Toward data-centric experimentation in prehospital AI: a customizable pipeline

Tamara Krafft, Bernhard Bauer

Angaben zur Veröffentlichung / Publication details:

Krafft, Tamara, and Bernhard Bauer. 2026. “Toward data-centric experimentation in prehospital AI: a customizable pipeline.” In *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies, March 2-4, 2026, Marbella, Spain, volume 3*, edited by Janina Bahnemann, Sebastian Fudickar, Francesco Mercaldo, Jordi Solé-Casals, Hui Liu, Eleni Kaldoudi, and Hugo Gamboa, 351–63. Setúbal: SciTePress. https://doi.org/10.5220/0014619300004070.

Toward Data-Centric Experimentation in Prehospital AI: A Customizable Pipeline

Tamara Krafft[a] and Bernhard Bauer[b]
Software Methodologies for Distributed Systems, University of Augsburg, Universitätsstraße 6a, Augsburg, Germany

Keywords: Prehospital Emergency Care, Machine Learning, Artificial Intelligence, Data-Centric AI, Missing Data, Informative Missingness, Early Warning Systems, Sepsis Prediction.

Abstract: Early warning in prehospital emergency care is critical for improving patient outcomes. However, applying machine learning in this context remains challenging due to the nature of prehospital data, which are often incomplete, irregularly sampled, and affected by plausibility issues. This paper introduces a data-centric experimentation pipeline that treats these challenges and their mitigation strategies as explicit experimental variables rather than hidden preprocessing choices. The pipeline structures time-series experimentation into traceable stages that enable systematic assessment of how data-quality decisions influence downstream model behavior. We illustrate the feasibility of our pipeline through a representative case study on sepsis prediction involving 96,551 encounters (4.01% prevalence). The evaluation systematically varies two experimental dimensions, imputation and usage of missingness-derived features, evaluated under differing levels of missingness. Selected results show that missingness is the dominant source of performance degradation, and that imputation and missingness-derived features provide complementary benefits. These findings demonstrate how the proposed pipeline enables transparent, reproducible exploration of data-quality trade-offs, laying the groundwork for broader data-centric experimentation in prehospital machine learning.

1 INTRODUCTION

Early initiation of interventions in prehospital emergency care is associated with improved patient outcomes and reduced mortality rates (Bleyer et al., 2011). For time-sensitive, nonspecific conditions such as sepsis, early identification of patient deterioration is crucial for timely intervention and escalation of care (Evans et al., 2021).

Emergency medical services (EMS) are experiencing ongoing digitalization. For example, the number of EMS activations recorded in the National EMS Information System (NEMSIS), which collects de-identified, standardized electronic care reports from U.S. EMS agencies, increased by approximately 67% from 2019 (36,119,969 activations) to 2024 (60,298,684 activations) (NEMSIS, 2019; NEMSIS, 2024). These developments create new opportunities to apply machine learning (ML) for outcome prediction, early risk detection, and decision support in prehospital settings.

However, prehospital data are characterized by unique challenges that differentiate them from in-hospital datasets, including high levels of missingness, plausibility violations, and irregular sampling intervals arising from heterogeneous devices and dynamic emergency contexts. For instance, in the NEMSIS 2024 dataset, 61.97% of encounters with captured vital sign data have at least one timestamp where heart rate, respiratory rate, systolic blood pressure, or oxygen saturation (SpO_2) is missing. Among encounters with repeated recordings, the median standard deviation of sampling intervals is approximately 3.0 minutes, with an interquartile range of 1.6 to 5.6 minutes, indicating substantial variability.[1]

Traditional ML pipelines primarily focus on enhancing model performance by optimizing learning parameters and treat data as a static resource that fuels the training process (Jarrahi et al., 2023). As a consequence, they tend to produce brittle systems when confronted with noisy and incomplete real-world data (Jarrahi et al., 2023). Moreover, data quality issues

[a] https://orcid.org/0009-0002-0505-5724
[b] https://orcid.org/0000-0002-7931-1105

[1] The reported numbers were computed by the authors using the NEMSIS 2024 dataset; analysis scripts are provided in the supplementary material, see Appendix A.

Krafft, T. and Bauer, B.
Toward Data-Centric Experimentation in Prehospital AI: A Customizable Pipeline.
DOI: 10.5220/0014619300004070

In *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2026) - Volume 3: HEALTHINF*, pages 351-363
ISBN: 978-989-758-802-0; ISSN: 2184-4305

are often handled through opaque preprocessing and cleaning steps and rarely examined as experimental factors in their own right (Bazo-Alvarez et al., 2021). In contrast, prior work on informative missingness (Lipton et al., 2016; Che et al., 2018; Mercaldo and Blume, 2020) suggests that missing data can encode clinically relevant signals, motivating a more data-centric perspective, in which data imperfection (DI) and mitigation strategies are treated as explicit design choices and experimental factors.

Despite growing interest in ML for EMS, applications remain relatively scarce and fragmented compared to in-hospital early warning systems. While studies such as (Ward et al., 2025) demonstrate the feasibility of time-series prediction, standardized early warning systems, such as a prehospital Early Warning Score (EWS) or an equivalent to the in-hospital, FDA-cleared eCART (Churpek et al., 2024), have not been established. Advancing this field, therefore, requires systematic experimentation under realistic prehospital data conditions.

Taking these challenges into consideration, this paper addresses two research questions:

1. How can data quality challenges inherent in prehospital time-series data be effectively *tackled* and potentially *leveraged*?
2. How can we design a *systematic*, *reproducible*, and *data-centric* experimentation approach to support robust ML research in this domain?

To contribute to these questions, we (i) illustrate the current landscape of prehospital ML and the methodological possibilities inherited from in-hospital research in Section 2, (ii) provide an overview of DI profiling approaches in multivariate time-series data in Section 3, (iii) propose a customizable data-centric pipeline (DCP) for ML experimentation in EMS that treats DI as a primary experimental dimension in Section 4, and (iv) present a case study that instantiates the DCP for early sepsis prediction in Section 5, analyzing the effects of two experimental knobs: imputation strategy and DI-derived features.

2 RELATED WORK

Despite growing interest and progress in digitalization, machine learning in prehospital care remains underexplored and substantially less mature than in-hospital early warning systems. A tabular summary of representative prehospital studies is provided in Table 2 in Appendix B.

A majority of these EMS studies rely on single-timestamp predictors and employ conventional ML models, such as logistic regression, tree-based methods, and support vector machines. Deep neural networks are rare and primarily used in triage and trauma tools. Outcomes commonly include short-term mortality, ICU admission or escalation of care, diagnosis of specific time-critical conditions, or the need for lifesaving interventions. Data sources predominantly comprise vital signs, demographics, and disease-specific features such as mechanism of injury, symptom descriptors, or ECG findings. DI handling is heterogeneous but generally limited to complete-case exclusion, mean/median imputation, or multiple imputation, with a few studies relying on model-native missingness handling. Only isolated works explicitly encode informative missingness.

A small subset explicitly models vital sign time-series, but demonstrates clear potential for EMS applications. (Ward et al., 2025) use minutes-scale vital sign trajectories and gradient-boosted models with missingness indicators to predict 7-day mortality. (Liu et al., 2014) integrate continuous vital sign dynamics into hybrid rule/ML systems to anticipate lifesaving interventions in trauma patients. Findings by (Krafft et al., 2025), who employ logistic regression with trend-based EWS features, suggest that short-term EWS trend features carry additional prognostic signal in early EMS care. (Weidman et al., 2025) leverage continuous waveform segments during air transport and train histogram-gradient-boosting models to predict imminent interventions.

In-hospital vital sign time-series modeling for early warning provides a rich methodological landscape that EMS research can leverage. Two broad paradigms recur: feature-engineered learning over aggregated time-series (Rangan et al., 2022; Li et al., 2022; Choi et al., 2023; Akel et al., 2021), and sequence models, that directly learn temporal dependencies (Shamout et al., 2019; Kwon et al., 2018; Choi et al., 2022a; Cheng et al., 2022; Chae et al., 2022; Silva et al., 2021; Lee et al., 2024; Senthil Pandi et al., 2024; Sim et al., 2025; Jehangir and Li, 2025; Kim et al., 2025; Su et al., 2022). Furthermore, several recurring design patterns from hospital-based studies are directly relevant for EMS exploration. Minimal-variable models achieve strong performance with small sets of variables (Akel et al., 2021; Kwon et al., 2018; Rangan et al., 2022), aligning well with sparse prehospital data. Volatility and quantiles may carry meaningful predictive signals (Alghatani et al., 2021). Lag–lead window tuning aligns the observation window and prediction horizon with operational needs (Choi et al., 2023; Rangan et al., 2022). Forecast-then-score strategies decouple physiologic modeling from event prediction (Silva et al., 2021;

Amer et al., 2020; Jehangir and Li, 2025). Threshold-reach prediction (Senthil Pandi et al., 2024; Thiele et al., 2025) maps directly to EMS decision triggers.

The literature on prehospital ML demonstrates feasibility across diverse outcomes but remains dominated by single-timestamp features and simplistic handling of DI. In contrast, in-hospital research illustrates methodological possibilities regarding temporal modeling and outcome formulation. These observations directly motivate the customizable data-centric pipeline introduced in Section 4, which is designed to bridge the gap between the methodological richness of in-hospital time-series modeling and the practical realities of prehospital EMS data.

3 DATA IMPERFECTION

This section examines intrinsic data imperfection in prehospital numerical time-series data and outlines associated analytical and operational considerations.

3.1 Types of Data Imperfection

We categorize numerical DI observed in multivariate vital sign time-series data $X = \{x_t \in \mathbb{R}^d \mid t \in \mathbb{N}, d \in \mathbb{N}\}$, where t indexes time and d corresponds to the set of measured vital signs, into three types: *missingness*, *plausibility violations*, and *irregular sampling*.

Firstly, missing values occur when a measurement is unavailable at a given timestamp and can be represented by a binary indicator, where 1 indicates an observed value and 0 a missing value (Che et al., 2018). This representation enables the construction of missingness masks, $M = \{m_t \in \{0,1\}^d \mid t \in \mathbb{N}, d \in \mathbb{N}\}$, that facilitate downstream analysis. Secondly, we define plausibility violations as recorded values that are implausible according to one or more criteria. Univariate violations encompass logically or physiologically out-of-range values, such as an SpO_2 exceeding 100% or a heart rate above 300 beats per minute. Temporal violations include sudden jumps or non-physiological rate-of-change. Multivariate inconsistencies arise when relationships among variables are violated, for example, physiologically incompatible combinations of heart rate and blood pressure. Plausibility violations can likewise be encoded as masks. Lastly, irregularity affects the timestamp structure itself (Zhang et al., 2023). Intra-series intervals between successive measurements may vary, depending on clinician judgement or contextual constraints. Although irregularity can induce missingness, the two represent distinct imperfection types and should be treated separately to preserve temporal dynamics.

3.2 Mechanisms and Predictive Value

Missing data mechanisms are commonly described in terms of the statistical relationship between the probability of missingness and the underlying data, $p(M \mid X)$. Let X_{obs} and X_{mis} be the observed and unobserved subsets of X. (Rubin, 1976) defines three categories: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). MCAR occurs when missingness is independent of both observed and unobserved variables, $p(M \mid X) = p(M)$, such as in the case of random device malfunctions. MAR assumes missingness depends on observed variables, $p(M \mid X) = p(M \mid X_{\text{obs}})$. For example, certain vital signs may be omitted while clinicians attend to other documented priorities or interventions. If neither MCAR nor MAR holds, the mechanism is MNAR. MNAR describes situations in which the probability of missing data depends on unobserved data, including the missing value itself, i.e., $p(M \mid X_{\text{obs}}, X_{\text{mis}}) \neq p(M \mid X_{\text{obs}})$, such as failed blood pressure readings in severe hypotension.

Clinical data are rarely MCAR. Qian et al. (Qian et al., 2024) identify protocol-driven, resource-related, condition-dependent, and value-dependent patterns that map onto MAR and MNAR and reflect clinical workflow influences. Protocol-driven patterns arise from scheduled measurements, for example, EMS may be instructed to read blood pressure every n minutes, introducing structured gaps. Resource-related patterns reflect measurement omissions due to situational constraints like transport. Condition-dependent patterns emerge when measurement frequency or quality is influenced by the patient's physiological state. Value-dependent imperfection occurs when previously observed abnormal readings trigger additional assessments or repeated measurements. These patterns highlight that missingness in clinical data is often systematically linked to observable or unobservable clinical states rather than occurring unpredictably, and motivate a growing body of research demonstrating that missingness can carry predictive signal.

For example, Che et al. introduce the GRU-D model, which incorporates features derived from missingness, and outperforms baseline Gated Recurrent Unit (GRU) models in mortality prediction tasks (Che et al., 2018). Lipton et al. demonstrate that last observation carried forward (LOCF) imputation degraded Long Short-Term Memory (LSTM) performance in multi-class diagnosis prediction relative to zero imputation, suggesting that missingness masks

hold informative signal (Lipton et al., 2016). Mercaldo et al. train separate submodels for different missingness patterns, achieving superior results compared to standard imputation strategies (Mercaldo and Blume, 2020). Pérez-Lebel et al. reported significant performance gains by appending missingness indicators to feature sets, showing that features with high missing rates can still carry substantial predictive value (Perez-Lebel et al., 2022).

These findings illustrate that missingness may encode physiological, contextual, or workflow-related signals. Importantly, this perspective can be extended beyond missingness to implausibility and irregularity, as their occurrence is likewise entangled with clinical decision making and operational constraints. Plausibility violations can arise from device behavior, measurement conditions, or environmental factors. Irregular sampling similarly reflects the prehospital assessment, where measurement intervals vary according to on-scene events rather than fixed schedules. Because such patterns originate from operational and clinical processes, understanding and characterizing them is essential for assessing their potential influence on downstream modeling.

3.3 Characterizing Data Imperfection

As described, missingness and plausibility violations can be represented as masks, enabling their systematic characterization. A structured subspace of the analytical design space for examining DI in multivariate time-series data is illustrated in Figure 1. We organize characterization methods along two principal axes: (i) *intra-variable* analyses, which operate on a single variable, versus *inter-variable* analyses, which capture cross-variable relationships, and (ii) *"symmetric"* analyses, which relate masks to masks, versus *"asymmetric"* analyses, which relate masks to observed values. Lagged and windowed variants further extend these analyses along the temporal dimension.

Intra-variable methods diagnose the structure of DI within an individual variable. Column-level statistics summarize basic properties, such as the percentage of missing or implausible values at the case or global level, which can reveal subgroup differences and general workflow trends. Gap-based analyses quantify the lengths of imperfect segments and their relationships to the subsequent observation. Markov chain summaries and symmetric lagged autocorrelation characterize how the current mask state of a variable at time i depends on its preceding or subsequent mask states at times $i \pm n$. Asymmetric univariate approaches, such as windowed significance analysis, relate missing or implausible values at time i to observations within a predefined temporal window $i \pm n$, enabling detection of localized associations between imperfection and measured values.

Inter-variable analyses examine how DI co-occurs across variables and provide insight into patterned dependencies. Row-based statistics summarize properties such as the row-level percentage of missing or implausible values, highlighting systematic multivariable DI patterns. These patterns can be tested for MCAR using Little's MCAR test (Little, 1988). Symmetric inter-variable analyses correlate masks across variables to reveal synchronized gaps or co-occurring implausible values. Asymmetric analyses relate the DI pattern of one variable to the observed values of another, indicating whether DI preferentially arises under specific physiological states. Lagged variants extend both perspectives to capture temporal dependencies, for instance, whether DI in one variable tends to precede that in another. Although MNAR cannot be tested directly, feature-wise MAR–MNAR likelihood-ratio tests, as proposed by Alasal et al. (Alasal et al., 2025), can provide evidence for the presence of MNAR.

Irregular sampling affects the temporal grid itself and therefore requires adapted analyses. Methods analogous to gap-based univariate characterization can be applied by examining the distribution of time intervals between sampling timestamps instead of between observed values. Case-level summaries capture measurement regularity within individual records, while global distributions reveal system-level or workflow-level timing patterns. Additional temporal characterizations, such as Markov chain summaries of interval transitions, can further expose structured measurement dynamics.

3.4 Mitigating Data Imperfection

Several mitigation strategies exist to address imperfections in time-series data and are often necessary to make the data suitable for downstream ML tasks. Plausibility violations typically require an initial outlier detection step, after which values can be removed or corrected using domain-specific rule-based approaches (Ortiz et al., 2024). Imputation methods address missingness and range from simple statistical techniques to advanced deep learning methods (Emmanuel et al., 2021; Wang et al., 2025). Recent sequence models, such as the Self-Attention-based Imputation for Time Series (SAITS) (Du et al., 2023), natively perform imputation and leverage missingness patterns as part of their architecture. Irregular sampling can be mitigated through temporal resampling or more sophisticated modeling techniques that op-

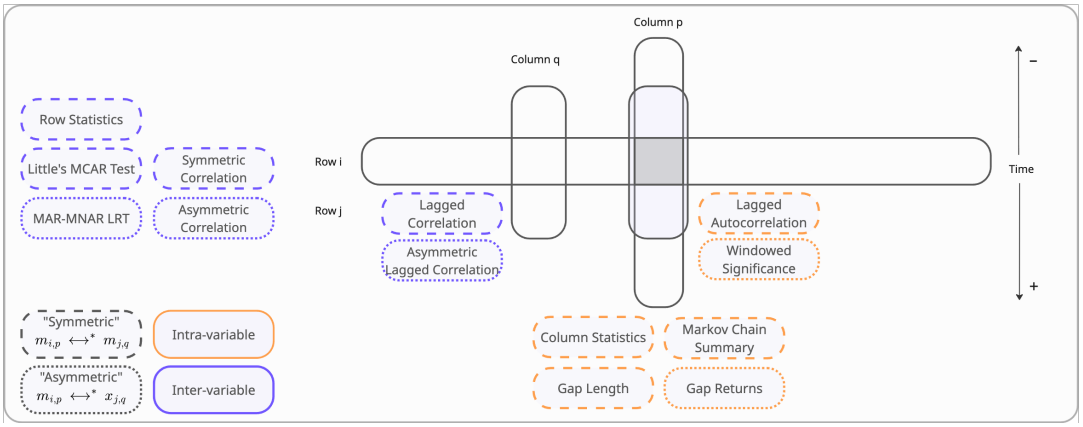

Figure 1: Overview of analytical dimensions for characterizing data imperfection in multivariate time-series data. Columns represent physiological variables, and rows represent time steps. The notation $\longleftrightarrow^{*}$ denotes a relation defined over all index pairs (i,p) and (j,q), including cases $i=j$ and $p=q$. $x_{i,p}$ denotes the observed value of variable p at time step i, and $m_{i,p}$ represents the corresponding binary missingness indicator.

erate on irregular grids (Zhang et al., 2023). Each mitigation strategy entails a trade-off between cost and bias. Simpler methods are often computationally inexpensive but rely on stronger assumptions, while more complex techniques reduce bias at the expense of model complexity and computational overhead. For instance, simple imputation methods such as mean or LOCF implicitly assume MCAR behavior and are inappropriate when missingness is informative, while deep-learning–based imputers can accommodate MAR or MNAR, but incur higher computational cost and require more data (Emmanuel et al., 2021; Wang et al., 2025).

Together, the taxonomy of imperfection, mechanisms, characterization, and mitigation provides the foundation for a data-centric perspective. In Section 4, we build on these insights, treating DI types and their mitigations as explicit experimental dimensions.

4 CUSTOMIZABLE PIPELINE

We introduce a customizable, data-centric pipeline (DCP) for prehospital multivariate time-series data that structures the end-to-end workflow into eight stages. The overall organization of the stages is informed by established process models such as CRISP-ML(Q) and KDD, which outline full-lifecycle workflows from problem formulation to model assessment (Studer et al., 2021; Fayyad et al., 1996). In contrast, our DCP is deliberately scoped to the experimental pipeline and does not address deployment or operational integration. Aligned with the data-centric artificial intelligence (AI) perspective of Jarrahi et al. (Jarrahi et al., 2023), our DCP prioritizes systematic data profiling, preparation, and logging over model-centric optimization. We distinguish between structural design choices (e.g., representation and model class) and experimental knobs (e.g., imputation, filtering, and DI-derived features). Our contribution lies in operationalizing data-centric principles specifically for the complexities of time-series data through this explicit set of knobs that elevate DI-related decisions to first-class experimental variables, making them configurable, transparent, and empirically comparable.

4.1 Stage Overview

The DCP and its individual stages are outlined below and illustrated in Figure 2.

Stage 0 (Data Acquisition and Ingestion) concerns the collection and standardization of raw EMS data. Decisions include selecting data sources (e.g., field monitors, electronic patient care reports, and hospital outcomes) and data formats. Harmonized ingestion with data versioning (snapshots or incremental updates) ensures reproducibility and traceability, and enables longitudinal assessment of model robustness.

Stage 1 (Prediction Task Formulation) defines the clinical prediction problem, which can be specified based on available data or, in prospective studies, prior to data acquisition. Key decisions include the target, horizon (e.g., 15-minute, upon hospital arrival), and granularity (e.g., binary, multi-class, regression). Targets may be framed at different levels: physiological (e.g., vital-sign trajectory prediction), clinical (e.g., sepsis diagnosis), or actionable (e.g., reaching a critical threshold). Stage 1 defines the prediction function to be learned, $f: X \rightarrow Y$, which an-

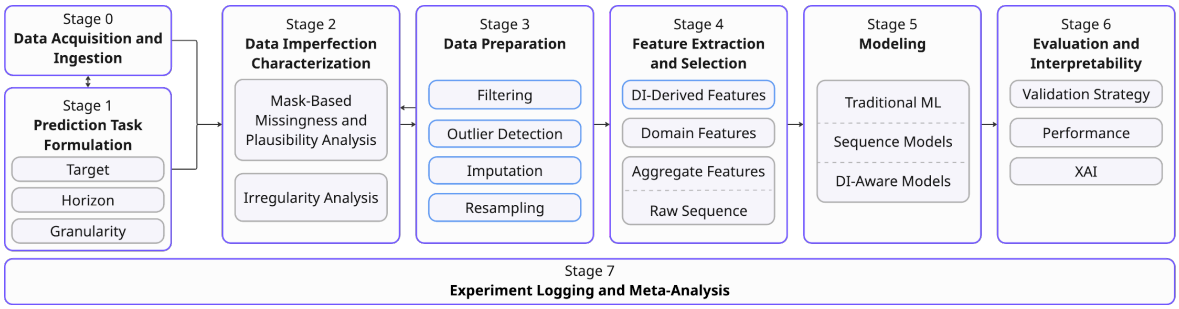

Figure 2: Customizable data-centric pipeline for time-series modeling. Blue elements denote experimental knobs for DI-centric experimentation; grey elements show structural and methodological design choices; dashed lines indicate alternatives.

chors all subsequent structural and experimental design choices and enables stratified DI analyses.

Stage 2 (Data Imperfection Characterization) applies a predefined characterization protocol, as outlined in Section 3.3, to quantify patterns of missingness, plausibility violations, and irregularity. Resulting task-specific DI profiles inform downstream structural design choices and experimental knobs. Stage 2 may be repeated after Stage 3 to evaluate how data preparation decisions affect DI characteristics.

Stage 3 (Data Preparation) applies preprocessing steps to render the data suitable for modeling while aiming to preserve clinically meaningful signals. Decisions include filtering encounters or variables (e.g., based on excessive missingness), and detecting plausibility violations using predefined rules, followed by appropriate mitigation. Missingness and irregularity are mitigated by selecting suitable methods, as outlined in Section 3.4, or deferred to models capable of native DI handling in Stage 5. Because each preparation step encodes assumptions about the underlying data, Stage 3 makes these assumptions explicit and enables systematic evaluation of their downstream impact.

Stage 4 (Feature Extraction and Selection) derives predictive representations from the prepared data. A key structural design choice is the feature representation, ranging from statistical aggregates and trend descriptors to raw sequence inputs. Domain-specific clinical features are treated as part of the study's structural feature definition and may capture clinical proxies (e.g., EWS) or contextual information beyond time-series data (e.g., administered medication). In contrast, DI-derived features, such as masks, temporal gap/run statistics, or indicators reflecting mitigation choices, constitute experimental add-ons derived from Stage 2 and the preparation steps in Stage 3. By making these elements explicit, Stage 4 enables systematic testing of whether and how DI patterns contribute actionable predictive signal.

Stage 5 (Modeling) implements ML algorithms that learn the prediction function defined in Stage 1 using the representations produced in Stage 4. Choices span traditional ML methods (e.g., logistic regression, random forests, gradient boosting), sequence models designed to capture temporal dependencies (e.g., GRU, LSTM, transformers), and DI-aware architectures that natively handle or exploit DI patterns, as discussed in Section 3.4. Model choice interacts directly with Stage 3 and Stage 4 decisions, for example, XGBoost (Chen and Guestrin, 2016) can parse missing values but cannot leverage temporal structure, whereas vanilla sequence models require resolved irregularity or explicit temporal encodings. Therefore, Stage 5 enables systematic comparison by either holding the model constant to isolate the effects of upstream design choices or varying models and hyperparameters to assess robustness.

Stage 6 (Evaluation and Interpretability) assesses model performance and clinical utility. Because DI patterns vary across time and EMS systems, validation strategies must assess robustness to distribution shifts. In addition to standard metrics, task-specific clinical measures (e.g., the minimum prediction lead time at which a performance threshold is achieved or sensitivity at fixed specificity) capture EMS relevance. Evaluation should examine subgroup performance to identify potential fairness concerns, e.g., stratified by DI patterns. Finally, global or local explainable AI approaches address interpretability and allow for assessing how DI-aware representations support predictions. The metrics recorded at this stage form the basis for Stage 7 meta-analyses.

Stage 7 (Experiment Logging and Meta-Analysis) ensures full reproducibility by logging and versioning experimental configurations, artifacts, and results across all stages. Experiment logging enables the outlined systematic comparison of alternative DI-handling strategies, facilitates ablation studies, and supports meta-analyses of how data-quality decisions influence downstream model behavior.

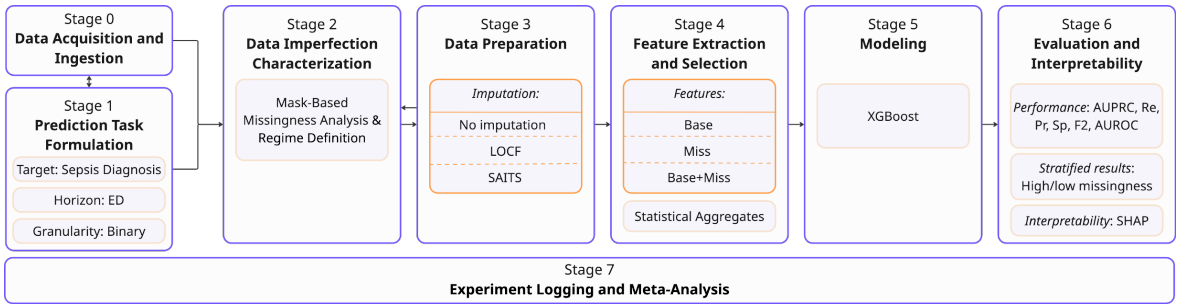

Figure 3: Instantiation of the proposed data-centric pipeline for early sepsis prediction.

5 EVALUATION

To demonstrate how the proposed DCP enables systematic, data-centric experimentation, we present a case study that instantiates the DCP for early sepsis prediction during EMS care and is illustrated in Figure 3. While the DCP supports multiple data imperfection types, this instantiation focuses exclusively on missingness. The objective is to quantify the impact of different missingness-handling strategies on model performance under varying *computational cost constraints* and *missingness levels*, which we define as two complementary scenarios reflecting common EMS research and deployment conditions:

1. **Scenario 1 (Computational Cost Constraints):** In resource-limited settings, computationally expensive deep-learning imputation may be infeasible. We therefore examine how simple and costly imputation strategies compare, and whether DI-aware feature design can mitigate the need for heavy preprocessing.
2. **Scenario 2 (Missingness Levels):** Because missingness levels vary widely across EMS encounters, we assess configuration performance under low- and high-missingness conditions.

5.1 Pipeline Instantiation

Stage 0: The NEMSIS 2024 dataset forms the basis for the analysis. For each encounter, the first 20 minutes of available vital sign recordings are extracted. Encounters are included if the interval between the first and last recorded vital is at least 20 minutes and at least five timestamped measurement points fall within this window[2]. Predictors include heart rate, respiratory rate, systolic blood pressure, and SpO_2.

Stage 1: The prediction task is defined as ICD-10-coded sepsis at the emergency department.

Stage 2: The final cohort comprised 96,551 encounters (3,878 sepsis cases; 4.01% prevalence). Global missingness rates for this cohort are 52.94% for systolic blood pressure, 47.23% for respiratory rate, 42.64% for SpO_2, and 28.46% for heart rate. Further outcome-stratified missingness statistics are reported in the supplementary material.

Following the DCP, we vary two experimental dimensions:

- **Stage 3 - Imputation strategy**: *No* imputation versus *LOCF*[3] versus *SAITS*[4].
- **Stage 4 - Feature selection**: Baseline statistical aggregates (*Base*) versus missingness-derived aggregates (*Miss*) versus the combined feature set (*Base+Miss*).

Stage 5: We fix the model class to XGBoost, which natively handles missing values. A complete feature specification and logged hyperparameters are provided in the supplementary material.

Stage 6: We apply an 80/20 train-test split and, to operationalize Scenario 2, quantify per-encounter missingness as the proportion of missing measurements across all vital signs and time steps and divide the test set into two subcohorts using the global median as the split threshold. Predicted probabilities are converted to binary decisions using F_2-optimal thresholds, prioritizing recall due to the clinical importance of avoiding missed sepsis cases. Given the low sepsis prevalence, we report AUROC as a threshold-independent discrimination metric but emphasize AUPRC, recall, precision, and F_2 as indicators of clinical utility. Each evaluation metric is accompanied by stratified bootstrap confidence intervals based on 1,000 resamples. In addition to performance metrics, we quantified the contribution of individual predictors by deriving SHAP values.

[2] A sensitivity analysis varying the minimum encounter duration, window length, and minimum measurement count is provided in the supplementary material.

[3] Missing values at the beginning of an encounter are replaced with the encounter-level mean.

[4] Implementation of (Du et al., 2023): https://github.com/WenjieDu/PyPOTS

Table 1: Performance of XGBoost across experiments. Metrics are means with a 95% confidence interval. Decision thresholds were optimized by maximizing the F_2 score. 'Imp' denotes the applied imputation strategy. Analyses were conducted on the full test cohort and on low- and high-missingness subcohorts. Within each subcohort, the best value for each metric is indicated in bold.

Feature Set	Subcohort	Imp	Threshold	AUROC	AUPRC	Precision	Recall	F_2	Specificity
Base	Full	No	0.0578	0.777 (0.763–0.791)	0.159 (0.141–0.177)	0.122 (0.110–0.143)	0.604 (0.519–0.654)	0.337 (0.318–0.355)	0.818 (0.787–0.869)
Base	Low-miss	No	0.0410	0.769 (0.749–0.789)	0.146 (0.125–0.171)	0.136 (0.113–0.163)	0.525 (0.449–0.606)	0.331 (0.305–0.358)	0.856 (0.808–0.896)
Base	High-miss	No	0.0243	0.702 (0.677–0.726)	0.116 (0.098–0.139)	0.098 (0.080–0.130)	0.498 (0.379–0.595)	0.272 (0.248–0.295)	0.806 (0.733–0.888)
Base	Full	LOCF	0.0678	0.789 (0.776–0.803)	0.162 (0.145–0.180)	0.139 (0.124–0.153)	0.576 (0.536–0.626)	0.353 (0.335–0.372)	0.851 (0.817–0.869)
Base	Low-miss	LOCF	0.0642	0.805 (0.787–0.821)	**0.181 (0.155–0.206)**	0.149 (0.131–0.170)	0.600 (0.538–0.657)	0.372 (0.346–0.396)	0.854 (0.826–0.882)
Base	High-miss	LOCF	0.0524	0.774 (0.754–0.792)	0.148 (0.126–0.171)	0.122 (0.108–0.152)	0.612 (0.499–0.657)	0.338 (0.311–0.362)	0.816 (0.799–0.888)
Base	Full	SAITS	0.0754	0.788 (0.773–0.802)	**0.174 (0.155–0.194)**	0.148 (0.122–0.173)	0.537 (0.472–0.618)	0.349 (0.330–0.369)	**0.868 (0.820–0.898)**
Base	Low-miss	SAITS	0.0759	0.793 (0.774–0.811)	0.178 (0.152–0.207)	**0.151 (0.126–0.188)**	0.559 (0.468–0.638)	0.360 (0.336–0.386)	**0.864 (0.825–0.916)**
Base	High-miss	SAITS	0.0754	0.782 (0.762–0.803)	**0.174 (0.148–0.203)**	0.148 (0.109–0.183)	0.527 (0.439–0.665)	0.343 (0.315–0.370)	0.868 (0.784–0.916)
Miss	Full	No	0.0356	0.539 (0.522–0.556)	0.055 (0.048–0.064)	0.042 (0.039–0.048)	**0.887 (0.539–0.980)**	0.176 (0.167–0.185)	0.151 (0.037–0.551)
Miss	Low-miss	No	0.0390	0.550 (0.525–0.574)	0.058 (0.047–0.070)	0.047 (0.039–0.056)	**0.706 (0.478–0.977)**	0.182 (0.168–0.197)	0.376 (0.042–0.642)
Miss	High-miss	No	0.0342	0.528 (0.505–0.551)	0.054 (0.044–0.066)	0.042 (0.038–0.048)	**0.885 (0.534–1.000)**	0.175 (0.161–0.187)	0.152 (0.001–0.567)
Base+Miss	Full	No	0.0625	0.786 (0.773–0.800)	0.166 (0.148–0.186)	0.128 (0.114–0.154)	0.598 (0.503–0.657)	0.344 (0.325–0.362)	0.828 (0.800–0.877)
Base+Miss	Low-miss	No	0.0510	0.779 (0.759–0.797)	0.163 (0.138–0.189)	0.138 (0.117–0.168)	0.549 (0.463–0.610)	0.342 (0.315–0.368)	0.853 (0.822–0.899)
Base+Miss	High-miss	No	0.0269	0.720 (0.696–0.743)	0.122 (0.104–0.145)	0.107 (0.092–0.128)	0.509 (0.428–0.572)	0.289 (0.264–0.314)	0.822 (0.793–0.871)
Base+Miss	Full	LOCF	0.0693	0.790 (0.777–0.803)	0.166 (0.148–0.184)	0.134 (0.118–0.148)	0.567 (0.519–0.627)	0.344 (0.326–0.361)	0.846 (0.809–0.868)
Base+Miss	Low-miss	LOCF	0.0654	0.803 (0.785–0.819)	0.173 (0.149–0.199)	0.140 (0.125–0.159)	0.611 (0.549–0.666)	0.365 (0.340–0.390)	0.841 (0.816–0.870)
Base+Miss	High-miss	LOCF	0.0675	0.773 (0.753–0.793)	0.158 (0.133–0.185)	0.133 (0.105–0.191)	0.531 (0.404–0.630)	0.328 (0.301–0.354)	0.852 (0.781–0.928)
Base+Miss	Full	SAITS	0.0736	**0.797 (0.783–0.811)**	0.174 (0.156–0.193)	**0.149 (0.133–0.173)**	0.568 (0.500–0.617)	**0.362 (0.342–0.382)**	0.863 (0.840–0.899)
Base+Miss	Low-miss	SAITS	0.0703	**0.806 (0.788–0.822)**	0.178 (0.154–0.205)	0.148 (0.130–0.179)	0.616 (0.523–0.672)	**0.376 (0.351–0.401)**	0.849 (0.818–0.897)
Base+Miss	High-miss	SAITS	0.0795	**0.787 (0.766–0.807)**	0.172 (0.147–0.200)	**0.157 (0.130–0.184)**	0.520 (0.464–0.580)	**0.354 (0.325–0.382)**	**0.882 (0.852–0.907)**

Stage 7: Artifacts are logged at multiple stages, including missingness statistics, feature sets, model hyperparameters, and evaluation metrics, which enable the subsequent analysis.

5.2 Evaluation Results

Model performance across all configurations and missingness levels is summarized in Table 1, and detailed results are provided in the supplementary material. Regarding the two introduced scenarios, three key findings emerge. The first two findings focus on Scenario 1 while also reflecting missingness-level effects relevant to Scenario 2. The third finding addresses Scenario 2 explicitly.

(i) Comparing simple and computationally intensive imputation suggests modest overall performance differences, with SAITS consistently improving precision and performance under high missingness (e.g., Base–SAITS vs. Base–LOCF: precision 0.148 vs. 0.122, AUPRC 0.174 vs. 0.148). In the remaining settings, confidence intervals largely overlap, indicating limited benefit despite the vastly higher computational cost of SAITS (LOCF: 0.33 seconds; SAITS 100-epoch training and imputation on an NVIDIA RTX 6000 GPU: 1593.31 seconds).

(ii) Comparing the relative effects of imputation (Base–LOCF, Base–SAITS) and missingness-derived features (Base+Miss–No) shows that imputation is the primary driver of performance, particularly under high missingness. In the high-missingness subcohort, Base+Miss–No improves on the non-imputed baseline but remains limited (AUPRC 0.122), whereas Base–LOCF and Base–SAITS achieve higher AUPRC (0.148 and 0.174), indicating that missingness-derived features do not substitute for imputation when missingness is severe. In the full cohort, Base+Miss–No (AUPRC 0.166) is below Base–SAITS (0.174) but comparable to Base–LOCF (0.162), suggesting that missingness-derived features provide predictive signal and can partially offset the absence of imputation. When both strategies are combined, Base+Miss–LOCF improves on Base–LOCF in the high-missingness subcohort (AUPRC 0.158 vs. 0.148; Precision 0.133 vs. 0.122). The SHAP analysis of Base+Miss–LOCF (Appendix C) further indicates the contribution of missingness-derived features, for example, by the feature heartrate_time_since_imperfect_mean, which captures the mean time since the last missing heart rate. For Base+Miss–SAITS, the effect is smaller but is reflected in slightly higher AUROC and F_2.

(iii) Across all configurations, performance is consistently higher in the low- than in the high-missingness subcohort (e.g., Base+Miss–No: $\Delta AUPRC = +0.041$ and $\Delta F_2 = +0.053$), confirming that missingness remains a key limiting factor to model performance, regardless of the handling strategy.

6 DISCUSSION

The proposed pipeline for prehospital ML provides a systematic structure for characterizing, handling, and leveraging DI. Beyond its conceptual organization, the pipeline provides practical experimentation benefits by making DI handling decisions explicit, testable, and attributable. This section discusses the research questions and future work.

The DCP addresses the first research question by regarding DI, which includes missingness, plausibility violations, and irregularity, as first-class experimental entities. Stage 2 of the pipeline enforces explicit DI characterization, providing insights into patterns that are otherwise difficult to consider in multivariate time-series. This characterization allows to *tackle* DI in Stages 3 and 4 by testing alternative preparation strategies and incorporating DI-derived representations, thereby enabling models to *leverage* workflow- or physiology-dependent patterns. In contrast to general process models (Studer et al., 2021; Fayyad et al., 1996), which position DI exclusively within data cleaning, the DCP turns DI types into explicit, traceable variables and enables downstream assessment of whether DI patterns contain clinically relevant signals. The instantiation validates the pipeline's feasibility by allowing us to isolate the relative effects of imputation and missingness-derived features under fixed modeling conditions, to assess their interaction across missingness regimes, and to quantify cost–performance trade-offs. The evaluation demonstrates that DI-derived features contribute a consistent but modest complementary signal alongside physiological features, and that the benefits of imputation become particularly visible under high-missingness conditions, insights made possible precisely because DI-related decisions were defined, isolated, and tracked.

The second research question concerns the architectural properties of the DCP. *Systematic experimentation* is enabled by the eight-stage decomposition of the DCP, which exposes design decisions at each step and opens up a structured multidimensional space. The instantiation explored a 3×3 configuration grid, yielding 7 valid configurations, as the Miss-only feature set does not require imputation. Although this is only a small subset, the pipeline structure ensured that all configurations were directly comparable, enabling interpretable differences in model behavior to be attributed to data-centric choices rather than model variance. *Reproducibility* is supported by Stage 7, which mandates explicit logging of configurations, DI characterization outputs, preprocessing artifacts, and model parameters. The instantiation illustrates how this logging supports meta-analysis of DI-handling strategies and prevents the silent propagation of DI-related assumptions, which is a common challenge in generic pipelines (Jarrahi et al., 2023). The DCP is *data-centric* by design, with early stages focusing exclusively on data rather than models, and with systematic profiling, representation, and propagation of DI rather than filtering it away. The evaluation also incorporated a missingness-stratified analysis, which indicated that higher levels of missingness may constrain model performance.

Several limitations remain. The DCP currently focuses on uniform-timestamp missingness, plausibility violations, and irregularity. Additional classes, such as inter-series discrepancy (e.g., (Zhang et al., 2023)), should be incorporated to better capture the complexity of real-world prehospital data. Furthermore, the instantiation explores only a small subset of the DCP's experimental space and focuses on missingness. Notably, a large instantiation and complete evaluation of the DCP is beyond the scope of a single study. The DCP is designed to support progressively expanding evaluations as the research community accumulates domain insights and additional datasets. Future work will incrementally validate the feasibility and utility of the DCP through broader experimentation. Important next steps include incorporating domain-informed plausibility rules and utilizing sequence models, which may uncover more substantial DI-related effects with fine-grained temporal details. Likewise, validation across datasets and outcomes will be essential to advance ML-driven early warning in prehospital emergency care.

Finally, although designed with EMS in mind, the DCP is conceptually applicable to any ML setting involving imperfect time-series data, including wearables and remote monitoring. Its systematic DI-centric structure offers a foundation for developing robust, transparent, and clinically meaningful models across diverse real-world data environments.

7 CONCLUSIONS

This paper introduces a customizable data-centric experimentation pipeline for prehospital machine learn-

ing that treats data imperfection as a first-class experimental variable. The pipeline structures time-series experimentation into explicit, traceable stages and highlights data-quality decision points to support systematic, reproducible exploration. We instantiated the pipeline for early sepsis prediction on the NEMSIS 2024 dataset to validate its feasibility to facilitate systematic, data-centric experimentation. Two experimental dimensions were examined: imputation strategy and missingness-derived features. The evaluation insights showed that missingness-derived features contribute modest but complementary gains when combined with physiological information, imputation yields measurable performance improvements, and missingness remains the main limiting factor to performance. Overall, the proposed data-centric pipeline enables explicit examination of data-quality trade-offs and can guide the design of robust, reproducible, and clinically meaningful early warning models. Future work will extend data-centric experimentation to additional imperfection types, handling strategies, modeling paradigms, and EMS datasets.

ACKNOWLEDGEMENTS

The content reproduced from the NEMSIS Database remains the property of the National Highway Traffic Safety Administration (NHTSA). The NHTSA is not responsible for any claims arising from works based on the original Data, Text, Tables, or Figures.

The readability of this text has been improved using AI technologies.

REFERENCES

Abe, D., Inaji, M., Hase, T., Takahashi, S., Sakai, R., et al. (2022). A Prehospital Triage System to Detect Traumatic Intracranial Hemorrhage Using Machine Learning Algorithms. *JAMA Network Open*, 5(6).

Akel, M., Carey, K., Winslow, C., Churpek, M., and Edelson, D. (2021). Less is more: Detecting clinical deterioration in the hospital with machine learning using only age, heart rate, and respiratory rate. *Resuscitation*, 168:6–10.

Alasal, L. M., Hammarlund, E. U., Pienta, K. J., Rönnstrand, L., and Kazi, J. U. (2025). XeroGraph: enhancing data integrity in the presence of missing values with statistical and predictive analysis. *Bioinformatics Advances*, 5(1).

Alghatani, K., Ammar, N., Rezgui, A., and Shaban-Nejad, A. (2021). Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation. *JMIR Medical Informatics*, 9(5).

Amer, A. Y. A., Wouters, F., Vranken, J., Boer, D. d. K.-d., Smit-Fun, V., et al. (2020). Vital Signs Prediction and Early Warning Score Calculation Based on Continuous Monitoring of Hospitalised Patients Using Wearable Technology. *Sensors*, 20(22):6593.

Bazo-Alvarez, J. C., Morris, T. P., Carpenter, J. R., and Petersen, I. (2021). Current Practices in Missing Data Handling for Interrupted Time Series Studies Performed on Individual-Level Data: A Scoping Review in Health Research. *Clinical Epidemiology*, 13(0):603–613.

Bleyer, A. J., Vidya, S., Russell, G. B., Jones, C. M., Sujata, L., et al. (2011). Longitudinal analysis of one million vital signs in patients in an academic medical center. *Resuscitation*, 82(11):1387–1392.

Bourke-Matas, E., Doan, T., Bowles, K., and Bosley, E. (2024). A prediction model for prehospital clinical deterioration: The use of early warning scores. *Academic Emergency Medicine*, 31(11):1139–1149.

Chae, M., Gil, H.-W., Cho, N.-J., and Lee, H. (2022). Machine Learning-Based Cardiac Arrest Prediction for Early Warning System. *Mathematics*, 10(12):2049.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1):6085.

Chen, Q., Qin, Y., Jin, Z., Zhao, X., He, J., et al. (2024). Enhancing Performance of the National Field Triage Guidelines Using Machine Learning: Development of a Prehospital Triage Model to Predict Severe Trauma. *Journal of Medical Internet Research*, 26.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Cheng, C.-Y., Kung, C.-T., Chen, F.-C., Chiu, I.-M., Lin, C.-H. R., et al. (2022). Machine learning models for predicting in-hospital mortality in patient with sepsis: Analysis of vital sign dynamics. *Frontiers in Medicine*, 9.

Choi, A., Choi, S. Y., Chung, K., Chung, H. S., Song, T., et al. (2023). Development of a machine learning-based clinical decision support system to predict clinical deterioration in patients visiting the emergency department. *Scientific Reports*, 13(1):8561.

Choi, A., Chung, K., Chung, S. P., Lee, K., and Hyun, H. o. (2022a). Advantage of Vital Sign Monitoring Using a Wireless Wearable Device for Predicting Septic Shock in Febrile Patients in the Emergency Department: A Machine Learning-Based Analysis. *Sensors*, 22(18):7054.

Choi, A., Kim, M. J., Sung, J. M., Kim, S., Lee, J., et al. (2022b). Development of Prediction Models for Acute Myocardial Infarction at Prehospital Stage with Machine Learning Based on a Nationwide Database. *Journal of Cardiovascular Development and Disease*, 9(12):430.

Churpek, M. M., Carey, K. A., Snyder, A., Winslow, C. J., Gilbert, E., et al. (2024). Multicenter Development and Prospective Validation of eCARTv5: A Gradi-

ent Boosted Machine Learning Early Warning Score. *medRxiv*.

Du, W., Côté, D., and Liu, Y. (2023). Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., et al. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8(1):140.

Evans, L., Rhodes, A., Alhazzani, W., Antonelli, M., Coopersmith, C. M., et al. (2021). Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive Care Medicine*, 47(11):1181–1247.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–37.

Hasan, M., Bath, P., Marincowitz, C., Sutton, L., Pilbery, R., et al. (2022). Pre-hospital prediction of adverse outcomes in patients with suspected COVID-19: Development, application and comparison of machine learning and deep learning methods. *Computers in Biology and Medicine*, 151(Pt A):106024.

Hayashi, Y., Shimada, T., Hattori, N., Shimazui, T., Yoshida, Y., et al. (2021). A prehospital diagnostic algorithm for strokes using machine learning: a prospective observational study. *Scientific Reports*, 11(1):20519.

Jarrahi, M. H., Memariani, A., and Guha, S. (2023). The Principles of Data-Centric AI. *Communications of the ACM*, 66(8):84–92.

Jehangir, B. and Li, W. (2025). Enhancing Early Warning Systems: Predicting Next Vital Signs Using Recurrent Neural Networks and Attention Models. In *Annual Hawaii International Conference on System Sciences*.

Kang, D.-Y., Cho, K.-J., Kwon, O., Kwon, J.-m., Jeon, K.-H., et al. (2020). Artificial intelligence algorithm to predict the need for critical care in prehospital emergency medical services. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 28(1):17.

Kauppi, W., Imberg, H., Herlitz, J., Molin, O., Axelsson, C., et al. (2025). Advancing a machine learning-based decision support tool for pre-hospital assessment of dyspnoea by emergency medical service clinicians: a retrospective observational study. *BMC Emergency Medicine*, 25(1):2.

Kim, D., You, S., So, S., Lee, J., Yook, S., et al. (2018). A data-driven artificial intelligence model for remote triage in the prehospital environment. *PLoS ONE*, 13(10).

Kim, J.-h., Cho, E. Y., Choi, Y., Won, J.-Y., Cheon, S. H., et al. (2025). Deep Learning–Based Early Warning Systems in Hospitalized Patients at Risk of Code Blue Events and Length of Stay: Retrospective Real-World Implementation Study. *JMIR Medical Informatics*, 13.

Kitano, S., Ogawa, K., Igarashi, Y., Nishimura, K., Osawa, S., et al. (2023). Development of a Machine Learning Model to Predict Cardiac Arrest during Transport of Trauma Patients. *Journal of Nippon Medical School*, 90(2):186–193.

Krafft, T., Stieler, F., and Bauer, B. (2025). Early warning score trend analysis: A data-driven approach for emergency medical services. In *2025 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–7.

Kwon, J., Lee, Y., Lee, Y., Lee, S., and Park, J. (2018). An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest. *Journal of the American Heart Association*, 7(13).

Larsson, A., Berg, J., Gellerfors, M., and Wärnberg, M. G. (2021). The advanced machine learner XGBoost did not reduce prehospital trauma mistriage compared with logistic regression: a simulation study. *BMC Medical Informatics and Decision Making*, 21(1):192.

Lee, H.-Y., Kuo, P.-C., Qian, F., Li, C.-H., Hu, J.-R., et al. (2024). Prediction of In-Hospital Cardiac Arrest in the Intensive Care Unit: Machine Learning–Based Multimodal Approach. *JMIR Medical Informatics*, 12.

Li, Y., Ye, W., Yang, K., Zhang, S., He, X., et al. (2022). Prediction of cardiac arrest in critically ill patients based on bedside vital signs monitoring. *Computer Methods and Programs in Biomedicine*, 214:106568.

Lindskou, T. A., Ward, L. M., Søvsø, M. B., Mogensen, M. L., and Christensen, E. F. (2023). Prehospital Early Warning Scores to Predict Mortality in Patients Using Ambulances. *JAMA Network Open*, 6(8).

Lipton, Z. C., Kale, D. C., and Wetzel, R. (2016). Modeling Missing Data in Clinical Time Series with RNNs. In *Machine Learning for Healthcare*.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, 83(404):1198–1202.

Liu, N. T., Holcomb, J. B., Wade, C. E., Darrah, M. I., and Salinas, J. (2014). Utility of Vital Signs, Heart Rate Variability and Complexity, and Machine Learning for Identifying the Need for Lifesaving Interventions in Trauma Patients. *Shock*, 42(2):108–114.

Majouni, S., Tennankore, K., and Abidi, S. S. R. (2025). Machine Learning-Based Early Prediction of Hospitalization in Hemodialysis Patients During Ambulance Transport to the Emergency Department.

Marsden, M. E. R., Perkins, Z. B., Pisirir, E., Marsh, W., Kyrimi, E., et al. (2025). Early clinical evaluation of a machine-learning system for risk prediction of trauma-induced coagulopathy in the prehospital setting. *Emergency Medicine Journal*, 42(10):654–661.

Mercaldo, S. F. and Blume, J. D. (2020). Missing data and prediction: the pattern submodel. *Biostatistics*, 21(2):236–252.

Nederpelt, C. J., Mokhtari, A. K., Alser, O., Tsiligkaridis, T., Roberts, J., et al. (2021). Development of a field artificial intelligence triage tool. *Journal of Trauma and Acute Care Surgery*, 90(6):1054–1060.

NEMSIS (2019). 2019 National EMS Data Report. Technical report, National Highway Traffic Safety Administration Office. Available at: https://nemsis.org/wp-

content/uploads/2025/02/NEMSIS-End-of-Year-Report-2019-9-13-22.pdf (Accessed: 2025-11-23).

NEMSIS (2024). 2024 National EMS Data Report. Technical report, National Highway Traffic Safety Administration Office. Available at: https://nemsis.org/wp-content/uploads/2025/09/NEMSIS-End-of-Year-Report-2024-9-24-25.pdf (Accessed: 2025-11-23).

Ortiz, B. L., Gupta, V., Kumar, R., Jalin, A., Cao, X., et al. (2024). Data Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care. *JMIR mHealth and uHealth*, 12.

Perez-Lebel, A., Varoquaux, G., Morvan, M. L., Josse, J., and Poline, J.-B. (2022). Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*, 11.

Phimphisan, S., Wichaiyo, W., Saengprajak, N., Jantu, W., Sirigit, W., et al. (2025). Artificial Intelligence for EMS Triage: A Data-Driven Approach to Emergency Patient Prioritization in Kalasin Province, Thailand. *Engineering, Technology & Applied Science Research*, 15(4):24204–24210.

Qian, L., Yang, Y., Du, W., Wang, J., Dobsoni, R., et al. (2024). Beyond Random Missingness: Clinically Rethinking for Healthcare Time Series Imputation.

Rangan, E. S., Pathinarupothi, R. K., Anand, K. J. S., and Snyder, M. P. (2022). Performance effectiveness of vital parameter combinations for early warning of sepsis—an exhaustive study using machine learning. *JAMIA Open*, 5(4).

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Senthil Pandi, S., Kumaragurubaran, T., Vijay Raj, S. R., and Vigneshwaran, R. (2024). Predictive Modelling of Critical Vital Signs in ICU Patients by Machine Learning: An Early Warning System for Improved Patient Outcomes. *2024 3rd International Conference for Innovation in Technology (INOCON)*, 00:1–6.

Shamout, F. E., Zhu, T., Sharma, P., Watkinson, P. J., and Clifton, D. A. (2019). Deep Interpretable Early Warning System for the Detection of Clinical Deterioration. *IEEE Journal of Biomedical and Health Informatics*, 24(2):437–446.

Shirakawa, T., Sonoo, T., Ogura, K., Fujimori, R., Hara, K., et al. (2020). Institution-Specific Machine Learning Models for Prehospital Assessment to Predict Hospital Admission: Prediction Model Development Study. *JMIR Medical Informatics*, 8(10).

Silva, D. B. d., Schmidt, D., Costa, C. A. d., Righi, R. d. R., and Eskofier, B. (2021). DeepSigns: A predictive model based on Deep Learning for the early detection of patient health deterioration. *Expert Systems with Applications*, 165:113905.

Sim, T., Hahn, S., Kim, K.-J., Cho, E.-Y., Jeong, Y., et al. (2025). Preserving Informative Presence: How Missing Data and Imputation Strategies Affect the Performance of an AI-Based Early Warning Score. *Journal of Clinical Medicine*, 14(7):2213.

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., et al. (2021). Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction*, 3(2):392–413.

Su, C.-F., Chiu, S.-I., Jang, J.-S. R., and Lai, F. (2022). Improved inpatient deterioration detection in general wards by using time-series vital signs. *Scientific Reports*, 12(1):11901.

Takeda, M., Oami, T., Hayashi, Y., Shimada, T., Hattori, N., et al. (2022). Prehospital diagnostic algorithm for acute coronary syndrome using machine learning: a prospective observational study. *Scientific Reports*, 12(1):14593.

Tamminen, J., Kallonen, A., Hoppu, S., and Kalliomäki, J. (2021). Machine learning model predicts short-term mortality among prehospital patients: A prospective development study from Finland. *Resuscitation Plus*, 5:100089.

Thiele, D., Rodseth, R., Friedland, R., Berger, F., Mathew, C., et al. (2025). Machine Learning Models for the Early Real-Time Prediction of Deterioration in Intensive Care Units—A Novel Approach to the Early Identification of High-Risk Patients. *Journal of Clinical Medicine*, 14(2):350.

Wang, J., Du, W., Yang, Y., Qian, L., Cao, W., et al. (2025). Deep learning for multivariate time series imputation: A survey. In Kwok, J., editor, *34th International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 10696–10704. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Ward, L. M., Lindskou, T. A., Mogensen, M. L., Christensen, E. F., and Søvsø, M. B. (2025). Machine learning to improve predictive performance of prehospital early warning scores. *Scientific Reports*, 15(1):21459.

Weidman, A. C., Malakouti, S., Salcido, D. D., Zikmund, C., Patel, R., and Others (2025). A Machine Learning Trauma Triage Model for Critical Care Transport. *JAMA Network Open*, 8(6).

Zhang, J., Zheng, S., Cao, W., Bian, J., and Li, J. (2023). Warpformer: A Multi-scale Modeling Approach for Irregular Clinical Time Series. In *9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3273–3285.

APPENDIX

A. Supplementary Material (Online)

All materials are available on our Open Science Framework page: https://osf.io/ke8sy/overview?view_only=86488bba2f764c78bc12d3afd44ef5ec

B. Overview of Representative Prehospital Machine Learning Studies

A summary of representative prehospital studies is provided in Table 2.

Table 2: Overview of utilized temporal resolutions, modeling approaches, prediction targets, and data imperfection handling in EMS deterioration prediction studies.

Dimension	Topic	Representative studies
Temporal resolution	Single timepoint	(Abe et al., 2022; Bourke-Matas et al., 2024; Chen et al., 2024; Choi et al., 2022b; Hasan et al., 2022; Hayashi et al., 2021; Kang et al., 2020; Kauppi et al., 2025; Kim et al., 2018; Kitano et al., 2023; Larsson et al., 2021; Lindskou et al., 2023; Majouni et al., 2025; Marsden et al., 2025; Nederpelt et al., 2021; Phimphisan et al., 2025; Shirakawa et al., 2020; Takeda et al., 2022; Tamminen et al., 2021)
	Time-series	(Krafft et al., 2025; Liu et al., 2014; Ward et al., 2025; Weidman et al., 2025)
Modelling approach	Logistic Regression	(Abe et al., 2022; Bourke-Matas et al., 2024; Choi et al., 2022b; Hayashi et al., 2021; Kauppi et al., 2025; Kim et al., 2018; Kitano et al., 2023; Krafft et al., 2025; Lindskou et al., 2023; Shirakawa et al., 2020; Takeda et al., 2022; Ward et al., 2025)
	Gradient Boosting	(Abe et al., 2022; Chen et al., 2024; Choi et al., 2022b; Hasan et al., 2022; Hayashi et al., 2021; Kauppi et al., 2025; Kitano et al., 2023; Majouni et al., 2025; Shirakawa et al., 2020; Takeda et al., 2022; Ward et al., 2025; Weidman et al., 2025)
	Random Forest	(Abe et al., 2022; Hayashi et al., 2021; Kim et al., 2018; Kitano et al., 2023; Majouni et al., 2025; Shirakawa et al., 2020; Takeda et al., 2022; Tamminen et al., 2021; Ward et al., 2025)
	Support Vector Machine	(Abe et al., 2022; Hasan et al., 2022; Hayashi et al., 2021; Majouni et al., 2025; Takeda et al., 2022)
	Deep Learning	(Kang et al., 2020; Kim et al., 2018; Liu et al., 2014; Nederpelt et al., 2021; Phimphisan et al., 2025; Shirakawa et al., 2020)
Prediction target	Short-term mortality	(Kauppi et al., 2025; Lindskou et al., 2023; Tamminen et al., 2021; Ward et al., 2025)
	ICU admission / escalation of care	(Bourke-Matas et al., 2024; Chen et al., 2024; Kang et al., 2020; Kauppi et al., 2025; Phimphisan et al., 2025; Tamminen et al., 2021; Ward et al., 2025)
	Time-critical diagnoses	(Abe et al., 2022; Choi et al., 2022b; Hayashi et al., 2021; Kitano et al., 2023; Marsden et al., 2025; Takeda et al., 2022)
	Lifesaving interventions	(Liu et al., 2014; Nederpelt et al., 2021; Phimphisan et al., 2025; Weidman et al., 2025)
DI handling	Complete-case analysis	(Kim et al., 2018; Larsson et al., 2021; Tamminen et al., 2021)
	Mean / Median imputation	(Abe et al., 2022; Choi et al., 2022b; Kang et al., 2020; Majouni et al., 2025; Nederpelt et al., 2021; Phimphisan et al., 2025)
	Multiple imputation	(Bourke-Matas et al., 2024; Hasan et al., 2022; Hayashi et al., 2021; Kitano et al., 2023)
	Model-native handling	(Chen et al., 2024; Marsden et al., 2025; Shirakawa et al., 2020)
	Informative missingness	(Shirakawa et al., 2020; Ward et al., 2025; Weidman et al., 2025)
	Outlier filtering	(Larsson et al., 2021; Phimphisan et al., 2025)

C. SHAP Summary Plot of the Base+Miss–LOCF Model

Figure 4 presents the SHAP summary plot for the Base+Miss–LOCF configuration in the test cohort.

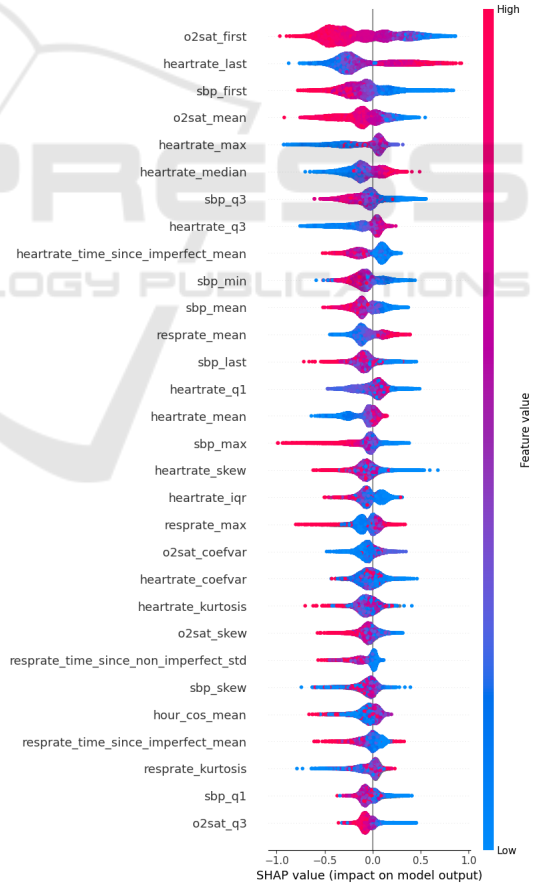

Figure 4: SHAP summary plot for the Base+Miss–LOCF configuration in the full test cohort, depicting the relative importance and direction of influence of the 20 most contributory features.

D. Abbreviations

Table 3 summarizes all abbreviations used in the paper for reference.

Table 3: Abbreviations used in the paper.

Acronym	Meaning
AI	Artificial Intelligence
AUROC	Area Under the Receiver Operating Characteristic Curve
AUPRC	Area Under the Precision–Recall Curve
CRISP-ML(Q)	Cross-Industry Standard Process for ML (with Quality)
DCP	Data-Centric Pipeline
DI	Data Imperfection
eCART	Electronic Cardiac Arrest Risk Triage (FDA-cleared EWS)
ECG	Electrocardiography
EMS	Emergency Medical Services
EWS	Early Warning Score
FDA	U.S. Food and Drug Administration
F_2	F_2 score (recall-weighted F-measure)
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
GRU-D	GRU with Decay
ICD-10	International Classification of Diseases, 10th Revision
ICU	Intensive Care Unit
KDD	Knowledge Discovery in Databases
LOCF	Last Observation Carried Forward
LSTM	Long Short-Term Memory
MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Not At Random
ML	Machine Learning
NEMSIS	National EMS Information System
SAITS	Self-Attention-based Imputation for Time Series
SHAP	SHapley Additive exPlanations
SpO_2	Peripheral oxygen saturation
XGBoost	Extreme Gradient Boosting