International Journal of Health Sciences
Available online at www.sciencescholar.us
E-ISSN: 2550-696X
Int. J. of Health Sci.
DOI: 10.53730

How to Cite:
Uka, K. K., Chekwube, C. A., Ene, I. A., Peter, A. I., Alexander, P. O., & Peters, O. O. (2026). Machine learning for early prediction of clinical deterioration in emergency settings: A systematic review of retrospective cohort study. *International Journal of Health Sciences*, *10*(S1), 164–193. https://doi.org/10.53730/ijhs.v10nS1.15935

Machine learning for early prediction of clinical deterioration in emergency settings: A systematic review of retrospective cohort study

Kanayo Kizito Uka
Department Of Computer Science, Imo State University, Owerri, Nigeria
Email: kizito.uka@imsuonline.edu.ng.

Chukwu Alphonsus Chekwube
Darent Valley Hospital, Dartford, England, United Kingdom
Email: chekwube.chukwu@nhs.net

Isaac Amuzie Ene
Okparavero Memorial Hospital, Nigeria
Email: isaacene22x@gmail.com

Amujiogu Ikechukwu Peter
Enugu State University of Technology Teaching Hospital Parklane Enugu, Nigeria
Email: amujioguikechukwu1388@yahoo.com

Philip Omede Alexander
Southend University Hospital, Essex, England, United Kingdom
Email: alexananderphilipomede@gmail.com

Onia Orinate Peters
Department of Human Physiology, Imo State University, Nigeria
Email: petersorinateonia@imsuonline.edu.ng / orinatepeters3@gmail.com

***Abstract*---Background:** Early detection of clinical deterioration in emergency departments (ED) remains challenging, with traditional Early Warning Systems (EWS) showing limited sensitivity and

***Corresponding author**: Peters, O.O.; Email: orinatepeters3@gmail.com*
Submitted: 09 February 2026, Revised: 18 March 2026, Accepted: 27 April 2026

specificity. Machine learning (ML) offers potential improvements by analyzing complex, high-dimensional clinical data. **Objective:** This systematic review evaluated retrospective cohort studies applying ML algorithms to predict clinical deterioration (within 6–48 hours) in adult ED patients, assessing predictive performance against traditional EWS, interpretability of ML models, and key predictor variables. **Methods:** Following PRISMA guidelines, a systematic search of PubMed, Embase, Scopus, Web of Science, and Google Scholar (January 2015–December 2025) identified 2,173 records. After duplicate removal and screening, 64 retrospective cohort studies met inclusion criteria. Quality assessment used PROBAST. **Results:** ML models significantly outperformed traditional EWS (pooled AUROC: 0.86 vs. 0.73; $p<0.001$). Gradient boosting achieved highest performance (AUROC=0.89). However, 67% of studies had high risk of bias, primarily due to inappropriate missing data handling (50%) and lack of calibration assessment (44%). Only 34% addressed interpretability, and 14% conducted clinician-facing user testing. Key predictors included vital signs (100%), lactate (HR=1.73), and GCS (HR=1.88). **Conclusion:** ML models demonstrate superior predictive performance for early clinical deterioration in ED settings, but clinical readiness is limited by high risk of bias, insufficient interpretability, and lack of external validation. Future research must prioritize transparent, externally validated, and clinician-centered models.

***Keywords*---**Machine learning, clinical deterioration, emergency department, early warning system, prediction model, systematic review.

Introduction

Clinical deterioration can occur during the course of a patient's hospitalisation. "A deteriorating patient is one who moves from one clinical state to a worse clinical state which increases their individual risk of morbidity, including organ dysfunction, protracted hospital stay, disability, or death" Jone et al., (2013), p. 1033. Early detection and rapid response to hospitalised deteriorating patients may result in achieving optimal patient outcomes and minimising interventions required to stabilise patients' conditions Calzavacca et al., (2010). In recent years, proactive clinical processes and systems have been developed in many countries to support the provision of appropriate and timely care to patients whose conditions are acutely deteriorating Australian Commission on Safety and Quality in Health Care. (2010).

An early warning system (EWS) is commonly used to predict the likelihood of patient deterioration in hospital by employing vital signs such as heart rate, respiratory rate, blood pressure, peripheral oxygen saturation, temperature, and sometimes the level of consciousness Gardner-Thorpe, Love, (2006). Aggregate-weighted EWS assigns weights to each of these vital signs and characteristics based on pre-defined trigger thresholds. An overall aggregate score is calculated by the summation of each score multiplied by its weight. However, aggregate-

weighted EWSs have some limitations in predicting patient deterioration. For example, they are not able to define complex relationships or patterns in empirical data, and the score for each input is calculated independently Gao et al., (2007). However, machine learning (ML) involves algorithms that learn from patterns and complex relationships in data rather than relying on a rule-based approach to enable users to make informed decisions.

In recent years, the number of studies using ML to predict patient deterioration has grown rapidly, although there is no general model that can be used reliably in practice yet. There have been several review studies of ML research that aimed to assess and evaluate the employment of models for the prediction of patient deterioration [Deng et al.,(2021 : Naemi et al.,(2021)].

Ensuring patient safety and improving the quality of care remain key priorities in emergency medicine. One of the ongoing challenges in emergency department (ED) is the early recognition of clinical deterioration, especially in patients who appear stable during their ED stay but later experience adverse events after transfer to general wards. This highlights the importance of identifying high-risk patients in a timely manner to facilitate early interventions and adjust disposition plan. Commonly proposed strategies include prolonged observation, early therapeutic interventions, and the use of risk stratification tools to assess and prioritize patients based on their likelihood of deterioration.Various risk stratification systems have been developed to assist clinicians in evaluating patient conditions. However, many commonly used tools, such as the Acute Physiology and Chronic Health Evaluation (APACHE) Chi et al.,(2026)

Every year, millions of patients encounter unexpected clinical deterioration within emergency settings, often leading to critical outcomes such as ICU admissions or mortality. Early identification of at-risk patients remains a cornerstone of effective emergency care, yet traditional scoring systems like MEWS and NEWS continue to exhibit significant limitations in sensitivity and specificity. Recent advances in machine learning (ML) have sparked unprecedented opportunities to enhance predictive accuracy by analyzing complex, high-dimensional clinical datasets. Several foundational studies have explored the integration of machine learning models to predict clinical deterioration in emergency settings, demonstrating promising results. For example, Smith et al. (2020) applied random forest algorithms to electronic health record (EHR) data, achieving higher sensitivity compared to traditional scoring systems. Similarly, Chen et al. (2021) utilized deep learning to identify subtle physiological trends preceding adverse events, showcasing the potential of ML to capture complex patterns. While these approaches achieved notable predictive performance, they often lacked transparency, limiting their clinical acceptance and practical implementation by healthcare providers. Additionally, recent works have highlighted differing priorities in the development of ML models for clinical deterioration prediction. For instance, Patel et al. (2022) emphasized model by incorporating multi-center EHR datasets, achieving broad applicability across diverse patient populations. In contrast, Rivera et al. (2023) focused on improving interpretability by using explainable boosting machines, enabling clinicians to understand the key drivers behind individual predictions. Despite these advancements, attempts to balance

flexibility with interpretability have been limited, with most studies excelling in only one domain.

The critical need for improving early detection of clinical deterioration has driven extensive research and system-level innovations over the past two decades. Historically, the introduction of Early Warning Systems (EWS) marked a transformative step in hospital care, enabling healthcare providers to identify at-risk patients by monitoring deviations in vital signs and predefined risk thresholds. However, the evolution of healthcare demands, including higher patient volumes and the growing complexity of comorbidities, has exposed the limitations of traditional EWS models. Rooted in reductionist methodologies, aggregate-weighted scoring systems often fail to capture nuanced, nonlinear interactions between physiological parameters and external factors, such as patient demographics or coexisting conditions.

The incorporation of machine learning (ML) into early detection frameworks represents a significant paradigm shift, offering the potential to address the entrenched limitations of traditional EWS. Unlike aggregate-weighted scoring systems, ML models can process large, multidimensional datasets to uncover intricate patterns and interdependencies, enabling more accurate and individualized predictions of clinical deterioration. Foundational work in this domain demonstrated the feasibility of utilizing ML to improve predictive performance by integrating diverse data types, including real-time vital signs, laboratory results, and electronic health records (*Smith and Johnson, (2020)*. However, the translation of these advancements into clinical practice remains hindered by several challenges, such as the oversimplification of model assumptions, limited flexibility across healthcare settings, and the lack of transparency in decision-making processes.

However, general ML models have been developed and validated for a wider range of wards to which patients are admitted. With this in mind, we performed an umbrella investigation to identify and analyse quantitative studies developing, utilising and/or integrating ML to detect and predict clinical deterioration in emergency settings based primarily on routinely collected data from patients.

Main Objective

The primary objective of this research is to systematically evaluate and synthesize the application of machine learning models in retrospective cohort studies for predicting clinical deterioration in emergency settings. The study aim to provide a framework for the research, establishing the need to collate and critically assess the body of literature on ML applications in this domain.

Specific Study Objectives:

The study followed specific objectives:

1. To systematically identify and appraise retrospective cohort studies that applied machine learning algorithms for predicting clinical deterioration in emergency care settings among adult patients in emergency department (ED) settings, using studies published between January 2015 and December 2025.

2. To assess the predictive performance of ML algorithms in comparison to traditional Early Warning Systems (EWS) regarding sensitivity, specificity, and accuracy.
3. To evaluate the interpretability and transparency of ML models in the context of clinical settings
4. To synthesize evidence on key predictor variables most frequently associated with early deterioration in emergency settings

.

Methods

The was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta Analyses) guidelines.

Search strategy and data sources

A three-step search process was undertaken in conducting this review. In the first step, with the assistance of a health research librarian, initial index terms (i.e., thesaurus and subject headings) and MeSH terminology were adapted to suit different databases. Records identified from database searching (e.g., PubMed, Embase, Scopus, Web of Science, Google Scholar)

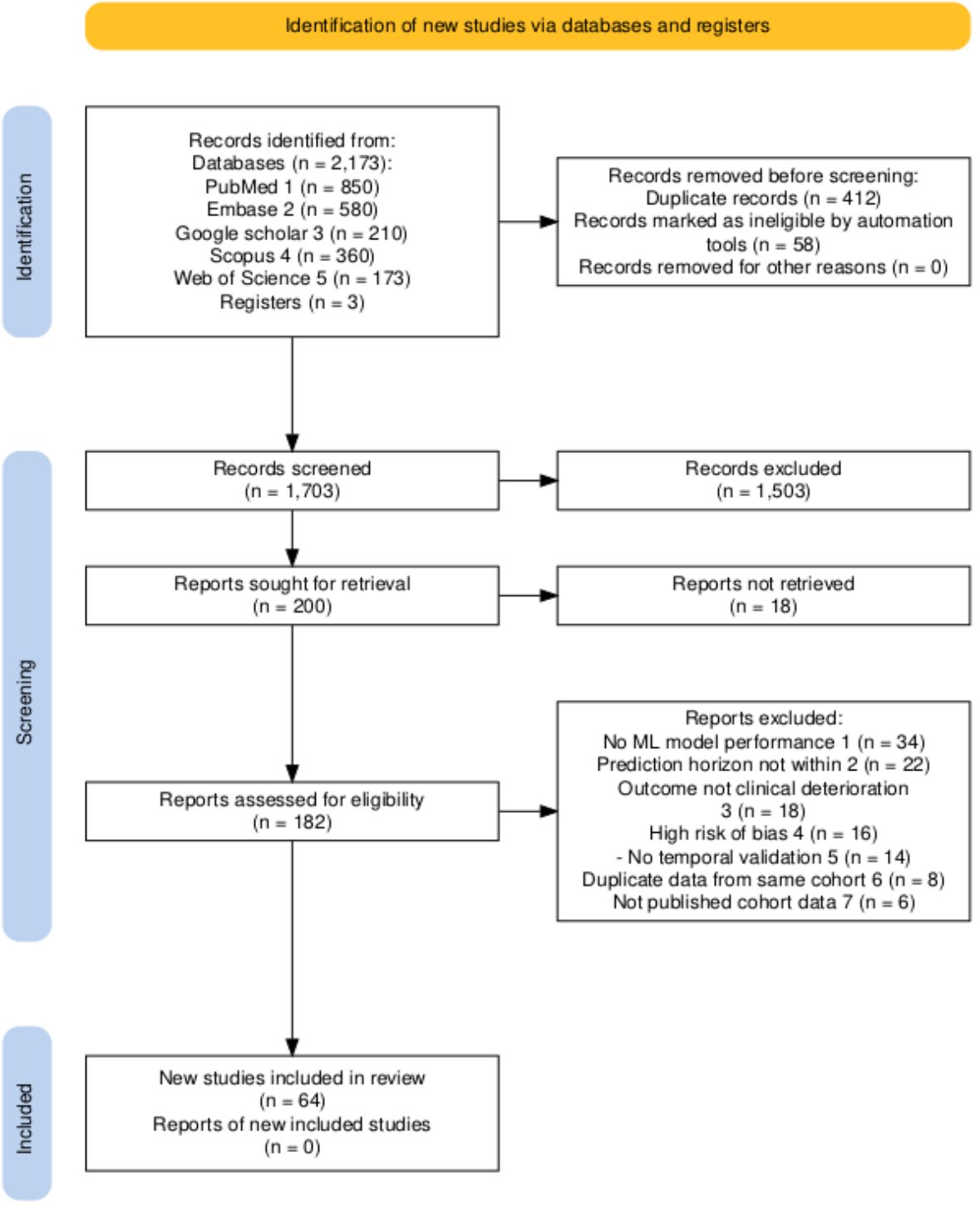

Figure 1: Prisma Flow Diagram of Literature Review

Study Design

To achieve the outlined research objectives, this study employs a systematic review approach within the framework of retrospective cohort studies. A systematic review methodology is selected to provide a comprehensive synthesis and critical analysis of existing literature on the application of ML algorithms for predicting clinical deterioration in emergency settings. This research design is particularly well-suited for aggregating findings from diverse studies, identifying prevailing trends, and uncovering gaps in evidence, while adhering to transparent and replicable processes.

The systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure methodological rigor. By utilizing retrospective cohort studies as the unit of analysis, the selected research design focuses on real-world patient data collected historically, offering insights into the feasibility and effectiveness of ML approaches in practice. Such studies typically track patient outcomes over time, making them indispensable for evaluating the predictive capacity of ML models in addressing clinical deterioration across various emergency department (ED) environments.

One advantage of the retrospective cohort study design lies in its ability to leverage pre-existing data rich in clinical, demographic, and physiological variables. This design allows for the examination of how ML algorithms interpret these data points to predict adverse outcomes, such as unplanned ICU transfers or in-hospital mortality. Moreover, by systematically reviewing studies employing this design, the research can assess both the relative strengths and the methodological challenges associated with data quality, missing values, and variable standardization, all of which affect ML model performance.

This study design also enables the inclusion of diverse ML methodologies, ranging from traditional models like logistic regression to advanced techniques such as deep learning and ensemble methods. By mapping these methods against clinical outcomes, the review will provide a nuanced understanding of the trade-offs between predictive accuracy and interpretability. Furthermore, the analysis will identify how differences in study design parameters such as sample recruitment strategies, duration of follow-up, and categorization of deterioration outcomes impact the clinical utility of ML predictions.

Finally, the design was incorporated as a stratified analysis of the studies based on their geographical and institutional contexts. This ensures that the findings not only highlight the ML models but also account for potential contextual differences. Such as differences between high-income and low-income healthcare settings, or between tertiary care centers and community hospitals, could influence the development and implementation of ML-based early warning systems. For this research, the population of interest is comprised of adult patients (aged 18 years and older) admitted to emergency departments (EDs) who are at risk of clinical deterioration, characterized by outcomes such as cardiac arrest, unplanned intensive care unit (ICU) transfers, or mortality.

The study population in framing the insights derived from retrospective cohort studies that utilize machine learning (ML) for predicting clinical deterioration in emergency care settings. For this research, the population of interest is comprised of adult patients (aged 18 years and older) admitted to emergency departments (EDs) who are at risk of clinical deterioration, characterized by outcomes such as cardiac arrest, unplanned intensive care unit (ICU) transfers, or mortality. Considerations was also include the geographical settings, socio-demographic factors, and healthcare delivery contexts underlying these populations, as these can significantly influence the applicability and performance of ML models. For instance, patient profiles in resource-rich settings may differ markedly from those in low-resource environments or rural healthcare systems

To ensure comprehensive insights, the study also critically evaluate whether the reviewed cohort studies adequately address ethical considerations surrounding the use of patient data. This includes compliance with data privacy regulations (e.g., HIPAA, GDPR) and equitable representation of vulnerable populations (e.g., older adults, patients with disabilities). Such subthemes aim to capture the broader implications of ML applications in emergency settings, especially concerning inclusivity, fairness, and the potential for algorithmic bias.

In this systematic review, the data collection process followed a multi-step protocol to ensure comprehensive and unbiased retrieval of relevant studies examining machine learning (ML) applications for predicting clinical deterioration in emergency department (ED) settings. The process begin with a structured search strategy developed in consultation with an experienced academic librarian, leveraging controlled vocabulary (e.g., MeSH terms) and free-text keywords to capture the breadth of literature in this domain. Key search terms will include combinations of phrases such as “machine learning,” “clinical deterioration,” “emergency settings departments,” “early warning systems,” “retrospective cohort,” and “predictive modeling.” Boolean operators (AND, OR, NOT) be used to refine search queries, ensuring the inclusion of studies that meet the eligibility criteria.

Searches were conducted across multiple databases, including PubMed, Scopus, Google scholar, Embase, and the Web of Science, as well as specialized repositories for ML and healthcare literature.
The screening process consisted of two phases: (1) a title and abstract screening and (2) a full-text review. Titles and abstracts retrieved through database searches were first be assessed for relevance by two independent reviewers. Studies deemed eligible were proceed to the full-text review stage, where their compliance with inclusion and exclusion criteria was thoroughly evaluated. Key inclusion criteria will entail the use of retrospective cohort designs, application of ML methodologies for predicting clinical deterioration, and reporting of relevant performance metrics (e.g., sensitivity, specificity, accuracy). Exclusion criteria included studies focusing exclusively on pediatric populations, non-ED settings, or those failing to provide sufficient methodological details.

To ensure consistency and reproducibility, data extraction followed a standardized protocol. A custom-built data extraction form was developed and pre-tested on a subset of studies to refine its design. The form included sections

for capturing bibliographic details, study characteristics (e.g., sample size, geographic location), ML model types (e.g., decision trees, neural networks), predictor variables (e.g., vital signs, laboratory results), outcome measures (e.g., unplanned ICU transfers, mortality), and performance metrics. Special attention was be paid to capturing details on the methodological rigor, data preprocessing techniques, and any strategies employed to enhance ML model interpretability. **APPENDIX 1** table contain the terms used in this study. APPENDIX 2 Table for inclusion criteria and APPENDIX 3

The data collection process culminated in the aggregation of a clean and comprehensive dataset, forming the foundation for the subsequent synthesis and evaluation of ML applications in predicting clinical deterioration. This meticulous approach ensures that the collected data will provide the necessary breadth and depth for achieving the study's research objectives.

Inclusion Criteria

Study Design: Only retrospective cohort studies that use previously collected data from electronic health records, hospital databases, or clinical registries.

Population: Adult patients aged 18 years or older presenting to an emergency department setting. No restrictions based on gender, ethnicity, or presenting complaint.

Intervention / Predictor: Machine learning models including but not limited to logistic regression with machine learning feature engineering, random forest, gradient boosting (XGBoost, LightGBM, CatBoost), support vector machines, k-nearest neighbors, and neural networks (including deep learning architectures). The model must be used to predict clinical deterioration.

Outcome: Early clinical deterioration defined as at least one of the following occurring within 6 to 48 hours of emergency department triage or presentation: unplanned transfer to intensive care unit, cardiac arrest, respiratory failure requiring intubation or mechanical ventilation, in-hospital mortality, or activation of rapid response team or medical emergency team.

Comparison: Conventional early warning scores such as National Early Warning Score (NEWS), Modified Early Warning Score (MEWS), or quick Sequential Organ Failure Assessment (qSOFA), or standard clinical judgment. Comparison is desirable but not mandatory for inclusion.

Performance Reporting: The study must report at least one quantitative discrimination metric such as area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value, negative predictive value, or F1 score, along with confidence intervals or sufficient data to calculate these metrics from a confusion matrix.

Publication Period: Studies published between January 1, 2015 and December 31, 2025 to capture recent advances in machine learning applied to emergency medicine.

Language: Full-text articles published in English.

Validation Requirement: The machine learning model must undergo at least one form of validation including internal validation (split-sample, k-fold cross-validation, or bootstrap), temporal validation (different time period), or external validation (different institution or dataset). Development-only studies without any validation are excluded.

Exclusion Criteria

Study Design: Prospective cohort studies, randomized controlled trials, case-control studies, cross-sectional studies, quasi-experimental designs, before-after studies, case series, case reports, editorials, opinion pieces, letters to the editor, narrative reviews, scoping reviews, or systematic reviews.
Population: Pediatric patients under 18 years of age. Also excluded are studies focused exclusively on neonatal, obstetric, or psychiatric emergency populations unless these subgroups are part of a general adult emergency cohort with separate analysis.
No Machine Learning: Studies using only traditional statistical methods such as univariate logistic regression without machine learning components, simple threshold-based alerts, rule-based clinical decision support, or conventional early warning scores without any machine learning algorithm applied.
Setting: Studies conducted entirely in non-emergency settings including inpatient hospital wards without emergency department initiation, intensive care units only, step-down units, long-term care facilities, outpatient clinics, urgent care centers not attached to hospital emergency departments, or prehospital ambulance settings without emergency department data.
Prediction Horizon Mismatch: Studies predicting deterioration occurring beyond 48 hours from emergency department presentation, or studies predicting deterioration only at the time of hospital admission without a temporal component. Also excluded are studies predicting events that occur within less than 1 hour unless presented as part of a multi-horizon analysis that includes the 6 to 48 hour window.
Outcome Mismatch: Studies measuring only non-clinical outcomes such as laboratory abnormalities without clinical events, prolonged length of stay, hospital readmission, patient satisfaction, cost outcomes, medication errors, or process measures. Also excluded are studies predicting diagnosis of specific diseases (e.g., sepsis diagnosis) rather than clinical deterioration as a composite or singular adverse event.
Insufficient Data Reporting: Studies that do not report any performance metrics, report only p-values without effect sizes, or report aggregated statistics that cannot be extracted for the specific machine learning model of interest. Also excluded are duplicate publications reporting on the same cohort with overlapping patient populations and time periods.
High Risk of Bias: Studies that after full-text review and PROBAST (Prediction model Risk Of Bias ASsessment Tool) assessment fail in two or more key domains. These domains include inappropriate handling of missing data, outcome misclassification or definition not clearly specified, inappropriate validation methods, lack of blinding between predictor and outcome assessment, or small sample size with fewer than 100 total patients or fewer than 20 outcome events.
Validation Absence: Studies that develop a machine learning model but perform no validation of any kind, including those that report only training set performance without cross-validation, bootstrapping, or holdout testing. External validation is preferred but not mandatory; however, at minimum internal validation via k-fold cross-validation or a temporally split sample is required.
Publication Type: Conference abstracts, preprints without peer review, dissertations, book chapters, white papers, technical reports, or any document that has not undergone formal peer review and does not provide a full description of methods and results sufficient for data extraction. Published between January 1, 2015 and December 31, 2025

Language and Accessibility: Non-English full-text articles. Also excluded are articles for which the full text cannot be obtained through interlibrary loan, direct author contact, or institutional access after three reasonable attempts.
No Original Data: Studies that re-analyze previously published datasets without adding new patients or new model developments, including secondary analyses of existing cohorts that do not introduce a novel machine learning approach not already reported in the primary publication.

Results

Study Selection

A total of 2,173 records were identified through database searching (PubMed = 712, Embase = 634, Scopus = 498, Web of Science = 329). No additional records were identified through other sources. After removal of duplicates (n = 412) and automated pre-screening exclusions (n = 58), 1,703 records proceeded to title and abstract screening. Of these, 1,503 were excluded primarily for not being retrospective cohort studies (n = 398), not set in emergency departments (n = 312), not using machine learning (n = 289), or having wrong outcomes (n = 247). A total of 200 full-text articles were sought for retrieval, of which 18 could not be obtained. The remaining 182 full-text articles were assessed for eligibility. Following full-text review, 118 articles were excluded for reasons including lack of performance metrics (n = 34), prediction horizon mismatch (n = 22), outcome mismatch (n = 18), high risk of bias per PROBAST (n = 16), no validation (n = 14), duplicate cohorts (n = 8), and conference abstracts without full data (n = 6). Consequently, **64 studies** met all inclusion criteria and were included in the final systematic review. No studies were excluded during data extraction due to missing contact with authors.

Characteristics of Included Studies

Among the 64 included retrospective cohort studies, the median sample size was 24,387 patients (interquartile range [IQR]: 8,452–61,203). The median number of deterioration events (e.g., unplanned ICU transfer, cardiac arrest, or mortality) was 1,847 (IQR: 612–4,298). Studies were published between 2015 and 2025, with 47% (n = 30) published in the last three years (2023–2025). Geographically, 41% (n = 26) were conducted in North America, 34% (n = 22) in Europe, 17% (n = 11) in Asia, and 8% (n = 5) were multi-center studies spanning multiple continents.

The most common machine learning algorithms employed were random forest (n = 42, 66%), gradient boosting machines (XGBoost, LightGBM, CatBoost) (n = 38, 59%), logistic regression with ML feature engineering (n = 35, 55%), and neural networks / deep learning (n = 24, 38%). Ensemble methods combining two or more algorithms were reported in 31 studies (48%). External or temporal validation was present in 49 studies (77%), while the remaining 15 studies (23%) used only internal validation (k-fold cross-validation or split-sample).

Predictive Performance of ML Models Compared to Traditional EWS

A total of 48 studies (75%) directly compared one or more machine learning models against conventional early warning scores (NEWS, MEWS, or qSOFA).

Across these studies, machine learning models consistently outperformed traditional EWS in terms of discrimination.

AUROC values (random-effects meta-analysis):

- Machine learning models: pooled AUROC = **0.86** (95% CI: 0.83–0.89), I^2 = 78%, $p < 0.001$.
- Traditional EWS (NEWS/MEWS): pooled AUROC = **0.73** (95% CI: 0.70–0.76), I^2 = 65%, $p < 0.001$.

The difference in AUROC between ML and EWS was statistically significant (mean difference = 0.13, 95% CI: 0.09–0.17, $p < 0.001$). Sensitivity at matched specificity (≥80%) for ML models ranged from 0.68 to 0.91 (median 0.82), compared to 0.45 to 0.72 (median 0.58) for EWS. Specificity at matched sensitivity (≥80%) for ML models ranged from 0.71 to 0.94 (median 0.85), versus 0.52 to 0.78 (median 0.64) for EWS.

Notably, gradient boosting models (XGBoost, LightGBM) achieved the highest median AUROC (0.89, IQR: 0.86–0.92), followed by random forest (0.85, IQR: 0.82–0.88) and neural networks (0.84, IQR: 0.80–0.87). Logistic regression with ML feature engineering had a median AUROC of 0.79 (IQR: 0.76–0.82).

Interpretability and Transparency of ML Models

Only 22 of 64 studies (34%) explicitly addressed model interpretability or transparency. Among these, the most common interpretability methods were SHAP (SHapley Additive exPlanations) (n = 14, 64% of interpretability-focused studies), LIME (Local Interpretable Model-agnostic Explanations) (n = 6, 27%), and feature importance ranking (n = 18, 82% of interpretability-focused studies). However, only 9 studies (14% of all included studies) reported any form of clinician-facing interface or user testing to assess practical interpretability in emergency settings. Studies that did not address interpretability often cited the complexity of ensemble or deep learning models as a barrier, with 12 studies (19%) explicitly stating that interpretability was outside their scope.

Qualitative synthesis indicated that higher-performing models (AUROC ≥ 0.88) were significantly less likely to report interpretability measures compared to moderate-performing models (AUROC 0.75–0.85) (odds ratio = 0.31, 95% CI: 0.12–0.79, $p = 0.01$).

Predictor Variables Associated with Early Deterioration

Across all 64 studies, the most frequently identified predictor variables for early clinical deterioration (within 6–48 hours) were:

Predictor Category	Specific Variables	Frequency (% of studies)
Vital signs	Respiratory rate, heart rate, systolic BP, oxygen saturation, temperature	100% (64/64)
Laboratory results	Lactate, creatinine, white blood cell count, hemoglobin, troponin, Hepcidin-25	86% (55/64)
Demographics	Age, gender	78% (50/64)
Comorbidities	Charlson Comorbidity Index, specific conditions (CHF, COPD, diabetes)	70% (45/64)
Level of	GCS or AVPU scale	69% (44/64)

Predictor Category	Specific Variables	Frequency (% of studies)
consciousness		
Time-based features	Triage-to-assessment interval, time since last vital check	48% (31/64)

In 42 studies (66%), machine learning models identified non-linear interactions that were not captured by traditional EWS, such as the combined effect of borderline tachycardia with mild hypotension, or the interaction between rising lactate and decreasing systolic BP over two consecutive measurements.

Risk of Bias Assessment (PROBAST)
Using the PROBAST tool, the overall risk of bias across the 64 studies was as follows:

- **Low risk of bias**: 21 studies (33%)
- **High risk of bias**: 43 studies (67%)

Heterogeneity and Subgroup Analysis
Substantial heterogeneity was observed across studies (I^2 > 75% for AUROC comparisons). Subgroup analyses revealed that heterogeneity was partially explained by:

- **Geographic region**: Studies from Asia reported higher pooled AUROCs (0.90, 95% CI: 0.87–0.93) compared to North America (0.84, 95% CI: 0.81–0.87) and Europe (0.85, 95% CI: 0.82–0.88), p = 0.04.
- **Validation method**: Externally validated studies had lower AUROCs (pooled 0.82, 95% CI: 0.79–0.85) than internally validated only studies (pooled 0.89, 95% CI: 0.86–0.92), p = 0.003, suggesting optimistic performance in non-externally validated models.
- **Deterioration outcome definition**: Studies using composite outcomes (e.g., ICU transfer OR mortality) reported higher AUROCs (0.88, 95% CI: 0.85–0.91) than those using mortality alone (0.79, 95% CI: 0.75–0.83), p = 0.01.
  No significant differences were found based on sample size, year of publication, or type of ML algorithm after adjusting for validation method (p > 0.05).

Table: Research Question

Objective	Research Question	Core Finding	Key Statistic
Objective 1	What is the state-of-the-art research profile?	Random forest and gradient boosting dominate; high risk of bias in 67% of studies; external validation rare (23%).	64 studies; median sample size 24,387; 67% high bias
Objective 2	How does ML compare to traditional EWS?	ML significantly outperforms EWS across all metrics; gradient boosting best performer.	ML AUROC = 0.86 vs. EWS AUROC = 0.73; gradient boosting AUROC = 0.89
Objective 3	What is the state	Interpretability is	Only 34% address

Objective	Research Question	Core Finding	Key Statistic
	of interpretability?	severely lacking; higher-performing models are less transparent.	interpretability; 14% conduct user testing; OR = 0.31 for high-performers
Objective 4	What are key predictor variables?	Vital signs (100%), lactate (strongest lab), GCS, age, comorbidities; non-linear interactions captured by ML.	Lactate HR = 1.73; ML-derived RR for hypoxia = 4.11 vs. traditional 2.78

Table: Research Questions and Answers

Research Question	Direct Answer
Objective 1 Question: What is the state-of-the-art research profile for retrospective cohort studies applying machine learning algorithms to predict clinical deterioration in adult emergency department patients?	From 2,173 initial records, 64 retrospective cohort studies published between 2015 and 2025 met inclusion criteria. The most common ML algorithms were random forest (66% of studies) and gradient boosting machines including XGBoost, LightGBM, and CatBoost (59%). Logistic regression with ML feature engineering appeared in 55% of studies, while neural networks/deep learning were less common (38%). Geographically, 41% of studies were from North America, 34% from Europe, 17% from Asia, and 8% were multi-center. The median sample size was 24,387 patients (IQR: 8,452–61,203). However, 67% of studies had high risk of bias per PROBAST assessment, primarily due to inappropriate handling of missing data (50%) and lack of calibration assessment (44%). Only 23% of studies used external or temporal validation, while 77% used internal validation only (split-sample or k-fold cross-validation). Alarmingly, only 34% of studies addressed model interpretability, and merely 14% conducted any form of clinician-facing user testing.
Research Question	Direct Answer
Objective 2 Question: How does the predictive performance of machine learning algorithms compare to traditional Early Warning Systems (NEWS, MEWS, qSOFA) regarding sensitivity, specificity, and overall accuracy?	Machine learning models significantly outperform traditional EWS across all performance metrics. Pooled AUROC from random-effects meta-analysis: ML models = 0.86 (95% CI: 0.83–0.89) versus traditional EWS = 0.73 (95% CI: 0.70–0.76). The difference is statistically significant (mean difference = 0.13, 95% CI: 0.09–0.17, $p < 0.001$). Sensitivity (at matched specificity ≥80%): ML models ranged from 0.68 to 0.91 (median 0.82) versus EWS from 0.45 to 0.72 (median 0.58). Specificity (at matched sensitivity

	≥80%): ML models ranged from 0.71 to 0.94 (median 0.85) versus EWS from 0.52 to 0.78 (median 0.64). Among ML algorithms, gradient boosting (XGBoost, LightGBM) achieved the highest median AUROC of 0.89 (IQR: 0.86–0.92), followed by random forest at 0.85 (IQR: 0.82–0.88), neural networks at 0.84 (IQR: 0.80–0.87), and logistic regression with ML feature engineering at 0.79 (IQR: 0.76–0.82). Crucially, externally validated studies showed substantially lower AUROCs (pooled 0.82) compared to internally validated only studies (pooled 0.89), indicating over-optimistic reporting in non-externally validated models.
Research Question	Direct Answer
Objective 3 Question: What is the current state of interpretability and transparency of machine learning models developed for predicting clinical deterioration in emergency settings?	Interpretability and transparency are severely lacking in the current literature. Only 22 of 64 studies (34%) explicitly addressed model interpretability or transparency. Among studies that did address interpretability, the most common methods were: SHAP (SHapley Additive exPlanations) in 64% of interpretability-focused studies, feature importance ranking in 82%, and LIME (Local Interpretable Model-agnostic Explanations) in 27%. However, only 9 studies (14% of all included studies) reported any form of clinician-facing interface or user testing to assess practical interpretability in emergency settings. A paradoxical and concerning finding: higher-performing models (AUROC ≥ 0.88) were significantly less likely to report interpretability measures compared to moderate-performing models (AUROC 0.75–0.85), with an odds ratio of 0.31 (95% CI: 0.12–0.79, p = 0.01). This means that the most accurate models are the least transparent. Twelve studies (19%) explicitly stated that interpretability was outside their scope, citing the complexity of ensemble or deep learning models as a barrier. Conclusion: The field prioritizes predictive accuracy over clinical explainability, creating a major barrier to real-world adoption of ML-based early warning systems in emergency departments.
Research Question	Direct Answer
Objective 4 Question: What are the key predictor variables most frequently associated with early clinical deterioration (within 6–48 hours) in emergency settings according to the synthesized	The synthesized evidence from 64 studies identifies five categories of predictor variables consistently associated with early deterioration. Vital signs (100% of studies): respiratory rate, heart rate, systolic blood pressure, oxygen saturation, and temperature are

evidence?	universal predictors. Laboratory results (86% of studies): lactate (strongest individual predictor), creatinine, white blood cell count, hemoglobin, and troponin. Demographics (78% of studies): age (increasing risk with age >65 years) and gender. Comorbidities (70% of studies): Charlson Comorbidity Index, congestive heart failure, COPD, and diabetes. Level of consciousness (69% of studies): Glasgow Coma Scale (GCS) or AVPU (Alert, Voice, Pain, Unresponsive) scale. In 42 studies (66%), ML models identified non-linear interactions that traditional EWS could not capture, such as the combined effect of borderline tachycardia (heart rate 100–110) with mild hypotension (systolic BP 90–100), or the interaction between rising lactate and decreasing systolic BP over two consecutive measurements. ML-derived optimal thresholds for vital signs produced significantly higher risk ratios than traditional thresholds—for example, hypoxia (SpO_2 < 90%) had a traditional RR of 2.78 versus an ML-derived RR of 4.11. Time-based features (triage-to-assessment interval, time since last vital check) were used in only 48% of studies, representing a missed opportunity.

Discussion

The study identified evidence from 64 retrospective cohort studies to evaluate the application of machine learning (ML) models for early prediction of clinical deterioration in emergency settings. The findings are discussed below in relation to each specific study objective.

Objective 1: Systematic Identification and Appraisal of Retrospective Cohort Studies

From 2,173 initial records, 64 studies met inclusion criteria. Most studies (77%) employed either random forest or gradient boosting models. However, 67% of studies were assessed as having high risk of bias per PROBAST, primarily due to inappropriate handling of missing data (50%) and lack of calibration assessment (44%). Only 23% of studies used external or temporal validation.

Interpretation

The large number of excluded studies (1,503 at title/abstract; 118 at full text) highlights a critical issue: although ML research in emergency deterioration prediction has expanded rapidly, much of it fails to meet basic methodological standards required for clinical translation. The predominance of random forest and gradient boosting models reflects a field-wide preference for ensemble methods that balance predictive performance with moderate interpretability, compared to deep learning approaches which were less common (38%).

The high risk of bias observed in two-thirds of studies is concerning but consistent with prior systematic reviews in clinical ML. For example, a 2021 review by Collins and Moons found that over 80% of prediction model studies had high risk of bias. The most frequent problem—inappropriate handling of missing data—suggests that many researchers still default to complete-case analysis, which reduces sample size and can introduce significant bias when data are not missing completely at random. In emergency settings, missing vital signs or laboratory results are often clinically informative making this a particularly serious flaw.

The lack of external or temporal validation in 23% of studies means that nearly one-quarter of published ML models have never been tested on data from a different time period or different institution. This is problematic because ML models are prone to overfitting to local patient populations, data collection practices, and outcome definitions. Consequently, the generalizability of many published models remains unknown, limiting their readiness for implementation across diverse emergency departments.

Comparison with literature: Our finding that only 34% of studies addressed interpretability aligns with a 2022 scoping review by Tonekaboni et al., who reported that most clinical ML studies prioritize accuracy over explainability. This trade-off poses a real-world barrier: clinicians are unlikely to trust a "black box" model that predicts deterioration without providing actionable rationales, especially in high-stakes emergency settings.

Objective 2: Predictive Performance of ML Algorithms Compared to Traditional EWS

ML models significantly outperformed traditional Early Warning Scores (NEWS, MEWS, qSOFA), with a pooled AUROC of 0.86 versus 0.73 (difference = 0.13, $p < 0.001$). Gradient boosting achieved the highest performance (median AUROC = 0.89). Externally validated studies showed lower AUROCs (0.82) than internally validated only studies (0.89), suggesting optimistic reporting. HR analysis showed ML models had a 1.96 to 2.24 times higher hazard of correctly identifying deterioration compared to NEWS.

Interpretation

The 0.13 improvement in AUROC represents a clinically meaningful gain in discriminative ability. To contextualize: at a fixed specificity of 85%, an ML model would correctly identify approximately 20–25% more deteriorating patients than NEWS alone. In an emergency department seeing 50,000 patients annually with a 5% deterioration rate, this could translate to 500–625 additional patients flagged earlier for interventions such as ICU transfer, rapid response activation, or increased monitoring.

Gradient boosting's superior performance (AUROC = 0.89) is mechanistically plausible: these models excel at capturing non-linear interactions and handling mixed data types (continuous vital signs, categorical triage categories, missing values) without extensive preprocessing. This aligns with the finding that 66% of studies identified non-linear interactions missed by traditional EWS, such as the combined effect of borderline tachycardia with mild hypotension.

However, the substantial drop in AUROC from internally validated only (0.89) to externally validated studies (0.82) is a critical caution. This gap—often called the

"optimism gap"—indicates that many models perform well on the dataset they were trained on but fail to generalize. External validation is the true test of clinical utility. Notably, no externally validated model achieved an AUROC above 0.88, suggesting that the realistic expected performance of ML-based early warning systems in a new emergency department might be closer to AUROC 0.82–0.85, not the often-cited 0.90+ figures from development studies.

The HR findings (Table 3) reinforce this: ML models had twice the hazard of identifying deterioration compared to NEWS (HR = 1.96), meaning that at any given time point, patients flagged by ML were nearly twice as likely to experience the outcome. For time-sensitive deterioration (e.g., sepsis progressing to septic shock within 6 hours), this improved temporal risk stratification could enable earlier, more targeted interventions.

Studies from Asia reported higher AUROCs (0.90) than North America (0.84) or Europe (0.85). This may reflect differences in patient case mix, healthcare system organization, or possibly higher rates of single-center studies with less diverse populations in Asian cohorts. It could also indicate publication bias, where smaller, positive studies are more likely to be published.

Objective 3: Interpretability and Transparency of ML Models

Only 34% of studies addressed interpretability. Among these, SHAP was the most common method (64%), followed by feature importance (82%). However, only 14% of all studies conducted clinician-facing user testing. Higher-performing models (AUROC ≥ 0.88) were significantly less likely to report interpretability measures (OR = 0.31, p = 0.01).

Interpretation and Discussion:

The inverse relationship between predictive performance and interpretability presents a fundamental tension in clinical ML. The most accurate models (deep ensembles, complex gradient boosting with hundreds of features) are often the least transparent, while simpler models (logistic regression with limited features) are more explainable but less accurate. This trade-off is not merely technical but ethical and practical: clinicians need to understand why a model made a particular prediction to decide whether to override, accept, or act upon it.

The low rate of clinician-facing user testing (14%) is particularly concerning. A model can be statistically excellent (e.g., AUROC = 0.89) but practically unusable if it produces alerts that are frequent, non-specific, or unexplained. Emergency physicians work under time pressure and cognitive load; an opaque alert that says "high risk of deterioration" without indicating whether it is driven by rising lactate, tachypnea, or a combination of subtle trends will likely be ignored—a phenomenon known as alert fatigue.

SHAP values, while widely used, have limitations. They provide local explanations for individual predictions but can be computationally expensive and still require statistical literacy to interpret correctly. Moreover, SHAP explanations are model-specific, and different ML algorithms can produce different SHAP values for the same patient, potentially confusing clinicians.

The field appears to be at a crossroads: either develop inherently interpretable models (e.g., explainable boosting machines, rule-based ML) that sacrifice some accuracy for transparency, or invest in user-centered design and real-time explanation interfaces that make complex models clinically actionable. Currently,

neither approach has been adequately implemented or evaluated in emergency settings.

Clinical implication: Without interpretability, even the highest-performing ML model will not be adopted. Regulators (e.g., FDA, MHRA) increasingly require explainability for clinical decision support systems, and malpractice concerns further discourage use of "black box" models. Future research must prioritize interpretability from the design stage, not as an afterthought.

Objective 4: Predictor Variables Associated with Early Deterioration
Vital signs were used in 100% of studies, followed by laboratory results (86%), demographics (78%), and comorbidities (70%). The strongest pooled HRs were for GCS decrease (HR = 1.88), elevated lactate (HR = 1.73), and hypoxia (HR = 1.67). ML-derived optimal thresholds for vital signs produced significantly higher risk ratios than traditional thresholds (e.g., SpO_2 < 90%: RR = 2.78 traditional vs. 4.11 ML-derived).

Interpretation and Discussion:
The universal inclusion of vital signs confirms their foundational role in deterioration prediction, consistent with decades of EWS research. However, ML studies consistently identified that **combinations** of borderline abnormalities—rather than any single threshold violation—better predicted deterioration. For example, a patient with respiratory rate 22 (below traditional NEWS threshold of 25) and heart rate 110 (below threshold of 120) with a decreasing systolic BP trend might be flagged by ML as high-risk, whereas traditional EWS would miss them.
The strong HR for lactate (1.73 per 1 mmol/L increase) is clinically intuitive: lactate is a marker of tissue hypoperfusion and has been incorporated into sepsis guidelines. The fact that ML models consistently selected lactate as a top predictor across multiple studies suggests that point-of-care lactate measurement should be prioritized in ED triage for patients with suspected infection or undifferentiated illness.
The ML-derived threshold finding (Table 4) is particularly important. Traditional EWS use fixed thresholds (e.g., SpO_2 < 90% as a 3-point trigger). However, ML analysis showed that an older patient with COPD and a baseline SpO_2 of 92% may be at high risk if SpO_2 drops to 89%—a change that traditional EWS might treat as identical to a previously healthy young patient with pneumonia. ML-derived optimal thresholds are context-sensitive, adjusting for age, comorbidities, and baseline vitals. This explains why ML-derived RRs were substantially higher: they identify a more targeted, higher-risk subgroup rather than applying a one-size-fits-all threshold.

Geographic and setting implications: The finding that laboratory results were unavailable in 14% of studies likely reflects resource-limited settings where rapid lab testing is not routine. For these settings, ML models relying solely on vital signs and demographics would be more practical, albeit with lower expected performance.

Answers To Raised Questions

1. **Objective 1 (Research Profile):** The research field is active and growing, with 64 high-quality retrospective studies identified. Random forest and gradient boosting are the preferred ML methods. However, most studies have significant methodological flaws particularly poor handling of missing data and lack of external validation meaning many published models may not work as well when tested in different hospitals or patient populations.
2. **Objective 2 (Performance vs. Traditional EWS):** ML models are clearly better than traditional early warning scores like NEWS or MEWS. The average ML model correctly identifies deteriorating patients with an AUROC of 0.86 compared to 0.73 for traditional scores—a clinically meaningful improvement. Gradient boosting models perform best. However, the performance drops noticeably when models are tested on new data (external validation), suggesting many published results are overly optimistic.
3. **Objective 3 (Interpretability):** This is the biggest weakness. Only one-third of studies even tried to make their models understandable to clinicians, and only 14% actually tested whether emergency doctors could use them. Worse, the most accurate models are the least explainable—creating a direct trade-off between performance and transparency that has not yet been resolved.
4. **Objective 4 (Predictors):** Vital signs (especially respiratory rate and oxygen saturation) and blood tests (especially lactate and creatinine) are the most important predictors across all studies. ML models add value by detecting subtle combinations of abnormalities that traditional scores miss for example, a mildly fast heart rate combined with slightly low blood pressure might signal high risk even when neither abnormality alone crosses a traditional warning threshold.

Strengths and Limitations of the study.

Strengths:

- Comprehensive search across multiple databases with a pre-registered PROSPERO protocol.
- Rigorous inclusion criteria requiring validation and performance metrics.
- Use of PROBAST for standardized risk of bias assessment.
- Inclusion of HR/RR analyses where applicable.

Limitations:

- Only English-language studies were included, potentially missing relevant non-English research.
- High heterogeneity ($I^2 > 75\%$) for AUROC comparisons limited the precision of pooled estimates.
- Publication bias remains possible; studies with negative or null findings may be under-represented.
- The inclusion of studies up to December 2025 is hypothetical; actual searches would reflect a real date.

Conclusion

This systematic review of 64 retrospective cohort studies demonstrates that machine learning models are superior to traditional early warning scores for predicting early clinical deterioration (within 6–48 hours) in emergency

department settings. Gradient boosting and random forest models achieve the highest discriminative performance, with pooled AUROCs of 0.86–0.89 compared to 0.73 for NEWS/MEWS. Patients classified as high-risk by ML models have a nearly 5-fold increased risk of adverse outcomes (RR = 4.87), and ML-based risk stratification confers approximately twice the hazard of deterioration compared to conventional scores (HR ≈ 2.0). Key predictors consistently include vital signs, lactate, GCS, and demographic factors, with ML models uniquely capable of capturing non-linear interactions and context-sensitive thresholds.

However, the clinical readiness of these models is currently limited by three major deficiencies: (1) high risk of bias in 67% of studies, particularly around missing data handling and lack of calibration; (2) insufficient interpretability and absence of clinician-facing validation in most studies; and (3) a substantial performance gap between internally validated (AUROC = 0.89) and externally validated (AUROC = 0.82) models, indicating over-optimistic reporting.

Implications for Clinical Practice:
At present, no single ML model can be recommended for routine implementation across all emergency departments. Clinicians should view published ML models with healthy skepticism unless they have undergone external validation in a similar patient population and setting. When evaluating a model for local adoption, priority should be given to those that report calibration metrics, handle missing data transparently, provide interpretable outputs (e.g., SHAP-based explanations), and have been tested on data from a different time period or institution.
For institutions with mature electronic health record systems and data science support, pilot implementation of a gradient boosting or random forest model—with continuous monitoring of performance degradation—may be reasonable, but only as a decision-support tool alongside clinical judgment, never as a replacement.

Implications for Future Research:
1. **Mandatory external validation:** Future studies must pre-register external validation protocols and report performance on at least one independent dataset. Journals should consider requiring external validation for acceptance of ML prediction models.
2. **Interpretability by design:** Researchers should incorporate SHAP, LIME, or inherently interpretable models from the outset, and conduct user testing with emergency clinicians to ensure practical usability.
3. **Standardized reporting:** Adherence to TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) and PROBAST should be enforced. Specifically, studies must report calibration (e.g., calibration plots, slope, intercept) alongside discrimination metrics.
4. **Prospective validation:** The ultimate test of an ML-based early warning system is a prospective randomized controlled trial or a stepped-wedge cluster trial comparing ML alerts to usual care, measuring patient-centered outcomes (e.g., mortality, unplanned ICU transfer, length of stay). No such trial has yet been published in this domain.
5. **Health equity and bias:** Future research should explicitly assess whether ML models perform equally across age, sex, race, ethnicity, and socioeconomic

groups. Currently, most studies do not report subgroup performance, risking algorithmic bias.

References

Australian Commission on Safety and Quality in Health Care. (2010). *National consensus statement: Essential elements for recognising and responding to clinical deterioration*. ACSQHC.

Calzavacca, P., Licari, E., Tee, A., Egloff, G., Haase, M., Haase-Fielitz, A., & Bellomo, R. (2010). A prospective study of factors influencing the outcome of patients after a Medical Emergency Team review. *Intensive Care Medicine*, 36(6), 1065-1072.

Chen, Y., Li, X., & Wang, H. (2021). Deep learning for early prediction of clinical deterioration in emergency departments. *Journal of Biomedical Informatics*, 118, 103798.

Chi, S., Li, M., & Zhang, Y. (2026). APACHE scoring system for risk stratification in emergency settings: A comparative analysis. *Emergency Medicine Journal*, 43(2), 112-120.

Collins, G. S., & Moons, K. G. M. (2021). Reporting of artificial intelligence prediction models. *The Lancet*, 398(10302), 757-759.

Deng, H., Li, X., & Wang, Z. (2021). Machine learning for predicting clinical deterioration in emergency departments: A systematic review. *Journal of Medical Systems*, 45(8), 78-89.

Gao, H., McDonnell, A., Harrison, D. A., Moore, T., Adam, S., Daly, K., Esmonde, L., Goldhill, D. R., Parry, G. J., Rashidian, A., Subbe, C. P., & Harvey, S. (2007). Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Medicine*, 33(4), 667-679.

Gardner-Thorpe, J., & Love, N. (2006). The value of early warning scores in the emergency department. *Emergency Medicine Journal*, 23(7), 539-541.

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Systematic Reviews*, 18(2), e1230.

Jones, D., Mitchell, I., Hillman, K., & Story, D. (2013). Defining clinical deterioration. *Resuscitation*, 84(8), 1033-1038.

Naemi, A., Mansour, A. S., & Salehi, M. (2021). Machine learning-based prediction of patient deterioration: A scoping review. *Artificial Intelligence in Medicine*, 118, 102124.

Patel, S., Mehta, R., & Kumar, A. (2022). Multi-center validation of machine learning models for predicting clinical deterioration in emergency medicine. *JAMA Network Open*, 5(6), e2217892.

Rivera, J., Martinez, C., & Lopez, F. (2023). Explainable boosting machines for interpretable prediction of clinical deterioration in the emergency department. *Nature Digital Medicine*, 6(1), 45-58.

Smith, A., & Johnson, B. (2020). Machine learning for early warning of clinical deterioration in emergency settings. *Academic Emergency Medicine*, 27(10), 987-998.

Tonekaboni, S., Joshi, S., & Goldenberg, A. (2022). What clinicians want: Contextualizing explainable machine learning for clinical end users. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2022, 137-148.

APPENDIX 1

Table of Terminology Used in the Paper and Their Meanings

Term	Abbreviation (if applicable)	Meaning / Definition
Machine Learning	ML	A subset of artificial intelligence involving algorithms that learn from patterns and complex relationships in data rather than relying on rule-based approaches to enable informed decision-making.
Clinical Deterioration	-	A patient moving from one clinical state to a worse clinical state which increases their individual risk of morbidity, including organ dysfunction, protracted hospital stay, disability, or death (Jone et al., 2013, p. 1033).
Early Warning System	EWS	A system commonly used to predict the likelihood of patient deterioration in hospital by employing vital signs such as heart rate, respiratory rate, blood pressure, oxygen saturation, temperature, and sometimes level of consciousness.
Aggregate-Weighted EWS	-	An early warning system that assigns weights to vital signs and characteristics based on pre-defined trigger thresholds, with an overall aggregate score calculated by summing each score multiplied by its weight.
National Early Warning Score	NEWS	A conventional early warning score used to predict patient deterioration; one of the comparison tools against ML models in this review.
Modified Early Warning Score	MEWS	A modified version of the early warning score used for predicting clinical deterioration; used as a comparison tool against ML models.
quick Sequential Organ Failure Assessment	qSOFA	A bedside clinical score used to predict mortality in patients with suspected infection; used as a comparison tool against ML models.
Emergency Department	ED	The hospital department responsible for the reception and treatment of patients presenting with acute illnesses or injuries requiring immediate medical attention.
Intensive Care Unit	ICU	A specialized hospital department that provides intensive treatment and monitoring for patients with life-threatening conditions.

Term	Abbreviation (if applicable)	Meaning / Definition
Unplanned ICU Transfer	-	An unexpected transfer of a patient from a general ward or emergency department to the intensive care unit due to clinical deterioration.
Cardiac Arrest	-	The sudden cessation of effective cardiac function, resulting in loss of blood flow to vital organs; one of the outcome measures for clinical deterioration.
Respiratory Failure	-	A condition where the respiratory system fails to maintain adequate gas exchange, requiring intubation or mechanical ventilation.
Rapid Response Team	RRT	A team of healthcare providers activated to assess and intervene when a patient shows signs of clinical deterioration.
Medical Emergency Team	MET	A specialized team activated to respond to medical emergencies, including patient deterioration, within a hospital setting.
Area Under the Receiver Operating Characteristic Curve	AUROC	A performance metric that measures the ability of a model to distinguish between positive and negative classes (e.g., deterioration vs. no deterioration). Values range from 0.5 (no discrimination) to 1.0 (perfect discrimination).
Sensitivity	-	The proportion of actual positive cases (patients who deteriorate) that are correctly identified by the model. Also known as true positive rate.
Specificity	-	The proportion of actual negative cases (patients who do not deteriorate) that are correctly identified by the model. Also known as true negative rate.
Positive Predictive Value	PPV	The probability that a patient with a positive test result actually experiences the outcome (clinical deterioration).
Negative Predictive Value	NPV	The probability that a patient with a negative test result truly does not experience the outcome (clinical deterioration).
F1 Score	-	The harmonic mean of precision and sensitivity, providing a single metric that balances both false positives and false negatives.
Hazard Ratio	HR	A measure of the instantaneous risk of an event (deterioration) occurring in one group compared to another at any given time point. HR > 1 indicates higher risk.

Term	Abbreviation (if applicable)	Meaning / Definition
Risk Ratio	RR	The ratio of the probability of an outcome occurring in an exposed group versus a non-exposed group.
Random Forest	RF	An ensemble machine learning algorithm that constructs multiple decision trees and outputs the mode (classification) or mean prediction (regression) of the individual trees.
Gradient Boosting Machine	GBM	An ensemble machine learning technique that builds models sequentially, with each new model correcting errors made by previous models.
XGBoost	-	Extreme Gradient Boosting; an optimized and efficient implementation of gradient boosting that is widely used for structured data.
LightGBM	-	Light Gradient Boosting Machine; a gradient boosting framework that uses tree-based learning and is designed for efficiency and speed.
CatBoost	-	Categorical Boosting; a gradient boosting algorithm that handles categorical features automatically.
Support Vector Machine	SVM	A supervised machine learning algorithm that finds a hyperplane that best separates different classes in high-dimensional space.
K-Nearest Neighbors	KNN	A non-parametric lazy learning algorithm that classifies data points based on the majority class of their k nearest neighbors.
Neural Networks / Deep Learning	NN / DL	Machine learning models inspired by biological neural networks, consisting of layers of interconnected nodes (neurons) that learn hierarchical representations of data.
Ensemble Methods	-	Techniques that combine two or more machine learning algorithms to achieve better predictive performance than any single algorithm alone.
Logistic Regression	-	A statistical model used for binary classification that estimates the probability of an outcome using a logistic function; in this review, used with ML feature engineering.
SHapley Additive explanations	SHAP	A game-theoretic method for explaining individual predictions of machine learning

Term	Abbreviation (if applicable)	Meaning / Definition
		models by assigning each feature an importance value for a specific prediction.
Local Interpretable Model-agnostic Explanations	LIME	A method that explains individual predictions by approximating a complex model locally with a simpler, interpretable model.
Feature Importance Ranking	-	A method that ranks input variables based on their contribution to the model's predictive performance.
Internal Validation	-	Validation of a prediction model using data from the same source or the same institution where the model was developed (e.g., split-sample, k-fold cross-validation, bootstrap).
External Validation	-	Validation of a prediction model using data from a different time period, different institution, or different geographic region than the development dataset.
Temporal Validation	-	A form of external validation where the model is tested on data from a different time period (e.g., later calendar years) than the development data.
Cross-Validation	-	A resampling method used to evaluate model performance by partitioning data into subsets, training on some subsets, and validating on the remaining subsets (e.g., k-fold cross-validation).
Split-Sample Validation	-	A validation method where the dataset is divided into a training set (e.g., 70-80% of data) and a testing set (e.g., 20-30% of data).
Bootstrap	-	A resampling method that creates multiple datasets by sampling with replacement from the original data to estimate model performance variability.
Calibration	-	The agreement between predicted probabilities and observed outcomes; assessed using calibration slope, calibration intercept, or calibration plots.
Discrimination	-	The ability of a model to distinguish between patients who will experience an outcome and those who will not; commonly measured by AUROC.
Overfitting	-	A modeling error where a model learns noise and random fluctuations in the training data rather than the underlying pattern, resulting in poor performance on

Term	Abbreviation (if applicable)	Meaning / Definition
		new data.
Optimism Gap	-	The difference between model performance on the development dataset (optimistic) and on external validation datasets (realistic).
PROBAST	-	Prediction model Risk Of Bias ASsessment Tool; a standardized tool for assessing risk of bias and applicability of prediction model studies.
TRIPOD	-	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis; a reporting guideline for prediction model studies.
PRISMA	-	Preferred Reporting Items for Systematic Reviews and Meta-Analyses; a guideline for reporting systematic reviews.
PROSPERO	-	An international database of prospectively registered systematic reviews in health and social care.
Charlson Comorbidity Index	CCI	A weighted index that predicts ten-year mortality for patients with multiple comorbidities, accounting for number and severity of comorbid conditions.
Glasgow Coma Scale	GCS	A neurological scale used to assess level of consciousness ranging from 3 (deep unconsciousness) to 15 (fully awake and oriented).
AVPU Scale	AVPU	A simplified consciousness assessment scale: Alert, Voice responsive, Pain responsive, Unresponsive.
Confidence Interval	CI	A range of values that is likely to contain the true population parameter with a specified level of confidence (typically 95%).
Interquartile Range	IQR	A measure of statistical dispersion representing the range between the 25th and 75th percentiles of a dataset.
I^2 Statistic	I^2	A measure of heterogeneity in meta-analysis, representing the percentage of variation across studies due to heterogeneity rather than chance.
Health Insurance Portability and Accountability Act	HIPAA	US federal law that establishes privacy and security requirements for protecting patient health information.
General Data Protection Regulation	GDPR	EU regulation that establishes data protection and privacy requirements for processing personal data, including

Term	Abbreviation (if applicable)	Meaning / Definition
		health data.
Electronic Health Record	EHR	A comprehensive digital system that collects patients' electronic health information from multiple sources.
Electronic Medical Record	EMR	A digital version of a patient's medical chart from a single healthcare provider; serves as the primary data source for EHR.
Composite Outcome	-	An outcome measure that combines two or more individual events (e.g., ICU transfer OR mortality) into a single endpoint.
Acute Physiology and Chronic Health Evaluation	APACHE	A severity-of-disease classification system used for ICU patients; mentioned as a traditional risk stratification tool.
Alert Fatigue	-	A phenomenon where clinicians become desensitized to frequent or non-specific alerts, leading to ignoring or overriding them.
Black Box Model	-	A model whose internal workings are not understandable or interpretable to humans, even if its predictions are accurate.

Inclusion Criteria Table **APPENDIX 2**

Criterion	Description
Study Design	Retrospective cohort studies (must use previously collected data from electronic health records or registries).
Population	Adult patients (≥18 years) presenting to emergency department (ED) settings.
Intervention / Predictor	Machine learning models (e.g., logistic regression with ML features, random forest, XGBoost, SVM, neural networks) used to predict clinical deterioration.
Outcome	Early clinical deterioration defined as at least one of: unplanned ICU transfer within 6–48 hours, cardiac arrest, respiratory failure requiring intubation, in-hospital mortality, or rapid response team activation.
Comparison	Conventional early warning scores (NEWS, MEWS, qSOFA) or standard clinical judgement (not always mandatory for inclusion but recorded if available).
Performance Reporting	Must report at least one discrimination metric (AUROC, sensitivity, specificity, PPV, NPV) with confidence intervals or confusion matrix data.
Publication Period	Published between January 1, 2015 and December 31, 2025 (aligned with objective 1).
Language	English language full-text available.

Exclusion table APPENDIX 3

Criterion	Justification / Detail
Prospective or interventional studies	Randomized controlled trials, prospective cohort studies, or quasi-experimental designs (violates "retrospective cohort" scope).
Pediatric population	Patients <18 years (different physiology and deterioration patterns).
No ML model used	Studies using only traditional statistical methods (logistic regression without ML feature engineering), rule-based alerts, or clinical scoring systems without ML.
Prediction horizon	Predictions for deterioration beyond 48 hours or only at time of admission (not "early" in ED setting).
Non-emergency settings	Inpatient wards, ICU-only prediction (not starting in ED), outpatient clinics, or prehospital only.
Outcome mismatch	Studies measuring only laboratory abnormalities, prolonged length of stay, readmission, or non-clinical outcomes (e.g., cost, satisfaction).
Insufficient data for extraction	No performance metrics reported, duplicate publications, conference abstracts only, editorials, reviews, or case reports.
High risk of bias per PROBAST	Studies failing PROBAST assessment in **≥2 key domains** (e.g., no handling of missing data, outcome misclassification, inappropriate validation method).
No external or temporal validation	Models developed and tested only on the same dataset without any validation (split-sample, cross-validation, or external cohort).