

Multimodal ICU In-hospital Mortality Prediction: A Scoping Review of Transparency, Calibration, and Translation Readiness

Alexander Bakumenko

abakume@clemson.edu

Clemson University <https://orcid.org/0000-0001-7212-9573>

Janine Hoelscher

Clemson University <https://orcid.org/0000-0003-3970-0613>

D. Hudson Smith

Clemson University <https://orcid.org/0000-0003-3041-4602>

Svetlana Bakumenko

Linnaeus University

Aaron J. Masino

Clemson University <https://orcid.org/0000-0002-2684-0548>

Jihad S. Obeid

Medical University of South Carolina <https://orcid.org/0000-0002-7193-7779>

Systematic Review

Keywords: ICU, In-hospital mortality, Multimodal learning, Electronic health records, Calibration, Decision curve analysis, Interpretability, Reproducibility, Scoping review

Posted Date: March 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9217081/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Multimodal ICU In-hospital Mortality Prediction: A Scoping Review of Transparency, Calibration, and Translation Readiness

Alexander Bakumenko^{a,*}, Janine Hoelscher^b, D. Hudson Smith^c, Svetlana Bakumenko^d, Aaron J. Masino^{e,f}, Jihad S. Obeid^g

^aClemson University, School of Computing, Charleston, SC, USA

^bClemson University, Department of Bioengineering, Clemson, SC, USA

^cClemson University, Department of Mathematical and Statistical Sciences, Clemson, SC, USA

^dLinnaeus University, Department of Informatics, Växjö, Sweden

^eClemson University, School of Computing, Clemson, SC, USA

^fClemson University, Institute for Human Genetics, Greenwood, SC, USA

^gMedical University of South Carolina, Department of Public Health Sciences, Charleston, SC, USA

Abstract

Background: Multimodal machine learning for intensive care unit (ICU) in-hospital mortality (IHM) prediction has rapidly expanded, yet reporting practices relevant to clinical readiness, including calibration, clinical utility, robustness to data missingness, and actionable transparency, remain heterogeneous. **Objective:** To map the landscape of multimodal IHM prediction studies and synthesize evidence on (i) fusion architectures and transparency implementations, (ii) evaluation beyond discrimination (calibration, clinical utility, and per-case uncertainty), and (iii) operational readiness (missing modality handling, prediction timeliness, external validation, subgroup evaluation, and reproducibility artifacts). **Methods:** We conducted a scoping review of peer-reviewed studies (English, January 2010 – November 2025) that fused at least two data modalities by provenance and predicted IHM in adult ICU populations. Searches were performed in PubMed, Scopus, and Google Scholar. Screening and extraction used an established AI-assisted workflow with two-reviewer verification. We summarized extracted fields from 115 included studies using descriptive statistics and narrative synthesis, including thematic analysis of author-reported gaps, limitations, and future directions. **Results:** Among 115 included studies, feature-level (early) fusion predominated (n = 80, 69.6%), followed by intermediate fusion (n = 30, 26.1%), late fusion (n = 4, 3.5%), and hybrid fusion (n=1, 0.9%). The modality landscape was dominated by physiologic time-series and coded/tabular EHR data; clinical text, imaging, and high-frequency waveforms each appeared in a minority of studies. Feature-level interpretability was reported in 21 studies (18.3%), with full per-case attribution in only 13 (11.3%); modality-level attribution was identified in 2 studies (1.7%), and missing-modality policies in 7 (6.1%). Calibration was assessed in 40 studies (34.8%), decision curve analysis in 20 (17.4%), and per-case uncertainty quantification in 1 (0.9%). External validation was performed in 28 studies (24.3%), prediction timeliness was evaluated in 22 (19.1%), and code repositories were provided in 29 (25.2%); no study reported a structured model card. **Conclusions:** The multimodal IHM literature emphasizes discrimination performance and early fusion, yet the evidence base remains immature for clinical translation, with systematic gaps in calibration, clinical utility, per-case transparency, missing-modality robustness, and reproducibility. These findings motivate standardized reporting and evaluation frameworks tailored to multimodal ICU prediction.

Keywords: ICU, In-hospital mortality, Multimodal learning, Electronic health records, Calibration, Decision curve analysis, Interpretability, Reproducibility, Scoping review

*Corresponding author.

Email addresses: abakume@clemson.edu (Alexander Bakumenko), janineh@clemson.edu (Janine Hoelscher), dane2@clemson.edu (D. Hudson Smith), svetlanabakumenko@outlook.com (Svetlana Bakumenko), amasino@clemson.edu (Aaron J. Masino), jobeid@musc.edu (Jihad S. Obeid)

1. Introduction

1.1. Clinical and informatics context

The intensive care unit (ICU) is among the highest-acuity settings in hospital medicine, managing patients with acute organ dysfunction under extreme time pressure and resource constraints. Adult ICU populations carry substantial in-hospital mortality, and early, accurate identification of patients at elevated risk is clinically valuable: it enables timely targeted interventions, more efficient allocation of specialist personnel and equipment, and earlier prognostic communication with families [1]. Historically, bedside risk stratification has relied on physiology-based severity scores. APACHE II [2] and SAPS II [3] aggregate vital signs, laboratory values, and clinical state into mortality risk estimates that remain foundational in critical care. In parallel, SOFA [4] summarizes the extent of organ dysfunction and is widely used to characterize physiologic severity and trajectories over time.

The release of large de-identified EHR databases, such as MIMIC-III [5], MIMIC-IV [6], and the eICU Collaborative Research Database [7], catalyzed a decade of reproducible machine learning research. Standardized preprocessing and benchmarking pipelines [8] enabled rigorous comparison across model families, and contemporary approaches have demonstrated meaningful discrimination gains over traditional severity scores on these benchmarks [8]. This performance trajectory has motivated increasing interest in multimodal IHM prediction, integrating data from two or more clinically distinct information sources that differ in provenance, data structure, and representation requirements [9, 10].

Clinical ICU data are inherently heterogeneous. Structured physiologic time-series (vital signs, laboratory results, ventilator parameters) capture the objective physiologic trajectory; clinical notes contain clinician assessments and qualitative observations; coded tabular data (diagnoses, procedures, severity indices) summarize clinical context; and in a growing minority of studies, medical imaging or high-frequency waveforms are incorporated. Each modality originates from a distinct source system, carries unique information content, and poses distinct computational challenges [11–13]. Combining these heterogeneous sources reflects the information a clinician synthesizes when assessing prognosis, and the multimodal paradigm is increasingly recognized as a natural fit for biomedical AI [11, 12].

1.2. Why "beyond discrimination metrics"

Despite sustained gains in predictive performance on benchmark datasets, the clinical adoption of machine learning mortality models remains limited [1]. A persistent gap exists between optimizing discrimination metrics (primarily AUROC and AUPRC) and delivering systems whose outputs clinicians can trust and act upon. Qualitative studies of clinician needs consistently document requirements that extend well beyond discrimination: transparent reasoning that can be audited at the bedside, probability estimates that are well calibrated, and predictable behavior when patient data are incomplete [14–16].

Calibration is the correspondence between predicted probability and observed outcome frequency, which is fundamental to any threshold-based clinical decision. Van Calster et al. characterized calibration as the "Achilles heel" of predictive analytics, documenting its systematic neglect across the prediction model literature [17]. A well-discriminating model with systematic miscalibration cannot safely support a "treat if predicted risk exceeds a specified percentage" protocol [18]. Decision curve analysis (DCA), which computes net clinical benefit across decision thresholds and compares it to treat-all and treat-none baselines, translates discrimination and calibration jointly into clinically interpretable value [18]. Despite methodological maturity and long-standing emphasis in the clinical prediction model methodology and reporting literature [18, 19], calibration reporting and DCA are frequently omitted from the machine learning literature, a gap recognized as a barrier to clinical translation [15, 17]. *Interpretability* presents a further challenge specific to multimodal systems. Unlike unimodal models, multimodal predictors can, in principle, provide not only feature-level attributions quantifying the influence of individual inputs on a prediction, but also modality-level attributions that quantify the contribution of each data source on a per-case basis [20, 21]. This second layer of transparency is clinically meaningful: it enables a clinician to understand whether a risk signal derives from physiologic deterioration in the time-series or from information in the clinical notes, and it is directly relevant to governance and accountability [15, 22]. However, many published fusion architectures combine heterogeneous sources through simple feature concatenation into a single shared model trunk, which does not inherently yield explicit per-case modality attribution [9, 10]. *Operational readiness* requires that models perform reliably outside their training data distribution. Missing data are common in real ICUs: modalities may be absent for clinical reasons (no imaging ordered, no clinical note charted within the prediction window), and feature values are irregularly sampled

with informative missingness patterns [23, 24]. External validation, testing a model on a cohort from a different time period, institution, or EHR system, is an essential but infrequently conducted step in establishing the generalizability of learned decision rules; systematic reviews of clinical prediction models have documented that external validation is reported in a minority of published studies [19, 25]. Without such evidence, high benchmark performance provides limited assurance of real-world utility. *Prediction timeliness* is also underexplored. An operational risk model must provide actionable output at a defined and clinically relevant horizon; the literature exhibits substantial heterogeneity in whether studies predict at fixed early windows, continuous horizons, or at ICU admission, limiting direct comparability across studies [26–28]. *Reproducibility* is the availability of code, data splits, preprocessing pipelines, and model specifications, which is likewise a prerequisite for independent replication and cumulative scientific progress [19].

The convergence of these dimensions (multimodal fusion architecture, feature-level and modality-level interpretability, calibration, clinical utility, external validation, missingness handling, prediction timeliness, and reproducibility) provides a practical lens for a multimodal IHM model to be rigorously evaluated for clinical deployment [14, 15, 17, 19]. Prior multimodal health AI reviews have characterized broad performance trends and algorithmic approaches across clinical domains [13], but no scoping review has systematically mapped how well the multimodal IHM literature addresses these dimensions as an integrated set.

1.3. Objectives

We map multimodal in-hospital mortality (IHM) prediction in adult ICU populations using EHR-derived data. The search was conducted in November 2025 and covers the last 15 years of research practice (search window: 2010 onward). We include studies that fuse structured ICU time-series (vitals and/or labs) and/or clinical notes with at least one additional modality. This requirement anchors the review in ICU EHR prediction settings and excludes multimodal mortality models that do not incorporate ICU physiologic data or bedside clinical assessment [5, 8, 12].

Our objectives are as follows:

1. **Catalog architectural and transparency approaches:** multimodal fusion architectures, data inputs, and reported transparency implementations, including feature-level interpretability and any modality-level attribution outputs (global or per-case contributions);
2. **Characterize evaluation practices beyond discrimination:** document the prevalence of probability calibration assessment, clinical utility evaluation (decision curve analysis and net benefit), and per-case uncertainty quantification;
3. **Chart operational readiness:** how studies handle missing features and missing modalities at inference time, address prediction timeliness (early-horizon windows or dynamic/streaming updates), evaluate generalizability through external validation and subgroup performance reporting, and document reproducibility artifacts (code, data splits, hyperparameters, and model documentation).

Altogether, we synthesize how these studies ensure transparency and reliability under real-world conditions, while demonstrating predictive performance. We also summarize author-identified gaps, limitations, and future directions to identify where current evidence is most and least ready for clinical evaluation and actionable use.

1.4. What this scoping review contributes

This review provides a field-level evidence map of 115 peer-reviewed multimodal IHM studies, representing one of the largest corpora assembled to date for the indicated aims. Unlike prior efforts focused primarily on model performance, our synthesis explicitly quantifies the prevalence of practices that bear on clinical readiness: calibration reporting, decision curve analysis, external validation, feature-level and modality-level interpretability, missing-modality policies, per-case uncertainty estimation, and reproducibility artifacts. We characterize the full landscape of fusion strategies (early, intermediate, late, and hybrid) and data provenance combinations, and provide a thematic synthesis of author-reported gaps and future directions. The resulting evidence map and gap analysis are intended to inform future research design and to motivate standardized reporting and evaluation criteria tailored to multimodal clinical prediction models in critical care, as a meaningful step toward realizing model translation to real-world clinical settings.

2. Methods

2.1. Study design and reporting standards

This scoping review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines and the Arksey & O’Malley methodological framework [29, 30]. Consistent with growing practice in evidence synthesis, this review incorporates LLM-assisted tools to support the screening and data extraction phases of the review process, with all AI-assisted outputs subject to mandatory human verification and refinement. The complete AI-assisted workflow, including model specifications and prompt templates, is described in §2.5 and §2.6, and summarized in Appendix D.

2.2. Eligibility criteria

We defined the following inclusion and exclusion eligibility criteria for the studies:

Inclusion Criteria:

- Population/setting: Adult ICU inpatients (hospital critical-care units).
- Outcome: In-hospital mortality (primary or among reported outcomes).
- Design: Empirical model development and validation (retrospective/prospective studies; RCTs if applicable).
- Data/approach: Fuses ≥ 2 data modalities by provenance and includes at least one of: (a) ICU physiologic data (vitals and/or labs) or (b) clinical notes (e.g., progress/discharge summaries).
- Reporting: ≥ 1 captured discrimination (e.g. AUPRC, AUC) metrics, calibration (e.g. ECE, Brier), decision curve/net benefit.
- Language/years: English; 2010–2025.
- Publication type: Peer-reviewed journals or full peer-reviewed conference papers.

Exclusion Criteria:

- Pediatric/neonatal only; post-discharge-only outcomes without in-hospital mortality (IHM).
- Simulation/methods-only without empirical patient data.
- Editorials/letters/protocols, theses/dissertations, preprints without peer review, reviews.
- Imaging-only without EHR clinical data for mortality modeling.

2.3. Operational taxonomies for multimodality, fusion level and algorithms

Multimodal learning utilizes computational models that integrate information from multiple “modalities”, distinct channels of information differing in generative process, statistical structure, and representation requirements [9]. In clinical AI, “multimodality” is often operationalized by heterogeneity of clinical information sources and data structures (free text, structured tabular data, indexed signals, images) rather than sensor physics alone, reflecting how clinical data are produced, stored, and consumed within health information systems [11–13].

To compare how studies combine heterogeneous ICU data sources, we pre-specified two complementary taxonomies: (i) a *provenance-based modality taxonomy* to determine whether a study is multimodal, and (ii) a *fusion-level taxonomy* to categorize where and how integration occurs within the pipeline. Complete coding and modality mappings appear in Appendix B.

2.3.1. Operational definition of modality and multimodality by data provenance

We establish multimodality by *data provenance*: a study is *multimodal* if it uses data from more than one pre-defined modality category (Appendix B). This aligns with biomedical multimodal frameworks that distinguish modalities by data structure and source system (e.g., tabular records, time-indexed signals, clinical text, medical imaging) and examine performance across these heterogeneous inputs [11–13]. A provenance-based definition preserves comparability across model families: heterogeneous sources may be fused through radically different architectures (concatenation into a single vector versus multi-branch neural encoders), yet the scientific question remains whether multiple clinically distinct information sources were used and how they were integrated [9, 11].

Our modality buckets separate ICU-relevant information streams that differ systematically in sampling regime, encoding priors, and typical storage pathways. Electronic health record (EHR) content commonly comprises intermittently measured physiologic variables and laboratory values, coded tabular descriptors (demographics, diagnoses), and narrative documentation [5]. High-frequency physiologic waveforms (e.g., ECG, arterial pressure at ≥ 100 Hz) are archived in specialized signal repositories requiring distinct processing assumptions from standard EHR time-series [31]. Medical imaging is managed through separate pipelines and storage systems (e.g., PACS), motivating distinct treatment even when downstream modeling yields embeddings [12]. This bucketization reflects clinically and technically meaningful heterogeneity beyond superficial file formats, consistent with the view that multimodal AI leverages complementary information spanning EHR, biosignals, and imaging data [12].

2.3.2. Fusion-level taxonomy by stage of integration

To characterize *how* multimodal information is combined, we apply a fusion-level taxonomy based on the integration stage. The view that multimodal fusion occurs at feature/observation, intermediate representation, or decision/prediction levels is established in multimodal information fusion and machine learning surveys [9, 10, 32], and explicitly adopted in biomedical multimodal frameworks describing common merging patterns (early/data-level versus intermediate/joint/model-level versus late/decision-level fusion) [11, 13]. We operationalized four mutually exclusive fusion categories (Appendix B): early fusion (feature-level), intermediate fusion (representation-level), late fusion (decision-level), and hybrid fusion (multi-stage). Hybrid was assigned when a single system exhibited integration at two or more stages; otherwise, a single fusion stage was assigned.

Critically, *fusion level does not determine whether a study is multimodal*. A multimodal study may use early fusion by concatenating features from distinct provenance buckets into a single model input; it remains multimodal because inputs originate from heterogeneous clinical sources [9, 13]. Separating these concepts prevents collapsing studies into “unimodal” solely because they implement single-stream architectures after preprocessing, thereby obscuring the scientific role of heterogeneous data sources.

2.3.3. Coding rules for ambiguous architectures: branches, embeddings, and trainability

Published multimodal systems frequently include preprocessing pipelines, pre-trained components, or offline feature extraction. We adopted two a priori coding rules for consistent classification across heterogeneous reporting styles. First, fusion level is determined by the *topology of the evaluated model or system*: intermediate fusion requires distinct modality-specific processing streams explicitly part of the pipeline before merging, whereas early fusion applies when the evaluated model receives a pre-merged vector/sequence processed by a single shared trunk [9, 10]. This “architectural branch” criterion directly operationalizes the stage-of-integration perspective [32]. Second, *trainability does not define fusion level*. A modality-specific stream remains a stream even if its encoder is frozen or pre-trained, because the architectural fact of separate representation learning (or provision) prior to merging distinguishes representation-level (intermediate) from feature-level (early) fusion [9, 10]. This rule is necessary in modern biomedical pipelines where pre-trained language models or vision backbones are frequently reused and optionally frozen; freezing changes optimization but not the fusion stage when a multi-stream topology is maintained. If a modeling pipeline computes text embeddings offline then trains a single classifier on concatenated embeddings with structured covariates, the evaluated model is early fusion because it treats embeddings as input features and contains no modality-specific branches [10, 32]. Conversely, if the modeling pipeline defines a text encoder stream and a structured-data stream within the modeled system and fuses their representations before prediction, it is intermediate fusion even if one stream is frozen [9, 10]. This distinction separates (i) multimodal *feature aggregation* pipelines from (ii) explicitly multimodal *representation-learning* architectures, while anchoring the multimodality decision to provenance buckets [11, 13]. Together, the provenance-based modality taxonomy and stage-based fusion taxonomy provide an auditable, theory-aligned framework for extracting comparable labels across diverse ICU mortality prediction studies.

2.3.4. Algorithms normalization

To enable consistent synthesis, reported algorithms were mapped to predefined canonical model families based on underlying methodology rather than implementation details, and usage was summarized at the study level with each study contributing at most one count per family (definitions in Appendix C).

2.4. Information sources and search strategy

Databases: PubMed, Scopus, and Google Scholar (first 5 results pages); English language; 2010–2025.

Core PubMed query:

```
("Intensive Care Units"[Mesh] OR "Critical Care"[Mesh] OR ICU[tiab] OR "intensive care"[tiab] OR "critical care"[tiab]) AND ("Hospital Mortality"[Mesh] OR "in-hospital mortality"[tiab] OR "inpatient mortality"[tiab] OR "hospital death"[tiab] OR (mortal*[tiab] AND (hospital[tiab] OR inpatient[tiab] OR ICU[tiab]))) AND (notes[tiab] OR "clinical notes"[tiab] OR "free text"[tiab] OR "discharge summary"[tiab] OR "discharge summaries"[tiab] OR "radiology report"[tiab] OR NLP[tiab] OR "natural language"[tiab] OR "natural language processing"[tiab] OR unstructured[tiab] OR "clinical text"[tiab] OR "time series"[tiab] OR "time-series"[tiab] OR vitals[tiab] OR physiolog*[tiab] OR labs[tiab] OR laboratory[tiab] OR "lab results"[tiab] OR "chart events"[tiab] OR "lab result"[tiab] OR "chart event"[tiab] OR diagnosis[tiab] OR diagnoses[tiab] OR ICD[tiab]) AND (EHR[tiab] OR "electronic health records"[tiab] OR "electronic health record"[tiab] OR "health record"[tiab] OR "health records"[tiab] OR "electronic medical records"[tiab] OR "electronic medical record"[tiab] OR "medical record"[tiab] OR "medical records"[tiab] OR EMR[tiab] OR "clinical data"[tiab] OR MIMIC[tiab] OR "MIMIC-III"[tiab] OR "MIMIC-IV"[tiab] OR "eICU"[tiab] OR "Information Mart for Intensive Care"[tiab] OR registry[tiab] OR database[tiab]) AND ("Prognosis"[Mesh] OR "Risk Assessment"[Mesh] OR "Models, Statistical"[Mesh] OR "Severity of Illness Index"[Mesh] OR APACHE[tiab] OR SAPS[tiab] OR SOFA[tiab] OR predict*[tiab] OR prognos*[tiab] OR "risk score"[tiab] OR "risk stratification"[tiab] OR "risk model"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "artificial intelligence"[tiab] OR "logistic regression"[tiab]) AND English[lang] AND ("2010/01/01"[Date - Publication] : "2025/11/30"[Date - Publication]) NOT ("Intensive Care Units, Neonatal"[Mesh] OR pediatric*[tiab] OR neonatal*[tiab] OR child[Mesh] OR infant[Mesh]) NOT animals[Mesh] NOT "Review"[Publication Type] NOT booksdocs[Filter]
```

The PubMed search strategy prioritizes ICU, hospital-anchored mortality, and EHR presence. Multimodality is enforced during screening to maximize recall without sacrificing the review's focus. Scopus and Google Scholar search strategies are adapted from the core query.

2.5. Screening and selection process

All citations retrieved from the databases were uploaded into Zotero (Zotero Foundation) for reference management and automated deduplication prior to screening. Following deduplication, records were exported, and title/abstract screening was performed using a custom-developed, LLM-assisted workflow implemented in Python, employing the same methodological approach implemented by the research team at Medical University of South Carolina and first described in the prior published systematic review [33]. Minor adaptations specific to the present review are noted in Appendix D. Full-text screening was subsequently conducted using the same LLM-assisted pipeline, with eligibility evaluated against the pre-specified criteria (§2.2).

LLM outputs were treated as an assistive, not authoritative, layer throughout. A precision-focused benchmark was conducted to assess screening accuracy: a random sample of screened records ($n = 100$) and extracted studies ($n = 30$) were independently reviewed in full by two human experts (AB, SB), who logged all discrepancies and resolved them by consensus. The benchmark confirmed acceptable precision prior to proceeding with full-corpus extraction. All final inclusion and exclusion decisions were reviewed and confirmed by human authors. Detailed prompt templates used for screenings are provided in Appendix D.

2.6. Data charting (extraction)

A standardized data extraction form was developed and iteratively refined to capture study characteristics, model design, and reporting elements relevant to interpretability, evaluation, operational readiness, and reproducibility. The following categories of information were charted during extraction:

- **Bibliographic:** Title, authors, year.
- **Dataset and cohort:** Dataset characteristics (including site, era, and EHR system), cohort size, event rate, and mortality details as reported.
- **Data inputs and modalities:** Standardized data inputs using mapping (described in Appendix B) of fused EHR data domains (e.g., Laboratory values, Continuous vital signs, Demographics, Diagnoses, Procedures, Clinical notes, Chest imaging) and modality categories (e.g., Physiologic time-series, Clinical text, Coded / Categorical / Tabular, Medical imaging pixels, High-frequency physiologic waveforms).
- **Fusion architecture and algorithms:** Classification of fusion strategy (e.g., Early, Intermediate, Late, Hybrid); details of specific technical approach for combining modalities; summary of the data flow and method; machine learning algorithms utilized (e.g., XGBoost, LSTM, Transformer).
- **Evaluation beyond discrimination:** List of reported performance metrics; presence of calibration evaluation; Decision Curve Analysis (DCA) and net benefit findings; presence of per-case uncertainty quantification.
- **Operation under real-world conditions:** External validation; evaluation of model performance across specific patient subgroups; timeliness and evaluation horizon (discrete/continuous); data missingness strategies, including handling missing features and missing modalities at inference time (policies for partial multimodal input).
- **Interpretability and attribution:** Feature-level XAI methods (e.g., SHAP, Integrated Gradients) and per-case feature-level attribution; modality-level attribution methods for quantifying the contribution of specific modalities to the model's prediction.
- **Reproducibility, documentation, and narrative synthesis:** Reproducibility indicators (code availability, pseudocode, split specification, data lineage, model cards, and hyperparameters); qualitative synthesis of reported gaps, limitations, and future directions; extracted notable quotations.

The structured extraction form was designed by human experts (AB, SB, JH) and subsequently translated into a structured LLM prompt to guide automated extraction (Appendix D). Extraction was performed using Google Gemini (Google DeepMind, Gemini Pro 3) via the Python-based pipeline. Following automated extraction, raw extracted data were cleaned with assistance from ChatGPT (OpenAI, GPT-5.1), encompassing standardization of text formatting (e.g., case and delimiter consistency), identification and removal of duplicated field entries, and logical regrouping of related information extracted from narrative passages. Extracted and cleaned data were compiled into a comprehensive tabular form of 26 columns listing all 115 included studies and their corresponding extracted fields (Appendix E). All automatically extracted values used in analyses were verified by human reviewers (AB, SB). Identified errors or ambiguities were corrected directly in the extraction table prior to synthesis.

2.7. Synthesis approach

Synthesis was primarily quantitative-descriptive: extracted fields were summarized as frequencies and proportions (n, %) for categorical elements (e.g., fusion level, calibration reporting), and distributions were described for continuous or ordinal fields. Summary statistics and figures were generated using Python. Qualitative synthesis addressed author-reported gaps, limitations, and future directions extracted verbatim from included studies, as well as other non-standardized narrative fields (e.g., fusion mechanism description, pipeline method overview). Thematic analysis was performed to identify recurring themes and map them to the review's aims, including transparency approaches, evaluation practices, and operational readiness.

3. Results

3.1. Study selection

The initial database search identified a total of 793 records (PubMed: n = 575; Scopus: n = 168; Google Scholar: n = 50). After removing duplicates (n=20), 773 records remained for title and abstract screening. Following the screening phase, 576 records were excluded as they did not meet the inclusion criteria. A total of 197 reports were sought for retrieval, of which one report could not be retrieved (n = 1). The remaining 196 reports were assessed for eligibility via full-text review. Of these, 81 studies were excluded. The primary reasons for exclusion were: ineligible

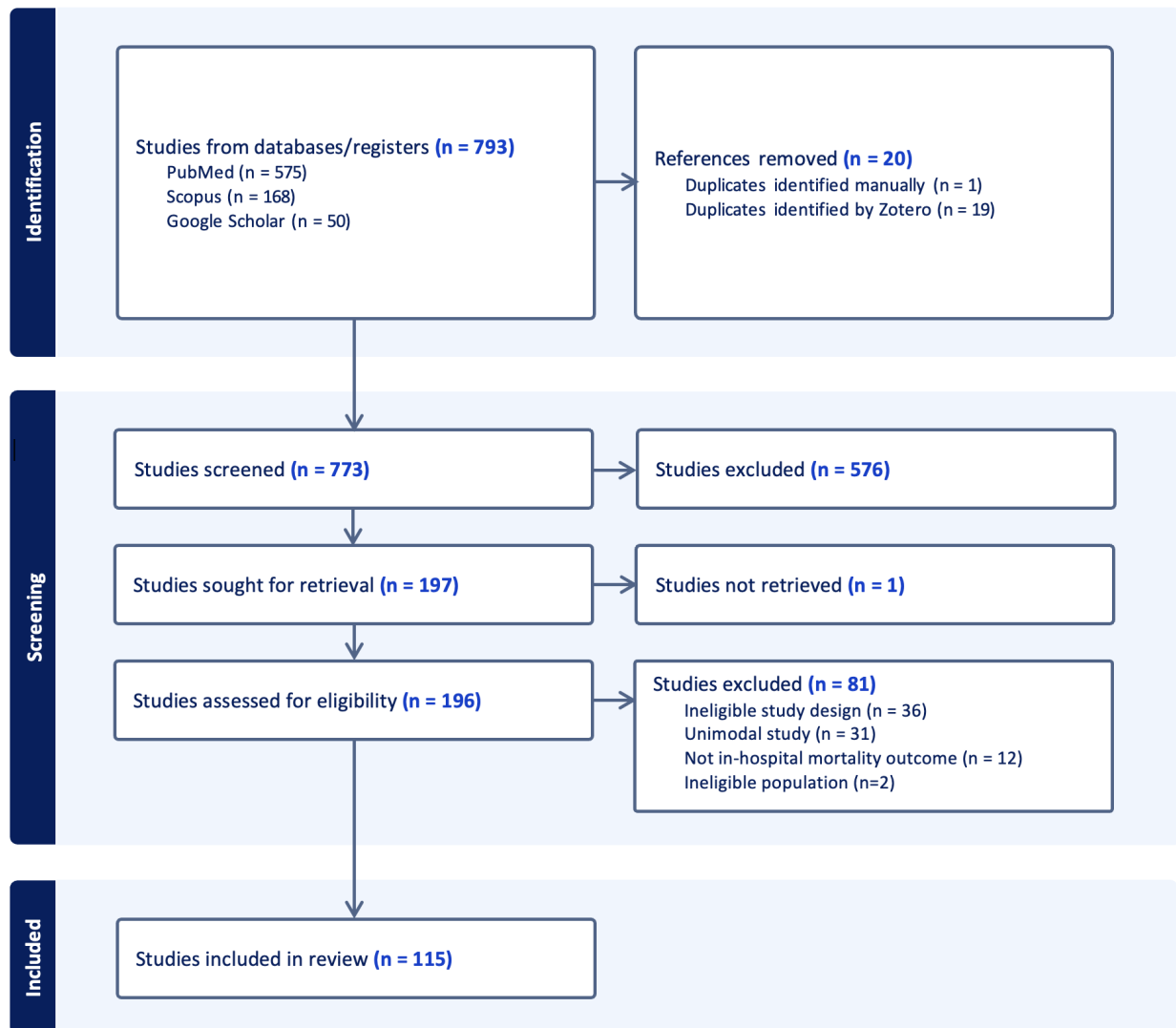


Figure 1: PRISMA flow diagram detailing the study selection process.

study design (e.g., conference abstracts, reviews, protocols, or studies lacking primary model development) ($n = 36$); use of a single information-type modality (e.g., physiologic time-series only or text only) ($n = 31$); outcome not in-hospital mortality (e.g., sepsis, discharge status) ($n = 12$); and ineligible population (e.g., non-ICU or pediatric settings) ($n = 2$). Ultimately, 115 studies were included in the final scoping review. The study selection process is illustrated in Fig. 1.

3.2. Overview of included studies

The 115 included studies span 2011–2025 (search window: 2010–2025) and collectively represent the current state of multimodal IHM prediction research in adult ICU populations, as defined by our search and taxonomy. Before detailing each dimension in turn, Figure 2 provides a field-level snapshot of the reporting practices documented across the corpus.

Across the literature, discriminative performance was the dominant evaluation target, whereas practices that more directly support clinical interpretation and deployment were inconsistently reported. As shown in Figure 2, calibration and decision curve analysis were reported in a minority of studies, per-case predictive uncertainty was rare, and

modality-level attribution was near-absent. Evidence relevant to operation under real-world conditions was likewise limited: external validation and subgroup analyses were performed in only a subset of studies, timeliness or horizon sensitivity was evaluated infrequently, and explicit missing-modality policies were uncommon. Reproducibility artifacts were uneven, with code repositories provided in a minority of studies and structured model cards absent across the corpus.

These patterns motivate the more granular synthesis that follows in §§3.3–3.9, and their implications for clinically evaluable multimodal ICU prediction are taken up in the Discussion (§4).

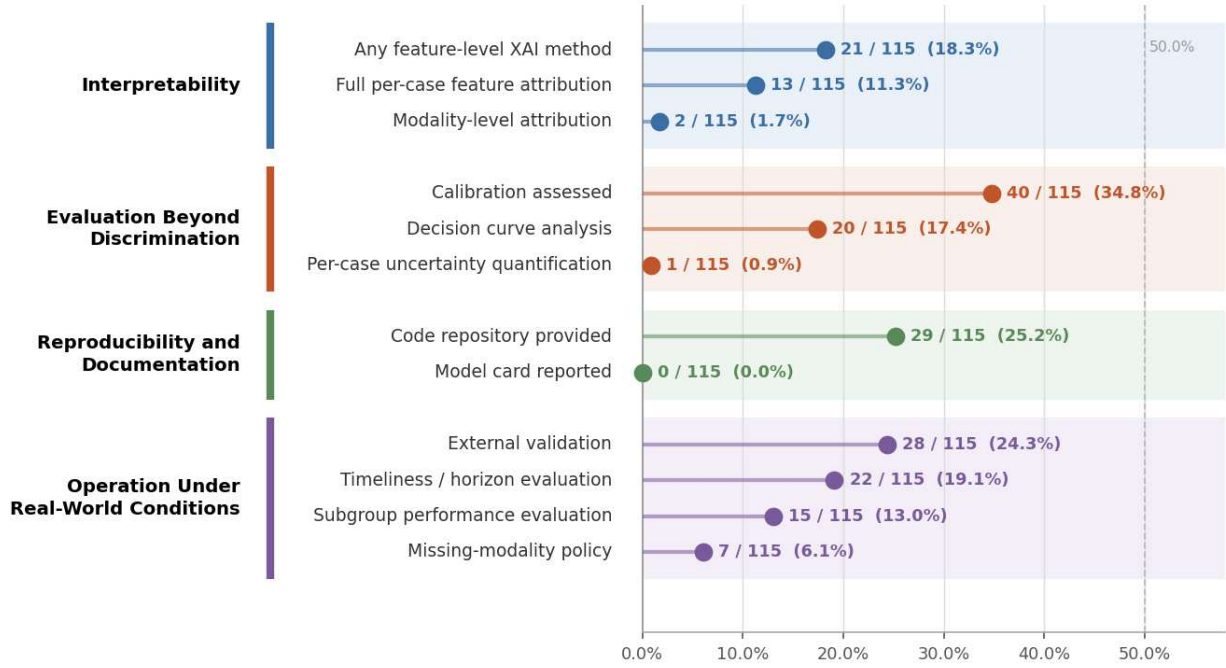


Figure 2: Prevalence of selected reporting practices across 115 included multimodal ICU in-hospital mortality prediction studies [34–148]. *Interpretability* covers feature-level explainability methods and modality-level attribution outputs. *Evaluation beyond discrimination* covers calibration assessment, decision curve analysis, and per-case predictive uncertainty quantification. *Reproducibility & documentation* covers code repository availability and structured model cards. *Operation under real-world conditions* covers external validation, subgroup performance evaluation, timeliness or prediction-horizon evaluation, and explicit missing-modality policies. All proportions are computed out of all 115 included studies.

3.3. Datasets, cohorts, and multimodal inputs

Publicly available critical care databases from the PhysioNet ecosystem were the dominant data sources for multimodal IHM prediction. The most frequently utilized logical dataset families were *MIMIC-IV* ($n = 46$, 40%) and *MIMIC-III* ($n = 44$, 38%), followed by the multi-center *eICU Collaborative Research Database* (eICU-CRD; $n = 19$, 17%). Older or specialized iterations such as *MIMIC-II* and *MIMIC-CXR* were less common ($n = 4$, 3% each). Despite the prevalence of these open-access benchmarks, a substantial portion of the literature ($n = 38$, 33%) relied on *Other* datasets, comprising unique single-center cohorts, private national registries, or competition-specific subsets. Reported event rates varied considerably depending on inclusion criteria and outcome definitions. In-hospital mortality rates typically ranged between approximately 10% and 25%, consistent with mixed adult ICU populations. Higher rates were observed in studies restricted to high-acuity subgroups, while lower rates generally corresponded to cohorts with broader inclusion criteria or shorter prediction horizons.

The landscape of multimodal inputs was dominated by structured EHR data domains. Among the 115 included studies, physiologic time-series were the most ubiquitous modality, serving as the backbone for nearly all predictive models. Specifically, *Laboratory values* ($n = 110$, 96%) and *Continuous vital signs* ($n = 93$, 81%) were the two most frequently fused data domains, distinguished by their disparate provenance (order-based LIS records vs. automated

bedside monitoring) and temporal resolution [11–13]. Other physiologic inputs included *Neurologic assessments* (e.g., GCS; $n = 29$, 25%), *Fluid intake/output* ($n = 21$, 18%), and respiratory-specific data such as *Ventilator parameters* ($n = 20$, 17%) and *Respiratory support* modes ($n = 16$, 14%). Coded and tabular data represented the second most common information source. *Demographics* were reported in the vast majority of studies ($n = 95$, 83%). Clinical context was frequently encoded via *Comorbidities* ($n = 46$, 40%), *Severity scores* (e.g., SOFA/APACHE; $n = 44$, 38%), and *Diagnoses* (ICD codes; $n = 32$, 28%). Unstructured and high-dimensional modalities appeared in a minority of studies. *Clinical notes* were the primary source of unstructured text ($n = 18$, 16%), followed by *Diagnostic reports* ($n = 7$, 6%). Medical imaging was largely restricted to *Chest imaging* ($n = 9$, 8%), with only isolated use of Cardiac or Other imaging modalities ($n = 1$, 1% each). High-frequency waveforms remained exceptionally rare, with only *Cardiac waveforms* ($n = 2$, 2%) and *Neurologic waveforms* ($n = 1$, 1%) reported. A detailed breakdown of data domains by category is provided in Table 1. Pairwise association rule mining identified distinct clustering among structured

Table 1: Frequency of fused data domains reported across included studies, grouped by data modality category (Appendix B).

Modality Category	Data Domain	Count (n)	% (of $n = 115$)
Physiologic time-series	Laboratory values	110	95.7
	Continuous vital signs	93	80.9
	Neurologic assessments	29	25.2
	Fluid intake/output	21	18.3
	Ventilator parameters	20	17.4
	Respiratory support	16	13.9
	Medication infusions	7	6.1
Coded / Tabular	Demographics	95	82.6
	Comorbidities	46	40.0
	Severity scores	44	38.3
	Diagnoses	32	27.8
	Medications	24	20.9
	Administrative	21	18.3
	Procedures	18	15.7
	Care intensity	4	3.5
Clinical text	Clinical notes	18	15.7
	Diagnostic reports	7	6.1
Medical imaging	Chest imaging	9	7.8
	Cardiac imaging	1	0.9
	Other imaging modalities	1	0.9
Waveforms	Cardiac waveforms	2	1.7
	Neurologic waveforms	1	0.9

coded domains. The strongest pairwise association was observed between *Procedures* and *Diagnoses* (Lift = 2.40), with additional strong clustering between *Medications* and *Comorbidities* (Lift = 1.56), and between *Severity scores* and *Comorbidities/Diagnoses* (Lift = 1.42/1.39). Full association-rule method parameters (minimum support 10%, minimum lift 1.1) and the complete ranked pairwise table are provided in Appendix A (Table A.6). Notably, *Clinical notes* exhibited no strong pairwise clustering (Lift < 1.1) with any specific structured domain. This independence suggests that unstructured text is predominantly utilized as a universal additive modality atop the standard physiologic baseline (Vitals/Labs), rather than co-occurring within specific niche clusters.

3.4. Fusion architectures and implementation patterns

Multimodal integration strategies were evaluated using 4 defined fusion levels [11, 13]. *Early fusion* (feature-level integration) was the predominant architectural paradigm, utilized in the majority of included studies ($n = 80$, 69.6%). This approach typically involved concatenating raw features or pre-processed vectors from different modalities into a single input vector prior to modeling. *Intermediate fusion* (representation-level integration), where modality-specific

encoders process distinct data streams before merging, was the second most common strategy ($n = 30$, 26.1%). *Late fusion* (decision-level integration), which aggregates predictions from independent unimodal models, was reported in a small minority of studies ($n = 4$, 3.5%). A single study ($n = 1$, 0.9%) employed a *Hybrid fusion* strategy, combining elements of early and intermediate integration within a hierarchical architecture.

The technical implementation of multimodal fusion was largely dictated by the chosen architectural strategy. In the predominant *early fusion* paradigm, the most common mechanism was feature concatenation, where heterogeneous inputs were aggregated into a single unified tabular array or vector prior to modeling. This approach was frequently paired with tree-based algorithms or multivariate regression; for instance, [34] aggregated real-time monitoring data with laboratory results into a unified feature set processed by a Gradient Boosting Decision Tree (GBDT). Similarly, [35] combined extracted clinical features from MIMIC-IV into a structured dataset for supervised learning, and [40] utilized LASSO regression to select features from a concatenated demographic and physiologic vector for nomogram construction. *Intermediate fusion* architectures ($n = 30$) generally leveraged joint representation learning, utilizing deep neural networks to merge embeddings from distinct modality-specific encoders before the classification layer. For example, the MedFuse architecture [53] employed separate branches, an LSTM for clinical time-series and a ResNet-34 for chest X-rays, fusing their feature vectors via a concatenation layer. More recent approaches have increasingly adopted attention mechanisms to model cross-modal dependencies; [118] utilized a time-aware self-attention mechanism to integrate structured time-series with clinical notes, while [69] employed a multi-branch transformer architecture to align real-time data streams with unstructured text. In contrast, *late fusion* strategies typically employed score-level aggregation or meta-learning. A representative example is the GRACE-ICU model [36], which processed clinical notes via a fine-tuned Clinical-Longformer to generate a risk probability; this score was subsequently combined with structured physiologic variables in a logistic regression meta-learner to produce the final prediction. Other late fusion approaches utilized ensemble voting mechanisms, such as [83], which aggregated probability outputs from multiple base classifiers via soft voting.

Beyond the topological point of fusion, the processing pipelines generally followed one of the architectural paradigms we define below. The most prevalent workflow was the pipeline, characterized by heavy reliance on manual feature engineering and selection prior to modeling. In these studies, high-dimensional data were aggregated into summary statistics (e.g., min/max/mean over 24 hours), reduced via LASSO or Recursive Feature Elimination, and processed by interpretable classifiers or nomograms. For instance, [40] and [38] employed LASSO regression to isolate independent prognostic factors from aggregated demographics and vitals, subsequently constructing logistic regression-based nomograms for clinical use. The next distinct paradigm involved multi-branch neural architectures, designed for end-to-end learning without manual feature extraction. These pipelines utilized parallel, modality-specific encoders to map raw data streams into latent space. The MedFuse architecture [53] exemplifies this approach, utilizing an LSTM branch for clinical time-series and a ResNet-34 branch for chest X-rays, projecting them into a shared sequence dimension. Similarly, [69] processed structured data and clinical notes via separate Recurrent Neural Network and BioBERT-Transformer branches, respectively, fusing the outputs only at the final fully connected layers. The third distinct approach focused on ensemble and meta-learning workflows to maximize predictive performance. Rather than relying on a single architecture, these pipelines trained multiple diverse base learners (e.g., Random Forest, LightGBM, XGBoost) and aggregated their decisions. A notable implementation by [45] stacked the outputs of three gradient-boosting models into a final logistic regression meta-classifier. More advanced iterations, such as [57], introduced self-supervised pre-training stages, aligning time-series and text representations via contrastive loss before fine-tuning the pipeline on the specific mortality prediction task.

3.4.1. Algorithms used across studies

To enable consistent synthesis across heterogeneous studies, reported predictive models were normalized into a hierarchical taxonomy. For high-level comparability with prior mortality prediction reviews, algorithms were aggregated into canonical Macro Families (Linear/Statistical, Tree-based, Neural Networks, and Other) following the classification scheme of Delpino et al. [149]. To capture methodological diversity specific to multimodal research, we retained granular Micro-level Architectural Families (e.g., differentiating Transformers, CNNs, and RNNs) as enumerated in multimodal precision health reviews [13]. Meta-learning frameworks (e.g., stacking) were categorized by their specific meta-learner algorithm rather than treated as a standalone family.

The distribution of algorithm families across the 115 included studies is summarized in Table 2. Because studies frequently evaluated multiple algorithms, Micro Family counts within a Macro Family are non-exclusive and do not

sum to the Macro total. Overall, the landscape reflects a continued reliance on interpretable baselines alongside substantial uptake of higher-capacity methods (tree ensembles and deep learning). Linear and Statistical models were the most frequently reported macro-family ($n = 71$, 61.7%), predominantly represented by logistic and regularized regression ($n = 71$, 61.7%). These were widely utilized as baseline comparators or as interpretable implementations (e.g., nomograms). Tree-based models were the second most prevalent family ($n = 54$, 47.0%). Within this category,

Table 2: Hierarchical distribution of algorithm families. Algorithms were mapped to Macro Families following Delpino et al. [149] and Micro Families following Kline et al. [13]. Percentages exceed 100% as studies frequently evaluated multiple algorithms.

Macro Family	Micro Architecture / Class	Count	% (of $n = 115$)
Linear / Statistical	<i>Total Macro Count</i>	71	61.7%
	Logistic / regularized regression	71	61.7%
Tree-based	<i>Total Macro Count</i>	54	47.0%
	Gradient-boosted trees	42	36.5%
	Random forest / ExtraTrees	42	36.5%
	Decision tree	19	16.5%
Neural Networks	<i>Total Macro Count</i>	47	40.9%
	RNN family (LSTM / GRU / variants)	29	25.2%
	MLP / feed-forward NN	20	17.4%
	Transformer / BERT / LLM family	16	13.9%
	CNN family	12	10.4%
	Neural network (unspecified)	6	5.2%
Other	<i>Total Macro Count</i>	56	48.7%
	Other (rare / unclear / paradigm-level)	38	33.0%
	Support Vector Machines	20	17.4%
	k-nearest neighbors	12	10.4%
	Naive Bayes / Bayesian classifiers	10	8.7%

ensemble methods dominated: gradient-boosted trees (e.g., XGBoost, LightGBM) and random forests were each reported in 36.5% ($n = 42$), significantly outpacing single decision trees ($n = 19$, 16.5%). Neural Networks were employed in 40.9% of studies ($n = 47$, 40.9%). Reflecting the multimodal nature of the review, architectures were diverse: Recurrent Neural Networks (RNN/LSTM/GRU), typically used for time-series data, were the most common deep learning subclass ($n = 29$, 25.2%). Transformer-based architectures (e.g., BERT, LLMs) appeared in 13.9% of studies ($n = 16$, 13.9%), surpassing Convolutional Neural Networks (CNNs) ($n = 12$, 10.4%). Other models ($n = 56$, 48.7%) largely comprised traditional machine learning baselines, including Support Vector Machines ($n = 20$, 17.4%) and k-Nearest Neighbors ($n = 12$, 10.4%), alongside niche or paradigm-level approaches categorized under Other ($n = 38$, 33.0%).

3.5. Transparency and interpretability

Feature-level interpretability methods were reported in 21 studies (18.3%). Of these, 13 studies (11.3%) implemented full per-case attribution, providing explicit feature contribution scores or visualizations for individual patient predictions. A further 8 studies (7.0%) reported partial per-case attribution, where interpretability was applied to only a subset of modalities or aggregated across samples rather than resolved to individual cases. The remaining 94 studies (81.7%) reported no feature-level interpretability.

The methods employed across the 21 studies were heterogeneous. SHAP was the most recurrent approach, used to generate force plots or waterfall plots attributing feature contributions at the individual prediction level [45, 64, 91, 108, 130]. Attention mechanisms served as the primary interpretability instrument in studies employing sequence or text encoders, with weights visualized at the token, concept, or time-step level for individual patients [48, 55, 65, 133]. Integrated Gradients were applied in studies combining text and structured modalities to attribute token-level importance for clinical notes [111, 114, 137]; one of these studies extended the approach by directly comparing machine-generated attributions against physician annotations at the token and feature level for 200 patient cases [114]. Grad-CAM was used in studies incorporating imaging to produce spatial heatmaps over chest X-ray regions [66, 88],

including one study pairing imaging-level Grad-CAM with Local-SHAP for structured variables on selected cases [66]. Nomogram-based implementations provided per-case interpretability by mapping individual feature values to weighted point scores [102, 127], and one study applied LIME-SP to derive patient-specific marker sets via local surrogate modeling [59].

Partial implementations frequently reflected deliberate or incidental modality-subset coverage. In several cases, interpretability was restricted to the NLP module while structured variables were presented only via a global nomogram [36], or Grad-CAM attribution was applied exclusively to the imaging modality without corresponding structured EHR feature attribution [88]. In others, SHAP analyses explicitly excluded temporal vital-sign features from the explainability reporting [129], or attribution results were reported as global feature rankings across the test set rather than resolved to individual cases [111, 137].

Modality-level attribution that quantifies the relative contribution of each data modality rather than individual features was rare. It was identified in only 2 studies (1.7%). One study implemented per-case modality attribution via an Uncertainty-Aware Weighting strategy, computing sample-specific modality weights from energy-based uncertainty scores derived from unimodal encoders, such that each patient’s prediction was associated with a dynamic importance weight for each modality branch (EHR time-series versus chest X-ray) [84]. The second study reported global modality attribution using Shapley values computed at the modality level to quantify the aggregate contribution of each modality to multimodal predictive performance, without resolving weights to individual cases [136].

3.6. Evaluation practices beyond discrimination

While discrimination performance was nearly universally reported via the Area Under the Receiver Operating Characteristic curve (AUROC; $n = 108$, 93.9%), performance evaluation via other metrics remained heterogeneous. Classification-based metrics were frequently utilized, including sensitivity ($n = 59$, 51.3%), F1-score ($n = 43$, 37.4%), accuracy ($n = 42$, 36.5%), precision ($n = 41$, 35.7%), and specificity ($n = 38$, 33.0%). Despite the prevalence of imbalanced classes in ICU mortality, the Area Under the Precision-Recall Curve (AUPRC) was reported in less than a third of the studies ($n = 31$, 27.0%). Other metrics utilized for granular error analysis included Negative Predictive Value (NPV; $n = 16$, 13.9%) and Cohen’s Kappa ($n = 7$, 6.1%). A small subset of studies utilized more specialized performance measures, such as the minimum of sensitivity and precision (Min(Se, P+); $n = 4$, 3.5%) and the Matthews Correlation Coefficient (MCC; $n = 2$, 1.7%). Among others, highly specialized metrics, including the Youden Index, Likelihood Ratios (NLR/PLR), and Balanced Accuracy, were reported in only a single study each ($n = 1$, 0.9%).

Evaluation practices beyond discrimination (calibration, clinical utility, and uncertainty quantification) were less consistently adopted across the included studies. Table 3 summarizes their prevalence.

Table 3: Evaluation practices beyond discrimination. Percentages are computed out of all included studies. Practices are non-exclusive except for external validation (Yes/No).

Practice	Count	% (of $n = 115$)
Any calibration assessment reported	40	34.8%
Decision curve analysis / net benefit reported	20	17.4%
Per-case uncertainty quantification reported	1	0.9%

Calibration, which assesses the agreement between predicted probabilities and observed outcome rates, was evaluated in 40 studies (34.8%). Among these, calibration plots (reliability diagrams) were the most widely used method ($n = 30$, 26.1%), followed by the Brier score ($n = 12$, 10.4%), the Hosmer–Lemeshow goodness-of-fit test ($n = 9$, 7.8%), calibration slope and intercept ($n = 5$, 4.3%), and the Expected Calibration Error ($n = 1$, 0.9%). The co-occurrence of multiple methods within individual studies accounts for the overlapping counts; no single method was consistently paired with another in a standardized reporting bundle.

Decision Curve Analysis (DCA), which evaluates clinical utility by quantifying net benefit across a range of decision thresholds, was performed in 20 studies (17.4%). In each instance, DCA was employed to compare the proposed model against conventional clinical scoring systems (most commonly SOFA, SAPS-II, APACHE, or OASIS) as well as against the default treat-all and treat-none strategies. The consistent finding across these studies was that the proposed models delivered superior or comparable net benefit over the conventional scores within practically relevant threshold ranges. Several studies extended the DCA to both training and external validation cohorts, reporting

superiority on held-out populations including eICU-CRD [95, 135]. A subset additionally reported Clinical Impact Curves alongside DCA to illustrate the alignment between the number of patients classified as high-risk and actual mortality events across the full range of decision thresholds [46, 59].

Per-case predictive uncertainty was computed in a single study ($n = 1$, 0.9%). This work proposed an Uncertainty-Aware Weighting strategy that used energy-based uncertainty scores derived from Helmholtz free energy to dynamically weight the contribution of each modality (chest X-ray and EHR time-series) on a per-sample basis during fusion, such that more uncertain modalities received reduced weight at inference [84]. No study in the corpus reported abstention policies, coverage-based uncertainty metrics, or explicit separation of aleatoric and epistemic uncertainty components.

3.7. Operational readiness: timeliness and missingness

Table 4 summarizes the prevalence of operational readiness practices across the included studies.

Table 4: Operational readiness practices. Percentages are computed out of all included studies. Within each coded dimension, categories (Yes/No/Partial) are mutually exclusive.

Practice	Count	% (of $n = 115$)
Timeliness / horizon evaluation reported	22	19.1%
Missing features addressed	77	67.0%
Missing features not addressed	37	32.2%
Missing features addressed (partial)	1	0.9%
Missing modality not addressed	108	93.9%
Missing modality policy reported (full)	6	5.2%
Missing modality policy reported (partial)	1	0.9%
External validation performed	28	24.3%
Subgroup performance evaluation reported	15	13.0%

Explicit evaluation of prediction timeliness or horizon sensitivity was reported in 22 studies (19.1%). The dominant approach was discrete horizon comparison, where models were evaluated at two or more fixed observation windows (e.g., 12h, 24h, 48h from admission) to assess how performance varied with data availability. Several studies compared daily snapshots across the first days of ICU admission, consistently finding that predictive performance improved as more longitudinal data accumulated [54, 79, 112, 113]. Others evaluated multiple fixed post-admission intervals ranging from hours to days or months, training or applying separate models for each target horizon [39, 82, 133]. A smaller subset moved beyond static snapshots toward streaming or dynamic update paradigms: models that generated predictions at regular sub-daily intervals (e.g., hourly or every 6 hours) throughout the ICU stay using a rolling or expanding observation window [41, 88, 105, 123]. The remaining 93 studies (80.9%) did not report any evaluation of prediction horizon or timeliness sensitivity.

Strategies for handling missing feature values within a given modality were reported as addressed in 77 studies (67.0%), not addressed in 37 studies (32.2%), and partially addressed in 1 study (0.9%). Reported approaches were heterogeneous and included simple imputation (e.g., mean/median/mode or clinically motivated default values), threshold-based exclusion of variables with high missingness followed by imputation for retained variables (e.g., exclusions at 25%–40% in extracted examples), and multiple imputation methods including chained-equation approaches such as MICE [38, 102, 129]. Time-series-specific strategies included carry-forward imputation and interpolation/resampling, sometimes augmented with explicit staleness/decay indicators [55, 67] or with missingness masks provided to the model [105, 126]. Some studies leveraged model-native missingness tolerance in gradient-boosted tree implementations (LightGBM/XGBoost) [34, 37]. More specialized strategies included Gaussian Process Regression for multivariate time-series interpolation [133] and GAN-based imputation via a GAIN module [146].

Explicit policies for handling the complete absence of one or more modalities at inference time were reported in only 7 studies (6.1%), while 108 studies (93.9%) did not address missing modality scenarios. Among the 7 studies that did, strategies varied in architectural formality. MedFuse processed variable-length sequences such that an absent chest X-ray reduced the fusion input to a single-element LSTM sequence rather than triggering a model failure

[53]. One study set all pixel values to zero for missing imaging modalities, preserving the fixed-input structure of the model [80]. A graph-based fusion architecture encoded modality presence or absence via binary adjacency matrices, allowing the model to adapt its aggregation to whichever modalities were available for a given patient [132]. Another architecture handled missing modalities by omitting the corresponding feature extractor and classifier branch during fusion without requiring structural modifications [136]. One study described its Transformer architecture as processing whichever modality tokens happened to be available without requiring resampling [77]. One further study described its framework as remaining applicable when some EHR modalities were lacking, though detailed masking logic for a fully missing branch was not provided [55]. One study reported that its dual-encoder pretraining design permitted independent use of each modality encoder after training [57].

External validation, defined as evaluation on a dataset drawn from a different site, time period, or patient population than the training cohort, was conducted in 28 studies (24.3%). The remaining majority ($n = 87$, 75.7%) reported performance on internal splits or held-out test sets from the same source cohort only. Subgroup analyses were reported in 15 studies (13.0%) and varied considerably in scope and motivation. Age-stratified performance was the most common form, assessing whether discrimination held across patient age bands [36, 53]. Disease-specific subgroups were also evaluated, including stratification by cancer type [38], infection source [131], and clinical diagnosis [56]. A smaller number of studies examined performance across sex [140] or race and ethnicity [37, 143], and one study conducted a geographic leave-one-country-out evaluation across nine European countries, reporting ICU mortality AUROCs ranging from 0.73 to 0.90 [123]. Subgroup analyses were predominantly discriminative, reporting AUROC or AUPRC within each stratum; none reported calibration or net benefit disaggregated by subgroup.

3.8. Reproducibility indicators

Table 5 summarizes the prevalence of reproducibility artifacts across the included studies.

Table 5: Reproducibility indicators. Percentages are computed out of all included studies. All indicators are binary (Yes/No) except where noted.

Indicator	Count	% (of $n = 115$)
Data lineage reported (dataset identified)	115	100.0%
Train/validation/test splits specified	79	68.7%
Model hyperparameters reported	59	51.3%
Code repository URL provided	29	25.2%
Pseudocode provided	8	7.0%
Model card reported	0	0.0%

Reproducibility indicators were unevenly distributed across the corpus. All 115 studies (100%) identified their data sources, meeting a basic requirement for reproducibility by enabling the underlying dataset or cohort to be located or requested. Beyond data lineage, the picture was more heterogeneous. Train, validation, and test split specifications were reported in 79 studies (68.7%), while 36 studies (31.3%) did not describe how data were partitioned, leaving exact replication undetermined. Hyperparameter configurations of the models were reported in 59 studies (51.3%) and absent in 56 studies (48.7%). Code availability was limited: only 29 studies (25.2%) provided a repository URL, while 86 studies (74.8%) released no exact implementation artifact. Pseudocode was provided in 8 studies (7.0%), offering partial algorithmic transparency in the absence of executable code. No study in the corpus reported a model card ($n = 0$, 0%).

3.9. Author-reported gaps, limitations, and future directions

Author-reported gaps, limitations, and future directions were synthesized thematically with emphasis on issues that directly affect evaluation and deployment readiness. Many studies motivated their models by pointing to practical limitations of legacy severity scores, describing them as imprecise, time-consuming to compute, and overly dependent on static admission-time snapshots [34, 40, 46]. In parallel, authors often framed timeliness as a core unmet need, arguing that static, single-timepoint prediction can miss evolving physiology and treatment response and that continuous or real-time risk updating is a necessary direction for ICU prognostic modeling [34, 41, 44, 79, 123, 131].

Beyond discrimination, a subset of studies explicitly identified incomplete evaluation as a limitation, including missing calibration assessment and the need for calibration plots, Brier score reporting, and local (re)calibration prior

to deployment [35, 96, 111, 128]. Several authors also emphasized that retrospective performance alone is insufficient for translation, calling for prospective evaluation and impact studies to assess whether model use changes clinician behavior and improves patient outcomes [36, 44, 86, 109]. Data quality constraints were frequently cited as practical barriers: missingness and sparsity were described as sources of selection bias (e.g., exclusions) or noise (e.g., simple imputation), with implications for robustness in real-world settings [41, 113, 119].

For multimodal systems specifically, some authors highlighted a structural mismatch between research assumptions and clinical reality: the assumption that all modalities are available at inference time was described as unrealistic, motivating architectures that can tolerate missing streams and scale to additional modalities while remaining stable under incomplete inputs [53, 88, 132]. Interpretability was repeatedly raised as a barrier to clinical trust, often framed as concern over “black-box” fused models; concrete future directions included more granular mechanisms (e.g., attention or instance-level analysis) and evaluating machine-generated explanations against clinician annotation [53, 108, 114, 122, 126, 133, 136]. Finally, generalizability was the most common limitation, with many studies acknowledging single-center retrospective development (frequently using MIMIC) and urging multicenter external validation, including attention to context transfer and lower-resourced settings [38, 85, 104, 109]. Demographic gaps were mentioned less often but included calls for more rigorous evaluation across non-white racial groups in fairness analyses and for more sex-disaggregated evidence [37, 140].

4. Discussion

4.1. Summary of evidence mapped to objectives

This scoping review mapped 115 peer-reviewed multimodal IHM prediction studies published between 2011 and 2025 (search window: 2010–2025) across three pre-specified objectives: (i) fusion architecture and transparency, (ii) evaluation beyond discrimination, and (iii) operational readiness. The resulting evidence base reveals a field that has achieved methodological maturity in constructing predictive pipelines but has not developed a correspondingly mature evaluation culture, specifically with respect to clinical translation readiness.

Objective 1: Fusion architectures and transparency. Early fusion, predominantly implemented as feature concatenation into a single shared model, accounts for 70.0% of the corpus. The multimodal input landscape was dominated by structured physiologic time-series, laboratory values (95.7%) and continuous vital signs (80.9%), and by coded/tabular EHR variables such as demographics (82.6%). Clinical notes (15.7%), medical imaging (e.g., chest imaging: 7.8%), and high-frequency waveforms (cardiac: 1.7%; neurologic: 0.9%) appeared in a minority of studies. This architectural dominance has a structural interpretive consequence: concatenation-based pipelines do not inherently produce per-case modality-level attribution, because the boundary between data sources is dissolved before modeling. The consequence is not merely technical; it limits the type of explanation that can be offered to a clinician reviewing a specific patient’s risk estimate. Intermediate fusion, where modality-specific encoders maintain separate representations before merging, is present in 26.0% of studies and constitutes the architecture suited to modality-level transparency. Yet even within the intermediate fusion group, explicit per-case modality attribution was reported in only 2 studies across the entire corpus (1.7%), indicating that the architectural prerequisite for such transparency is often present but not operationalized as an interpretability output. Feature-level interpretability methods of any kind were reported in only 21 studies (18.3%), with full per-case resolution in 13 (11.3%).

Together, these findings characterize a field where architecture choice is principally driven by predictive performance rather than by the transparency requirements of the intended clinical application.

Objective 2: Evaluation beyond discrimination. Calibration was assessed in 34.8% of included studies, meaning nearly two-thirds of the corpus provides no evidence that predicted probabilities correspond to observed outcome frequencies. Decision curve analysis, which translates model outputs into clinically interpretable net benefit and provides the most direct evidence of value for threshold-based use, was performed in 17.4%. Per-case uncertainty quantification was reported in a single study (0.9%). These figures are consistent and mutually reinforcing: the literature has developed a shared reporting norm around discrimination metrics and has not yet established analogous norms around reliability, utility, or uncertainty estimation.

Objective 3: Operational readiness. Missing-modality handling was identified as the largest operational gap. Only 7 studies (6.1%) specified any policy for what the model does when a modality is absent at inference time; the remaining 93.9% implicitly assume complete modality availability. In a real clinical ICU, modalities are routinely

absent for clinical or logistical reasons: imaging may not be ordered, notes may not have been charted within the prediction window, and specific assays may be unavailable. A model evaluated only under complete-modality conditions provides no evidence of behavior under the conditions it will actually encounter during deployment. Missing feature handling within modalities was addressed in 67.0% of studies, indicating that intra-modality missingness (feature-level) is broadly recognized as a technical challenge, while cross-modality missingness is largely unaddressed as an architectural concern. Timeliness evaluation (assessing whether performance holds at early or variable prediction horizons) was reported in 19.1% of studies. External validation was conducted in 24.3%, leaving three quarters of the corpus with performance evidence drawn exclusively from internal splits of the same source cohort. Subgroup evaluation was reported only in 13.0%, and in all cases restricted to discrimination metrics. On reproducibility, the corpus maintains a complete record of data sources (100.0%) and reasonable coverage of split specifications (68.7%) and hyperparameters (51.3%), but only 25.2% provide executable code and no study reported a structured model card. The absence of model cards across all 115 studies is notable, as structured documentation of model behavior, intended use, and known limitations is increasingly used as a practical artifact for clinical AI review and deployment planning.

4.2. Practical implications for “clinically evaluable” ICU mortality models

The included studies demonstrate that heterogeneous ICU data can be integrated across modalities using a wide range of fusion strategies and algorithm families, and the field’s primary emphasis on discrimination metrics reflects that predictive performance has been the dominant objective. The gaps documented here are not in methodological variety but in the evidence base required to judge whether a model is ready for clinical evaluation.

Calibration matters most directly for any protocol that assigns action at a defined probability threshold. A well-calibrated model enables a clinician or care team to reason about the practical meaning of a predicted risk value; an uncalibrated model with the same AUROC does not [17]. The current review finds that $\approx 45\%$ of the multimodal IHM literature provides no such evidence across selected studies. Given that calibration can be computed from the same held-out data used for discrimination reporting, its absence reflects a reporting norm rather than a technical barrier. Thus, the simplest near-term step for the field is to treat calibration as a required companion to discrimination (AUROC/AUPRC) in model evaluation and reporting [19].

Decision curve analysis provides a second, complementary dimension of evidence that calibration alone does not supply: it quantifies whether acting on a model’s predictions yields better patient outcomes, on average, than uniform clinical policies across practically relevant decision thresholds, and provides the most direct evidence of clinical value for any threshold-based use case [18]. The studies in this corpus that conducted DCA used it to compare model-guided thresholding with clinical scoring systems and treat-all/treat-none baselines across decision thresholds. The rarity of DCA ($\approx 17\%$), therefore, represents a missed opportunity: without such analyses, it remains difficult to translate discrimination and calibration results into evidence about net clinical benefit for plausible use cases.

Interpretability in a multimodal context carries a specific structural requirement that extends beyond what is sufficient for unimodal models. A clinician encountering a multimodal risk estimate can reasonably ask not only which features drove the prediction, but which data source was most informative for this patient on this occasion. This question is answerable only when the architecture preserves the modality boundary through to an attribution mechanism. The near-total absence of modality-level attribution in the corpus reported by 2 of 115 studies indicates that this question is rarely answerable from the explored literature, even when the model has the architectural capacity to support it. As clinical AI governance frameworks increasingly require that systems be auditable and that their behavior be explainable at the case level [15], this gap will become harder to overlook in clinical evaluation processes.

The missing-modality gap has a practical severity that may be underappreciated in retrospective research settings. Publicly available benchmark datasets yield well-curated, relatively complete records accounting for the large majority of dataset-study associations (MIMIC-III: $\approx 38\%$; MIMIC-IV: $\approx 40\%$). A model trained and evaluated on complete multimodal observations may fail silently or behave unreliably when a modality is missing in deployment. Specifying a missing-modality policy (e.g., graceful degradation to the available subset, a calibrated fallback, or explicit abstention) is a design decision that should be made before clinical evaluation begins.

Reproducibility is closely linked to all of the above: external validation, calibration replication, and missing-modality stress testing all require access to implementation details that most studies do not provide. With $\approx 75\%$ of studies releasing no code and no study providing a structured model card, the field’s cumulative evidence base cannot be readily audited, replicated, or built upon.

4.3. Research gaps and actionable future agenda

The gaps identified in this review are hierarchically structured. Some are immediately addressable within existing study designs; others require architectural changes or new evaluation paradigms.

The most immediately actionable gap is calibration and DCA reporting. Neither requires new data, new models, nor architectural change. Any study that produces predicted probabilities on a held-out set can compute and report a reliability diagram, Brier score, and decision curve. The field would benefit from adopting a minimal reporting bundle: discrimination, calibration, and net benefit, as a community norm for multimodal ICU prediction studies, consistent with evolving standards in clinical prediction model reporting [19].

The modality-level transparency gap requires architectural intentionality. Studies seeking to provide modality-level interpretability should design for it, choosing fusion architectures that preserve modality-specific representations through to an attribution layer rather than treating transparency as a post-hoc addition. Intermediate and late fusion architectures are structurally suited to this purpose; early-fusion concatenation pipelines generally are not. The choice of fusion level should therefore be informed not only by predictive performance targets but also by the transparency obligations of the intended clinical application.

Missing-modality robustness requires both evaluation and, in most cases, architectural provisions. Future studies should explicitly specify what the model does when a modality is absent and evaluate performance under realistic missing-modality scenarios rather than only under complete-data conditions. For models intended for high-acuity clinical use, a calibrated fallback that reverts to the best available subset of modalities, with quantified performance expectations for that reduced configuration, represents a minimum operational specification.

External validation is essential for establishing generalizability beyond the training institution and era. The concentration of the corpus on MIMIC-III and MIMIC-IV creates both an opportunity and a risk: benchmark standardization enables reproducible comparison, but it may mask the degree to which learned decision rules depend on EHR system-specific data structures, coding practices, and patient population characteristics. Studies that validate across time periods (e.g., CareVue versus MetaVision eras within MIMIC), institutions, or national registries provide substantially stronger evidence of deployability.

Subgroup evaluation should be extended beyond discrimination metrics. The existing subgroup analyses in this corpus evaluate AUROC within strata; none report calibration or net benefit by subgroup. Whether a model is well-calibrated for older patients, patients of specific diagnoses, or demographic subgroups with known disparities in clinical AI performance is a distinct and clinically important question that disaggregated discrimination cannot answer.

Retrospective studies are evaluation-focused rather than describing end-to-end deployment workflows; nonetheless, structured documentation can facilitate independent verification and downstream clinical evaluation. Model documentation, specifically structured model cards that summarize intended use, training data characteristics, performance across relevant subgroups, and known limitations, represents a low-burden reproducibility and governance artifact that no study in this corpus produced. As pre-trial evaluation frameworks for clinical AI systems increasingly expect such documentation [19], the field's current norm of omitting structured model documentation constitutes a gap that is largely within the control of research teams to close.

4.4. Limitations of this review

The review is restricted to English-language, peer-reviewed publications indexed in PubMed, Scopus, and Google Scholar, and the search window closes at November 2025. Studies published in other languages, disseminated as preprints only, or indexed in other databases were not captured. Multimodality was defined by data provenance: a study was included if it fused data from two or more distinct modality categories, regardless of whether the authors described their approach as "multimodal." This inclusive operationalization maximizes recall but means that some included studies employed relatively simple cross-domain feature combinations rather than architectures designed for multimodal representation learning. The AI-assisted screening and extraction workflow, while validated against a human-reviewed benchmark following the approach described in the prior systematic review [33], introduces the possibility that nuanced reporting in complex studies was extracted with less than perfect fidelity; all extracted values used in quantitative analyses were verified by human reviewers, but qualitative synthesis fields carry residual uncertainty. Next, although the search window spanned 2010–2025, the corpus is heavily concentrated in 2021–2025 ($n=98$, $\approx 85\%$), reflecting the field's rapid emergence following the widespread adoption of large-scale EHR benchmarks and deep learning architectures. Finally, year-by-year trend analysis was outside the scope of this work.

Declarations**Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

For transparency, Appendix D summarizes the AI-assisted workflow and includes prompt templates, the extraction schema, and the human verification procedure. Specific tools and model versions used were: Google Gemini (Google DeepMind, Gemini Pro 3) for LLM-assisted screening and structured data extraction; ChatGPT (OpenAI, GPT-5.1) for raw data cleaning and refinement. All AI-generated output was reviewed, revised, and verified by the human reviewers.

Appendix A. Association rule mining details

To characterize how domains are combined, we performed pairwise association rule mining on the reported data inputs using minimum support 10% and minimum lift 1.1.

Table A.6: Top pairwise associations between data domains ranked by Lift. High lift (strength of association relative to random chance) values (> 1) indicate domains that frequently appear together as a cluster.

Antecedent Domain	Consequent Domain	Support	Confidence	Lift
Procedures	Diagnoses	0.10	0.67	2.40
Diagnoses	Procedures	0.10	0.38	2.40
Medications	Comorbidities	0.13	0.63	1.56
Comorbidities	Medications	0.13	0.33	1.56
Comorbidities	Administrative	0.10	0.26	1.43
Administrative	Comorbidities	0.10	0.57	1.43
Comorbidities	Severity scores	0.22	0.54	1.42
Severity scores	Comorbidities	0.22	0.57	1.42

Appendix B. Multimodal and Fusion Taxonomy

The following taxonomy was used to categorize data provenance and fusion strategies. Multimodality was defined as the use of data from more than one of the five *Data Modality Categories*. Fusion level was determined by the topological stage of integration as defined in the *Fusion Strategy Classifications*.

Data Modality Categories (Provenance Buckets)

Data inputs were mapped to five canonical categories based on their original source system, aligning with provenance-based definitions in biomedical AI [11–13].

1. Physiologic time-series

- Continuous vitals: Standard bedside vital signs monitored continuously or intermittently
- Laboratory values: Blood/serum/urine tests ordered and resulted at intervals
- Ventilator parameters: Mechanical ventilation settings and measurements
- Respiratory support: Oxygen delivery methods and ventilation modes
- Medication infusions: Time-varying rates of continuous medication drips
- Intake/output: Fluid administration and elimination measurements
- Neurologic assessments: Consciousness scales and neurologic examination findings

2. Clinical text

- Clinical notes: Provider documentation of patient status and care plans
- Procedure notes: Documentation of interventions and operations performed
- Diagnostic reports: Interpretations of diagnostic tests and imaging studies
- Consultation notes: Specialist evaluations and recommendations

3. Coded / Categorical / Tabular

- Demographics: Patient identifying characteristics and insurance information
- Comorbidities: Pre-existing conditions and chronic disease burden indices
- Diagnoses: Coded medical conditions and diagnosis groupings
- Procedures: Coded interventions and procedure flags

- Medications: Medication lists and orders (non-time-varying)
- Severity scores: Illness severity and risk stratification indices
- Administrative: Admission/discharge information and billing data
- Care intensity: Location and level of care provided

4. High-frequency physiologic waveforms

- Cardiac waveforms: ECG and arterial pressure signals sampled at ≥ 100 Hz
- Respiratory waveforms: Breathing effort and gas exchange signals
- Plethysmography waveforms: Optical pulse and perfusion signals
- Neurologic waveforms: Brain electrical activity and intracranial pressure signals
- Hemodynamic waveforms: Central venous and pulmonary artery pressure signals

5. Medical imaging pixels

- Chest imaging: Radiographic and CT imaging of thorax
- Neuroimaging: CT and MRI of brain and spine
- Abdominal imaging: CT and ultrasound of abdomen and pelvis
- Vascular imaging: Angiographic studies of blood vessels
- Cardiac imaging: Echocardiography and functional cardiac studies
- Other imaging modalities: Nuclear medicine and specialized imaging techniques

Fusion Strategy Classifications

Fusion strategies were mutually exclusive and categorized based on the topological stage of integration [9, 10].

Fusion Category	Topological Definition [9]
Early Fusion (Feature-level)	Integration occurs at the input stage. Raw data or extracted features are concatenated into a single vector or sequence before processing by a shared model trunk.
Intermediate Fusion (Representation-level)	Integration occurs after modality-specific encoders. The architecture contains distinct processing branches for different modalities that merge into a joint representation layer.
Late Fusion (Decision-level)	Integration occurs at the prediction stage. Independent models generate separate predictions or probability scores which are aggregated (e.g., averaging, voting, meta-learning).
Hybrid Fusion	The architecture utilizes two or more of the above strategies simultaneously (e.g., Early fusion of structured data + Intermediate fusion with text).

Appendix C. Algorithm Taxonomy and Normalization Strategy

To ensure reproducibility, algorithms were categorized using a fixed dictionary mapping rules based on the reported model architecture.

Macro-Family Definitions

Macro categories follow the structural classification of *Delpino et al.* [149]:

- **Linear / Statistical:** Parametric models utilizing linear combinations of features, including logistic regression, penalized regression (LASSO/Ridge), and Cox proportional hazards.
- **Tree-based:** Non-parametric models relying on decision tree structures, including single trees, bagging ensembles (Random Forest), and boosting ensembles (GBM, XGBoost).
- **Neural Networks:** Deep learning architectures and feed-forward networks.
- **Other:** Instance-based learning (kNN), kernel methods (SVM), probabilistic classifiers (Naive Bayes), and other non-canonical approaches.

Micro-Family Mapping Rules

Within Macro families, specific architectures were isolated following *Kline et al.* [13]:

- **Neural Networks** were stratified into:
 - *Transformer / BERT / LLM family:* Models using self-attention mechanisms (e.g., BERT, GPT, Clinical-BERT).
 - *RNN family:* Sequence-based models including LSTM, GRU, and Bi-LSTM.
 - *CNN family:* Convolutional architectures (e.g., ResNet, Inception).
 - *MLP / FFNN:* Fully connected feed-forward networks.
- **Tree-based models** were stratified into:
 - *Gradient-boosted trees:* Sequential boosting algorithms (XGBoost, LightGBM, CatBoost, AdaBoost).
 - *Random Forest / ExtraTrees:* Parallel bagging ensembles.
 - *Decision Tree:* Single CART or C4.5 implementations.

Exclusion of "Ensembles" as a Family

Consistent with structural normalization, "Ensemble" was not treated as a distinct algorithm family. Heterogeneous ensembles (e.g., stacking) were categorized according to their meta-learner algorithm (e.g., a stacked model with a logistic regression meta-learner was counted under *Linear / Statistical*).

Appendix D. LLM-assisted screening and extraction workflow

We implemented an LLM-assisted pipeline to increase throughput and standardization during screening and data charting, following the same general approach described in the prior review automation work [33], with adaptations for the present multimodal ICU mortality review. Records were managed in Zotero for reference management and deduplication. PDFs were retrieved using Zotero's full-text retrieval when available and supplemented by manual retrieval for any missing articles. Titles/abstracts and full texts were processed by custom Python scripts, which interfaced with a Gemini model to apply predefined inclusion/exclusion criteria and to populate a structured extraction form. Although the workflow is model- and language-agnostic in design, this review included English-language studies only, consistent with the eligibility criteria.

The workflow followed the same staged structure for title/abstract screening, full-text screening, and extraction:

1. Calibration and prompt refinement. A small representative sample was reviewed by human experts to ensure consistent application of eligibility criteria and to refine the extraction schema and prompts used by the LLM.
2. Automated LLM processing with redundancy. For each record, the pipeline issued repeated independent LLM calls using the finalized prompt (three runs per record) and retained the majority decision for screening tasks. For extraction, the LLM produced structured outputs aligned to the extraction schema.

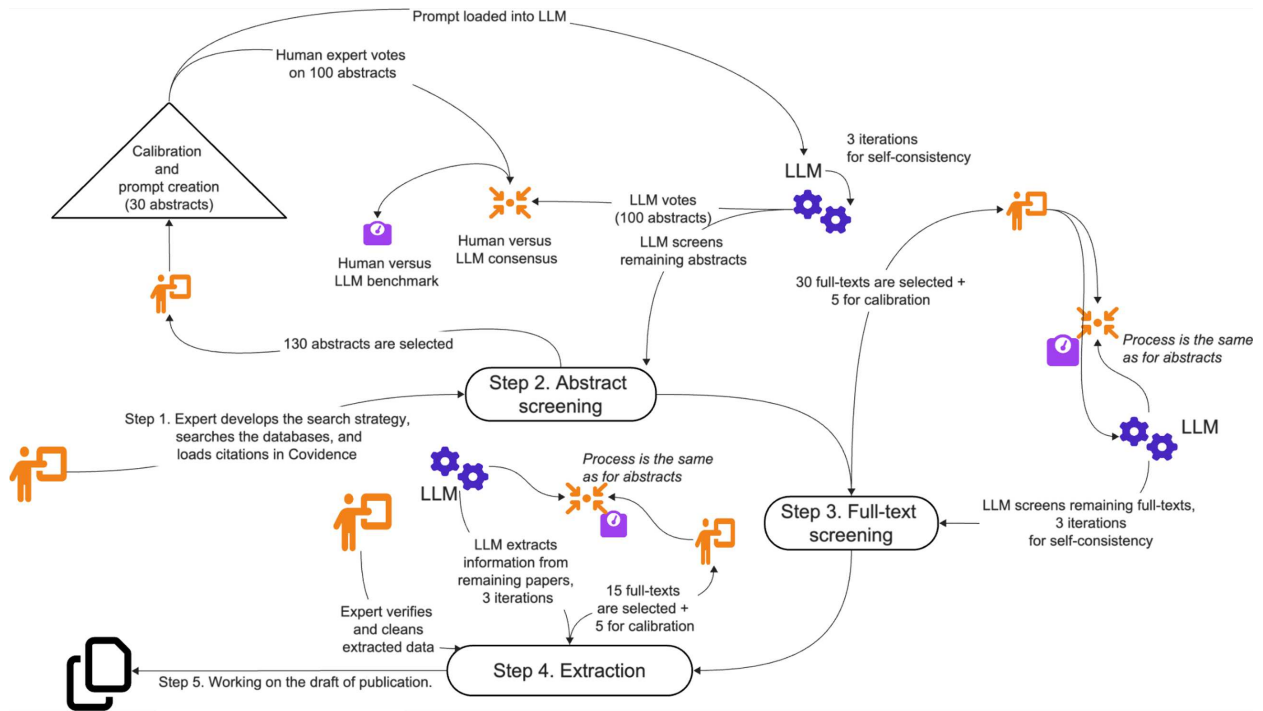


Figure D.3: AI-assisted workflow used to support screening and data charting in this scoping review (adopted from Scherbakov et al. [33]). The large language model (LLM) is used as an assistive layer for (i) title/abstract screening, (ii) full-text screening, and (iii) structured data extraction. At each stage, LLM outputs are subject to mandatory human verification and adjudication, and the final inclusion/exclusion decisions and extracted values used in analyses are confirmed by human reviewers.

- Human verification and benchmarking. LLM outputs were treated as assistive rather than authoritative. Final inclusion/exclusion decisions were confirmed by human reviewers. A precision-focused benchmark was conducted by independently reviewing a random sample of screened records ($n=100$) and thoroughly verifying extracted values for a subset of included studies ($n=30$); discrepancies were logged and resolved by consensus prior to proceeding with full-corpus extraction.

This workflow was used to support screening and structured data charting. Prompt templates used in each phase are provided in the Appendix for auditability.

Prompt templates used in this review

Screening 1 prompt (title + abstract):

```
You are screening TITLE + ABSTRACT for a scoping review on multimodal ICU
in-hospital mortality (IHM) prediction.

Use TAXONOMY_DEFINITIONS as the authoritative codebook for modality-by-
provenance definitions used in eligibility decisions.

Eligibility criteria

Inclusion (all required):
1. Adult ICU inpatient population (hospital critical-care units).
2. Outcome includes in-hospital mortality (primary or among reported outcomes).
3. Empirical model development and/or validation study.
4. Uses >=2 data modalities by provenance and includes at least one ICU-anchor
modality: (a) physiologic ICU data (vitals and/or labs) OR (b) clinical notes.
5. Reports at least one evaluation metric relevant to model performance
```

(e.g., discrimination, calibration, and/or decision-curve/net-benefit analysis).

- English-language publication (2010-2025).
- Peer-reviewed journal article or full peer-reviewed conference paper.

Exclusion (exclude if any true):

- Pediatric/neonatal-only cohort.
- Post-discharge-only outcomes without in-hospital mortality.
- Simulation/method-only paper without empirical patient data.
- Editorial/letter/protocol/thesis/dissertation/preprint-only/review article.
- Imaging-only modeling without EHR clinical data.

Decision rule:

- Include: all inclusion = Yes and all exclusion = No.
- Exclude: any inclusion = No OR any exclusion = Yes.
- Unclear: otherwise (if there is not enough information to determine inclusion or exclusion).

Instructions:

- Internally evaluate each criterion as Yes, No, or Unclear.
- Apply TAXONOMY_DEFINITIONS when evaluating multimodality/provenance terms.
- Apply the decision rule to produce one overall per-paper final decision.
- Output only the JSON object below (no explanatory prose, headings, or markdown).

Output format (JSON only; no extra text):

```
{
  "Final Decision": "Include|Exclude|Unclear",
  "Justification": "Concise decision rationale from title/abstract evidence."
}
```

Article

Title: {TITLE}

Abstract: {ABSTRACT}

Taxonomy definitions: {TAXONOMY_DEFINITIONS}

Screening 2 prompt (full text):

You are screening a full article text for a scoping review on multimodal ICU in-hospital mortality (IHM) prediction.

Use TAXONOMY_DEFINITIONS as the authoritative codebook for modality-by-provenance definitions used in eligibility decisions.

Eligibility criteria

Inclusion (all required):

- Adult ICU inpatient population (hospital critical-care units).
- Outcome includes in-hospital mortality (primary or among reported outcomes).
- Empirical model development and/or validation study.
- Uses ≥ 2 data modalities by provenance and includes at least one ICU-anchor modality: (a) physiologic ICU data (vitals and/or labs) OR (b) clinical notes.
- Reports at least one evaluation metric relevant to model performance (e.g., discrimination, calibration, and/or decision-curve/net-benefit analysis).
- English-language publication (2010-2025).
- Peer-reviewed journal article or full peer-reviewed conference paper.

Exclusion (exclude if any true):

- Pediatric/neonatal-only cohort.
- Post-discharge-only outcomes without in-hospital mortality.
- Simulation/method-only paper without empirical patient data.
- Editorial/letter/protocol/thesis/dissertation/preprint-only/review article.
- Imaging-only modeling without EHR clinical data.

Decision rule:

- Include: all inclusion = Yes and all exclusion = No.
- Exclude: any inclusion = No OR any exclusion = Yes.
- Unclear: otherwise (if there is not enough information to determine inclusion or exclusion).

Instructions:

1. Internally evaluate each criterion as Yes, No, or Unclear.
2. Apply TAXONOMY_DEFINITIONS when evaluating multimodality/provenance terms.
3. Apply the decision rule to produce one overall per-paper final decision.
4. Output only the JSON object below (no explanatory prose, headings, or markdown).

Output format (JSON only; no extra text):

```
{
  "Final Decision": "Include|Exclude|Unclear",
  "Justification": "Concise decision rationale from title/abstract evidence."
}
```

Article

Title: {TITLE}

Abstract: {ABSTRACT}

Full text: {FULL_TEXT}

Taxonomy definitions: {TAXONOMY_DEFINITIONS}

Data extraction/charting prompt:

You are extracting data for one included study in a scoping review on multimodal ICU in-hospital mortality prediction.

Instructions:

- Extract only what is reported in the article.
- Preserve study wording for narrative fields.
- Use TAXONOMY_DEFINITIONS as the coding dictionary for: "Fused Data Domains mapped", "Multimodal Fusion Level", and "Algorithms Micro Families".
- If a field is not reported, return "None".
- Return exactly one JSON object and no additional text.

Output schema:

```
{
  "Title": "",
  "Authors": "",
  "Year": "",
  "Reported Dataset Details": "",
  "Reported Cohort Details": "",
  "Fused Data Domains mapped": [],
  "Multimodal Fusion Level": "",
  "Fusion Description": "",
  "Pipeline Description": "",
  "Algorithms Micro Families": [],
  "Evaluation Metrics": [],
  "Calibration Evaluation": {
    "Brier_score_YN": "Y|N",
    "ECE_YN": "Y|N",
    "Reliability_diagram_YN": "Y|N",
    "Slope/intercept_YN": "Y|N",
    "Other": "None or concise text"
  },
  "Decision Curve Analysis": "None or concise text",
  "External Validation (Yes/No)": "Yes|No",
  "Timeliness Evaluation": "None or concise text",
  "Subgroups Evaluation": "None or concise text",
  "Feature-Level XAI": "None or concise text",
  "Modality-Level Attribution": "None or concise text",
  "Missing Features Strategy": "None or concise text",
}
```

```

"Missing Modality Policy": "None or concise text",
"Reported Gaps": "None or concise text",
"Reported Limitations": "None or concise text",
"Reported Future Directions": "None or concise text",
"Reproducibility": {
  "pseudocode_YN": "Y|N",
  "splits_specified_YN": "Y|N",
  "hyperparameters_YN": "Y|N",
  "model_card_YN": "Y|N",
  "data_lineage_YN": "Y|N",
  "code_YN": "Y|N"
},
"Notable Quotations": "None or one short verbatim quote",
"Per-case Uncertainty": "None or concise text"
}

Article
Title: {TITLE}
Abstract: {ABSTRACT}
Full text: {FULL_TEXT}
Taxonomy definitions: {TAXONOMY_DEFINITIONS}

```

Appendix E. Extraction table for the selected studies is available in the supplementary files

References

- [1] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, *Nature Medicine* 25 (1) (2019) 44–56. doi:10.1038/s41591-018-0300-7.
- [2] W. A. Knaus, E. A. Draper, D. P. Wagner, J. E. Zimmerman, APACHE II: a severity of disease classification system, *Critical Care Medicine* 13 (10) (1985) 818–829. doi:10.1097/00003246-198510000-00009.
- [3] J.-R. Le Gall, S. Lemeshow, F. Saulnier, A simplified acute physiology score (SAPS II) based on a European/North American multicenter study, *JAMA* 270 (24) (1993) 2957–2963. doi:10.1001/jama.1993.03510240069035.
- [4] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. Suter, L. G. Thijs, The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix), *Intensive care medicine* 22 (7) (1996) 707–710. doi:10.1007/BF01709751.
- [5] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, *Scientific data* 3 (1) (2016) 1–9. doi:10.1038/sdata.2016.35.
- [6] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al., Mimic-iv, a freely accessible electronic health record dataset, *Scientific data* 10 (1) (2023) 1. doi:10.1038/s41597-022-01899-x.
- [7] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, O. Badawi, The eICU collaborative research database, a freely available multi-center database for critical care research, *Scientific Data* 5 (2018) 180178. doi:10.1038/sdata.2018.178.
- [8] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, *Scientific Data* 6 (1) (2019) 96. doi:10.1038/s41597-019-0103-9.

- [9] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2) (2018) 423–443. doi:10.1109/TPAMI.2018.2798607.
- [10] D. Ramachandram, G. W. Taylor, Deep multimodal learning: A survey on recent advances and trends, *IEEE signal processing magazine* 34 (6) (2017) 96–108. doi:10.1109/MSP.2017.2738401.
- [11] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussieux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, D. Bertsimas, Integrated multimodal artificial intelligence framework for healthcare applications, *NPJ digital medicine* 5 (1) (2022) 149. doi:10.1038/s41746-022-00689-4.
- [12] J. N. Acosta, G. J. Falcone, P. Rajpurkar, E. J. Topol, Multimodal biomedical AI, *Nature Medicine* 28 (9) (2022) 1773–1784. doi:10.1038/s41591-022-01981-2.
- [13] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, Y. Luo, Multimodal machine learning in precision health: A scoping review, *npj Digital Medicine* 5 (2022) 171. doi:10.1038/s41746-022-00712-8.
- [14] S. Tonekaboni, S. Joshi, M. D. McCradden, A. Goldenberg, What clinicians want: contextualizing explainable machine learning for clinical end use, in: *Machine learning for healthcare conference*, PMLR, 2019, pp. 359–380.
- [15] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, J. H. Holmes, A manifesto on explainability for artificial intelligence in medicine, *Artificial Intelligence in Medicine* 133 (2022) 102423. doi:10.1016/j.artmed.2022.102423.
- [16] E. Begoli, T. Bhattacharya, D. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, *Nature Machine Intelligence* 1 (1) (2019) 20–23. doi:10.1038/s42256-018-0004-1.
- [17] B. Van Calster, D. J. McLernon, M. Van Smeden, L. Wynants, E. W. Steyerberg, Calibration: the Achilles heel of predictive analytics, *BMC Medicine* 17 (1) (2019) 230. doi:10.1186/s12916-019-1466-7.
- [18] A. J. Vickers, E. B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, *Medical Decision Making* 26 (6) (2006) 565–574. doi:10.1177/0272989X06295361.
- [19] G. S. Collins, J. B. Reitsma, D. G. Altman, K. G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement, *Journal of British Surgery* 102 (3) (2015) 148–158. doi:10.1136/bmj.g7594.
- [20] L. Wenderoth, K. Hemker, N. Simidjievski, M. Jamnik, Measuring cross-modal interactions in multimodal models, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, 2025, pp. 21501–21509.
- [21] J. Park, K. Park, D. Kim, J. Park, J. W. Choi, Layer-wise modality decomposition for interpretable multimodal sensor fusion, *arXiv preprint arXiv:2511.00859* (2025).
- [22] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* 1 (5) (2019) 206–215. doi:10.1038/s42256-019-0048-x.
- [23] R. H. Groenwold, Informative missingness in electronic health record systems: the curse of knowing, *Diagnostic and prognostic research* 4 (1) (2020) 8. doi:10.1186/s41512-020-00077-0.
- [24] A. Sharafoddini, J. A. Dubin, D. M. Maslove, J. Lee, A new insight into missing data in intensive care unit patient profiles: observational study, *JMIR medical informatics* 7 (1) (2019) e11605. doi:10.2196/11605.
- [25] G. S. Collins, J. A. de Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L.-M. Yu, K. G. Moons, et al., External validation of multivariable prediction models: a systematic review of methodological conduct and reporting, *BMC Medical Research Methodology* 14 (1) (2014) 40. doi:10.1186/1471-2288-14-40.

- [26] B. Xue, N. Shah, H. Yang, T. Kannampallil, P. R. O. Payne, C. Lu, A. S. Said, Multi-horizon predictive models for guiding extracorporeal resource allocation in critically ill covid-19 patients, *Journal of the American Medical Informatics Association* 30 (4) (2023) 656–667.
- [27] B. Wellner, J. Grand, E. Canzone, M. Coarr, P. W. Brady, J. Simmons, E. Kirkendall, N. Dean, M. Kleinman, P. Sylvester, Predicting unplanned transfers to the intensive care unit: a machine learning approach leveraging diverse clinical elements, *JMIR medical informatics* 5 (4) (2017) e45.
- [28] K. Stone, R. Zwigelaar, P. Jones, N. Mac Parthaláin, A systematic review of the prediction of hospital length of stay: Towards a unified framework, *PLOS digital health* 1 (4) (2022) e0000017.
- [29] H. Arksey, L. O'malley, Scoping studies: towards a methodological framework, *International journal of social research methodology* 8 (1) (2005) 19–32.
- [30] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, et al., Prisma extension for scoping reviews (prisma-scr): checklist and explanation, *Annals of internal medicine* 169 (7) (2018) 467–473.
- [31] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals, *Circulation* 101 (23) (2000) e215–e220. doi:10.1161/01.CIR.101.23.e215.
- [32] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, *Multimedia Systems* 16 (6) (2010) 345–379. doi:10.1007/s00530-010-0182-0.
- [33] D. Scherbakov, N. Hubig, V. Jansari, A. Bakumenko, L. A. Lenert, The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review, *Journal of the American Medical Informatics Association* 32 (6) (2025) 1071–1086.
- [34] H. Jiang, L. Su, H. Wang, D. Li, C. Zhao, N. Hong, Y. Long, W. Zhu, Noninvasive real-time mortality prediction in intensive care units based on gradient boosting method: Model development and validation study, *JMIR Medical Informatics* 9 (3) (2021) e23888. doi:10.2196/23888.
- [35] D. Q. Palmer, R. A. Gismondi, P. Gemal, G. S. Lomba, F. L. Seixas, Mortality prediction in heart failure patients: Machine learning versus get with the guidelines-heart failure (gwtg-hf) and acute decompensated heart failure national registry (adhere), *Cureus* 17 (8) (2025) e89226. doi:10.7759/cureus.89226.
- [36] X. Liu, W. Yeung, Z. Chen, S. Hao, Z. Yang, X. Sun, C. Liu, Z. Mao, M. Yan, W. Yan, et al., Grace-icu: A multimodal nomogram-based approach for illness severity assessment of older adults in the icu, *NPJ Digital Medicine* 8 (1) (2025) 519. doi:10.1038/s41746-025-01875-w.
- [37] J. A. Cummins, B. S. Gerber, M. I. Fukunaga, N. Henninger, C. I. Kiefe, F. Liu, In-hospital mortality prediction among intensive care unit patients with acute ischemic stroke: A machine learning approach, *Health Data Science* 5 (2025) 0179. doi:10.34133/hds.0179.
- [38] M. Wu, H. Gao, A prediction model for in-hospital mortality in intensive care unit patients with metastatic cancer, *Frontiers in Surgery* 10 (2023) 992936. doi:10.3389/fsurg.2023.992936.
- [39] H. Huang, Y. Liu, M. Wu, Y. Gao, X. Yu, Development and validation of a risk stratification model for predicting the mortality of acute kidney injury in critical care patients, *Annals of Translational Medicine* 9 (4) (2021) 323. doi:10.21037/atm-20-5723.
- [40] J. Xu, J. Weng, J. Yang, X. Shi, R. Hou, X. Zhou, Z. Zhou, Z. Wang, C. Chen, Development and validation of a nomogram to predict the mortality risk in elderly patients with arf, *PeerJ* 9 (2021) e11016. doi:10.7717/peerj.11016.

- [41] J. Ma, D. K. K. Lee, M. E. Perkins, M. A. Pisani, E. Pinker, Using the shapes of clinical data trajectories to predict mortality in icus, *Critical Care Explorations* 1 (7) (2019) e0010. doi:10.1097/CCE.000000000000010.
- [42] Y. Zhu, C. Ren, Z. Wang, X. Zheng, S. Xie, J. Feng, X. Zhu, Z. Li, L. Ma, C. Pan, Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation, in: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*, 2024, pp. 3549–3559. doi:10.1145/3627673.3679582.
- [43] E. C. Silveira, S. M. Pretti, B. A. Santos, C. F. S. Corrêa, L. M. Silva, F. F. de Melo, Prediction of hospital mortality in intensive care unit patients from clinical and laboratory data: A machine learning approach, *World journal of critical care medicine* 11 (5) (2022) 317. doi:10.5492/wjccm.v11.i5.317.
- [44] A. Awad, M. Bader-El-Den, J. McNicholas, J. Briggs, Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach, *International journal of medical informatics* 108 (2017) 185–195. doi:10.1016/j.ijmedinf.2017.10.002.
- [45] T. Huang, D. Le, L. Yuan, S. Xu, X. Peng, Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit, *Plos one* 18 (1) (2023) e0280606. doi:10.1371/journal.pone.0280606.
- [46] L. Gao, Y. Bian, S. Cao, W. Sang, Q. Zhang, Q. Yuan, F. Xu, Y. Chen, Development and validation of a simple-to-use nomogram for predicting in-hospital mortality in patients with acute heart failure undergoing continuous renal replacement therapy, *Frontiers in Medicine* 8 (2021) 678252. doi:10.3389/fmed.2021.678252.
- [47] X.-z. Xing, H.-j. Wang, C.-l. Huang, Q.-h. Yang, S.-n. Qu, H. Zhang, H. Wang, Y. Gao, Q.-l. Xiao, K.-l. Sun, Prognosis of patients with shock receiving vasopressors, *World journal of emergency medicine* 4 (1) (2013) 59. doi:10.5847/wjem.j.issn.1920-8642.2013.01.011.
- [48] G. Harerimana, J. W. Kim, B. Jang, A deep attention model to forecast the length of stay and the in-hospital mortality right on admission from icd codes and demographic data, *Journal of biomedical informatics* 118 (2021) 103778. doi:10.1016/j.jbi.2021.103778.
- [49] S. Patel, G. Singh, S. Zarbiv, K. Ghiassi, J.-S. Rachoin, Mortality prediction using sao2/fio2 ratio based on eicu database analysis, *Critical care research and practice* 2021 (1) (2021) 6672603. doi:10.1155/2021/6672603.
- [50] B. Yang, S. Xu, D. Wang, Y. Chen, Z. Zhou, C. Shen, Acei/arb medication during icu stay decrease all-cause in-hospital mortality in critically ill patients with hypertension: A retrospective cohort study based on machine learning, *Frontiers in Cardiovascular Medicine* 8 (2022) 787740. doi:10.3389/fcvm.2021.787740.
- [51] J. Y. Chen, G. Y. Jin, L. H. Zeng, B. Q. Ma, H. Chen, N. Y. Gu, K. Qiu, F. Tian, L. Pan, W. Hu, D. C. Liang, The establishment and validation of a prediction model for traumatic intracranial injury patients: a reliable nomogram, *Frontiers in Neurology* 14 (2023) 1165020. doi:10.3389/fneur.2023.1165020.
- [52] K. Pang, L. Li, W. Ouyang, X. Liu, Y. Tang, Establishment of icu mortality risk prediction models with machine learning algorithm using mimic-iv database, *Diagnostics* 12 (5) (2022) 1068. doi:10.3390/diagnostics12051068.
- [53] N. Hayat, K. J. Geras, F. E. Shamout, Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images, in: *Proceedings of the 7th Machine Learning for Healthcare Conference*, Vol. 182, PMLR, 2022, pp. 714–738.
- [54] L. A. G. Celi, R. J. Tang, M. C. Villarroel, G. A. Davidzon, W. T. Lester, H. C. Chueh, A clinical database-driven approach to decision support: Predicting mortality among patients with acute kidney injury, *Journal of healthcare engineering* 2 (1) (2011) 97–110. doi:10.1260/2040-2295.2.1.97.

- [55] C. Wang, X. Yang, M. Sun, Y. Gu, J. Niu, W. Zhang, Multimodal fusion network for icu patient outcome prediction, *Neural Networks* 180 (2024) 106672. doi:10.1016/j.neunet.2024.106672.
- [56] L. Huang, Z. Dou, F. Fang, B. Zhou, P. Zhang, R. Jiang, Prediction of mortality in intensive care unit with short-term heart rate variability: Machine learning-based analysis of the mimic-iii database, *Computers in Biology and Medicine* 186 (2025) 109635. doi:10.1016/j.combiomed.2024.109635.
- [57] R. King, T. Yang, B. J. Mortazavi, Multimodal pretraining of medical time series and notes (2023) 244–255.
- [58] R. M. L. Roepke, H. B. L. Janzantti, M. B. Cantamessa, L. F. Machado, G. D. Luckemeyer, J. V. Gandolfi, B. A. M. P. Besen, S. M. Lobo, Predictive performance of saps-3, sofa score, and procalcitonin for hospital mortality in covid-19 viral sepsis: A cohort study, *Life* 15 (8) (2025) 1161. doi:10.3390/life15081161.
- [59] E. Jamshidi, A. Asgary, N. Tavakoli, A. Zali, S. Setareh, H. Esmaily, S. H. Jamaldini, A. Daaee, A. Babajani, M. A. Sendani Kashi, et al., Using machine learning to predict mortality for covid-19 patients on day 0 in the icu, *Frontiers in digital health* 3 (2022) 681608. doi:10.3389/fdgth.2021.681608.
- [60] M. Lin, S. Wang, Y. Ding, L. Zhao, F. Wang, Y. Peng, An empirical study of using radiology reports and images to improve icu-mortality prediction, in: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), IEEE, 2021, pp. 497–498. doi:10.1109/ichi52183.2021.00088.
- [61] E. Papiol, R. Ferrer, J. C. Ruiz-Rodríguez, E. Díaz, R. Zaragoza, M. Borges-Sa, J. Berrueta, J. Gómez, M. Bodí, S. Sancho, et al., Machine learning-based identification of risk factors for icu mortality in 8902 critically ill patients with pandemic viral infection, *Journal of Clinical Medicine* 14 (15) (2025) 5383. doi:10.3390/jcm14155383.
- [62] H. Xie, Q. Luo, T. Huang, Usefulness of blood–urea–nitrogen to serum albumin ratio for in-hospital mortality predictions in atrial fibrillation patients admitted to the intensive care unit: A retrospective analysis from mimic-iv database, *Reviews in Cardiovascular Medicine* 26 (7) (2025) 36596. doi:10.31083/RCM36596.
- [63] Y. Hu, Q. Cao, H. Wang, Y. Yang, Y. Xiong, X. Li, Q. Zhou, Prognostic nutritional index predicts acute kidney injury and mortality of patients in the coronary care unit, *Experimental and Therapeutic Medicine* 21 (2) (2021) 123. doi:10.3892/etm.2020.9555.
- [64] J. A. Neyra, V. Ortiz-Soriano, L. J. Liu, T. D. Smith, X. Li, D. Xie, B. Adams-Huet, O. W. Moe, R. D. Toto, J. Chen, Prediction of mortality and major adverse kidney events in critically ill patients with acute kidney injury, *American Journal of Kidney Diseases* 81 (1) (2023) 36–47. doi:10.1053/j.ajkd.2022.06.004.
- [65] Y. Lin, L. Yu, Y. Fu, X. Zheng, A predictive model for clinical health risk using multimodal electronic health record data, in: Proceedings of the 2023 7th International Conference on Medical and Health Informatics, 2023, pp. 85–90. doi:10.1145/3608298.3608316.
- [66] J. Lin, J. Yang, M. Yin, Y. Tang, L. Chen, C. Xu, S. Zhu, J. Gao, L. Liu, X. Liu, et al., Development and validation of multimodal models to predict the 30-day mortality of icu patients based on clinical parameters and chest x-rays, *Journal of Imaging Informatics in Medicine* 37 (4) (2024) 1312–1322. doi:10.1007/s10278-024-01066-1.
- [67] Y. Wang, W. Li, Integrating multimodal ehr data for mortality prediction in icu sepsis patients, *Statistics in Medicine* 44 (10-12) (2025) e70060. doi:10.1002/sim.70060.
- [68] D. Li, W. Xing, J. Zhao, C. Shi, F. Wang, Multimodal deep learning for predicting in-hospital mortality in heart failure patients using longitudinal chest x-rays and electronic health records, *The International Journal of Cardiovascular Imaging* 41 (3) (2025) 427–440. doi:10.1007/s10554-025-03322-z.
- [69] H. Saleh, M. McCann, S. El-Sappagh, J. G. Breslin, Transformerfusionnet: a real-time multimodal framework for icu heart failure mortality prediction using big data streaming, in: 2024 International Conference on Computer and Applications (ICCA), IEEE, 2024, pp. 1–6. doi:10.1109/ICCA62237.2024.10927742.

- [70] W. Yang, H. Zou, M. Wang, Q. Zhang, S. Li, H. Liang, Mortality prediction among icu inpatients based on mimic-iii database results from the conditional medical generative adversarial network, *Heliyon* 9 (2) (2023). doi:10.1016/j.heliyon.2023.e13200.
- [71] H. Yang, L. Kuang, F. Xia, Multimodal temporal-clinical note network for mortality prediction, *Journal of Biomedical Semantics* 12 (1) (2021) 3. doi:10.1186/s13326-021-00235-3.
- [72] H. Zhu, J. Bai, N. Li, X. Li, D. Liu, D. L. Buckeridge, Y. Li, Fedweight: mitigating covariate shift of federated learning on electronic health records data through patients re-weighting, *npj Digital Medicine* 8 (1) (2025) 286. doi:10.1038/s41746-025-01661-8.
- [73] Z. Chen, Y. Dai, Y. Chen, H. Chen, H. Wu, L. Zhang, Prediction of mortality risk in critically ill patients with systemic lupus erythematosus: a machine learning approach using the mimic-iv database, *Lupus Science & Medicine* 12 (1) (2025). doi:10.1136/lupus-2024-001397.
- [74] H. Jin, N. Ashrafi, K. Alaei, E. Pishgar, G. Placencia, M. Pishgar, A novel multi-task teacher-student architecture with self-supervised pretraining for 48-hour vasoactive-inotropic trend analysis in sepsis mortality prediction, *IEEE Journal of Biomedical and Health Informatics* (2025). doi:10.1109/JBHI.2025.3609667.
- [75] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek, H. M. El-Bakry, Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model, *IEEE Access* 8 (2020) 133541–133564. doi:10.1109/ACCESS.2020.3010556.
- [76] F. Tang, C. Xiao, F. Wang, J. Zhou, Predictive modeling in urgent care: a comparative study of machine learning approaches, *JAMIA open* 1 (1) (2018) 87–98. doi:10.1093/jamiaopen/ooy011.
- [77] B. Shickel, B. Silva, T. Ozrazgat-Baslanti, Y. Ren, K. Khezeli, Z. Guan, P. J. Tighe, A. Bihorac, P. Rashidi, Multi-dimensional patient acuity estimation with longitudinal ehr tokenization and flexible transformer networks, *Frontiers in Digital Health* 4 (2022) 1029191. doi:10.3389/fdgth.2022.1029191.
- [78] S. Purushotham, C. Meng, Z. Che, Y. Liu, Benchmarking deep learning models on large healthcare datasets, *Journal of biomedical informatics* 83 (2018) 112–134. doi:10.1016/j.jbi.2018.04.007.
- [79] S. Zhao, G. Tang, P. Liu, Q. Wang, G. Li, Z. Ding, Improving mortality risk prediction with routine clinical data: a practical machine learning model based on eicu patients, *International Journal of General Medicine* (2023) 3151–3161doi:10.2147/IJGM.S391423.
- [80] F. Khader, J. N. Kather, G. Müller-Franzes, T. Wang, T. Han, S. Tayebi Arasteh, K. Hamesch, K. Bressemer, C. Haarburger, J. Stegmaier, et al., Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data, *Scientific Reports* 13 (1) (2023) 10666. doi:10.1038/s41598-023-37835-1.
- [81] G. Ali, G. A. Al-Kafi, J. T. Faiza, M. I. Hoque, S. A. Suha, Revolutionizing intensive care: A machine learning based approach for icu patients' in-hospital mortality prediction, in: *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, IEEE, 2024, pp. 711–716. doi:10.1109/ICEEICT62016.2024.10534482.
- [82] C.-C. Chiu, C.-M. Wu, T.-N. Chien, L.-J. Kao, C. Li, H.-L. Jiang, Applying an improved stacking ensemble model to predict the mortality of icu patients with heart failure, *Journal of Clinical Medicine* 11 (21) (2022) 6460. doi:10.3390/jcm11216460.
- [83] D. Su, J. Zheng, Y.-K. Shao, J.-Y. Liu, X.-X. Liu, K. Yu, B.-H. Feng, H. Mei, S. Qin, Developing and validating a machine learning-based model for predicting in-hospital mortality among icu-admitted heart failure patients: A study utilizing the mimic-iii database, *Digital Health* 11 (2025) 20552076251335705. doi:10.1177/20552076251335705.

- [84] J. Guo, Y. Cheng, W. He, Y. Zhang, R. Feng, X. Zhang, Uncertainty-aware dynamic fusion for multimodal clinical prediction tasks, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10889595.
- [85] S. Huang, Y. Zhou, Y. Liang, S. Ye, A. Zhu, J. Li, X. Bai, C. Yue, Y. Feng, Machine-learning-derived online prediction models of outcomes for patients with cholelithiasis-induced acute cholangitis: development and validation in two retrospective cohorts, *EClinicalMedicine* 76 (2024). doi:10.1016/j.eclinm.2024.102820.
- [86] L. Liu, W.-W. Lai, B.-W. Li, S.-H. Wang, M.-M. Yu, Y.-C. Liu, Y.-F. Chai, Prediction model of in-hospital mortality risk in intensive care unit patients with cardiac arrest: a multicenter retrospective cohort study based on an ensemble model, *Frontiers in Cardiovascular Medicine* 12 (2025) 1582636. doi:10.3389/fcvm.2025.1582636.
- [87] D. Dimopoulos, V. Danilatou, T. Kostoulas, Length of stay & mortality prediction for patients suffering from stroke in icu: A multimodal approach, in: Proceedings of the 13th Hellenic Conference on Artificial Intelligence, 2024, pp. 1–8. doi:10.1145/3688671.3688787.
- [88] S. Elsharief, S. Shurrah, B. Al Jorf, L. J. L. López, F. E. Shamout, Medmod: Multimodal benchmark for medical prediction tasks with electronic health records and chest x-ray scans, *Proceedings of Machine Learning Research* 287 (2025) 1–23.
- [89] F. Yang, L. Gao, C. Xu, Q. Wang, W. Gao, Association between advanced lung cancer inflammation index and in-hospital mortality in icu patients with community-acquired pneumonia: A retrospective analysis of the mimic-iv database, *Aging Medicine* 7 (4) (2024) 344–351. doi:10.1002/agm2.12334.
- [90] P. Liang, F. Yu, Value of crp, pct, and nlr in prediction of severity and prognosis of patients with bloodstream infections and sepsis, *Frontiers in Surgery* 9 (2022) 857218. doi:10.3389/fsurg.2022.857218.
- [91] Z. Liu, G. Jiang, L. Zhang, P. Shrestha, Y. Hu, Y. Zhu, G. Li, Y. Xiong, L. Zhan, The future of critical care: Ai-powered mortality prediction for acute variceal gastrointestinal bleeding and acute non-variceal gastrointestinal bleeding patients, *Frontiers in Medicine* 12 (2025) 1580094. doi:10.3389/fmed.2025.1580094.
- [92] X. Nie, Y. Cai, J. Liu, X. Liu, J. Zhao, Z. Yang, M. Wen, L. Liu, Mortality prediction in cerebral hemorrhage patients using machine learning algorithms in intensive care units, *Frontiers in neurology* 11 (2021) 610531. doi:10.3389/fneur.2020.610531.
- [93] Y. Zhu, J. Zhang, G. Wang, R. Yao, C. Ren, G. Chen, X. Jin, J. Guo, S. Liu, H. Zheng, et al., Machine learning prediction models for mechanically ventilated patients: Analyses of the mimic-iii database, *Frontiers in Medicine* 8 (2021) 662340. doi:10.3389/fmed.2021.662340.
- [94] W. Bai, B. Hao, L. Xu, J. Qin, W. Xu, L. Qin, Frailty index based on laboratory tests improves prediction of short-and long-term mortality in patients with critical acute myocardial infarction, *Frontiers in Medicine* 9 (2022) 1070951. doi:10.3389/fmed.2022.1070951.
- [95] G.-G. Zhang, J.-H. Hao, Q. Yong, Q.-Q. Nie, G.-Q. Yuan, Z.-Q. Zheng, J.-Q. Li, Lactate-to-albumin ratio is associated with in-hospital mortality in patients with spontaneous subarachnoid hemorrhage and a nomogram model construction, *Frontiers in Neurology* 13 (2022) 1009253. doi:10.3389/fneur.2022.1009253.
- [96] Z. Chen, T. Li, S. Guo, D. Zeng, K. Wang, Machine learning-based in-hospital mortality risk prediction tool for intensive care unit patients with heart failure, *Frontiers in Cardiovascular Medicine* 10 (2023) 1119699. doi:10.3389/fcvm.2023.1119699.
- [97] A. Nandakumar, S. Sudeep, A. C. Sreemohan, S. Vijayakumar, G. J. Sudhakaran, G. Gutjahr, R. K. Pathinaruporthi, S. Balachandran, S. Chandra, S. S. Purushothaman, et al., Developing augmented pro-sofa and pro-saps models by integrating biomarkers pct, nlr, and crp with sofa and saps-iii scores, *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine* 28 (10) (2024) 935. doi:10.5005/jp-journals-10071-24807.

- [98] A. Morish, A. Alsaigh, E. Almaghrabi, H. Alenzi, N. M. Ahmed, A. AlAhdal, A. A. Alsaigh, H. M. Alenzi, Quality performance evaluation of the largest covid-19-designated intensive care unit in the western region of saudi arabia, *Cureus* 15 (5) (2023). doi:10.7759/cureus.39800.
- [99] X. Zhang, J. Ni, H. Zhang, M. Diao, A nomogram to predict in-hospital mortality of gastrointestinal bleeding patients in the intensive care unit, *Frontiers in Medicine* 10 (2023) 1204099. doi:10.3389/fmed.2023.1204099.
- [100] Y. Guo, Y. Leng, C. Gao, Blood urea nitrogen-to-albumin ratio may predict mortality in patients with traumatic brain injury from the mimic database: A retrospective study, *Bioengineering* 11 (1) (2024) 49. doi:10.3390/bioengineering11010049.
- [101] Y. An, R. Qiu, L. Guo, X. Chen, Attention-based multimodal fusion with adversarial network for in-hospital mortality prediction, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2024, pp. 1789–1795. doi:10.1109/BIBM62325.2024.10821815.
- [102] Z. Liu, Y. Yang, H. Song, J. Luo, A prediction model with measured sentiment scores for the risk of in-hospital mortality in acute pancreatitis: a retrospective cohort study, *Annals of Translational Medicine* 10 (12) (2022) 676. doi:10.21037/atm-22-1613.
- [103] A. Alkhatib, L. L. Price, R. Esteitie, P. LaCamera, A predictive model for acute respiratory distress syndrome mortality using red cell distribution width, *Critical care research and practice* 2020 (1) (2020) 3832683. doi:10.1155/2020/3832683.
- [104] R. L. Amorim, L. M. Oliveira, L. M. Malbouisson, M. M. Nagumo, M. Simoes, L. Miranda, E. Bor-Seng-Shu, A. Beer-Furlan, A. F. De Andrade, A. M. Rubiano, et al., Prediction of early tbi mortality using a machine learning approach in a lmic population, *Frontiers in neurology* 10 (2020) 1366. doi:10.3389/fneur.2019.01366.
- [105] A. Al-Dailami, H. Kuang, J. Wang, Predicting length of stay in icu and mortality with temporal dilated separable convolution and context-aware feature fusion, *Computers in Biology and Medicine* 151 (2022) 106278. doi:10.1016/j.combiomed.2022.106278.
- [106] C. V. C. Palamim, M. N. Boschiero, F. A. L. Marson, Epidemiological profile and risk factors associated with death in patients receiving invasive mechanical ventilation in an adult intensive care unit from brazil: a retrospective study, *Frontiers in Medicine* 10 (2023) 1064120. doi:10.3389/fmed.2023.1064120.
- [107] L. Liu, R. Chen, J. Qu, C. Liu, Y. Li, D. Wu, Multi-scale spatiotemporal dynamic graph neural network for early prediction of mortality risks in heart failure patients, *IEEE Journal of Biomedical and Health Informatics* (2025). doi:10.1109/JBHI.2025.3574566.
- [108] P.-S. Hung, P.-R. Lin, H.-H. Hsu, Y.-C. Huang, S.-H. Wu, C.-T. Kor, Explainable machine learning-based risk prediction model for in-hospital mortality after continuous renal replacement therapy initiation, *Diagnostics* 12 (6) (2022) 1496. doi:10.3390/diagnostics12061496.
- [109] L. A. Celi, S. Galvin, G. Davidzon, J. Lee, D. Scott, R. Mark, A database-driven decision support system: Customized mortality prediction, *Journal of Personalized Medicine* 2 (4) (2012) 138–148. doi:10.3390/jpm2040138.
- [110] J. Guo, H. Cheng, Z. Wang, M. Qiao, J. Li, J. Lyu, Factor analysis based on shapley additive explanations for sepsis-associated encephalopathy in icu mortality prediction using xgboost - a retrospective study based on two large database, *Frontiers in Neurology* 14 (2023) 1290117. doi:10.3389/fneur.2023.1290117.
- [111] Z. Gao, X. Liu, Y. Kang, P. Hu, X. Zhang, W. Yan, M. Yan, P. Yu, Q. Zhang, W. Xiao, et al., Improving the prognostic evaluation precision of hospital outcomes for heart failure using admission notes and clinical tabular data: Multimodal deep learning model, *Journal of Medical Internet Research* 26 (2024) e54363. doi:10.2196/54363.

- [112] H. Yang, J. Li, C. Zhang, A. P. Sierra, B. Shen, Predictive model for daily risk alerts in sepsis patients in the icu: visualization and clinical analysis of risk indicators, *Precision Clinical Medicine* 8 (2025) pbaf003. doi:10.1093/pcmedi/pbaf003.
- [113] A. Abdollahi, N. Ashrafi, X. Ma, J. Zhang, D. Wu, T. Wu, Z. Ye, M. Pishgar, Advanced predictive modeling for enhanced mortality prediction in icu stroke patients using clinical data, *Plos one* 20 (5) (2025) e0323441. doi:10.1371/journal.pone.0323441.
- [114] A. M. Ravichandran, J. Grune, N. Feldhus, A. Burchardt, R. Roller, S. Möller, Xai for better exploitation of text in medical decision support, in: *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 2024, pp. 506–513.
- [115] S. Lee, G. Jang, C. Kim, S. Park, K. Yoo, J. Kim, S. Kim, J. Kang, Enhancing clinical outcome predictions through auxiliary loss and sentence-level self-attention, in: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2023, pp. 1210–1217. doi:10.1109/BIBM58861.2023.10385852.
- [116] Y. Wang, J. Rong, Z. Wei, X. Bai, Y. Deng, Predicting icu mortality in heart failure patients based on blood tests and vital signs, *Frontiers in Cardiovascular Medicine* 12 (2025) 1590367. doi:10.3389/fcvm.2025.1590367.
- [117] H. P. Bagaswoto, N. Taufiq, B. Y. Setianto, A simplified risk scoring system to predict mortality in cardiovascular intensive care unit, *Cardiology research* 10 (4) (2019) 216. doi:10.14740/cr884.
- [118] S. Chen, D. Wang, C. Che, Z. Wei, Z. Zhong, A heterogeneous data fusion mortality prediction model based on time-aware self-attention mechanism, *International Journal of Machine Learning and Cybernetics* 16 (9) (2025) 6381–6397. doi:10.1007/s13042-025-02627-8.
- [119] B. Huang, D. Liang, R. Zou, X. Yu, G. Dan, H. Huang, H. Liu, Y. Liu, Mortality prediction for patients with acute respiratory distress syndrome based on machine learning: a population-based study, *Annals of translational medicine* 9 (9) (2021) 794. doi:10.21037/atm-20-6624.
- [120] A. Zalewski, W. Long, A. E. Johnson, R. G. Mark, L.-w. H. Lehman, Estimating patient's health state using latent structure inferred from clinical time series and text, in: *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, IEEE, 2017, pp. 449–452. doi:10.1109/BHI.2017.7897302.
- [121] A. Abkhoo, E. Shaker, M.-M. Mehrabinejad, J. Azadbakht, N. Sadighi, F. Salahshour, Factors predicting outcome in intensive care unit-admitted covid-19 patients: Using clinical, laboratory, and radiologic characteristics, *Critical Care Research and Practice* 2021 (1) (2021) 9941570. doi:10.1155/2021/9941570.
- [122] Y. Deng, S. Liu, Y. Wang, Ziyao; Wang, Y. Jiang, B. Liu, Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients, *Frontiers in Medicine* 9 (2022) 933037. doi:10.3389/fmed.2022.933037.
- [123] B. Mamandipoor, R. R. Bruno, B. Wernly, G. Wolff, J. Fjølner, A. Artigas, B. B. Pinto, et al., Covid-19 machine learning model predicts outcomes in older patients from various european countries, between pandemic waves, and in a cohort of asian, african, and american patients, *PLOS Digital Health* 1 (12) (2022) e0000136. doi:10.1371/journal.pdig.0000136.
- [124] E. Sakhaee, A. Amirahmadi, M. Mahdiani, M. Shojaei, H. Hassanian-Moghaddam, R. Bauer, N. Zamani, H. Pakdaman, K. Gharagozli, Developing a novel prediction model in opioid overdose using machine learning; a pilot analytical study, *Health science reports* 5 (5) (2022) e767. doi:10.1002/hsr2.767.
- [125] L. Yang, D. He, D. Huang, Z. Zhang, Z. Liang, Development and validation of nomogram for hospital mortality in immunocompromised patients with severe pneumonia in intensive care units: a single-center, retrospective cohort study, *International Journal of General Medicine* (2022) 451–463doi:10.2147/IJGM.S344544.

- [126] Y. Feng, Z. Xu, L. Gan, N. Chen, B. Yu, T. Chen, F. Wang, Dcmn: Double core memory network for patient outcome prediction with multimodal data, in: 2019 IEEE International Conference on Data Mining (ICDM), IEEE, 2019, pp. 200–209. doi:10.1109/ICDM.2019.00030.
- [127] K. Zou, S. Huang, W. Ren, H. Xu, W. Zhang, X. Shi, L. Shi, X. Zhong, Y. Peng, M. L. , X. Tang, Development and validation of a dynamic nomogram for predicting in-hospital mortality in patients with acute pancreatitis: A retrospective cohort study in the intensive care unit, *International Journal of General Medicine* (2023) 2541–2553doi:10.2147/IJGM.S409812.
- [128] E. M. Tan, R. Kashyap, I. C. Olson, J. C. O’Horo, Validation of a retrospective computing model for mortality risk in the intensive care unit, *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* 4 (5) (2020) 575–582. doi:10.1016/j.mayocpiqo.2020.09.001.
- [129] F. B. Emdad, S. Tian, E. Nandy, K. Hanna, Z. He, Towards interpretable multimodal predictive models for early mortality prediction of hemorrhagic stroke patients, *AMIA Summits on Translational Science Proceedings* 2023 (2023) 128.
- [130] T. Gao, Z. Nong, Y. Luo, M. Mo, Z. Chen, Z. Yang, L. Pan, Machine learning-based prediction of in-hospital mortality for critically ill patients with sepsis-associated acute kidney injury, *Renal failure* 46 (1) (2024) 2316267. doi:10.1080/0886022X.2024.2316267.
- [131] M. Pieroni, I. Olier, S. Ortega-Martorell, B. W. Johnston, I. D. Welters, In-hospital mortality of sepsis differs depending on the origin of infection: An investigation of predisposing factors, *Frontiers in Medicine* 9 (2022) 915224. doi:10.3389/fmed.2022.915224.
- [132] A. Al-Dailami, H. Kuang, J. Wang, Multimodal representation learning based on personalized graph-based fusion for mortality prediction using electronic medical records, *Big Data Mining and Analytics* 8 (4) (2025) 933–950. doi:10.26599/BDMA.2024.9020099.
- [133] Z. M. Ibrahim, D. Bean, T. Searle, L. Qian, H. Wu, A. Shek, Z. Kraljevic, J. Galloway, S. Norton, J. T. Teo, et al., A knowledge distillation ensemble framework for predicting short-and long-term hospitalization outcomes from electronic health records data, *IEEE Journal of Biomedical and Health Informatics* 26 (1) (2021) 423–435. doi:10.1109/JBHI.2021.3089287.
- [134] K. Bauchmuller, R. Condliffe, J. Southern, C. Billings, A. Charalampopoulos, C. A. Elliot, A. Hameed, D. G. Kiely, I. Sabroe, A. R. Thompson, et al., Critical care outcomes in patients with pre-existing pulmonary hypertension: insights from the aspire registry, *ERJ open research* 7 (2) (2021). doi:10.1183/23120541.00046-2021.
- [135] X.-W. Tang, W.-S. Ren, S. Huang, K. Zou, H. Xu, X.-M. Shi, W. Zhang, L. Shi, M.-H. Lü, Development and validation of a nomogram for predicting in-hospital mortality of intensive care unit patients with liver cirrhosis, *World Journal of Hepatology* 16 (4) (2024) 625. doi:10.4254/wjh.v16.i4.625.
- [136] Y. Wang, C. Yin, P. Zhang, Multimodal risk prediction with physiological signals, medical images and clinical notes, *Heliyon* 10 (5) (2024). doi:10.1016/j.heliyon.2024.e26772.
- [137] W. Lyu, X. Dong, R. Wong, S. Zheng, K. Abell-Hart, F. Wang, C. Chen, A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction, in: *AMIA Annual Symposium Proceedings*, Vol. 2022, American Medical Informatics Association, 2023, p. 719.
- [138] M. Mirtchouk, B. Srikishan, S. Kleinberg, Hierarchical information criterion for variable abstraction, in: *Machine Learning for Healthcare Conference*, PMLR, 2021, pp. 440–460.
- [139] Q. Yang, W. Liu, J. Yu, J. Jiang, T. Xu, Y. Zhou, Effect of prealbumin level on mortality in heatstroke patients, *Experimental and therapeutic medicine* 17 (4) (2019) 3053–3060. doi:10.3892/etm.2019.7298.

- [140] A. Kautzky-Willer, M. Kaleta, S. D. Lindner, M. Leutner, S. Thurner, P. Klimek, Sex differences in clinical characteristics and outcomes of patients with sars-cov-2-infection admitted to intensive care units in austria, *Journal of Personalized Medicine* 12 (4) (2022) 517. doi : 10.3390/jpm12040517.
- [141] Y. Liu, R. Sun, H. Jiang, G. Liang, Z. Huang, L. Qi, J. Lu, Development and validation of a predictive model for in-hospital mortality in patients with sepsis-associated liver injury, *Annals of Translational Medicine* 10 (18) (2022) 997. doi : 10.21037/atm-22-4319.
- [142] R.-E. Ko, J. Cho, M.-K. Shin, S. W. Oh, Y. Seong, J. Jeon, K. Jeon, S. Paik, et al., Machine learning-based mortality prediction model for critically ill cancer patients admitted to the intensive care unit (canicu), *Cancers* 15 (3) (2023) 569. doi : 10.3390/cancers15030569.
- [143] J. Hou, B. Haslund-Gourley, J. Diray-Arce, A. Hoch, N. Rouphael, P. M. Becker, A. D. Augustine, A. Ozonoff, et al., Baseline predictors for 28-day covid-19 severity and mortality among hospitalized patients: results from the impacc study, *Frontiers in Medicine* 12 (2025) 1604388. doi : 10.3389/fmed.2025.1604388.
- [144] Y. Wu, G. Zhu, Association between coagulation disorder scores and in-hospital mortality in arf patients: a retrospective analysis from the mimic-iv database, *Frontiers in Medicine* 10 (2023) 1184166. doi : 10.3389/fmed.2023.1184166.
- [145] D. Zhang, C. Yin, J. Zeng, X. Yuan, P. Zhang, Combining structured and unstructured data for predictive models: a deep learning approach, *BMC medical informatics and decision making* 20 (1) (2020) 280. doi : 10.1186/s12911-020-01297-6.
- [146] K. Zhang, K. Niu, Y. Zhou, W. Tai, G. Lu, Medct-bert: Multimodal mortality prediction using medical convtransformer-bert model, in: *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2023, pp. 700–707. doi : 10.1109/ICTAI59109.2023.00109.
- [147] S. Du, N. Su, Z. Yu, J. Li, Y. Jiang, L. Zeng, J. Hu, A prediction model for prognosis of nephrotic syndrome with tuberculosis in intensive care unit patients: a nomogram based on the mimic-iv v2.2 database, *Frontiers in Medicine* 11 (2024) 1413541. doi : 10.3389/fmed.2024.1413541.
- [148] C. Fang, Y. Bai, Y. Xin, L. Zhang, J. Ma, Factors linked to lung cancer in mimic-iv database, *Journal of Thoracic Disease* 17 (5) (2025) 2765–2777. doi : 10.21037/jtd-2024-1998.
- [149] F. M. Delpino, L. P. Pimenta, D. F. Gonzalez, A. Victor, C. Fonseca Araujo, K. De Araujo-Moura, J. J. Miranda, S. R. R. Batista, A. Dias Porto Chiavegatto Filho, B. P. Nunes, Global performance of machine learning models to predict all-cause mortality: systematic review and meta-analysis, *Scientific reports* (2025).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Extractiontable.pdf](#)