Majalah Kesehatan Indonesia
Volume 7, Issue 3, 2026, pages 169–192
P-ISSN 2745-6498, E-ISSN 2745-8008

Early Warning System through Artificial Intelligence in Intensive Care: A Diagnostic Meta-Analysis of Mortality, Deterioration of Consciousness, and Neurological Outcome

Stefanus Erdana Putra[1*)], Baarid Luqman Hamidi[2], Benedictus[3], Muhammad Hafizhan[4], Tyasno Koeshermanto[5], Retnaningsih[6], Abdulloh Machin[7]

[1] Department of Neurology, Faculty of Medicine, Universitas Kristen Duta Wacana
[2] Department of Neurology, Faculty of Medicine, Universitas Sebelas Maret
[3] Department of Pathological Anatomy, Faculty of Medicine, Universitas Indonesia
[4] Profesi Ners, Keperawatan, Pltekkes Kemenkes Surakarta Department of Neurology, Ananda Babelan General Hospital, Bekasi
[5] Medical Emergency Unit, Hj Anna Lasmanah Regional General Hospital, Banjarnegara
[6] Department of Neurology, Faculty of Medicine, Universitas Diponegoro
[7] Department of Neurology, Faculty of Medicine, Universitas Airlangga

ARTICLE INFO

Article history:

Received 24 April 2026
Accepted 05 July 2026
Published 13 July 2026

Keyword:

Artificial Intelligence
Deterioration Of Consciousness
Early Warning Systems
Mortality
Neurocritical Care
Neurological Outcome

**) corresponding author*

dr. Stefanus Erdana Putra, Sp.N.
Jalan Dr. Wahidin Sudirohusodo No. 5-25, Kotabaru, Kecamatan Gondokusuman, Kota Yogyakarta, Daerah Istimewa Yogyakarta – Indonesia 55224
Email: stefanuserdanaputra@staff.ukdw.ac.id
DOI: 10.47679/makein. 2026347

ABSTRACT

Prognostication in neurocritical care remains challenging due to patient heterogeneity and the limitations of static conventional scores. This systematic review and meta-analysis evaluated the diagnostic accuracy of artificial intelligence (AI) models as retrospective predictive and prognostic frameworks, assessing their potential utility as clinical early warning systems (EWS). Conducted via PubMed, ScienceDirect, and SCOPUS per PRISMA guidelines, the study evaluated deterioration of consciousness, mortality, and neurological outcomes. Out of 20 included studies, 18 were meta-analyzed. For mortality prediction, AI demonstrated a pooled sensitivity of 0.569, specificity of 0.810, diagnostic odds ratio (DOR) of 9.620, and area under the curve (AUC) of 0.698. In predicting deterioration in consciousness, AI achieved a pooled sensitivity of 0.648, specificity of 0.901, DOR of 38.346, and AUC of 0.796. For neurological outcomes, the pooled sensitivity was 0.864, specificity 0.865, DOR 45.566, and AUC 0.858. AI models demonstrate high accuracy in predicting long-term neurological outcomes and acceptable performance in predicting short-term deterioration in consciousness, but remain highly limited and offer no clear advantage over conventional scoring for ICU mortality. These retrospective findings provide a robust foundation for future prospective designs, though widespread clinical integration is not yet warranted.

This open access article is under the CC–BY-SA license.

INTRODUCTION

Acute neurological diseases represent a significant and growing burden in critical care, frequently necessitating admission to the intensive care unit (ICU). Patients with severe traumatic brain injury, stroke, status epilepticus, or subarachnoid hemorrhage require highly complex and nuanced management to mitigate primary brain injury, prevent secondary brain injury, and optimize the potential for neurological recovery. The inherent fragility of the brain and its susceptibility to various systemic insults make meticulous monitoring and timely intervention paramount for improving patient outcomes (Rajajee et al., 2023).

However, prognostication in neurological ICU patients remains a challenging endeavor. This difficulty stems from multiple factors, including the inherent heterogeneity of the neurological injuries, the dynamic nature of the neurological function, and the profound impact of extracranial comorbidities. Accurately predicting future neurological outcomes and mortality is crucial for guiding therapeutic intensity, facilitating family discussions, and making complex, often difficult, decisions regarding the withdrawal of life-sustaining therapy (WLST). Inaccurate or delayed prognostication can lead to suboptimal resource utilization, prolonged hospital stays, increased healthcare costs, and significant emotional distress for patients and their families (Finley Caulfield et al., 2022).

A striking observation from a recent retrospective study in the United States underscores these limitations: only 9% of patients underwent at least one neurodiagnostic test, and 16% of deaths occurred on or after the third day of admission (Elmer et al., 2023). This finding suggests a potential underutilization of diagnostic tools and a delay in identifying patients at high risk of deterioration or those who might benefit from earlier, more aggressive interventions or earlier discussions about palliation. Such delays can result in missed opportunities for therapeutic windows and prolonged care for patients with poor prognoses, raising ethical and economic concerns.

Historically, neuro-prognostication has relied heavily on conventional scoring systems, such as the Glasgow Coma Scale (GCS), the Acute Physiology and Chronic Health Evaluation (APACHE II), and the Sequential Organ Failure Assessment (SOFA). Although these established clinical scores provide valuable guidance, they have major limitations for real-time prognostication. They rely on static, intermittent clinical and laboratory inputs, failing to capture the dynamic, rapidly evolving physiology of critically ill patients (Thakur et al., 2023). Furthermore, conventional scoring systems lack the computational capacity to integrate diverse data streams, such as continuous electroencephalography (EEG), high-dimensional neuroimaging, and detailed electronic health records (EHR), into personalized, real-time predictions (Al-Mufti et al., 2019).

Rapid advancements in artificial intelligence (AI), particularly in machine learning (ML) and deep learning algorithms, have presented a transformative opportunity to address these prognostic challenges in recent years. AI models can analyze vast datasets, identify intricate patterns, and learn complex relationships that may not be apparent to human observers (Yuan et al., 2025). When applied to critical care data, these algorithms can potentially generate early warning systems (EWS) that alert clinicians to impending physiological deterioration, changes in consciousness, or adverse neurological outcomes, thereby enabling proactive rather than reactive care (Yap et al., 2025).

The application of AI in healthcare, particularly in the ICU, has garnered significant attention because of its potential to enhance predictive accuracy, optimize resource allocation, and improve patient safety (Abe et al., 2025). For neurological patients, in whom subtle changes can have profound implications, an AI-driven EWS could be particularly beneficial. Such systems can integrate real-time physiological parameters, electronic health record data, and even neuroimaging features to provide continuous risk assessments, flagging patients who require immediate attention or warrant re-evaluation of their management plans (Cherifa & Pirracchio, 2019; Kim et al., 2024).

To evaluate these models, this study structures its analysis around three core outcomes: acute deterioration of consciousness, long-term functional neurological outcomes, and ICU mortality. Rather than analyzing these endpoints in parallel, this meta-analysis defines a clear, clinically logical relationship among them. Acute deterioration of consciousness is positioned as a proximal, early bedside indicator of acute neurological decline. In contrast, long-term functional neurological outcomes (assessed at 3 to 12 months) and ICU mortality are positioned as distal, final endpoints that reflect the cumulative impact of primary and secondary brain injury. Evaluating these three outcomes together provides a comprehensive synthesis of AI's predictive capabilities across different clinical horizons.

Despite growing interest, there is currently no quantitative synthesis evaluating the diagnostic accuracy of AI across these distinct clinical horizons and input modalities. Existing literature is diverse and highly heterogeneous, leaving several key clinical and methodological gaps unresolved. Specifically, the overall diagnostic accuracy of these algorithms remains unclear; their performance across different clinical endpoints has not been compared; and the extent to which performance varies with input data modalities (EHR, EEG, and neuroimaging) is unknown. Most importantly, the clinical readiness of these retrospective models for real-world deployment remains unestablished (Kumar et al., 2025).

To address these gaps, this systematic review and diagnostic meta-analysis was designed with a clear, operational objective: to calculate the pooled diagnostic and predictive performance (sensitivity, specificity, and diagnostic odds ratios) of AI models in neurocritical care, and to systematically explore source-level heterogeneity based on clinical outcomes, patient populations, and input modalities. By doing so, this study aims to provide a robust, evidence-based assessment of AI's current capabilities, its limitations, and its clinical readiness for integration into real-time early warning systems.

METHODS

This study was a systematic review and diagnostic meta-analysis evaluating the predictive utility of AI for mortality, deterioration in consciousness, and neurological outcomes among patients admitted to the ICU. The methodology of this research strictly adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Page et al., 2021), ensuring comprehensive and transparent reporting of the review process. The "EWS" term is a broad descriptor for any AI tool that provides a prospective risk assessment to aid clinical decision-making. Since the prediction horizons of the included studies might vary from real-time or short-term alerts for imminent physiological decline to longer-term prognostication for outcomes at hospital discharge or at 3 to 6 months, to address this heterogeneity in the analysis, findings would be stratified or discussed by prediction window (e.g., short-term vs. long-term) to provide a more nuanced understanding of the results.

Search Strategy

A systematic search was performed in three major electronic databases: PubMed, ScienceDirect, and Scopus. To ensure broad literature capture, the search was designed to identify relevant published journal articles from their inception to May 2026, with no restrictions on publication period. The research question for this systematic review and diagnostic meta-analysis was framed using the PICO framework, i.e., P (patient/population): patients in the neurocritical care unit, including adult and pediatric populations with various acute neurological diseases; I (intervention): machine learning or deep learning algorithms evaluated as predictive or prognostic models; C (comparison): reference standards, including established clinical scores (APACHE, SOFA, GCS) or validated functional scales (Glasgow Outcome Scale (GOS), Glasgow Outcome Scale-Extended (GOS-E), Modified Rankin Scale (mRS)); O (outcome): the predictive accuracy of AI models for three key outcomes: mortality, deterioration of consciousness, and overall neurological outcome. The search strategy combined Medical Subject Headings (MeSH) terms and free-text keywords

related to artificial intelligence and neurological outcomes in critical care settings. The primary keywords used were: "artificial intelligence," "mortality," "deterioration of consciousness," "neurological outcome," and "intensive care unit."

Eligibility Criteria

This systematic review and diagnostic meta-analysis included studies of various designs, including cohort designs (prospective or retrospective) and randomized controlled trials (RCTs) validating an AI or machine learning model for at least one of the target outcomes, studies reporting raw diagnostic performance data (true positives, false positives, false negatives, true negatives) or providing sensitivity and specificity metrics alongside exact sample sizes, enabling the reconstruction of 2x2 contingency tables, and studies validating models on human clinical data. The review focused on English-language studies that investigated the use of AI or machine learning models for prognostication and reported at least one of the predefined outcomes: mortality, deterioration of consciousness, or neurological outcome in patients with acute neurological diseases in the intensive care unit. Given the field's heterogeneity, studies involving adult and pediatric patients in neurological ICUs were included to synthesize the full breadth of available evidence. Studies focusing solely on AI development, purely technical papers without clinical validation, or those unrelated to neurological prognostication in the ICU were excluded from the review.

The target clinical outcomes were defined using clear diagnostic boundaries to ensure clinical homogeneity. Mortality was defined as all-cause mortality within defined clinical horizons and validated using clinical instruments such as ICU mortality, in-hospital mortality, short-term (28- or 30-day) mortality, or long-term (90-day to 1-year) mortality. Deterioration of consciousness was defined as an acute clinical decline in mental status, arousal, or cognitive engagement during the ICU stay and validated using clinical instruments such as dropping Glasgow Coma Scale (GCS) scores, emerging delirium (positive Confusion Assessment Method for the Intensive Care Unit or CAM-ICU), changes in the Richmond Agitation-Sedation Scale (RASS), or suppression patterns on EEG. Otherwise, the neurological outcome was defined as long-term functional, physical, and cognitive recovery, typically assessed 3 to 12 months post-injury and validated using clinical instruments such as GOS, GOS-E, mRS, or Clinical Performance Category (CPC).

Study Selection Process

The study selection process followed a systematic, multiphase approach, as depicted in the PRISMA flowchart (Figure 1). Initially, all records retrieved from the electronic database search were compiled, and duplicate entries were removed (n = 11). Subsequently, four authors screened the titles and abstracts of the remaining articles (n = 60) independently based on predefined keywords and eligibility criteria. If a consensus could not be reached, disputes were adjudicated by senior investigators within the research team to ensure objectivity.

In the eligibility phase, the full texts of potentially relevant articles (n = 27) were retrieved and thoroughly assessed against the inclusion and exclusion criteria. The reasons for exclusion at this stage included studies not associated with artificial intelligence (n = 19), studies not related to neurological outcomes (n = 14), and studies that lacked a clinical validation cohort or a non-prognostic design (n = 7). Studies identified only as abstracts were also excluded (n = 0). Ultimately, 20 studies met all inclusion criteria and were included in the qualitative synthesis and subsequent meta-analysis.

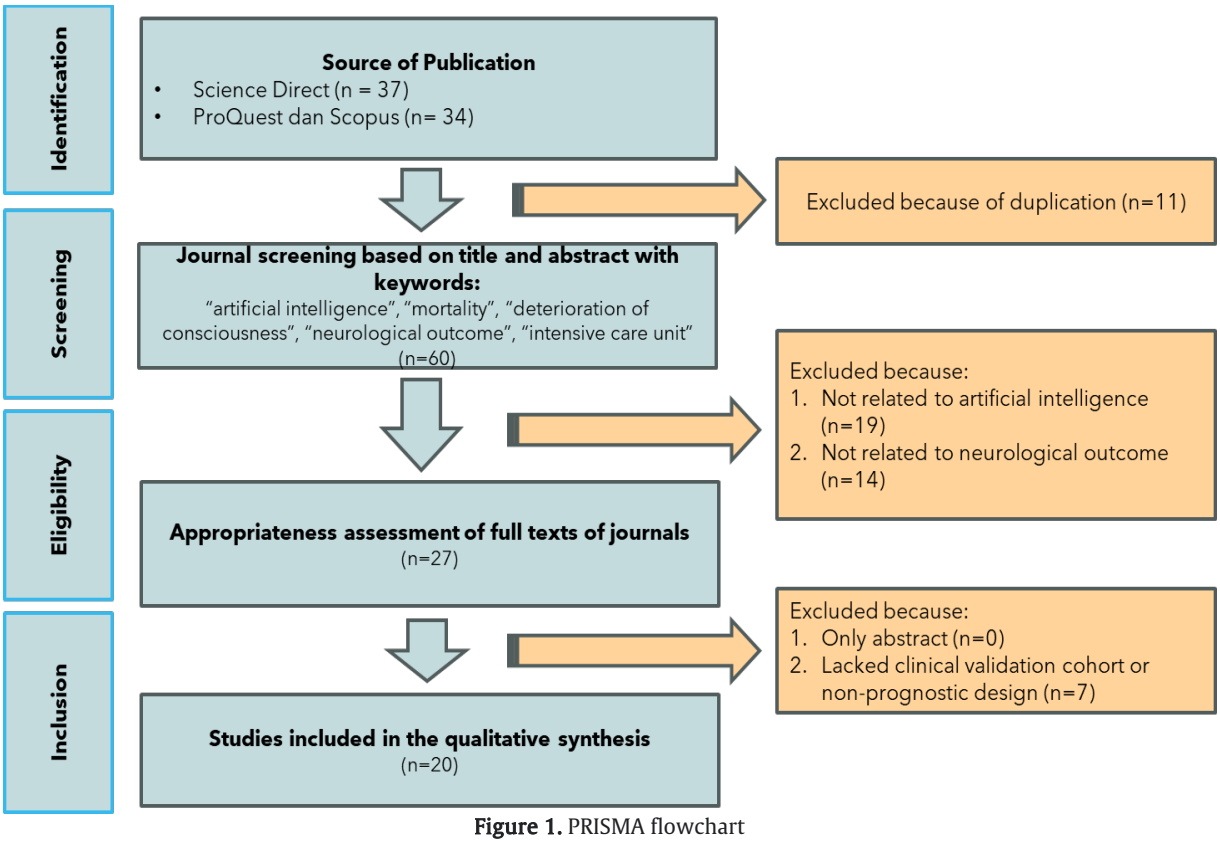


Figure 1. PRISMA flowchart

Data Extraction

Four authors independently extracted relevant data from each included study using a standardized data extraction form. The extracted information included study characteristics (e.g., first author, year of publication, country, and study design), patient demographics (sample size, age distribution (adult vs. pediatric), and specific neurological diagnoses), details of the AI model characteristics used (e.g., algorithm family, input data modalities (EHR, continuous EEG, or neuroimaging), and clinical prediction horizons), and diagnostic accuracy parameters (e.g., true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), optimal cut-off values, reference standards, and total event rates). For studies reporting multiple AI models evaluated on the same patient cohort, a strict selection hierarchy was applied to preserve statistical independence and prevent artificial inflation of weights. Only the primary model designated by the original authors or the model with the highest validation performance (highest AUC) was extracted for the main meta-analysis. A consensus was reached through consultation with three senior investigators within the research team, resolving any discrepancies in data extraction.

Quality Assessment and Risk of Bias

The risk of bias within the individual included studies was independently assessed by four reviewers using a standardized scoring system developed by the Cochrane Collaboration. This tool evaluates bias across several domains, including sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other potential sources of bias.

Data Analysis

The statistical analysis was conducted using R software (version 4.3.2). The primary diagnostic meta-analysis was performed using the meta package in R. For each primary study, raw 2x2 contingency tables were reconstructed. Standard descriptive pooling was conducted using the meta package (version 8.3-0), with a standard continuity correction of 0.5 applied to all zero-cell tables to stabilize estimation. A random-effects model was employed to calculate the pooled sensitivity, specificity, and diagnostic odds ratio (DOR), along with their respective 95% confidence intervals (CIs), accounting for anticipated heterogeneity. Additional diagnostic accuracy metrics, including the summary receiver operating characteristic (SROC) curve, Area Under the Curve (AUC), and positive and negative likelihood ratios, or the primary pooling of diagnostic accuracy parameters, were performed using Split Component Synthesis (SCS), implemented via the SCS meta function.

The SCS method represents a major advance over traditional bivariate random-effects models, particularly in the presence of high clinical and methodological heterogeneity (Furuya-Kanamori et al., 2021). The SCS method operates in two distinct phases. It first pools the study-specific diagnostic odds ratios on the natural-log scale (ln(DOR)) using the robust inverse-variance heterogeneity (IVhet) estimator. The IVhet model does not assume that true effects are normally distributed; instead, it uses the weights from a random-effects model to compute robust confidence intervals, effectively minimizing undercoverage and reducing bias (Schwarzer et al., 2015). Consequently, it then mathematically splits the pooled ln(DOR) into its component parts, logit sensitivity (logit(Se)) and logit specificity (logit(Sp)):

$$\ln(\mathrm{DOR}) = \mathrm{logit}(\mathrm{Se}) + \mathrm{logit}(\mathrm{Sp})$$

$$\mathrm{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

$$\mathrm{DOR} = \frac{\mathrm{Sensitivity}/(1-\mathrm{Sensitivity})}{(1-\mathrm{Specificity})/\mathrm{Specificity}}$$

Simulation studies demonstrate that the SCS method outperforms traditional bivariate models. Traditional models are prone to overestimation and undercoverage when between-study heterogeneity is high. In contrast, the SCS estimator maintains nominal coverage probabilities and provides highly reliable pooled estimates. While the resulting confidence intervals are often narrower than those from standard bivariate models, this reflects the superior mathematical efficiency and reduced bias of the SCS method, rather than artificial overconfidence (Furuya-Kanamori et al., 2021). Between-study heterogeneity was evaluated using the Cochran Q-test and quantified with the I^2 statistic; $I^2 > 50\%$ indicates substantial heterogeneity. Where appropriate ($k \geq 10$ studies), publication bias was evaluated using Egger's regression test for funnel plot asymmetry, combined with the ln(DOR) metric to prevent false-positive indications of bias.

RESULTS OF STUDY

Study Selection and Characteristics

The systematic database search retrieved 60 unique citations. Following a systematic search and rigorous screening, 20 studies were identified as meeting the predefined inclusion criteria for the qualitative synthesis. Of these, 18 studies provided sufficient data to include diagnostic accuracy in the quantitative meta-analysis. The included studies represented a total of 115,221 patients, with individual study sizes ranging from 25 to 24,885, reflecting the high diversity of the literature. A summary of the included studies, including their authors, year of publication, study location, participant characteristics, measured outcomes, AI models used, and underlying mechanisms, is presented in Table 1.

Diagnostic Accuracy of AI in Predicting Mortality

The quantitative synthesis of AI models for predicting ICU and short-term mortality included data from 10 clinical validation cohorts. The SCS pooling yielded a random-effects diagnostic odds ratio (DOR) of 9.620 (95% CI: 6.224–14.871), with an Area Under the Curve (AUC) in the summary receiver operating characteristic (SROC) of 0.698 (95% CI: 0.656–0.737), as shown in Figure 2. The pooled sensitivity was moderate at 0.569 (95% CI: 0.566–0.573), while the pooled specificity was substantially higher at 0.810 (95% CI: 0.807–0.813). From a clinical standpoint, this moderate sensitivity (0.569) is a major limitation for early warning systems. A sensitivity of approximately 57% means the model fails to identify more than 40% of patients who subsequently die. In an ICU setting, such a high false-negative rate is highly dangerous; it means a large proportion of high-risk patients remain undetected, potentially delaying lifesaving clinical

interventions. Conversely, the high specificity (0.810) indicates that the models are highly reliable at identifying patients at low risk of mortality, which is valuable for avoiding unnecessary diagnostic testing.

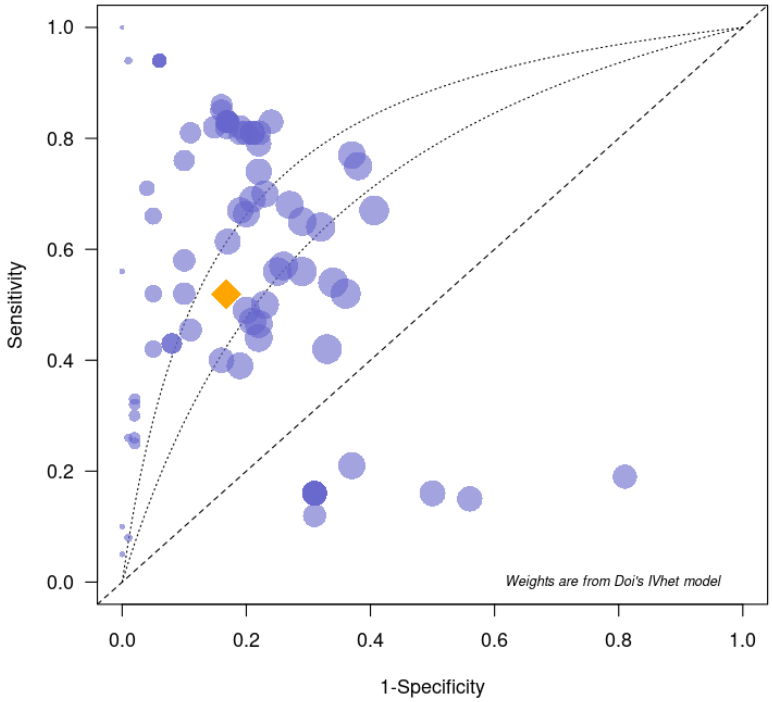


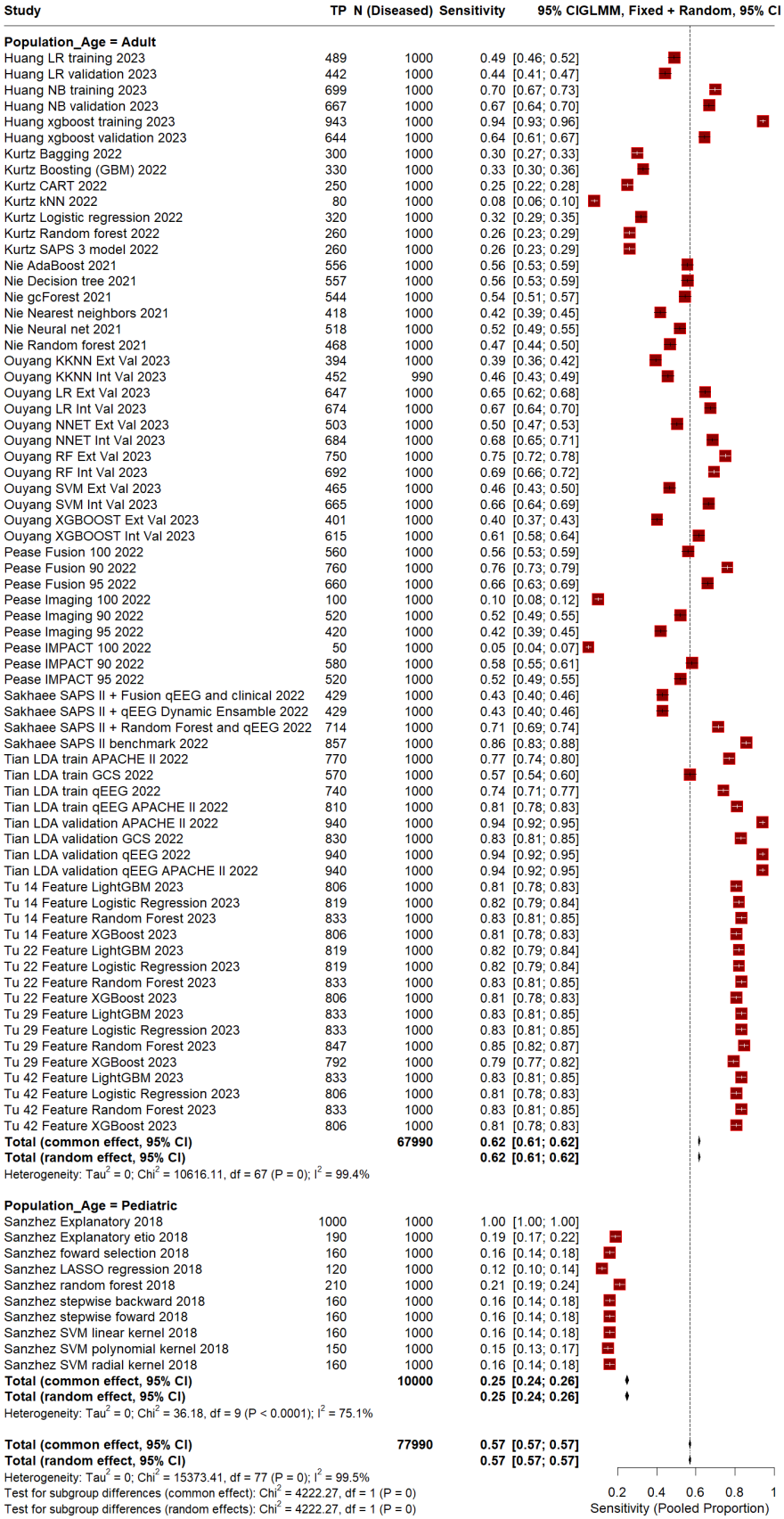
Figure 2. SROC curve in mortality prediction

Their pooled likelihood ratios further illuminate the clinical utility of these models.[1] The positive likelihood ratio (LR^+) was 3.086 (95% CI: 2.253–4.228), and the negative likelihood ratio (LR^-) was 0.579 (95% CI: 0.465–0.720). In diagnostic medicine, an LR^+ between 2 and 5 provides only a minimal-to-moderate shift in post-test probability, while an LR^- of 0.58 is insufficient to rule out mortality safely. These findings show that current AI models for mortality prediction are not yet sufficiently robust to serve as the sole basis for clinical decision-making. Instead, they must be used as adjunctive tools alongside established clinical criteria.

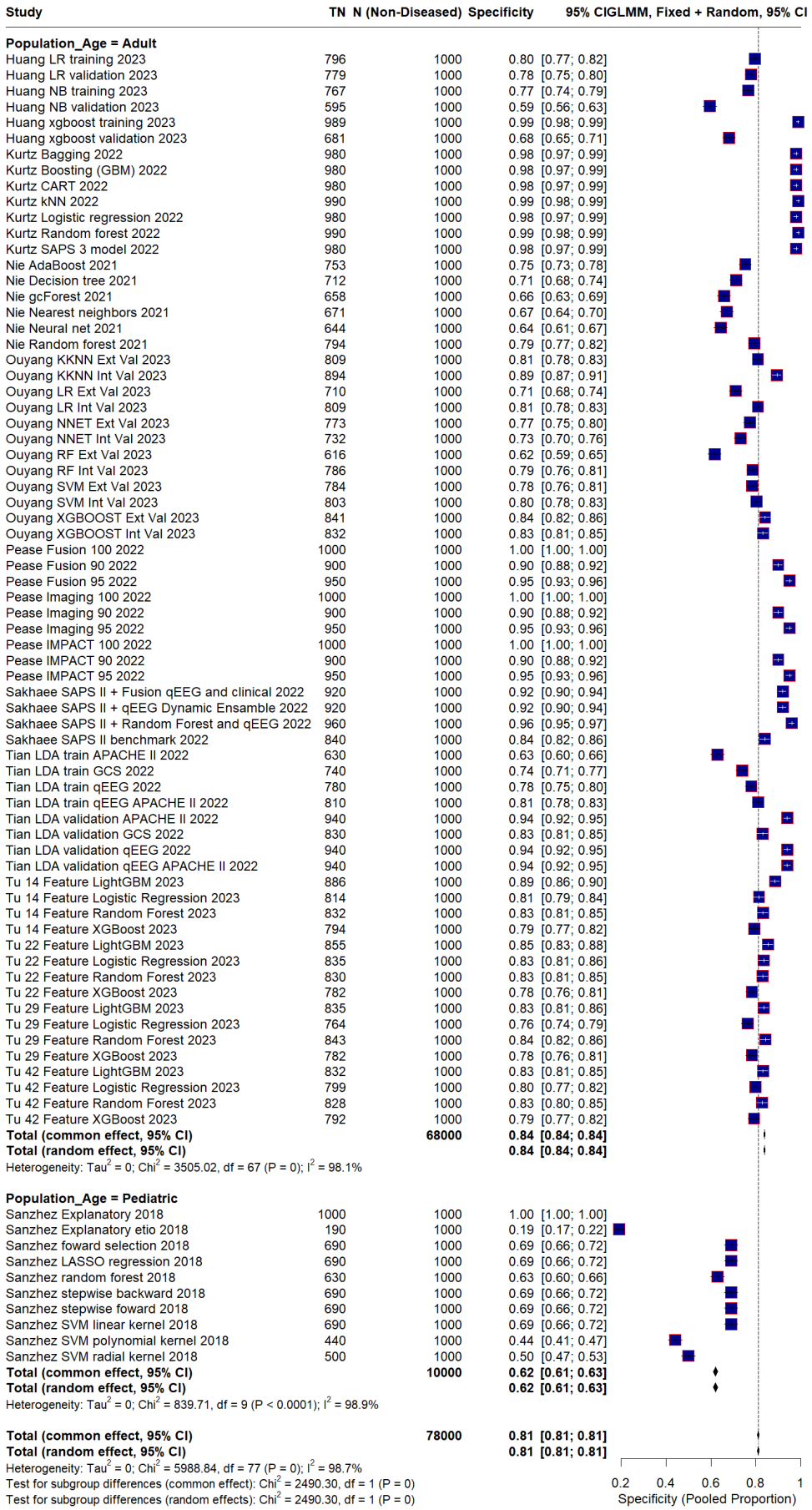
Substantial heterogeneity was observed across studies (I^2=99.4% for DOR, 99.5% for sensitivity, and 98.7% for specificity). Egger's test revealed significant funnel plot asymmetry (p = 0.0009), suggesting potential publication bias. To address heterogeneity, an initial subgroup analysis was conducted by patient age. For adult populations, the pooled sensitivity was 0.617 (95% CI, 0.613–0.621) and specificity was 0.838 (95% CI, 0.836–0.841), with a strong pooled DOR of 13.710 (95% CI, 10.091–18.627). In contrast, for pediatric populations, performance was substantially lower, with a pooled sensitivity of 0.247 (95% CI, 0.239–0.256), specificity of 0.621 (95% CI, 0.611–0.631), and a DOR of 1.083 (95% CI, 0.063–18.509). The forest plot detailing individual study results is shown in Figures 3a and 3b.

A further subgroup analysis was conducted by AI input modality, and the test for subgroup differences was significant for all three measures (p<0.05). For sensitivity, studies using electronic health records (EHRs) or clinical data showed the highest pooled sensitivity, at 0.613 (95% CI, 0.608–0.617), likely because they integrate a broad range of systemic physiological and laboratory parameters. Models based on electroencephalography (EEG) data had a sensitivity of 0.520 (95% CI, 0.513–0.526), while models using imaging data had the lowest sensitivity at 0.463 (95% CI, 0.453–0.473). In terms of specificity, models based on imaging data performed best, with a pooled specificity of 0.950 (95% CI, 0.945–0.954). The EHR or clinical subgroup achieved a specificity of 0.813 (95% CI, 0.809–0.816), while the EEG subgroup showed the lowest specificity at 0.748 (95% CI, 0.742–0.753). The DOR analysis also highlighted differences across modalities (test for subgroup differences: Q = 7.96, p = 0.019). Imaging models had the highest random-effects DOR of 25.006 (95% CI, 13.879–45.052), indicating that structural brain injuries on admission computed tomography (CT) or magnetic resonance imaging (MRI) are strong, specific predictors of survival. It was followed by EHR- or clinical-model analyses with a DOR of 10.043 (95% CI, 7.240–13.932). EEG models had the lowest DOR at 6.066 (95% CI, 1.428–25.755), indicating that the choice of input modality significantly influences the diagnostic performance of AI models for mortality prediction. The forest plot detailing individual study results is shown in Figures 3c and 3d.

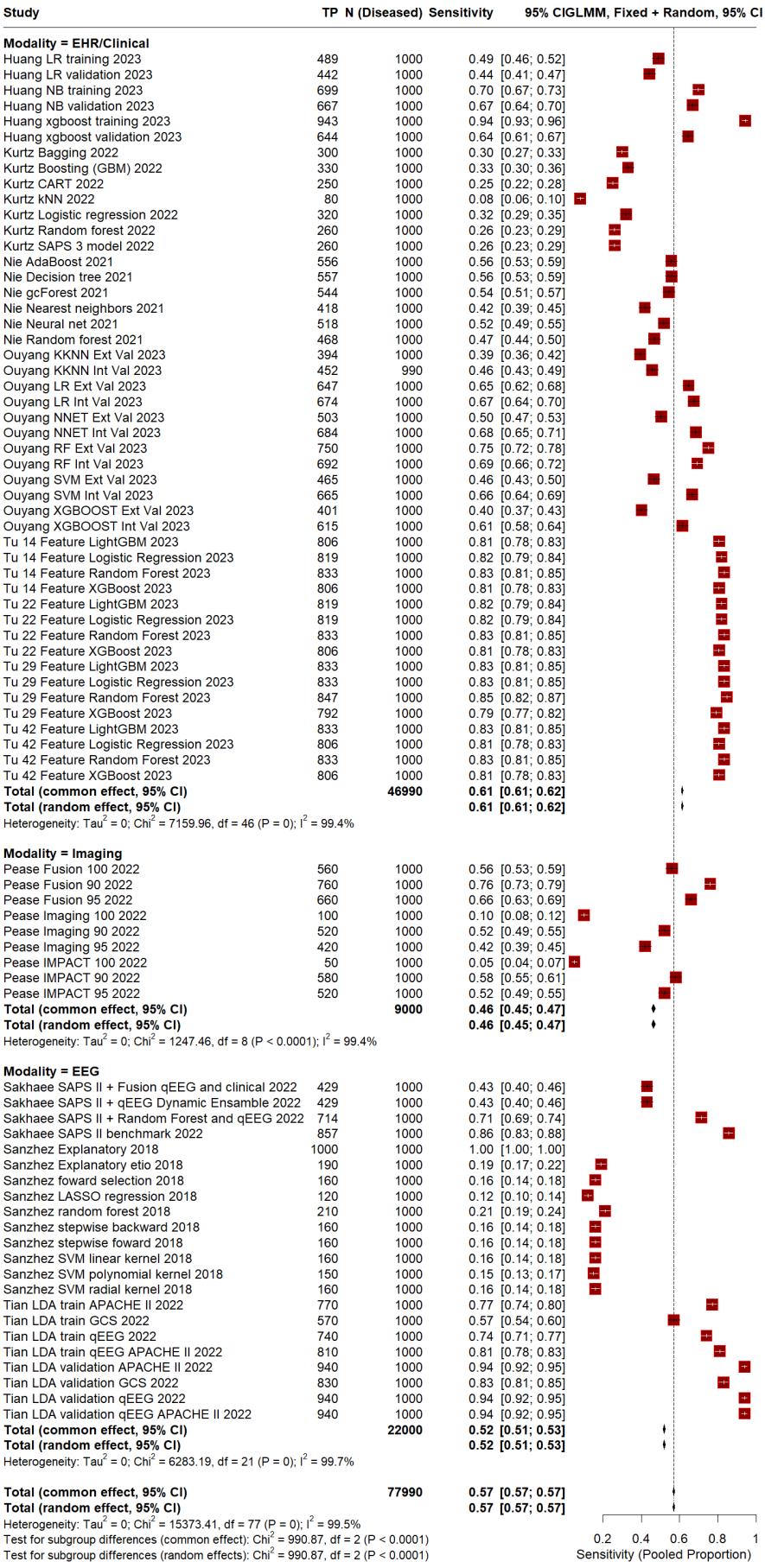
The narrowness of the 95% confidence intervals for sensitivity and specificity reported in these pools deserves critical evaluation. While the underlying study-level heterogeneity is extremely high (I^2>98%), the narrowness of the pooled intervals is mathematically driven by the massive cumulative sample sizes of several included cohorts (e.g., Kurtz et al., 2022, with 17,263 patients; Munjal et al., 2023, with 10,078 patients; Bishara et al., 2022, with 24,885 patients; Wong et al., 2018, with 18,223 patients). Under standard inverse-variance weighting, these massive studies carry disproportionate weight, which significantly reduces the pooled standard error and leads to narrower confidence intervals (Shim, 2022). This is further stabilized by the SCS method, which has been shown in simulation studies to provide superior, more robust coverage probabilities than traditional bivariate models, ensuring that these narrow widths reflect statistical efficiency rather than model overconfidence (Furuya-Kanamori et al., 2021).


Study
TP N (Diseased) Sensitivity
95% CIGLMM, Fixed + Random, 95% CI
Population_Age = Adult
Huang LR training 2023 489 1000 0.49 [0.46; 0.52]
Huang LR validation 2023 442 1000 0.44 [0.41; 0.47]
Huang NB training 2023 699 1000 0.70 [0.67; 0.73]
Huang NB validation 2023 667 1000 0.67 [0.64; 0.70]
Huang xgboost training 2023 943 1000 0.94 [0.93; 0.96]
Huang xgboost validation 2023 644 1000 0.64 [0.61; 0.67]
Kurtz Bagging 2022 300 1000 0.30 [0.27; 0.33]
Kurtz Boosting (GBM) 2022 330 1000 0.33 [0.30; 0.36]
Kurtz CART 2022 250 1000 0.25 [0.22; 0.28]
Kurtz kNN 2022 80 1000 0.08 [0.06; 0.10]
Kurtz Logistic regression 2022 320 1000 0.32 [0.29; 0.35]
Kurtz Random forest 2022 260 1000 0.26 [0.23; 0.29]
Kurtz SAPS 3 model 2022 260 1000 0.26 [0.23; 0.29]
Nie AdaBoost 2021 556 1000 0.56 [0.53; 0.59]
Nie Decision tree 2021 557 1000 0.56 [0.53; 0.59]
Nie gcForest 2021 544 1000 0.54 [0.51; 0.57]
Nie Nearest neighbors 2021 418 1000 0.42 [0.39; 0.45]
Nie Neural net 2021 518 1000 0.52 [0.49; 0.55]
Nie Random forest 2021 468 1000 0.47 [0.44; 0.50]
Ouyang KKNN Ext Val 2023 394 1000 0.39 [0.36; 0.42]
Ouyang KKNN Int Val 2023 452 990 0.46 [0.43; 0.49]
Ouyang LR Ext Val 2023 647 1000 0.65 [0.62; 0.68]
Ouyang LR Int Val 2023 674 1000 0.67 [0.64; 0.70]
Ouyang NNET Ext Val 2023 503 1000 0.50 [0.47; 0.53]
Ouyang NNET Int Val 2023 684 1000 0.68 [0.65; 0.71]
Ouyang RF Ext Val 2023 750 1000 0.75 [0.72; 0.78]
Ouyang RF Int Val 2023 692 1000 0.69 [0.66; 0.72]
Ouyang SVM Ext Val 2023 465 1000 0.46 [0.43; 0.50]
Ouyang SVM Int Val 2023 665 1000 0.66 [0.64; 0.69]
Ouyang XGBOOST Ext Val 2023 401 1000 0.40 [0.37; 0.43]
Ouyang XGBOOST Int Val 2023 615 1000 0.61 [0.58; 0.64]
Pease Fusion 100 2022 560 1000 0.56 [0.53; 0.59]
Pease Fusion 90 2022 760 1000 0.76 [0.73; 0.79]
Pease Fusion 95 2022 660 1000 0.66 [0.63; 0.69]
Pease Imaging 100 2022 100 1000 0.10 [0.08; 0.12]
Pease Imaging 90 2022 520 1000 0.52 [0.49; 0.55]
Pease Imaging 95 2022 420 1000 0.42 [0.39; 0.45]
Pease IMPACT 100 2022 50 1000 0.05 [0.04; 0.07]
Pease IMPACT 90 2022 580 1000 0.58 [0.55; 0.61]
Pease IMPACT 95 2022 520 1000 0.52 [0.49; 0.55]
Sakhaee SAPS II + Fusion qEEG and clinical 2022 429 1000 0.43 [0.40; 0.46]
Sakhaee SAPS II + qEEG Dynamic Ensamble 2022 429 1000 0.43 [0.40; 0.46]
Sakhaee SAPS II + Random Forest and qEEG 2022 714 1000 0.71 [0.69; 0.74]
Sakhaee SAPS II benchmark 2022 857 1000 0.86 [0.83; 0.88]
Tian LDA train APACHE II 2022 770 1000 0.77 [0.74; 0.80]
Tian LDA train GCS 2022 570 1000 0.57 [0.54; 0.60]
Tian LDA train qEEG 2022 740 1000 0.74 [0.71; 0.77]
Tian LDA train qEEG APACHE II 2022 810 1000 0.81 [0.78; 0.83]
Tian LDA validation APACHE II 2022 940 1000 0.94 [0.92; 0.95]
Tian LDA validation GCS 2022 830 1000 0.83 [0.81; 0.85]
Tian LDA validation qEEG 2022 940 1000 0.94 [0.92; 0.95]
Tian LDA validation qEEG APACHE II 2022 940 1000 0.94 [0.92; 0.95]
Tu 14 Feature LightGBM 2023 806 1000 0.81 [0.78; 0.83]
Tu 14 Feature Logistic Regression 2023 819 1000 0.82 [0.79; 0.84]
Tu 14 Feature Random Forest 2023 833 1000 0.83 [0.81; 0.85]
Tu 14 Feature XGBoost 2023 806 1000 0.81 [0.78; 0.83]
Tu 22 Feature LightGBM 2023 819 1000 0.82 [0.79; 0.84]
Tu 22 Feature Logistic Regression 2023 819 1000 0.82 [0.79; 0.84]
Tu 22 Feature Random Forest 2023 833 1000 0.83 [0.81; 0.85]
Tu 22 Feature XGBoost 2023 806 1000 0.81 [0.78; 0.83]
Tu 29 Feature LightGBM 2023 833 1000 0.83 [0.81; 0.85]
Tu 29 Feature Logistic Regression 2023 833 1000 0.83 [0.81; 0.85]
Tu 29 Feature Random Forest 2023 847 1000 0.85 [0.82; 0.87]
Tu 29 Feature XGBoost 2023 792 1000 0.79 [0.77; 0.82]
Tu 42 Feature LightGBM 2023 833 1000 0.83 [0.81; 0.85]
Tu 42 Feature Logistic Regression 2023 806 1000 0.81 [0.78; 0.83]
Tu 42 Feature Random Forest 2023 833 1000 0.83 [0.81; 0.85]
Tu 42 Feature XGBoost 2023 806 1000 0.81 [0.78; 0.83]
Total (common effect, 95% CI) 67990 0.62 [0.61; 0.62]
Total (random effect, 95% CI) 0.62 [0.61; 0.62]
Heterogeneity: Tau² = 0; Chi² = 10616.11, df = 67 (P = 0); I² = 99.4%
Population_Age = Pediatric
Sanzhez Explanatory 2018 1000 1000 1.00 [1.00; 1.00]
Sanzhez Explanatory etio 2018 190 1000 0.19 [0.17; 0.22]
Sanzhez foward selection 2018 160 1000 0.16 [0.14; 0.18]
Sanzhez LASSO regression 2018 120 1000 0.12 [0.10; 0.14]
Sanzhez random forest 2018 210 1000 0.21 [0.19; 0.24]
Sanzhez stepwise backward 2018 160 1000 0.16 [0.14; 0.18]
Sanzhez stepwise foward 2018 160 1000 0.16 [0.14; 0.18]
Sanzhez SVM linear kernel 2018 160 1000 0.16 [0.14; 0.18]
Sanzhez SVM polynomial kernel 2018 150 1000 0.15 [0.13; 0.17]
Sanzhez SVM radial kernel 2018 160 1000 0.16 [0.14; 0.18]
Total (common effect, 95% CI) 10000 0.25 [0.24; 0.26]
Total (random effect, 95% CI) 0.25 [0.24; 0.26]
Heterogeneity: Tau² = 0; Chi² = 36.18, df = 9 (P < 0.0001); I² = 75.1%
Total (common effect, 95% CI) 77990 0.57 [0.57; 0.57]
Total (random effect, 95% CI) 0.57 [0.57; 0.57]
Heterogeneity: Tau² = 0; Chi² = 15373.41, df = 77 (P = 0); I² = 99.5%
Test for subgroup differences (common effect): Chi² = 4222.27, df = 1 (P = 0)
Test for subgroup differences (random effects): Chi² = 4222.27, df = 1 (P = 0)
0.2
0.4
0.6
0.8
1
Sensitivity (Pooled Proportion)

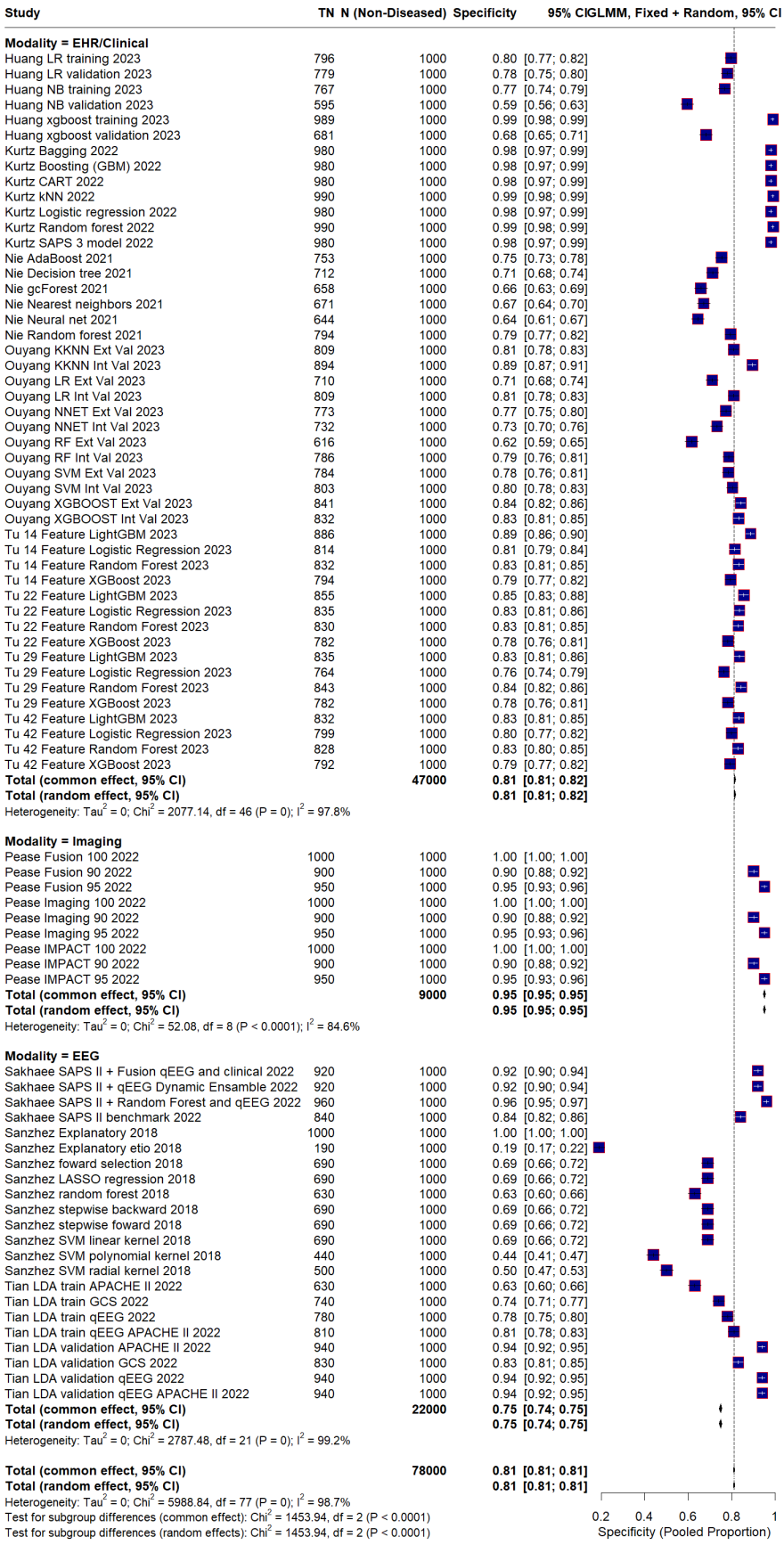
(a)


Study
TN
N (Non-Diseased)
Specificity
95% CIGLMM, Fixed + Random, 95% CI
Population_Age = Adult
Huang LR training 2023 796 1000 0.80 [0.77; 0.82]
Huang LR validation 2023 779 1000 0.78 [0.75; 0.80]
Huang NB training 2023 767 1000 0.77 [0.74; 0.79]
Huang NB validation 2023 595 1000 0.59 [0.56; 0.63]
Huang xgboost training 2023 989 1000 0.99 [0.98; 0.99]
Huang xgboost validation 2023 681 1000 0.68 [0.65; 0.71]
Kurtz Bagging 2022 980 1000 0.98 [0.97; 0.99]
Kurtz Boosting (GBM) 2022 980 1000 0.98 [0.97; 0.99]
Kurtz CART 2022 980 1000 0.98 [0.97; 0.99]
Kurtz kNN 2022 990 1000 0.99 [0.98; 0.99]
Kurtz Logistic regression 2022 980 1000 0.98 [0.97; 0.99]
Kurtz Random forest 2022 990 1000 0.99 [0.98; 0.99]
Kurtz SAPS 3 model 2022 980 1000 0.98 [0.97; 0.99]
Nie AdaBoost 2021 753 1000 0.75 [0.73; 0.78]
Nie Decision tree 2021 712 1000 0.71 [0.68; 0.74]
Nie gcForest 2021 658 1000 0.66 [0.63; 0.69]
Nie Nearest neighbors 2021 671 1000 0.67 [0.64; 0.70]
Nie Neural net 2021 644 1000 0.64 [0.61; 0.67]
Nie Random forest 2021 794 1000 0.79 [0.77; 0.82]
Ouyang KKNN Ext Val 2023 809 1000 0.81 [0.78; 0.83]
Ouyang KKNN Int Val 2023 894 1000 0.89 [0.87; 0.91]
Ouyang LR Ext Val 2023 710 1000 0.71 [0.68; 0.74]
Ouyang LR Int Val 2023 809 1000 0.81 [0.78; 0.83]
Ouyang NNET Ext Val 2023 773 1000 0.77 [0.75; 0.80]
Ouyang NNET Int Val 2023 732 1000 0.73 [0.70; 0.76]
Ouyang RF Ext Val 2023 616 1000 0.62 [0.59; 0.65]
Ouyang RF Int Val 2023 786 1000 0.79 [0.76; 0.81]
Ouyang SVM Ext Val 2023 784 1000 0.78 [0.76; 0.81]
Ouyang SVM Int Val 2023 803 1000 0.80 [0.78; 0.83]
Ouyang XGBOOST Ext Val 2023 841 1000 0.84 [0.82; 0.86]
Ouyang XGBOOST Int Val 2023 832 1000 0.83 [0.81; 0.85]
Pease Fusion 100 2022 1000 1000 1.00 [1.00; 1.00]
Pease Fusion 90 2022 900 1000 0.90 [0.88; 0.92]
Pease Fusion 95 2022 950 1000 0.95 [0.93; 0.96]
Pease Imaging 100 2022 1000 1000 1.00 [1.00; 1.00]
Pease Imaging 90 2022 900 1000 0.90 [0.88; 0.92]
Pease Imaging 95 2022 950 1000 0.95 [0.93; 0.96]
Pease IMPACT 100 2022 1000 1000 1.00 [1.00; 1.00]
Pease IMPACT 90 2022 900 1000 0.90 [0.88; 0.92]
Pease IMPACT 95 2022 950 1000 0.95 [0.93; 0.96]
Sakhaee SAPS II + Fusion qEEG and clinical 2022 920 1000 0.92 [0.90; 0.94]
Sakhaee SAPS II + qEEG Dynamic Ensamble 2022 920 1000 0.92 [0.90; 0.94]
Sakhaee SAPS II + Random Forest and qEEG 2022 960 1000 0.96 [0.95; 0.97]
Sakhaee SAPS II benchmark 2022 840 1000 0.84 [0.82; 0.86]
Tian LDA train APACHE II 2022 630 1000 0.63 [0.60; 0.66]
Tian LDA train GCS 2022 740 1000 0.74 [0.71; 0.77]
Tian LDA train qEEG 2022 780 1000 0.78 [0.75; 0.80]
Tian LDA train qEEG APACHE II 2022 810 1000 0.81 [0.78; 0.83]
Tian LDA validation APACHE II 2022 940 1000 0.94 [0.92; 0.95]
Tian LDA validation GCS 2022 830 1000 0.83 [0.81; 0.85]
Tian LDA validation qEEG 2022 940 1000 0.94 [0.92; 0.95]
Tian LDA validation qEEG APACHE II 2022 940 1000 0.94 [0.92; 0.95]
Tu 14 Feature LightGBM 2023 886 1000 0.89 [0.86; 0.90]
Tu 14 Feature Logistic Regression 2023 814 1000 0.81 [0.79; 0.84]
Tu 14 Feature Random Forest 2023 832 1000 0.83 [0.81; 0.85]
Tu 14 Feature XGBoost 2023 794 1000 0.79 [0.77; 0.82]
Tu 22 Feature LightGBM 2023 855 1000 0.85 [0.83; 0.88]
Tu 22 Feature Logistic Regression 2023 835 1000 0.83 [0.81; 0.86]
Tu 22 Feature Random Forest 2023 830 1000 0.83 [0.81; 0.85]
Tu 22 Feature XGBoost 2023 782 1000 0.78 [0.76; 0.81]
Tu 29 Feature LightGBM 2023 835 1000 0.83 [0.81; 0.86]
Tu 29 Feature Logistic Regression 2023 764 1000 0.76 [0.74; 0.79]
Tu 29 Feature Random Forest 2023 843 1000 0.84 [0.82; 0.86]
Tu 29 Feature XGBoost 2023 782 1000 0.78 [0.76; 0.81]
Tu 42 Feature LightGBM 2023 832 1000 0.83 [0.81; 0.85]
Tu 42 Feature Logistic Regression 2023 799 1000 0.80 [0.77; 0.82]
Tu 42 Feature Random Forest 2023 828 1000 0.83 [0.80; 0.85]
Tu 42 Feature XGBoost 2023 792 1000 0.79 [0.77; 0.82]
Total (common effect, 95% CI) 68000 0.84 [0.84; 0.84]
Total (random effect, 95% CI) 0.84 [0.84; 0.84]
Heterogeneity: Tau² = 0; Chi² = 3505.02, df = 67 (P = 0); I² = 98.1%
Population_Age = Pediatric
Sanzhez Explanatory 2018 1000 1000 1.00 [1.00; 1.00]
Sanzhez Explanatory etio 2018 190 1000 0.19 [0.17; 0.22]
Sanzhez foward selection 2018 690 1000 0.69 [0.66; 0.72]
Sanzhez LASSO regression 2018 690 1000 0.69 [0.66; 0.72]
Sanzhez random forest 2018 630 1000 0.63 [0.60; 0.66]
Sanzhez stepwise backward 2018 690 1000 0.69 [0.66; 0.72]
Sanzhez stepwise foward 2018 690 1000 0.69 [0.66; 0.72]
Sanzhez SVM linear kernel 2018 690 1000 0.69 [0.66; 0.72]
Sanzhez SVM polynomial kernel 2018 440 1000 0.44 [0.41; 0.47]
Sanzhez SVM radial kernel 2018 500 1000 0.50 [0.47; 0.53]
Total (common effect, 95% CI) 10000 0.62 [0.61; 0.63]
Total (random effect, 95% CI) 0.62 [0.61; 0.63]
Heterogeneity: Tau² = 0; Chi² = 839.71, df = 9 (P < 0.0001); I² = 98.9%
Total (common effect, 95% CI) 78000 0.81 [0.81; 0.81]
Total (random effect, 95% CI) 0.81 [0.81; 0.81]
Heterogeneity: Tau² = 0; Chi² = 5988.84, df = 77 (P = 0); I² = 98.7%
Test for subgroup differences (common effect): Chi² = 2490.30, df = 1 (P = 0)
Test for subgroup differences (random effects): Chi² = 2490.30, df = 1 (P = 0)
0.2
0.4
0.6
0.8
1
Specificity (Pooled Proportion)

(b)


Study
TP N (Diseased) Sensitivity
95% CIGLMM, Fixed + Random, 95% CI
Modality = EHR/Clinical
Huang LR training 2023 489 1000 0.49 [0.46; 0.52]
Huang LR validation 2023 442 1000 0.44 [0.41; 0.47]
Huang NB training 2023 699 1000 0.70 [0.67; 0.73]
Huang NB validation 2023 667 1000 0.67 [0.64; 0.70]
Huang xgboost training 2023 943 1000 0.94 [0.93; 0.96]
Huang xgboost validation 2023 644 1000 0.64 [0.61; 0.67]
Kurtz Bagging 2022 300 1000 0.30 [0.27; 0.33]
Kurtz Boosting (GBM) 2022 330 1000 0.33 [0.30; 0.36]
Kurtz CART 2022 250 1000 0.25 [0.22; 0.28]
Kurtz kNN 2022 80 1000 0.08 [0.06; 0.10]
Kurtz Logistic regression 2022 320 1000 0.32 [0.29; 0.35]
Kurtz Random forest 2022 260 1000 0.26 [0.23; 0.29]
Kurtz SAPS 3 model 2022 260 1000 0.26 [0.23; 0.29]
Nie AdaBoost 2021 556 1000 0.56 [0.53; 0.59]
Nie Decision tree 2021 557 1000 0.56 [0.53; 0.59]
Nie gcForest 2021 544 1000 0.54 [0.51; 0.57]
Nie Nearest neighbors 2021 418 1000 0.42 [0.39; 0.45]
Nie Neural net 2021 518 1000 0.52 [0.49; 0.55]
Nie Random forest 2021 468 1000 0.47 [0.44; 0.50]
Ouyang KKNN Ext Val 2023 394 1000 0.39 [0.36; 0.42]
Ouyang KKNN Int Val 2023 452 990 0.46 [0.43; 0.49]
Ouyang LR Ext Val 2023 647 1000 0.65 [0.62; 0.68]
Ouyang LR Int Val 2023 674 1000 0.67 [0.64; 0.70]
Ouyang NNET Ext Val 2023 503 1000 0.50 [0.47; 0.53]
Ouyang NNET Int Val 2023 684 1000 0.68 [0.65; 0.71]
Ouyang RF Ext Val 2023 750 1000 0.75 [0.72; 0.78]
Ouyang RF Int Val 2023 692 1000 0.69 [0.66; 0.72]
Ouyang SVM Ext Val 2023 465 1000 0.46 [0.43; 0.50]
Ouyang SVM Int Val 2023 665 1000 0.66 [0.64; 0.69]
Ouyang XGBOOST Ext Val 2023 401 1000 0.40 [0.37; 0.43]
Ouyang XGBOOST Int Val 2023 615 1000 0.61 [0.58; 0.64]
Tu 14 Feature LightGBM 2023 806 1000 0.81 [0.78; 0.83]
Tu 14 Feature Logistic Regression 2023 819 1000 0.82 [0.79; 0.84]
Tu 14 Feature Random Forest 2023 833 1000 0.83 [0.81; 0.85]
Tu 14 Feature XGBoost 2023 806 1000 0.81 [0.78; 0.83]
Tu 22 Feature LightGBM 2023 819 1000 0.82 [0.79; 0.84]
Tu 22 Feature Logistic Regression 2023 819 1000 0.82 [0.79; 0.84]
Tu 22 Feature Random Forest 2023 833 1000 0.83 [0.81; 0.85]
Tu 22 Feature XGBoost 2023 806 1000 0.81 [0.78; 0.83]
Tu 29 Feature LightGBM 2023 833 1000 0.83 [0.81; 0.85]
Tu 29 Feature Logistic Regression 2023 833 1000 0.83 [0.81; 0.85]
Tu 29 Feature Random Forest 2023 847 1000 0.85 [0.82; 0.87]
Tu 29 Feature XGBoost 2023 792 1000 0.79 [0.77; 0.82]
Tu 42 Feature LightGBM 2023 833 1000 0.83 [0.81; 0.85]
Tu 42 Feature Logistic Regression 2023 806 1000 0.81 [0.78; 0.83]
Tu 42 Feature Random Forest 2023 833 1000 0.83 [0.81; 0.85]
Tu 42 Feature XGBoost 2023 806 1000 0.81 [0.78; 0.83]
Total (common effect, 95% CI) 46990 0.61 [0.61; 0.62]
Total (random effect, 95% CI) 0.61 [0.61; 0.62]
Heterogeneity: Tau² = 0; Chi² = 7159.96, df = 46 (P = 0); I² = 99.4%
Modality = Imaging
Pease Fusion 100 2022 560 1000 0.56 [0.53; 0.59]
Pease Fusion 90 2022 760 1000 0.76 [0.73; 0.79]
Pease Fusion 95 2022 660 1000 0.66 [0.63; 0.69]
Pease Imaging 100 2022 100 1000 0.10 [0.08; 0.12]
Pease Imaging 90 2022 520 1000 0.52 [0.49; 0.55]
Pease Imaging 95 2022 420 1000 0.42 [0.39; 0.45]
Pease IMPACT 100 2022 50 1000 0.05 [0.04; 0.07]
Pease IMPACT 90 2022 580 1000 0.58 [0.55; 0.61]
Pease IMPACT 95 2022 520 1000 0.52 [0.49; 0.55]
Total (common effect, 95% CI) 9000 0.46 [0.45; 0.47]
Total (random effect, 95% CI) 0.46 [0.45; 0.47]
Heterogeneity: Tau² = 0; Chi² = 1247.46, df = 8 (P < 0.0001); I² = 99.4%
Modality = EEG
Sakhaee SAPS II + Fusion qEEG and clinical 2022 429 1000 0.43 [0.40; 0.46]
Sakhaee SAPS II + qEEG Dynamic Ensamble 2022 429 1000 0.43 [0.40; 0.46]
Sakhaee SAPS II + Random Forest and qEEG 2022 714 1000 0.71 [0.69; 0.74]
Sakhaee SAPS II benchmark 2022 857 1000 0.86 [0.83; 0.88]
Sanzhez Explanatory 2018 1000 1000 1.00 [1.00; 1.00]
Sanzhez Explanatory etio 2018 190 1000 0.19 [0.17; 0.22]
Sanzhez foward selection 2018 160 1000 0.16 [0.14; 0.18]
Sanzhez LASSO regression 2018 120 1000 0.12 [0.10; 0.14]
Sanzhez random forest 2018 210 1000 0.21 [0.19; 0.24]
Sanzhez stepwise backward 2018 160 1000 0.16 [0.14; 0.18]
Sanzhez stepwise foward 2018 160 1000 0.16 [0.14; 0.18]
Sanzhez SVM linear kernel 2018 160 1000 0.16 [0.14; 0.18]
Sanzhez SVM polynomial kernel 2018 150 1000 0.15 [0.13; 0.17]
Sanzhez SVM radial kernel 2018 160 1000 0.16 [0.14; 0.18]
Tian LDA train APACHE II 2022 770 1000 0.77 [0.74; 0.80]
Tian LDA train GCS 2022 570 1000 0.57 [0.54; 0.60]
Tian LDA train qEEG 2022 740 1000 0.74 [0.71; 0.77]
Tian LDA train qEEG APACHE II 2022 810 1000 0.81 [0.78; 0.83]
Tian LDA validation APACHE II 2022 940 1000 0.94 [0.92; 0.95]
Tian LDA validation GCS 2022 830 1000 0.83 [0.81; 0.85]
Tian LDA validation qEEG 2022 940 1000 0.94 [0.92; 0.95]
Tian LDA validation qEEG APACHE II 2022 940 1000 0.94 [0.92; 0.95]
Total (common effect, 95% CI) 22000 0.52 [0.51; 0.53]
Total (random effect, 95% CI) 0.52 [0.51; 0.53]
Heterogeneity: Tau² = 0; Chi² = 6283.19, df = 21 (P = 0); I² = 99.7%
Total (common effect, 95% CI) 77990 0.57 [0.57; 0.57]
Total (random effect, 95% CI) 0.57 [0.57; 0.57]
Heterogeneity: Tau² = 0; Chi² = 15373.41, df = 77 (P = 0); I² = 99.5%
Test for subgroup differences (common effect): Chi² = 990.87, df = 2 (P < 0.0001)
Test for subgroup differences (random effects): Chi² = 990.87, df = 2 (P < 0.0001)
0.2
0.4
0.6
0.8
1
Sensitivity (Pooled Proportion)

(c)



(d)

Figure 3. (a) Pooled sensitivity for population subgroup in predicting mortality; (b) pooled specificity for population subgroup in predicting mortality; (c) pooled sensitivity for the modality′s subgroup in predicting mortality; (d) pooled specificities for the modality′s subgroup in predicting mortality

Diagnostic Accuracy of AI in Predicting Neurological Outcome

In predicting neurological outcomes (assessed at 3 to 12 months using the mRS or GOS), the models demonstrated their strongest performance, with the meta-analysis indicating a pooled random-effects DOR of 45.566 (95% CI, 15.952–130.162). The pooled sensitivity for neurological outcome prediction was 0.864 (95% CI, 0.851–0.876), and the pooled specificity was 0.865 (95% CI, 0.852–0.876). Due to the limited number of studies, a funnel plot asymmetry test could not be performed reliably (Egger's test was not applicable). The pooled LR^+ was 6.080 (95% CI, 2.667–13.857), and the pooled LR^- was 0.165 (95% CI, 0.073–0.377). The overall performance showed significant heterogeneity (I^2 = 98.1% for DOR, 96.6% for sensitivity, and 96.2% for specificity). This strong diagnostic performance is clinically highly significant. The low LR^- (0.17) means that if the AI model predicts a favorable long-term neurological outcome, the probability of a poor outcome is extremely low. This high predictive accuracy makes these models highly valuable for clinical counseling, helping clinicians provide families with reliable, evidence-based reassurance during difficult discussions. The ROC curve analysis of neurological outcome prediction yielded an AUC of 0.858 (95% CI, 0.772–0.916), as shown in Figure 6.

The analysis of neurological outcomes was based on studies that exclusively included adult participants. The pooled random-effects sensitivity for this subgroup was 0.864 (95% CI, 0.851–0.876), and the specificity was 0.865 (95% CI, 0.852–0.876). The pooled random-effects DOR was 45.566 (95% CI, 15.952–130.162). The forest plot detailing individual study results is shown in Figures 7a and 7b.

A subgroup analysis by AI input modality was conducted, and the test for subgroup differences was highly significant across all three measures (p<0.001). For sensitivity, models using EHR or clinical data achieved higher performance (0.896; 95% CI, 0.882–0.909) than imaging models (0.800; 95% CI, 0.774–0.824). Similarly, for specificity, EHR or clinical models were higher at 0.897 (95% CI, 0.883–0.910) compared to imaging models at 0.800 (95% CI, 0.774–0.824). The random-effects DOR analysis (test for subgroup differences Q = 62.640, p < 0.001) also indicated that EHR or clinical models achieved a substantially higher DOR of 77.029 (95% CI, 55.844–106.251) than Imaging models, which had a DOR of 16.000 (95% CI, 12.852–19.920). This finding indicates that long-term functional recovery is highly dependent on systemic clinical factors, such as age, baseline comorbidities, and hospital complications, which are captured extensively in EHR databases. The forest plot detailing individual study results is shown in Figures 7c and 7d.

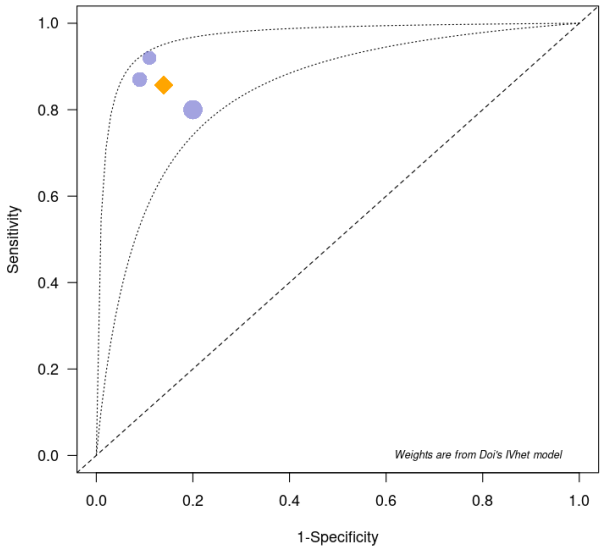
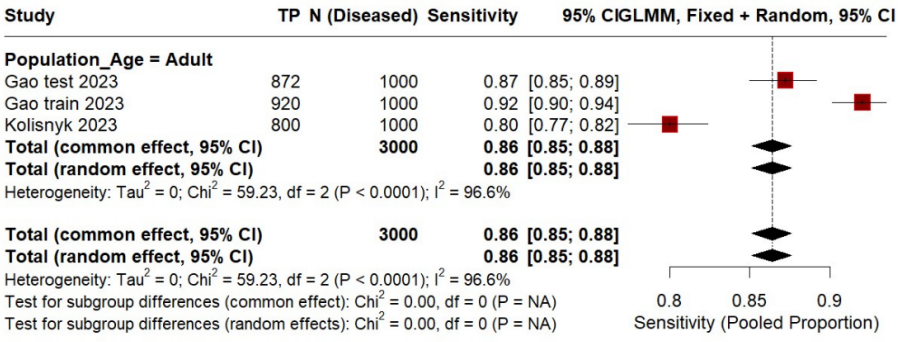
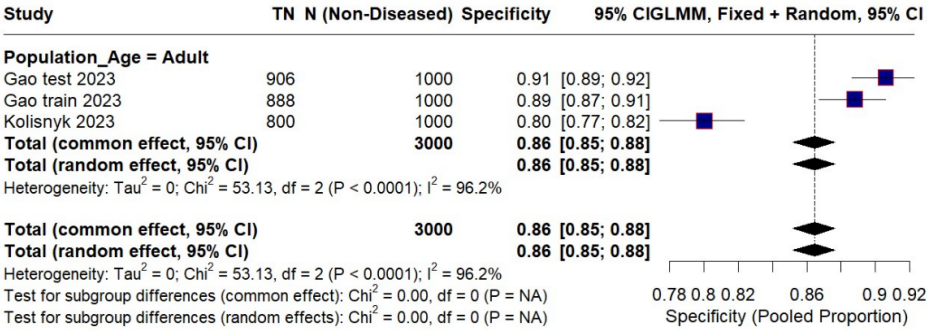


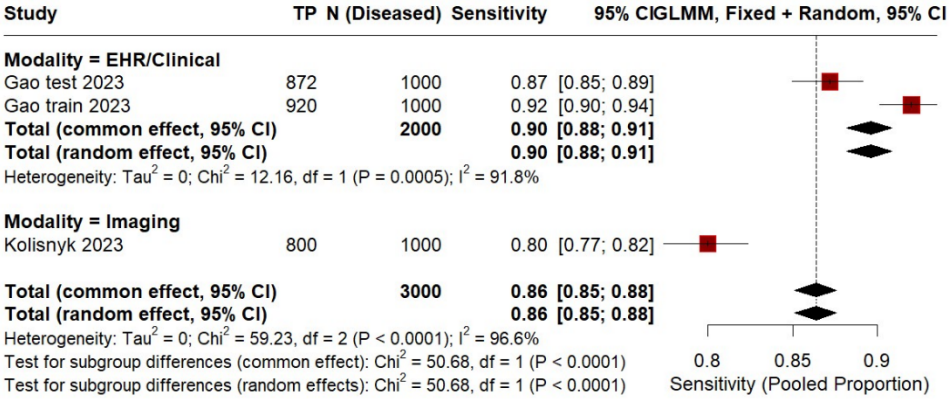
Figure 6. SROC curve in neurological outcome prediction



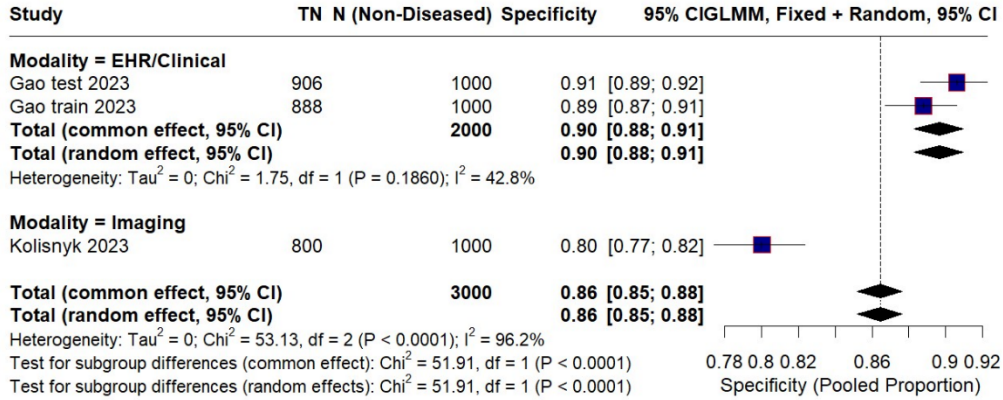
(a)



(b)



(c)



(d)

Figure 7. (a) Pooled sensitivity for population subgroup in neurological outcome prediction; (b) pooled specificity for population subgroup in neurological outcome prediction; (c) pooled sensitivity for the modality's subgroup in neurological outcome prediction; (d) pooled specificities for the modality's subgroup in neurological outcome prediction

Overall, the meta-analysis demonstrated that AI-based early warning systems exhibit varying diagnostic accuracies across the three evaluated outcomes. Although sensitivity remained moderate for mortality and deterioration in consciousness, specificity was notably high across all outcomes. The AUC values, ranging from 0.698 for mortality to 0.858 for neurological outcomes, suggest that AI models have reasonable to good discriminatory ability in these critical care scenarios.

DISCUSSION

The application of AI in critical care, particularly in the ICU, holds immense promise for improving prognostication and resource allocation. Our findings highlight the varying diagnostic accuracy of AI-driven prognostication in neurocritical care, depending on the clinical endpoint evaluated, suggesting that although AI performs satisfactorily in some areas, its utility in others requires further refinement and validation. For mortality prediction, our meta-analysis revealed an overall AUC of 0.698 and a DOR of 9.620. While indicating some predictive capacity, these figures suggest that current AI models still do not demonstrate a significant advantage over traditional methods for predicting ICU mortality. This finding aligns with observations by several researchers who highlight the complex, multifactorial nature of mortality in critically ill patients, often influenced by numerous confounding variables and dynamic changes in patient status, which challenge even sophisticated AI algorithms (Kang & Yoon, 2023; Weissman & Liu, 2021). The limited sensitivity suggests that current AI systems might miss many patients at high risk of mortality.

In contrast, the AI achieved acceptable results in predicting deterioration of consciousness, with an AUC of 0.796 and a notably higher DOR of 38.346. The specificity was extremely high, indicating that the AI systems were highly effective at identifying patients who did not experience a decline in consciousness. Although the sensitivity remained moderate, the higher overall accuracy and DOR suggest greater clinical utility in this domain. This improved performance may be attributable to the more direct physiological signals, such as EEG and neuroimaging features, that AI models can leverage to detect subtle changes indicative of impending neurological decline (Amann et al., 2022; Lee et al., 2022). Timely prediction of consciousness deterioration can enable earlier interventions, potentially mitigate secondary brain injury, and improve patient outcomes (Yuan et al., 2025).

Therefore, the most striking finding is the superior performance of AI in predicting long-term functional neurological outcomes (AUC: 0.858) and acute deterioration of consciousness (AUC: 0.796) compared to its relatively

limited performance in predicting short-term mortality (AUC: 0.698). To understand this performance gap, it is helpful to look at the underlying physiology of these endpoints. Long-term functional recovery (such as the mRS at 3 to 12 months) and acute changes in consciousness are directly tied to the structural and functional integrity of the central nervous system. These outcomes are strongly reflected in neurodiagnostic data, including admission CT/MRI scans and continuous EEG recordings. Deep learning models, particularly convolutional neural networks (CNNs), are highly effective at analyzing these high-dimensional inputs. By identifying subtle patterns of structural injury or microvascular ischemia, these algorithms can make highly accurate predictions about neurological recovery (Courville et al., 2023; Guo et al., 2022; Mattia et al., 2022). The high predictive accuracy in this domain could significantly aid clinicians in providing families with informed prognoses and making critical decisions regarding rehabilitation and long-term care planning.

In contrast, mortality in the ICU is a highly complex, multifactorial endpoint. While acute brain injury is a major factor, death in the ICU is frequently driven by non-neurological, systemic complications. These include septic shock, acute respiratory distress syndrome (ARDS), multi-organ dysfunction syndrome (MODS), and hospital-acquired infections. These systemic complications are highly dynamic and influenced by active clinical interventions, which are difficult to capture in retrospective baseline models. Consequently, models trained on admission data struggle to accurately predict mortality because they cannot account for unpredictable downstream systemic events (Kang & Yoon, 2023; Weissman & Liu, 2021). Furthermore, compared with traditional scoring systems such as APACHE II or SOFA, AI models for mortality prediction do not demonstrate a clear clinical advantage (Thakur et al., 2023).

While AI excels at integrating complex, real-time data, traditional scores remain highly effective for general mortality risk stratification. This finding suggests that AI's primary value in the ICU is not in predicting survival, but in augmenting clinical scores to predict neurological recovery and acute changes in consciousness. Besides, our review shows that AI's primary value may not be in replacing conventional scores but in augmenting them, especially for predicting subtle changes in consciousness or longer-term neurological outcomes, which are often not captured effectively by traditional scoring systems. The continuous, dynamic nature of AI analysis can provide granularity and ongoing risk assessment that traditional tools lack, enabling truly proactive care in the neurocritical care setting (Dang et al., 2021; Vitt & Mainali, 2024).

The included studies used diverse AI models, ranging from traditional machine learning algorithms, such as logistic regression, support vector machines (SVMs), and random forests, to more complex deep learning approaches, such as neural networks and CNNs. Each model possesses unique strengths and weaknesses in terms of interpretability, data requirements, and computational intensity (Porcellato et al., 2025; Subudhi et al., 2021). The choice of an AI model often depends on the specific clinical question, the types of available data, and the computational resources available. However, the inherent variability in data quality, patient heterogeneity, and confounding factors across different ICU environments is a significant challenge. AI deep learning will never be 100% accurate, primarily due to inter-examiner differences and varying environmental conditions. Clinicians must underscore the limitations of achieving universal applicability and perfect predictive accuracy (Zhou et al., 2021).

Despite these advancements, the clinical application of AI for prognostication in critically ill neurological patients faces several hurdles. The "black box" nature of some complex AI models and the intensive training required can erode clinicians' trust and limit interpretability, hindering their integration into daily practice (Henzler et al., 2025). Furthermore, the generalizability of models trained on specific patient cohorts or healthcare systems to diverse populations and clinical settings is a concern. The efficiency and applicability of each AI method varied significantly across studies, influenced by factors such as dataset size, feature selection, and predicted outcome. This heterogeneity necessitates the careful validation of AI tools in real-world prospective studies before their widespread implementation (Kim et al., 2024; Mudgal et al., 2022).

When evaluating AI models for real-world ICU deployment, clinicians must look beyond mathematical performance (such as AUC) to evaluate actual clinical utility (Liu et al., 2026). This requires a careful analysis of the clinical consequences of false-positive and false-negative predictions. In neurocritical care, a false-negative prediction, where the AI model fails to identify a patient at high risk of deterioration, is highly dangerous. If an EWS fails to alert the clinical team, a patient experiencing subclinical seizures or early brain herniation may remain undetected, delaying lifesaving interventions. Therefore, to be clinically safe, an EWS must maintain high sensitivity to ensure that very few high-risk patients are missed (Allanled Siauta et al., 2023).

On the other hand, a high false-positive rate carries its own severe clinical risks, primarily driven by alarm fatigue. Intensive care units are already noisy environments, with clinicians bombarded by hundreds of alerts from telemetry, ventilators, and infusion pumps every day. If an AI-driven EWS frequently generates false alarms for stable patients, clinicians will quickly suffer from alarm fatigue. Over time, this cognitive overload can lead clinicians to ignore or disable the system, defeating its purpose and potentially putting patients at risk (Allanled Siauta et al., 2023). To demonstrate true clinical utility, future AI models must undergo decision curve analysis (DCA). DCA evaluates the net clinical benefit of using a prediction model across a range of threshold probabilities, comparing it to "treat all" or "treat none" strategies. This analysis helps determine whether using the AI tool to guide clinical decisions achieves a higher net benefit for patients, ensuring that the benefit of early detection is not outweighed by the clinical cost of false alarms (Liu et al., 2026).

Ethical considerations and potential biases are paramount in deploying AI in critical care. AI models are trained on historical data, and if these data reflect existing disparities in healthcare access, treatment, or outcomes across different demographic groups, the AI might perpetuate or even amplify these biases (Cross et al., 2024; Obermeyer et al., 2019). Otherwise, a major ethical concern when deploying AI for neuro-prognostication is the risk of algorithmic bias and self-fulfilling prophecies. In neurocritical care, prognostic predictions directly influence critical clinical decisions, most notably the WLST. When an AI model is trained on historical ICU data, it inevitably learns from prior clinical decisions, including cases in which life-sustaining therapy was withdrawn (Finley Caulfield et al., 2022). If the historical data contains cases where clinicians withdrew therapy from patients with severe injuries because they assumed the prognosis was poor, the AI model will learn to associate those injuries with mortality. When applied to new patients, the

model will predict a high probability of death for similar injuries. This creates a dangerous, self-fulfilling prophecy. If clinicians see a high mortality prediction from the AI model, they may be more likely to recommend WLST. The patient then dies, "confirming" the AI's prediction and feeding that biased outcome back into the hospital's database. Over time, this loop can entrench and amplify historical clinical biases, potentially leading to premature withdrawal of therapy for patients who might have survived with aggressive treatment (Sounderajah et al., 2025). Clinicians must remain aware of this limitation and ensure that AI predictions are never used as the sole basis for WLST decisions.

For instance, an AI model trained predominantly on data from a specific ethnic group or socioeconomic background might perform sub-optimally when applied to a different, underrepresented population, leading to inequitable care. Ensuring fairness, accountability, and transparency in AI development and deployment is crucial for preventing adverse outcomes and building trust among patients and healthcare providers (Char et al., 2018; Liebig et al., 2024). Thus, another major barrier to clinical implementation is the lack of external validation across diverse healthcare systems. The vast majority of the included studies were developed and validated in single centers or within high-income healthcare systems with extensive digital infrastructures. AI models are highly sensitive to "domain shift," where changes in patient demographics, local clinical practices, or ICU equipment can severely degrade model performance (Liu et al., 2026). An algorithm developed in a highly resourced hospital in the United States may perform poorly when applied to a different patient population or clinical workflow in a nationally accredited facility in Indonesia. Without rigorous, prospective external validation across diverse clinical settings, these models cannot be considered ready for widespread clinical use (Shim, 2022).

The successful adoption of AI in clinical settings also relies heavily on Explainable AI (XAI). Clinicians must understand why an AI model makes a particular prediction, not just what the Prediction is, especially when dealing with life-and-death decisions in an ICU. XAI methods can provide insights into the features that influence an AI's output, enabling clinicians to critically evaluate recommendations and integrate them with their expertise and contextual knowledge (Adadi & Berrada, 2018; Ueda et al., 2024). Without interpretability, the "black box" nature of complex algorithms can lead to a lack of trust and reluctance to incorporate AI into routine clinical workflows, regardless of its statistical accuracy.

Beyond diagnostic accuracy, the economic impact and cost-effectiveness of integrating AI early warning systems into the ICU warrant careful consideration. Although the initial investment in AI infrastructure and model development can be substantial, the potential for improved patient outcomes, reduced length of stay, optimized resource allocation, and the prevention of costly complications could yield significant long-term savings (Holzinger et al., 2017; Topol, 2019). Economic evaluations, including cost-benefit and cost-effectiveness analyses, are crucial for demonstrating the tangible value of AI in critical care and justifying its widespread adoption within healthcare budgets (Khanna et al., 2022). Such analyses must consider both direct healthcare costs and indirect societal benefits, including improved patient quality of life and reduced caregiver burden.

While this diagnostic meta-analysis provides valuable insights, several limitations must be noted. Firstly, significant clinical and methodological heterogeneity was observed across the included studies, driven by variations in patient populations, technical algorithms, and input data modalities. Secondly, there was substantial heterogeneity in outcome definitions and prediction horizons, ranging from short-term bedside changes (7-day delirium) to long-term functional recovery (12-month GOS). Thirdly, the exceptionally high performance of EEG-based models must be interpreted with caution, as it was calculated from a small number of studies with highly selected, small sample sizes, which are at high risk of overfitting. Fourthly, there is a clear geographical imbalance in the literature, with most studies conducted in high-resource healthcare systems, limiting the generalizability of the findings to resource-constrained settings (Ahuja, 2019). Last but not least, in resource-limited intensive care units, such as many hospitals in developing countries, deploying these AI models is highly challenging. These algorithms frequently require expensive, high-performance infrastructure, such as continuous quantitative EEG monitoring, advanced 3D MRI imaging, and integrated, real-time EHR databases. In facilities where these tools are unavailable, the findings of this meta-analysis cannot be directly applied, highlighting a major barrier to global clinical implementation.

CONCLUSIONS AND RECOMMENDATION

In conclusion, artificial intelligence demonstrates varying diagnostic accuracy and clinical potential across different neurocritical care outcomes. The strength of the evidence differs significantly depending on the clinical endpoint being evaluated. AI models demonstrate strong, satisfactory performance in predicting long-term functional neurological outcomes and acceptable accuracy in predicting short-term deterioration of consciousness. However, their performance in predicting ICU mortality remains limited, offering no clear advantage over established traditional clinical scoring systems. Given these limitations, current AI models are not yet ready for widespread, independent clinical deployment in ICUs. Widespread clinical integration is not yet warranted. Instead, these retrospective findings provide a robust statistical foundation for future clinical research and highlight the need for rigorous, prospective validation.

To advance the field and bridge the gap between retrospective research and bedside utility, future clinical AI research should implement several technical recommendations. Firstly, prospective, multicenter validation studies should be conducted to evaluate AI models in real time, ensuring they can perform reliably across diverse clinical workflows and patient populations. Secondly, head-to-head comparisons between AI models and established clinical scores (such as APACHE II, SOFA, and GCS) should be performed to demonstrate clear, incremental predictive value. Thirdly, DCA should be integrated into model evaluation to quantify the actual net clinical benefit and assess the real-world impact of false predictions. Last but not least, future clinical prediction studies using machine learning should strictly adhere to the 27-item checklist of the TRIPOD+AI statement to ensure transparent, standardized reporting across all model development and validation phases (Collins et al., 2024). Similarly, studies evaluating AI-centered diagnostic accuracy should follow the STARD-AI reporting guideline (Sounderajah et al., 2025). Adhering to these international reporting standards is critical to minimizing research waste, improving transparency, and building the clinical trust needed for future integration.

Acknowledgments

The authors sincerely thank the Faculty of Medicine, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia, for covering the article processing charges for this publication, and the Neurointensive Working Group of the Indonesian Neurological Association for their support during the writing process of this manuscript.

DECLARATION

Ethics Approval and Consent to Participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest Statement

The authors declare that they have no competing interests.

Funding

The authors received no financial support for this article's research, authorship, and/or publication.

Artificial Intelligence-Assisted Technology

The author used Claude (Anthropic, San Francisco, CA, USA) to refine phrasing and linguistic nuance in the English-language presentation of this manuscript. All intellectual content, arguments, and conclusions are solely the author's own. The author takes full responsibility for the integrity of the published work.

Authors' contributions.

Conceptualization and methodology, S.E.P., B.L.H., and B.B.; Investigation, S.E.P., M.H., M.H., and T.K.; Formal analysis, B.B. and T.K.; Visualization and writing – original draft, B.B., S.E.P., and M.H.; Writing – review and editing, B.L.H., T.K., R.R., and A.M.; Funding acquisition, S.E.P. and B.L.H.; Supervision, R.R. and A.M. All authors have read and agreed to the final version of the manuscript.

ABOUT THE AUTHORS

Stefanus Erdana Putra is a clinical neurologist and academic staff member at the Department of Neurology, Faculty of Medicine, Universitas Kristen Duta Wacana, and Bethesda Lempuyangwangi General Hospital, Yogyakarta. His clinical practice and research focus primarily on clinical neurology, neurodegenerative conditions, and general neurological disorders.

Baarid Luqman Hamidi is a neurologist and consultant in neuro-intensive care (*Neurocritical Care*). He currently serves as a lecturer and researcher at the Department of Neurology, Faculty of Medicine, Universitas Sebelas Maret, Surakarta, where he actively investigates critical care neurology and acute neurological interventions.

Benedictus is a medical professional and medical resident currently affiliated with the Department of Pathological Anatomy at the Faculty of Medicine, Universitas Indonesia, Jakarta. His academic work and clinical interests center around histopathology, oncological pathology, and the cellular mechanisms underpinning human diseases.

Muhammad Hafizhan is a neurologist who manages acute and chronic neurological conditions in secondary clinical settings. He is currently a clinical practitioner at the Department of Neurology, Ananda Babelan General Hospital, Bekasi, with a focus on clinical neurology and community neuro-health outcomes.

Tyasno Koeshermanto is a general practitioner dedicated to frontline emergency healthcare delivery. He is currently based at the Medical Emergency Unit of Hj Anna Lasmanah Regional General Hospital in Banjarnegara, where his professional focus centers on trauma management, acute medical triage, and emergency medicine.

Retnaningsih is a senior lecturer, researcher, and consultant in neurocritical care and intensive care medicine. Serving at the Department of Neurology, Faculty of Medicine, Universitas Diponegoro, Semarang, she has published extensive research that bridges neuro-intensive care protocols, public health management, and critical patient outcomes.

Abdulloh Machin is a doctor of medicine and consultant neurologist specializing in neuro-intensive care. He is currently an active faculty member and clinician in the Department of Neurology, Faculty of Medicine, Universitas Airlangga, Surabaya, where his research agenda focuses on neurocritical care mechanisms, stroke management, and advanced clinical neurology.

REFERENCES

Abe, D., Inaji, M., Hase, T., Suehiro, E., Shiomi, N., Yatsushige, H., Hirota, S., Hasegawa, S., Karibe, H., Miyata, A., Kawakita, K., Haji, K., Aihara, H., Yokobori, S., Maeda, T., Onuki, T., Oshio, K., Komoribayashi, N., Suzuki, M., & Maehara, T. (2025). A machine learning model to predict neurological deterioration after mild traumatic brain injury in older adults. *Frontiers in Neurology*, *15*, 1502153. https://doi.org/10.3389/fneur.2024.1502153

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahuja, A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, *7*, e7702. https://doi.org/10.7717/peerj.7702

Allanled Siauta, V., Inayah, I., & Rohayani, L. (2023). Design of SBAR Communication Application Based on National Hospital Accreditation Standards and Imogene King's Theory in a Digital-Based Handover at Advent Bandung Hospital. *Majalah Kesehatan Indonesia*, *4*(1), 55–64. https://doi.org/10.47679/makein.2023188

Al-Mufti, F., Dodson, V., Lee, J., Wajswol, E., Gandhi, C., Scurlock, C., Cole, C., Lee, K., & Mayer, S. A. (2019). Artificial intelligence in neurocritical care. *Journal of the Neurological Sciences*, *404*, 1–4. https://doi.org/10.1016/j.jns.2019.06.024

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., Gilbert, T. K., Hagendorff, T., Holm, S., Livne, M., Spezzatti, A., Strümke, I., Zicari, R. V., Madai, V. I., & on behalf of the Z-Inspection initiative. (2022). To explain or not to explain?–Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*,

1(2), e0000016. https://doi.org/10.1371/journal.pdig.0000016

Amiri, M., Raimondo, F., Fisher, P. M., Cacic Hribljan, M., Sidaros, A., Othman, M. H., Zibrandtsen, I., Bergdal, O., Fabritius, M. L., Hansen, A. E., Hassager, C., Højgaard, J. L. S., Jensen, H. R., Knudsen, N. V., Laursen, E. L., Møller, J. E., Nersesjan, V., Nicolic, M., Sigurdsson, S. T., … Kondziella, D. (2024). Multimodal Prediction of 3- and 12-Month Outcomes in ICU Patients with Acute Disorders of Consciousness. *Neurocritical Care*, *40*(2), 718–733. https://doi.org/10.1007/s12028-023-01816-z

Bishara, A., Chiu, C., Whitlock, E. L., Douglas, V. C., Lee, S., Butte, A. J., Leung, J. M., & Donovan, A. L. (2022). Postoperative delirium prediction using machine learning models and preoperative electronic health record data. *BMC Anesthesiology*, *22*(1), 8. https://doi.org/10.1186/s12871-021-01543-y

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care–Addressing Ethical Challenges. *New England Journal of Medicine*, *378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Cherifa, M., & Pirracchio, R. (2019). What every intensivist should know about Big Data and targeted machine learning in the intensive care unit. *Revista Brasileira de Terapia Intensiva*, *31*(4), 444–446. https://doi.org/10.5935/0103-507X.20190069

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Courville, E., Kazim, S. F., Vellek, J., Tarawneh, O., Stack, J., Roster, K., Roy, J., Schmidt, M., & Bowers, C. (2023). Machine learning algorithms for predicting outcomes of traumatic brain injury: A systematic review and meta-analysis. *Surgical Neurology International*, *14*, 262. https://doi.org/10.25259/SNI_312_2023

Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, *3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

Dang, J., Lal, A., Flurin, L., James, A., Gajic, O., & Rabinstein, A. A. (2021). Predictive modeling in neurocritical care using causal artificial intelligence. *World Journal of Critical Care Medicine*, *10*(4), 112–119. https://doi.org/10.5492/wjccm.v10.i4.112

Elmer, J., Steinberg, A., & Callaway, C. W. (2023). Paucity of neuroprognostic testing after cardiac arrest in the United States. *Resuscitation*, *188*, 109762. https://doi.org/10.1016/j.resuscitation.2023.109762

Finley Caulfield, A., Mlynash, M., Eyngorn, I., Lansberg, M. G., Afjei, A., Venkatasubramanian, C., Buckwalter, M. S., & Hirsch, K. G. (2022). Prognostication of ICU Patients by Providers with and without Neurocritical Care Training. *Neurocritical Care*, *37*(1), 190–199. https://doi.org/10.1007/s12028-022-01467-6

Furuya-Kanamori, L., Kostoulas, P., & Doi, S. A. R. (2021). A new method for synthesizing test accuracy data outperformed the bivariate method. *Journal of Clinical Epidemiology*, *132*, 51–58. https://doi.org/10.1016/j.jclinepi.2020.12.015

Guo, R., Zhang, R., Liu, R., Liu, Y., Li, H., Ma, L., He, M., You, C., & Tian, R. (2022). Machine learning-based approaches for prediction of patients' functional outcome and mortality after spontaneous intracerebral hemorrhage. *Journal of Personalized Medicine*, *12*(1), 112. https://doi.org/10.3390/jpm12010112

Henzler, D., Schmidt, S., Koçar, A., Herdegen, S., Lindinger, G. L., Maris, M. T., Bak, M. A. R., Willems, D. L., Tan, H. L., Lauerer, M., Nagel, E., Hindricks, G., Dagres, N., & Konopka, M. J. (2025). Healthcare professionals' perspectives on artificial intelligence in patient care: A systematic review of hindering and facilitating factors on different levels. *BMC Health Services Research*, *25*(1), 633. https://doi.org/10.1186/s12913-025-12664-2

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). *What do we need to build explainable AI systems for the medical domain?* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1712.09923

Huang, J., Jin, W., Duan, X., Liu, X., Shu, T., Fu, L., Deng, J., Chen, H., Liu, G., Jiang, Y., & Liu, Z. (2023). Twenty-eight-day in-hospital mortality prediction for elderly patients with ischemic stroke in the intensive care unit: Interpretable machine learning models. *Frontiers in Public Health*, *10*, 1086339. https://doi.org/10.3389/fpubh.2022.1086339

Kang, C.-Y., & Yoon, J. H. (2023). Current challenges in adopting machine learning to critical care and emergency medicine. *Clinical and Experimental Emergency Medicine*, *10*(2), 132–137. https://doi.org/10.15441/ceem.23.041

Khanna, N. N., Maindarkar, M. A., Viswanathan, V., Fernandes, J. F. E., Paul, S., Bhagawati, M., Ahluwalia, P., Ruzsa, Z., Sharma, A., Kolluri, R., Singh, I. M., Laird, J. R., Fatemi, M., Alizad, A., Saba, L., Agarwal, V., Sharma, A., Teji, J. S., Al-Maini, M., … Suri, J. S. (2022). Economics of Artificial Intelligence in Healthcare: Diagnosis vs. Treatment. *Healthcare*, *10*(12), 2493. https://doi.org/10.3390/healthcare10122493

Kim, K. A., Kim, H., Ha, E. J., Yoon, B. C., & Kim, D.-J. (2024). Artificial Intelligence-Enhanced Neurocritical Care for Traumatic Brain Injury: Past, Present and Future. *Journal of Korean Neurosurgical Society*, *67*(5), 493–509. https://doi.org/10.3340/jkns.2023.0195

Kolisnyk, M., Kazazian, K., Rego, K., Novi, S. L., Wild, C. J., Gofton, T. E., Debicki, D. B., Owen, A. M., & Norton, L. (2023). Predicting neurologic recovery after severe acute brain injury using resting-state networks. *Journal of Neurology*, *270*(12), 6071–6080. https://doi.org/10.1007/s00415-023-11941-6

Kumar, S., Kumari, S., Jaiswal, M., & Madhukar, S. K. (2025). Neurosurgical ICU Outcome Prediction using Artificial Intelligence: A Retrospective Observational Study. *Journal of Trauma Intensive Care STIC*, *1*(2), 26–28. https://doi.org/10.5005/jtric-11018-0016

Kurtz, P., Peres, I. T., Soares, M., Salluh, J. I. F., & Bozza, F. A. (2022). Hospital Length of Stay and 30-Day Mortality Prediction in Stroke: A Machine Learning Analysis of 17,000 ICU Admissions in Brazil. *Neurocritical Care*, *37*(S2), 313–321. https://doi.org/10.1007/s12028-022-01486-3

Lee, M., Sanz, L. R. D., Barra, A., Wolff, A., Nieminen, J. O., Boly, M., Rosanova, M., Casarotto, S., Bodart, O., Annen, J., Thibaut, A., Panda, R., Bonhomme, V., Massimini, M., Tononi, G., Laureys, S., Gosseries, O., & Lee, S.-W. (2022). Quantifying arousal and awareness in altered states of consciousness using interpretable deep learning. *Nature Communications*, *13*(1), 1064. https://doi.org/10.1038/s41467-022-28451-0

Liebig, L., Jobin, A., Güttel, L., & Katzenbach, C. (2024). Situating AI policy: Controversies covered and the normalisation of AI. *Big Data & Society*, *11*(4),

20539517241299725. https://doi.org/10.1177/20539517241299725

Liu, L., Zhu, Q., Zong, Y., Chen, X., Zhang, W., & Wang, J. (2026). Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *28*, e78714. https://doi.org/10.2196/78714

Mattia, G. M., Sarton, B., Villain, E., Vinour, H., Ferre, F., Buffieres, W., Le Lann, M.-V., Franceries, X., Peran, P., & Silva, S. (2022). Multimodal MRI-Based Whole-Brain Assessment in Patients In Anoxoischemic Coma by Using 3D Convolutional Neural Networks. *Neurocritical Care*, *37*(S2), 303–312. https://doi.org/10.1007/s12028-022-01525-z

Miyazawa, Y., Katsuta, N., Nara, T., Nojiri, S., Naito, T., Hiki, M., Ichikawa, M., Takeshita, Y., Kato, T., Okumura, M., & Tobita, M. (2024). Identification of risk factors for the onset of delirium associated with COVID-19 by mining nursing records. *PLOS ONE*, *19*(1), e0296760. https://doi.org/10.1371/journal.pone.0296760

Mudgal, S. K., Agarwal, R., Chaturvedi, J., Gaur, R., & Ranjan, N. (2022). Real-world application, challenges and implication of artificial intelligence in healthcare: An essay. *The Pan African Medical Journal*, *43*, 3. https://doi.org/10.11604/pamj.2022.43.3.33384

Muller, E., Shock, J. P., Bender, A., Kleeberger, J., Högen, T., Rosenfelder, M., Bah, B., & Lopez-Rolon, A. (2019). Outcome prediction with serial neuron-specific enolase and machine learning in anoxic-ischaemic disorders of consciousness. *Computers in Biology and Medicine*, *107*, 145–152. https://doi.org/10.1016/j.compbiomed.2019.02.006

Munjal, N. K., Clark, R. S. B., Simon, D. W., Kochanek, P. M., & Horvat, C. M. (2023). Interoperable and explainable machine learning models to predict morbidity and mortality in acute neurological injury in the pediatric intensive care unit: Secondary analysis of the TOPICC study. *Frontiers in Pediatrics*, *11*, 1177470. https://doi.org/10.3389/fped.2023.1177470

Nie, X., Cai, Y., Liu, J., Liu, X., Zhao, J., Yang, Z., Wen, M., & Liu, L. (2021). Mortality Prediction in Cerebral Hemorrhage Patients Using Machine Learning Algorithms in Intensive Care Units. *Frontiers in Neurology*, *11*, 610531. https://doi.org/10.3389/fneur.2020.610531

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Ouyang, Y., Cheng, M., He, B., Zhang, F., Ouyang, W., Zhao, J., & Qu, Y. (2023). Interpretable machine learning models for predicting in-hospital death in patients in the intensive care unit with cerebral infarction. *Computer Methods and Programs in Biomedicine*, *231*, 107431. https://doi.org/10.1016/j.cmpb.2023.107431

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pease, M., Gonzalez-Martinez, J., Puccio, A., Nwachuku, E., Castellano, J. F., Okonkwo, D. O., & Elmer, J. (2022). Risk Factors and Incidence of Epilepsy after Severe Traumatic Brain Injury. *Annals of Neurology*, *92*(4), 663–669. https://doi.org/10.1002/ana.26443

Porcellato, E., Lanera, C., Ocagli, H., & Danielis, M. (2025). Exploring Applications of Artificial Intelligence in Critical Care Nursing: A Systematic Review. *Nursing Reports*, *15*(2), 55. https://doi.org/10.3390/nursrep15020055

Rajajee, V., Muehlschlegel, S., Wartenberg, K. E., Alexander, S. A., Busl, K. M., Chou, S. H. Y., Creutzfeldt, C. J., Fontaine, G. V., Fried, H., Hocker, S. E., Hwang, D. Y., Kim, K. S., Madzar, D., Mahanes, D., Mainali, S., Meixensberger, J., Montellano, F., Sakowitz, O. W., Weimar, C., … Varelas, P. N. (2023). Guidelines for Neuroprognostication in Comatose Adult Survivors of Cardiac Arrest. *Neurocritical Care*, *38*(3), 533–563. https://doi.org/10.1007/s12028-023-01688-3

Sakhaee, E., Amirahmadi, A., Mahdiani, M., Shojaei, M., Hassanian-Moghaddam, H., Bauer, R., Zamani, N., Pakdaman, H., & Gharagozli, K. (2022). Developing a novel prediction model in opioid overdose using machine learning; a pilot analytical study. *Health Science Reports*, *5*(5), e767. https://doi.org/10.1002/hsr2.767

Sánchez Fernández, I., Sansevere, A. J., Gaínza-Lein, M., Kapur, K., & Loddenkemper, T. (2018). Machine Learning for Outcome Prediction in Electroencephalograph (EEG)-Monitored Children in the Intensive Care Unit. *Journal of Child Neurology*, *33*(8), 546–553. https://doi.org/10.1177/0883073818773230

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-Analysis with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-21416-0

Shim, S. R. (2022). Meta-analysis of diagnostic test accuracy studies with multiple thresholds for data integration. *Epidemiology and Health*, *44*, e2022083. https://doi.org/10.4178/epih.e2022083

Sounderajah, V., Guni, A., Liu, X., Collins, G. S., Karthikesalingam, A., Markar, S. R., Golub, R. M., Denniston, A. K., Shetty, S., Moher, D., Bossuyt, P. M., Darzi, A., Ashrafian, H., STARD-AI Steering Committee, Acharya, A., Mateen, B. A., Kelly, C., Ting, D., Treanor, D., … Saria, S. (2025). The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine*, *31*(10), 3283–3289. https://doi.org/10.1038/s41591-025-03953-8

Subudhi, S., Verma, A., Patel, A. B., Hardin, C. C., Khandekar, M. J., Lee, H., McEvoy, D., Stylianopoulos, T., Munn, L. L., Dutta, S., & Jain, R. K. (2021). Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *Npj Digital Medicine*, *4*(1), 87. https://doi.org/10.1038/s41746-021-00456-x

Sun, H., Kimchi, E., Akeju, O., Nagaraj, S. B., McClain, L. M., Zhou, D. W., Boyle, E., Zheng, W.-L., Ge, W., & Westover, M. B. (2019). Automated tracking of level of consciousness and delirium in critical illness using deep learning. *Npj Digital Medicine*, *2*(1), 89. https://doi.org/10.1038/s41746-019-0167-0

Thakur, R., Naga Rohith, V., & Arora, J. K. (2023). Mean SOFA score in comparison with APACHE II score in predicting mortality in surgical patients with sepsis. *Cureus, 15*(3), e36653. https://doi.org/10.7759/cureus.36653

Tian, J., Zhou, Y., Liu, H., Qu, Z., Zhang, L., & Liu, L. (2022). Quantitative EEG parameters can improve the predictive value of the non-traumatic neurological ICU patient prognosis through the machine learning method. *Frontiers in Neurology*, *13*, 897734. https://doi.org/10.3389/fneur.2022.897734

Topol, E. J. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.

Tu, K.-C., Tau, E. N. T., Chen, N.-C., Chang, M.-C., Yu, T.-C., Wang, C.-C., Liu, C.-F., & Kuo, C.-L. (2023). Machine Learning Algorithm Predicts Mortality Risk in Intensive Care Unit

for Patients with Traumatic Brain Injury. *Diagnostics*, *13*(18), 3016. https://doi.org/10.3390/diagnostics13183016

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology*, *42*(1), 3–15. https://doi.org/10.1007/s11604-023-01474-3

Vitt, J. R., & Mainali, S. (2024). Artificial Intelligence and Machine Learning Applications in Critically Ill Brain Injured Patients. *Seminars in Neurology*, *44*(03), 342–356. https://doi.org/10.1055/s-0044-1785504

Weissman, G. E., & Liu, V. X. (2021). Algorithmic prognostication in critical care: A promising but unproven technology for supporting difficult decisions. *Current Opinion in Critical Care*, *27*(5), 500–505. https://doi.org/10.1097/MCC.0000000000000855

Wong, A., Young, A. T., Liang, A. S., Gonzales, R., Douglas, V. C., & Hadley, D. (2018). Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. *JAMA Network Open*, *1*(4), e181018. https://doi.org/10.1001/jamanetworkopen.2018.1018

Yap, X.-H. V., Tu, K.-C., Chen, N.-C., Wang, C.-C., Chen, C.-J., Liu, C.-F., Eric Nya, T.-T., & Kuo, C.-L. (2025). Developing a high-performance AI model for spontaneous intracerebral hemorrhage mortality prediction using machine learning in ICU settings. *BMC Medical Informatics and Decision Making*, *25*(1), 149. https://doi.org/10.1186/s12911-025-02984-y

Yuan, S., Yang, Z., Li, J., Wu, C., & Liu, S. (2025). AI-Powered early warning systems for clinical deterioration significantly improve patient outcomes: A meta-analysis. *BMC Medical Informatics and Decision Making*, *25*(1), 203. https://doi.org/10.1186/s12911-025-03048-x

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., & Summers, R. M. (2021). A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proceedings of the IEEE*, *109*(5), 820–838. https://doi.org/10.1109/JPROC.2021.3054390

Correspondence All inquiries and requests for additional materials should be directed to the Corresponding Author.

Publisher's Note Utan Kayu Publishing maintains a neutral stance regarding territorial claims depicted in published maps and does not endorse or reject the institutional affiliations stated by the authors.

Open Access This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0), which permits others to share, adapt, and redistribute the material in any medium or format, even for commercial purposes, provided appropriate credit is given to the original author(s) and the source, a link to the license is provided, and any changes made are indicated. If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/.

© The Author(s) 2026

ADDITIONAL INFORMATION

Table 1. Summary of Included Studies

First Author, Year	Study Location	Study Design	Participant Condition	Sample Size	Outcome Measured	Outcome Time Horizon	AI Model	AI Model Mechanism	Training/ Validation /External Test Splits
Huang et al., 2023	United States	Observational cohort	Elderly patients with ischemic stroke	1,236 patients	28-day in-hospital mortality	28 days	Naïve Bayes (NB), eXtreme Gradient Boosting (xgboost), and logistic regression (LR)	Utilizing a variety of clinical features, such as vital signs, demographics, laboratory tests, and comorbidities, to predict the 28-day in-hospital mortality of elderly patients with ischemic stroke.	Training: 80% / Validation: 20%
Kurtz et al., 2022	Brazil	Observational cohort	Ischemic or nontraumatic hemorrhagic stroke patient	17,263 patients	Hospital length of stay and 30-day in-hospital mortality	30 days	Support vector regression, k-nearest neighborhood, gradient boosting machine (GBM), random forests (RF), classification and regression trees (CART), bagging, and linear regression	Analyzing demographic data, comorbidities, acute disease characteristics, organ support, and laboratory data using a data-driven methodology.	Internal-external cross-validation; specific ratios not provided
Munjal et al., 2023	United States	Retrospective cohort	Acute neurological injury patients admitted to the pediatric intensive care unit (PICU)	10,078 patients	Mortality and morbidity rates among patients with suspected acute neurological injury	"Death vs. Survival" / "Death or Morbidity" / "New Morbidity" (timeframe not specified)	Ensemble regressor containing Random Forest, Gradient Boosting, and Support Vector Machine learners	Analyzing the admission physiology data to identify patterns and relationships that could predict mortality and morbidity outcomes. Feature importance analysis using SHapley Additive exPlanations (SHAP) was conducted to understand the contribution of different variables (such as vital signs and laboratory values) to the predictive models.	Training: 85% / Test: 15% (10-fold cross-validation)
Nie et al., 2021	China	Retrospective cohort	Cerebral hemorrhage patient	760 patients	In-hospital mortality	In-hospital mortality (timeframe not specified)	Nearest neighbors, decision tree, neural net, AdaBoost, random forest, and gcForest	Extracting features from thousands of variables in the database, selecting the most critical variables for analysis, and using various classifiers to improve the probability of good discrimination performance. These models were compared with the APACHE II score to determine their performance in predicting mortality.	Nor specified

First Author, Year	Study Location	Study Design	Participant Condition	Sample Size	Outcome Measured	Outcome Time Horizon	AI Model	AI Model Mechanism	Training/ Validation /External Test Splits
Ouyang et al., 2023	China	Retrospective cohort	Cerebral infarction patients	4,338 patients	In-hospital death	In-hospital death (timeframe not specified)	Logistic regression, neural network, support vector machine, weighted k-nearest neighbors, random forest, and extreme gradient boosting	The random forest algorithm creates multiple decision trees and combines their predictions to make a final prediction. The models were trained on the eRI dataset and tested on the MIMIC-III dataset.	eRI (Training), MIMIC-III (Test)
Pease et al., 2022	United States	Retrospective cohort	Traumatic brain injury patients	392 patients	Mortality and unfavorable outcomes at 6 months after severe traumatic brain injury (sTBI)	6 months	Convolutional neural network (CNN) model	The CNN model was trained using transfer learning and curriculum learning techniques on admission head CT scans. The model was then combined with clinical input for a holistic fusion model. The models were evaluated using an independent internal test set and an external cohort of patients from the TRACK-TBI study.	Independent internal test set, external TRACK-TBI cohort
Sakhaee et al., 2022	Iran	Retrospective cohort	Opioid overdose patients	32 patients	Prediction of opioid overdose mortality	Mortality (timeframe not specified)	KNN, Decision Tree, Random Forest, AdaBoost, MLP, SVM, logistic regression, and SAPS II are examples of disease severity classification systems. Additionally, the study employed the MSC approach for qEEG data analysis	Analyze and process clinical and qEEG data to predict patient mortality. The study employed three levels of data analysis to enhance mortality prediction, including constructing a pool of classifiers, selecting the most effective classifiers based on their performance, and aggregating their decisions for testing. Furthermore, the study employed feature and model fusion at different levels to enhance predictive performance.	Leave-one-out cross-validation (LOOCV)
Sánchez Fernández et al., 2018	United States	Observational cohort	Children undergoing continuous electroencephalography (cEEG) in the intensive care unit (ICU)	414 patients	In-hospital mortality	In-hospital mortality (timeframe not specified)	Stepwise forward selection, stepwise backward elimination, backward elimination, forward selection, least absolute shrinkage and selection operator (LASSO) regression, random forest, support	Machine learning models were developed automatically based on available data, with no human intervention. The models employed different algorithms to select predictive variables, striking a balance between predictive power and model complexity. The stepwise selection or elimination model	Least absolute shrinkage and selection operator (LASSO)

First Author, Year	Study Location	Study Design	Participant Condition	Sample Size	Outcome Measured	Outcome Time Horizon	AI Model	AI Model Mechanism	Training/ Validation /External Test Splits
							vector machine with linear kernel, support vector machine with polynomial kernel, and support vector machine with radial kernel	selects the best variable to include or exclude from the regression model by minimizing the Akaike information criterion (AIC). The LASSO regression model includes a penalty term that penalizes model complexity, which tends to shrink less important variables toward zero, resulting in a sparse and more interpretable model. The random forests model fits several decorrelated decision trees to the data and "averages" them, yielding a model with as little bias and variance as possible.	
Tian et al., 2022	China	Retrospective cohort	Non-traumatic neurological ICU patients who were older than 18 years and admitted to the neurological ICU	110 patients	3-month mortality	3 months	Linear discriminant analysis (LDA) based on quantitative EEG parameters	The LDA model was used to analyze the EEG parameters, clinical data, and APACHE II score to predict 3-month mortality in non-traumatic patients in the neurological ICU.	Training: 80 patients /Validation: 30 patients
Tu et al., 2023	Taiwan	Retrospective cohort	Traumatic brain injury patient	2,260 patients	Prediction of mortality risk	Mortality risk (timeframe not specified)	LightGBM	Analyzing 42 selected features to identify suitable models for predicting mortality risk. The model was trained using ICU data and various feature combinations to achieve accurate predictions.	Training/Validation: 70% / Test: 30% (cross-validation)
Amiri et al., 2024	Denmark	Observational cohort	Patients with disorders of consciousness (DoC) who were admitted to the ICU. The conditions underlying the participants varied and included traumatic brain injury (TBI), ischemic stroke, subarachnoid or intracerebral	123 patients	Level of consciousness of the participants at the time of study enrollment and ICU discharge	3 and 12-month outcomes	Support vector machine (SVM) and random forest algorithms	Analyzing EEG and fMRI features to predict the consciousness levels of patients. The models were trained on a combination of EEG and fMRI data to classify patients as either ≤ unresponsive wakefulness syndrome (UWS) or ≥ minimally conscious state (MCS).	Although it is not explicitly stated, it mentions LOOCV as a prevalent strategy.

First Author, Year	Study Location	Study Design	Participant Condition	Sample Size	Outcome Measured	Outcome Time Horizon	AI Model	AI Model Mechanism	Training/ Validation /External Test Splits
			hemorrhage, cardiac arrest, epilepsy, and other medical or neurological disorders						
Bishara et al., 2022	United States	Retrospective cohort	All encounters in patients aged 18 and over undergoing surgery or a procedure requiring anesthesia care and staying in the hospital at least overnight	24,885 patients	Postoperative delirium (POD)	7 days	eXtreme Gradient Boosting (XGBoost) algorithm and a neural network	The XGBoost algorithm was trained on the training dataset to predict POD using preoperative risk features. The neural network was also trained on the training dataset to find complex interactions between variables. Both models were fine-tuned using the validation dataset.	Training: 80% / Test: 20% (with a validation set of 20% of the training data)
Lee et al., 2022	Belgium	Observational cohort	Individuals with severe brain injuries, including those in a vegetative state (UWS), minimally conscious state (MCS), and MCS with some signs of consciousness	34 patients with severe brain injury and 15 patients under general anesthesia	Explainable consciousness indicator (ECI), which is designed to quantify arousal and awareness in altered states of consciousness	During neurocritical care (timeframe not specified)	Convolutional neural network (CNN) with spatiotemporal information	The CNN model works by learning to classify EEG responses to transcranial magnetic stimulation, allowing it to calculate the explainable consciousness indicator (ECI) that disentangles the components of consciousness, specifically arousal and awareness, in altered states of consciousness.	Not specified
Mattia et al., 2022	France	Retrospective cohort	Patients who had experienced a coma induced by cardiac arrest (CA)	29 coma patients and 34 controls	Neurological outcome of the patients, which was assessed 3 months after hospital admission using the coma recovery scale revised (CRS-R)	3 months	Convolutional neural network (CNN)	The CNN in this study processed high-dimensional arrays of multimodal MRI data, including structural MRI (sMRI) and functional MRI (fMRI) data. The CNN utilized convolutional and pooling layers to extract features from the input data and then classify the data to predict neurological outcomes in patients with anoxic-ischemic coma. The CNN was trained to automatically extract relevant features from the MRI data without manual feature engineering, enabling the identification of weak signals in	10-time repeated tenfold cross-validation

First Author, Year	Study Location	Study Design	Participant Condition	Sample Size	Outcome Measured	Outcome Time Horizon	AI Model	AI Model Mechanism	Training/ Validation /External Test Splits
								complex, multimodal neuroimaging datasets.	
Miyazawa et al., 2024	Japan	Retrospective cohort	Patients with COVID-19	273 patients	Incidence of delirium	7 days before and after delirium onset	Natural language processing (NLP) and machine learning (ML) techniques, specifically using logistic regression (LR), support vector machine (SVM), decision tree (DT), and random forest (RF) as ML methods	The NLP and ML techniques were employed to analyze free-text nursing records, enabling the identification of hidden risk factors, such as changes in elimination-related behaviors and conditions, as well as an increased frequency of nurse calls and sensor alerts, which were not represented in a structured format in electronic health records.	Not applicable (statistical analysis)
Muller et al., 2019	Germany	Retrospective cohort	Anoxic-ischaemic disorders of consciousness (AI-DOC) following cardiac arrest	126 patients	Neurological outcome, assessed 1 year after onset using the clinical performance category (CPC) scale	1 year	Logistic regression, support vector machines, nearest neighbors clustering, and naive Bayes classification	The machine learning models were trained on the available data, including serial NSE measurements, demographic and clinical variables, and outcome data. The models were then used to predict the probability of poor neurological outcome in new patients based on their NSE measurements and other variables. The K-nearest neighbors (KNN) imputation method was used to fill in missing data, and the models were compared to standard single-day and difference-between-day methods. The naive Bayes classifier was the most accurate in predicting outcomes.	Not applicable (statistical analysis)
Sun et al., 2019	United States	Observational cohort	ICU patients who met specific criteria, including being 18 years or older, on mechanical ventilation, and having at least one RASS or CAM-ICU	174 patients (RASS), 121 (CAM-ICU)	Level of consciousness (LOC) and delirium	Continuous tracking of the level of consciousness and delirium (timeframe not specified)	Convolutional neural network (CNN) followed by long-short-term memory (LSTM)	The CNN extracted valuable information from each 4-second segment of the EEG waveform, while the LSTM provided temporal context. The model was trained to perform ordinal regression for RASS and binary classification for CAM-ICU. The ordinal regression learned a continuous "z-score" and	Testing patients pooled from all folds

First Author, Year	Study Location	Study Design	Participant Condition	Sample Size	Outcome Measured	Outcome Time Horizon	AI Model	AI Model Mechanism	Training/ Validation /External Test Splits
			assessment during EEG recording					thresholds, which could be used to discretize the z-score into RASS levels. The binary classification outputted the probability of delirium (CAM-ICU positive).	
Wong et al., 2018	United States	Retrospective cohort	Newly hospitalized patients without known cognitive impairment	18,223 admissions	Prediction of incident delirium risk based on electronic health data available on admission	Hospital-acquired delirium (timeframe not specified)	Gradient boosting machine, penalized logistic regression, random forest, and age >79 years (AWOL), failure to spell world backward, disorientation to place, and higher nurse-rated illness severity	The machine learning models in the study analyzed 796 clinical variables from electronic health records within 24 hours of admission to predict the risk of incident delirium in newly hospitalized patients without a known history of cognitive impairment.	Training: 14,227 / Test: 3,996
Guo et al., 2022	China	Retrospective cohort	Intracerebral hemorrhage (ICH) patients	751 patients	Functional outcome at 3 months after ICH, as assessed by the modified Rankin Scale (mRS)	90 days	Random forest algorithm	Constructing multiple decision trees and combining their predictions to make a final prediction. The algorithm incorporated 12 selected variables, including inflammation, changes in patient condition during treatment, and various clinical and laboratory parameters, to predict the likelihood of a poor functional outcome 3 months after ICH.	Not specified
Kolisnyk et al., 2023	Canada	Observational cohort	Unresponsive critically ill patients with severe brain injuries of various etiologies, including traumatic brain injury, stroke, cardiac arrest, infection, hepatic failure, and status epilepticus	25 patients	Neurologic recovery of the participants within 6 months following their brain injury, assessed using the Glasgow Outcome Scale (GOS) score	6 months	Nearest centroid classifier	This model computed the mean value (centroid) of training examples for both labels (good and bad outcomes) across each dimension (ten resting-state networks from MRI) in the data. Then, on the test set, the Euclidean distance between each centroid and the patient data was computed, and test examples with smaller distances to the centroid were assigned the corresponding label.	Not applicable (statistical analysis)

This page has been intentionally left blank