

Artificial intelligence approach for predicting suicide-related behaviour in emergency departments

Received: 17 April 2026

Accepted: 22 July 2026

Published online: 02 August 2026

Cite this article as: Mokheleli T., Makaba T., Ndayizigamiye P. *et al.* Artificial intelligence approach for predicting suicide-related behaviour in emergency departments. *Sci Rep* (2026). <https://doi.org/10.1038/s41598-026-64073-y>

Tsholofelo Mokheleli, Tebogo Makaba, Patrick Ndayizigamiye & Nompumelelo Ndlovu

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Artificial Intelligence Approach for Predicting Suicide-Related Behaviour in Emergency Departments

Tsholofelo Mokheleli^{1*} [0000-0001-9046-4084], Tebogo Makaba² [0000-0003-3710-2513],
Patrick Ndayizigamiye¹ [0000-0001-5721-6070], Nompumelelo Ndlovu³ [0000-0003-0604-5114]

^{1*} Centre for Applied Data Science, University of Johannesburg, Johannesburg, 2006, South Africa.

² Department of Applied Information Systems, University of Johannesburg, Johannesburg, 2006, South Africa.

³ Department of Nursing, University of Johannesburg, Johannesburg, South Africa.

*Corresponding author(s). E-mail(s): tsholofelom@uj.ac.za;

Contributing authors: tbokaba@uj.ac.za; pndayizigamiye@uj.ac.za; mpumin@uj.ac.za.

Abstract: Suicide remains a major public health challenge, with emergency departments (EDs) representing a critical point of contact for individuals at risk. However, conventional suicide screening approaches are often time-intensive, disclosure-dependent, and difficult to implement consistently in high-acuity settings. This study developed and evaluated an artificial intelligence (AI)-based model for predicting suicide-related behaviour (SRB) within 30 days of ED presentation using routinely collected structured triage data available at admission.

Five supervised learning models, logistic regression, random forest, extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and a multilayer perceptron (MLP), were developed and evaluated using patient-level train-validation-test splits. Performance was assessed using a multidimensional framework including area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), Recall, F2-score, Matthews Correlation Coefficient (MCC), Balanced Accuracy, threshold optimisation, decision-curve analysis, bootstrap confidence intervals, and SHapley Additive exPlanations (SHAP).

Among the evaluated models, LightGBM showed the most favourable balance between overall discrimination, minority-class discrimination, sensitivity, and decision stability, achieving an AUROC of 0.88, an AUPRC of 0.25, a Recall of 0.79, and an F2-score of 0.41 on the independent test set. Threshold optimisation and decision-curve analysis further supported the model's clinical utility, while SHAP analyses identified prior psychiatric history, age, and physiological variables as important contributors to risk prediction.

These findings demonstrate that short-term SRB risk can be predicted using structured ED triage information available at presentation. Rather than replacing clinicians, the proposed model aims to provide transparent, scalable decision support to support prioritisation, follow-up planning, and early intervention. Although external validation and prospective studies remain necessary, the study provides a deployment-oriented framework for integrating AI-assisted suicide risk prediction into emergency care workflows.

Keywords: Artificial Intelligence; Clinical Decision Support; Emergency Department; Explainable Machine Learning; Suicide Related Behaviour

1. Introduction

Suicide constitutes a persistent global public health challenge and remains one of the leading causes of premature mortality worldwide (Erdem et al., 2025), affecting populations across diverse contexts (Hughes et al., 2025; Machado, 2025). Beyond individual loss, suicide puts a significant psychological and economic strain on communities and healthcare infrastructure. Clinically, suicidal behaviour is recognised as a critical phenomenon that may occur across a range of mental disorders and clinical presentations, rather than as a disorder confined to a single diagnostic category (American Psychiatric Association, 2022). Among healthcare settings, the emergency department (ED) often serves as the first, and sometimes only, point of contact for individuals experiencing acute psychological distress (Nugent et al., 2024). Individuals in suicidal crisis often present to EDs with co-occurring conditions such as substance intoxication, injury, or acute psychiatric symptoms (Rheinberger et al., 2022).

Global evidence indicates that the risk of suicide or suicide attempt is highest within 30 days following discharge from an ED or inpatient psychiatric unit (Haglund et al., 2019; Thompson et al., 2025). This period represents a vulnerable transition from structured clinical supervision to less monitored outpatient care. Although patients presenting with acute suicidality are often admitted for stabilisation, discharge is frequently associated with

reduced monitoring intensity and inconsistent follow-up. Up to 70% of individuals discharged after a suicide attempt fail to attend their first outpatient appointment, and approximately 37% of individuals without a prior psychiatric diagnosis who die by suicide had visited an ED within a year preceding death (Department of Public Health and Human Services, 2025; Olfson et al., 2021). These findings position the ED not merely as a point of care, but as a critical intervention window where timely identification of suicide-related behaviour (SRB) could enable early intervention and potentially prevent subsequent harm.

Despite this, suicide risk assessment in EDs remains predominantly reliant on subjective and time-intensive screening approaches, including clinician judgment, structured interviews, and standardised tools such as the Ask-Suicide Screening Questions (ASQ), Suicidal Ideation Questionnaire (SIQ), and Columbia-Suicide Severity Rating Scale (C-SSRS) (Scudder et al., 2022). While clinically validated and aligned with the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR) conceptualisations, these approaches depend heavily on patient self-disclosure and clinician availability, both of which are constrained in a high-acuity emergency environment (Dobscha et al., 2023; Peralta, 2024). In addition, psychological distress often manifests indirectly through physiological, behavioural, or contextual indicators that are not fully captured by traditional screening tools (Kleiman et al., 2021). The resulting under-identification of at-risk individuals persists, limiting the ability of clinicians to prioritise patients based on risk severity in real-time.

Recent advances in artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), present an opportunity to augment clinical decision-making in suicide risk assessment (McHugh & Large, 2020; Shah, 2022). Predictive models trained on structured triage data, such as vital signs, demographic information, visit context, and psychiatric history, can estimate the likelihood of SRB prior to formal assessment. Unlike post-hoc approaches relying on clinical notes or discharge summaries (Garriga et al., 2022) structured triage data is available at the point of admission, enabling real-time risk estimation within ED workflows (Kishore et al., 2023). Such models have the potential to identify subtle but clinically meaningful patterns associated with elevated suicide risk (El Arab & Al Moosa, 2025). When combined with explainability techniques, these systems can provide transparent, interpretable insights that complement, rather than replace, clinical expertise (Joyce et al., 2023).

However, despite growing interest in AI-driven suicide prediction, several limitations constrain the clinical applicability of existing approaches. A substantial body of research has focused on unstructured data sources, particularly social media and textual clinical narratives, using linguistic features, sentiment analysis, and topic modelling to infer psychological states (Homan et al., 2022; Mokheleli et al., 2026; Rabani et al., 2023). While these approaches demonstrate predictive capability, they are often limited by privacy concerns, sampling biases, and the absence of clinically validated ground truth labels (Hanna et al., 2025; Humbert-Droz et al., 2022). DL models applied to such data frequently require large-scale datasets and complex representations, reducing their practicality in structured clinical environments and highly imbalanced datasets typical of ED settings (Gaur et al., 2024).

In contrast, structured electronic health records (EHRs) and ED triage variables, including vital signs, demographic indicators, presenting complaints, and psychiatric history, are routinely collected and available during patient intake (Kishore et al., 2023; Sithiprawiat et al., 2025). Despite their availability and clinical relevance, these data are rarely leveraged in a systematic, model-driven manner for suicide risk assessment in real-time ED workflows (Bentley et al., 2025). Existing studies that incorporate structured triage data remain limited in number and scope (Abugroun et al., 2025; Sithiprawiat et al., 2025), leaving a gap in understanding whether such data can support reliable and scalable early risk prediction. Addressing this gap is necessary for developing deployable decision-support systems that align with routine clinical processes. Another limitation lies in the lack of model interpretability. Although many AI models achieve high discrimination performance, their internal reasoning is often difficult to understand, limiting trust and adoption in clinical settings where transparency and accountability are essential. Black-box models introduce ethical and operational concerns, particularly in high-stakes domains such as mental health care.

To address these limitations, this study develops an interpretable AI-based triage model for predicting SRB within 30 days of ED admission. Using structured clinical data from the Medical Information Mart for Intensive Care (MIMIC-IV) database, the study evaluates multiple supervised learning models, including interpretable linear models, ensemble tree-based methods, and a deep neural network architecture, to identify those that balance predictive performance with clinical interpretability. Model explanations are generated using SHapley Additive

exPlanations (SHAP), enabling transparent assessment of how individual demographic, clinical, and contextual features contribute to estimated suicide risk at both global and patient-specific levels.

In this study, *prediction* and *screening* are conceptually distinct. Screening refers to clinician- or tool-based processes (such as ASQ, C-SSRS) that identify current or self-reported suicidal ideation or behaviour at the point of assessment, whereas prediction involves using computational models to estimate the probability of a future outcome based on available data. This study focuses on prediction; the proposed model is intended as a clinical decision-support tool that estimates the probability of documented SRB within 30 days using information available at ED presentation. It is designed to complement, rather than replace, established suicide screening instruments and clinician assessment by supporting early risk stratification and prioritisation.

The study is guided by the following research questions (RQs):

1. Which AI model provides the best balance between predictive performance, minority-class detection, clinical utility, and interpretability?
2. How interpretable and clinically actionable are these models in explaining SRB risk?
3. Which clinical and contextual features contribute most strongly to SRB prediction at triage?

This study makes several contributions to clinical AI and emergency mental healthcare. First, it reframes suicide risk prediction as a presentation-time decision problem by demonstrating that short-term SRB risk can be estimated using routinely collected structured data available at ED triage. Second, it establishes a reproducible framework for integrating ED and hospital records from MIMIC-IV and evaluates model performance using patient-level stratified partitions that better reflect real-world deployment conditions and minimise information leakage. Third, it introduces a clinically grounded evaluation model for SRB prediction in ED settings that extends beyond conventional performance assessment. As shown in the literature review section, many previous studies that rely primarily on discrimination metrics, the proposed model integrates minority-class discrimination (AUPRC), overall discrimination (AUROC), sensitivity, decision stability (MCC), threshold optimisation, bootstrap confidence intervals, patient-level validation, and explainability to support clinically meaningful model selection and deployment. Fourth, the study demonstrates that high-performing ensemble models can remain clinically interpretable by integrating SHAP-based global and local explanations. By treating explainability as a design requirement rather than a post hoc addition, the model provides transparent, patient-specific insights that support clinical accountability and trust. The remainder of this paper is structured as follows: Section 2 reviews related literature on AI in suicide prediction and triage modelling; Section 3 describes the methodology; Section 4 presents the results, discusses implications, outlines limitations and future work; and Section 5 concludes the study.

2. Literature review

This literature review synthesises existing evidence on suicide risk assessment in ED settings, focusing on the role of EDs as critical points of contact for SRB, the limitations of existing triage-based assessment approaches, the predictive value of routinely collected structured ED data, advances in AI for suicide risk prediction, associated evaluation challenges in rare-event clinical settings, and the growing importance of explainability for clinical adoption. These themes establish the conceptual and empirical foundation for developing an interpretable, structured-data-driven AI model for SRB prediction at ED triage.

Suicide-Related Behaviour in Emergency Department Settings

EDs are recognised as critical settings for suicide prevention as they frequently represent the first, and sometimes only, healthcare contact for individuals experiencing suicidal crises. SRB presentations show identifiable demographic, temporal, and clinical patterns that are observable through routine triage processes. For example, Wang et al. (2026) reported that among adolescents presenting following suicide attempts, drug overdose was the predominant method, presentations were disproportionately female, and arrivals clustered during evening hours. Similarly, McCabe et al. (2026) showed that overdose-related self-harm involving opioids and stimulants frequently required hospital admission and intensive care. These findings highlight that ED clinicians operate under considerable time pressure while simultaneously managing behavioural crises and acute physiological instability.

SRB presentations are also influenced by broader contextual and environmental factors. Cortellazzo Wiel et al. (2025) reported increases in alcohol- and substance-related ED presentations among adolescents during and after

the COVID-19 pandemic, proving how societal stressors can indirectly shape suicide vulnerability. Beyond the acute encounter, evidence consistently identifies the post-contact period as a critical window for recurrent suicidal behaviour. Using logistic regression, Shi et al. (2026) demonstrated that the recency of a previous suicide attempt was a strong and graded predictor of reattempt within one year, with the highest risk observed immediately after a recent attempt. Similarly, Carson et al. (2024) showed that structured EHR variables could predict SRB within 90 days following psychiatric discharge, supporting the notion that suicide risk extends beyond a single encounter and evolves over time.

Evidence further suggests that many individuals at risk do not present with explicit psychiatric complaints. Thomas et al. (2025) argued that trauma exposure, impulsivity, and behavioural sensitisation contribute to suicide vulnerability and noted that many patients engaging in suicidal behaviour lack a documented psychiatric diagnosis. Therefore, reliance on self-disclosure or psychiatric history alone may underestimate risk at triage. This challenge is also evident in paediatric populations. Pagliaccio et al. (2026) found that suicide thoughts and behaviours were common among children presenting with behavioural concerns, with prior SRB strongly predicting repeat ED visits. These findings suggest that suicide vulnerability is often embedded within broader behavioural and contextual presentations rather than isolated psychiatric complaints.

Although early identification is essential, detection alone is insufficient without effective follow-up. Sano et al. (2026) reported that family psychoeducation following ED admission was associated with reduced recurrence among patients who completed at least two sessions, highlighting the importance of continuity of care after discharge. The literature positions the ED as a high-leverage environment for suicide prevention, characterised by identifiable presentation patterns, high clinical acuity, hidden psychiatric vulnerability, and elevated short-term recurrence risk. These characteristics highlight the importance of developing triage-compatible approaches that can identify risk using information available at the point of presentation.

Limitations of Suicide Risk Assessment at ED Triage

Despite the ED's central role in suicide prevention, triage-based suicide risk assessment remains constrained by several structural and operational limitations. Many individuals at risk of suicide do not present with explicit psychiatric complaints or have a documented mental health history, limiting the effectiveness of approaches that depend on self-disclosure or prior diagnoses. Thomas et al. (2025) argued that trauma exposure and behavioural sensitisation may underlie suicidal crises even in the absence of recognised psychiatric illness. Similar challenges have been reported in paediatric populations, where Pagliaccio et al. (2026) found that important drivers of suicide thoughts and behaviours, including family conflict, bullying, and social stressors, were often documented retrospectively rather than captured during triage. These findings suggest that conventional screening approaches may under-detect risk, particularly when psychosocial factors remain hidden during the initial encounter.

Limitations in routine coding further compound these challenges. Edgcomb et al. (2024) demonstrated that reliance on diagnosis codes alone missed a substantial proportion of suicide-related ED visits, whereas random forest (RF) models incorporating structured triage variables, such as chief concerns and vital signs, substantially improved detection. Similarly, Robinson et al. (2020) showed that the Victorian Self-Harm Monitoring System identified approximately 25% more self-harm presentations than administrative coding alone. These findings indicate that coding-based surveillance underestimates the burden of SRB and highlight the value of routinely collected structured clinical information.

Although narrative clinical notes can capture hidden psychosocial risk factors, their use introduces additional challenges. Tsui et al. (2021) reported that combining structured data with NLP-derived features improved the prediction of first-time suicide attempts, achieving an AUC of 0.932 with XGBoost. However, the study also highlighted concerns regarding computational complexity, single-site generalisability, and potential data leakage when notes describing the suicide attempt were included. These limitations reduce the feasibility of NLP-intensive approaches for real-time triage support.

Beyond predictive performance, successful implementation depends on clinician trust and usability. Interian et al. (2025) argued that precision suicide prevention requires decision-support tools that align with clinical reasoning and support personalised interventions. In fast-paced ED environments, black-box systems that lack transparency are unlikely to influence triage decisions regardless of their predictive accuracy. The literature suggests that ED suicide risk assessment is constrained by reliance on self-disclosure, incomplete coding, operational pressures, and limited interpretability, motivating the development of transparent, triage-compatible models that use routinely available structured data.

Structured ED Data for Suicide Risk Prediction

Routinely collected structured ED data provide a practical foundation for suicide risk prediction as they are consistently available, ethically compliant, and naturally integrated into real-time clinical workflows. Edgcomb et al. (2024) provided strong empirical support for this approach, demonstrating that models incorporating structured variables such as chief concerns, demographics, and vital signs substantially outperformed ICD-only approaches across demographic groups. Specific structured features are also supported by converging clinical evidence. Shi et al. (2026) showed that the recency of a previous suicide attempt is a powerful predictor of subsequent risk, while McCabe et al. (2026) demonstrated that overdose characteristics and hospital disposition reflect the severity and lethality of self-harm presentations. Wang et al. (2026) further showed that demographic and temporal patterns of SRB are embedded within routine triage data, supporting their role as predictive rather than merely descriptive variables.

Contextual indicators captured during routine care may also provide valuable risk signals. Cortellazzo Wiel et al. (2025) reported shifts in adolescent intoxication-related presentations during periods of societal disruption, suggesting that substance-related encounters can act as markers of heightened psychological vulnerability. At the same time, Pagliaccio et al. (2026) and Tsui et al. (2021) highlighted that some psychosocial drivers are embedded in narrative text, illustrating the trade-off between contextual richness and operational feasibility. System-level evidence further suggests that ED workflows already generate clinically deployable data streams. Robinson et al. (2020) demonstrated that routine ED data can support real-time surveillance, while Carson et al. (2024) showed that structured history and healthcare utilisation variables can predict short-term post-discharge suicide risk.

The literature supports structured ED data as a scalable, clinically relevant, and ethically appropriate foundation for suicide risk prediction. Although structured variables may not capture the full psychosocial context available in clinical narratives, their availability, standardisation, and compatibility with existing workflows make them particularly suitable for deployment-oriented suicide risk assessment frameworks.

AI-Based Suicide Risk Prediction: Performance and Evaluation Challenges

Across suicide prediction studies, ensemble-based ML models, particularly RF, XGBoost, and boosting algorithms, have consistently demonstrated strong predictive performance. Tsui et al. (2021) provided an early benchmark, reporting an AUC of 0.932 using XGBoost to predict suicide attempts across multiple time horizons. Similar findings have been reported in adolescent populations. Li et al. (2025) found that LightGBM achieved the highest performance (AUC = 0.79) among four competing models, while Y. Wang et al. (2025) showed that RF performed best in predicting suicide attempts among adolescents with depression, achieving AUC values of 0.736 in males and 0.720 in females. Ntakolia et al. (2025) also reported LightGBM as the best-performing model (AUC = 0.895) for identifying clinically significant suicide risk in patients with major depressive disorder. These studies suggest that boosting and ensemble methods are effective for structured mental health data, although performance may vary across populations and outcome definitions.

However, predictive performance depends not only on the algorithm but also on how model utility is evaluated. Ahn et al. (2025) showed that distinguishing suicide attempts from suicidal ideation remains challenging, with LR achieving AUC values ranging from 0.705 to 0.944 depending on the target population. Similarly, Kim and Lee (2025) reported that RF achieved a higher recall (0.76) than Logistic Regression (LR) (0.48) when predicting suicide risk among adults with ADHD and depression, while H. Wang et al. (2025) obtained a recall of 97% using gradient boosting machines. These findings show that no single metric adequately captures clinical utility and that model evaluation requires balancing discrimination with sensitivity to rare events. In suicide prevention, where missed cases may have serious clinical consequences, model evaluation must extend beyond conventional discrimination metrics to incorporate minority-class detection, decision stability, and clinically informed threshold selection that reflects real-world clinical decision-making.

Beyond discrimination, several studies have examined the clinical characteristics associated with suicidal behaviour. Bazrafshan and Sayehmiri (2024) identified older age, lower education, socioeconomic disadvantage, and physical methods of self-harm as important predictors, with RF achieving an AUC of 0.79. Meanwhile, Nordin et al. (2025) demonstrated that multiclass risk stratification combined with XAI could achieve an accuracy of 0.81 and a recall of 0.80, demonstrating a shift from binary prediction toward more risk categorisation and interpretable decision support.

Some studies have also explored DL approaches. (Gaur et al., 2024) developed an artificial neural network for identifying students at risk of suicidal tendencies, while Tadesse et al. (2019) employed CNN-LSTM architectures to detect suicidal ideation from Reddit posts. Although these approaches demonstrate the potential of DL, they have primarily been applied to unstructured or specialised datasets and typically require larger sample sizes and more complex feature representations. Accordingly, their advantages over well-tuned ensemble and linear models remain less established in structured clinical settings. Overall, the literature suggests that model selection should be guided not only by predictive accuracy but also by outcome definition, class imbalance, clinical interpretability, and deployment requirements.

Explainability in Suicide Risk Prediction

Explainability has emerged as a critical requirement for AI-assisted suicide risk prediction because clinicians must understand the rationale underlying model outputs before incorporating them into decision-making. Several studies have demonstrated that SHAP provides clinically meaningful explanations when applied to structured healthcare data. For example, Edgcomb et al. (2024) showed that SHAP explanations integrated with ED models identified chief concerns and age as key contributors to risk, illustrating how routinely collected variables can be translated into interpretable clinical signals.

Beyond global feature ranking, recent studies have shown that SHAP can reveal subgroup differences and patient-specific mechanisms. Li et al. (2025), Ahn et al. (2025), and Y. Wang et al. (2025) demonstrated that SHAP explanations can identify progression pathways from suicidal ideation to attempts, quantify interaction effects, and characterise gender-specific risk factors. Nordin et al. (2025) further extended these approaches to multiclass suicide risk stratification, using SHAP summary and force plots to provide both population-level and individual-level explanations. Similarly, Ntakolia et al. (2025) showed that SHAP is compatible with high-performing boosting models, identifying depression severity and attachment insecurity as major risk drivers, while employment and family support acted as protective factors.

Although simpler feature importance approaches remain useful (Kim & Lee, 2025), SHAP offers richer global and local explanations that can support case-level interpretation and auditability. Thomas et al. (2025) argued that model explanations should align with plausible psychiatric and behavioural mechanisms rather than merely highlight statistical associations. These studies establish explainability as a critical bridge between predictive performance and clinical trust, supporting the integration of AI-assisted suicide risk prediction into emergency care workflows.

Synthesis and Research Gap

The reviewed literature converges on three consistent insights: (i) EDs are a high-leverage point for suicide prevention due to recognisable SRB presentation patterns and elevated short-term risk following ED presentation or discharge; (ii) triage-time detection is operationally difficult as screening is time-intensive, disclosure-dependent, and inconsistently applied; and (iii) many prior predictive efforts are constrained by either retrospective signals (for example, post-assessment ICD coding) or deployment-heavy signals (for example, NLP over clinical narratives), which limit immediate decision support in high-throughput ED settings. What remains practically under-served in this evidence base is a triage-compatible approach that can operate at presentation using data streams in the ED that are already standardised, routinely captured, and governed for operational use.

This study bridges that triage-time gap by demonstrating that structured ED triage variables available on arrival (demographics, vitals, time context, brief chief-complaint-derived flags, and prior psychiatric history markers) contain sufficient signal to support SRB risk prediction within 30 days, without depending on full narrative text or delayed documentation. In doing so, the study moves beyond descriptive ED epidemiology and post-discharge modelling by framing the problem as a presentation-time prediction task and validating it using patient-level splits to reduce leakage risk, conditions that more closely resemble deployment realities for clinical decision support.

A closely related gap concerns clinical usability and trust, which the literature repeatedly identifies as barriers to adoption when models are unclear or when explanation is treated as an afterthought. This study addresses that gap by integrating explainability into the evaluation framework through SHAP-based global and local explanations, demonstrating that interpretability can be treated as a design requirement rather than a post hoc addition. This supports clinician-facing accountability and transparency while enabling the use of high-performing ensemble models. This combination demonstrates that explainability can be treated as a design requirement in an ED triage model, not merely a post-hoc add-on, supporting clinician-facing accountability at the point of care.

The literature also highlights a misalignment between common evaluation practices and the clinical realities of rare-event prediction, where false negatives carry substantial consequences, and no single metric adequately captures model utility. Although AUROC remains the most frequently reported metric in suicide prediction studies and provides an overall measure of discrimination, it may present an overly optimistic view of performance in highly imbalanced settings where positive outcomes are rare. Metrics such as AUPRC, Recall, F2-score, and MCC provide complementary perspectives on minority-class detection, the sensitivity-precision trade-off, and decision stability (Martin et al., 2025; McDermott et al., 2024; Sitthiprawiat et al., 2025). This study addresses that evaluation gap by adopting a clinically grounded multi-dimensional evaluation framework that integrates overall discrimination (AUROC), minority-class discrimination (AUPRC), sensitivity (Recall), decision stability (MCC), threshold optimisation, decision-curve analysis, bootstrap confidence intervals, patient-level validation, and explainability. Rather than selecting models based on a single metric, the model supports selection based on overall clinical utility and operational feasibility. In this way, the study contributes a pathway from predictive performance to deployment-oriented decision-making in ED settings, where the relevant question is not merely whether a model predicts accurately, but whether it provides meaningful clinical benefit while maintaining acceptable alert burden and avoiding missed high-risk patients.

3. Methodology

This section presents the methodological framework for developing and evaluating an explainable AI model to estimate short-term SRB risk following ED presentation. The framework was designed to support presentation-time risk stratification using routinely collected structured triage data while maintaining alignment with real-world clinical workflows.

Data Source

This study used the MIMIC-IV (version 3.1) database (Johnson et al., 2024), a publicly available but credential-restricted clinical data resource developed by the Massachusetts Institute of Technology's (MIT) Laboratory for Computational Physiology in collaboration with Beth Israel Deaconess Medical Centre. MIMIC-IV integrates multiple relational modules containing comprehensive, de-identified health records from patients admitted to hospitals and ICUs between 2008 and 2019. The database captures detailed information on demographics, admission characteristics, physiological measurements, laboratory results, diagnoses, procedures, medications, and outcomes. For this study, data were drawn from the hospital (*hosp*) module to provide longitudinal clinical context, and from the emergency department (*ED*) module to capture triage-level variables relevant to acute presentations.

Access to MIMIC-IV and MIMIC-IV-ED was obtained through the PhysioNet platform and the Google BigQuery public dataset interface, both of which support secure, structured querying of large-scale healthcare data. As the database contains sensitive health information, access is restricted to credentialed researchers who have completed the required Collaborative Institutional Training Initiative (CITI) course on human-subjects research ethics and have agreed to the PhysioNet Data Use Agreement. Ethical use of the data is governed by the Institutional Review Boards (IRBs) of MIT and Beth Israel Deaconess Medical Centre, which have approved MIMIC-IV as a fully de-identified dataset compliant with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbour provisions. As such, this secondary analysis did not require additional IRB approval.

Cohort Definition

The cohort was derived using a structured query written in Structured Query Language (SQL) and executed on the Google BigQuery instance of the MIMIC-IV dataset. The logic integrated data across the *hosp* and *ED* module; the tables *diagnoses_icd*, *admissions*, *patients*, *edstays*, and *triage* were used to construct a unified dataset of ED encounters. The SQL pipeline was executed in modular stages to ensure quality, reliability and error control.

As shown in Table 1, the algorithm began by identifying suicide-related admissions from the *diagnoses_icd* table using both International Classification of Diseases (ICD) versions 9 and 10. As summarised in Table 1, the outcome definition included ICD-9 codes for suicide and self-inflicted injury (E95*) and ICD-10 codes representing intentional self-harm (X71-X83), poisoning and toxic effects associated with self-harm (T36-T65 with intent modifiers), suicidal ideation (R45.851), and suicide attempt (T14.91). These codes were used to label encounters as positive for SRB (SRB = 1). Although previous studies have shown that ICD coding may under-identify some suicide-related presentations, particularly when compared with enhanced surveillance or clinician-reviewed approaches (Edgcomb et al., 2024; Robinson et al., 2020), ICD-based definitions remain the most widely

used and reproducible method for constructing large-scale retrospective suicide prediction cohorts and therefore provide a practical reference standard within MIMIC-IV. The overall cohort construction and feature engineering process are outlined in Table 2. All remaining admissions that did not contain any SRB-related ICD codes during the study period were assigned to the negative cohort (SRB = 0). These admissions comprised the complete set of eligible ED encounters that satisfied the same inclusion criteria and linkage constraints (ED stay within 24 hours of admission) but lacked SRB diagnostic codes. No additional exclusion criteria were applied beyond the absence of SRB codes, ensuring that the negative cohort reflected routine ED presentations encountered in real-world triage settings.

Table 1. ICD-9 and ICD-10 code mapping used to identify cases of SRB

ICD Version	Codes Used for SRB Identification	Description
ICD-9-CM	E95*	Suicide and self-inflicted injury
ICD-10-CM	X71–X83	Intentional self-harm
ICD-10-CM	T36–T65 with intent modifiers	Poisoning and toxic effects associated with self-harm
ICD-10-CM	R45.851	Suicidal ideation
ICD-10-CM	T14.91	Suicide attempt

To minimise temporal leakage, predictor variables were restricted to information available at or before the index (current) ED presentation. ED stays were linked to hospital admissions only when the ED intake time occurred within the 24-hour window preceding admission. Predictor variables included triage vital signs, demographic characteristics, temporal context variables, selected non-SRB chief complaint categories, and psychiatric and substance-use history indicators derived exclusively from diagnoses recorded during admissions preceding the index encounter. Therefore, no diagnoses, discharge information, procedures, medications, laboratory results, or clinical notes generated after the index triage encounter were used as predictors. The SRB outcome was defined independently using ICD-9 and ICD-10 diagnosis codes associated with the linked hospital admission. In addition, patient-level partitioning was employed during model development to minimise information leakage and better approximate real-world deployment. Thus, the model should be interpreted as estimating the risk of a documented SRB event within 30 days using information available at ED presentation, rather than exploiting post-triage clinical information.

Each admission was linked to its corresponding ED stay (*edstays*) and triage record (*triage*) using the *subject_id* and *hadm_id* identifiers, restricted to ED encounters occurring within 24 hours before hospital admission. From the triage records, routinely collected physiological and contextual variables were extracted, including temperature, heart rate, respiratory rate, oxygen saturation, systolic and diastolic blood pressure, and the free-text chief complaint. These variables served as the basis for subsequent feature engineering and model development, using information available in the ED presentation.

Direct suicide-related indicators (*flag_direct_srb*) captured explicit references to suicidal intent or self-harm documented during ED triage, including terms such as *suicidal ideation*, *self-harm*, and *attempt* (Radunz et al., 2025). Overdose-related indicators (*flag_overdose*) represented presentations involving overdose, poisoning, or ingestion, which are recognised pathways associated with suicidal behaviour in emergency settings (Tu et al., 2024). Psychiatric crisis indicators (*flag_psych_crisis*) captured acute psychological distress, as reflected in terms related to psychiatric emergencies and crises, while accounting for common spelling variations encountered in routine triage documentation (Degirmenci & Mercan, 2025). These rule-based indicators were used to characterise common presentation patterns and reduce variability arising from abbreviations and inconsistent wording in free-text chief complaints.

Demographic variables (*age* and *gender*) were obtained from the *patients* table, while psychiatric history indicators, including prior mood disorders (*prior_flag_mood*), anxiety disorders (*prior_flag_anxiety*), and substance use disorders (*prior_flag_sud*), were derived exclusively from diagnoses recorded in hospital admissions preceding the index encounter using the *diagnoses_icd* table. The final cohort comprised 204,470 ED encounters corresponding to 106,714 unique patients. Among these encounters, approximately 3.4% were labelled as experiencing SRB within 30 days of the index ED presentation, while the remaining 96.6% constituted the non-SRB cohort (see Figure 1). This distribution reflects the inherent rarity of SRB in routine ED populations and presents a clinically relevant class-imbalance challenge, requiring evaluation strategies that prioritise minority-class detection alongside overall discrimination.

Table 2. Algorithm for cohort and feature construction for 30-day SRB prediction using MIMIC-IV

Input:
Admissions A, Diagnoses D, ED stays E, ED triage T, Patients P
Output:
Dataset C with structured ED triage predictors and binary outcome <code>srb_outcome_30d</code>
1: Identify all SRB-related admissions:
2: <code>SRB_EVENTS</code> $\leftarrow$ admissions with ICD-9 E95* or ICD-10 X71-X83, T36-T65 with intentional self-harm modifiers, R45.851, or T14.91
3: Store <code>subject_id</code> , <code>hadm_id</code> , and SRB admission time
4: Construct index ED cohort:
5: For each hospital admission $a \in A$:
6: Link to ED stay $e \in E$ where $e.subject_id = a.subject_id$ and $e.intime$ occurs within 24 hours before $a.admittime$
7: Join triage record $t \in T$ using <code>stay_id</code>
8: Extract triage variables: temperature, heartrate, resprate, o2sat, sbp, dbp, chief complaint, <code>admittime</code> , <code>stay_id</code>
9: Derive chief complaint indicators:
10: Generate candidate SRB-proximal indicators: <code>flag_direct_srb</code> , <code>flag_overdose</code> , <code>flag_psych_crisis</code> , <code>cc_suicide</code> , <code>cc_overdose</code> , <code>cc_psych</code> , <code>cc_substance</code> , <code>cc_psychiatric</code>
11: Generate non-SRB complaint indicators: <code>cc_pain</code> , <code>cc_injury</code>
12: Exclude SRB-proximal complaint variables from final modelling to reduce label leakage risk
13: Add demographic variables:
14: Join with Patients P using <code>subject_id</code>
15: Compute age at admission using: <code>anchor_age + (year(admittime) - anchor_year)</code>
16: Include gender
17: Add prior clinical history:
18: For each index admission r :
19: Search diagnoses from previous admissions for the same subject where $prior_admittime < r.admittime$
20: Create: <code>prior_flag_sud</code> , <code>prior_flag_mood</code> , <code>prior_flag_anxiety</code>
21: Missing prior-history indicators are set to 0
22: Engineer physiological and temporal features:
23: <code>shock_index</code> $\leftarrow$ <code>heartrate / sbp</code>
24: <code>pulse_pressure</code> $\leftarrow$ <code>sbp - dbp</code>
25: Extract <code>admission_hour</code> , <code>admission_dayofweek</code> , <code>admission_month</code>

26:	Create <code>is_night_admission</code> and <code>is_weekend_admission</code>
27:	Create cyclic encodings: <code>hour_sin</code> , <code>hour_cos</code> , <code>month_sin</code> , <code>month_cos</code>
28:	Define 30-day SRB outcome:
29:	For each index ED admission <code>r</code> :
30:	Search <code>SRB_EVENTS</code> for the same <code>subject_id</code>
31:	If any SRB event occurs between <code>r.admittime</code> and <code>r.admittime + 30 days</code> : <code>srb_outcome_30d</code> $\leftarrow$ 1
	Else: <code>srb_outcome_30d</code> $\leftarrow$ 0
32:	Finalise dataset:
33:	Return dataset C containing identifiers, triage-available predictors, engineered features, prior-history indicators, and <code>srb_outcome_30d</code>



Figure 1. Distribution of suicide-related behaviour outcomes in the study cohort.

Prediction Target and Clinical Rationale

The prediction target was a binary outcome, *srb_outcome_30d*, indicating whether a patient experienced documented SRB within 30 days following the index ED presentation. Patients were labelled as positive (SRB =

1) if an SRB-related diagnosis, identified using predefined ICD-9 and ICD-10 codes, was recorded within 30 days of the initial ED encounter. Patients without a documented SRB event during this interval were assigned to the negative class (SRB = 0). The task represents a 30-day post-presentation risk prediction problem based solely on information available at triage.

The 30-day horizon was selected because the period immediately following an ED encounter is recognised as a particularly vulnerable phase for individuals at risk of suicide (Haglund et al., 2019; Thompson et al., 2025). From a clinical perspective, the objective is not to diagnose an ongoing suicidal crisis at triage, but rather to identify patients who are likely to experience documented SRB during the subsequent month. Such predictions may support discharge planning, prioritisation of follow-up, safety interventions, and the allocation of limited mental health resources. Compared with shorter horizons, such as 7 or 14 days, a 30-day follow-up period captures a broader interval of elevated risk while remaining sufficiently close to the index presentation to retain clinical relevance. Therefore, the selected horizon represents a pragmatic compromise between immediate post-discharge risk and longer-term vulnerability. However, using a single prediction horizon is a limitation, and future work should evaluate model performance across multiple clinically relevant time intervals, including 7, 14, 30, and 90 days. Accordingly, the proposed model should be interpreted as a short-term post-ED risk-prediction model rather than a real-time suicide-screening instrument. It is intended to support downstream clinical decision-making, follow-up planning, and prioritisation of assessment resources by estimating the probability of documented SRB occurring within 30 days of ED presentation.

Data Preparation

The preprocessing workflow was designed to prepare structured ED triage variables for model development while preserving the integrity of model evaluation. The feature set included demographics, prior psychiatric history, physiological vital signs, derived physiological measures, temporal admission context, and structured chief complaint indicators. Missing values were retained from the original MIMIC-IV extraction and handled during preprocessing, while categorical variables were appropriately encoded and numerical variables standardised where required. All preprocessing procedures were derived exclusively from the training subset and subsequently applied unchanged to the validation and test sets to prevent information leakage and optimistic bias. The overall process comprised feature preprocessing, imputation, encoding, scaling, and patient-level stratified data splitting.

Missing Data Imputation

Missing values were present only in physiological vital sign variables, including temperature (8.50%), heart rate (6.59%), respiratory rate (7.57%), oxygen saturation (7.57%), systolic blood pressure (6.80%), and diastolic blood pressure (7.01%). All other variables were complete. Given the relatively low proportion of missing values and the clinical nature of these measurements, missing values were imputed using the median from the training dataset. Median imputation was selected as it is reliable against skewed distributions and outliers (Song et al., 2026), which are common characteristics of clinical data (Gao et al., 2025). Given the modest degree of missingness, median imputation was unlikely to substantially affect model performance while preserving the full cohort for analysis.

Feature Engineering (FE)

Two derived physiological indicators were computed from the recorded vital signs. The *shock index* was calculated as the ratio of heart rate to systolic blood pressure ($\text{heart rate} / \text{SBP}$). Shock index is a widely used marker of haemodynamic instability and acute physiological stress (Yu et al., 2025). In the context of SRB, abnormal shock index values are unlikely to represent direct markers of suicide risk but may instead act as proxies for acute medical severity associated with overdose, intoxication, poisoning, or other physiological disturbances present at ED admission (Ghonem et al., 2024).

Pulse pressure was calculated as the difference between systolic and diastolic blood pressure ($\text{SBP} - \text{DBP}$). Pulse pressure reflects cardiovascular dynamics and circulatory status (Lu et al., 2025). Similar to the shock index, it should not be interpreted as a direct indicator of suicide risk, but rather as a proxy for acute physiological compromise or medical severity that may accompany poisoning, overdose, or severe distress requiring emergency care.

Several temporal variables were derived from the admission timestamp to capture operational and behavioural patterns associated with ED utilisation. *admission_hour*, *admission_dayofweek*, and *admission_month* were first

extracted to represent day, weekly, and seasonal patterns in patient presentations. To appropriately model the cyclical nature of time, the hour and month variables were transformed using sine and cosine encoding (*hour_sin*, *hour_cos*, *month_sin*, *month_cos*), which preserve temporal continuity (23:00 and 00:00) and avoid artificial ordering effects that may impair model learning (Khazem & Kanso, 2025). This transformation enables the models to capture periodic temporal relationships more effectively than conventional ordinal representations. Following this transformation, the *admission_hour*, *admission_dayofweek*, and *admission_month* variables were removed to reduce redundancy and potential multicollinearity.

Two additional binary indicators were constructed to represent contextual admission patterns. *is_night_admission* identifies presentations during night shifts, when staffing levels and patient characteristics may differ from those during daytime periods, while *is_weekend_admission* captures admissions on weekends, which may be associated with different patterns of substance use, social stressors, or reduced access to outpatient services. These engineered features allow the models to incorporate temporal context while maintaining interpretability.

Chief Complaint FE

ED chief complaints were originally recorded as free-text entries containing significant variation in spelling, abbreviations, and phrasing (Karagounis et al., 2020). To facilitate their use in predictive modelling, complaint text was standardised through case normalisation and keyword-based matching to derive structured complaint indicators. Initially, several clinically relevant complaint categories were explored, including suicide-related presentations, overdose or poisoning, psychiatric complaints, substance-related presentations, pain-related complaints, and injury-related presentations. However, to minimise the risk of information leakage and ensure that prediction relied on clinically generalisable triage information rather than direct indicators of SRB, complaint features closely related to the outcome were excluded from the final models. Therefore, only non-SRB complaint categories representing pain-related and injury-related presentations were retained and encoded as binary indicators (*cc_pain* and *cc_injury*). This approach enabled the incorporation of routinely available chief complaint information while reducing dependence on explicit suicide-related descriptors and improving the realism of the prediction setting.

Scaling

All continuous variables (vital signs and age) were standardised using z-score normalisation (refer to Equation 1) through Scikit-learn's StandardScaler. Standardisation ensures that each variable contributes equally to distance-based calculations in the model, preventing variables with larger numeric ranges (such as blood pressure) from dominating those with smaller scales (such as temperature) (Buvaneswari et al., 2025).

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where:

- x is the original feature value,
- μ is the mean of the feature and
- σ is the standard deviation of the feature.

Encoding

Categorical variables were encoded to facilitate ML while preserving their clinical interpretation. Gender was one-hot encoded, producing separate binary indicators for female and male. Binary contextual variables, including *is_weekend_admission* and *is_night_admission*, were retained in their original form.

Final Feature Set

The final modelling dataset consisted of 204,470 ED encounters represented by structured triage variables capturing demographic characteristics, physiological vital signs, prior clinical history, temporal admission context, and selected chief complaint indicators. All variables are available at triage and are therefore suitable for

a 30-day SRB risk prediction. The final dataset included 21 predictor variables, a subject identifier and one outcome variable, summarised in Table 3.

Table 3. Final features included in the modelling dataset

Category	Feature	Description
Identifier	subject_id *	Unique patient identifier
Demographics	gender	Patient gender (binary encoded)
	Age	Patient age at admission
Vital Signs	temperature	Body temperature (Fahrenheits)
	heartrate	Heart rate (beats per minute)
	resprate	Respiratory rate (breaths per minute)
	o2sat	Oxygen saturation (%)
	sbp	Systolic blood pressure (mmHg)
	dbp	Diastolic blood pressure (mmHg)
	shock_index	Heart rate divided by systolic blood pressure
	pulse_pressure	Difference between systolic and diastolic blood pressure
Prior clinical history	prior_flag_sud	Prior substance use disorder
	prior_flag_mood	Prior mood disorder
	prior_flag_anxiety	Prior anxiety disorder
Temporal Context	is_night_admission	Indicator for night-time ED admission
	is_weekend_admission	Indicator for weekend ED admission
	hour_sin	Cyclical encoding of admission hour (sine)
	hour_cos	Cyclical encoding of admission hour (cosine)
	month_sin	Cyclical encoding of admission month (sine)
	month_cos	Cyclical encoding of admission month (cosine)
Chief Complaint Indicators	cc_pain	Pain-related complaint
	cc_injury	Injury-related complaint
Outcome Variable	srb_outcome_30d	Suicide-related behaviour within 30 days of ED admission

* subject_id was excluded from model training.

Patient-Level Stratified Split

To prevent information leakage arising from multiple encounters for the same patient, all admissions associated with a given *subject_id* were assigned to a single data partition. The dataset was divided into training (64%), validation (16%), and test (20%) subsets using patient-level stratified sampling to preserve the prevalence of SRB outcomes across splits. This partitioning strategy balances three competing requirements common in clinical prediction studies. First, a sufficiently large training set was required to support model fitting and hyperparameter optimisation. Second, a dedicated validation set was retained to enable model selection and threshold optimisation without contaminating the final evaluation. Third, an independent test set was preserved to provide an unbiased assessment of model performance on previously unseen patients. Three-way data splitting strategies that allocate approximately 60-70% of observations to training and the remainder to validation and testing are widely recommended in clinical ML studies (Riley et al., 2020, 2025), particularly when patient-level independence must be maintained.

By combining patient-level assignment with stratified sampling, model development, optimisation, and evaluation were conducted on mutually exclusive patient cohorts while maintaining consistent outcome distributions across subsets, thereby improving generalisability and model performance (Mokhtari et al., 2025).

Table 4 summarises the distribution of encounters and SRB prevalence across the experimental partitions. The resulting split allocated 130,673 encounters to training, 32,595 to validation, and 41,202 to testing, corresponding to 68,296, 17,075, and 21,343 unique patients, respectively. The prevalence of SRB within 30 days of the index ED encounter was 3.38% in the training set, 3.28% in the validation set, and 3.63% in the test set, closely reflecting the overall prevalence observed in the cohort. This consistency confirms that the class distribution was preserved across all partitions and supports reliable model development and evaluation. Each encounter was represented by

21 structured features, including demographics, vital signs, temporal admission variables, prior psychiatric history indicators, and non-SRB chief complaint categories derived from information available at ED presentation.

Table 4. Dataset distribution and SRB prevalence across experimental splits.

Subset	Encounters (n)	Unique Patients (n)	Positive SRB Cases (n)	Prevalence (%)	Feature Dimension
Training	130,673	68,296	4,416	3.38 %	21
Validation	32,595	17,075	1,070	3.28 %	21
Test	41,202	21,343	1,494	3.63 %	21
Total	204,470	106,714	6,980	3.41%	21

Model Development

Classifier Selection

Five supervised learning algorithms were used to develop predictive models for SRB within 30 days of ED admission: LR, RF, XGBoost, LightGBM, and a MLP neural network. The selection of these models was guided by prior suicide-risk prediction literature and the clinical requirements of emergency triage, rather than by model capacity alone (Nock et al., 2022; Piloni et al., 2024; Xu et al., 2022). The models represent complementary learning paradigms spanning linear, bagging-based, boosting-based, and neural network approaches.

LR was selected as a baseline model as it remains one of the most widely used and clinically accepted approaches for risk stratification in healthcare, including suicide and self-harm prediction studies (Nock et al., 2022; Xu et al., 2022). Its transparent linear structure allows direct interpretation of feature coefficients (Alghazzawi et al., 2025; Xu et al., 2022), supports clinical reasoning and aligns with established statistical risk-assessment practices in emergency medicine. Prior studies have shown that LR often performs competitively with more complex models when applied to structured EHR data (Nock et al., 2022; Tasnim et al., 2024), particularly under strong class imbalance.

RF was included to capture non-linear relationships and feature interactions that linear classifiers cannot model, while maintaining strong performance and interpretability. RF has been frequently applied in suicide risk and mental health prediction using structured clinical data, where it demonstrates stable performance under noisy measurements and heterogeneous patient populations (Piloni et al., 2024; Tasnim et al., 2024). Its ensemble nature reduces variance through bootstrap aggregation and provides an intermediate level of complexity between linear and boosted models (Teimoori & Keyvanpour, 2025).

XGBoost was selected as a representative high-performance ensemble method that has consistently demonstrated strong predictive accuracy in clinical risk prediction tasks, including suicide-related outcomes (Zheng et al., 2022). XGBoost is particularly well-suited to structured tabular data and offers built-in mechanisms for handling class imbalance (Gomes et al., 2025; Sitthiprawiat et al., 2025). Its inclusion enables evaluation of whether increased algorithmic complexity yields meaningful gains over simpler, more interpretable approaches in the ED triage setting.

LightGBM was included as an additional gradient boosting method designed for efficient training and scalability in large, structured datasets. Compared with traditional boosting algorithms, LightGBM employs histogram-based tree learning and leaf-wise tree growth, thereby improving computational efficiency while maintaining high predictive performance (Sutone & Malhotra, 2025). Prior studies in healthcare analytics have demonstrated that LightGBM often achieves competitive or superior performance relative to other ensemble methods when applied to structured clinical data (Ntakolia et al., 2025; Sutone & Malhotra, 2025).

To complement the tree-based approaches, an *MLP neural network* was included to evaluate the potential contribution of DL for structured clinical prediction. MLPs are feedforward neural networks capable of modelling complex non-linear relationships through multiple hidden layers and non-linear activation functions (Oweimieotu et al., 2024; Zheng & Liu, 2024). Although DL architectures are commonly applied to unstructured data such as text or images, MLPs have also been used for tabular clinical prediction tasks to capture interactions that traditional models may not fully represent (Gaur et al., 2024).

LR, RF, XGBoost, LightGBM, and MLP represent complementary modelling strategies spanning linear statistical models, ensemble tree methods, and neural networks. This selection enables a principled comparison of interpretability, predictive performance, and algorithmic complexity within the context of 30-day SRB risk prediction in ED triage settings. This model diversity enables evaluation of whether increased algorithmic complexity yields meaningful improvements over interpretable baseline models for rare-event suicide risk prediction in ED settings.

Weighted Learning

Given the class imbalance present in the dataset, weighted learning was applied to ensure the minority class contributed proportionally to model training. For LR, RF, and LightGBM, imbalance was addressed using *class_weight = balanced*, which automatically assigns weights inversely proportional to class frequencies (Zhu et al., 2025). For XGBoost, class imbalance was handled through the *scale_pos_weight* parameter, defined as the ratio of negative to positive samples (see Equation 2), thereby increasing the penalty for misclassifying SRB cases (Prasher et al., 2025). For MLP, class imbalance was addressed through cost-sensitive learning using inverse-prevalence class weights incorporated directly into the binary cross-entropy loss function during training (Diaz-Vico et al., 2018). MLP was implemented using TensorFlow Keras with early stopping and a manually defined hyperparameter search over network size and learning rate. This ensured that the neural network received the same level of imbalance-specific treatment as the other candidate models. No oversampling or undersampling techniques were applied, thereby preserving the natural distribution of outcomes and avoiding the introduction of synthetic samples or selection bias.

$$\text{scale_pos_weight} = \frac{N_{\text{negative}}}{N_{\text{positive}}} \quad (2)$$

where:

- N_{negative} is the number of negative (non-SRB) samples, and
- N_{positive} is the number of positive (SRB) samples.

Model Tuning and Evaluation

Hyperparameter optimisation was conducted to identify model configurations that best captured the underlying data distribution and improved minority-class detection performance. For LR, RF, XGBoost and LightGBM, hyperparameter tuning was performed using RandomizedSearchCV, which efficiently explores high-dimensional parameter spaces by sampling a fixed number of combinations rather than exhaustively evaluating all possibilities (Almadhoun & Burhanuddin, 2025). This approach was selected to reduce computational cost while retaining strong optimisation performance. Each search used 10 randomly sampled configurations with three-fold cross-validation. As the SRB prediction problem is highly imbalanced, the F2-score was used as the optimisation objective to prioritise sensitivity while maintaining reasonable precision. For the MLP model, a manual hyperparameter search was performed using a predefined search space (see Table 5). Candidate architectures were evaluated using early stopping, and the model with the highest validation AUPRC was selected. Tables 5 and 6 summarise the hyperparameter search spaces and the final selected configurations for all models.

Table 5. Hyperparameter search strategy and search spaces

Model	Search Strategy	Search Space	Optimisation Objective
LR	RandomizedSearchCV (10 iterations, 3-fold CV)	Solver: liblinear, lbfgs, saga; Penalty: L1, L2; C: 10^{-4} to 10^2	F2-score
RF	RandomizedSearchCV (10 iterations, 3-fold CV)	n_estimators: 100, 200, 300; max_depth: 5, 10, 15, 20; min_samples_leaf: 2, 4, 8 min_samples_split: 5, 10, 15; max_features: sqrt, log2	F2-score
XGBoost	RandomizedSearchCV (10 iterations, 3-fold CV)	max_depth: 3, 4, 5, 6; learning_rate: 0.01, 0.05, 0.1; n_estimators: 100, 200, 300; gamma: 0, 0.2, 1;	F2-score

		subsample: 0.7, 0.8, 1; colsample_bytree: 0.6, 0.7, 0.8; min_child_weight: 1, 3, 5	
LightGBM	RandomizedSearchCV (10 iterations, 3-fold CV)	num_leaves: 7, 15, 31; max_depth: 3, 5, 10, -1; learning_rate: 0.01, 0.05, 0.1; n_estimators: 100, 200, 300; min_child_samples: 20, 50, 100; subsample: 0.7, 0.8, 1; colsample_bytree: 0.6, 0.7, 0.8	F2-score
MLP	Manual search + Early stopping	hidden_units: 64, 128; learning_rate: 0.001, 0.01	Validation AUPRC

Table 6. Final selected hyperparameter configurations

Model	Final selected hyperparameter configuration
LR	solver = lbfgs; penalty = L2; C = 10
RF	n_estimators = 200; max_depth = 15; min_samples_split = 10; min_samples_leaf = 2; max_features = log2
XGBoost	subsample = 0.8; n_estimators = 300; min_child_weight = 5; max_depth = 6; learning_rate = 0.1; gamma = 0; colsample_bytree = 0.7
LightGBM	subsample = 0.7; num_leaves = 31; n_estimators = 300; min_child_samples = 20; max_depth = 10; learning_rate = 0.05; colsample_bytree = 0.7
MLP	hidden_units = 64; learning_rate = 0.01; early stopping enabled

Model evaluation incorporated a group of complementary metrics to assess predictive discrimination, clinical utility, and performance stability for the binary prediction of SRB within 30 days of ED admission. Discrimination was assessed using AUROC and AUPRC (Shukla & Singh, 2025; Sitthiprawiat et al., 2025), both of which are appropriate for imbalanced datasets. Clinical utility was assessed using recall, precision, and the F2-score, reflecting the trade-off between identifying high-risk patients and limiting unnecessary alerts. Performance stability was evaluated using the Matthews Correlation Coefficient (MCC) and Balanced Accuracy, which remain informative under class imbalance by accounting for both positive and negative classifications. Given the relatively low prevalence of SRB, model evaluation focused on AUROC, AUPRC, Recall, Precision, F2-score, MCC, and Balanced Accuracy to provide a comprehensive assessment of overall discrimination, minority-class detection, and decision stability. Rather than relying on a single metric, model selection considered the balance among these complementary measures to identify a clinically useful and reliable prediction model.

These metrics were derived from the confusion matrix, which summarises the relationships between true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as shown in Table 7. All models were evaluated under identical validation and test conditions to ensure fair comparison and reproducibility.

Table 7. Binary confusion matrix

True Positives (TP)	False Negatives (FN)
False Positive (FP)	True Negatives (TN)

Where:

- TP: Correctly predicted positive cases
- FP: Negative cases incorrectly predicted as positive
- FN: Positive cases incorrectly predicted as negative
- TN: Correctly predicted negative cases

Threshold Optimisation

In probabilistic classifiers such as LR, RF, XGBoost, LightGBM, and MLP, each prediction represents the estimated probability that an instance belongs to the positive class. However, clinical deployment requires a decision threshold to translate these probabilities into actionable binary outcomes, “at risk” or “not at risk”, which

in the context of this study correspond to “SRB” or “non-SRB”. The threshold determines the sensitivity-specificity trade-off, primarily shaping the model’s clinical behaviour. By default, most ML classifiers apply a decision threshold of 0.5, meaning that any predicted probability greater than or equal to 0.5 is classified as positive (Kim et al., 2021). While this approach is suitable for balanced datasets, it is less than ideal for rare-event contexts, where class imbalance requires tailored threshold optimisation.

In suicide-related triage, this decision boundary is not purely statistical but clinically strategic. A lower threshold increases recall (sensitivity), capturing more true positives at the expense of generating more false alerts (Wynants et al., 2019). In contrast, a higher threshold improves precision (Positive Predictive Value) but risks missing individuals in crisis (Wynants et al., 2019). Given that the cost of a missed case is far greater than that of an unnecessary alert, threshold selection in this study prioritised maximising the F2-score, which weights recall more heavily than precision. This approach ensures that the model emphasises patient safety by identifying as many SRB-positive cases as possible while maintaining an acceptable level of alert burden in the ED triage setting.

Mathematically, for a classifier that outputs probabilities $\hat{P}_i$, the predicted class $\hat{y}_i$ is defined as shown in Equation (3):

$$\hat{y}_i = \begin{cases} 1, & \text{if } \hat{P}_i \geq \tau \\ 0, & \text{if } \hat{P}_i < \tau \end{cases} \quad (3)$$

Where τ represents the decision threshold. The optimal τ^* was determined by evaluating candidate thresholds between 0 and 1 on the validation set and selecting the operating point that maximised the F2-score, as shown in Equation (4)

$$F\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}}, \text{ where } \beta = 2 \quad (4)$$

This threshold optimisation was performed after model selection to identify a clinically appropriate operating point that balanced sensitivity with alert burden. The F2-score was used because it places greater emphasis on recall while still accounting for precision, making it suitable for determining the deployment threshold in rare-event clinical prediction without relying solely on overall classification accuracy.

Bootstrap Confidence Intervals

To quantify uncertainty in performance estimates under class imbalance, 95% confidence intervals were computed using bootstrap resampling with 1,000 iterations. Bootstrap methods are widely recommended for assessing the stability and uncertainty of clinical prediction models and provide strong interval estimates without relying on distributional assumptions (Efron & Tibshirani, 1994; Mokhtar et al., 2023). For each model, bootstrap samples were generated from the test set with replacement, and performance metrics were recalculated for each resample. Within each bootstrap iteration, the optimal probability threshold was determined by maximising the F2-score, after which prevalence, AUROC, AUPRC, Recall, Precision, F2-score, Balanced Accuracy, MCC, and the optimal threshold were recorded. As threshold optimisation was repeated within each bootstrap iteration, the resulting confidence intervals reflect variability in both model performance and the selected operating threshold. Therefore, these estimates are not direct confidence intervals around the fixed-threshold performance reported in the primary analysis. Instead, they represent the expected range of performance when the operating threshold is re-optimised under repeated sampling, providing a more reliable assessment of model stability for deployment under class-imbalanced conditions. The 2.5th and 97.5th percentiles of the resulting metric distributions were used to derive 95% confidence intervals.

Explainability

To ensure model transparency and clinical interpretability, the study employed the SHAP framework to quantify both global and local feature contributions. SHAP was selected for its ability to provide consistent, model-agnostic attributions. Depending on the architecture of the best model, the appropriate SHAP explainer was applied: *TreeExplainer* for tree-based models (RF, XGBoost, and LightGBM), *LinearExplainer* for LR, and *KernelExplainer* for the MLP neural network. Global interpretability was assessed through two complementary SHAP visualisations. The summary (beeswarm) plot illustrated the direction and magnitude of feature effects across all test samples, showing which features increased or decreased the predicted risk of SRB (Stenwig et al.,

2022). The feature importance (bar) plot summarised the average absolute SHAP values across all features, providing a clear ranking of overall global impact (Stenwig et al., 2022). These visualisations revealed which clinical and contextual variables most strongly influenced model predictions in this dataset.

Intrinsic coefficient analysis was not applicable because the final model (LightGBM) was a tree-based ensemble. Accordingly, interpretability relied on global and local SHAP explanations and LightGBM feature importance. These approaches provide population-level and patient-level insight into the factors influencing model predictions. This enabled verification of consistency between SHAP attributions and traditional statistical feature effects, thereby strengthening the model's interpretability and trustworthiness. Local interpretability was achieved through SHAP force plots, which visualised individual test cases, particularly those with positive outcomes, to demonstrate how specific feature values combined to influence the predicted SRB probability (Stenwig et al., 2022). These instance-level explanations highlight the contribution of key clinical features for a single patient, supporting potential future use of the model as a decision-support tool in emergency triage settings.

To improve interpretability and reduce redundancy, highly overlapping chief complaint-derived variables were excluded from the final feature set. This design choice also helped minimise the risk of label leakage and reduced the potential influence of multicollinearity on feature attribution. Although the final model is relatively strong in terms of correlated predictors, feature importance rankings and SHAP values should be interpreted as measures of predictive association rather than evidence of independent causal effects. Thus, the reported explanations identify variables that contribute to risk prediction, but they do not imply that those variables are direct causes of SRB.

Experimental Pipeline Overview

The experimental workflow followed a structured and reproducible pipeline encompassing data preprocessing, model training, evaluation, and explainability analysis. Figure 2 presents an overview of the complete experimental analysis. All experiments were conducted in Python (version 3.12.13) using Google Colab with hardware acceleration provided by a T4 GPU. The implementation utilised Scikit-learn (version 1.6.1), NumPy (2.0.2), Pandas (2.2.2), Matplotlib (3.10.0) and Seaborn (version 0.13.2) for preprocessing, model development, visualisation and hyperparameter optimisation; XGBoost (version 3.2.0) and LightGBM (version 4.6.0) for model development; TensorFlow (version 2.20.0) for the MLP model; and SHAP for model explainability.

To ensure reproducibility, a fixed random seed (42) was applied throughout the experimental workflow. Hyperparameter optimisation was performed using RandomizedSearchCV with 10 iterations and three-fold cross-validation, with model selection guided by the F2-score to prioritise clinically important minority-class detection.

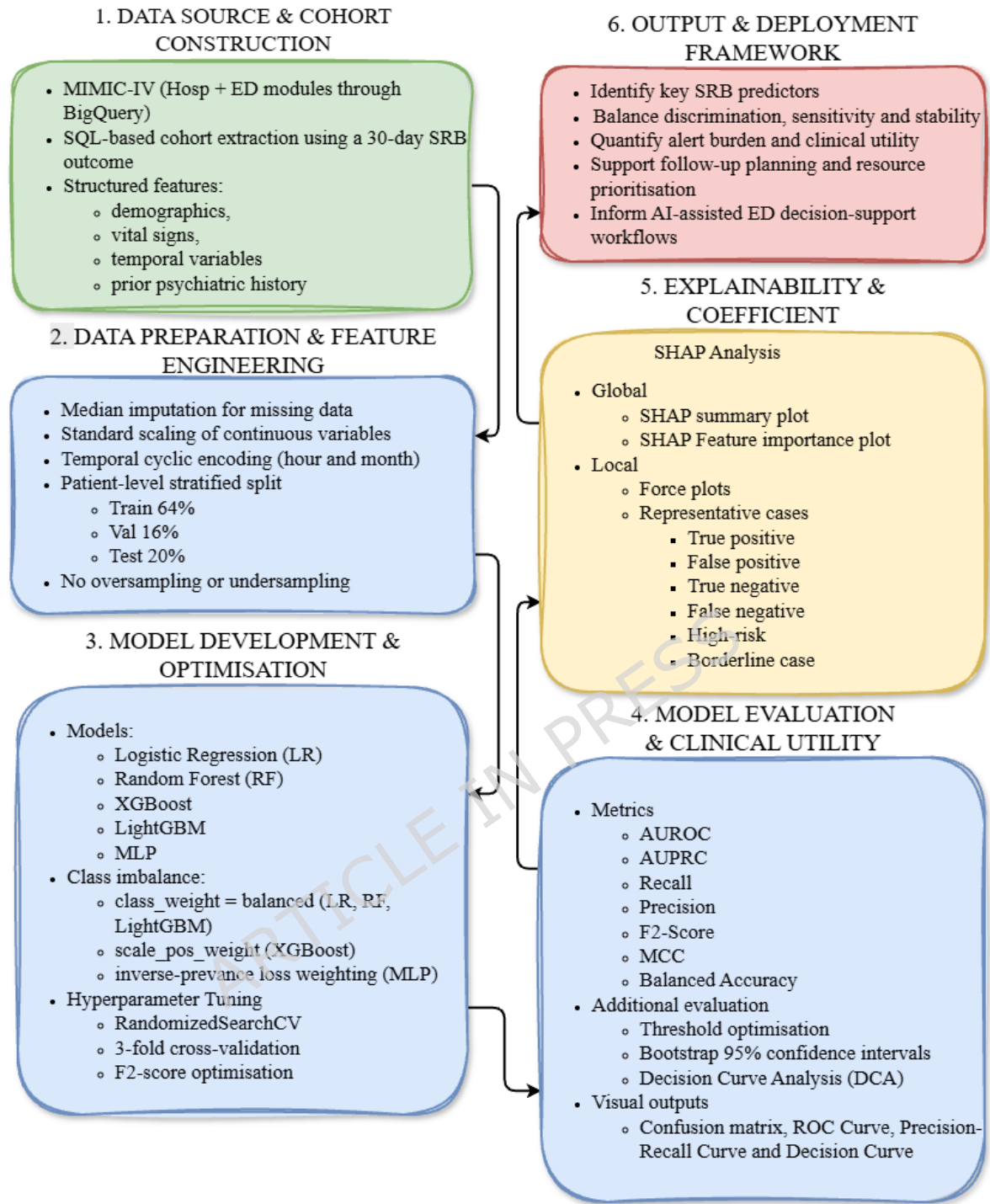


Figure 2. Overview of the experimental pipeline for AI-based triage model development

4. Results and Discussion

This section presents the findings and discusses the proposed AI-based triage model for 30-day SRB prediction.

Comparative Analysis

Table 8 summarises the validation and test performance of the five evaluated models. Across all models, AUROC values ranged from 0.87 to 0.89, while AUPRC values were between 0.23 and 0.25. This pattern indicates that the structured ED triage variables contain meaningful discriminatory signals for 30-day SRB prediction, but also shows that prediction remains difficult because the positive class is relatively uncommon and clinically heterogeneous. The gap between AUROC and AUPRC is important: although AUROC suggests stable overall

separation between SRB and non-SRB encounters, AUPRC provides a more realistic indication of minority-class detection under class imbalance (Shukla & Singh, 2025; Sitthiprawiat et al., 2025). Therefore, model comparison was not based solely on AUROC.

No single algorithm dominated across all evaluation dimensions. On the validation set, the MLP achieved the highest AUROC (0.89), while LightGBM achieved the highest AUPRC (0.25). RF produced the highest F2-score (0.40), whereas the MLP and LR achieved the highest recall (0.83). Similar trends were observed on the independent test set. LightGBM achieved the highest AUROC (0.88) and AUPRC (0.25), while RF and XGBoost achieved the highest F2-scores (0.42). The MLP and LR achieved slightly higher recall (0.82 and 0.81, respectively), but at the expense of lower precision and weaker overall minority-class discrimination. These findings suggest that no single metric or algorithm adequately captures clinical utility. Models with higher recall identified more future SRB cases, whereas models with higher precision or F2-score reduced alert burden at the expense of missed cases. As a result, model selection required balancing discrimination, sensitivity, precision, and operational burden. This trade-off is crucial in suicide-risk prediction, as false negatives may represent missed opportunities for intervention, whereas false positives may increase workload through additional screening or psychiatric review (Thangada & Kasoju, 2024; Walsh et al., 2025). For this reason, recall and F2-score were interpreted alongside AUPRC, MCC, and Balanced Accuracy rather than treated as isolated endpoints. This approach aligns with clinical decision-support evaluation, where model utility depends on discrimination and on the sensitivity-alert burden balance (Rana et al., 2024).

Table 8. Comparative performance of the models on validation and test datasets.

Model	Subset	AUROC	AUPRC	F2-Score	Precision	Recall	MCC	Bal Acc
MLP	Validation	0.89	0.25	0.37	0.12	0.83	0.26	0.81
LightGBM	Validation	0.88	0.25	0.39	0.13	0.80	0.27	0.81
LR	Validation	0.88	0.24	0.37	0.11	0.83	0.26	0.81
RF	Validation	0.88	0.23	0.40	0.16	0.62	0.28	0.76
XGBoost	Validation	0.87	0.24	0.39	0.14	0.70	0.27	0.78
LightGBM	Test	0.88	0.25	0.41	0.14	0.79	0.28	0.80
MLP	Test	0.87	0.25	0.39	0.13	0.82	0.27	0.80
XGBoost	Test	0.87	0.24	0.42	0.16	0.71	0.29	0.78
LR	Test	0.87	0.23	0.38	0.12	0.81	0.26	0.80
RF	Test	0.87	0.23	0.42	0.18	0.64	0.29	0.77

Bal Acc = Balanced Accuracy

The MLP results are both clinically and methodologically informative. The incorporation of inverse-prevalence class weighting enabled the neural network to achieve competitive discrimination and recall, demonstrating that cost-sensitive learning can effectively address class imbalance in structured clinical data. However, the tree-based ensemble methods consistently achieved comparable or superior minority-class performance, supporting previous observations that gradient boosting approaches remain highly effective for tabular healthcare prediction tasks (Ren et al., 2025).

LightGBM was selected as the final predictive model as it provided the most favourable overall balance across the evaluation framework. Although RF and XGBoost achieved marginally higher F2-scores on the test set, LightGBM achieved the highest test AUROC and AUPRC while maintaining strong recall (0.79), MCC (0.28), and Balanced Accuracy (0.80). This suggests that LightGBM offered the best compromise between global discrimination, minority-class detection, and decision stability. Its performance also remained consistent across the validation and test sets, reducing concern that model selection was driven by validation-set artefacts. Therefore, the final model choice reflects a multidimensional clinical evaluation strategy rather than optimisation of a single metric. To further quantify uncertainty and assess model stability under class imbalance, bootstrap-derived 95% confidence intervals for the final model.

As shown in Table 9, the bootstrap confidence intervals were relatively narrow across the validation and test sets, indicating stable discrimination and minority-class detection. In addition, the overlap between the validation and test intervals suggests limited evidence of overfitting. The confidence intervals for the optimal threshold also show relatively low variability, supporting the stability of the selected operating point under class imbalance.

Table 9. Bootstrap 95% confidence intervals of the final model (LightGBM)

Metric	Validation	Test
AUROC	0.88 (0.88 - 0.89)	0.88 (0.87 - 0.89)
AUPRC	0.25 (0.22 - 0.28)	0.26 (0.23 - 0.28)
F2-Score	0.42 (0.40 - 0.44)	0.44 (0.42 - 0.46)
Precision	0.17 (0.14 - 0.22)	0.19 (0.17 - 0.21)
Recall	1.00 (1.00 - 1.00)	1.00 (1.00 - 1.00)
MCC	0.30 (0.28 - 0.32)	0.31 (0.30 - 0.33)
Balanced Accuracy	0.77 (0.74 - 0.81)	0.77 (0.76 - 0.79)
Prevalence	0.03 (0.03 - 0.03)	0.04 (0.03 - 0.04)
Optimal Threshold	0.68 (0.57 - 0.77)	0.70 (0.65 - 0.72)

The comparative analysis results show that AI-assisted 30-day SRB prediction using structured ED triage data is feasible but should not be interpreted as a standalone screening mechanism. Although the models provide clinically useful risk-stratification signals, the modest precision values indicate that predictions are best viewed as prioritisation cues that can trigger additional assessment within existing workflows. Consistent with this objective, model selection was based on a multidimensional evaluation framework incorporating discrimination, minority-class detection, decision stability, and clinical interpretability. Within this framework, LightGBM achieved the most favourable overall balance and was therefore selected for subsequent threshold optimisation, decision-curve analysis, and explainability analyses.

Performance of the Selected Model: LightGBM

It is important to distinguish model selection from deployment configuration. Model selection assesses an algorithm's ability to predict 30-day SRB risk, whereas deployment requires defining an operating threshold that translates predicted probabilities into clinically actionable alerts. In high-acuity ED settings, thresholds that maximise sensitivity may generate excessive false-positive alerts, potentially contributing to alert fatigue and reduced clinician trust (Walsh et al., 2025; Wynants et al., 2019). Therefore, threshold selection represents an operational decision that balances sensitivity with alert burden rather than a criterion for model selection.

Following model selection, threshold optimisation was performed to identify a clinically appropriate operating point. The selected threshold of 0.6979 yielded an F2-score of 0.44 on the test set, with a recall of 0.64 and a precision of 0.19 (refer to Table 10). This operating point represents a compromise between identifying high-risk patients and maintaining a manageable alert burden. Lower thresholds can increase sensitivity but may substantially increase false-positive alerts, whereas higher thresholds improve precision at the expense of missed cases (Scarffe et al., 2024). Therefore, the selected threshold should be interpreted as a clinically informed operating point rather than a universally optimal threshold. In suicide risk prediction, this balance is important as excessive alert generation may contribute to clinician workload and alert fatigue, while overly conservative thresholds increase the likelihood of failing to identify patients requiring further assessment (Kitchen et al., 2026).

Table 10. Threshold-optimised performance of the selected LightGBM model.

Metrics	Value
Selected Threshold	0.6979
F2-Score	0.44
Recall (Sensitivity)	0.64
Precision	0.19

Confusion Matrix Analysis

Figure 3 presents the confusion matrix of the threshold-optimised LightGBM model evaluated on the independent test set. At the selected operating threshold, the model correctly identified 957 future SRB encounters while 537 SRB cases were missed, corresponding to a recall of 0.64. In addition, the model generated 4,013 false-positive alerts while correctly classifying 35,695 non-SRB encounters. These results highlight the fundamental challenge of suicide-risk prediction: despite strong discrimination, the rarity and heterogeneity of SRB mean that improving case detection inevitably increases the number of patients requiring further assessment. Operationally, the model

generated 4,970 alerts across 41,202 ED encounters, corresponding to approximately 121 alerts per 1,000 presentations. Although only a minority of alerts represented confirmed SRB events, the relatively low precision should be interpreted within the context of emergency medicine and psychiatric practice. Suicide risk assessment is inherently probabilistic, and many patients identified as high risk may still require additional screening or clinical review despite not experiencing documented SRB during the 30-day follow-up period. Furthermore, some apparent false positives may reflect undocumented suicidal crises or limitations of ICD-based outcome definitions rather than true misclassification.

The confusion matrix also highlights the cost of increasing specificity. At the selected operating point, 36% of future SRB encounters were missed. While higher thresholds would reduce alert burden, they would also increase the number of missed high-risk patients. On the other hand, lower thresholds would improve sensitivity but generate substantially more alerts, potentially contributing to alert fatigue and reduced clinician trust (Scarffe et al., 2024; Walsh et al., 2025; Wynants et al., 2019). Similar trade-offs have been reported in early-warning systems for sepsis, clinical deterioration, and psychiatric crises, where model utility depends not only on predictive performance but also on operational sustainability (van der Vegt et al., 2026).

From a psychiatric and emergency medicine perspective, these findings support the role of AI-assisted risk prediction as a prioritisation mechanism rather than a replacement for established assessment procedures. In routine ED practice, patients flagged as high risk would undergo secondary evaluation using clinician judgment and validated screening instruments, such as the ASQ or C-SSRS, before management decisions are made. Therefore, the confusion matrix should not be interpreted solely in terms of classification errors but rather as illustrating the balance between case detection and assessment workload. Within this context, the selected operating point provides a clinically reasonable compromise, enabling early identification of a substantial proportion of future SRB cases while maintaining an alert burden that may be manageable within existing workflows.

ARTICLE IN PRESS

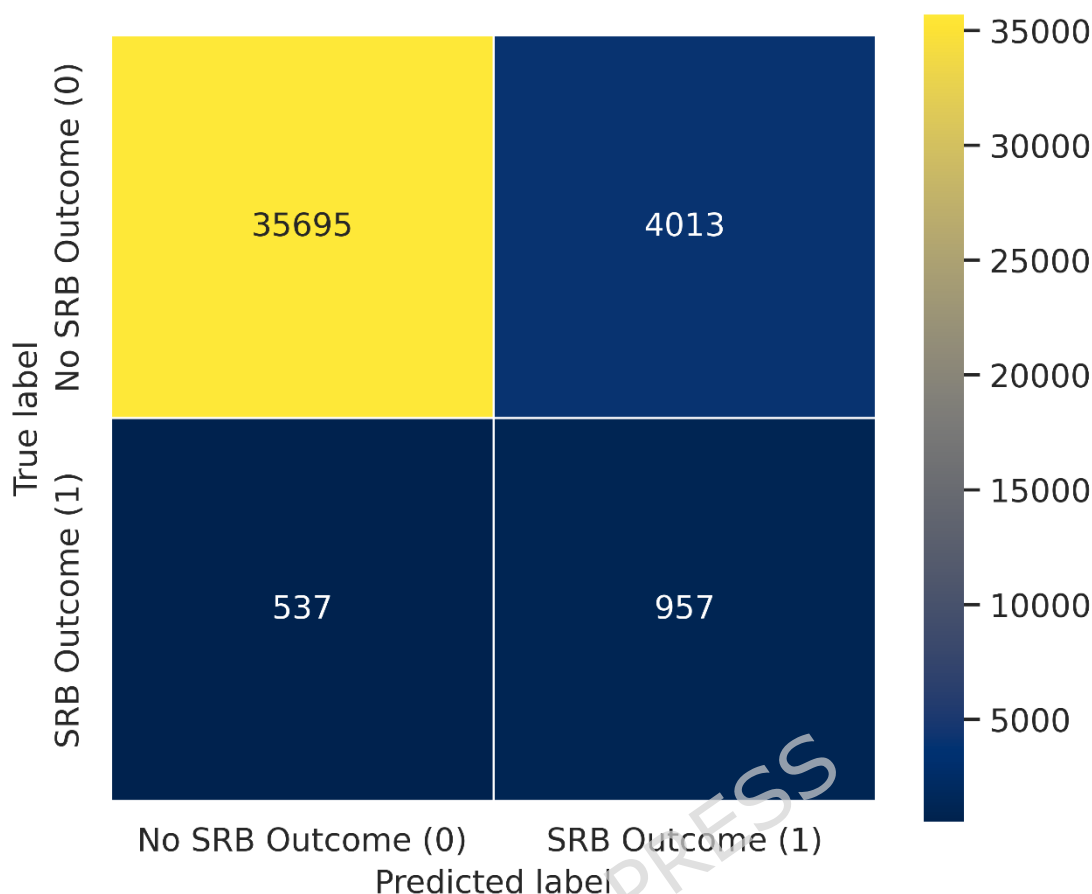


Figure 3. Confusion matrix for the LightGBM model at the optimal triage threshold.

Receiver Operating Characteristic (ROC) and Precision-Recall (PR) Curves

Figures 4 and 5 present the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for the five evaluated models on the independent test dataset. These curves extend the point-estimate comparison by showing model behaviour across the full range of possible decision thresholds. This is important because the clinical usefulness of a 30-day SRB risk model depends on its ability to rank patients by risk and on how precision and sensitivity change as the operating threshold is varied.

The ROC curves in Figure 4 show the overlap across models, with AUROC values ranging from 0.8675 to 0.8788. LightGBM achieved the highest AUROC (0.8788), followed by MLP (0.8741), XGBoost (0.8697), LR (0.8676), and RF (0.8675). However, the small separation between curves indicates that AUROC alone provides limited grounds for model selection. In imbalanced clinical prediction tasks, AUROC primarily reflects ranking performance across thresholds and can obscure the practical difficulty of identifying rare positive outcomes, as the large number of true negatives strongly contributes to apparent discrimination (Shukla & Singh, 2025; Sitthiprawiat et al., 2025). Therefore, the ROC results were interpreted as evidence of general separability, not as sufficient evidence of clinical utility.

The PR curves in Figure 5 provide a complementary view by focusing directly on the relationship between precision and recall. The baseline AUPRC was 0.0363, corresponding to the SRB prevalence in the test set. All models performed better than this baseline, indicating that structured ED triage variables improved the identification of future SRB cases beyond random classification. LightGBM achieved the highest AUPRC (0.2549), followed by MLP (0.2467), XGBoost (0.2389), LR (0.2314), and RF (0.2255). These values represent an improvement over the prevalence baseline, but they also show that the absolute precision remains modest. This is clinically important because even a model with useful risk-ranking ability will still generate a considerable number of false-positive alerts when deployed in a low-prevalence setting.

Figures 4 and 5 support the multidimensional model-selection strategy adopted in this study. The ROC curves show that all models achieved broadly similar global discrimination, while the PR curves reveal clearer differences

in minority-class detection. LightGBM did not dominate every metric, but it showed the most favourable threshold-wide balance by combining the highest AUROC and AUPRC with competitive recall, MCC, and Balanced Accuracy. Therefore, the curve-based analysis supports the broader evaluation framework, which focuses on discrimination, minority-class detection, sensitivity, decision stability, and operational feasibility.

Clinically, these findings reinforce that the proposed model should be interpreted as a risk-prioritisation tool rather than a standalone screening instrument. The PR curves show that predicted high-risk groups are enriched for future SRB cases, but not sufficiently precise to justify automatic intervention without clinical review. In emergency medicine and psychiatric assessment, this is an appropriate role for decision support: the model can help prioritise patients for additional assessment, while final clinical decisions remain dependent on clinician judgement and established suicide-risk assessment procedures.

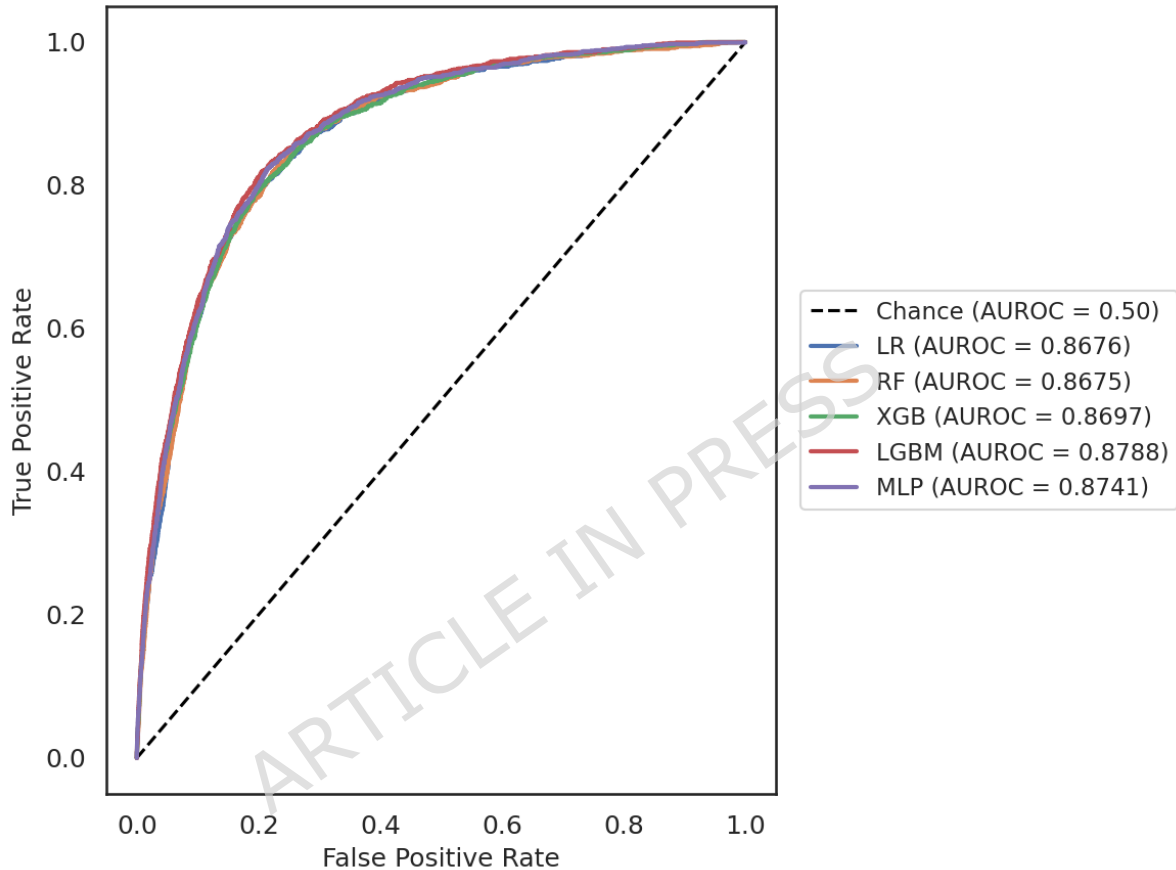


Figure 4. Receiver operating characteristic (ROC) curves for LR, RF, XGBoost, LightGBM, and MLP on the test set.

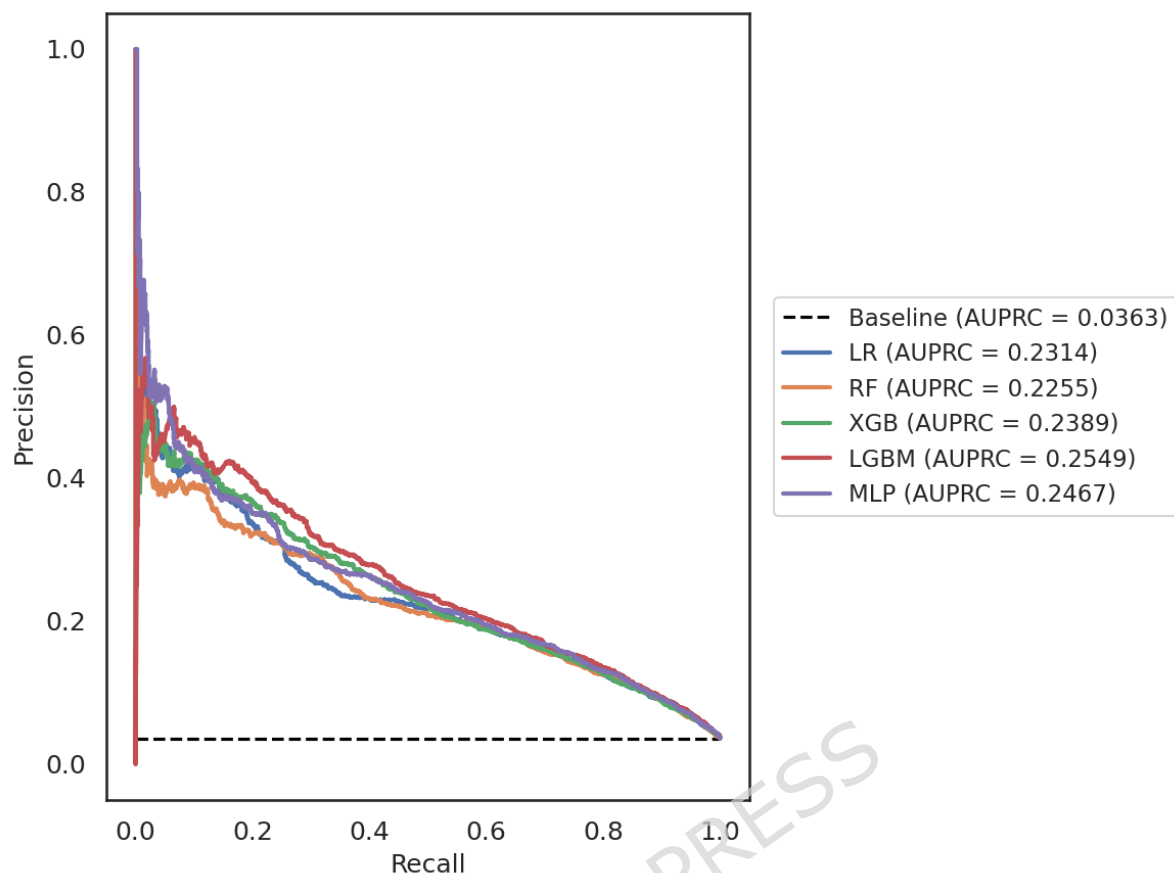


Figure 5. Precision-recall (PR) curves for LR, RF, XGBoost, LightGBM, and MLP on the test set.

Decision Curve Analysis (DCA)

Decision Curve Analysis (DCA) was performed to evaluate whether the improved discrimination achieved by the selected LightGBM model translates into meaningful clinical utility. Unlike conventional performance metrics, DCA explicitly accounts for the trade-off between correctly identifying future SRB cases and the burden of unnecessary escalation, thereby providing a clinically oriented assessment of model usefulness (Vickers & Holland, 2021). Figure 6 illustrates the net benefit of the LightGBM model relative to the Treat All and Treat None strategies across threshold probabilities ranging from 0.2% to 5%. This range reflects clinically plausible risk tolerances for low-prevalence outcomes, in which even small predicted probabilities may justify additional assessment or follow-up (Wynants et al., 2019). Across the entire range examined, the LightGBM model consistently provided greater net benefit than both reference strategies. This indicates that model-guided risk stratification yields more favourable clinical decisions than either indiscriminately escalating all patients or ignoring predicted risk altogether.

The model's advantage became more apparent as the threshold increased. While the Treat All strategy rapidly lost utility due to the growing penalty associated with excessive false-positive alerts, the LightGBM model maintained a positive net benefit through more selective risk stratification. This behaviour complements the confusion matrix analysis, which demonstrated that the operational challenge lies not only in detecting future SRB cases, but also in maintaining an acceptable alert burden. Therefore, DCA provides additional evidence that the selected model achieves a clinically reasonable balance between sensitivity and resource utilisation. Although the absolute net benefit values are necessarily small because SRB is a rare outcome, their consistent superiority over the Treat All and Treat None strategies suggests that the model may support more efficient allocation of psychiatric assessment resources. These findings further support the model's role as a prioritisation mechanism that enriches the population referred for secondary evaluation. From an emergency medicine perspective, such selective escalation is particularly valuable in resource-constrained environments, where universal psychiatric assessment is impractical and excessive alert volumes may contribute to clinician fatigue (Kitchen et al., 2026). The DCA findings support the broader objective of this study: providing clinically actionable risk stratification that

complements existing workflows while improving the efficiency of downstream assessment and follow-up planning.

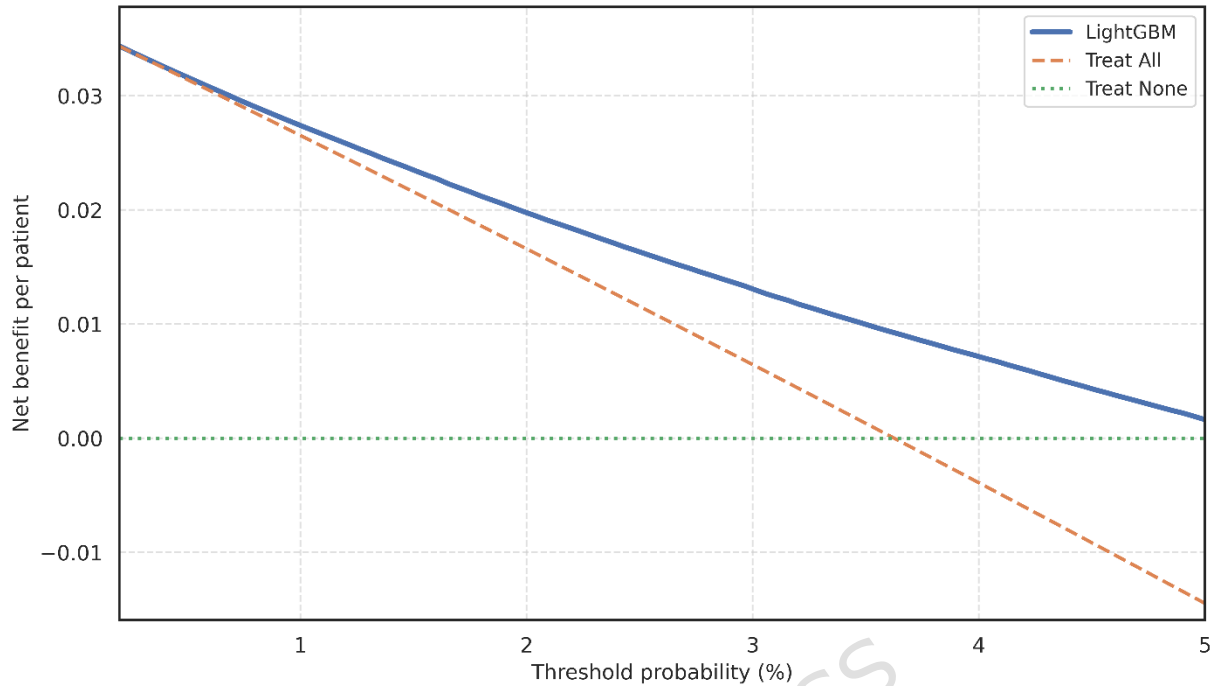


Figure 6. Comparison of net benefit between the AI-based triage model, “Treat All,” and “Treat None” strategies across decision thresholds (0.2 to 5%)

Explainability and Feature importance

This section presents the interpretability of the LightGBM model using both model-agnostic explanations (SHAP) and model-intrinsic feature importance.

Global Feature Impact

Figure 7 presents the SHAP summary plot for the selected LightGBM model, providing a global view of how individual features influence 30-day SRB risk. The distribution of SHAP values shows that model predictions are driven by a combination of demographic factors, psychiatric history, physiological variables, temporal patterns, and presenting complaints rather than by any single variable. This supports the multidimensional nature of suicidal behaviour, which arises from interactions between psychiatric vulnerability, acute physiological stress, and contextual factors.

Age emerged as the most influential predictor. Lower age values were generally associated with positive SHAP values, indicating increased predicted SRB risk. Although this pattern may partly reflect the age distribution of the cohort, it is also consistent with epidemiological evidence showing that adolescents and young adults experience higher rates of suicidal ideation, self-harm, and suicide attempts than older adults (CDC, 2025; Thangada & Kasoju, 2024). Therefore, the observed age effect is likely attributable to both underlying population characteristics and genuine clinical risk patterns rather than dataset bias alone. Prior mood disorder history (*prior_flag_mood*) showed strong positive contributions to predicted risk, consistent with psychiatric literature identifying mood disorders as one of the strongest predictors of suicidal behaviour. Similarly, a history of substance use disorder (*prior_flag_sud*) and anxiety (*prior_flag_anxiety*) contributed positively to risk predictions, reflecting well-established associations between psychiatric comorbidity, impaired emotional regulation, and increased suicide vulnerability. These findings suggest that the model captures clinically meaningful longitudinal risk factors rather than relying solely on the current presentation.

Several physiological variables, including temperature, blood pressure, respiratory rate, oxygen saturation, and shock index, also influenced predictions. Their effects were heterogeneous, indicating that no single vital sign acts as a direct marker of SRB. Instead, these variables may capture acute medical instability, intoxication, overdose, withdrawal syndromes, or heightened autonomic activation, all of which are frequently encountered in

emergency psychiatric presentations. This observation aligns with emergency medicine literature, where abnormal physiological parameters often accompany acute behavioural crises rather than serving as specific indicators of suicide risk (Nabar et al., 2026). Among presenting complaints, pain-related presentations (*cc_pain*) were notable. Although pain itself is not a direct indicator of suicidal behaviour, chronic pain and severe acute pain are recognised risk factors for depression, hopelessness, and suicidal ideation. Therefore, pain-related complaints may represent an indirect pathway through which psychological distress manifests in emergency settings. Temporal variables showed relatively smaller effects, suggesting that admission timing contributes modestly to risk stratification but is less informative than demographic, psychiatric, and physiological factors.

The SHAP summary plot demonstrates that the model does not rely on isolated features or simple heuristics. Instead, predictions arise from the interaction of demographic vulnerability, psychiatric history, physiological status, and presenting context. The observed feature patterns are broadly consistent with current psychiatric and emergency medicine knowledge, supporting the clinical plausibility and transparency of the proposed model. Such explainability is particularly important for deployment in ED triage, where clinician trust and accountability are essential for the safe integration of AI-assisted decision support (Rana et al., 2024).

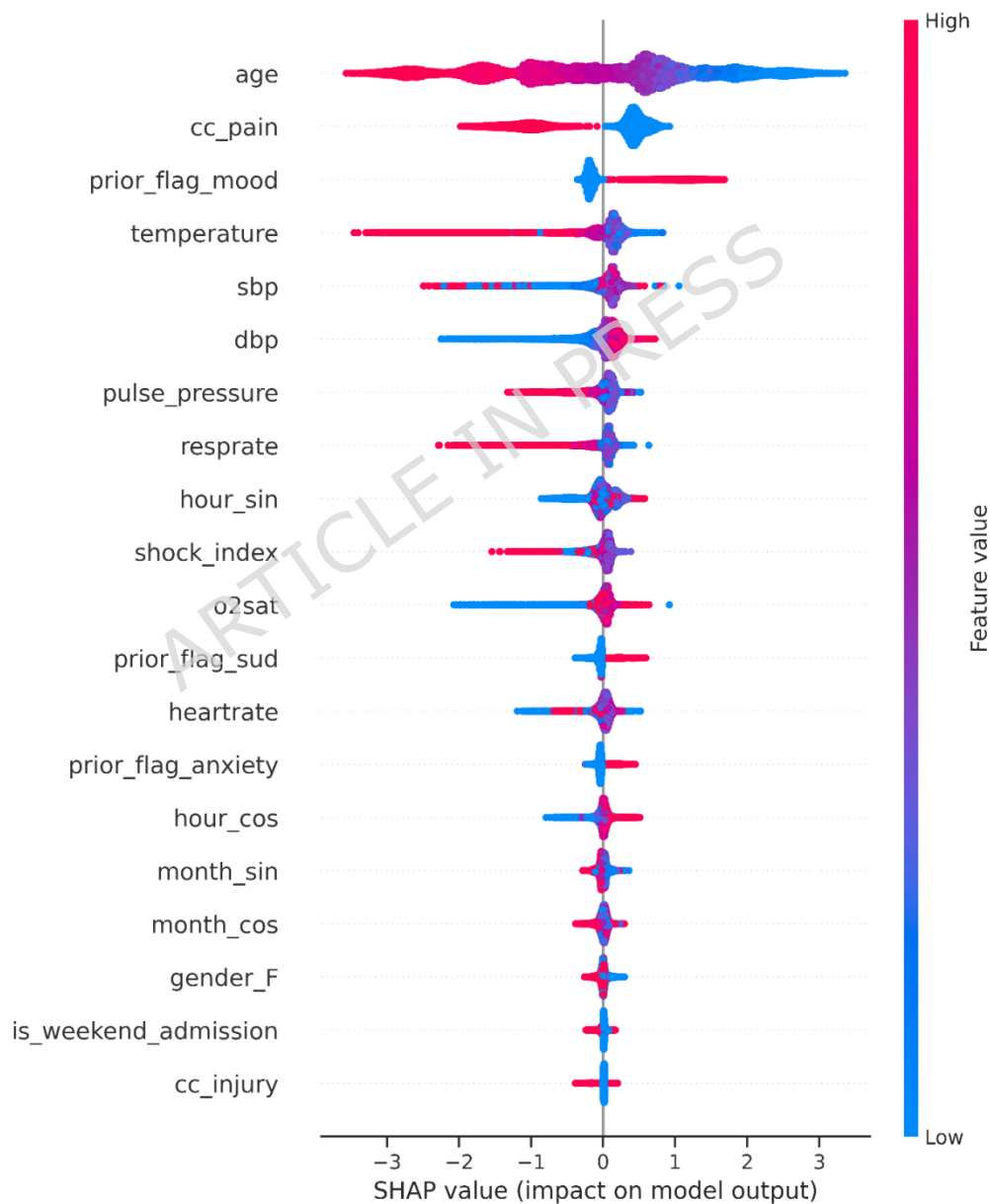


Figure 7. SHAP Summary plot (Global Impact) for LightGBM

Global Feature Importance

Figure 8 presents the mean absolute SHAP values of the selected LightGBM model and provides a complementary ranking of feature importance. Consistent with the SHAP summary plot (Figure 7), age showed the largest average contribution to model predictions, followed by pain-related presentations and prior mood disorder history. Physiological variables, including temperature, blood pressure, pulse pressure, respiratory rate, shock index, and oxygen saturation, also contributed substantially, highlighting the importance of the acute clinical state alongside psychiatric vulnerability. The close agreement between Figures 7 and 8 demonstrates the internal consistency of the explainability analysis and suggests that model predictions arise from a clinically plausible combination of demographic, psychiatric, and physiological factors. However, mean absolute SHAP values quantify the magnitude of feature influence but do not indicate causality or the direction of association; these should instead be interpreted using the SHAP summary plot.

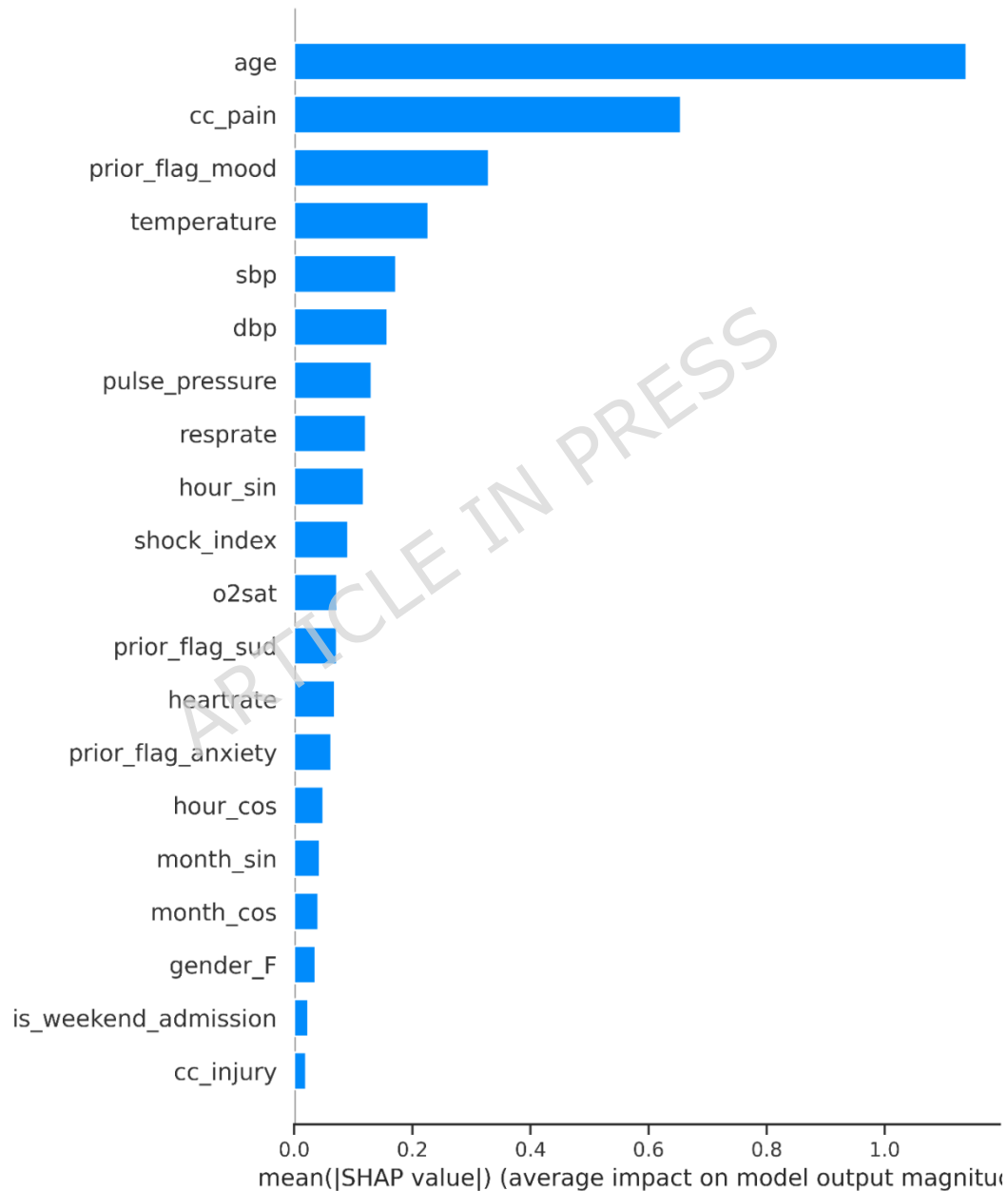


Figure 8. SHAP Feature Importance (Global Magnitude) for LightGBM

Local Explanation

To complement the global explanations, SHAP force plots were used to examine representative patient-level predictions. One encounter was randomly selected from each prediction category (true positive, false positive, true negative, and false negative) to illustrate how the model combined demographic, psychiatric, and physiological variables to generate individual risk estimates. In addition, two clinically informative examples were included: a high-risk case with a very high predicted probability and a borderline case with a predicted probability approximately equal to the selected operating threshold. These cases provide insight into correct classifications, model errors, and prediction behaviour around the clinical decision boundary. The selected encounters are summarised in Table 11.

Table 11. Representative local explanations used for SHAP analysis

Case	Index	True Label	Predicted Label	Predicted Probability	Optimal Threshold
True Positive	24820	1	1	96.6%	69.8%
False Positive	9910	0	1	96.8%	69.8%
True Negative	12178	0	0	0.02%	69.8%
False Negative	9469	1	0	69.7%	69.8%
High-Risk	9910	0	1	96.8%	69.8%
Borderline	27531	1	1	69.8%	69.8%

True-Positive Case

For the true-positive patient (Figure 9), younger age exerted the strongest positive contribution to risk, followed by prior mood disorder, anxiety history, and substance use disorder. Additional contributions were made by blood pressure, respiratory rate, and temperature. These effects acted in the same direction and, collectively, shifted the prediction from the baseline value towards a very high risk probability (96.6%). The explanation suggests that the model identified a coherent accumulation of risk factors rather than relying on a single variable, indicating that risk prediction emerged from the combined effect of psychiatric history and physiological state.



Figure 9. SHAP force plot for the true-positive case (Patient #24820)

False-Positive and High-Risk Case

The high-risk case corresponds to the false-positive encounter with the highest predicted probability and was included separately to illustrate the model's reasoning under an extreme-confidence prediction. The false-positive and high-risk case (Figure 10) produced the highest predicted probability (96.8%), despite the absence of a documented SRB outcome within 30 days. The explanation closely resembled the true-positive case, with younger age, prior mood disorder, anxiety, substance use disorder, and elevated shock index all contributing positively to risk. Only systolic blood pressure and weekend admission provided modest protective effects. The similarity between this profile and the true-positive case suggests that the model assigned high risk because the patient showed characteristics commonly associated with SRB. The prediction appears to represent a clinically plausible high-risk profile rather than an arbitrary error, highlighting the difficulty of distinguishing future events from latent risk.

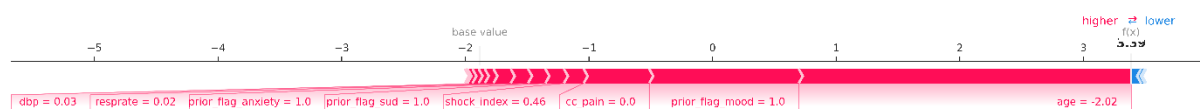


Figure 10. SHAP force plot for the false-positive high-risk case (Patient #9910).

True-Negative Case

For the true-negative patient (Figure 11), the prediction was dominated by several protective factors. Older age had the strongest negative effect, while systolic and diastolic blood pressure, respiratory rate, and chief complaint

pain further reduced the predicted risk. Although prior substance use history contributed positively, its effect was insufficient to offset the stronger protective influences. The resulting prediction probability was extremely low (0.02%), demonstrating that the model could recognise patients with an overall low risk profile despite the presence of isolated risk factors.

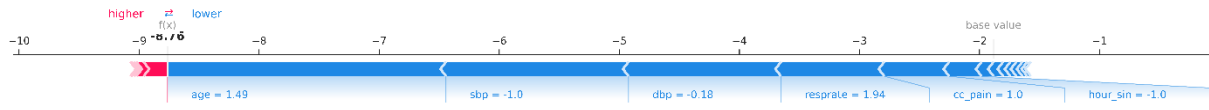


Figure 11. SHAP force plot for the true-negative case (Patient #12178)

False-Negative Case

The false-negative case (Figure 12) highlights the complexity of individual predictions. Younger age, prior mood disorder, anxiety, substance use disorder, and several physiological variables all increased risk. However, temperature exerted a strong negative contribution, partially counteracting these effects and reducing the final probability below the selected threshold. The predicted probability (69.67%) differed from the operating threshold (69.79%) by only a small margin. Therefore, this misclassification appears to reflect uncertainty near the decision boundary rather than a complete failure of the model. The force plot suggests that competing feature effects produced a finely balanced prediction in which small changes in the operating threshold could alter the classification outcome.

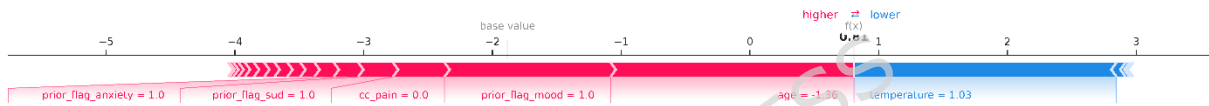


Figure 12. SHAP force plot for the false-negative case (Patient #9469)

Borderline Case

The borderline patient (Figure 13) illustrates how the model behaves in uncertain regions of the feature space. Younger age and the absence of pain contributed strongly to increasing risk, whereas the absence of prior mood disorder and substance use disorder exerted protective effects. The remaining physiological variables made relatively small positive contributions, resulting in a probability that was almost exactly equal to the selected threshold (69.79%). Unlike the true-positive case, risk was not driven by multiple strong psychiatric factors but by the accumulation of several weaker contributors. This example demonstrates that the model does not produce binary decisions from a single dominant variable; instead, predictions arise from the interaction of multiple factors, whose combined effects determine whether an encounter falls above or below the operating threshold.

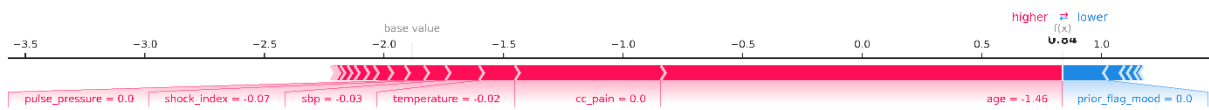


Figure 13. SHAP force plot for the borderline case (Patient #27531)

The local explanations reveal that model predictions were driven by combinations of factors rather than isolated variables. Correct predictions and misclassifications often shared similar risk profiles, with differences arising from the relative magnitude and interaction of feature contributions. The false-negative and borderline cases further demonstrate that uncertainty is concentrated near the decision threshold, suggesting that some apparent errors reflect the probabilistic nature of risk prediction rather than inconsistent model behaviour. These findings support the interpretability of the LightGBM model and show that individual predictions are traceable to clinically understandable combinations of variables. These cases are illustrative rather than representative and intended to demonstrate the model's learned decision pathways.

Intrinsic Feature Importance

To complement the SHAP analyses, the intrinsic feature importance scores of the final LightGBM model were examined (Table 12). Unlike SHAP values, which quantify both the direction and magnitude of feature contributions, LightGBM feature importance reflects how frequently and effectively variables are used during tree construction. Therefore, these scores provide information about the relevance of the predictor but not whether a variable increases or decreases risk.

The importance ranking showed strong agreement with the global SHAP analyses, supporting the model's internal consistency. Age emerged as the most influential predictor, followed by temperature, systolic blood pressure, shock index, diastolic blood pressure, pulse pressure, and heart rate. The prominence of these variables suggests that the model relied heavily on routinely collected physiological measurements, indicating that short-term SRB risk prediction may partly capture manifestations of acute physiological stress rather than depending exclusively on psychiatric history. Temporal variables, including hour and month encodings, contributed moderately to prediction, whereas psychiatric history variables, such as prior mood disorder, substance use disorder, and anxiety history, exhibited lower importance scores. This finding should not be interpreted as evidence that psychiatric factors are unimportant. Instead, it suggests that psychiatric history serves as contextual information that modifies risk alongside physiological and demographic characteristics, rather than as the primary driver of prediction. Similarly, chief complaint variables, including pain and injury, contributed less than age and vital signs, indicating that the model integrated multiple weak signals rather than relying on a few highly dominant features.

The close correspondence between LightGBM feature importance and SHAP results strengthens confidence that the identified predictors are not artefacts of a particular explanation method. While Gini importance provides a global ranking of feature relevance, SHAP analysis further demonstrated how these variables interact and contribute to individual predictions. These complementary explanations support the transparency of the proposed model and provide a more comprehensive understanding of how routinely available ED triage variables contribute to 30-day SRB risk prediction.

Table 12. LightGBM intrinsic feature importance scores.

Feature	Importance
age	1157
temperature	889
sbp	823
shock_index	815
dbp	769
pulse_pressure	715
heartrate	673
hour_sin	425
o2sat	424
resprate	412
hour_cos	352
month_sin	295
month_cos	289
prior_flag_mood	196
prior_flag_sud	160
cc_pain	142
prior_flag_anxiety	122
gender_F	114
is_weekend_admission	88
cc_injury	54
gender_M	50
is_night_admission	36

Clinical Interpretation

Consistent with the objective of developing an interpretable and clinically actionable model for early suicide risk identification, the explainability analyses demonstrate that routinely collected ED triage variables contain meaningful signals associated with 30-day SRB. The model did not rely on a single dominant predictor, but rather

integrated demographic characteristics, physiological measurements, presenting complaints, temporal information, and psychiatric history to predict risk. This multidimensional pattern reflects the complex and heterogeneous nature of suicidal behaviour, which rarely manifests through a single clinical indicator.

The explainability results suggest that risk prediction in the ED extends beyond traditional physiology-based triage. Although vital signs remain central to emergency assessment, the SHAP analyses showed that contextual and psychiatric variables modify risk even when physiological instability is absent. This finding is consistent with psychiatric and emergency medicine literature, indicating that suicidal crises frequently present through indirect behavioural and psychosocial manifestations rather than obvious physiological deterioration (Hughes et al., 2023; Shiraly et al., 2024). Accordingly, reliance on conventional triage alone may fail to identify some individuals at elevated risk.

The proposed model is therefore intended as an augmentative clinical information system rather than an autonomous screening tool. In high-acuity environments characterised by time pressure, incomplete information, and competing priorities, the model provides transparent risk predictions that may help clinicians recognise patterns warranting further evaluation. The output should be interpreted as a prioritisation signal that supports secondary assessment rather than as a diagnosis or replacement for clinical judgement. This approach aligns with principles of trustworthy and responsible AI in healthcare, emphasising transparency, human oversight, and clinician accountability (Rana et al., 2024; Wynants et al., 2019).

A practical advantage of the model is that it relies exclusively on structured variables routinely available at triage, requiring no additional questionnaires or specialised data collection. This enhances feasibility for real-time deployment and may be particularly valuable in resource-constrained settings where psychiatric expertise is limited. Furthermore, the consistency between global SHAP patterns, intrinsic LightGBM feature importance, and local patient-level explanations suggests that the model generates clinically coherent predictions rather than relying on false or forged associations. Such transparency is essential for fostering clinician trust and facilitating integration into existing emergency care workflows. Overall, the findings support the role of explainable AI as an additional safety layer within ED triage. By transforming routinely collected data into interpretable risk estimates, the model has the potential to improve risk visibility, support earlier psychiatric assessment, and reduce missed opportunities for suicide prevention while maintaining clinician-centred decision-making.

Answering Research Questions (RQs)

RQ 1: Which AI model provides the best balance between predictive performance, minority-class detection, clinical utility, and interpretability?

The findings demonstrate that no single performance metric was sufficient for selecting an AI model for 30-day SRB prediction. Although all five models showed comparable discrimination, differences emerged in minority-class detection, decision stability, threshold behaviour, and explainability. These findings reinforce the need for model selection in rare-event clinical prediction to be based on a multidimensional evaluation framework rather than discrimination alone. Among the evaluated algorithms, LightGBM provided the most favourable overall balance. While other models achieved marginal advantages for individual metrics, LightGBM consistently combined strong discrimination, effective minority-class detection, stable threshold behaviour, favourable clinical utility, and interpretable explanations. Bootstrap confidence intervals further demonstrated stable performance across validation and test cohorts, while SHAP analyses confirmed that the model's predictions were driven by clinically plausible risk factors. Decision curve analysis also showed that the selected model provided greater clinical utility than treat-all or treat-none strategies across clinically relevant threshold probabilities.

Therefore, within the context of routinely collected structured ED triage data, LightGBM represented the most appropriate model for estimating 30-day SRB risk. Rather than identifying a universally superior algorithm, this study demonstrates that clinically deployable suicide risk prediction requires balancing predictive performance with interpretability, operational feasibility, and clinical decision support, consistent with current recommendations for trustworthy AI in healthcare (Riley et al., 2020; Walsh et al., 2025; Wynants et al., 2019).

RQ 2: How interpretable and clinically actionable are these models in explaining SRB risk?

The findings demonstrate that the selected LightGBM model provides transparent and clinically meaningful explanations of 30-day SRB risk. The combination of global and local explainability showed that predictions were driven by clinically plausible interactions among demographic characteristics, psychiatric history, physiological

measurements, and presentation context rather than by isolated predictors. The agreement between SHAP analyses, intrinsic feature importance, and patient-level explanations further supports the internal consistency and reliability of the explanation framework. From a clinical perspective, the model generated patient-specific risk estimates traceable to identifiable contributing factors, enabling clinicians to understand why a patient was assigned a higher or lower risk. Local explanations also showed that prediction uncertainty was concentrated around the clinical decision threshold, whereas high-confidence predictions were supported by coherent combinations of established risk factors. This level of transparency supports clinician verification, facilitates auditability, and enables model outputs to be interpreted alongside existing clinical assessments rather than replacing clinical judgement.

The findings indicate that explainability and predictive performance can be achieved simultaneously. By combining strong prediction with transparent reasoning, the proposed model provides clinically actionable risk estimates that support early suicide risk stratification while maintaining human oversight, consistent with contemporary recommendations for trustworthy and explainable AI in healthcare (Molnar, 2025; Walsh et al., 2025).

RQ 3: Which clinical and contextual features contribute most strongly to SRB prediction at triage?

The findings indicate that short-term SRB risk at ED triage is influenced by the interaction of demographic, physiological, psychiatric, and contextual factors rather than any single predictor. Across the global and local explainability analyses, age consistently emerged as the most influential contributor, while physiological measurements, including blood pressure, shock index, temperature, heart rate, respiratory rate, and oxygen saturation, provided important information for estimating acute risk. Psychiatric history, particularly prior mood, anxiety, and substance use disorders, further modified individual risk estimates, reflecting the interaction between underlying vulnerability and acute clinical presentation. The results also demonstrate that routinely collected ED triage data contain clinically meaningful information beyond conventional physiology-based assessment. Rather than relying solely on psychiatric history or presenting complaint, the model integrated multiple complementary sources of information to predict patient-level SRB risk. This aligns with contemporary suicide prevention frameworks, which recognise suicidal behaviour as a multifactorial process arising from the interaction of predisposing vulnerabilities, acute physiological status, and contextual stressors (Haglund et al., 2019; Hughes et al., 2023; Walsh et al., 2025).

The findings show that structured ED triage variables provide a clinically relevant and operationally feasible foundation for early suicide risk prediction, supporting timely prioritisation of patients for further mental health assessment while remaining compatible with routine emergency care workflows.

Implications of the study

Theoretical Implications

This study provides empirical evidence that routinely collected structured ED triage variables contain sufficient information to support short-term SRB risk prediction without requiring unstructured clinical narratives or specialised assessments. The findings extend existing suicide prediction literature by demonstrating that demographic, physiological, psychiatric, and contextual variables jointly contribute to risk prediction, supporting contemporary conceptualisations of suicidal behaviour as a multifactorial and dynamic process rather than the consequence of isolated risk factors. The results further suggest that structured triage data can operationalise clinically meaningful risk signals at the earliest stage of emergency care, providing a theoretical foundation for triage-based suicide prediction models.

The study also contributes to the growing body of literature on explainable AI in healthcare. By combining global SHAP analysis, intrinsic LightGBM feature importance, and patient-level explanations, the findings demonstrate that predictive performance and interpretability are not mutually exclusive. The consistency observed across global and local explanations provides evidence that explainable AI can support transparent and clinically coherent reasoning, thereby contributing to theories of trustworthy and human-centred AI. Rather than viewing explainability as a post hoc add-on, the results suggest that interpretability can be integrated into the model development and evaluation process to enhance transparency, auditability, and clinician trust.

Furthermore, the findings challenge the conventional practice of evaluating clinical prediction models primarily through discrimination metrics such as AUROC. Despite similar AUROC values across models, important differences emerged in minority-class detection, threshold behaviour, and clinical utility. This study therefore

supports a multidimensional evaluation paradigm for rare-event prediction, where AUPRC, F2-score, decision-curve analysis, threshold optimisation, bootstrap confidence intervals, and explainability complement traditional discrimination metrics. Such an approach provides a more comprehensive theoretical framework for evaluating AI systems intended for safety-critical healthcare applications.

The study contributes to the theoretical understanding of AI deployment in emergency psychiatry by emphasising that model selection and deployment configuration represent distinct processes. The findings demonstrate that the most suitable predictive model is not necessarily the one that optimises a single metric, but rather the one that provides the most favourable balance between discrimination, minority-class detection, decision stability, interpretability, and operational utility. This perspective supports emerging views of clinical AI as a socio-technical decision-support model, in which predictive models function as clinician-supportive tools integrated into existing workflows rather than autonomous diagnostic systems.

Practical Implications

From a practical perspective, the study demonstrates the feasibility of using structured triage variables available at the time of patient presentation to predict short-term SRB risk. Because these variables are routinely collected during standard ED assessment, the proposed approach has the potential to support integration into clinical decision-support systems without imposing substantial additional documentation burden. Furthermore, explainability techniques such as SHAP can provide transparent patient-level explanations, potentially improving clinicians' understanding and trust in model outputs.

From an operational perspective, the selected LightGBM model achieved a favourable balance between discrimination, minority-class detection, and decision stability. Threshold optimisation further demonstrated how sensitivity and alert burden can be balanced in high-throughput ED settings, where excessive false-positive alerts may contribute to alert fatigue and reduce the effectiveness of decision-support systems. Thus, threshold calibration should be adapted to local clinical priorities and resource availability. Despite these potential benefits, successful implementation cannot be assumed. Practical deployment would require clinician acceptance, workflow redesign, user-interface considerations, governance mechanisms, and clear delineation of responsibility for acting on model recommendations. Medicolegal and liability concerns, together with the risk of alert fatigue, represent important challenges that may influence adoption. Therefore, the proposed model should be viewed as a clinician-support tool intended to augment, rather than replace, clinical judgement. Prospective validation, usability studies, and implementation research will be necessary before routine deployment can be considered.

By leveraging routinely collected triage information and explainable ML, the model provides a scalable foundation for supporting early identification of patients at elevated short-term risk of SRB, particularly in resource-constrained healthcare settings. The proposed model is intended to complement rather than replace established suicide screening and assessment instruments, such as the ASQ, C-SSRS, and SIQ. Because the model relies exclusively on information available at ED presentation, it may provide an initial estimate of short-term SRB risk before a comprehensive psychiatric assessment. Patients identified as elevated risk could subsequently undergo formal screening and clinical evaluation, allowing AI-assisted prioritisation to function alongside existing standards of care. To minimise alert fatigue, risk estimates should be interpreted within a hierarchical framework rather than as binary alerts. For example, low-risk patients may proceed through routine care pathways, intermediate-risk patients may prompt additional screening or observation, and high-risk patients may trigger urgent psychiatric consultation and enhanced safety procedures. Such tiered strategies may reduce unnecessary interruptions while preserving sensitivity for patients at greatest risk.

As illustrated in Figure 14, risk predictions can be incorporated into a tiered clinical workflow, enabling interventions tailored to each risk level. This hierarchical approach may reduce unnecessary alerts and support more efficient allocation of limited psychiatric resources while maintaining clinician involvement in all decision-making processes.

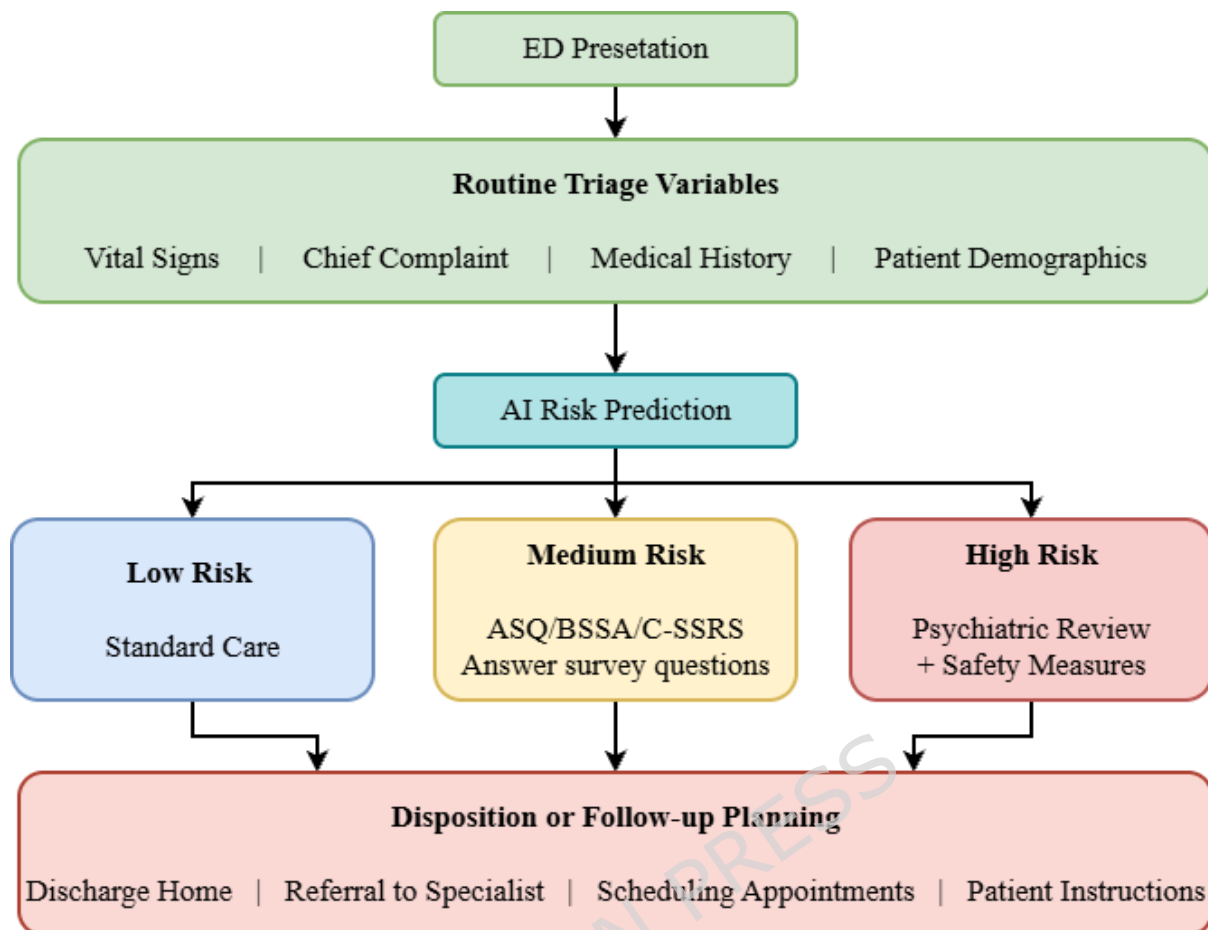


Figure 14. The proposed integration of the LightGBM-based 30-day SRB risk prediction model within ED workflows

Ethical Considerations

The deployment of AI systems in clinical settings, particularly for sensitive outcomes such as SRB, requires careful ethical consideration. In this study, the proposed LightGBM model is explicitly conceptualised as a clinical decision-support tool rather than a diagnostic system. Clinical responsibility remains with healthcare professionals, and the model is intended solely to support prioritisation and follow-up planning within existing ED workflows. The model does not make autonomous clinical decisions and should not be used in isolation to determine patient care. Instead, its outputs are designed to complement established screening protocols and clinician expertise by providing early, data-driven predictions of 30-day SRB risk based on routinely collected triage information. This human-in-the-loop approach aligns with current best practices for trustworthy AI in healthcare and emphasises accountability, transparency, and clinician oversight.

The use of de-identified MIMIC-IV data ensured patient privacy and complied with the ethical requirements governing access to the database. Although informed consent was waived because the dataset is retrospective and anonymised, prospective deployment would require careful consideration of patient privacy, data governance, transparency, and institutional policies governing AI-assisted clinical decision support. From an operational perspective, the model is designed to support risk stratification rather than definitive diagnosis, ensuring that elevated-risk predictions are followed by additional clinical assessment rather than automatic intervention. Responsibility for patient management, therefore, remains with healthcare professionals, while AI predictions serve to augment clinical judgement. This approach helps mitigate potential harms associated with false-positive alerts and missed high-risk cases.

Fairness and equity represent additional ethical considerations. Because the model was developed using a single-centre retrospective dataset, predictions may reflect the underlying demographic and clinical patterns of the source population. Although subgroup-specific fairness analyses were beyond the scope of the present study, future work should evaluate model performance across demographic and clinical subgroups and investigate performance among patients without documented psychiatric histories to ensure equitable detection of vulnerable individuals.

The proposed model aligns with the principles of beneficence, non-maleficence, transparency, and clinical accountability by augmenting, rather than replacing, human decision-making in emergency care.

Limitations and Future Work

While the study demonstrated strong discrimination, explainability and deployment-oriented evaluation, several limitations should be acknowledged. First, the data originate from the MIMIC-IV database, representing a single healthcare network in the United States. This may limit the generalisability of the findings to other populations, particularly in low- and middle-income countries. Patient demographics, clinical workflows, documentation practices, and coding conventions may differ across hospitals and regions. Although patient-level splitting and bootstrap confidence intervals strengthened the study's internal validity, they do not guarantee transportability to other institutions or populations. Therefore, the findings should be interpreted as demonstrating internal validity within the MIMIC-IV setting rather than universal generalisability.

Second, the patient-level splitting was used to mitigate data leakage; the study used retrospective data and did not include prospective validation. As a result, real-world clinical deployment may reveal additional workflow, usability, or operational constraints not captured in retrospective analyses. Third, the study relied solely on structured features. While this approach improves interpretability and operational readiness, it excludes linguistic and contextual signals captured in clinical notes or patient narratives, which may further enhance risk detection and provide richer contextual understanding. Fourth, another limitation relates to the use of ICD-9 and ICD-10 diagnostic codes to define SRB outcomes. Although ICD-based definitions are widely used in EHR-based suicide prediction studies and provide a reproducible outcome measure, previous studies have shown that administrative coding may under-identify self-harm and suicide-related presentations. Therefore, some patients experiencing SRB may have been misclassified as non-SRB cases. Because no independent reference standard or manual chart review was available, the extent of this misclassification could not be directly quantified in the present cohort. Such under-detection may introduce label noise during model development and lead to conservative estimates of model performance by weakening the true distinction between positive and negative cases. Even so, the use of ICD codes enabled consistent cohort construction and aligns with established approaches employed in large-scale retrospective studies. Fifth, SRB cases are a relatively uncommon event. Although class weighting and threshold optimisation improved minority-class detection, class imbalance remains an inherent challenge, and no model achieved perfect sensitivity. Thus, some high-risk patients may still be missed.

Future work should proceed in phases. The first phase should focus on external, temporal, and multi-centre validation, including model recalibration and local threshold optimisation to assess transportability across healthcare systems. The second phase should investigate multimodal approaches that combine structured ED variables with unstructured clinical narratives and other contextual data. The third phase should involve prospective clinical studies evaluating AI-assisted risk prediction within routine ED workflows, including clinician acceptance, alert fatigue, and ethical considerations. Finally, formal health-economic evaluations are required to quantify the costs and benefits of AI-assisted workflows, including psychiatric consultation workload, false-positive burden, missed high-risk cases, resource utilisation, and potential reductions in adverse outcomes. Such studies are essential for establishing the practical value and cost-effectiveness of AI-supported suicide risk prediction systems.

5. Conclusion

This study demonstrated that the risk of experiencing documented SRB within 30 days following an ED encounter can be predicted using routinely collected, structured triage variables available at the time of presentation, without relying on delayed documentation or unstructured clinical narratives. Multiple ML and DL models were evaluated using a clinically grounded multidimensional framework incorporating discrimination, minority-class detection, decision stability, threshold optimisation, decision-curve analysis, bootstrap confidence intervals, and explainability. Among the evaluated models, LightGBM provided the most favourable balance between overall discrimination, minority-class performance, sensitivity, and decision stability, making it the most suitable candidate for deployment within the present model. The findings highlight that model selection in rare-event clinical prediction should not rely on a single metric, as no individual measure fully captures clinical utility. Instead, deployment decisions require balancing sensitivity, minority-class discrimination, alert burden, and operational feasibility.

Explainability analyses demonstrated that clinically meaningful variables contributed to risk estimation and provided transparent insights into model behaviour. By integrating uncertainty quantification, threshold

optimisation, decision-curve analysis, and SHAP-based explanations, the study contributes a deployment-oriented framework that bridges predictive performance, interpretability, and clinical applicability. Several limitations remain, including the use of retrospective single-centre data, reliance on ICD-based outcome definitions, and the absence of prospective validation. Future work should prioritise external validation, multimodal data integration, prospective workflow evaluation, and health-economic analyses to establish the transportability, clinical impact, and cost-effectiveness of AI-assisted suicide risk prediction. The study provides evidence that routinely collected ED triage data contain sufficient information to support early identification of patients at risk of SRB within 30 days and offers a transparent model for integrating AI-assisted risk prediction into emergency care. Overall, the proposed model demonstrates that hospitals can estimate short-term suicide-related behaviour risk at the point of emergency department presentation using routinely collected structured triage data. By supporting early patient prioritisation, follow-up planning, and resource allocation, the framework provides a practical foundation for integrating transparent AI-assisted decision support into routine emergency mental healthcare.

Ethics approval and consent to participate

This study utilised the MIMIC-IV database, a publicly available, de-identified clinical dataset. The MIMIC-IV database was approved by the Institutional Review Boards of the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (BIDMC), and informed consent was waived due to the de-identified nature of the data. Access to the dataset was obtained after completion of the required data use training and credentialing process via PhysioNet. As this study involved secondary analysis of fully anonymised data, additional ethical approval was not required. Due to the retrospective nature of the study, the requirement for obtaining informed consent was waived.

Data availability

The data that support the findings of this study are publicly available from the MIMIC-IV database on PhysioNet (DOI: <https://doi.org/10.13026/kpb9-mt58>). Access to the dataset is subject to credentialed approval; researchers must complete the required training and sign a data use agreement with PhysioNet prior to access. The specific data extraction queries and derived variables used in this study are available from the corresponding author upon request.

Code availability

The code used for data preprocessing, model development, and evaluation is available from the corresponding author upon reasonable request. To support reproducibility, the SQL extraction scripts used to construct the study cohort and derive structured variables from MIMIC-IV will be provided as supplementary material. Due to institutional and data-use constraints, the direct public release of certain components may be restricted, but reasonable efforts will be made to facilitate replication of the study.

Author contributions

T.M. contributed to conceptualisation, data curation, formal analysis, methodology, and writing of the original draft, as well as review and editing of the manuscript. T.M.II contributed to methodology, investigation, formal analysis, and review and editing of the manuscript. P.N. contributed to methodology, investigation, supervision, formal analysis, and review and editing of the manuscript. N.N. contributed to methodology, investigation, supervision, and review and editing of the manuscript. All authors reviewed and approved the final manuscript.

Funding

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Acknowledgements

The authors acknowledge the PhysioNet team for providing access to the MIMIC-IV database and supporting reproducible research in critical care and emergency medicine.

References

- Abugroun, A., Awadalla, S., Singh, S., & Fang, M. C. (2025). Development of an emergency department triage tool to predict admission or discharge for older adults. *International Journal of Emergency Medicine*, 18(1), 26. <https://doi.org/10.1186/s12245-025-00825-3>
- Ahn, H., Kim, H.-Y., Choi, K.-B., Jo, H., Song, S., & Kim, J. (2025). Comparison of factors distinguishing suicide attempters among suicidal ideators and factors associated with suicidal ideation and attempts using machine learning models in South Korean adolescents. *Journal of Psychiatric Research*, 191, 790–797. <https://doi.org/10.1016/j.jpsychires.2025.10.020>
- Alghazzawi, D., Ullah, H., Tabassum, N., Badri, S. K., & Asghar, M. Z. (2025). Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique. *Scientific Reports*, 15(1), 1111. <https://doi.org/10.1038/s41598-024-84275-6>
- Almadhoun, M. B., & Burhanuddin, M. (2025). Optimizing Feature Selection and Machine Learning Algorithms for Early Detection of Prediabetes Risk: Comparative Study. *JMIR Bioinformatics and Biotechnology*, 6, e70621–e70621. <https://doi.org/10.2196/70621>
- American Psychiatric Association. (2022). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR)*. American Psychiatric Association Publishing. <https://doi.org/10.1176/appi.books.9780890425787>
- Bayat Mokhtari, L., Sanjari, S., Kikha, A., Mohammadi Soliemani, M. R., Abbasi, H., & Jahan, M. R. (2025). A Cross-sectional Study of the Correlations of Sleep Quality, Anxiety, and Social Support with Academic Engagement: The Mediating Role of Mental Health. *Shiraz E-Medical Journal*, 26(10). <https://doi.org/10.5812/semj-160489>
- Bazrafshan, M., & Sayehmiri, K. (2024). Predicting suicidal behavior outcomes: an analysis of key factors and machine learning models. *BMC Psychiatry*, 24(1), 841. <https://doi.org/10.1186/s12888-024-06273-2>
- Bentley, K. H., Kennedy, C. J., Khadse, P. N., Brooks Stephens, J. R., Madsen, E. M., Flics, M. J., Lee, H., Smoller, J. W., & Burke, T. A. (2025). Clinician Suicide Risk Assessment for Prediction of Suicide Attempt in a Large Health Care System. *JAMA Psychiatry*, 82(6), 599. <https://doi.org/10.1001/jamapsychiatry.2025.0325>
- Buwaneswari, P. R., G K, M., Alasady, H., Alagarraja, K., & Soni, M. (2025). Anomaly-Based Intrusion Detection Systems by using Machine Learning based Stacking Ensemble Model. *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS)*, 1–5. <https://doi.org/10.1109/ICICACS65178.2025.10967781>
- Carson, N. J., Yang, X., Mullin, B., Stettenbauer, E., Waddington, M., Zhang, A., Williams, P., Rios Perez, G. E., & Cook, B. L. (2024). Predicting adolescent suicidal behavior following inpatient discharge using structured and unstructured data. *Journal of Affective Disorders*, 350, 382–387. <https://doi.org/10.1016/j.jad.2023.12.059>
- CDC. (2025, March 26). *Facts About Suicide*. Centers for Disease Control and Prevention. <https://www.cdc.gov/suicide/facts/index.html>
- Cortellazzo Wiel, L., D'Agostin, M., Curatola, A., Castagno, E., Giangreco, M., Bondone, C., Rulli, E., Ramlakhan, S., Congiu, G., Amaddeo, A., & Cozzi, G. (2025). Trend of emergency department attendances for alcohol and substances intoxication in adolescents across the COVID-19 pandemic: a multicenter international study. *European Journal of Pediatrics*, 185(1), 40. <https://doi.org/10.1007/s00431-025-06699-0>
- Degirmenci, S., & Mercan, N. (2025). Psychiatric emergencies: epidemiological analysis and healthcare professionals' experiences. *BMC Emergency Medicine*, 25(1), 109. <https://doi.org/10.1186/s12873-025-01268-y>
- Department of Public Health and Human Services. (2025, October 27). *Suicide Prevention in the Emergency Department*. Montana Department of Public Health and Human Services. <https://dphhs.mt.gov/suicideprevention/toolkit/emergencydepartment>
- Diaz-Vico, D., Figueiras-Vidal, A. R., & Dorronsoro, J. R. (2018). Deep MLPs for Imbalanced Classification. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–7. <https://doi.org/10.1109/IJCNN.2018.8489504>
- Dobscha, S. K., Newell, S. B., Elliott, V. J., Rynerson, A. L., Rabin, S., Bahraini, N., Post, E. P., & Denneson, L. M. (2023). Primary Care and Mental Health Staff Perspectives on Universal Suicide Risk Screening and Care Coordination. *Psychiatric Services*, 74(3), 305–311. <https://doi.org/10.1176/appi.ps.20220087>
- Edgecomb, J. B., Olde Loohuis, L., Tseng, C., Klonhaus, A. M., Choi, K. R., Ponce, C. G., & Zima, B. T. (2024). Electronic Health Record Phenotyping of Pediatric Suicide-Related Emergency Department Visits. *JAMA Network Open*, 7(10), e2442091. <https://doi.org/10.1001/jamanetworkopen.2024.42091>

- Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC. <https://doi.org/10.1201/9780429246593>
- El Arab, R. A., & Al Moosa, O. A. (2025). The role of AI in emergency department triage: An integrative systematic review. *Intensive and Critical Care Nursing*, 89, 104058. <https://doi.org/10.1016/j.iccn.2025.104058>
- Erdem, P. C., Gica, S., Cinar, B., & Karamustafalioglu, N. (2025). The effect of impoverishment of thought in early psychosis: Influence on suicide risk via depression and aggression. *Dusunen Adam: Journal of Psychiatry and Neurological Sciences*, 136–147. <https://doi.org/10.14744/DAJPNS.2025.00285>
- Gao, Y.-H., Huang, X.-W., & Meng, Y.-M. (2025). A systematic evaluation of data preprocessing and model optimization for machine learning algorithms: Using sphalerite trace element data as an example. *Journal of Asian Earth Sciences*, 292, 106728. <https://doi.org/10.1016/j.jseaes.2025.106728>
- Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*, 28(6), 1240–1248. <https://doi.org/10.1038/s41591-022-01811-5>
- Gaur, V., Maggu, G., Bairwa, K., Chaudhury, S., Dhamija, S., & Ali, T. (2024). Artificial intelligence in suicide prevention: Utilizing deep learning approach for early detection. *Industrial Psychiatry Journal*, 33(2), 226–233. https://doi.org/10.4103/ipj.ipj_20_24
- Ghonem, M. M., Abdelnoor, A. A., & Hodeib, A. A. (2024). Shock and modified shock indices in predicting poisoning severity and outcomes in acute aluminum phosphide poisoned patients. *Toxicology Research*, 13(1). <https://doi.org/10.1093/toxres/tfad124>
- Gomes, S., Dhanoa, H., Assheton, P., Carr, E., Roland, D., & Deep, A. (2025). Predicting sepsis treatment decisions in the paediatric emergency department using machine learning: the AiSEPTRON study. *BMJ Paediatrics Open*, 9(1), e003273. <https://doi.org/10.1136/bmjpo-2024-003273>
- Haglund, A., Lysell, H., Larsson, H., Lichtenstein, P., & Runeson, B. (2019). Suicide Immediately After Discharge From Psychiatric Inpatient Care. *The Journal of Clinical Psychiatry*, 80(2). <https://doi.org/10.4088/JCP.18m12172>
- Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. H. (2025). Ethical and Bias Considerations in Artificial Intelligence/Machine Learning. *Modern Pathology*, 38(3), 100686. <https://doi.org/10.1016/J.MODPAT.2024.100686>
- Homan, S., Gabi, M., Klee, N., Bachmann, S., Moser, A. M., Duri, M., Michel, S., Bertram, A. M., Maatz, A., Seiler, G., Stark, E., & Kleim, B. (2022). Linguistic features of suicidal thoughts and behaviors: A systematic review. *Clinical Psychology Review*, 95, 102161. <https://doi.org/10.1016/J.CPR.2022.102161>
- Hughes, C. L., Liu, B., & Qin, P. (2025). Socioeconomic, demographic, and occupational risk factors for suicide amongst males: a systematic review and meta-analysis. *Journal of Affective Disorders*, 391, 119986. <https://doi.org/10.1016/j.jad.2025.119986>
- Hughes, J. L., Horowitz, L. M., Ackerman, J. P., Adrian, M. C., Campo, J. V., & Bridge, J. A. (2023). Suicide in young people: screening, risk assessment, and intervention. *BMJ*, e070630. <https://doi.org/10.1136/bmj-2022-070630>
- Humbert-Droz, M., Mukherjee, P., & Gevaert, O. (2022). Strategies to Address the Lack of Labeled Data for Supervised Machine Learning Training With Electronic Health Records: Case Study for the Extraction of Symptoms From Clinical Notes. *JMIR Medical Informatics*, 10(3), e32903. <https://doi.org/10.2196/32903>
- Interian, A., Kimbrel, N. A., Goodman, M. S., Constans, J. I., Dobscha, S. K., Ilgen, M., Owen, R. R., Strickland, S. M., & Marx, B. P. (2025). Precision Medicine for Suicide Prevention. *The Journal of Clinical Psychiatry*, 87(1). <https://doi.org/10.4088/JCP.25com16004>
- Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., & Mark, R. (2024, October 11). *MIMIC-IV*. PhysioNet. <https://doi.org/https://doi.org/10.13026/kpb9-mt58>
- Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *Npj Digital Medicine*, 6(1), 6. <https://doi.org/10.1038/s41746-023-00751-9>
- Karagounis, S., Sarkar, I. N., & Chen, E. S. (2020). Coding Free-Text Chief Complaints from a Health Information Exchange: A Preliminary Study. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 2020, 638–647.
- Khazem, S., & Kanso, H. (2025). *Cyclical Temporal Encoding and Hybrid Deep Ensembles for Multistep Energy Forecasting*.

- Kim, J., Lim, H., Ahn, J.-H., Lee, K. H., Lee, K. S., & Koo, K. C. (2021). Optimal Triage for COVID-19 Patients Under Limited Health Care Resources With a Parsimonious Machine Learning Prediction Model and Threshold Optimization Using Discrete-Event Simulation: Development Study. *JMIR Medical Informatics*, 9(11), e32726. <https://doi.org/10.2196/32726>
- Kim, S., & Lee, K. (2025). Machine learning-based prediction of suicide risk using adult attention-deficit/hyperactivity disorder symptoms and depression indicators: insights from a nationally representative south korean survey. *Psychiatry Research*, 352, 116702. <https://doi.org/10.1016/j.psychres.2025.116702>
- Kishore, K., Braitberg, G., Holmes, N. E., & Bellomo, R. (2023). Early prediction of hospital admission of emergency department patients. *Emergency Medicine Australasia*, 35(4), 572–588. <https://doi.org/10.1111/1742-6723.14169>
- Kitchen, C., Belouali, A., Nestadt, P. S., Wilcox, H. C., & Kharrazi, H. (2026). Evaluating the role of cost and fairness in machine learning models of suicide risk. *Scientific Reports*, 16(1), 19246. <https://doi.org/10.1038/s41598-026-50095-z>
- Kleiman, E. M., Bentley, K. H., Maimone, J. S., Lee, H.-I. S., Kilbury, E. N., Fortgang, R. G., Zuromski, K. L., Huffman, J. C., & Nock, M. K. (2021). Can passive measurement of physiological distress help better predict suicidal thinking? *Translational Psychiatry*, 11(1), 611. <https://doi.org/10.1038/s41398-021-01730-y>
- Li, J., Zhao, Y., Yang, C., Li, W., Huang, Z., & Fan, C. (2025). Prediction of suicidal attempts among Chinese adolescents with mood disorders: a clinical study using a machine learning approach. *BMC Psychiatry*, 25(1), 703. <https://doi.org/10.1186/s12888-025-07145-z>
- Lu, H., Kondo, T., Claggett, B. L., Vaduganathan, M., Neuen, B. L., Beldhuis, I. E., Jhund, P. S., Mc Causland, F. R., Anand, I. S., Pfeffer, M. A., Pitt, B., Zannad, F., Zile, M. R., McMurray, J. J. V., Solomon, S. D., & Desai, A. S. (2025). Systolic Blood Pressure and Pulse Pressure in Heart Failure. *Journal of the American College of Cardiology*, 85(7), 710–722. <https://doi.org/10.1016/j.jacc.2024.11.007>
- Machado, D. B. (2025). Rethinking suicide prevention: insights from the global south for a new global agenda. *Nature Mental Health*, 3(9), 982–990. <https://doi.org/10.1038/s44220-025-00491-7>
- Martin, B., Bennett, T. D., DeWitt, P. E., Russell, S., & Sanchez-Pinto, L. N. (2025). Use of the Area Under the Precision-Recall Curve to Evaluate Prediction Models of Rare Critical Illness Events. *Pediatric Critical Care Medicine*, 26(6), e855–e859. <https://doi.org/10.1097/PCC.0000000000003752>
- McCabe, D. J., Gibbs, H., Falise, A., Hagemann, J., Kroll, C., Culbreth, R., Hendrickson, R., Campleman, S., Wax, P., Brent, J., & Aldy, K. (2026). Level of care following suspected opioid or stimulant overdose with self-harm intent. *The American Journal of Emergency Medicine*, 99, 264–269. <https://doi.org/10.1016/j.ajem.2025.10.016>
- McDermott, M., Zhang, H., Hansen, L., Angelotti, G., & Gallifant, J. (2024). A Closer Look at AUROC and AUPRC under Class Imbalance. *Advances in Neural Information Processing Systems* 37, 44102–44163. <https://doi.org/10.52202/079017-1400>
- McHugh, C. M., & Large, M. M. (2020). Can machine-learning methods really help predict suicide? *Current Opinion in Psychiatry*, 33(4), 369–374. <https://doi.org/10.1097/YCO.0000000000000609>
- Mokheleli, T., Makaba, T., Ndayizigamiye, P., Ndlovu, N., & Twinomurinzi, H. (2026). Explainable Machine Learning for Suicide Risk Assessment Using Social Media Data. *IEEE Access*, 14, 77656–77676. <https://doi.org/10.1109/ACCESS.2026.3694163>
- Mokhtar, S. F., Md Yusof, Z., & Sapiri, H. (2023). Confidence Intervals by Bootstrapping Approach: A Significance Review. *Malaysian Journal of Fundamental and Applied Sciences*, 19(1), 30–42. <https://doi.org/10.11113/mjfas.v19n1.2660>
- Molnar, Christoph. (2025). *Interpretable machine learning: a guide for making black box models explainable*. Christoph Molnar.
- Nabar, M. M. Y., Slipak, D., Algieri, D., Álvarez, S., & Progetti, M. A. (2026). Standardization of Documentation in Prehospital Psychiatric Emergencies: Analysis of a Specific form Implemented for the Psychiatric Ambulance in the City of Buenos Aires. *Open Journal of Emergency Medicine*, 14(02), 101–109. <https://doi.org/10.4236/ojem.2026.142008>
- Nock, M. K., Millner, A. J., Ross, E. L., Kennedy, C. J., Al-Suwaidi, M., Barak-Corren, Y., Castro, V. M., Castro-Ramirez, F., Lauricella, T., Murman, N., Petukhova, M., Bird, S. A., Reis, B., Smoller, J. W., & Kessler, R. C. (2022). Prediction of Suicide Attempts Using Clinician Assessment, Patient Self-report, and Electronic Health Records. *JAMA Network Open*, 5(1), e2144373. <https://doi.org/10.1001/jamanetworkopen.2021.44373>

- Ntakolia, C., Yotsidi, V., Rannou, I., & Gournellis, R. (2025). Interpretable Machine Learning approach for predicting clinically significant suicide risk: A case study of patients with major depressive disorder in Greece. *Psychiatry Research*, 351, 116607. <https://doi.org/10.1016/j.psychres.2025.116607>
- Nugent, S. M., Anderson, J., & Young, S. K. (2024). Behavioural mental health interventions delivered in the emergency department for suicide, overdose and psychosis: a scoping review. *BMJ Open*, 14(3), e080023. <https://doi.org/10.1136/bmjopen-2023-080023>
- Olfson, M., Gao, Y. N., Xie, M., Wiesel Cullen, S., & Marcus, S. C. (2021). Suicide Risk Among Adults With Mental Health Emergency Department Visits With and Without Suicidal Symptoms. *The Journal of Clinical Psychiatry*, 82(6). <https://doi.org/10.4088/JCP.20m13833>
- Oweimieotu, A. E., Edje, A. E., Akazue, , Maureen Ifeanyi, & Asuai, C. (2024). Designing a Hybrid Genetic Algorithm Trained Feedforward Neural Network for Mental Health Disorder Detection. *Advances in Multidisciplinary & Scientific Research Journal Publications*, 12(1), 49–62. <https://doi.org/10.22624/AIMS/DIGITAL/V11N4P4>
- Pagliaccio, D., Kirshenbaum, J. S., Greenblatt, J., Mroczkowski, M. M., Chernick, L. S., Dayan, P. S., & Auerbach, R. P. (2026). Suicide Risk in 8- to 12-Year-Old Children: Retrospective Review of Emergency Department Electronic Health Records. *JAACAP Open*, 4(1), 90–100. <https://doi.org/10.1016/j.jaacop.2025.08.002>
- Peralta, D. (2024). AI and suicide risk prediction: Facebook live and its aftermath. *AI & SOCIETY*, 39(4), 2155–2167. <https://doi.org/10.1007/s00146-023-01651-y>
- Pigoni, A., Delvecchio, G., Turtulici, N., Madonna, D., Pietrini, P., Cecchetti, L., & Brambilla, P. (2024). Machine learning and the prediction of suicide in psychiatric populations: a systematic review. *Translational Psychiatry*, 14(1), 140. <https://doi.org/10.1038/s41398-024-02852-9>
- Prasher, S., Nelson, L., & Gafar, M. (2025). NIPP: Non-Invasive PCOS Prediction using XG-boost Machine Learning Model. *International Journal of Information Technology and Computer Science*, 17(1), 82–95. <https://doi.org/10.5815/ijitcs.2025.01.06>
- Rabani, S. T., Ud Din Khanday, A. M., Khan, Q. R., Hajam, U. A., Imran, A. S., & Kastrati, Z. (2023). Detecting suicidality on social media: Machine learning at rescue. *Egyptian Informatics Journal*, 24(2), 291–302. <https://doi.org/10.1016/J.EIJ.2023.04.003>
- Radunz, M., Johnson, C., O'Dea, B., & Wade, T. D. (2025). Interventions for self-harm and suicidality in paediatric emergency departments: a meta-review. *European Child & Adolescent Psychiatry*, 34(9), 2649–2658. <https://doi.org/10.1007/s00787-025-02706-9>
- Rana, R., Higgins, N., Haque, K. N., Burke, K., Turner, K., & Stedman, T. (2024). Feasibility of Mental Health Triage Call Priority Prediction Using Machine Learning. *Nursing Reports*, 14(4), 4162–4172. <https://doi.org/10.3390/nursrep14040303>
- Ren, W., Zhao, T., Huang, Y., & Honavar, V. (2025). *Deep Learning within Tabular Data: Foundations, Challenges, Advances and Future Directions*.
- Rheinberger, D., Wang, J., McGillivray, L., Shand, F., Torok, M., Maple, M., & Wayland, S. (2022). Understanding Emergency Department Healthcare Professionals' Perspectives of Caring for Individuals in Suicidal Crisis: A Qualitative Study. *Frontiers in Psychiatry*, 13. <https://doi.org/10.3389/fpsy.2022.918135>
- Riley, R. D., Ensor, J., Snell, K. I. E., Archer, L., Whittle, R., Dhiman, P., Alderman, J., Liu, X., Kirton, L., Manson-Whitton, J., van Smeden, M., Moons, K. G., Nirantharakumar, K., Cazier, J.-B., Denniston, A. K., Van Calster, B., & Collins, G. S. (2025). Importance of sample size on the quality and utility of AI-based prediction models for healthcare. *The Lancet Digital Health*, 7(6), 100857. <https://doi.org/10.1016/j.landig.2025.01.013>
- Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, m441. <https://doi.org/10.1136/bmj.m441>
- Robinson, J., Witt, K., Lamblin, M., Spittal, M. J., Carter, G., Verspoor, K., Page, A., Rajaram, G., Rozova, V., Hill, N. T. M., Pirkis, J., Bleeker, C., Pleban, A., & Knott, J. C. (2020). Development of a Self-Harm Monitoring System for Victoria. *International Journal of Environmental Research and Public Health*, 17(24), 9385. <https://doi.org/10.3390/ijerph17249385>
- Sano, T., Yonemoto, N., Narita, K., Nemoto, K., Ishibashi, R., Hirayasu, Y., & Kawanishi, C. (2026). Effectiveness of psychoeducation for family members of suicide attempters in preventing recurrent suicidal behaviour: A secondary analysis of the ACTION-J Study. *Psychiatry Research*, 356, 116865. <https://doi.org/10.1016/j.psychres.2025.116865>

- Scarffe, A., Coates, A., Brand, K., & Michalowski, W. (2024). Decision threshold models in medical decision making: a scoping literature review. *BMC Medical Informatics and Decision Making*, 24(1), 273. <https://doi.org/10.1186/s12911-024-02681-2>
- Scudder, A., Rosin, R., Baltich Nelson, B., Boudreaux, E. D., & Larkin, C. (2022). Suicide Screening Tools for Pediatric Emergency Department Patients: A Systematic Review. *Frontiers in Psychiatry*, 13. <https://doi.org/10.3389/fpsyt.2022.916731>
- Shah, V. (2022). AI in Mental Health: Predictive Analytics and Intervention Strategies. *Journal of Environmental Sciences and Technology*, 1(2). <https://doi.org/10.5281/zenodo.10779085>
- Shi, T., Merai, R., Farmer, C., Wei, A., Miller, I. W., Camargo, C. A., Pao, M., Bridge, J. A., Boudreaux, E. D., & Horowitz, L. M. (2026). The association between recency of suicide attempt and reattempt in adult emergency department patients. *General Hospital Psychiatry*, 98, 157–158. <https://doi.org/10.1016/j.genhosppsych.2025.12.017>
- Shiraly, R., Jazayeri, S. A., Seifaei, A., Jeihooni, A. K., & Griffiths, M. D. (2024). Suicidal thoughts and behaviors among untreated illicit substance users: a population-based study. *Harm Reduction Journal*, 21(1), 96. <https://doi.org/10.1186/s12954-024-01015-9>
- Shukla, S. S. P., & Singh, M. P. (2025). Enhancing suicidal ideation detection through advanced feature selection and stacked deep learning models. *Applied Intelligence*, 55(5), 303. <https://doi.org/10.1007/s10489-025-06256-0>
- Sithiprawiat, P., Wittayachamnankul, B., Sirikul, W., & Laohavisudhi, K. (2025). Development and internal validation of an AI-based emergency triage model for predicting critical outcomes in emergency department. *Scientific Reports*, 15(1), 31212. <https://doi.org/10.1038/s41598-025-17180-1>
- Song, M., Kim, J., Jang, E.-C., & Kwon, S. (2026). Interpretable Machine Learning for Emergency Department Triage: Clinical Insights from 133,198 Patients Using the Korean Triage and Acuity Scale (KTAS). *Diagnostics*, 16(6), 954. <https://doi.org/10.3390/diagnostics16060954>
- Stenwig, E., Salvi, G., Rossi, P. S., & Skjærvold, N. K. (2022). Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Medical Research Methodology*, 22(1), 53. <https://doi.org/10.1186/s12874-022-01540-w>
- Sutone, R., & Malhotra, R. (2025). Lifestyle and Behavioral Data-Driven Mental Health Prediction: Insights from Machine Learning Models. 2025 IEEE 6th India Council International Subsections Conference (INDISCON), 1–6. <https://doi.org/10.1109/INDISCON66021.2025.11253883>
- Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of Suicide Ideation in Social Media Forums Using Deep Learning. *Algorithms*, 13(1), 7. <https://doi.org/10.3390/a13010007>
- Tasnim, S. S., Mohtasim, S. M. A., Priya, S. A., Nahar, R., Ferdous, J., & Elma, S. A. (2024). Optimizing Predictive Models for Drug-Induced Suicidal Risk: A Data-Balanced Framework with Explainable AI. 2024 27th International Conference on Computer and Information Technology (ICCIT), 1475–1480. <https://doi.org/10.1109/ICCIT64611.2024.11021794>
- Teimoory, G. K., & Keyvanpour, M. R. (2025). An Explainable Ai Model for Diabetes Prediction Using Random Forest. 2025 11th International Conference on Web Research (ICWR), 264–269. <https://doi.org/10.1109/ICWR65219.2025.11006200>
- Thangada, M. S., & Kasoju, R. (2024). A systematic review of suicide risk management strategies in primary care settings. *Frontiers in Psychiatry*, 15. <https://doi.org/10.3389/fpsyt.2024.1440738>
- Thomas, Y. T., Swann, A. C., Murtaza, S., Kosten, T. R., Murphy, N., Mathew, S. J., Verrico, C. D., Davis, J., & Moukaddam, N. (2025). Understanding traumatic injuries and suicide risk: A conceptual exploration for the emergency department. *Journal of Affective Disorders*, 391, 119977. <https://doi.org/10.1016/j.jad.2025.119977>
- Thompson, C., Sieffien, W., Iwajomo, T., Mason, J., de Oliveira, C., Rudoler, D., Husain, I., Kurdyak, P., & Zaheer, J. (2025). Suicide following discharge from inpatient psychiatric care: A retrospective case control study. *Psychiatry Research*, 351, 116645. <https://doi.org/10.1016/j.psychres.2025.116645>
- Tsui, F. R., Shi, L., Ruiz, V., Ryan, N. D., Biernesser, C., Iyengar, S., Walsh, C. G., & Brent, D. A. (2021). Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. *JAMIA Open*, 4(1). <https://doi.org/10.1093/jamiaopen/ooab011>
- Tu, L., Liu, Y., Zhu, H., Zhang, Q., & Ji, X. (2024). Characteristics of drug overdose suicide attempts presenting to the psychiatric emergency department of Beijing Anding Hospital. *BMC Public Health*, 24(1), 1597. <https://doi.org/10.1186/s12889-024-19095-4>

- van der Vegt, A. H., Campbell, V. K., Webb, R., Venkatesh, B., Lane, P. J., Wilks, K., McPhail, S., Rice, M., Isaacs, T., Abdel-Hafez, A., Whebell, S., Irwin, A., Schnetler, R. J., Shetty, A., & Scott, I. A. (2026). A novel, standardised approach to balancing effectiveness, efficiency and utility of surveillance AI prediction models for hospitalised patients using sepsis prediction as an exemplar. *Journal of the American Medical Informatics Association*, 33(2), 304–315. <https://doi.org/10.1093/jamia/ocaf192>
- Vickers, A. J., & Holland, F. (2021). Decision curve analysis to evaluate the clinical benefit of prediction models. *The Spine Journal*, 21(10), 1643–1648. <https://doi.org/10.1016/j.spinee.2021.02.024>
- Walsh, C. G., Ripperger, M. A., Novak, L., Reale, C., Anders, S., Spann, A., Kolli, J., Robinson, K., Chen, Q., Isaacs, D., Acosta, L. M. Y., Phibbs, F., Fielstein, E., Wilimitis, D., Musacchio Schafer, K., Hilton, R., Albert, D., Shelton, J., Stroh, J., ... Johnson, K. B. (2025). Risk Model–Guided Clinical Decision Support for Suicide Screening. *JAMA Network Open*, 8(1), e2452371. <https://doi.org/10.1001/jamanetworkopen.2024.52371>
- Wang, H., Yuan, H., Zhang, Y., Wang, Q., Gao, Z., & Zhao, M. (2025). Suicide risk prediction for Korean adolescents based on machine learning. *Scientific Reports*, 15(1), 14921. <https://doi.org/10.1038/s41598-025-99626-0>
- Wang, W., Cao, Y., Chen, F., Qiu, J., Zhang, X., Huang, Z., Wu, Y., Ji, X., & Wu, H. (2026). Clinical profiles and demographics of suicide attempts in early-to-mid adolescents in emergency departments: A retrospective study. *The American Journal of Emergency Medicine*, 101, 54–60. <https://doi.org/10.1016/j.ajem.2025.12.019>
- Wang, Y., Chen, S., Liu, J., Zhang, B., Zhu, Z., Zou, X., Zhou, Y., & Niu, B. (2025). Unveiling sex difference in factors associated with suicide attempt among Chinese adolescents with depression: a machine learning-based study. *Journal of Mental Health*, 34(4), 409–419. <https://doi.org/10.1080/09638237.2025.2478374>
- Wynants, L., van Smeden, M., McLernon, D. J., Timmerman, D., Steyerberg, E. W., & Van Calster, B. (2019). Three myths about risk thresholds for prediction models. *BMC Medicine*, 17(1), 192. <https://doi.org/10.1186/s12916-019-1425-3>
- Xu, W., Su, C., Li, Y., Rogers, S., Wang, F., Chen, K., & Aseltine, R. (2022). Improving suicide risk prediction via targeted data fusion: proof of concept using medical claims data. *Journal of the American Medical Informatics Association*, 29(3), 500–511. <https://doi.org/10.1093/jamia/ocab209>
- Yu, B., Park, C. G., Lee, K., & Jo, Y. Y. (2025). Utility of the Shock Index as a Prognostic Predictor in Patients Undergoing Emergency Surgery for Trauma: A Single Center, Retrospective Study. *Journal of Clinical Medicine*, 14(19), 6783. <https://doi.org/10.3390/jcm14196783>
- Zheng, S., Zeng, W., Xin, Q., Ye, Y., Xue, X., Li, E., Liu, T., Yan, N., Chen, W., & Yin, H. (2022). Can cognition help predict suicide risk in patients with major depressive disorder? A machine learning study. *BMC Psychiatry*, 22(1), 580. <https://doi.org/10.1186/s12888-022-04223-4>
- Zheng, Y., & Liu, X. (2024). An Optimized Multilayer Perceptron Model for Disease Risk Prediction in Psychiatric and Psychological Nursing. *Proceedings of the 2024 5th International Symposium on Artificial Intelligence for Medicine Science*, 426–430. <https://doi.org/10.1145/3706890.3706964>
- Zhu, Z., Wei, N., Guo, J., Yue, C., Chen, C., Zhang, Z., Wu, S., Su, J., & Song, B. (2025). Predicting the risk of threatened abortion using machine learning methods: a comparative study. *BMC Pregnancy and Childbirth*, 25(1), 901. <https://doi.org/10.1186/s12884-025-08030-z>