± 0.010	0.856 ± 0.007
Random (20%)	0.624 ± 0.009	0.622 ± 0.011	0.642 ± 0.018	0.864 ± 0.003
Task anchor (5%)	0.563 ± 0.003	0.604 ± 0.001	0.610 ± 0.004	0.834 ± 0.002
Task anchor (10%)	0.597 ± 0.004	0.614 ± 0.003	0.629 ± 0.004	0.845 ± 0.002
Task anchor (20%)	0.621 ± 0.004	0.629 ± 0.003	0.640 ± 0.005	0.859 ± 0.002
Train. dyn. (5%)	0.622 ± 0.002	0.591 ± 0.001	0.631 ± 0.004	0.820 ± 0.001
Train. dyn. (10%)	0.622 ± 0.003	0.606 ± 0.002	0.637 ± 0.004	0.831 ± 0.003
Train. dyn. (20%)	0.625 ± 0.002	0.623 ± 0.003	0.644 ± 0.005	0.845 ± 0.002

Table 75: Training subjects — MC-MED. Task-level test AUROC (mean ± seed SD) underlying the AUROC figures. Dashes mark combinations absent from that plotted grid. Budget values follow the plotted export; MIMIC-III uses training-window quotas.

Decomp. 60m / Lactate NC / Pending hyperK / Pending anemia

Condition / budget	Decomp. 60m	Lactate NC	Pending hyperK	Pending anemia
Diversity (10%)	0.624 ± 0.013	0.516 ± 0.018	0.496 ± 0.051	0.488 ± 0.027
Diversity (20%)	0.629 ± 0.015	0.502 ± 0.014	0.536 ± 0.025	0.550 ± 0.030
Diversity (50%)	0.637 ± 0.010	0.502 ± 0.029	0.557 ± 0.029	0.573 ± 0.011
Div. cluster (10%)	0.626 ± 0.013	0.514 ± 0.016	0.484 ± 0.058	0.569 ± 0.010
Div. cluster (20%)	0.629 ± 0.009	0.504 ± 0.019	0.488 ± 0.042	0.595 ± 0.011
Div. cluster (50%)	0.634 ± 0.007	0.491 ± 0.018	0.541 ± 0.018	0.569 ± 0.026
Full cohort (100%)	0.642 ± 0.004	0.508 ± 0.021	0.555 ± 0.029	0.563 ± 0.031
Grad. match (10%)	0.547 ± 0.004	0.490 ± 0.007	0.554 ± 0.045	0.488 ± 0.022
Grad. match (20%)	0.542 ± 0.023	0.510 ± 0.019	0.483 ± 0.032	0.472 ± 0.021
Grad. match (50%)	0.590 ± 0.009	0.534 ± 0.009	0.515 ± 0.044	0.531 ± 0.016
NLDR (10%)	0.612 ± 0.011	0.488 ± 0.020	0.480 ± 0.054	0.545 ± 0.008
NLDR (20%)	0.626 ± 0.007	0.479 ± 0.024	0.502 ± 0.042	0.541 ± 0.025
NLDR (50%)	0.639 ± 0.010	0.486 ± 0.025	0.558 ± 0.029	0.552 ± 0.019
Noise filter (10%)	0.602 ± 0.017	0.508 ± 0.010	0.508 ± 0.047	0.547 ± 0.022
Noise filter (20%)	0.603 ± 0.009	0.492 ± 0.025	0.551 ± 0.028	0.554 ± 0.024
Noise filter (50%)	0.597 ± 0.011	0.519 ± 0.015	0.527 ± 0.018	0.566 ± 0.012
Random (10%)	0.617 ± 0.018	0.496 ± 0.013	0.528 ± 0.051	0.543 ± 0.028
Random (20%)	0.625 ± 0.011	0.503 ± 0.019	0.488 ± 0.042	0.546 ± 0.028
Random (50%)	0.631 ± 0.015	0.513 ± 0.021	0.546 ± 0.044	0.547 ± 0.019
Task anchor (10%)	0.513 ± 0.010	0.500 ± 0.009	0.437 ± 0.007	0.510 ± 0.015
Task anchor (20%)	0.516 ± 0.018	0.516 ± 0.023	0.441 ± 0.018	0.504 ± 0.025
Task anchor (50%)	0.576 ± 0.016	0.491 ± 0.015	0.512 ± 0.020	0.494 ± 0.009
Train. dyn. (10%)	0.627 ± 0.007	0.491 ± 0.010	0.534 ± 0.023	0.594 ± 0.012
Train. dyn. (20%)	0.629 ± 0.005	0.494 ± 0.018	0.510 ± 0.035	0.601 ± 0.013
Train. dyn. (50%)	0.636 ± 0.004	0.498 ± 0.009	0.563 ± 0.024	0.581 ± 0.019

Table 76: Transfer summaries — MIMIC-III → MIMIC-IV. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC
Full fine-tuning	0.678 ± 0.007	0.568 ± 0.006	0.357 ± 0.007
Head only	0.638 ± 0.011	0.585 ± 0.015	0.359 ± 0.018
Last block + head	0.668 ± 0.013	0.580 ± 0.010	0.357 ± 0.011
Random init	0.665 ± 0.010	0.561 ± 0.007	0.341 ± 0.007

Table 77: Transfer summaries — VitalDB → MIMIC-IV. Updated export; values retain the supplied mean and SD precision.

Configuration	Val. AUROC	Test AUROC	Test AUPRC
Full fine-tuning	0.655 ± 0.007	0.557 ± 0.016	0.358 ± 0.013
Head only	0.656 ± 0.007	0.565 ± 0.012	0.371 ± 0.013
Last block + head	0.669 ± 0.014	0.572 ± 0.005	0.372 ± 0.007
Random init	0.664 ± 0.019	0.574 ± 0.015	0.376 ± 0.015

Table 78: MIMIC-IV transfer AUROC for Figure 36. Original inner-OOF validation, original selected-inner-fit test, revised adaptation-refit test, and feature-control test results are separated. Values are means ± seed SD; EHR-only uses patient-bootstrap SD.

Method	Original val.	Original test	Revised test	Control test
MIMIC-III source; Acute det.				
Full fine-tuning	0.726 ± 0.017	0.548 ± 0.002	0.544 ± 0.016	–
Last block + head	0.717 ± 0.014	0.533 ± 0.010	0.541 ± 0.021	–
Head only	0.695 ± 0.020	0.547 ± 0.013	0.541 ± 0.022	–
Random init.	0.731 ± 0.026	0.543 ± 0.011	0.555 ± 0.020	–
EHR only	–	–	–	0.587 ± 0.042
Waveform features	–	–	–	0.533 ± 0.011
EHR + waveform features	–	–	–	0.582 ± 0.009
MIMIC-III source; AKI 2/3				
Full fine-tuning	0.612 ± 0.070	0.410 ± 0.017	0.439 ± 0.018	–
Last block + head	0.565 ± 0.059	0.440 ± 0.022	0.451 ± 0.033	–
Head only	0.551 ± 0.075	0.445 ± 0.025	0.460 ± 0.032	–
Random init.	0.577 ± 0.024	0.405 ± 0.017	0.401 ± 0.020	–
EHR only	–	–	–	0.525 ± 0.054
Waveform features	–	–	–	0.446 ± 0.028
EHR + waveform features	–	–	–	0.562 ± 0.026
MIMIC-III source; Circ. det.				
Full fine-tuning	0.730 ± 0.019	0.677 ± 0.022	0.708 ± 0.018	–
Last block + head	0.716 ± 0.025	0.695 ± 0.028	0.715 ± 0.029	–
Head only	0.682 ± 0.028	0.708 ± 0.025	0.722 ± 0.026	–
Random init.	0.711 ± 0.014	0.640 ± 0.008	0.684 ± 0.017	–
EHR only	–	–	–	0.662 ± 0.038
Waveform features	–	–	–	0.636 ± 0.036
EHR + waveform features	–	–	–	0.663 ± 0.017
MIMIC-III source; Delirium / neuro				
Full fine-tuning	0.631 ± 0.019	0.531 ± 0.014	0.533 ± 0.013	–
Last block + head	0.658 ± 0.039	0.524 ± 0.021	0.535 ± 0.034	–
Head only	0.614 ± 0.034	0.554 ± 0.018	0.558 ± 0.027	–
Random init.	0.622 ± 0.032	0.536 ± 0.004	0.528 ± 0.014	–
EHR only	–	–	–	0.646 ± 0.042
Waveform features	–	–	–	0.510 ± 0.010
EHR + waveform features	–	–	–	0.627 ± 0.015
MIMIC-III source; Future SOFA				
Full fine-tuning	0.692 ± 0.031	0.616 ± 0.021	0.618 ± 0.037	–
Last block + head	0.686 ± 0.007	0.627 ± 0.014	0.655 ± 0.024	–
Head only	0.650 ± 0.019	0.623 ± 0.017	0.645 ± 0.036	–
Random init.	0.682 ± 0.022	0.617 ± 0.021	0.636 ± 0.024	–
EHR only	–	–	–	0.689 ± 0.038

Table 78: MIMIC-IV transfer AUROC for Figure 36. Original inner-OOF validation, original selected-inner-fit test, revised adaptation-refit test, and feature-control test results are separated. Values are means ± seed SD; EHR-only uses patient-bootstrap SD. (Continued.)

Method	Original val.	Original test	Revised test	Control test
Waveform features	–	–	–	0.682 ± 0.018
EHR + waveform features	–	–	–	0.689 ± 0.017
VitalDB source; Acute det.				
Full fine-tuning	0.737 ± 0.024	0.547 ± 0.018	0.547 ± 0.014	–
Last block + head	0.728 ± 0.022	0.550 ± 0.015	0.556 ± 0.026	–
Head only	0.728 ± 0.028	0.551 ± 0.019	0.556 ± 0.020	–
Random init.	0.739 ± 0.023	0.557 ± 0.019	0.563 ± 0.006	–
EHR only	–	–	–	0.587 ± 0.042
Waveform features	–	–	–	0.519 ± 0.029
EHR + waveform features	–	–	–	0.569 ± 0.011
VitalDB source; AKI 2/3				
Full fine-tuning	0.480 ± 0.039	0.404 ± 0.019	0.410 ± 0.021	–
Last block + head	0.480 ± 0.037	0.431 ± 0.018	0.445 ± 0.013	–
Head only	0.472 ± 0.044	0.434 ± 0.029	0.439 ± 0.024	–
Random init.	0.509 ± 0.045	0.392 ± 0.011	0.406 ± 0.023	–
EHR only	–	–	–	0.525 ± 0.054
Waveform features	–	–	–	0.380 ± 0.031
EHR + waveform features	–	–	–	0.530 ± 0.018
VitalDB source; Circ. det.				
Full fine-tuning	0.709 ± 0.042	0.648 ± 0.009	0.654 ± 0.014	–
Last block + head	0.735 ± 0.027	0.661 ± 0.011	0.664 ± 0.029	–
Head only	0.741 ± 0.012	0.654 ± 0.021	0.655 ± 0.031	–
Random init.	0.702 ± 0.028	0.671 ± 0.010	0.685 ± 0.027	–
EHR only	–	–	–	0.662 ± 0.038
Waveform features	–	–	–	0.620 ± 0.017
EHR + waveform features	–	–	–	0.692 ± 0.008
VitalDB source; Delirium / neuro				
Full fine-tuning	0.640 ± 0.042	0.534 ± 0.011	0.529 ± 0.021	–
Last block + head	0.684 ± 0.050	0.536 ± 0.014	0.536 ± 0.006	–
Head only	0.653 ± 0.049	0.534 ± 0.015	0.533 ± 0.018	–
Random init.	0.652 ± 0.047	0.552 ± 0.008	0.568 ± 0.028	–
EHR only	–	–	–	0.646 ± 0.042
Waveform features	–	–	–	0.525 ± 0.015
EHR + waveform features	–	–	–	0.656 ± 0.013
VitalDB source; Future SOFA				
Full fine-tuning	0.710 ± 0.006	0.643 ± 0.019	0.644 ± 0.034	–
Last block + head	0.719 ± 0.036	0.650 ± 0.014	0.660 ± 0.015	–
Head only	0.688 ± 0.026	0.641 ± 0.015	0.639 ± 0.028	–

Table 78: MIMIC-IV transfer AUROC for Figure 36. Original inner-OOF validation, original selected-inner-fit test, revised adaptation-refit test, and feature-control test results are separated. Values are means ± seed SD; EHR-only uses patient-bootstrap SD. (Continued.)

Method	Original val.	Original test	Revised test	Control test
Random init.	0.717 ± 0.023	0.639 ± 0.006	0.649 ± 0.024	–
EHR only	–	–	–	0.689 ± 0.038
Waveform features	–	–	–	0.685 ± 0.016
EHR + waveform features	–	–	–	0.700 ± 0.017

Table 79: MIMIC-IV transfer AUPRC for Figure 37. Original inner-OOF validation, original selected-inner-fit test, revised adaptation-refit test, and feature-control test results are separated. Values are means ± seed SD; EHR-only uses patient-bootstrap SD.

Method	Original val.	Original test	Revised test	Control test
MIMIC-III source; Acute det.				
Full fine-tuning	0.594 ± 0.068	0.408 ± 0.007	0.409 ± 0.009	–
Last block + head	0.561 ± 0.016	0.396 ± 0.015	0.402 ± 0.027	–
Head only	0.565 ± 0.050	0.407 ± 0.008	0.396 ± 0.019	–
Random init.	0.606 ± 0.042	0.413 ± 0.014	0.413 ± 0.012	–
EHR only	–	–	–	0.451 ± 0.067
Waveform features	–	–	–	0.400 ± 0.006
EHR + waveform features	–	–	–	0.460 ± 0.006
MIMIC-III source; AKI 2/3				
Full fine-tuning	0.543 ± 0.069	0.432 ± 0.015	0.458 ± 0.013	–
Last block + head	0.528 ± 0.078	0.447 ± 0.023	0.443 ± 0.034	–
Head only	0.503 ± 0.099	0.450 ± 0.026	0.447 ± 0.029	–
Random init.	0.540 ± 0.024	0.436 ± 0.011	0.434 ± 0.021	–
EHR only	–	–	–	0.494 ± 0.081
Waveform features	–	–	–	0.449 ± 0.023
EHR + waveform features	–	–	–	0.503 ± 0.019
MIMIC-III source; Circ. det.				
Full fine-tuning	0.415 ± 0.048	0.354 ± 0.026	0.365 ± 0.034	–
Last block + head	0.370 ± 0.031	0.369 ± 0.055	0.382 ± 0.057	–
Head only	0.348 ± 0.036	0.374 ± 0.053	0.381 ± 0.064	–
Random init.	0.375 ± 0.020	0.291 ± 0.005	0.304 ± 0.015	–
EHR only	–	–	–	0.333 ± 0.061
Waveform features	–	–	–	0.335 ± 0.025
EHR + waveform features	–	–	–	0.334 ± 0.016
MIMIC-III source; Delirium / neuro				
Full fine-tuning	0.376 ± 0.029	0.292 ± 0.008	0.293 ± 0.010	–
Last block + head	0.393 ± 0.060	0.291 ± 0.014	0.299 ± 0.025	–
Head only	0.318 ± 0.026	0.312 ± 0.020	0.318 ± 0.024	–
Random init.	0.378 ± 0.039	0.302 ± 0.005	0.292 ± 0.007	–
EHR only	–	–	–	0.393 ± 0.068
Waveform features	–	–	–	0.296 ± 0.016
EHR + waveform features	–	–	–	0.382 ± 0.009
MIMIC-III source; Future SOFA				
Full fine-tuning	0.308 ± 0.026	0.270 ± 0.015	0.259 ± 0.027	–
Last block + head	0.285 ± 0.017	0.259 ± 0.021	0.261 ± 0.025	–
Head only	0.288 ± 0.030	0.255 ± 0.016	0.254 ± 0.017	–
Random init.	0.306 ± 0.041	0.262 ± 0.011	0.262 ± 0.012	–
EHR only	–	–	–	0.348 ± 0.063

Table 79: MIMIC-IV transfer AUPRC for Figure 37. Original inner-OOF validation, original selected-inner-fit test, revised adaptation-refit test, and feature-control test results are separated. Values are means ± seed SD; EHR-only uses patient-bootstrap SD. (Continued.)

Method	Original val.	Original test	Revised test	Control test
Waveform features	–	–	–	0.324 ± 0.031
EHR + waveform features	–	–	–	0.338 ± 0.009
VitalDB source; Acute det.				
Full fine-tuning	0.602 ± 0.039	0.428 ± 0.021	0.430 ± 0.024	–
Last block + head	0.614 ± 0.041	0.429 ± 0.006	0.433 ± 0.020	–
Head only	0.597 ± 0.022	0.432 ± 0.016	0.437 ± 0.018	–
Random init.	0.592 ± 0.019	0.438 ± 0.009	0.448 ± 0.009	–
EHR only	–	–	–	0.451 ± 0.067
Waveform features	–	–	–	0.396 ± 0.011
EHR + waveform features	–	–	–	0.448 ± 0.010
VitalDB source; AKI 2/3				
Full fine-tuning	0.445 ± 0.035	0.404 ± 0.014	0.401 ± 0.015	–
Last block + head	0.473 ± 0.025	0.429 ± 0.017	0.435 ± 0.025	–
Head only	0.454 ± 0.035	0.444 ± 0.024	0.442 ± 0.022	–
Random init.	0.483 ± 0.055	0.406 ± 0.008	0.409 ± 0.024	–
EHR only	–	–	–	0.494 ± 0.081
Waveform features	–	–	–	0.411 ± 0.027
EHR + waveform features	–	–	–	0.495 ± 0.011
VitalDB source; Circ. det.				
Full fine-tuning	0.398 ± 0.068	0.347 ± 0.011	0.349 ± 0.013	–
Last block + head	0.413 ± 0.051	0.356 ± 0.010	0.366 ± 0.017	–
Head only	0.408 ± 0.039	0.354 ± 0.017	0.366 ± 0.024	–
Random init.	0.427 ± 0.033	0.348 ± 0.011	0.363 ± 0.022	–
EHR only	–	–	–	0.333 ± 0.061
Waveform features	–	–	–	0.319 ± 0.019
EHR + waveform features	–	–	–	0.367 ± 0.008
VitalDB source; Delirium / neuro				
Full fine-tuning	0.418 ± 0.046	0.314 ± 0.014	0.313 ± 0.022	–
Last block + head	0.399 ± 0.060	0.312 ± 0.013	0.311 ± 0.014	–
Head only	0.404 ± 0.053	0.317 ± 0.013	0.309 ± 0.016	–
Random init.	0.398 ± 0.039	0.326 ± 0.009	0.350 ± 0.029	–
EHR only	–	–	–	0.393 ± 0.068
Waveform features	–	–	–	0.328 ± 0.022
EHR + waveform features	–	–	–	0.420 ± 0.018
VitalDB source; Future SOFA				
Full fine-tuning	0.330 ± 0.025	0.311 ± 0.009	0.297 ± 0.022	–
Last block + head	0.347 ± 0.050	0.312 ± 0.005	0.315 ± 0.004	–
Head only	0.332 ± 0.064	0.304 ± 0.011	0.299 ± 0.018	–

Table 79: MIMIC-IV transfer AUPRC for Figure 37. Original inner-OOF validation, original selected-inner-fit test, revised adaptation-refit test, and feature-control test results are separated. Values are means ± seed SD; EHR-only uses patient-bootstrap SD. (Continued.)

Method	Original val.	Original test	Revised test	Control test
Random init.	0.372 ± 0.038	0.313 ± 0.007	0.311 ± 0.024	–
EHR only	–	–	–	0.348 ± 0.063
Waveform features	–	–	–	0.372 ± 0.026
EHR + waveform features	–	–	–	0.348 ± 0.016

Table 80: Architectures — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
DLinear	Acute det.	0.039	0.119	0.328
FITS	Acute det.	0.038	0.192	0.345
ResNet1D	Acute det.	0.039	0.092	0.299
InceptionTime	Acute det.	0.040	0.096	0.290
TCN	Acute det.	0.041	0.107	0.296
BiLSTM	Acute det.	0.038	0.122	0.289
Transformer	Acute det.	0.040	0.131	0.321
S4	Acute det.	0.038	0.131	0.347
Mamba-3	Acute det.	0.041	0.157	0.356
DLinear	Dyn. mort.	0.017	0.242	0.463
FITS	Dyn. mort.	0.016	0.302	0.220
ResNet1D	Dyn. mort.	0.021	0.182	0.338
InceptionTime	Dyn. mort.	0.025	0.153	0.303
TCN	Dyn. mort.	0.025	0.123	0.266
BiLSTM	Dyn. mort.	0.018	0.161	0.172
Transformer	Dyn. mort.	0.017	0.304	0.392
S4	Dyn. mort.	0.017	0.171	0.283
Mamba-3	Dyn. mort.	0.016	0.380	0.646
DLinear	Resp. esc.	0.033	0.282	0.305
FITS	Resp. esc.	0.035	0.262	0.498
ResNet1D	Resp. esc.	0.037	0.219	0.373
InceptionTime	Resp. esc.	0.037	0.228	0.376
TCN	Resp. esc.	0.034	0.170	0.339
BiLSTM	Resp. esc.	0.037	0.189	0.395
Transformer	Resp. esc.	0.039	0.235	0.320
S4	Resp. esc.	0.036	0.195	0.369
Mamba-3	Resp. esc.	0.035	0.229	0.455
DLinear	Circ. det.	0.046	0.154	0.239
FITS	Circ. det.	0.046	0.137	0.184
ResNet1D	Circ. det.	0.047	0.160	0.215
InceptionTime	Circ. det.	0.050	0.133	0.278
TCN	Circ. det.	0.050	0.192	0.117
BiLSTM	Circ. det.	0.045	0.198	0.235
Transformer	Circ. det.	0.047	0.135	0.158
S4	Circ. det.	0.047	0.113	0.295
Mamba-3	Circ. det.	0.046	0.099	0.216
DLinear	Fluid resp.	0.315	0.110	0.429
FITS	Fluid resp.	0.313	0.050	0.400
ResNet1D	Fluid resp.	0.303	0.085	0.410
InceptionTime	Fluid resp.	0.309	0.097	0.441

Table 80: Architectures — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
TCN	Fluid resp.	0.325	0.081	0.396
BiLSTM	Fluid resp.	0.305	0.120	0.404
Transformer	Fluid resp.	0.314	0.117	0.431
S4	Fluid resp.	0.303	0.129	0.471
Mamba-3	Fluid resp.	0.306	0.173	0.452
DLinear	AKI 2/3	0.103	0.169	0.488
FITS	AKI 2/3	0.113	0.163	0.334
ResNet1D	AKI 2/3	0.112	0.100	0.342
InceptionTime	AKI 2/3	0.113	0.155	0.456
TCN	AKI 2/3	0.112	0.097	0.353
BiLSTM	AKI 2/3	0.097	0.238	0.394
Transformer	AKI 2/3	0.096	0.269	0.193
S4	AKI 2/3	0.106	0.081	0.309
Mamba-3	AKI 2/3	0.090	0.166	0.160
DLinear	Vent. lib.	0.157	0.251	0.509
FITS	Vent. lib.	0.149	0.270	0.460
ResNet1D	Vent. lib.	0.156	0.209	0.555
InceptionTime	Vent. lib.	0.152	0.190	0.554
TCN	Vent. lib.	0.152	0.166	0.563
BiLSTM	Vent. lib.	0.154	0.280	0.591
Transformer	Vent. lib.	0.156	0.236	0.378
S4	Vent. lib.	0.144	0.203	0.488
Mamba-3	Vent. lib.	0.159	0.270	0.518
DLinear	Sepsis-3	0.038	0.229	0.291
FITS	Sepsis-3	0.037	0.215	0.283
ResNet1D	Sepsis-3	0.041	0.272	0.303
InceptionTime	Sepsis-3	0.043	0.156	0.258
TCN	Sepsis-3	0.043	0.170	0.226
BiLSTM	Sepsis-3	0.043	0.199	0.263
Transformer	Sepsis-3	0.040	0.220	0.271
S4	Sepsis-3	0.046	0.173	0.294
Mamba-3	Sepsis-3	0.036	0.195	0.227

Table 81: Architectures — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
DLinear	Acute det.	126.5	3.318	0.108	0.000
FITS	Acute det.	119.6	3.311	0.102	0.000
ResNet1D	Acute det.	122.0	3.263	0.104	0.000
InceptionTime	Acute det.	155.9	3.416	0.133	0.000
TCN	Acute det.	140.0	3.297	0.120	0.000
BiLSTM	Acute det.	117.1	3.409	0.100	0.000
Transformer	Acute det.	127.9	3.395	0.109	0.000
S4	Acute det.	149.5	3.304	0.128	0.000
Mamba-3	Acute det.	465.9	3.429	0.398	0.000
DLinear	Dyn. mort.	169.7	3.971	0.130	0.000
FITS	Dyn. mort.	141.2	3.277	0.108	0.000
ResNet1D	Dyn. mort.	148.6	3.263	0.114	0.000
InceptionTime	Dyn. mort.	171.7	3.270	0.132	0.000
TCN	Dyn. mort.	156.6	3.297	0.120	0.000
BiLSTM	Dyn. mort.	164.8	4.235	0.126	0.000
Transformer	Dyn. mort.	158.0	3.242	0.121	0.000
S4	Dyn. mort.	207.9	3.902	0.159	0.000
Mamba-3	Dyn. mort.	1715.8	4.360	1.315	0.000
DLinear	Resp. esc.	135.7	3.318	0.107	0.000
FITS	Resp. esc.	125.0	3.332	0.098	0.000
ResNet1D	Resp. esc.	118.7	3.423	0.093	0.000
InceptionTime	Resp. esc.	160.8	3.402	0.127	0.000
TCN	Resp. esc.	130.4	3.466	0.103	0.000
BiLSTM	Resp. esc.	117.4	3.409	0.092	0.000
Transformer	Resp. esc.	139.1	3.395	0.110	0.000
S4	Resp. esc.	141.4	3.311	0.111	0.000
Mamba-3	Resp. esc.	257.3	3.582	0.203	0.000
DLinear	Circ. det.	120.3	3.256	0.121	0.000
FITS	Circ. det.	127.0	3.277	0.128	0.000
ResNet1D	Circ. det.	111.0	3.263	0.112	0.000
InceptionTime	Circ. det.	144.0	3.271	0.145	0.000
TCN	Circ. det.	139.3	3.297	0.140	0.000
BiLSTM	Circ. det.	126.6	3.339	0.127	0.000
Transformer	Circ. det.	147.6	3.242	0.148	0.000
S4	Circ. det.	138.1	3.242	0.139	0.000
Mamba-3	Circ. det.	175.1	3.311	0.176	0.000
DLinear	Fluid resp.	38.3	3.256	0.662	0.000
FITS	Fluid resp.	37.9	3.867	0.655	0.000
ResNet1D	Fluid resp.	38.4	3.478	0.664	0.000
InceptionTime	Fluid resp.	40.6	3.645	0.702	0.000

Table 81: Architectures — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
TCN	Fluid resp.	39.3	3.673	0.680	0.000
BiLSTM	Fluid resp.	92.7	4.235	1.602	0.000
Transformer	Fluid resp.	39.7	3.242	0.686	0.000
S4	Fluid resp.	43.0	3.902	0.744	0.000
Mamba-3	Fluid resp.	47.6	5.569	0.823	0.000
DLinear	AKI 2/3	146.0	3.256	0.112	0.000
FITS	AKI 2/3	144.3	3.277	0.110	0.000
ResNet1D	AKI 2/3	146.9	3.263	0.112	0.000
InceptionTime	AKI 2/3	189.9	3.270	0.145	0.000
TCN	AKI 2/3	168.0	3.297	0.128	0.000
BiLSTM	AKI 2/3	178.0	4.235	0.136	0.000
Transformer	AKI 2/3	180.0	3.839	0.137	0.000
S4	AKI 2/3	208.6	3.903	0.159	0.000
Mamba-3	AKI 2/3	340.8	5.507	0.260	0.000
DLinear	Vent. lib.	47.6	3.971	0.512	0.000
FITS	Vent. lib.	46.3	3.860	0.498	0.000
ResNet1D	Vent. lib.	45.6	3.457	0.490	0.000
InceptionTime	Vent. lib.	51.1	3.610	0.550	0.000
TCN	Vent. lib.	48.7	3.610	0.524	0.000
BiLSTM	Vent. lib.	48.4	4.235	0.520	0.000
Transformer	Vent. lib.	45.5	3.395	0.489	0.000
S4	Vent. lib.	47.9	3.902	0.515	0.000
Mamba-3	Vent. lib.	58.6	5.507	0.630	0.000
DLinear	Sepsis-3	92.2	3.971	0.179	0.000
FITS	Sepsis-3	87.8	3.860	0.171	0.000
ResNet1D	Sepsis-3	80.1	3.478	0.156	0.000
InceptionTime	Sepsis-3	107.8	3.645	0.209	0.000
TCN	Sepsis-3	89.9	3.673	0.175	0.000
BiLSTM	Sepsis-3	93.4	4.237	0.181	0.000
Transformer	Sepsis-3	93.6	3.862	0.182	0.000
S4	Sepsis-3	101.9	3.902	0.198	0.000
Mamba-3	Sepsis-3	150.3	5.569	0.292	0.000

Table 82: Architectures — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Transformer	CL07-002	0.039	0.013	0.027
Transformer	H5-002	0.032	0.014	0.009
Transformer	H5-030	0.079	0.053	0.039
Transformer	L2-30	0.101	0.110	0.061
Transformer	OX01-IOH	0.133	0.021	0.028
Transformer	VA1-CO2	0.106	0.029	0.027
Transformer	VA1-STIM	0.060	0.028	0.015
Transformer	VCS05	0.609	0.010	0.009
BiLSTM	CL07-002	0.039	0.017	0.027
BiLSTM	H5-002	0.031	0.013	0.009
BiLSTM	H5-030	0.076	0.054	0.039
BiLSTM	L2-30	0.080	0.091	0.074
BiLSTM	OX01-IOH	0.138	0.021	0.029
BiLSTM	VA1-CO2	0.107	0.027	0.026
BiLSTM	VA1-STIM	0.058	0.028	0.014
BiLSTM	VCS05	0.598	0.015	0.010
Mamba-3	CL07-002	0.038	0.016	0.028
Mamba-3	H5-002	0.030	0.015	0.009
Mamba-3	H5-030	0.076	0.053	0.039
Mamba-3	L2-30	0.085	0.102	0.074
Mamba-3	OX01-IOH	0.136	0.024	0.029
Mamba-3	VA1-CO2	0.102	0.024	0.027
Mamba-3	VA1-STIM	0.059	0.029	0.014
Mamba-3	VCS05	0.604	0.010	0.009
InceptionTime	CL07-002	0.037	0.019	0.028
InceptionTime	H5-002	0.030	0.014	0.009
InceptionTime	H5-030	0.071	0.063	0.042
InceptionTime	L2-30	0.087	0.097	0.076
InceptionTime	OX01-IOH	0.128	0.029	0.029
InceptionTime	VA1-CO2	0.106	0.033	0.028
InceptionTime	VA1-STIM	0.058	0.035	0.014
InceptionTime	VCS05	0.597	0.007	0.010
TCN	CL07-002	0.038	0.016	0.029
TCN	H5-002	0.031	0.014	0.009
TCN	H5-030	0.072	0.059	0.042
TCN	L2-30	0.082	0.090	0.078
TCN	OX01-IOH	0.125	0.027	0.030
TCN	VA1-CO2	0.111	0.026	0.027
TCN	VA1-STIM	0.057	0.035	0.014
TCN	VCS05	0.590	0.009	0.010

Table 82: Architectures — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
ResNet1D	CL07-002	0.039	0.021	0.029
ResNet1D	H5-002	0.033	0.016	0.009
ResNet1D	H5-030	0.074	0.069	0.040
ResNet1D	L2-30	0.086	0.103	0.058
ResNet1D	OX01-IOH	0.139	0.023	0.029
ResNet1D	VA1-CO2	0.111	0.032	0.027
ResNet1D	VA1-STIM	0.057	0.037	0.014
ResNet1D	VCS05	0.591	0.010	0.010
S4	CL07-002	0.038	0.015	0.028
S4	H5-002	0.030	0.014	0.009
S4	H5-030	0.078	0.056	0.038
S4	L2-30	0.087	0.108	0.074
S4	OX01-IOH	0.131	0.024	0.028
S4	VA1-CO2	0.100	0.031	0.026
S4	VA1-STIM	0.060	0.028	0.014
S4	VCS05	0.596	0.010	0.010
FITS	CL07-002	0.039	0.017	0.027
FITS	H5-002	0.030	0.014	0.009
FITS	H5-030	0.076	0.060	0.038
FITS	L2-30	0.084	0.109	0.074
FITS	OX01-IOH	0.139	0.022	0.028
FITS	VA1-CO2	0.106	0.032	0.026
FITS	VA1-STIM	0.061	0.034	0.014
FITS	VCS05	0.601	0.011	0.009
DLinear	CL07-002	0.040	0.017	0.027
DLinear	H5-002	0.032	0.015	0.009
DLinear	H5-030	0.077	0.056	0.038
DLinear	L2-30	0.085	0.106	0.074
DLinear	OX01-IOH	0.136	0.026	0.029
DLinear	VA1-CO2	0.104	0.025	0.027
DLinear	VA1-STIM	0.061	0.032	0.014
DLinear	VCS05	0.599	0.010	0.009

Table 83: Architectures — VitalDB. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	ms/sample	Fail (%)
Transformer	Shared telemetry	32.0	0.414	0.000
BiLSTM	Shared telemetry	30.1	0.390	0.000
Mamba-3	Shared telemetry	41.0	0.531	0.000
InceptionTime	Shared telemetry	33.5	0.433	0.000
TCN	Shared telemetry	31.8	0.412	0.000
ResNet1D	Shared telemetry	30.0	0.388	0.000
S4	Shared telemetry	32.4	0.419	0.000
FITS	Shared telemetry	30.7	0.398	0.000
DLinear	Shared telemetry	31.9	0.413	0.000

Table 84: Architectures — MC-MED. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Transformer	Decomp. 60m	0.144	0.358	0.536
Transformer	Pending hyperK	0.047	0.333	0.509
Transformer	Lactate NC	0.411	0.099	0.526
Transformer	Pending anemia	0.060	0.340	0.229
BiLSTM	Decomp. 60m	0.141	0.351	0.266
BiLSTM	Pending hyperK	0.049	0.319	0.521
BiLSTM	Lactate NC	0.387	0.128	0.415
BiLSTM	Pending anemia	0.051	0.350	0.348
Mamba-3	Decomp. 60m	0.141	0.351	0.379
Mamba-3	Pending hyperK	0.047	0.291	0.246
Mamba-3	Lactate NC	0.372	0.119	0.532
Mamba-3	Pending anemia	0.062	0.321	0.312
InceptionTime	Decomp. 60m	0.138	0.342	0.366
InceptionTime	Pending hyperK	0.047	0.408	0.250
InceptionTime	Lactate NC	0.388	0.126	0.479
InceptionTime	Pending anemia	0.054	0.371	0.211
TCN	Decomp. 60m	0.144	0.349	0.562
TCN	Pending hyperK	0.047	0.414	0.587
TCN	Lactate NC	0.390	0.125	0.474
TCN	Pending anemia	0.060	0.374	0.046
ResNet1D	Decomp. 60m	0.164	0.330	0.634
ResNet1D	Pending hyperK	0.052	0.362	0.560
ResNet1D	Lactate NC	0.406	0.126	0.447
ResNet1D	Pending anemia	0.053	0.345	0.048
S4	Decomp. 60m	0.141	0.335	0.251
S4	Pending hyperK	0.049	0.283	0.468
S4	Lactate NC	0.392	0.089	0.451
S4	Pending anemia	0.064	0.275	0.245
FITS	Decomp. 60m	0.137	0.362	0.414
FITS	Pending hyperK	0.048	0.240	0.461
FITS	Lactate NC	0.371	0.124	0.539
FITS	Pending anemia	0.051	0.328	0.373
DLinear	Decomp. 60m	0.141	0.354	0.405
DLinear	Pending hyperK	0.049	0.312	0.477
DLinear	Lactate NC	0.376	0.068	0.552
DLinear	Pending anemia	0.049	0.295	0.280

Table 85: Architectures — MC-MED. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	ms/sample	Fail (%)
Transformer	Decomp. 60m	216.2	2.413	0.000
Transformer	Pending hyperK	227.1	2.431	0.000
Transformer	Lactate NC	138.0	26.954	0.000
Transformer	Pending anemia	190.8	3.726	0.000
BiLSTM	Decomp. 60m	639.1	7.132	0.000
BiLSTM	Pending hyperK	637.6	6.823	0.000
BiLSTM	Lactate NC	108.0	21.093	0.000
BiLSTM	Pending anemia	358.8	7.007	0.000
Mamba-3	Decomp. 60m	598.0	6.674	0.000
Mamba-3	Pending hyperK	600.7	6.428	0.000
Mamba-3	Lactate NC	80.6	15.738	0.000
Mamba-3	Pending anemia	555.5	10.850	0.000
InceptionTime	Decomp. 60m	200.1	2.233	0.000
InceptionTime	Pending hyperK	579.3	6.199	0.000
InceptionTime	Lactate NC	83.4	16.294	0.000
InceptionTime	Pending anemia	542.5	10.596	0.000
TCN	Decomp. 60m	557.1	6.218	0.000
TCN	Pending hyperK	559.9	5.992	0.000
TCN	Lactate NC	92.8	18.123	0.000
TCN	Pending anemia	544.6	10.637	0.000
ResNet1D	Decomp. 60m	553.4	6.176	0.000
ResNet1D	Pending hyperK	559.8	5.991	0.000
ResNet1D	Lactate NC	85.0	16.603	0.000
ResNet1D	Pending anemia	526.2	10.278	0.000
S4	Decomp. 60m	602.1	6.720	0.000
S4	Pending hyperK	557.1	5.962	0.000
S4	Lactate NC	85.7	16.740	0.000
S4	Pending anemia	523.6	10.227	0.000
FITS	Decomp. 60m	544.6	6.078	0.000
FITS	Pending hyperK	561.3	6.007	0.000
FITS	Lactate NC	121.2	23.668	0.000
FITS	Pending anemia	243.8	4.762	0.000
DLinear	Decomp. 60m	554.2	6.185	0.000
DLinear	Pending hyperK	555.8	5.948	0.000
DLinear	Lactate NC	85.0	16.605	0.000
DLinear	Pending anemia	204.3	3.990	0.000

Table 86: Tokenizers — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Conv1D	Acute det.	0.040	0.131	0.321
Conv1D	Dyn. mort.	0.025	0.214	0.290
Conv1D	Resp. esc.	0.039	0.235	0.320
Conv1D	Circ. det.	0.050	0.161	0.225
Conv1D	Fluid resp.	0.323	0.112	0.458
Conv1D	AKI 2/3	0.113	0.102	0.426
Conv1D	Vent. lib.	0.156	0.236	0.378
Conv1D	Sepsis-3	0.038	0.230	0.223
STraTS	Sepsis-3	0.040	0.161	0.263
SeFT	Sepsis-3	0.047	0.166	0.227
mTAND	Sepsis-3	0.035	0.170	0.273
Crossformer	Sepsis-3	0.037	0.254	0.299
iTransformer	Sepsis-3	0.037	0.307	0.272
FFT	Sepsis-3	0.041	0.107	0.189
STraTS	Acute det.	0.041	0.117	0.214
SeFT	Acute det.	0.037	0.098	0.267
mTAND	Acute det.	0.036	0.123	0.204
Crossformer	Acute det.	0.039	0.104	0.226
iTransformer	Acute det.	0.034	0.153	0.193
FFT	Acute det.	0.048	0.078	0.229
STraTS	Dyn. mort.	0.020	0.260	0.177
SeFT	Dyn. mort.	0.030	0.197	0.165
mTAND	Dyn. mort.	0.029	0.119	0.174
Crossformer	Dyn. mort.	0.018	0.383	0.167
iTransformer	Dyn. mort.	0.022	0.324	0.202
FFT	Dyn. mort.	0.019	0.171	0.139
STraTS	Resp. esc.	0.035	0.232	0.323
SeFT	Resp. esc.	0.034	0.161	0.295
mTAND	Resp. esc.	0.035	0.298	0.410
Crossformer	Resp. esc.	0.044	0.245	0.283
iTransformer	Resp. esc.	0.041	0.278	0.169
FFT	Resp. esc.	0.042	0.150	0.233
STraTS	Circ. det.	0.050	0.095	0.210
SeFT	Circ. det.	0.047	0.155	0.181
mTAND	Circ. det.	0.050	0.157	0.161
Crossformer	Circ. det.	0.050	0.173	0.142
iTransformer	Circ. det.	0.045	0.131	0.133
FFT	Circ. det.	0.053	0.112	0.248
STraTS	Fluid resp.	0.303	0.140	0.398
SeFT	Fluid resp.	0.312	0.133	0.474

Table 86: Tokenizers — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
mTAND	Fluid resp.	0.308	0.135	0.515
Crossformer	Fluid resp.	0.319	0.135	0.398
iTransformer	Fluid resp.	0.305	0.158	0.430
FFT	Fluid resp.	0.313	0.136	0.422
STraTS	AKI 2/3	0.088	0.406	0.367
SeFT	AKI 2/3	0.088	0.276	0.637
mTAND	AKI 2/3	0.091	0.265	0.191
Crossformer	AKI 2/3	0.089	0.199	0.329
iTransformer	AKI 2/3	0.089	0.283	0.514
FFT	AKI 2/3	0.107	0.054	0.226
STraTS	Vent. lib.	0.168	0.066	0.440
SeFT	Vent. lib.	0.155	0.175	0.426
mTAND	Vent. lib.	0.175	0.237	0.443
Crossformer	Vent. lib.	0.154	0.244	0.590
iTransformer	Vent. lib.	0.148	0.185	0.516
FFT	Vent. lib.	0.174	0.217	0.480

Table 87: Tokenizers — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Conv1D	Acute det.	127.9	3.395	0.109	0.000
Conv1D	Dyn. mort.	148.3	3.395	0.114	0.000
Conv1D	Resp. esc.	139.1	3.395	0.110	0.000
Conv1D	Circ. det.	134.5	3.395	0.135	0.000
Conv1D	Fluid resp.	38.7	3.395	0.668	0.000
Conv1D	AKI 2/3	158.7	3.395	0.121	0.000
Conv1D	Vent. lib.	45.5	3.395	0.489	0.000
Conv1D	Sepsis-3	76.1	3.395	0.148	0.000
STraTS	Sepsis-3	370.8	7.535	0.720	0.000
SeFT	Sepsis-3	120.0	4.881	0.233	0.000
mTAND	Sepsis-3	691.2	17.233	1.343	0.000
Crossformer	Sepsis-3	367.2	7.320	0.713	0.000
iTransformer	Sepsis-3	109.2	3.638	0.212	0.000
FFT	Sepsis-3	104.7	3.402	0.203	0.000
STraTS	Acute det.	800.2	7.535	0.684	0.000
SeFT	Acute det.	233.7	4.885	0.200	0.000
mTAND	Acute det.	1491.1	17.240	1.275	0.000
Crossformer	Acute det.	790.4	7.320	0.676	0.000
iTransformer	Acute det.	197.7	3.617	0.169	0.000
FFT	Acute det.	184.5	3.409	0.158	0.000
STraTS	Dyn. mort.	888.2	7.535	0.681	0.000
SeFT	Dyn. mort.	249.6	4.885	0.191	0.000
mTAND	Dyn. mort.	1683.2	17.247	1.290	0.000
Crossformer	Dyn. mort.	870.4	7.327	0.667	0.000
iTransformer	Dyn. mort.	215.2	3.626	0.165	0.000
FFT	Dyn. mort.	206.3	3.416	0.158	0.000
STraTS	Resp. esc.	856.6	7.535	0.674	0.000
SeFT	Resp. esc.	245.3	4.883	0.193	0.000
mTAND	Resp. esc.	1608.9	17.233	1.267	0.000
Crossformer	Resp. esc.	842.9	7.320	0.664	0.000
iTransformer	Resp. esc.	203.3	3.646	0.160	0.000
FFT	Resp. esc.	194.4	3.402	0.153	0.000
STraTS	Circ. det.	685.6	7.535	0.690	0.000
SeFT	Circ. det.	204.7	4.881	0.206	0.000
mTAND	Circ. det.	1292.4	17.233	1.300	0.000
Crossformer	Circ. det.	677.4	7.320	0.681	0.000
iTransformer	Circ. det.	170.2	3.582	0.171	0.000
FFT	Circ. det.	163.4	3.402	0.164	0.000
STraTS	Fluid resp.	68.4	7.535	1.183	0.000
SeFT	Fluid resp.	43.3	4.881	0.749	0.000

Table 87: Tokenizers — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
mTAND	Fluid resp.	101.9	17.233	1.761	0.000
Crossformer	Fluid resp.	68.7	7.320	1.188	0.000
iTransformer	Fluid resp.	40.0	3.638	0.691	0.000
FFT	Fluid resp.	40.0	3.402	0.691	0.000
STraTS	AKI 2/3	884.3	7.535	0.675	0.000
SeFT	AKI 2/3	242.8	4.878	0.185	0.000
mTAND	AKI 2/3	1656.8	17.233	1.265	0.000
Crossformer	AKI 2/3	869.1	7.320	0.664	0.000
iTransformer	AKI 2/3	208.2	3.582	0.159	0.000
FFT	AKI 2/3	201.3	3.402	0.154	0.000
STraTS	Vent. lib.	93.1	7.535	1.001	0.000
SeFT	Vent. lib.	50.1	4.881	0.539	0.000
mTAND	Vent. lib.	147.1	17.233	1.581	0.000
Crossformer	Vent. lib.	94.6	7.320	1.017	0.000
iTransformer	Vent. lib.	46.6	3.582	0.501	0.000
FFT	Vent. lib.	46.3	3.402	0.498	0.000

Table 88: Tokenizers — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Conv1D	CL07-002	0.039	0.013	0.027
Conv1D	H5-002	0.032	0.014	0.009
Conv1D	H5-030	0.079	0.053	0.039
Conv1D	L2-30	0.101	0.110	0.061
Conv1D	OX01-IOH	0.133	0.021	0.028
Conv1D	VA1-CO2	0.106	0.029	0.027
Conv1D	VA1-STIM	0.060	0.028	0.015
Conv1D	VCS05	0.609	0.010	0.009
SeFT	CL07-002	0.032	0.020	0.032
SeFT	H5-002	0.022	0.013	0.011
SeFT	H5-030	0.074	0.059	0.039
SeFT	L2-30	0.103	0.160	0.045
SeFT	OX01-IOH	0.125	0.026	0.031
SeFT	VA1-CO2	0.107	0.036	0.031
SeFT	VA1-STIM	0.052	0.027	0.013
SeFT	VCS05	0.352	0.016	0.012
Crossformer	CL07-002	0.036	0.015	0.030
Crossformer	H5-002	0.021	0.014	0.012
Crossformer	H5-030	0.072	0.043	0.049
Crossformer	L2-30	0.094	0.116	0.064
Crossformer	OX01-IOH	0.096	0.031	0.036
Crossformer	VA1-CO2	0.072	0.030	0.038
Crossformer	VA1-STIM	0.042	0.019	0.020
Crossformer	VCS05	0.583	0.009	0.011
STraTS	CL07-002	0.037	0.018	0.029
STraTS	H5-002	0.025	0.016	0.009
STraTS	H5-030	0.074	0.056	0.038
STraTS	L2-30	0.096	0.147	0.050
STraTS	OX01-IOH	0.132	0.024	0.029
STraTS	VA1-CO2	0.121	0.034	0.025
STraTS	VA1-STIM	0.058	0.026	0.012
STraTS	VCS05	0.447	0.014	0.012
iTransformer	CL07-002	0.039	0.017	0.028
iTransformer	H5-002	0.018	0.011	0.013
iTransformer	H5-030	0.066	0.039	0.052
iTransformer	L2-30	0.079	0.077	0.082
iTransformer	OX01-IOH	0.079	0.037	0.043
iTransformer	VA1-CO2	0.076	0.034	0.039
iTransformer	VA1-STIM	0.041	0.019	0.020
iTransformer	VCS05	0.545	0.011	0.012

Table 88: Tokenizers — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
FFT	CL07-002	0.038	0.020	0.029
FFT	H5-002	0.032	0.014	0.008
FFT	H5-030	0.081	0.049	0.038
FFT	L2-30	0.138	0.082	0.043
FFT	OX01-IOH	0.083	0.028	0.043
FFT	VA1-CO2	0.102	0.031	0.025
FFT	VA1-STIM	0.060	0.024	0.015
FFT	VCS05	0.607	0.009	0.011
mTAND	CL07-002	0.040	0.017	0.027
mTAND	H5-002	0.023	0.020	0.013
mTAND	H5-030	0.068	0.059	0.045
mTAND	L2-30	0.088	0.127	0.059
mTAND	OX01-IOH	0.135	0.023	0.030
mTAND	VA1-CO2	0.106	0.038	0.034
mTAND	VA1-STIM	0.041	0.030	0.021
mTAND	VCS05	0.504	0.013	0.012

Table 89: Tokenizers — VitalDB. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	ms/sample	Fail (%)
Conv1D	Shared telemetry	32.0	0.414	0.000
SeFT	Shared telemetry	33.8	0.437	0.000
Crossformer	Shared telemetry	32.7	0.424	0.000
STraTS	Shared telemetry	–	–	0.000
iTransformer	Shared telemetry	–	–	0.000
FFT	Shared telemetry	–	–	0.000
mTAND	Shared telemetry	–	–	0.000

Table 90: Tokenizers — MC-MED. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Conv1D	Decomp. 60m	0.133	0.307	0.158
Conv1D	Pending hyperK	0.042	0.287	0.220
Conv1D	Lactate NC	0.359	0.100	0.544
Conv1D	Pending anemia	0.057	0.324	0.383
SeFT	Decomp. 60m	0.132	0.332	0.528
SeFT	Pending hyperK	0.037	0.451	0.596
SeFT	Lactate NC	0.365	0.120	0.525
SeFT	Pending anemia	0.047	0.446	0.513
Crossformer	Decomp. 60m	0.137	0.105	0.218
Crossformer	Pending hyperK	0.047	0.136	0.237
Crossformer	Lactate NC	0.391	0.236	0.483
Crossformer	Pending anemia	0.037	0.083	0.448
STraTS	Decomp. 60m	0.127	0.366	0.422
STraTS	Pending hyperK	0.036	0.496	0.781
STraTS	Lactate NC	0.346	0.102	0.595
STraTS	Pending anemia	0.043	0.393	0.560
iTransformer	Decomp. 60m	0.135	0.154	0.265
iTransformer	Pending hyperK	0.045	0.184	0.221
iTransformer	Lactate NC	0.361	0.254	0.490
iTransformer	Pending anemia	0.036	0.069	0.391
FFT	Decomp. 60m	0.144	0.358	0.536
FFT	Pending hyperK	0.047	0.333	0.509
FFT	Lactate NC	0.411	0.099	0.526
FFT	Pending anemia	0.060	0.340	0.229
mTAND	Decomp. 60m	0.133	0.269	0.537
mTAND	Pending hyperK	0.041	0.492	0.590
mTAND	Lactate NC	0.375	0.135	0.474
mTAND	Pending anemia	0.045	0.443	0.211

Table 91: Tokenizers — MC-MED. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	ms/sample	Fail (%)
Conv1D	Decomp. 60m	228.9	2.554	0.000
Conv1D	Pending hyperK	231.4	2.477	0.000
Conv1D	Lactate NC	114.2	22.313	0.000
Conv1D	Pending anemia	175.2	3.421	0.000
SeFT	Decomp. 60m	175.9	1.963	0.000
SeFT	Pending hyperK	183.5	1.964	0.000
SeFT	Lactate NC	77.4	15.119	0.000
SeFT	Pending anemia	164.9	3.220	0.000
Crossformer	Decomp. 60m	183.7	2.050	0.000
Crossformer	Pending hyperK	182.4	1.952	0.000
Crossformer	Lactate NC	89.0	17.380	0.000
Crossformer	Pending anemia	152.0	2.969	0.000
STraTS	Decomp. 60m	175.0	1.953	0.000
STraTS	Pending hyperK	185.0	1.980	0.000
STraTS	Lactate NC	83.2	16.253	0.000
STraTS	Pending anemia	169.9	3.318	0.000
iTransformer	Decomp. 60m	167.6	1.870	0.000
iTransformer	Pending hyperK	182.2	1.950	0.000
iTransformer	Lactate NC	107.0	20.905	0.000
iTransformer	Pending anemia	150.9	2.947	0.000
FFT	Decomp. 60m	216.2	2.413	0.000
FFT	Pending hyperK	227.1	2.431	0.000
FFT	Lactate NC	138.0	26.954	0.000
FFT	Pending anemia	190.8	3.726	0.000
mTAND	Decomp. 60m	586.6	6.547	0.000
mTAND	Pending hyperK	622.3	6.660	0.000
mTAND	Lactate NC	120.2	23.477	0.000
mTAND	Pending anemia	392.0	7.657	0.000

Table 92: Observation length — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means.

Condition	Task	AUPRC	ECE	AURC
10 min	Task mean	0.119	0.164	0.322
10 min	Acute det.	0.090	0.118	0.331
10 min	AKI 2/3	0.161	0.060	0.223
10 min	Circ. det.	0.053	0.171	0.220
10 min	Fluid resp.	0.358	0.090	0.427
10 min	Dyn. mort.	0.021	0.151	0.206
10 min	Resp. esc.	0.055	0.226	0.336
10 min	Sepsis-3	0.046	0.290	0.394
10 min	Vent. lib.	0.170	0.204	0.436
20 min	Task mean	0.118	0.154	0.319
20 min	Acute det.	0.085	0.102	0.319
20 min	AKI 2/3	0.158	0.088	0.231
20 min	Circ. det.	0.054	0.166	0.213
20 min	Fluid resp.	0.353	0.073	0.421
20 min	Dyn. mort.	0.022	0.162	0.209
20 min	Resp. esc.	0.058	0.177	0.341
20 min	Sepsis-3	0.046	0.260	0.358
20 min	Vent. lib.	0.166	0.203	0.460
60 min	Task mean	0.123	0.167	0.328
60 min	Acute det.	0.102	0.111	0.305
60 min	AKI 2/3	0.165	0.052	0.221
60 min	Circ. det.	0.053	0.187	0.249
60 min	Fluid resp.	0.358	0.087	0.416
60 min	Dyn. mort.	0.020	0.140	0.195
60 min	Resp. esc.	0.073	0.279	0.453
60 min	Sepsis-3	0.042	0.270	0.404
60 min	Vent. lib.	0.168	0.210	0.381

Table 93: Observation length — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means.

Condition	Task	Mem. (%)	ms/sample	Fail (%)
10 min	Shared telemetry	5.240	0.230	0.000
20 min	Shared telemetry	5.150	0.230	0.000
60 min	Shared telemetry	4.570	0.210	0.000

Table 94: Observation length — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means.

Condition	Task	AUPRC	ECE	AURC
15 min	Task mean	0.078	0.039	0.030
15 min	CL07-002	0.038	–	–
15 min	H5-002	0.028	–	–
15 min	H5-030	0.075	–	–
15 min	L2-30	0.096	–	–
15 min	OX01-IOH	0.144	–	–
15 min	VA1-CO2	0.105	–	–
15 min	VA1-STIM	0.061	–	–
30 min	Task mean	0.081	0.038	0.029
30 min	CL07-002	0.038	–	–
30 min	H5-002	0.025	–	–
30 min	H5-030	0.074	–	–
30 min	L2-30	0.099	–	–
30 min	OX01-IOH	0.159	–	–
30 min	VA1-CO2	0.104	–	–
30 min	VA1-STIM	0.067	–	–
5 min	Task mean	0.076	0.041	0.030
5 min	CL07-002	0.038	–	–
5 min	H5-002	0.030	–	–
5 min	H5-030	0.076	–	–
5 min	L2-30	0.091	–	–
5 min	OX01-IOH	0.132	–	–
5 min	VA1-CO2	0.109	–	–
5 min	VA1-STIM	0.058	–	–
60 min	Task mean	0.081	0.035	0.028
60 min	CL07-002	0.037	–	–
60 min	H5-002	0.024	–	–
60 min	H5-030	0.074	–	–
60 min	L2-30	0.108	–	–
60 min	OX01-IOH	0.160	–	–
60 min	VA1-CO2	0.102	–	–
60 min	VA1-STIM	0.063	–	–

Table 95: Observation length — VitalDB. Training-cost and failure values for the corresponding EES figure; entries are the plotted means.

Condition	Task	ms/sample	Fail (%)
15 min	Shared telemetry	0.420	0.000
30 min	Shared telemetry	0.420	0.000
5 min	Shared telemetry	0.420	0.000
60 min	Shared telemetry	0.420	0.000

Table 96: Observation length — MC-MED. Predictive and confidence values for the corresponding EES figure; entries are the plotted means.

Condition	Task	AUPRC	ECE	AURC
10 min	Task mean	0.198	0.087	0.268
10 min	Decomp. 60m	0.276	–	–
10 min	Lactate NC	0.397	–	–
10 min	Pending hyperK	0.066	–	–
10 min	Pending anemia	0.052	–	–
15 min	Task mean	0.221	0.096	0.266
15 min	Decomp. 60m	0.317	–	–
15 min	Lactate NC	0.469	–	–
15 min	Pending hyperK	0.052	–	–
15 min	Pending anemia	0.046	–	–
5 min	Task mean	0.181	0.103	0.238
5 min	Decomp. 60m	0.264	–	–
5 min	Lactate NC	0.367	–	–
5 min	Pending hyperK	0.046	–	–
5 min	Pending anemia	0.047	–	–

Table 97: Observation length — MC-MED. Training-cost and failure values for the corresponding EES figure; entries are the plotted means.

Condition	Task	ms/step	Mem. (MiB)	Fail (%)
10 min	Shared telemetry	–	43212.5	0.000
10 min	Decomp. 60m	19.54	–	–
10 min	Lactate NC	16.87	–	–
10 min	Pending hyperK	19.82	–	–
10 min	Pending anemia	18.95	–	–
15 min	Shared telemetry	–	28901.5	0.000
15 min	Decomp. 60m	25.99	–	–
15 min	Lactate NC	25.17	–	–
15 min	Pending hyperK	25.96	–	–
15 min	Pending anemia	23.97	–	–
5 min	Shared telemetry	–	31176.2	0.000
5 min	Decomp. 60m	13.11	–	–
5 min	Lactate NC	14.76	–	–
5 min	Pending hyperK	14.06	–	–
5 min	Pending anemia	14.90	–	–

Table 98: Forecast gap — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means.

Condition	Task	AUPRC	ECE	AURC
2 h	Sepsis-3	0.036	0.311	0.379
24 h	Dyn. mort.	0.018	0.249	0.355
3 h	AKI 2/3	0.151	0.169	0.344
3 h	Dyn. mort.	0.027	0.181	0.259
4 h	Sepsis-3	0.032	0.265	0.307
6 h	AKI 2/3	0.140	0.193	0.313
6 h	Sepsis-3	0.030	0.220	0.266
12 h	AKI 2/3	0.120	0.222	0.292
12 h	Dyn. mort.	0.017	0.342	0.253

Table 99: Forecast gap — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means.

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
2 h	Sepsis-3	49.3	7.600	0.100	0.000
24 h	Dyn. mort.	252.8	3.800	0.200	0.000
3 h	AKI 2/3	294.5	3.800	0.240	0.000
3 h	Dyn. mort.	323.8	3.800	0.250	0.000
4 h	Sepsis-3	46.4	7.350	0.100	0.000
6 h	AKI 2/3	279.8	3.800	0.240	0.000
6 h	Sepsis-3	47.1	7.110	0.100	0.000
12 h	AKI 2/3	250.9	3.800	0.250	0.000
12 h	Dyn. mort.	314.0	3.800	0.240	0.000

Table 100: Forecast gap — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means.

Condition	Task	AUPRC	ECE	AURC
15 min	Task mean	0.071	0.035	0.029
15 min	CL07-002	0.036	0.015	0.025
15 min	H5-002	0.022	0.017	0.008
15 min	H5-030	0.077	0.048	0.039
15 min	L2-30	0.114	0.087	0.058
15 min	OX01-IOH	0.101	0.022	0.027
15 min	VA1-CO2	0.100	0.029	0.028
15 min	VA1-STIM	0.048	0.026	0.015
30 min	Task mean	0.077	0.034	0.030
30 min	CL07-002	0.034	0.021	0.021
30 min	H5-002	0.027	0.013	0.008
30 min	H5-030	0.073	0.053	0.035
30 min	L2-30	0.137	0.072	0.067
30 min	OX01-IOH	0.113	0.032	0.028
30 min	VA1-CO2	0.099	0.024	0.034
30 min	VA1-STIM	0.053	0.021	0.018
45 min	Task mean	0.068	0.030	0.030
45 min	CL07-002	0.029	0.021	0.018
45 min	H5-002	0.025	0.015	0.006
45 min	H5-030	0.080	0.042	0.039
45 min	L2-30	0.111	0.066	0.062
45 min	OX01-IOH	0.104	0.023	0.032
45 min	VA1-CO2	0.082	0.026	0.031
45 min	VA1-STIM	0.047	0.019	0.022
5 min	Task mean	0.100	0.033	0.027
5 min	CL07-002	0.044	0.016	0.029
5 min	H5-002	0.034	0.008	0.010
5 min	H5-030	0.072	0.061	0.039
5 min	L2-30	0.112	0.084	0.061
5 min	OX01-IOH	0.223	0.020	0.020
5 min	VA1-CO2	0.102	0.027	0.024
5 min	VA1-STIM	0.114	0.017	0.009
60 min	Task mean	0.071	0.034	0.029
60 min	CL07-002	0.036	0.018	0.023
60 min	H5-002	0.022	0.014	0.006
60 min	H5-030	0.078	0.046	0.032
60 min	L2-30	0.086	0.070	0.064
60 min	OX01-IOH	0.103	0.029	0.037
60 min	VA1-CO2	0.090	0.030	0.033
60 min	VA1-STIM	0.087	0.029	0.011

Table 100: Forecast gap — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. (Continued.)

Condition	Task	AUPRC	ECE	AURC
75 min	Task mean	0.082	0.035	0.029
75 min	CL07-002	0.035	0.026	0.019
75 min	H5-002	0.022	0.015	0.007
75 min	H5-030	0.065	0.033	0.035
75 min	L2-30	0.085	0.071	0.068
75 min	OX01-IOH	0.127	0.029	0.034
75 min	VA1-CO2	0.124	0.041	0.029
75 min	VA1-STIM	0.113	0.030	0.013

Table 101: Forecast gap — MC-MED. Predictive and confidence values for the corresponding EES figure; entries are the plotted means.

Condition	Task	AUPRC	ECE	AURC
0 min	Decomp. 60m	0.264	0.068	0.279
0 min	Pending hyperK	0.046	0.051	0.202
0 min	Pending anemia	0.047	0.035	0.050
10 min	Decomp. 60m	0.214	0.066	0.201
15 min	Pending hyperK	0.046	0.072	0.323
15 min	Pending anemia	0.070	0.046	0.133
30 min	Decomp. 60m	0.157	0.073	0.229
30 min	Lactate NC	0.367	0.258	0.423
60 min	Lactate NC	0.385	0.153	0.419
60 min	Pending hyperK	0.077	0.081	0.257
60 min	Pending anemia	0.200	0.078	0.200
90 min	Lactate NC	0.399	0.247	0.390

Table 102: Forecast gap — MC-MED. Training-cost and failure values for the corresponding EES figure; entries are the plotted means.

Condition	Task	Time (s)	Mem. (MiB)	Fail (%)
0 min	Decomp. 60m	2.9	39238.0	0.000
0 min	Pending hyperK	15.9	46850.0	0.000
0 min	Pending anemia	13.7	28876.0	0.000
10 min	Decomp. 60m	5.5	35301.0	0.000
15 min	Pending hyperK	7.5	46461.0	0.000
15 min	Pending anemia	5.5	21562.0	0.000
30 min	Decomp. 60m	3.8	28188.0	0.000
30 min	Lactate NC	2.0	9741.0	0.000
60 min	Lactate NC	1.5	9606.0	0.000
60 min	Pending hyperK	6.4	21782.0	0.000
60 min	Pending anemia	0.6	8324.0	0.000
90 min	Lactate NC	1.6	8381.0	0.000

Table 103: Channel selection — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Clinical	Acute det.	0.041	0.143	0.339
MI	Acute det.	0.035	0.250	0.356
Lasso	Acute det.	0.036	0.152	0.334
SQI	Acute det.	0.038	0.219	0.396
Temporal MI	Acute det.	0.030	0.245	0.677
Random	Acute det.	0.033	0.229	0.339
All channels	Acute det.	0.040	0.131	0.321
Clinical	Dyn. mort.	0.026	0.243	0.292
MI	Dyn. mort.	0.015	0.423	0.372
Lasso	Dyn. mort.	0.017	0.291	0.270
SQI	Dyn. mort.	0.023	0.233	0.248
Temporal MI	Dyn. mort.	0.014	0.455	0.882
Random	Dyn. mort.	0.016	0.482	0.693
All channels	Dyn. mort.	0.025	0.214	0.290
Clinical	Resp. esc.	0.037	0.223	0.468
MI	Resp. esc.	0.031	0.280	0.409
Lasso	Resp. esc.	0.034	0.204	0.459
SQI	Resp. esc.	0.038	0.276	0.477
Temporal MI	Resp. esc.	0.030	0.343	0.351
Random	Resp. esc.	0.031	0.447	0.607
All channels	Resp. esc.	0.039	0.235	0.320
Clinical	Circ. det.	0.047	0.231	0.138
MI	Circ. det.	0.040	0.293	0.140
Lasso	Circ. det.	0.044	0.213	0.260
SQI	Circ. det.	0.044	0.215	0.171
Temporal MI	Circ. det.	0.038	0.321	0.489
Random	Circ. det.	0.029	0.426	0.678
All channels	Circ. det.	0.050	0.161	0.225
Clinical	Fluid resp.	0.308	0.086	0.563
MI	Fluid resp.	0.307	0.074	0.452
Lasso	Fluid resp.	0.307	0.093	0.452
SQI	Fluid resp.	0.316	0.081	0.454
Temporal MI	Fluid resp.	0.307	0.083	0.540
Random	Fluid resp.	0.307	0.113	0.378
All channels	Fluid resp.	0.323	0.112	0.458
Clinical	AKI 2/3	0.114	0.222	0.223
MI	AKI 2/3	0.095	0.258	0.428
Lasso	AKI 2/3	0.106	0.260	0.257
SQI	AKI 2/3	0.116	0.194	0.241
Temporal MI	AKI 2/3	0.101	0.235	0.188

Table 103: Channel selection — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Random	AKI 2/3	0.093	0.289	0.239
All channels	AKI 2/3	0.113	0.102	0.426
Clinical	Vent. lib.	0.156	0.323	0.659
MI	Vent. lib.	0.147	0.278	0.457
Lasso	Vent. lib.	0.171	0.309	0.529
SQI	Vent. lib.	0.160	0.221	0.542
Temporal MI	Vent. lib.	0.146	0.177	0.346
Random	Vent. lib.	0.137	0.361	0.424
All channels	Vent. lib.	0.156	0.236	0.378
Clinical	Sepsis-3	0.037	0.355	0.374
MI	Sepsis-3	0.029	0.374	0.437
Lasso	Sepsis-3	0.037	0.214	0.252
SQI	Sepsis-3	0.037	0.248	0.378
Temporal MI	Sepsis-3	0.029	0.337	0.648
Random	Sepsis-3	0.028	0.358	0.369
All channels	Sepsis-3	0.038	0.230	0.223

Table 104: Channel selection — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Clinical	Acute det.	117.4	1.646	0.100	0.000
MI	Acute det.	120.0	1.646	0.103	0.000
Lasso	Acute det.	118.5	1.646	0.101	0.000
SQI	Acute det.	118.8	1.648	0.102	0.000
Temporal MI	Acute det.	119.4	1.647	0.102	0.000
Random	Acute det.	123.2	1.647	0.105	0.000
All channels	Acute det.	127.9	3.395	0.109	0.000
Clinical	Dyn. mort.	137.8	2.376	0.106	0.000
MI	Dyn. mort.	139.3	2.376	0.107	0.000
Lasso	Dyn. mort.	139.8	2.376	0.107	0.000
SQI	Dyn. mort.	143.4	2.376	0.110	0.000
Temporal MI	Dyn. mort.	137.6	2.376	0.105	0.000
Random	Dyn. mort.	141.4	2.376	0.108	0.000
All channels	Dyn. mort.	148.3	3.395	0.114	0.000
Clinical	Resp. esc.	128.4	2.376	0.101	0.000
MI	Resp. esc.	131.9	2.376	0.104	0.000
Lasso	Resp. esc.	127.9	2.376	0.101	0.000
SQI	Resp. esc.	131.4	2.376	0.103	0.000
Temporal MI	Resp. esc.	132.2	2.376	0.104	0.000
Random	Resp. esc.	134.0	2.376	0.105	0.000
All channels	Resp. esc.	139.1	3.395	0.110	0.000
Clinical	Circ. det.	108.4	2.376	0.109	0.000
MI	Circ. det.	105.5	2.376	0.106	0.000
Lasso	Circ. det.	106.1	2.376	0.107	0.000
SQI	Circ. det.	105.6	2.376	0.106	0.000
Temporal MI	Circ. det.	107.7	2.376	0.108	0.000
Random	Circ. det.	107.7	2.376	0.108	0.000
All channels	Circ. det.	134.5	3.395	0.135	0.000
Clinical	Fluid resp.	15.5	2.649	0.268	0.000
MI	Fluid resp.	12.8	2.565	0.221	0.000
Lasso	Fluid resp.	22.7	2.649	0.392	0.000
SQI	Fluid resp.	11.8	2.460	0.204	0.000
Temporal MI	Fluid resp.	14.8	2.649	0.256	0.000
Random	Fluid resp.	16.7	2.649	0.289	0.000
All channels	Fluid resp.	38.7	3.395	0.668	0.000
Clinical	AKI 2/3	222.9	2.649	0.170	0.000
MI	AKI 2/3	164.6	2.460	0.126	0.000
Lasso	AKI 2/3	162.8	2.460	0.124	0.000
SQI	AKI 2/3	218.5	2.649	0.167	0.000
Temporal MI	AKI 2/3	134.0	2.376	0.102	0.000

Table 104: Channel selection — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Random	AKI 2/3	215.0	2.649	0.164	0.000
All channels	AKI 2/3	158.7	3.395	0.121	0.000
Clinical	Vent. lib.	13.1	2.376	0.141	0.000
MI	Vent. lib.	15.5	2.460	0.167	0.000
Lasso	Vent. lib.	15.4	2.460	0.165	0.000
SQI	Vent. lib.	17.4	2.565	0.187	0.000
Temporal MI	Vent. lib.	17.9	2.565	0.192	0.000
Random	Vent. lib.	17.0	2.460	0.183	0.000
All channels	Vent. lib.	45.5	3.395	0.489	0.000
Clinical	Sepsis-3	57.0	2.376	0.111	0.000
MI	Sepsis-3	88.3	2.649	0.172	0.000
Lasso	Sepsis-3	78.2	2.565	0.152	0.000
SQI	Sepsis-3	70.1	2.460	0.136	0.000
Temporal MI	Sepsis-3	78.4	2.565	0.152	0.000
Random	Sepsis-3	70.9	2.460	0.138	0.000
All channels	Sepsis-3	76.1	3.395	0.148	0.000

Table 105: Channel selection — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Budget-specific rows are shown; the figure pools available 2/4/8-channel results per policy.

Condition	Task	AUPRC	ECE	AURC
All channels	VCS05	0.605	0.010	0.009
All channels	CL07-002	0.039	0.014	0.027
All channels	H5-002	0.032	0.015	0.009
All channels	H5-030	0.080	0.051	0.039
All channels	L2-30	0.105	0.113	0.066
All channels	OX01-IOH	0.132	0.021	0.028
All channels	VA1-CO2	0.107	0.030	0.026
All channels	VA1-STIM	0.061	0.027	0.014
Lasso (2)	CL07-002	0.034	0.016	0.030
Lasso (2)	H5-002	0.021	0.018	0.012
Lasso (2)	H5-030	0.070	0.056	0.044
Lasso (2)	L2-30	0.084	0.146	0.057
Lasso (2)	OX01-IOH	0.088	0.022	0.041
Lasso (2)	VA1-CO2	0.078	0.038	0.033
Lasso (2)	VA1-STIM	0.038	0.030	0.017
Lasso (2)	VCS05	0.439	0.018	0.015
Lasso (4)	CL07-002	0.036	0.017	0.029
Lasso (4)	H5-002	0.026	0.018	0.011
Lasso (4)	H5-030	0.076	0.056	0.041
Lasso (4)	L2-30	0.099	0.148	0.057
Lasso (4)	OX01-IOH	0.102	0.023	0.037
Lasso (4)	VA1-CO2	0.096	0.037	0.030
Lasso (4)	VA1-STIM	0.045	0.030	0.016
Lasso (4)	VCS05	0.527	0.012	0.013
Lasso (8)	CL07-002	0.038	0.018	0.027
Lasso (8)	H5-002	0.031	0.017	0.009
Lasso (8)	H5-030	0.079	0.056	0.039
Lasso (8)	L2-30	0.118	0.131	0.057
Lasso (8)	OX01-IOH	0.120	0.023	0.030
Lasso (8)	VA1-CO2	0.108	0.031	0.028
Lasso (8)	VA1-STIM	0.056	0.032	0.015
Lasso (8)	VCS05	0.606	0.009	0.009
MI (2)	CL07-002	0.035	0.018	0.030
MI (2)	H5-002	0.021	0.019	0.012
MI (2)	H5-030	0.066	0.056	0.047
MI (2)	L2-30	0.093	0.146	0.057
MI (2)	OX01-IOH	0.095	0.025	0.039
MI (2)	VA1-CO2	0.068	0.039	0.037
MI (2)	VA1-STIM	0.040	0.028	0.017
MI (2)	VCS05	0.451	0.015	0.015

Table 105: Channel selection — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Budget-specific rows are shown; the figure pools available 2/4/8-channel results per policy. (Continued.)

Condition	Task	AUPRC	ECE	AURC
MI (4)	CL07-002	0.036	0.018	0.030
MI (4)	H5-002	0.025	0.016	0.010
MI (4)	H5-030	0.072	0.054	0.044
MI (4)	L2-30	0.104	0.140	0.055
MI (4)	OX01-IOH	0.111	0.022	0.035
MI (4)	VA1-CO2	0.090	0.036	0.031
MI (4)	VA1-STIM	0.048	0.031	0.015
MI (4)	VCS05	0.537	0.011	0.011
MI (8)	CL07-002	0.039	0.018	0.027
MI (8)	H5-002	0.031	0.013	0.009
MI (8)	H5-030	0.078	0.058	0.039
MI (8)	L2-30	0.097	0.122	0.063
MI (8)	OX01-IOH	0.124	0.019	0.030
MI (8)	VA1-CO2	0.105	0.033	0.028
MI (8)	VA1-STIM	0.056	0.030	0.015
MI (8)	VCS05	0.596	0.010	0.009
Random (2)	CL07-002	0.035	0.018	0.030
Random (2)	H5-002	0.019	0.018	0.013
Random (2)	H5-030	0.068	0.057	0.047
Random (2)	L2-30	0.097	0.150	0.057
Random (2)	OX01-IOH	0.077	0.023	0.045
Random (2)	VA1-CO2	0.062	0.040	0.039
Random (2)	VA1-STIM	0.036	0.028	0.020
Random (2)	VCS05	0.340	0.023	0.021
Random (4)	CL07-002	0.037	0.019	0.029
Random (4)	H5-002	0.024	0.017	0.011
Random (4)	H5-030	0.073	0.057	0.043
Random (4)	L2-30	0.104	0.146	0.057
Random (4)	OX01-IOH	0.090	0.026	0.041
Random (4)	VA1-CO2	0.083	0.036	0.033
Random (4)	VA1-STIM	0.042	0.027	0.018
Random (4)	VCS05	0.488	0.011	0.014
Random (8)	CL07-002	0.038	0.019	0.027
Random (8)	H5-002	0.029	0.017	0.010
Random (8)	H5-030	0.078	0.055	0.040
Random (8)	L2-30	0.108	0.123	0.063
Random (8)	OX01-IOH	0.114	0.021	0.032
Random (8)	VA1-CO2	0.102	0.031	0.028
Random (8)	VA1-STIM	0.054	0.027	0.017
Random (8)	VCS05	0.579	0.010	0.010

Table 105: Channel selection — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Budget-specific rows are shown; the figure pools available 2/4/8-channel results per policy. (Continued.)

Condition	Task	AUPRC	ECE	AURC
SQI (2)	CL07-002	0.029	0.016	0.032
SQI (2)	H5-002	0.016	0.018	0.014
SQI (2)	H5-030	0.067	0.056	0.044
SQI (2)	L2-30	0.098	0.156	0.049
SQI (2)	OX01-IOH	0.057	0.023	0.047
SQI (2)	VA1-CO2	0.072	0.040	0.035
SQI (2)	VA1-STIM	0.032	0.029	0.020
SQI (2)	VCS05	0.511	0.012	0.018
SQI (4)	CL07-002	0.041	0.016	0.024
SQI (4)	H5-002	0.018	0.018	0.013
SQI (4)	H5-030	0.062	0.054	0.050
SQI (4)	L2-30	0.131	0.154	0.045
SQI (4)	OX01-IOH	0.128	0.023	0.029
SQI (4)	VA1-CO2	0.102	0.036	0.028
SQI (4)	VA1-STIM	0.044	0.029	0.017
SQI (4)	VCS05	0.533	0.013	0.014
SQI (8)	CL07-002	0.037	0.016	0.027
SQI (8)	H5-002	0.025	0.016	0.010
SQI (8)	H5-030	0.076	0.053	0.046
SQI (8)	L2-30	0.094	0.124	0.057
SQI (8)	OX01-IOH	0.125	0.022	0.028
SQI (8)	VA1-CO2	0.102	0.034	0.028
SQI (8)	VA1-STIM	0.053	0.030	0.014
SQI (8)	VCS05	0.603	0.011	0.009
Temporal MI (2)	CL07-002	0.036	0.013	0.030
Temporal MI (2)	H5-002	0.022	0.016	0.011
Temporal MI (2)	H5-030	0.067	0.055	0.046
Temporal MI (2)	L2-30	0.093	0.152	0.054
Temporal MI (2)	OX01-IOH	0.095	0.019	0.039
Temporal MI (2)	VA1-CO2	0.071	0.038	0.037
Temporal MI (2)	VA1-STIM	0.042	0.029	0.016
Temporal MI (2)	VCS05	0.510	0.014	0.017
Temporal MI (4)	CL07-002	0.037	0.019	0.028
Temporal MI (4)	H5-002	0.027	0.016	0.010
Temporal MI (4)	H5-030	0.073	0.054	0.042
Temporal MI (4)	L2-30	0.101	0.147	0.055
Temporal MI (4)	OX01-IOH	0.111	0.023	0.034
Temporal MI (4)	VA1-CO2	0.090	0.037	0.032
Temporal MI (4)	VA1-STIM	0.051	0.031	0.015
Temporal MI (4)	VCS05	0.565	0.010	0.013

Table 105: Channel selection — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Budget-specific rows are shown; the figure pools available 2/4/8-channel results per policy. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Temporal MI (8)	CL07-002	0.038	0.016	0.027
Temporal MI (8)	H5-002	0.030	0.017	0.009
Temporal MI (8)	H5-030	0.076	0.052	0.040
Temporal MI (8)	L2-30	0.110	0.124	0.057
Temporal MI (8)	OX01-IOH	0.127	0.022	0.029
Temporal MI (8)	VA1-CO2	0.105	0.036	0.028
Temporal MI (8)	VA1-STIM	0.055	0.028	0.014
Temporal MI (8)	VCS05	0.602	0.011	0.009

Table 106: Channel selection — VitalDB. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Budget-specific rows are shown; the figure pools available 2/4/8-channel results per policy.

Condition	Task	Time (s)	ms/sample	Fail (%)
All channels	Shared telemetry	40.1	0.521	0.000
Lasso (2)	Shared telemetry	33.3	0.432	0.000
Lasso (4)	Shared telemetry	33.4	0.433	0.000
Lasso (8)	Shared telemetry	33.6	0.435	0.000
MI (2)	Shared telemetry	33.5	0.434	0.000
MI (4)	Shared telemetry	33.6	0.435	0.000
MI (8)	Shared telemetry	33.4	0.434	0.000
Random (2)	Shared telemetry	33.3	0.431	0.000
Random (4)	Shared telemetry	33.8	0.438	0.000
Random (8)	Shared telemetry	33.2	0.430	0.000
SQI (2)	Shared telemetry	33.4	0.432	0.000
SQI (4)	Shared telemetry	33.7	0.436	0.000
SQI (8)	Shared telemetry	33.4	0.432	0.000
Temporal MI (2)	Shared telemetry	33.6	0.435	0.000
Temporal MI (4)	Shared telemetry	33.6	0.435	0.000
Temporal MI (8)	Shared telemetry	33.3	0.431	0.000

Table 107: Missingness — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Zero-mask+Δ	Acute det.	0.037	0.238	0.372
GRU-D	Acute det.	0.043	0.096	0.291
Lead-agn. SSL	Acute det.	0.046	0.077	0.292
Rect. flow	Acute det.	0.039	0.123	0.277
TrigFlow	Acute det.	0.039	0.105	0.284
Drifting	Acute det.	0.039	0.146	0.308
GRIN	Acute det.	0.042	0.098	0.247
Raindrop	Acute det.	0.050	0.080	0.277
Pos. kriging	Acute det.	0.038	0.134	0.329
Zero-mask+Δ	Dyn. mort.	0.023	0.216	0.280
GRU-D	Dyn. mort.	0.031	0.125	0.298
Lead-agn. SSL	Dyn. mort.	0.020	0.162	0.378
Rect. flow	Dyn. mort.	0.027	0.218	0.338
TrigFlow	Dyn. mort.	0.026	0.223	0.369
Drifting	Dyn. mort.	0.026	0.217	0.318
GRIN	Dyn. mort.	0.026	0.247	0.226
Raindrop	Dyn. mort.	0.020	0.233	0.254
Pos. kriging	Dyn. mort.	0.026	0.166	0.340
Zero-mask+Δ	Resp. esc.	0.038	0.189	0.297
GRU-D	Resp. esc.	0.037	0.104	0.372
Lead-agn. SSL	Resp. esc.	0.043	0.164	0.365
Rect. flow	Resp. esc.	0.038	0.175	0.312
TrigFlow	Resp. esc.	0.034	0.149	0.455
Drifting	Resp. esc.	0.038	0.152	0.340
GRIN	Resp. esc.	0.036	0.117	0.425
Raindrop	Resp. esc.	0.036	0.146	0.533
Pos. kriging	Resp. esc.	0.040	0.105	0.368
Zero-mask+Δ	Circ. det.	0.056	0.243	0.212
GRU-D	Circ. det.	0.055	0.140	0.276
Lead-agn. SSL	Circ. det.	0.055	0.184	0.223
Rect. flow	Circ. det.	0.056	0.231	0.242
TrigFlow	Circ. det.	0.056	0.231	0.241
Drifting	Circ. det.	0.056	0.230	0.242
GRIN	Circ. det.	0.058	0.158	0.207
Raindrop	Circ. det.	0.056	0.150	0.202
Pos. kriging	Circ. det.	0.057	0.192	0.235
Zero-mask+Δ	Fluid resp.	0.318	0.066	0.362
GRU-D	Fluid resp.	0.318	0.137	0.494
Lead-agn. SSL	Fluid resp.	0.318	0.087	0.506
Rect. flow	Fluid resp.	0.322	0.068	0.418

Table 107: Missingness — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
TrigFlow	Fluid resp.	0.320	0.070	0.407
Drifting	Fluid resp.	0.319	0.110	0.509
GRIN	Fluid resp.	0.315	0.111	0.492
Raindrop	Fluid resp.	0.323	0.066	0.418
Pos. kriging	Fluid resp.	0.316	0.089	0.476
Zero-mask+Δ	AKI 2/3	0.110	0.124	0.379
GRU-D	AKI 2/3	0.112	0.126	0.366
Lead-agn. SSL	AKI 2/3	0.108	0.125	0.410
Rect. flow	AKI 2/3	0.115	0.187	0.193
TrigFlow	AKI 2/3	0.118	0.162	0.216
Drifting	AKI 2/3	0.109	0.172	0.306
GRIN	AKI 2/3	0.114	0.206	0.340
Raindrop	AKI 2/3	0.110	0.090	0.281
Pos. kriging	AKI 2/3	0.110	0.189	0.374
Zero-mask+Δ	Vent. lib.	0.156	0.227	0.394
GRU-D	Vent. lib.	0.164	0.211	0.407
Lead-agn. SSL	Vent. lib.	0.159	0.217	0.460
Rect. flow	Vent. lib.	0.156	0.199	0.423
TrigFlow	Vent. lib.	0.157	0.183	0.321
Drifting	Vent. lib.	0.154	0.183	0.472
GRIN	Vent. lib.	0.162	0.243	0.596
Raindrop	Vent. lib.	0.159	0.184	0.536
Pos. kriging	Vent. lib.	0.157	0.197	0.494
Zero-mask+Δ	Sepsis-3	0.039	0.174	0.308
GRU-D	Sepsis-3	0.041	0.147	0.269
Lead-agn. SSL	Sepsis-3	0.041	0.149	0.237
Rect. flow	Sepsis-3	0.040	0.164	0.341
TrigFlow	Sepsis-3	0.038	0.254	0.385
Drifting	Sepsis-3	0.040	0.159	0.342
GRIN	Sepsis-3	0.042	0.100	0.256
Raindrop	Sepsis-3	0.041	0.123	0.283
Pos. kriging	Sepsis-3	0.040	0.135	0.292
Architecture only	Acute det.	0.040	0.131	0.321
Architecture only	Dyn. mort.	0.021	0.244	0.329
Architecture only	Resp. esc.	0.039	0.235	0.320
Architecture only	Circ. det.	0.055	0.246	0.223
Architecture only	Fluid resp.	0.323	0.112	0.458
Architecture only	AKI 2/3	0.113	0.102	0.426
Architecture only	Vent. lib.	0.156	0.236	0.378
Architecture only	Sepsis-3	0.040	0.158	0.267

Table 108: Missingness — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Zero-mask+Δ	Acute det.	196.7	3.520	0.168	0.000
GRU-D	Acute det.	445.7	3.607	0.381	0.000
Lead-agn. SSL	Acute det.	565.7	6.031	0.484	0.000
Rect. flow	Acute det.	370.3	2.422	0.317	0.000
TrigFlow	Acute det.	406.8	2.495	0.348	0.000
Drifting	Acute det.	366.4	2.422	0.313	0.000
GRIN	Acute det.	342.2	4.301	0.293	0.000
Raindrop	Acute det.	291.0	5.899	0.249	0.000
Pos. kriging	Acute det.	214.9	4.350	0.184	0.000
Zero-mask+Δ	Dyn. mort.	384.2	4.164	0.294	0.000
GRU-D	Dyn. mort.	562.5	3.648	0.431	0.000
Lead-agn. SSL	Dyn. mort.	658.2	6.031	0.504	0.000
Rect. flow	Dyn. mort.	412.8	2.418	0.316	0.000
TrigFlow	Dyn. mort.	646.4	2.692	0.495	0.000
Drifting	Dyn. mort.	417.8	2.418	0.320	0.000
GRIN	Dyn. mort.	388.7	4.301	0.298	0.000
Raindrop	Dyn. mort.	324.9	5.899	0.249	0.000
Pos. kriging	Dyn. mort.	437.6	4.911	0.335	0.000
Zero-mask+Δ	Resp. esc.	210.3	3.517	0.166	0.000
GRU-D	Resp. esc.	523.9	3.600	0.412	0.000
Lead-agn. SSL	Resp. esc.	607.2	6.031	0.478	0.000
Rect. flow	Resp. esc.	457.1	2.456	0.360	0.000
TrigFlow	Resp. esc.	438.8	2.492	0.346	0.000
Drifting	Resp. esc.	399.4	2.418	0.314	0.000
GRIN	Resp. esc.	360.2	4.301	0.284	0.000
Raindrop	Resp. esc.	306.0	5.899	0.241	0.000
Pos. kriging	Resp. esc.	224.7	4.350	0.177	0.000
Zero-mask+Δ	Circ. det.	188.6	3.523	0.190	0.000
GRU-D	Circ. det.	489.3	3.607	0.492	0.000
Lead-agn. SSL	Circ. det.	509.0	6.031	0.512	0.000
Rect. flow	Circ. det.	329.9	2.422	0.332	0.000
TrigFlow	Circ. det.	362.1	2.496	0.364	0.000
Drifting	Circ. det.	448.6	2.573	0.451	0.000
GRIN	Circ. det.	296.9	4.304	0.299	0.000
Raindrop	Circ. det.	514.4	6.767	0.517	0.000
Pos. kriging	Circ. det.	211.7	4.352	0.213	0.000
Zero-mask+Δ	Fluid resp.	7.3	1.479	0.127	0.000
GRU-D	Fluid resp.	17.8	1.513	0.308	0.000
Lead-agn. SSL	Fluid resp.	12.9	2.533	0.223	0.000
Rect. flow	Fluid resp.	26.7	1.907	0.462	0.000

Table 108: Missingness — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
TrigFlow	Fluid resp.	14.5	1.815	0.250	0.000
Drifting	Fluid resp.	13.4	1.795	0.232	0.000
GRIN	Fluid resp.	11.5	2.841	0.199	0.000
Raindrop	Fluid resp.	26.9	2.498	0.465	0.000
Pos. kriging	Fluid resp.	7.5	1.601	0.130	0.000
Zero-mask+Δ	AKI 2/3	241.6	3.517	0.185	0.000
GRU-D	AKI 2/3	717.8	3.600	0.548	0.000
Lead-agn. SSL	AKI 2/3	788.0	6.031	0.602	0.000
Rect. flow	AKI 2/3	568.0	2.569	0.434	0.000
TrigFlow	AKI 2/3	454.9	2.492	0.347	0.000
Drifting	AKI 2/3	412.4	2.418	0.315	0.000
GRIN	AKI 2/3	378.8	4.301	0.289	0.000
Raindrop	AKI 2/3	594.0	6.761	0.454	0.000
Pos. kriging	AKI 2/3	232.5	4.343	0.178	0.000
Zero-mask+Δ	Vent. lib.	9.9	1.774	0.107	0.000
GRU-D	Vent. lib.	25.7	1.779	0.277	0.000
Lead-agn. SSL	Vent. lib.	18.4	2.798	0.197	0.000
Rect. flow	Vent. lib.	24.5	2.196	0.264	0.000
TrigFlow	Vent. lib.	24.6	2.151	0.264	0.000
Drifting	Vent. lib.	18.6	2.063	0.200	0.000
GRIN	Vent. lib.	16.6	3.107	0.179	0.000
Raindrop	Vent. lib.	12.3	2.474	0.132	0.000
Pos. kriging	Vent. lib.	10.1	1.875	0.109	0.000
Zero-mask+Δ	Sepsis-3	41.7	6.574	0.081	0.000
GRU-D	Sepsis-3	171.4	6.861	0.333	0.000
Lead-agn. SSL	Sepsis-3	90.6	7.600	0.176	0.000
Rect. flow	Sepsis-3	90.4	6.862	0.176	0.000
TrigFlow	Sepsis-3	145.0	7.424	0.282	0.000
Drifting	Sepsis-3	131.1	7.380	0.255	0.000
GRIN	Sepsis-3	78.7	7.931	0.153	0.000
Raindrop	Sepsis-3	54.6	7.296	0.106	0.000
Pos. kriging	Sepsis-3	42.4	6.676	0.082	0.000
Architecture only	Acute det.	127.9	3.395	0.109	0.000
Architecture only	Dyn. mort.	142.1	3.395	0.109	0.000
Architecture only	Resp. esc.	139.1	3.395	0.110	0.000
Architecture only	Circ. det.	131.2	3.395	0.132	0.000
Architecture only	Fluid resp.	38.7	3.395	0.668	0.000
Architecture only	AKI 2/3	158.7	3.395	0.121	0.000
Architecture only	Vent. lib.	45.5	3.395	0.489	0.000
Architecture only	Sepsis-3	89.1	3.395	0.173	0.000

Table 109: Missingness — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Architecture only	CL07-002	0.039	0.013	0.027
Architecture only	H5-002	0.032	0.014	0.009
Architecture only	H5-030	0.079	0.053	0.039
Architecture only	L2-30	0.101	0.110	0.061
Architecture only	OX01-IOH	0.133	0.021	0.028
Architecture only	VA1-CO2	0.106	0.029	0.027
Architecture only	VA1-STIM	0.060	0.028	0.015
Drifting	CL07-002	0.039	0.016	0.026
Drifting	H5-002	0.027	0.019	0.012
Drifting	H5-030	0.075	0.058	0.045
Drifting	L2-30	0.110	0.118	0.056
Drifting	OX01-IOH	0.129	0.023	0.029
Drifting	VA1-CO2	0.105	0.036	0.034
Drifting	VA1-STIM	0.046	0.030	0.021
GRIN	CL07-002	0.039	0.018	0.026
GRIN	H5-002	0.027	0.019	0.012
GRIN	H5-030	0.076	0.059	0.045
GRIN	L2-30	0.116	0.121	0.057
GRIN	OX01-IOH	0.125	0.025	0.029
GRIN	VA1-CO2	0.103	0.037	0.033
GRIN	VA1-STIM	0.049	0.029	0.020
GRU-D	CL07-002	0.039	0.017	0.026
GRU-D	H5-002	0.027	0.017	0.012
GRU-D	H5-030	0.075	0.055	0.045
GRU-D	L2-30	0.115	0.121	0.055
GRU-D	OX01-IOH	0.128	0.022	0.029
GRU-D	VA1-CO2	0.114	0.036	0.035
GRU-D	VA1-STIM	0.044	0.029	0.022
Lead-agn. SSL	CL07-002	0.039	0.015	0.026
Lead-agn. SSL	H5-002	0.027	0.018	0.012
Lead-agn. SSL	H5-030	0.074	0.059	0.046
Lead-agn. SSL	L2-30	0.117	0.116	0.056
Lead-agn. SSL	OX01-IOH	0.133	0.024	0.029
Lead-agn. SSL	VA1-CO2	0.105	0.039	0.034
Lead-agn. SSL	VA1-STIM	0.046	0.027	0.021
Pos. kriging	CL07-002	0.041	0.015	0.025
Pos. kriging	H5-002	0.026	0.018	0.012
Pos. kriging	H5-030	0.074	0.059	0.045
Pos. kriging	L2-30	0.101	0.126	0.060
Pos. kriging	OX01-IOH	0.127	0.024	0.028

Table 109: Missingness — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Pos. kriging	VA1-CO2	0.101	0.036	0.034
Pos. kriging	VA1-STIM	0.045	0.032	0.021
Raindrop	CL07-002	0.039	0.017	0.025
Raindrop	H5-002	0.029	0.015	0.010
Raindrop	H5-030	0.076	0.052	0.042
Raindrop	L2-30	0.093	0.113	0.069
Raindrop	OX01-IOH	0.135	0.021	0.028
Raindrop	VA1-CO2	0.103	0.027	0.030
Raindrop	VA1-STIM	0.058	0.027	0.014
Rect. flow	CL07-002	0.039	0.019	0.026
Rect. flow	H5-002	0.026	0.018	0.012
Rect. flow	H5-030	0.075	0.060	0.045
Rect. flow	L2-30	0.108	0.128	0.056
Rect. flow	OX01-IOH	0.126	0.024	0.029
Rect. flow	VA1-CO2	0.108	0.038	0.034
Rect. flow	VA1-STIM	0.046	0.033	0.020
TrigFlow	CL07-002	0.038	0.017	0.026
TrigFlow	H5-002	0.027	0.018	0.012
TrigFlow	H5-030	0.074	0.058	0.045
TrigFlow	L2-30	0.110	0.117	0.058
TrigFlow	OX01-IOH	0.129	0.021	0.029
TrigFlow	VA1-CO2	0.110	0.033	0.034
TrigFlow	VA1-STIM	0.051	0.029	0.019
Zero-mask+Δ	CL07-002	0.040	0.018	0.025
Zero-mask+Δ	H5-002	0.025	0.017	0.011
Zero-mask+Δ	H5-030	0.075	0.063	0.045
Zero-mask+Δ	L2-30	0.100	0.115	0.074
Zero-mask+Δ	OX01-IOH	0.134	0.020	0.030
Zero-mask+Δ	VA1-CO2	0.101	0.035	0.034
Zero-mask+Δ	VA1-STIM	0.049	0.030	0.020
Architecture only	VCS05	–	0.010	0.009
Drifting	VCS05	–	0.009	0.011
GRIN	VCS05	–	0.010	0.010
GRU-D	VCS05	–	0.011	0.011
Lead-agn. SSL	VCS05	–	0.015	0.011
Pos. kriging	VCS05	–	0.010	0.011
Raindrop	VCS05	–	0.009	0.010
Rect. flow	VCS05	–	0.013	0.011
TrigFlow	VCS05	–	0.014	0.010
Zero-mask+Δ	VCS05	–	0.012	0.010

Table 110: Missingness — VitalDB. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Fail (%)
Architecture only	Shared telemetry	0.000
Drifting	Shared telemetry	0.000
GRIN	Shared telemetry	0.000
GRU-D	Shared telemetry	0.000
Lead-agn. SSL	Shared telemetry	0.000
Pos. kriging	Shared telemetry	0.000
Raindrop	Shared telemetry	0.000
Rect. flow	Shared telemetry	0.000
TrigFlow	Shared telemetry	0.000
Zero-mask+Δ	Shared telemetry	0.000

Table 111: Training subjects — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Noise filter (10%)	Acute det.	0.037	0.187	0.227
Noise filter (20%)	Acute det.	0.038	0.149	0.234
Noise filter (30%)	Acute det.	0.039	0.191	0.339
Diversity (10%)	Acute det.	0.032	0.227	0.360
Diversity (20%)	Acute det.	0.035	0.187	0.321
Diversity (30%)	Acute det.	0.033	0.242	0.340
Div. cluster (10%)	Acute det.	0.031	0.129	0.314
Div. cluster (20%)	Acute det.	0.031	0.190	0.430
Div. cluster (30%)	Acute det.	0.031	0.199	0.442
NLDR (10%)	Acute det.	0.034	0.142	0.241
NLDR (20%)	Acute det.	0.034	0.141	0.279
NLDR (30%)	Acute det.	0.034	0.198	0.364
Grad. match (10%)	Acute det.	0.033	0.197	0.282
Grad. match (20%)	Acute det.	0.033	0.133	0.234
Grad. match (30%)	Acute det.	0.031	0.230	0.373
Train. dyn. (10%)	Acute det.	0.036	0.204	0.236
Train. dyn. (20%)	Acute det.	0.032	0.140	0.303
Train. dyn. (30%)	Acute det.	0.031	0.178	0.336
Task anchor (10%)	Acute det.	0.028	0.436	0.467
Task anchor (20%)	Acute det.	0.027	0.338	0.431
Task anchor (30%)	Acute det.	0.027	0.380	0.526
Noise filter (10%)	Dyn. mort.	0.017	0.214	0.087
Noise filter (20%)	Dyn. mort.	0.019	0.351	0.143
Noise filter (30%)	Dyn. mort.	0.024	0.372	0.154
Diversity (10%)	Dyn. mort.	0.017	0.323	0.208
Diversity (20%)	Dyn. mort.	0.018	0.220	0.244
Diversity (30%)	Dyn. mort.	0.018	0.306	0.297
Div. cluster (10%)	Dyn. mort.	0.020	0.260	0.160
Div. cluster (20%)	Dyn. mort.	0.017	0.248	0.249
Div. cluster (30%)	Dyn. mort.	0.019	0.291	0.406
NLDR (10%)	Dyn. mort.	0.017	0.317	0.197
NLDR (20%)	Dyn. mort.	0.018	0.128	0.313
NLDR (30%)	Dyn. mort.	0.020	0.306	0.179
Grad. match (10%)	Dyn. mort.	0.016	0.214	0.232
Grad. match (20%)	Dyn. mort.	0.014	0.272	0.239
Grad. match (30%)	Dyn. mort.	0.015	0.262	0.349
Train. dyn. (10%)	Dyn. mort.	0.016	0.014	0.013
Train. dyn. (20%)	Dyn. mort.	0.017	0.244	0.118
Train. dyn. (30%)	Dyn. mort.	0.019	0.239	0.143
Task anchor (10%)	Dyn. mort.	0.015	0.481	0.392

Table 111: Training subjects — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Task anchor (20%)	Dyn. mort.	0.014	0.527	0.440
Task anchor (30%)	Dyn. mort.	0.013	0.425	0.382
Noise filter (10%)	Resp. esc.	0.032	0.287	0.440
Noise filter (20%)	Resp. esc.	0.032	0.344	0.382
Noise filter (30%)	Resp. esc.	0.034	0.275	0.621
Diversity (10%)	Resp. esc.	0.032	0.282	0.426
Diversity (20%)	Resp. esc.	0.032	0.257	0.518
Diversity (30%)	Resp. esc.	0.033	0.238	0.362
Div. cluster (10%)	Resp. esc.	0.033	0.228	0.214
Div. cluster (20%)	Resp. esc.	0.033	0.319	0.506
Div. cluster (30%)	Resp. esc.	0.034	0.226	0.357
NLDR (10%)	Resp. esc.	0.031	0.251	0.387
NLDR (20%)	Resp. esc.	0.032	0.258	0.372
NLDR (30%)	Resp. esc.	0.032	0.208	0.294
Grad. match (10%)	Resp. esc.	0.032	0.243	0.329
Grad. match (20%)	Resp. esc.	0.033	0.296	0.426
Grad. match (30%)	Resp. esc.	0.036	0.255	0.367
Train. dyn. (10%)	Resp. esc.	0.027	0.326	0.204
Train. dyn. (20%)	Resp. esc.	0.028	0.204	0.365
Train. dyn. (30%)	Resp. esc.	0.030	0.215	0.513
Task anchor (10%)	Resp. esc.	0.032	0.249	0.324
Task anchor (20%)	Resp. esc.	0.033	0.332	0.303
Task anchor (30%)	Resp. esc.	0.033	0.305	0.425
Noise filter (10%)	Circ. det.	0.048	0.342	0.118
Noise filter (20%)	Circ. det.	0.045	0.230	0.167
Noise filter (30%)	Circ. det.	0.048	0.237	0.139
Diversity (10%)	Circ. det.	0.036	0.263	0.178
Diversity (20%)	Circ. det.	0.036	0.220	0.200
Diversity (30%)	Circ. det.	0.039	0.240	0.219
Div. cluster (10%)	Circ. det.	0.044	0.261	0.192
Div. cluster (20%)	Circ. det.	0.054	0.219	0.237
Div. cluster (30%)	Circ. det.	0.048	0.228	0.265
NLDR (10%)	Circ. det.	0.041	0.217	0.162
NLDR (20%)	Circ. det.	0.051	0.313	0.154
NLDR (30%)	Circ. det.	0.049	0.273	0.136
Grad. match (10%)	Circ. det.	0.034	0.110	0.175
Grad. match (20%)	Circ. det.	0.031	0.219	0.386
Grad. match (30%)	Circ. det.	0.035	0.099	0.272
Train. dyn. (10%)	Circ. det.	0.044	0.406	0.178
Train. dyn. (20%)	Circ. det.	0.044	0.428	0.219

Table 111: Training subjects — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Train. dyn. (30%)	Circ. det.	0.042	0.370	0.160
Task anchor (10%)	Circ. det.	0.028	0.471	0.515
Task anchor (20%)	Circ. det.	0.024	0.433	0.479
Task anchor (30%)	Circ. det.	0.022	0.401	0.524
Noise filter (10%)	Fluid resp.	0.303	0.175	0.478
Noise filter (20%)	Fluid resp.	0.313	0.160	0.512
Noise filter (30%)	Fluid resp.	0.319	0.123	0.494
Diversity (10%)	Fluid resp.	0.316	0.141	0.454
Diversity (20%)	Fluid resp.	0.324	0.139	0.445
Diversity (30%)	Fluid resp.	0.314	0.122	0.440
Div. cluster (10%)	Fluid resp.	0.330	0.132	0.458
Div. cluster (20%)	Fluid resp.	0.335	0.139	0.512
Div. cluster (30%)	Fluid resp.	0.324	0.148	0.538
NLDR (10%)	Fluid resp.	0.311	0.147	0.544
NLDR (20%)	Fluid resp.	0.310	0.132	0.468
NLDR (30%)	Fluid resp.	0.312	0.086	0.466
Grad. match (10%)	Fluid resp.	0.312	0.117	0.397
Grad. match (20%)	Fluid resp.	0.316	0.102	0.360
Grad. match (30%)	Fluid resp.	0.311	0.118	0.462
Train. dyn. (10%)	Fluid resp.	0.319	0.096	0.285
Train. dyn. (20%)	Fluid resp.	0.307	0.142	0.305
Train. dyn. (30%)	Fluid resp.	0.310	0.160	0.374
Task anchor (10%)	Fluid resp.	0.310	0.238	0.370
Task anchor (20%)	Fluid resp.	0.309	0.293	0.396
Task anchor (30%)	Fluid resp.	0.316	0.199	0.401
Noise filter (10%)	AKI 2/3	0.106	0.218	0.520
Noise filter (20%)	AKI 2/3	0.108	0.095	0.298
Noise filter (30%)	AKI 2/3	0.105	0.200	0.523
Diversity (10%)	AKI 2/3	0.108	0.162	0.188
Diversity (20%)	AKI 2/3	0.105	0.221	0.375
Diversity (30%)	AKI 2/3	0.107	0.172	0.387
Div. cluster (10%)	AKI 2/3	0.112	0.173	0.278
Div. cluster (20%)	AKI 2/3	0.108	0.152	0.172
Div. cluster (30%)	AKI 2/3	0.116	0.122	0.311
NLDR (10%)	AKI 2/3	0.100	0.175	0.316
NLDR (20%)	AKI 2/3	0.104	0.153	0.262
NLDR (30%)	AKI 2/3	0.103	0.116	0.204
Grad. match (10%)	AKI 2/3	0.103	0.126	0.113
Grad. match (20%)	AKI 2/3	0.102	0.108	0.120
Grad. match (30%)	AKI 2/3	0.098	0.130	0.118

Table 111: Training subjects — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Train. dyn. (10%)	AKI 2/3	0.105	0.416	0.304
Train. dyn. (20%)	AKI 2/3	0.108	0.359	0.237
Train. dyn. (30%)	AKI 2/3	0.105	0.295	0.224
Task anchor (10%)	AKI 2/3	0.092	0.475	0.658
Task anchor (20%)	AKI 2/3	0.097	0.398	0.508
Task anchor (30%)	AKI 2/3	0.090	0.265	0.429
Noise filter (10%)	Vent. lib.	0.147	0.226	0.462
Noise filter (20%)	Vent. lib.	0.146	0.328	0.668
Noise filter (30%)	Vent. lib.	0.141	0.261	0.552
Diversity (10%)	Vent. lib.	0.151	0.288	0.550
Diversity (20%)	Vent. lib.	0.161	0.313	0.650
Diversity (30%)	Vent. lib.	0.152	0.341	0.703
Div. cluster (10%)	Vent. lib.	0.141	0.336	0.607
Div. cluster (20%)	Vent. lib.	0.151	0.205	0.279
Div. cluster (30%)	Vent. lib.	0.153	0.277	0.427
NLDR (10%)	Vent. lib.	0.152	0.282	0.562
NLDR (20%)	Vent. lib.	0.158	0.227	0.482
NLDR (30%)	Vent. lib.	0.152	0.236	0.523
Grad. match (10%)	Vent. lib.	0.137	0.282	0.581
Grad. match (20%)	Vent. lib.	0.140	0.269	0.498
Grad. match (30%)	Vent. lib.	0.139	0.223	0.492
Train. dyn. (10%)	Vent. lib.	0.142	0.253	0.362
Train. dyn. (20%)	Vent. lib.	0.134	0.284	0.487
Train. dyn. (30%)	Vent. lib.	0.144	0.319	0.525
Task anchor (10%)	Vent. lib.	0.140	0.142	0.216
Task anchor (20%)	Vent. lib.	0.146	0.200	0.281
Task anchor (30%)	Vent. lib.	0.154	0.250	0.264
Noise filter (10%)	Sepsis-3	0.033	0.316	0.136
Noise filter (20%)	Sepsis-3	0.031	0.088	0.304
Noise filter (30%)	Sepsis-3	0.032	0.258	0.198
Diversity (10%)	Sepsis-3	0.033	0.301	0.357
Diversity (20%)	Sepsis-3	0.040	0.361	0.335
Diversity (30%)	Sepsis-3	0.037	0.260	0.345
Div. cluster (10%)	Sepsis-3	0.031	0.267	0.196
Div. cluster (20%)	Sepsis-3	0.033	0.232	0.231
Div. cluster (30%)	Sepsis-3	0.037	0.314	0.355
NLDR (10%)	Sepsis-3	0.034	0.238	0.140
NLDR (20%)	Sepsis-3	0.032	0.283	0.319
NLDR (30%)	Sepsis-3	0.035	0.226	0.153
Grad. match (10%)	Sepsis-3	0.032	0.255	0.380

Table 111: Training subjects — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Grad. match (20%)	Sepsis-3	0.036	0.292	0.190
Grad. match (30%)	Sepsis-3	0.030	0.316	0.519
Train. dyn. (10%)	Sepsis-3	0.025	0.041	0.027
Train. dyn. (20%)	Sepsis-3	0.025	0.084	0.130
Train. dyn. (30%)	Sepsis-3	0.023	0.176	0.063
Task anchor (10%)	Sepsis-3	0.021	0.491	0.492
Task anchor (20%)	Sepsis-3	0.023	0.445	0.388
Task anchor (30%)	Sepsis-3	0.024	0.482	0.466
Full cohort (100%)	Acute det.	0.040	0.131	0.321
Full cohort (100%)	Dyn. mort.	0.025	0.214	0.290
Full cohort (100%)	Resp. esc.	0.039	0.235	0.320
Full cohort (100%)	Circ. det.	0.050	0.161	0.225
Full cohort (100%)	Fluid resp.	0.323	0.112	0.458
Full cohort (100%)	AKI 2/3	0.113	0.102	0.426
Full cohort (100%)	Vent. lib.	0.156	0.236	0.378
Full cohort (100%)	Sepsis-3	0.038	0.230	0.223

Table 112: Training subjects — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Noise filter (10%)	Acute det.	30.5	3.391	0.261	0.000
Noise filter (20%)	Acute det.	42.8	3.391	0.183	0.000
Noise filter (30%)	Acute det.	55.6	3.391	0.158	0.000
Diversity (10%)	Acute det.	29.6	3.391	0.253	0.000
Diversity (20%)	Acute det.	43.7	3.391	0.186	0.000
Diversity (30%)	Acute det.	57.4	3.391	0.164	0.000
Div. cluster (10%)	Acute det.	33.8	3.391	0.260	0.000
Div. cluster (20%)	Acute det.	46.5	3.391	0.188	0.000
Div. cluster (30%)	Acute det.	60.0	3.391	0.166	0.000
NLDR (10%)	Acute det.	31.3	3.391	0.267	0.000
NLDR (20%)	Acute det.	43.4	3.391	0.185	0.000
NLDR (30%)	Acute det.	56.9	3.391	0.162	0.000
Grad. match (10%)	Acute det.	30.8	3.391	0.263	0.000
Grad. match (20%)	Acute det.	44.1	3.391	0.188	0.000
Grad. match (30%)	Acute det.	57.1	3.391	0.163	0.000
Train. dyn. (10%)	Acute det.	30.7	3.391	0.262	0.000
Train. dyn. (20%)	Acute det.	43.9	3.391	0.188	0.000
Train. dyn. (30%)	Acute det.	58.6	3.391	0.167	0.000
Task anchor (10%)	Acute det.	30.1	3.391	0.257	0.000
Task anchor (20%)	Acute det.	43.7	3.391	0.186	0.000
Task anchor (30%)	Acute det.	57.2	3.391	0.163	0.000
Noise filter (10%)	Dyn. mort.	65.9	3.805	0.505	0.000
Noise filter (20%)	Dyn. mort.	91.9	3.805	0.352	0.000
Noise filter (30%)	Dyn. mort.	114.4	3.780	0.292	0.000
Diversity (10%)	Dyn. mort.	35.8	3.391	0.274	0.000
Diversity (20%)	Dyn. mort.	51.0	3.392	0.194	0.000
Diversity (30%)	Dyn. mort.	68.0	3.391	0.174	0.000
Div. cluster (10%)	Dyn. mort.	39.9	3.391	0.260	0.000
Div. cluster (20%)	Dyn. mort.	56.8	3.391	0.194	0.000
Div. cluster (30%)	Dyn. mort.	70.2	3.391	0.172	0.000
NLDR (10%)	Dyn. mort.	38.3	3.391	0.293	0.000
NLDR (20%)	Dyn. mort.	53.4	3.391	0.205	0.000
NLDR (30%)	Dyn. mort.	78.8	3.416	0.201	0.000
Grad. match (10%)	Dyn. mort.	60.2	3.780	0.457	0.000
Grad. match (20%)	Dyn. mort.	51.7	3.391	0.198	0.000
Grad. match (30%)	Dyn. mort.	68.2	3.391	0.174	0.000
Train. dyn. (10%)	Dyn. mort.	68.8	3.780	0.523	0.000
Train. dyn. (20%)	Dyn. mort.	58.7	3.416	0.225	0.000
Train. dyn. (30%)	Dyn. mort.	73.4	3.416	0.187	0.000
Task anchor (10%)	Dyn. mort.	36.8	3.391	0.281	0.000

Table 112: Training subjects — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Task anchor (20%)	Dyn. mort.	54.0	3.391	0.206	0.000
Task anchor (30%)	Dyn. mort.	78.8	3.416	0.201	0.000
Noise filter (10%)	Resp. esc.	32.6	3.391	0.256	0.000
Noise filter (20%)	Resp. esc.	48.2	3.391	0.189	0.000
Noise filter (30%)	Resp. esc.	58.6	3.391	0.154	0.000
Diversity (10%)	Resp. esc.	33.0	3.391	0.258	0.000
Diversity (20%)	Resp. esc.	47.1	3.391	0.185	0.000
Diversity (30%)	Resp. esc.	63.4	3.391	0.166	0.000
Div. cluster (10%)	Resp. esc.	33.8	3.391	0.250	0.000
Div. cluster (20%)	Resp. esc.	49.6	3.391	0.185	0.000
Div. cluster (30%)	Resp. esc.	63.3	3.391	0.159	0.000
NLDR (10%)	Resp. esc.	33.6	3.393	0.264	0.000
NLDR (20%)	Resp. esc.	47.0	3.392	0.185	0.000
NLDR (30%)	Resp. esc.	59.4	3.391	0.156	0.000
Grad. match (10%)	Resp. esc.	32.9	3.391	0.258	0.000
Grad. match (20%)	Resp. esc.	46.1	3.391	0.181	0.000
Grad. match (30%)	Resp. esc.	61.8	3.391	0.162	0.000
Train. dyn. (10%)	Resp. esc.	32.9	3.391	0.257	0.000
Train. dyn. (20%)	Resp. esc.	46.8	3.391	0.184	0.000
Train. dyn. (30%)	Resp. esc.	61.4	3.391	0.161	0.000
Task anchor (10%)	Resp. esc.	32.3	3.391	0.253	0.000
Task anchor (20%)	Resp. esc.	47.0	3.391	0.185	0.000
Task anchor (30%)	Resp. esc.	61.4	3.391	0.160	0.000
Noise filter (10%)	Circ. det.	31.3	3.391	0.314	0.000
Noise filter (20%)	Circ. det.	40.8	3.391	0.205	0.000
Noise filter (30%)	Circ. det.	53.5	3.391	0.179	0.000
Diversity (10%)	Circ. det.	29.9	3.391	0.298	0.000
Diversity (20%)	Circ. det.	41.5	3.391	0.208	0.000
Diversity (30%)	Circ. det.	54.3	3.391	0.182	0.000
Div. cluster (10%)	Circ. det.	32.5	3.391	0.256	0.000
Div. cluster (20%)	Circ. det.	43.3	3.391	0.191	0.000
Div. cluster (30%)	Circ. det.	53.9	3.391	0.168	0.000
NLDR (10%)	Circ. det.	29.6	3.391	0.297	0.000
NLDR (20%)	Circ. det.	40.5	3.391	0.203	0.000
NLDR (30%)	Circ. det.	52.1	3.391	0.173	0.000
Grad. match (10%)	Circ. det.	29.9	3.391	0.299	0.000
Grad. match (20%)	Circ. det.	40.3	3.391	0.203	0.000
Grad. match (30%)	Circ. det.	51.0	3.391	0.171	0.000
Train. dyn. (10%)	Circ. det.	29.4	3.391	0.294	0.000
Train. dyn. (20%)	Circ. det.	40.6	3.391	0.204	0.000

Table 112: Training subjects — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Train. dyn. (30%)	Circ. det.	51.8	3.391	0.172	0.000
Task anchor (10%)	Circ. det.	30.2	3.391	0.302	0.000
Task anchor (20%)	Circ. det.	39.7	3.391	0.200	0.000
Task anchor (30%)	Circ. det.	51.2	3.391	0.169	0.000
Noise filter (10%)	Fluid resp.	4.9	3.200	0.853	0.000
Noise filter (20%)	Fluid resp.	6.1	3.196	0.519	0.000
Noise filter (30%)	Fluid resp.	6.2	3.172	0.359	0.000
Diversity (10%)	Fluid resp.	5.1	3.176	0.873	0.000
Diversity (20%)	Fluid resp.	5.6	3.197	0.482	0.000
Diversity (30%)	Fluid resp.	6.3	3.203	0.363	0.000
Div. cluster (10%)	Fluid resp.	5.5	3.169	0.824	0.000
Div. cluster (20%)	Fluid resp.	5.8	3.172	0.473	0.000
Div. cluster (30%)	Fluid resp.	6.8	3.204	0.373	0.000
NLDR (10%)	Fluid resp.	5.0	3.210	0.862	0.000
NLDR (20%)	Fluid resp.	5.7	3.177	0.491	0.000
NLDR (30%)	Fluid resp.	6.2	3.185	0.357	0.000
Grad. match (10%)	Fluid resp.	5.1	3.209	0.887	0.000
Grad. match (20%)	Fluid resp.	5.4	3.216	0.470	0.000
Grad. match (30%)	Fluid resp.	6.3	3.141	0.361	0.000
Train. dyn. (10%)	Fluid resp.	4.9	3.196	0.835	0.000
Train. dyn. (20%)	Fluid resp.	5.5	3.196	0.478	0.000
Train. dyn. (30%)	Fluid resp.	6.3	3.177	0.361	0.000
Task anchor (10%)	Fluid resp.	4.8	3.196	0.835	0.000
Task anchor (20%)	Fluid resp.	5.7	3.177	0.494	0.000
Task anchor (30%)	Fluid resp.	6.3	3.199	0.363	0.000
Noise filter (10%)	AKI 2/3	32.3	3.391	0.247	0.000
Noise filter (20%)	AKI 2/3	47.8	3.391	0.182	0.000
Noise filter (30%)	AKI 2/3	60.2	3.391	0.153	0.000
Diversity (10%)	AKI 2/3	32.0	3.391	0.244	0.000
Diversity (20%)	AKI 2/3	46.7	3.391	0.178	0.000
Diversity (30%)	AKI 2/3	61.6	3.391	0.157	0.000
Div. cluster (10%)	AKI 2/3	34.0	3.391	0.228	0.000
Div. cluster (20%)	AKI 2/3	46.6	3.391	0.173	0.000
Div. cluster (30%)	AKI 2/3	62.1	3.391	0.152	0.000
NLDR (10%)	AKI 2/3	32.8	3.392	0.249	0.000
NLDR (20%)	AKI 2/3	48.3	3.391	0.182	0.000
NLDR (30%)	AKI 2/3	62.8	3.391	0.160	0.000
Grad. match (10%)	AKI 2/3	32.3	3.391	0.246	0.000
Grad. match (20%)	AKI 2/3	46.9	3.391	0.177	0.000
Grad. match (30%)	AKI 2/3	61.2	3.391	0.155	0.000

Table 112: Training subjects — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Train. dyn. (10%)	AKI 2/3	33.7	3.391	0.256	0.000
Train. dyn. (20%)	AKI 2/3	46.9	3.391	0.179	0.000
Train. dyn. (30%)	AKI 2/3	62.5	3.391	0.159	0.000
Task anchor (10%)	AKI 2/3	35.0	3.391	0.261	0.000
Task anchor (20%)	AKI 2/3	46.8	3.391	0.178	0.000
Task anchor (30%)	AKI 2/3	62.4	3.391	0.158	0.000
Noise filter (10%)	Vent. lib.	5.5	3.190	0.589	0.000
Noise filter (20%)	Vent. lib.	7.1	3.234	0.381	0.000
Noise filter (30%)	Vent. lib.	8.3	3.277	0.296	0.000
Diversity (10%)	Vent. lib.	5.7	3.175	0.612	0.000
Diversity (20%)	Vent. lib.	7.3	3.242	0.391	0.000
Diversity (30%)	Vent. lib.	8.3	3.201	0.297	0.000
Div. cluster (10%)	Vent. lib.	6.2	2.475	0.556	0.000
Div. cluster (20%)	Vent. lib.	7.5	3.221	0.369	0.000
Div. cluster (30%)	Vent. lib.	8.2	3.219	0.282	0.000
NLDR (10%)	Vent. lib.	5.9	3.176	0.637	0.000
NLDR (20%)	Vent. lib.	6.5	2.939	0.350	0.000
NLDR (30%)	Vent. lib.	7.8	3.130	0.280	0.000
Grad. match (10%)	Vent. lib.	5.8	3.216	0.618	0.000
Grad. match (20%)	Vent. lib.	6.5	3.208	0.350	0.000
Grad. match (30%)	Vent. lib.	7.7	3.235	0.274	0.000
Train. dyn. (10%)	Vent. lib.	5.5	2.272	0.587	0.000
Train. dyn. (20%)	Vent. lib.	6.6	3.208	0.354	0.000
Train. dyn. (30%)	Vent. lib.	7.9	3.244	0.283	0.000
Task anchor (10%)	Vent. lib.	6.3	3.161	0.671	0.000
Task anchor (20%)	Vent. lib.	7.0	3.237	0.375	0.000
Task anchor (30%)	Vent. lib.	7.8	3.230	0.279	0.000
Noise filter (10%)	Sepsis-3	17.2	3.391	0.334	0.000
Noise filter (20%)	Sepsis-3	23.2	3.391	0.225	0.000
Noise filter (30%)	Sepsis-3	29.2	3.391	0.189	0.000
Diversity (10%)	Sepsis-3	17.2	3.392	0.332	0.000
Diversity (20%)	Sepsis-3	22.9	3.391	0.222	0.000
Diversity (30%)	Sepsis-3	28.2	3.391	0.182	0.000
Div. cluster (10%)	Sepsis-3	17.9	3.391	0.308	0.000
Div. cluster (20%)	Sepsis-3	23.8	3.393	0.224	0.000
Div. cluster (30%)	Sepsis-3	29.0	3.391	0.182	0.000
NLDR (10%)	Sepsis-3	17.3	3.391	0.335	0.000
NLDR (20%)	Sepsis-3	22.9	3.391	0.223	0.000
NLDR (30%)	Sepsis-3	28.7	3.391	0.183	0.000
Grad. match (10%)	Sepsis-3	17.2	3.391	0.333	0.000

Table 112: Training subjects — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Grad. match (20%)	Sepsis-3	22.5	3.393	0.218	0.000
Grad. match (30%)	Sepsis-3	28.0	3.391	0.181	0.000
Train. dyn. (10%)	Sepsis-3	17.2	3.391	0.334	0.000
Train. dyn. (20%)	Sepsis-3	22.7	3.398	0.221	0.000
Train. dyn. (30%)	Sepsis-3	28.2	3.391	0.183	0.000
Task anchor (10%)	Sepsis-3	17.1	3.391	0.332	0.000
Task anchor (20%)	Sepsis-3	23.5	3.391	0.223	0.000
Task anchor (30%)	Sepsis-3	28.4	3.391	0.184	0.000
Full cohort (100%)	Acute det.	127.9	3.395	0.109	0.000
Full cohort (100%)	Dyn. mort.	148.3	3.395	0.114	0.000
Full cohort (100%)	Resp. esc.	139.1	3.395	0.110	0.000
Full cohort (100%)	Circ. det.	134.5	3.395	0.135	0.000
Full cohort (100%)	Fluid resp.	38.7	3.395	0.668	0.000
Full cohort (100%)	AKI 2/3	158.7	3.395	0.121	0.000
Full cohort (100%)	Vent. lib.	45.5	3.395	0.489	0.000
Full cohort (100%)	Sepsis-3	76.1	3.395	0.148	0.000

Table 113: Training subjects — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Baseline (5%)	CL07-002	0.036	0.026	0.035
Baseline (5%)	H5-002	0.022	0.014	0.012
Baseline (5%)	H5-030	0.067	0.062	0.060
Baseline (5%)	L2-30	0.082	0.137	0.116
Baseline (5%)	OX01-IOH	0.087	0.042	0.048
Baseline (5%)	VA1-CO2	0.081	0.039	0.039
Baseline (5%)	VA1-STIM	0.050	0.029	0.021
Baseline (5%)	VCS05	0.501	0.010	0.016
Baseline (10%)	CL07-002	0.037	0.020	0.030
Baseline (10%)	H5-002	0.023	0.014	0.011
Baseline (10%)	H5-030	0.069	0.060	0.049
Baseline (10%)	L2-30	0.082	0.105	0.090
Baseline (10%)	OX01-IOH	0.097	0.028	0.038
Baseline (10%)	VA1-CO2	0.086	0.033	0.033
Baseline (10%)	VA1-STIM	0.052	0.025	0.016
Baseline (10%)	VCS05	0.532	0.012	0.015
Baseline (20%)	CL07-002	0.038	0.017	0.029
Baseline (20%)	H5-002	0.024	0.013	0.011
Baseline (20%)	H5-030	0.070	0.056	0.045
Baseline (20%)	L2-30	0.097	0.100	0.068
Baseline (20%)	OX01-IOH	0.107	0.023	0.034
Baseline (20%)	VA1-CO2	0.090	0.030	0.030
Baseline (20%)	VA1-STIM	0.057	0.028	0.015
Baseline (20%)	VCS05	0.560	0.008	0.013
Diversity (5%)	CL07-002	0.033	0.048	0.043
Diversity (5%)	H5-002	0.024	0.032	0.015
Diversity (5%)	H5-030	0.067	0.103	0.055
Diversity (5%)	L2-30	0.075	0.151	0.151
Diversity (5%)	OX01-IOH	0.098	0.046	0.041
Diversity (5%)	VA1-CO2	0.071	0.073	0.059
Diversity (5%)	VA1-STIM	0.042	0.064	0.036
Diversity (5%)	VCS05	0.459	0.014	0.015
Diversity (10%)	CL07-002	0.036	0.031	0.030
Diversity (10%)	H5-002	0.027	0.027	0.013
Diversity (10%)	H5-030	0.070	0.106	0.050
Diversity (10%)	L2-30	0.074	0.143	0.134
Diversity (10%)	OX01-IOH	0.111	0.049	0.035
Diversity (10%)	VA1-CO2	0.074	0.062	0.039
Diversity (10%)	VA1-STIM	0.050	0.076	0.027
Diversity (10%)	VCS05	0.518	0.012	0.016

Table 113: Training subjects — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Diversity (20%)	CL07-002	0.037	0.024	0.029
Diversity (20%)	H5-002	0.023	0.027	0.010
Diversity (20%)	H5-030	0.074	0.108	0.041
Diversity (20%)	L2-30	0.078	0.111	0.117
Diversity (20%)	OX01-IOH	0.114	0.043	0.034
Diversity (20%)	VA1-CO2	0.090	0.069	0.030
Diversity (20%)	VA1-STIM	0.056	0.074	0.028
Diversity (20%)	VCS05	0.539	0.012	0.015
Div. cluster (5%)	CL07-002	0.037	0.032	0.033
Div. cluster (5%)	H5-002	0.020	0.009	0.013
Div. cluster (5%)	H5-030	0.064	0.061	0.061
Div. cluster (5%)	L2-30	0.088	0.151	0.098
Div. cluster (5%)	OX01-IOH	0.098	0.032	0.040
Div. cluster (5%)	VA1-CO2	0.072	0.050	0.043
Div. cluster (5%)	VA1-STIM	0.044	0.039	0.023
Div. cluster (5%)	VCS05	0.468	0.017	0.018
Div. cluster (10%)	CL07-002	0.038	0.021	0.030
Div. cluster (10%)	H5-002	0.024	0.011	0.012
Div. cluster (10%)	H5-030	0.065	0.066	0.053
Div. cluster (10%)	L2-30	0.094	0.105	0.079
Div. cluster (10%)	OX01-IOH	0.109	0.028	0.036
Div. cluster (10%)	VA1-CO2	0.079	0.041	0.032
Div. cluster (10%)	VA1-STIM	0.048	0.039	0.018
Div. cluster (10%)	VCS05	0.519	0.013	0.017
Div. cluster (20%)	CL07-002	0.037	0.017	0.028
Div. cluster (20%)	H5-002	0.026	0.015	0.011
Div. cluster (20%)	H5-030	0.068	0.071	0.049
Div. cluster (20%)	L2-30	0.117	0.101	0.065
Div. cluster (20%)	OX01-IOH	0.120	0.028	0.033
Div. cluster (20%)	VA1-CO2	0.085	0.040	0.031
Div. cluster (20%)	VA1-STIM	0.052	0.037	0.017
Div. cluster (20%)	VCS05	0.549	0.011	0.013
Full cohort (100%)	CL07-002	0.039	0.015	0.027
Full cohort (100%)	H5-002	0.032	0.014	0.009
Full cohort (100%)	H5-030	0.079	0.057	0.039
Full cohort (100%)	L2-30	0.107	0.120	0.062
Full cohort (100%)	OX01-IOH	0.131	0.021	0.027
Full cohort (100%)	VA1-CO2	0.107	0.029	0.027
Full cohort (100%)	VA1-STIM	0.062	0.032	0.014
Full cohort (100%)	VCS05	0.607	0.009	0.009

Table 113: Training subjects — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Grad. match (5%)	CL07-002	0.039	0.036	0.041
Grad. match (5%)	H5-002	0.022	0.010	0.011
Grad. match (5%)	H5-030	0.074	0.014	0.045
Grad. match (5%)	L2-30	0.087	0.209	0.283
Grad. match (5%)	OX01-IOH	0.084	0.029	0.048
Grad. match (5%)	VA1-CO2	0.068	0.024	0.042
Grad. match (5%)	VA1-STIM	0.048	0.047	0.025
Grad. match (5%)	VCS05	0.550	0.015	0.010
Grad. match (10%)	CL07-002	0.038	0.014	0.026
Grad. match (10%)	H5-002	0.024	0.011	0.010
Grad. match (10%)	H5-030	0.075	0.011	0.045
Grad. match (10%)	L2-30	0.082	0.226	0.359
Grad. match (10%)	OX01-IOH	0.101	0.012	0.039
Grad. match (10%)	VA1-CO2	0.072	0.026	0.036
Grad. match (10%)	VA1-STIM	0.051	0.022	0.015
Grad. match (10%)	VCS05	0.572	0.015	0.009
Grad. match (20%)	CL07-002	0.038	0.012	0.028
Grad. match (20%)	H5-002	0.028	0.010	0.009
Grad. match (20%)	H5-030	0.085	0.011	0.038
Grad. match (20%)	L2-30	0.069	0.187	0.324
Grad. match (20%)	OX01-IOH	0.118	0.013	0.036
Grad. match (20%)	VA1-CO2	0.091	0.039	0.028
Grad. match (20%)	VA1-STIM	0.056	0.015	0.013
Grad. match (20%)	VCS05	0.589	0.014	0.009
NLDR (5%)	CL07-002	0.037	0.036	0.033
NLDR (5%)	H5-002	0.020	0.018	0.017
NLDR (5%)	H5-030	0.065	0.060	0.070
NLDR (5%)	L2-30	0.100	0.136	0.125
NLDR (5%)	OX01-IOH	0.089	0.042	0.046
NLDR (5%)	VA1-CO2	0.080	0.044	0.036
NLDR (5%)	VA1-STIM	0.043	0.034	0.025
NLDR (5%)	VCS05	0.464	0.015	0.017
NLDR (10%)	CL07-002	0.038	0.020	0.029
NLDR (10%)	H5-002	0.023	0.017	0.013
NLDR (10%)	H5-030	0.064	0.081	0.054
NLDR (10%)	L2-30	0.086	0.110	0.111
NLDR (10%)	OX01-IOH	0.101	0.031	0.037
NLDR (10%)	VA1-CO2	0.088	0.033	0.033
NLDR (10%)	VA1-STIM	0.052	0.042	0.018
NLDR (10%)	VCS05	0.510	0.011	0.013

Table 113: Training subjects — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
NLDR (20%)	CL07-002	0.036	0.016	0.029
NLDR (20%)	H5-002	0.025	0.017	0.010
NLDR (20%)	H5-030	0.065	0.085	0.048
NLDR (20%)	L2-30	0.085	0.094	0.077
NLDR (20%)	OX01-IOH	0.111	0.024	0.033
NLDR (20%)	VA1-CO2	0.090	0.035	0.030
NLDR (20%)	VA1-STIM	0.055	0.045	0.016
NLDR (20%)	VCS05	0.537	0.012	0.015
Noise filter (5%)	CL07-002	0.041	0.027	0.037
Noise filter (5%)	H5-002	0.021	0.011	0.014
Noise filter (5%)	H5-030	0.068	0.020	0.051
Noise filter (5%)	L2-30	0.083	0.093	0.097
Noise filter (5%)	OX01-IOH	0.079	0.031	0.045
Noise filter (5%)	VA1-CO2	0.067	0.021	0.039
Noise filter (5%)	VA1-STIM	0.047	0.022	0.016
Noise filter (5%)	VCS05	0.395	0.022	0.017
Noise filter (10%)	CL07-002	0.037	0.012	0.030
Noise filter (10%)	H5-002	0.020	0.009	0.013
Noise filter (10%)	H5-030	0.068	0.010	0.048
Noise filter (10%)	L2-30	0.081	0.074	0.083
Noise filter (10%)	OX01-IOH	0.094	0.013	0.037
Noise filter (10%)	VA1-CO2	0.078	0.015	0.035
Noise filter (10%)	VA1-STIM	0.058	0.019	0.013
Noise filter (10%)	VCS05	0.530	0.010	0.011
Noise filter (20%)	CL07-002	0.039	0.011	0.027
Noise filter (20%)	H5-002	0.025	0.012	0.010
Noise filter (20%)	H5-030	0.079	0.011	0.040
Noise filter (20%)	L2-30	0.095	0.061	0.053
Noise filter (20%)	OX01-IOH	0.113	0.011	0.033
Noise filter (20%)	VA1-CO2	0.091	0.027	0.029
Noise filter (20%)	VA1-STIM	0.062	0.020	0.013
Noise filter (20%)	VCS05	0.558	0.009	0.011
Task anchor (5%)	CL07-002	0.037	0.047	0.040
Task anchor (5%)	H5-002	0.021	0.035	0.013
Task anchor (5%)	H5-030	0.067	0.179	0.052
Task anchor (5%)	L2-30	0.088	0.506	0.658
Task anchor (5%)	OX01-IOH	0.078	0.057	0.069
Task anchor (5%)	VA1-CO2	0.101	0.071	0.036
Task anchor (5%)	VA1-STIM	0.048	0.049	0.023
Task anchor (5%)	VCS05	0.485	0.029	0.014

Table 113: Training subjects — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Task anchor (10%)	CL07-002	0.038	0.038	0.032
Task anchor (10%)	H5-002	0.023	0.040	0.012
Task anchor (10%)	H5-030	0.071	0.160	0.044
Task anchor (10%)	L2-30	0.108	0.302	0.442
Task anchor (10%)	OX01-IOH	0.088	0.057	0.041
Task anchor (10%)	VA1-CO2	0.095	0.074	0.033
Task anchor (10%)	VA1-STIM	0.054	0.054	0.019
Task anchor (10%)	VCS05	0.504	0.024	0.013
Task anchor (20%)	CL07-002	0.040	0.034	0.028
Task anchor (20%)	H5-002	0.025	0.036	0.011
Task anchor (20%)	H5-030	0.071	0.111	0.044
Task anchor (20%)	L2-30	0.104	0.121	0.204
Task anchor (20%)	OX01-IOH	0.104	0.044	0.034
Task anchor (20%)	VA1-CO2	0.104	0.061	0.029
Task anchor (20%)	VA1-STIM	0.056	0.056	0.017
Task anchor (20%)	VCS05	0.539	0.012	0.013
Train. dyn. (5%)	CL07-002	0.039	0.060	0.048
Train. dyn. (5%)	H5-002	0.022	0.035	0.012
Train. dyn. (5%)	H5-030	0.062	0.110	0.170
Train. dyn. (5%)	L2-30	0.087	0.804	0.907
Train. dyn. (5%)	OX01-IOH	0.109	0.081	0.046
Train. dyn. (5%)	VA1-CO2	0.071	0.079	0.045
Train. dyn. (5%)	VA1-STIM	0.051	0.098	0.044
Train. dyn. (5%)	VCS05	0.446	0.018	0.013
Train. dyn. (10%)	CL07-002	0.040	0.030	0.030
Train. dyn. (10%)	H5-002	0.021	0.022	0.011
Train. dyn. (10%)	H5-030	0.063	0.088	0.071
Train. dyn. (10%)	L2-30	0.103	0.473	0.786
Train. dyn. (10%)	OX01-IOH	0.103	0.044	0.038
Train. dyn. (10%)	VA1-CO2	0.081	0.057	0.034
Train. dyn. (10%)	VA1-STIM	0.055	0.079	0.018
Train. dyn. (10%)	VCS05	0.470	0.012	0.013
Train. dyn. (20%)	CL07-002	0.040	0.020	0.029
Train. dyn. (20%)	H5-002	0.022	0.015	0.010
Train. dyn. (20%)	H5-030	0.063	0.076	0.052
Train. dyn. (20%)	L2-30	0.120	0.130	0.362
Train. dyn. (20%)	OX01-IOH	0.102	0.027	0.034
Train. dyn. (20%)	VA1-CO2	0.089	0.032	0.032
Train. dyn. (20%)	VA1-STIM	0.055	0.046	0.016
Train. dyn. (20%)	VCS05	0.524	0.010	0.015

Table 114: Training subjects — VitalDB. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	ms/sample	Fail (%)
Baseline (5%)	Shared telemetry	32.1	0.416	0.000
Baseline (10%)	Shared telemetry	31.9	0.413	0.000
Baseline (20%)	Shared telemetry	32.0	0.414	0.000
Diversity (5%)	Shared telemetry	32.0	0.414	0.000
Diversity (10%)	Shared telemetry	31.9	0.413	0.000
Diversity (20%)	Shared telemetry	32.2	0.417	0.000
Div. cluster (5%)	Shared telemetry	32.0	0.415	0.000
Div. cluster (10%)	Shared telemetry	31.9	0.413	0.000
Div. cluster (20%)	Shared telemetry	32.1	0.415	0.000
Full cohort (100%)	Shared telemetry	32.0	0.414	0.000
Grad. match (5%)	Shared telemetry	31.9	0.414	0.000
Grad. match (10%)	Shared telemetry	31.9	0.413	0.000
Grad. match (20%)	Shared telemetry	31.9	0.413	0.000
NLDR (5%)	Shared telemetry	31.8	0.411	0.000
NLDR (10%)	Shared telemetry	32.0	0.414	0.000
NLDR (20%)	Shared telemetry	32.0	0.415	0.000
Noise filter (5%)	Shared telemetry	31.9	0.414	0.000
Noise filter (10%)	Shared telemetry	32.0	0.414	0.000
Noise filter (20%)	Shared telemetry	32.0	0.414	0.000
Task anchor (5%)	Shared telemetry	31.9	0.414	0.000
Task anchor (10%)	Shared telemetry	32.1	0.415	0.000
Task anchor (20%)	Shared telemetry	32.0	0.415	0.000
Train. dyn. (5%)	Shared telemetry	32.0	0.414	0.000
Train. dyn. (10%)	Shared telemetry	32.1	0.416	0.000
Train. dyn. (20%)	Shared telemetry	32.0	0.415	0.000

Table 115: Training subjects — MC-MED. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Predictive values are development-validation results.

Condition	Task	AUPRC
Grad. match (10%)	Lactate NC	0.343
Diversity (20%)	Pending anemia	0.055
NLDR (50%)	Pending hyperK	0.051
Grad. match (20%)	Lactate NC	0.331
Task anchor (50%)	Pending anemia	0.054
Diversity (50%)	Pending anemia	0.055
Noise filter (10%)	Lactate NC	0.406
Div. cluster (10%)	Pending hyperK	0.044
Task anchor (10%)	Pending anemia	0.057
Div. cluster (20%)	Pending hyperK	0.047
Diversity (50%)	Lactate NC	0.383
Task anchor (20%)	Pending anemia	0.061
Grad. match (20%)	Pending anemia	0.048
Div. cluster (50%)	Pending anemia	0.049
NLDR (20%)	Pending hyperK	0.047
NLDR (10%)	Pending anemia	0.054
Noise filter (10%)	Pending anemia	0.049
Diversity (20%)	Lactate NC	0.316
Div. cluster (50%)	Pending hyperK	0.062
NLDR (20%)	Lactate NC	0.415
Noise filter (50%)	Pending anemia	0.048
Grad. match (10%)	Pending anemia	0.048
NLDR (50%)	Lactate NC	0.330
Div. cluster (10%)	Lactate NC	0.336
Noise filter (20%)	Pending anemia	0.045
Diversity (50%)	Pending hyperK	0.054
Diversity (10%)	Pending hyperK	0.048
Grad. match (50%)	Lactate NC	0.314
Grad. match (50%)	Pending hyperK	0.072
NLDR (20%)	Pending anemia	0.046
Grad. match (50%)	Pending anemia	0.046
Noise filter (20%)	Lactate NC	0.313
Noise filter (50%)	Lactate NC	0.309
Diversity (10%)	Pending anemia	0.045
Div. cluster (50%)	Lactate NC	0.317
Diversity (20%)	Pending hyperK	0.076
Div. cluster (20%)	Pending anemia	0.043
Div. cluster (10%)	Pending anemia	0.045
Train. dyn. (10%)	Pending hyperK	0.042
Train. dyn. (50%)	Lactate NC	0.305

Table 115: Training subjects — MC-MED. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Predictive values are development-validation results. (Continued.)

Condition	Task	AUPRC
Task anchor (50%)	Pending hyperK	0.041
Train. dyn. (20%)	Pending hyperK	0.059
Train. dyn. (50%)	Pending hyperK	0.041
Grad. match (10%)	Pending hyperK	0.039
Diversity (10%)	Lactate NC	0.299
Noise filter (20%)	Decomp. 60m	0.164
Noise filter (50%)	Pending hyperK	0.052
Grad. match (20%)	Pending hyperK	0.039
NLDR (10%)	Lactate NC	0.307
Noise filter (50%)	Decomp. 60m	0.158
Train. dyn. (20%)	Decomp. 60m	0.162
Train. dyn. (20%)	Pending anemia	0.041
Train. dyn. (10%)	Pending anemia	0.042
Train. dyn. (50%)	Decomp. 60m	0.141
Train. dyn. (50%)	Pending anemia	0.041
Diversity (50%)	Decomp. 60m	0.148
Train. dyn. (20%)	Lactate NC	0.292
Div. cluster (20%)	Lactate NC	0.307
Grad. match (20%)	Decomp. 60m	0.144
Noise filter (20%)	Pending hyperK	0.050
NLDR (10%)	Pending hyperK	0.046
Grad. match (10%)	Decomp. 60m	0.140
Task anchor (10%)	Pending hyperK	0.051
Noise filter (10%)	Decomp. 60m	0.142
Div. cluster (50%)	Decomp. 60m	0.149
Train. dyn. (10%)	Lactate NC	0.287
Task anchor (20%)	Pending hyperK	0.044
NLDR (50%)	Pending anemia	0.040
Diversity (20%)	Decomp. 60m	0.135
Div. cluster (10%)	Decomp. 60m	0.149
Task anchor (50%)	Lactate NC	0.305
NLDR (50%)	Decomp. 60m	0.132
Task anchor (10%)	Lactate NC	0.303
Task anchor (20%)	Lactate NC	0.296
Div. cluster (20%)	Decomp. 60m	0.125
NLDR (10%)	Decomp. 60m	0.130
Task anchor (20%)	Decomp. 60m	0.134
Noise filter (10%)	Pending hyperK	0.050
Train. dyn. (10%)	Decomp. 60m	0.126
Task anchor (10%)	Decomp. 60m	0.121

Table 115: Training subjects — MC-MED. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Predictive values are development-validation results. (Continued.)

Condition	Task	AUPRC
NLDR (20%)	Decomp. 60m	0.122
Grad. match (50%)	Decomp. 60m	0.128
Diversity (10%)	Decomp. 60m	0.121
Task anchor (50%)	Decomp. 60m	0.132
Full cohort (100%)	Decomp. 60m	0.145
Full cohort (100%)	Lactate NC	0.309
Full cohort (100%)	Pending hyperK	0.074
Full cohort (100%)	Pending anemia	0.050

Table 116: Training subjects — MC-MED. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Predictive values are development-validation results.

Condition	Task	Time (s)	ms/sample	Fail (%)
Grad. match (10%)	Lactate NC	46.7	86.551	0.000
Diversity (20%)	Pending anemia	92.4	9.005	0.000
NLDR (50%)	Pending hyperK	152.1	3.228	0.000
Grad. match (20%)	Lactate NC	49.2	45.592	0.000
Task anchor (50%)	Pending anemia	101.4	3.953	0.000
Diversity (50%)	Pending anemia	139.1	5.423	0.000
Noise filter (10%)	Lactate NC	94.7	175.454	0.000
Div. cluster (10%)	Pending hyperK	87.1	9.210	0.000
Task anchor (10%)	Pending anemia	86.5	16.859	0.000
Div. cluster (20%)	Pending hyperK	88.7	4.663	0.000
Diversity (50%)	Lactate NC	100.6	37.254	0.000
Task anchor (20%)	Pending anemia	91.0	8.869	0.000
Grad. match (20%)	Pending anemia	84.8	8.262	0.000
Div. cluster (50%)	Pending anemia	141.9	5.518	0.000
NLDR (20%)	Pending hyperK	91.4	4.844	0.000
NLDR (10%)	Pending anemia	83.7	16.322	0.000
Noise filter (10%)	Pending anemia	87.3	17.022	0.000
Diversity (20%)	Lactate NC	47.5	43.961	0.000
Div. cluster (50%)	Pending hyperK	150.6	3.183	0.000
NLDR (20%)	Lactate NC	44.3	41.059	0.000
Noise filter (50%)	Pending anemia	93.6	3.649	0.000
Grad. match (10%)	Pending anemia	82.2	16.020	0.000
NLDR (50%)	Lactate NC	96.6	35.776	0.000
Div. cluster (10%)	Lactate NC	96.1	152.471	0.000
Noise filter (20%)	Pending anemia	135.9	13.242	0.000
Diversity (50%)	Pending hyperK	151.3	3.209	0.000
Diversity (10%)	Pending hyperK	87.7	9.304	0.000
Grad. match (50%)	Lactate NC	95.1	35.234	0.000
Grad. match (50%)	Pending hyperK	149.0	3.163	0.000
NLDR (20%)	Pending anemia	86.1	8.390	0.000
Grad. match (50%)	Pending anemia	143.1	5.578	0.000
Noise filter (20%)	Lactate NC	45.4	42.007	0.000
Noise filter (50%)	Lactate NC	44.8	16.576	0.000
Diversity (10%)	Pending anemia	105.1	20.488	0.000
Div. cluster (50%)	Lactate NC	97.5	35.722	0.000
Diversity (20%)	Pending hyperK	104.2	5.532	0.000
Div. cluster (20%)	Pending anemia	86.4	8.268	0.000
Div. cluster (10%)	Pending anemia	84.2	15.711	0.000
Train. dyn. (10%)	Pending hyperK	84.3	8.950	0.000
Train. dyn. (50%)	Lactate NC	48.0	17.780	0.000

Table 116: Training subjects — MC-MED. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Predictive values are development-validation results. (Continued.)

Condition	Task	Time (s)	ms/sample	Fail (%)
Task anchor (50%)	Pending hyperK	105.1	2.232	0.000
Train. dyn. (20%)	Pending hyperK	90.0	4.775	0.000
Train. dyn. (50%)	Pending hyperK	106.0	2.251	0.000
Grad. match (10%)	Pending hyperK	89.8	9.518	0.000
Diversity (10%)	Lactate NC	48.3	89.535	0.000
Noise filter (20%)	Decomp. 60m	115.8	6.421	0.000
Noise filter (50%)	Pending hyperK	101.8	2.162	0.000
Grad. match (20%)	Pending hyperK	90.5	4.796	0.000
NLDR (10%)	Lactate NC	49.5	91.712	0.000
Noise filter (50%)	Decomp. 60m	109.4	2.426	0.000
Train. dyn. (20%)	Decomp. 60m	87.6	4.858	0.000
Train. dyn. (20%)	Pending anemia	84.3	8.221	0.000
Train. dyn. (10%)	Pending anemia	82.3	16.051	0.000
Train. dyn. (50%)	Decomp. 60m	128.1	2.841	0.000
Train. dyn. (50%)	Pending anemia	97.6	3.806	0.000
Diversity (50%)	Decomp. 60m	196.1	4.351	0.000
Train. dyn. (20%)	Lactate NC	93.7	86.769	0.000
Div. cluster (20%)	Lactate NC	46.5	40.104	0.000
Grad. match (20%)	Decomp. 60m	91.1	5.050	0.000
Noise filter (20%)	Pending hyperK	92.7	4.920	0.000
NLDR (10%)	Pending hyperK	90.8	9.609	0.000
Grad. match (10%)	Decomp. 60m	88.0	9.753	0.000
Task anchor (10%)	Pending hyperK	135.5	14.380	0.000
Noise filter (10%)	Decomp. 60m	89.5	9.926	0.000
Div. cluster (50%)	Decomp. 60m	150.0	3.317	0.000
Train. dyn. (10%)	Lactate NC	47.8	88.609	0.000
Task anchor (20%)	Pending hyperK	89.0	4.722	0.000
NLDR (50%)	Pending anemia	137.6	5.365	0.000
Diversity (20%)	Decomp. 60m	90.4	5.012	0.000
Div. cluster (10%)	Decomp. 60m	91.1	9.812	0.000
Task anchor (50%)	Lactate NC	48.6	17.999	0.000
NLDR (50%)	Decomp. 60m	146.8	3.257	0.000
Task anchor (10%)	Lactate NC	50.5	93.432	0.000
Task anchor (20%)	Lactate NC	49.4	45.734	0.000
Div. cluster (20%)	Decomp. 60m	135.7	7.441	0.000
NLDR (10%)	Decomp. 60m	87.7	9.720	0.000
Task anchor (20%)	Decomp. 60m	91.5	5.077	0.000
Noise filter (10%)	Pending hyperK	86.7	9.203	0.000
Train. dyn. (10%)	Decomp. 60m	95.8	10.617	0.000
Task anchor (10%)	Decomp. 60m	94.6	10.484	0.000

Table 116: Training subjects — MC-MED. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. Predictive values are development-validation results. (Continued.)

Condition	Task	Time (s)	ms/sample	Fail (%)
NLDR (20%)	Decomp. 60m	110.3	6.118	0.000
Grad. match (50%)	Decomp. 60m	200.9	4.456	0.000
Diversity (10%)	Decomp. 60m	87.7	9.710	0.000
Task anchor (50%)	Decomp. 60m	99.8	2.215	0.000

Table 117: Ehr fusion — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Self-attention	Acute det.	0.036	0.139	0.355
Co-attention	Acute det.	0.037	0.085	0.315
Feature attn	Acute det.	0.038	0.117	0.308
Concat	Acute det.	0.037	0.140	0.322
Gated	Acute det.	0.036	0.194	0.306
Bilinear	Acute det.	0.037	0.118	0.269
Tensor	Acute det.	0.038	0.132	0.326
FiLM	Acute det.	0.039	0.127	0.318
Late weighted	Acute det.	0.037	0.182	0.378
Self-attention	Dyn. mort.	0.029	0.130	0.391
Co-attention	Dyn. mort.	0.030	0.108	0.190
Feature attn	Dyn. mort.	0.027	0.234	0.333
Concat	Dyn. mort.	0.030	0.216	0.328
Gated	Dyn. mort.	0.029	0.220	0.341
Bilinear	Dyn. mort.	0.030	0.225	0.309
Tensor	Dyn. mort.	0.029	0.204	0.406
FiLM	Dyn. mort.	0.028	0.228	0.264
Late weighted	Dyn. mort.	0.026	0.207	0.322
Self-attention	Resp. esc.	0.035	0.198	0.378
Co-attention	Resp. esc.	0.037	0.260	0.422
Feature attn	Resp. esc.	0.038	0.222	0.452
Concat	Resp. esc.	0.036	0.159	0.281
Gated	Resp. esc.	0.035	0.147	0.401
Bilinear	Resp. esc.	0.034	0.155	0.440
Tensor	Resp. esc.	0.034	0.174	0.416
FiLM	Resp. esc.	0.035	0.180	0.477
Late weighted	Resp. esc.	0.037	0.187	0.396
Self-attention	Circ. det.	0.051	0.142	0.223
Co-attention	Circ. det.	0.052	0.121	0.246
Feature attn	Circ. det.	0.055	0.241	0.211
Concat	Circ. det.	0.055	0.223	0.226
Gated	Circ. det.	0.050	0.214	0.277
Bilinear	Circ. det.	0.050	0.219	0.326
Tensor	Circ. det.	0.051	0.228	0.242
FiLM	Circ. det.	0.055	0.241	0.208
Late weighted	Circ. det.	0.056	0.256	0.197
Self-attention	Fluid resp.	0.314	0.123	0.464
Co-attention	Fluid resp.	0.315	0.092	0.426
Feature attn	Fluid resp.	0.317	0.082	0.517
Concat	Fluid resp.	0.315	0.088	0.388

Table 117: Ehr fusion — MIMIC-III. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Gated	Fluid resp.	0.313	0.162	0.492
Bilinear	Fluid resp.	0.319	0.158	0.444
Tensor	Fluid resp.	0.325	0.110	0.474
FiLM	Fluid resp.	0.317	0.090	0.521
Late weighted	Fluid resp.	0.317	0.124	0.565
Self-attention	AKI 2/3	0.111	0.133	0.328
Co-attention	AKI 2/3	0.112	0.117	0.264
Feature attn	AKI 2/3	0.112	0.186	0.216
Concat	AKI 2/3	0.116	0.159	0.194
Gated	AKI 2/3	0.116	0.109	0.313
Bilinear	AKI 2/3	0.112	0.133	0.300
Tensor	AKI 2/3	0.111	0.102	0.406
FiLM	AKI 2/3	0.113	0.153	0.337
Late weighted	AKI 2/3	0.109	0.147	0.267
Self-attention	Vent. lib.	0.155	0.304	0.470
Co-attention	Vent. lib.	0.156	0.146	0.338
Feature attn	Vent. lib.	0.161	0.289	0.416
Concat	Vent. lib.	0.157	0.231	0.377
Gated	Vent. lib.	0.156	0.333	0.659
Bilinear	Vent. lib.	0.156	0.180	0.431
Tensor	Vent. lib.	0.152	0.227	0.540
FiLM	Vent. lib.	0.160	0.279	0.474
Late weighted	Vent. lib.	0.157	0.244	0.526
Self-attention	Sepsis-3	0.040	0.163	0.303
Co-attention	Sepsis-3	0.040	0.156	0.281
Feature attn	Sepsis-3	0.040	0.160	0.271
Concat	Sepsis-3	0.038	0.240	0.402
Gated	Sepsis-3	0.039	0.150	0.277
Bilinear	Sepsis-3	0.040	0.172	0.341
Tensor	Sepsis-3	0.040	0.172	0.332
FiLM	Sepsis-3	0.039	0.153	0.270
Late weighted	Sepsis-3	0.040	0.145	0.240
Waveform only	Acute det.	0.040	0.131	0.321
Waveform only	Dyn. mort.	0.021	0.244	0.329
Waveform only	Resp. esc.	0.039	0.235	0.320
Waveform only	Circ. det.	0.055	0.246	0.223
Waveform only	Fluid resp.	0.323	0.112	0.458
Waveform only	AKI 2/3	0.113	0.102	0.426
Waveform only	Vent. lib.	0.156	0.236	0.378
Waveform only	Sepsis-3	0.040	0.158	0.267

Table 118: Ehr fusion — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Self-attention	Acute det.	226.6	2.106	0.194	0.000
Co-attention	Acute det.	275.6	1.889	0.236	0.000
Feature attn	Acute det.	267.2	1.933	0.229	0.000
Concat	Acute det.	296.4	3.642	0.254	0.000
Gated	Acute det.	345.6	3.648	0.296	0.000
Bilinear	Acute det.	322.2	1.933	0.276	0.000
Tensor	Acute det.	271.6	1.933	0.232	0.000
FiLM	Acute det.	349.3	3.642	0.299	0.000
Late weighted	Acute det.	304.9	3.648	0.261	0.000
Self-attention	Dyn. mort.	259.3	2.113	0.199	0.000
Co-attention	Dyn. mort.	302.8	1.885	0.232	0.000
Feature attn	Dyn. mort.	296.1	1.934	0.227	0.000
Concat	Dyn. mort.	333.5	3.648	0.256	0.000
Gated	Dyn. mort.	374.4	3.648	0.287	0.000
Bilinear	Dyn. mort.	353.1	1.934	0.271	0.000
Tensor	Dyn. mort.	305.0	1.933	0.234	0.000
FiLM	Dyn. mort.	340.1	3.648	0.261	0.000
Late weighted	Dyn. mort.	360.7	3.648	0.276	0.000
Self-attention	Resp. esc.	239.0	2.095	0.188	0.000
Co-attention	Resp. esc.	285.9	1.885	0.225	0.000
Feature attn	Resp. esc.	282.9	1.927	0.223	0.000
Concat	Resp. esc.	318.0	3.635	0.250	0.000
Gated	Resp. esc.	319.0	3.636	0.251	0.000
Bilinear	Resp. esc.	332.3	1.929	0.262	0.000
Tensor	Resp. esc.	282.9	1.927	0.223	0.000
FiLM	Resp. esc.	316.6	3.635	0.249	0.000
Late weighted	Resp. esc.	328.0	3.637	0.258	0.000
Self-attention	Circ. det.	370.5	2.383	0.373	0.000
Co-attention	Circ. det.	435.4	2.127	0.438	0.000
Feature attn	Circ. det.	439.1	2.202	0.442	0.000
Concat	Circ. det.	463.9	4.155	0.467	0.000
Gated	Circ. det.	472.4	4.155	0.475	0.000
Bilinear	Circ. det.	516.1	2.202	0.519	0.000
Tensor	Circ. det.	427.0	2.202	0.429	0.000
FiLM	Circ. det.	449.9	4.155	0.453	0.000
Late weighted	Circ. det.	480.8	4.155	0.484	0.000
Self-attention	Fluid resp.	17.5	1.862	0.303	0.000
Co-attention	Fluid resp.	21.0	1.723	0.363	0.000
Feature attn	Fluid resp.	20.9	1.750	0.361	0.000
Concat	Fluid resp.	19.9	1.750	0.344	0.000

Table 118: Ehr fusion — MIMIC-III. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	Mem. (%)	ms/sample	Fail (%)
Gated	Fluid resp.	19.8	1.750	0.342	0.000
Bilinear	Fluid resp.	24.7	1.752	0.427	0.000
Tensor	Fluid resp.	22.0	1.750	0.381	0.000
FiLM	Fluid resp.	20.2	1.750	0.349	0.000
Late weighted	Fluid resp.	20.9	1.750	0.361	0.000
Self-attention	AKI 2/3	456.1	2.378	0.348	0.000
Co-attention	AKI 2/3	561.6	2.127	0.429	0.000
Feature attn	AKI 2/3	546.6	2.177	0.417	0.000
Concat	AKI 2/3	602.1	4.105	0.460	0.000
Gated	AKI 2/3	598.5	4.105	0.457	0.000
Bilinear	AKI 2/3	673.6	2.180	0.514	0.000
Tensor	AKI 2/3	549.8	2.177	0.420	0.000
FiLM	AKI 2/3	571.5	4.105	0.436	0.000
Late weighted	AKI 2/3	374.9	3.664	0.286	0.000
Self-attention	Vent. lib.	22.9	2.183	0.246	0.000
Co-attention	Vent. lib.	16.7	1.766	0.179	0.000
Feature attn	Vent. lib.	29.9	2.082	0.321	0.000
Concat	Vent. lib.	31.6	2.082	0.340	0.000
Gated	Vent. lib.	28.1	2.082	0.302	0.000
Bilinear	Vent. lib.	36.2	2.083	0.389	0.000
Tensor	Vent. lib.	30.1	2.082	0.324	0.000
FiLM	Vent. lib.	29.2	2.082	0.314	0.000
Late weighted	Vent. lib.	32.4	2.082	0.348	0.000
Self-attention	Sepsis-3	362.9	81.566	0.705	0.000
Co-attention	Sepsis-3	343.0	78.853	0.666	0.000
Feature attn	Sepsis-3	236.3	65.758	0.459	0.000
Concat	Sepsis-3	426.2	79.049	0.828	0.000
Gated	Sepsis-3	455.4	81.367	0.885	0.000
Bilinear	Sepsis-3	393.9	78.302	0.765	0.000
Tensor	Sepsis-3	443.7	79.049	0.862	0.000
FiLM	Sepsis-3	417.7	79.049	0.811	0.000
Late weighted	Sepsis-3	211.1	39.152	0.410	0.000
Waveform only	Acute det.	127.9	3.395	0.109	0.000
Waveform only	Dyn. mort.	142.1	3.395	0.109	0.000
Waveform only	Resp. esc.	139.1	3.395	0.110	0.000
Waveform only	Circ. det.	131.2	3.395	0.132	0.000
Waveform only	Fluid resp.	38.7	3.395	0.668	0.000
Waveform only	AKI 2/3	158.7	3.395	0.121	0.000
Waveform only	Vent. lib.	45.5	3.395	0.489	0.000
Waveform only	Sepsis-3	89.1	3.395	0.173	0.000

Table 119: Ehr fusion — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
EHR only	CL07-002	0.042	0.015	0.026
EHR only	H5-002	0.082	0.008	0.006
EHR only	H5-030	0.089	0.035	0.037
EHR only	L2-30	0.200	0.067	0.031
EHR only	OX01-IOH	0.093	0.026	0.036
EHR only	VA1-CO2	0.117	0.027	0.021
EHR only	VA1-STIM	0.199	0.016	0.007
EHR only	VCS05	0.673	0.010	0.006
Co-attention	CL07-002	0.041	0.019	0.026
Co-attention	H5-002	0.040	0.011	0.008
Co-attention	H5-030	0.082	0.050	0.037
Co-attention	L2-30	0.134	0.114	0.049
Co-attention	OX01-IOH	0.154	0.019	0.023
Co-attention	VA1-CO2	0.106	0.029	0.024
Co-attention	VA1-STIM	0.112	0.023	0.010
Co-attention	VCS05	0.672	0.010	0.006
Gated	CL07-002	0.043	0.019	0.025
Gated	H5-002	0.085	0.009	0.005
Gated	H5-030	0.091	0.041	0.037
Gated	L2-30	0.201	0.074	0.033
Gated	OX01-IOH	0.152	0.029	0.022
Gated	VA1-CO2	0.125	0.026	0.020
Gated	VA1-STIM	0.182	0.017	0.008
Gated	VCS05	0.677	0.008	0.005
Late weighted	CL07-002	0.041	0.012	0.025
Late weighted	H5-002	0.086	0.006	0.006
Late weighted	H5-030	0.091	0.031	0.036
Late weighted	L2-30	0.205	0.067	0.032
Late weighted	OX01-IOH	0.151	0.021	0.021
Late weighted	VA1-CO2	0.123	0.020	0.020
Late weighted	VA1-STIM	0.181	0.018	0.008
Late weighted	VCS05	0.670	0.009	0.006
Concat	CL07-002	0.041	0.015	0.025
Concat	H5-002	0.089	0.008	0.005
Concat	H5-030	0.092	0.035	0.036
Concat	L2-30	0.190	0.077	0.034
Concat	OX01-IOH	0.152	0.025	0.022
Concat	VA1-CO2	0.124	0.020	0.020
Concat	VA1-STIM	0.188	0.018	0.007
Concat	VCS05	0.677	0.007	0.005

Table 119: Ehr fusion — VitalDB. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	AUPRC	ECE	AURC
Tensor	CL07-002	0.042	0.014	0.026
Tensor	H5-002	0.086	0.010	0.005
Tensor	H5-030	0.090	0.036	0.037
Tensor	L2-30	0.192	0.079	0.035
Tensor	OX01-IOH	0.150	0.024	0.022
Tensor	VA1-CO2	0.123	0.026	0.021
Tensor	VA1-STIM	0.182	0.023	0.008
Tensor	VCS05	0.675	0.007	0.006
Waveform only	CL07-002	0.037	0.017	0.027
Waveform only	H5-002	0.025	0.011	0.009
Waveform only	H5-030	0.076	0.053	0.039
Waveform only	L2-30	0.097	0.145	0.064
Waveform only	OX01-IOH	0.161	0.019	0.023
Waveform only	VA1-CO2	0.103	0.024	0.026
Waveform only	VA1-STIM	0.057	0.024	0.014
Waveform only	VCS05	0.556	0.016	0.010

Table 120: Ehr fusion — VitalDB. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	ms/sample	Fail (%)
EHR only	CL07-002	2.1	0.027	–
EHR only	H5-002	2.1	0.027	–
EHR only	H5-030	2.1	0.027	–
EHR only	L2-30	2.0	0.026	–
EHR only	OX01-IOH	2.0	0.026	–
EHR only	VA1-CO2	2.1	0.027	–
EHR only	VA1-STIM	2.0	0.027	–
EHR only	VCS05	2.0	0.027	–
Co-attention	CL07-002	11.7	0.152	–
Co-attention	H5-002	11.8	0.153	–
Co-attention	H5-030	11.7	0.153	–
Co-attention	L2-30	11.4	0.148	–
Co-attention	OX01-IOH	11.8	0.153	–
Co-attention	VA1-CO2	11.5	0.150	–
Co-attention	VA1-STIM	11.6	0.150	–
Co-attention	VCS05	12.2	0.158	–
Gated	CL07-002	11.1	0.145	–
Gated	H5-002	11.3	0.148	–
Gated	H5-030	11.3	0.148	–
Gated	L2-30	11.1	0.144	–
Gated	OX01-IOH	11.3	0.147	–
Gated	VA1-CO2	11.2	0.145	–
Gated	VA1-STIM	11.3	0.147	–
Gated	VCS05	11.7	0.152	–
Late weighted	CL07-002	11.0	0.144	–
Late weighted	H5-002	11.3	0.147	–
Late weighted	H5-030	11.1	0.144	–
Late weighted	L2-30	11.0	0.143	–
Late weighted	OX01-IOH	11.1	0.144	–
Late weighted	VA1-CO2	11.2	0.146	–
Late weighted	VA1-STIM	11.0	0.144	–
Late weighted	VCS05	11.5	0.150	–
Concat	CL07-002	10.6	0.138	–
Concat	H5-002	10.8	0.141	–
Concat	H5-030	10.9	0.142	–
Concat	L2-30	10.8	0.140	–
Concat	OX01-IOH	10.9	0.142	–
Concat	VA1-CO2	10.8	0.141	–
Concat	VA1-STIM	10.9	0.142	–
Concat	VCS05	11.4	0.148	–

Table 120: Ehr fusion — VitalDB. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables. (Continued.)

Condition	Task	Time (s)	ms/sample	Fail (%)
Tensor	CL07-002	11.2	0.146	–
Tensor	H5-002	11.6	0.151	–
Tensor	H5-030	11.2	0.146	–
Tensor	L2-30	10.9	0.142	–
Tensor	OX01-IOH	11.1	0.144	–
Tensor	VA1-CO2	11.2	0.145	–
Tensor	VA1-STIM	11.0	0.143	–
Tensor	VCS05	11.4	0.148	–
EHR only	Shared telemetry	–	–	0.000
Co-attention	Shared telemetry	–	–	0.000
Gated	Shared telemetry	–	–	0.000
Late weighted	Shared telemetry	–	–	0.000
Concat	Shared telemetry	–	–	0.000
Tensor	Shared telemetry	–	–	0.000
Waveform only	Shared telemetry	–	–	0.000

Table 121: Ehr fusion — MC-MED. Predictive and confidence values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	AUPRC	ECE	AURC
Co-attention	Decomp. 60m	0.243	0.068	0.178
Concat	Decomp. 60m	0.247	0.055	0.237
EHR only	Decomp. 60m	0.246	0.168	0.277
Gated	Decomp. 60m	0.243	0.054	0.241
Late	Decomp. 60m	0.260	0.062	0.327
Waveform only	Decomp. 60m	0.210	0.061	0.279
Co-attention	Lactate NC	0.376	–	–
Concat	Lactate NC	0.379	–	–
EHR only	Lactate NC	0.376	–	–
Gated	Lactate NC	0.406	–	–
Late	Lactate NC	0.400	–	–
Waveform only	Lactate NC	0.380	–	–
Co-attention	Pending hyperK	0.070	–	–
Concat	Pending hyperK	0.067	–	–
EHR only	Pending hyperK	0.082	–	–
Gated	Pending hyperK	0.065	–	–
Late	Pending hyperK	0.068	–	–
Waveform only	Pending hyperK	0.052	–	–
Co-attention	Pending anemia	0.082	–	–
Concat	Pending anemia	0.106	–	–
EHR only	Pending anemia	0.131	–	–
Gated	Pending anemia	0.103	–	–
Late	Pending anemia	0.114	–	–
Waveform only	Pending anemia	0.074	–	–

Table 122: Ehr fusion — MC-MED. Training-cost and failure values for the corresponding EES figure; entries are the plotted means. This figure-matched export is separate from the updated summary tables.

Condition	Task	Time (s)	Fail (%)
Co-attention	Decomp. 60m	5019.9	0.000
Concat	Decomp. 60m	641.6	0.000
EHR only	Decomp. 60m	250.0	0.000
Gated	Decomp. 60m	978.8	0.000
Late	Decomp. 60m	795.3	0.000
Waveform only	Decomp. 60m	125.6	0.000
Co-attention	Lactate NC	177.5	0.000
Concat	Lactate NC	166.6	0.000
EHR only	Lactate NC	82.3	0.000
Gated	Lactate NC	211.0	0.000
Late	Lactate NC	159.2	0.000
Waveform only	Lactate NC	30.7	0.000
Co-attention	Pending hyperK	413.9	0.000
Concat	Pending hyperK	300.3	0.000
EHR only	Pending hyperK	179.5	0.000
Gated	Pending hyperK	535.1	0.000
Late	Pending hyperK	403.2	0.000
Waveform only	Pending hyperK	168.9	0.000
Co-attention	Pending anemia	398.1	0.000
Concat	Pending anemia	348.3	0.000
EHR only	Pending anemia	162.8	0.000
Gated	Pending anemia	473.3	0.000
Late	Pending anemia	387.0	0.000
Waveform only	Pending anemia	140.6	0.000

Table 123: Numerical comparisons for Figure 48. Panel letters follow the figure; validation and test estimators retain its original definitions.

Setting	Condition	Validation AUROC	Test AUROC
(a) Forecast validation: encoder–tokenizer grid			
fits	seft	0.606 ± 0.003	–
fits	Conv1D	0.635 ± 0.001	–
resnet1d	Conv1D	0.623 ± 0.002	–
mamba3	crossformer	0.585 ± 0.004	–
dlinear	Conv1D	0.634 ± 0.001	–
dlinear	seft	0.608 ± 0.003	–
bilstm	seft	0.600 ± 0.003	–
transformer	crossformer	0.573 ± 0.002	–
fits	crossformer	0.582 ± 0.004	–
bilstm	crossformer	0.586 ± 0.003	–
s4	Conv1D	0.632 ± 0.002	–
tcn	crossformer	0.572 ± 0.003	–
resnet1d	seft	0.617 ± 0.002	–
mamba3	Conv1D	0.634 ± 0.002	–
bilstm	Conv1D	0.632 ± 0.002	–
inceptiontime	crossformer	0.574 ± 0.004	–
inceptiontime	Conv1D	0.613 ± 0.002	–
tcn	Conv1D	0.618 ± 0.002	–
s4	seft	0.609 ± 0.003	–
s4	crossformer	0.621 ± 0.003	–
mamba3	seft	0.606 ± 0.003	–
dlinear	crossformer	0.590 ± 0.002	–
tcn	seft	0.609 ± 0.000	–
inceptiontime	seft	0.606 ± 0.002	–
resnet1d	crossformer	0.586 ± 0.002	–
transformer	Conv1D	0.635 ± 0.003	–
transformer	seft	0.603 ± 0.001	–
(b) Transformer/Conv1D input history			
5 min	History	0.623 ± 0.003	0.628 ± 0.003
15 min	History	0.629 ± 0.002	0.625 ± 0.003
30 min	History	0.633 ± 0.002	0.635 ± 0.003
60 min	History	0.622 ± 0.001	0.640 ± 0.002
(c) Subject budgets: fold-mean validation			
5%	Concurrent BIS	0.780	–
5%	Forecast	0.530	–
10%	Concurrent BIS	0.814	–
10%	Forecast	0.533	–
20%	Concurrent BIS	0.839	–

Table 123: Numerical comparisons for Figure 48. Panel letters follow the figure; validation and test estimators retain its original definitions. (Continued.)

Setting	Condition	Validation AUROC	Test AUROC
20%	Forecast	0.544	–
100%	Concurrent BIS	0.866	–
100%	Forecast	0.607	–
(d) Channel budgets: three-pair validation means			
All channels	0 channels	0.600	–
Random	2 channels	0.541	–
Random	4 channels	0.563	–
Random	8 channels	0.585	–
Temporal MI	2 channels	0.564	–
Temporal MI	4 channels	0.580	–
Temporal MI	8 channels	0.595	–
(e) Natural missingness: three-pair means			
architecture only	Natural	0.601	0.603
drifting model	Natural	0.578	0.581
grin	Natural	0.578	0.583
grud	Natural	0.577	0.581
lead agnostic ssl	Natural	0.575	0.579
positional kriging	Natural	0.572	0.575
raindrop	Natural	0.596	0.603
rectified flow	Natural	0.578	0.582
trigflow	Natural	0.580	0.585
zero mask delta	Natural	0.580	0.584
(f) Native masks: probability-ensemble test stress			
Natural	Ensemble	–	0.619
25% sensor dropout	Ensemble	–	0.605
50% sensor dropout	Ensemble	–	0.583
Modality dropout	Ensemble	–	0.603

Table 124: Original validation gap study for Figure 49: macro and task means, seed SD where supplied, and subject support. These are the earlier fixed-recipe results, separate from the updated main-figure gap sweep.

Gap (min)	Task	AUROC	AUPRC	Subjects	Positive
5	Macro	0.643	0.078	–	–
15	Macro	0.622	0.069	–	–
30	Macro	0.613	0.067	–	–
45	Macro	0.591	0.059	–	–
60	Macro	0.575	0.055	–	–
75	Macro	0.561	0.052	–	–
5	OX01-IOH	0.719 ± 0.006	0.111 ± 0.007	2443	577
15	OX01-IOH	0.680 ± 0.008	0.087 ± 0.002	2261	478
30	OX01-IOH	0.662 ± 0.012	0.086 ± 0.002	1962	390
45	OX01-IOH	0.603 ± 0.009	0.060 ± 0.002	1591	270
60	OX01-IOH	0.562 ± 0.013	0.053 ± 0.003	1269	213
75	OX01-IOH	0.544 ± 0.008	0.049 ± 0.003	986	163
5	CL07-002	0.590 ± 0.007	0.044 ± 0.001	2447	461
15	CL07-002	0.560 ± 0.008	0.042 ± 0.003	2237	368
30	CL07-002	0.512 ± 0.010	0.032 ± 0.003	1856	264
45	CL07-002	0.520 ± 0.017	0.031 ± 0.002	1478	190
60	CL07-002	0.518 ± 0.011	0.029 ± 0.001	1139	139
75	CL07-002	0.491 ± 0.025	0.028 ± 0.004	906	100
5	H5-002	0.681 ± 0.005	0.028 ± 0.002	4175	435
15	H5-002	0.653 ± 0.005	0.024 ± 0.001	3884	407
30	H5-002	0.659 ± 0.009	0.026 ± 0.001	3379	322
45	H5-002	0.635 ± 0.009	0.024 ± 0.001	2955	261
60	H5-002	0.607 ± 0.014	0.021 ± 0.001	2585	209
75	H5-002	0.600 ± 0.017	0.023 ± 0.003	2251	175
5	H5-030	0.538 ± 0.008	0.063 ± 0.002	3437	1098
15	H5-030	0.556 ± 0.005	0.063 ± 0.002	3124	963
30	H5-030	0.582 ± 0.006	0.065 ± 0.001	2638	759
45	H5-030	0.558 ± 0.011	0.057 ± 0.002	2201	600
60	H5-030	0.567 ± 0.006	0.060 ± 0.002	1835	472
75	H5-030	0.573 ± 0.007	0.058 ± 0.002	1513	359
5	VA1-CO2-DROP	0.667 ± 0.004	0.107 ± 0.002	3580	846
15	VA1-CO2-DROP	0.655 ± 0.005	0.091 ± 0.001	3101	659
30	VA1-CO2-DROP	0.642 ± 0.011	0.090 ± 0.004	2331	468
45	VA1-CO2-DROP	0.621 ± 0.013	0.079 ± 0.008	1750	324
60	VA1-CO2-DROP	0.600 ± 0.006	0.070 ± 0.002	1313	226
75	VA1-CO2-DROP	0.549 ± 0.012	0.053 ± 0.003	984	167
5	L2-30	0.582 ± 0.017	0.113 ± 0.008	1297	368
15	L2-30	0.594 ± 0.006	0.120 ± 0.005	1361	382
30	L2-30	0.579 ± 0.014	0.110 ± 0.007	1359	375
45	L2-30	0.590 ± 0.006	0.113 ± 0.003	1271	333

Table 124: Original validation gap study for Figure 49: macro and task means, seed SD where supplied, and subject support. These are the earlier fixed-recipe results, separate from the updated main-figure gap sweep. (Continued.)

Gap (min)	Task	AUROC	AUPRC	Subjects	Positive
60	L2-30	0.567 ± 0.012	0.106 ± 0.008	1122	283
75	L2-30	0.561 ± 0.015	0.102 ± 0.003	953	234
5	VA1-STIM-RESPONSE	0.723 ± 0.013	0.082 ± 0.012	2238	339
15	VA1-STIM-RESPONSE	0.659 ± 0.012	0.054 ± 0.002	2003	286
30	VA1-STIM-RESPONSE	0.658 ± 0.010	0.056 ± 0.007	1623	203
45	VA1-STIM-RESPONSE	0.610 ± 0.025	0.048 ± 0.005	1252	166
60	VA1-STIM-RESPONSE	0.605 ± 0.025	0.050 ± 0.009	943	122
75	VA1-STIM-RESPONSE	0.611 ± 0.003	0.051 ± 0.003	710	94

Table 125: Original natural-input validation means for Figure 50(a,b). The waveform branch uses the 60-minute Mamba-3/Conv1D recipe.

Method	AUROC	AUPRC	ECE
waveform only	0.615	0.069	0.046
ehr only	0.648	0.091	0.034
simple concat	0.665	0.095	0.035
gated	0.661	0.096	0.034
late weighted logits	0.672	0.097	0.034
ehr query coattention	0.642	0.082	0.039
tensor fusion	0.659	0.095	0.033

Table 126: Fixed-checkpoint control contrasts for Figure 50(c). Differences are control minus natural-input validation scores on the 0–1 metric scale.

Method	Control	ΔAUROC	ΔAUPRC
ehr query coattention	ehr branch absent	-0.047	-0.020
ehr query coattention	ehr values permuted within split	-0.049	-0.021
ehr query coattention	ehr values zeroed keep masks	-0.046	-0.021
ehr query coattention	waveform branch absent	-0.144	-0.039
ehr query coattention	waveform values zeroed keep masks	-0.037	-0.013
gated	ehr branch absent	-0.080	-0.038
gated	ehr values permuted within split	-0.099	-0.042
gated	ehr values zeroed keep masks	-0.083	-0.038
gated	waveform branch absent	-0.026	-0.011
gated	waveform values zeroed keep masks	-0.016	-0.004
late weighted logits	ehr branch absent	-0.076	-0.037
late weighted logits	ehr values permuted within split	-0.096	-0.042
late weighted logits	ehr values zeroed keep masks	-0.084	-0.039
late weighted logits	waveform branch absent	-0.031	-0.009
late weighted logits	waveform values zeroed keep masks	-0.022	-0.005
simple concat	ehr branch absent	-0.078	-0.037
simple concat	ehr values permuted within split	-0.099	-0.041
simple concat	ehr values zeroed keep masks	-0.087	-0.038
simple concat	waveform branch absent	-0.028	-0.009
simple concat	waveform values zeroed keep masks	-0.016	-0.003
tensor fusion	ehr branch absent	-0.079	-0.035
tensor fusion	ehr values permuted within split	-0.091	-0.039
tensor fusion	ehr values zeroed keep masks	-0.074	-0.034
tensor fusion	waveform branch absent	-0.034	-0.009
tensor fusion	waveform values zeroed keep masks	-0.015	-0.005