Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28426

A Closed-Loop Framework for Dynamic, Interpretable, and Actionable Modeling of Youth Suicidal Ideation

Chengling Shi

Department of Pharmacy, Shanghai University of Medicine & Health Sciences, Shanghai, China
lune133@outlook.com

Abstract. An end-to-end closed-loop architecture is proposed to resolve systemic disconnections in digital suicide-risk pipelines by unifying dynamic data capture, calibrated short-horizon prediction, explainable attribution, and proportionate action. Centering within-person change over hours to days, the design targets the temporal mismatch that renders static assessments inadequate for adolescents and young adults. High-frequency Ecological Momentary Assessment (EMA) and passive signals are distilled into time-aware features (trends, volatility, persistence, baseline deviations, clinically motivated interactions), mapped to next-day risk, with subject-level validation and routine probability calibration, and translated via SHAP (Shapley Additive explanations) into dialogue-ready explanations that identify current drivers and buffers. Action policies couple JITAI(s) to active drivers with hysteresis to prevent alert oscillation, while guardrails embed fairness auditing, drift monitoring, and youth-appropriate governance. A staged protocol—feasibility, prospective generalization with calibration/fairness audits, micro-randomized trials of JITAI(s), and multi-site transportability—operationalizes evaluation. The goal is not to replace clinical judgment but to equip it with accurate, legible information at moments when trajectories can still change, so clinicians and adolescents can co-design feasible steps even under non-ideal conditions.

Keywords: Suicidal ideation, Ecological Momentary Assessment, Machine learning, Explainable AI, Psychological disorders

1. Introduction

Youth suicide remains a persistent global public health crisis, primarily because the ability to predict imminent risk is severely limited [1]. The dominant assessment paradigm relies on static, cross-sectional instruments that, while valuable for identifying long-term vulnerabilities [2], are fundamentally mismatched with the temporal dynamics of suicidal ideation (SI). SI is a highly dynamic state that fluctuates over hours or days, yet static tools only provide disconnected snapshots, inevitably missing the critical motion—such as rising instability or sudden shifts—that signals escalating risk [3-5]. This temporal mismatch between static assessment and dynamic risk represents a critical gap in care.

The ubiquity of smartphones and wearable devices presents a paradigm-shifting opportunity to create this 'film' of experience. Ecological Momentary Assessment (EMA) elicits brief, in-the-

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28426

moment self-reports on thoughts, feelings, and context, minimizing recall bias. Passive sensing can complement this with objective, continuous data streams on behavioral patterns like sleep fragmentation, physical activity rhythms, and social communication [6]. However, raw data alone do not constitute an alarm. They require distillation into meaningful features, models that can learn complex patterns, and explanations that make the model's reasoning visible.

This is where Machine Learning (ML) and Explainable AI (XAI) become indispensable. ML models, particularly tree-based ensembles, are adept at capturing the non-linear relationships and high-order interactions characteristic of psychological phenomena [7]. Yet, their "black box" nature is a significant barrier to clinical trust and adoption. A clinician cannot act on a risk score without understanding its basis [8]. Explainable AI (XAI), especially methods like SHAP (Shapley Additive explanations), bridges this chasm between prediction and understanding. By attributing a model's output to specific input features [9], XAI facilitates a dialogue about risk, transforming an algorithmic verdict into a negotiable, collaborative insight.

However, this framework's stance is fundamentally capability-oriented, a principle holding that the ultimate value of clinical technology lies not in passive prediction, but in actively creating the conditions for individuals to take feasible, self-regulatory actions, particularly under non-ideal circumstances. The ultimate value of technology lies not just in predicting who is at risk, but in creating conditions for action when environments are imperfect and human support is intermittent. The pipeline is designed as a mechanism for translating constraints into supports—reducing friction, offloading cognitive burden, and providing lightweight prompts commensurate with the day's risk drivers to help individuals restore self-regulation.

This paper outlines a closed-loop system that embodies this principle. It shifts the focus from asking "Who is at risk in general?" to answering "Why is this person at risk today, and what is a proportionate, enabling action?" In this architecture, high-frequency data informs calibrated risk probabilities; SHAP-based explanations articulate the "why"; and this insight empowers clinicians and young people to co-design the next step. New data re-enters the loop, sustaining a cycle of learning and support. The goal is not to replace clinical judgment, but to support it with timely, interpretable, and compassionate information.

The following sections review related literature to identify key gaps and detail the proposed five-layer architecture with its design principles. A hypothetical case study then illustrates the framework in action, followed by a staged evaluation protocol and a discussion of clinical, ethical, and implementation implications.

2. Related work: established insights and persistent gaps

2.1. The dynamic nature of suicidal ideation and its measurement with EMA

A growing body of literature has firmly established that SI and its proximal correlates are highly variable [10]. EMA protocols, which typically involve prompting participants 4-8 times per day via smartphone, have been central to documenting this variability. By sampling experience in situ, EMA captures rapid fluctuations in internal states with high ecological validity [11]. Convergent findings show that momentary elevations in negative affect, hopelessness, perceived burdensomeness, and thwarted belongingness—core constructs of the Interpersonal Theory of Suicide—often precede short-horizon increases in SI [12]. Sleep disturbance has likewise emerged as a critical and modifiable risk factor: both poorer subjective sleep quality and actigraphy-derived sleep fragmentation are associated with next-day elevations in ideation [4].

DOI: 10.54254/2755-2721/2025.LD28426

A crucial insight from this work is the primacy of within-person dynamics. For near-term forecasting, deviations from an individual's own baseline generally carry more predictive signal than comparisons to cross-sectional norms. This motivates models that learn person-specific risk and resilience signatures rather than treating absolute symptom levels as sufficient statistics. A key practical question is the iatrogenic risk of repeated assessment. While objective EMA data from monitored settings show no systematic harm and can even be validating for some participants, retrospective reports indicate a minority perceive adverse effects (e.g., ~22% reported mood worsening). These perceptions are more common in high-risk subgroups with higher baseline burdens of BPD traits or PTSD, suggesting heterogeneous tolerability that necessitates additional safeguards [13]. One plausible reconciliation is methodological and psychological. Retrospective memory biases can over-weight salient difficult moments relative to the full time series, and the heightened self-focus that makes EMA diagnostically useful may be misinterpreted by some participants as the cause of distress rather than a mirror of ongoing fluctuations. Taken together, the pattern suggests heterogeneity in tolerability: while population-level harms are not evident in controlled contexts, certain high-risk subgroups experience the assessment process as more taxing and therefore warrant additional safeguards (e.g., enhanced consent language, opt-out micro-pauses, just-in-time debriefs, and closer monitoring).

In sum, EMA—especially when coupled with models that encode within-person baselines and short-horizon temporal structure—offers a timely, individualized lens on near-term suicide risk. These momentary precursors align with the Interpersonal Theory of Suicide and argue for a dual commitment: leverage EMA's dynamic signal for precision assessment and intervention while building safeguards that accommodate variability in experience (e.g., enhanced consent language, optional micro-pauses, brief just-in-time debriefs, closer monitoring).

2.2. Advances in predictive modeling for mental health

The application of ML to suicide risk prediction has accelerated and, in many settings, now matches or exceeds traditional statistical baselines while better accommodating complex, nonlinear relations among predictors [14]. For structured, multimodal tabular inputs, tree-based ensembles such as Random Forests and Gradient Boosting Machines (e.g., XGBoost, LightGBM) remain strong defaults because they natively handle mixed data types, missingness, and higher-order interactions with minimal preprocessing. When risk depends on how signals unfold rather than on their levels alone, sequence models—Long Short-Term Memory networks and, increasingly, Transformers—can capture temporal motifs and dependencies that are invisible to static feature sets [5]. When the clinical question is explicitly "time to next event," survival analysis provides the right objective: from Cox models to Random Survival Forests and modern deep survival networks, these approaches model hazard or survival functions directly and support individualized risk trajectories over specified horizons [10].

Promising retrospective performance does not guarantee transportability. Models trained in one cohort often degrade when deployed in another because of dataset or domain shift, population heterogeneity, and differences in measurement practice. Accordingly, person-level validation is essential: leave-one-subject-out or grouped cross-validation avoids record-level leakage and yields more honest estimates of generalization. In early-stage, small-sample development, carefully specified k-fold or nested cross-validation helps stabilize selection and hyperparameter tuning while preserving subject independence [15]. Beyond discrimination, calibrated probabilities are non-negotiable in high-stakes use: a predicted 20% risk should materialize as an observed 20% event rate

DOI: 10.54254/2755-2721/2025.LD28426

within strata, which typically requires explicit calibration (e.g., isotonic or Platt scaling) evaluated on held-out subjects [7].

Finally, continuous data streams from digital behavioral health platforms and EMA pipelines make it feasible to train models that learn within-person baselines, update as new data arrive, and output short-horizon, well-calibrated risks linked to proportionate actions. In practice, a pragmatic stack often pairs tree-based ensembles for tabular context features with sequence or survival components for dynamics, all wrapped in a validation plan that enforces subject-level separation and routine recalibration. This combination addresses the signal's complexity while guarding against the translational pitfalls that have historically limited clinical impact.

2.3. The imperative of explainability in clinical AI

Alongside gains in predictive accuracy, transparency has become a prerequisite for clinical use. Among XAI tools, SHAP is widely adopted because it grounds explanation in cooperative game theory: each feature receives a Shapley value—the marginal contribution of that feature to the prediction—so that attributions sum to the model output (local accuracy), increase when a feature's true influence rises (consistency), and vanish for truly irrelevant inputs (missingness) [12]. This yields a unified attribution framework that supports both population-level audit and case-level reasoning without changing the underlying model.

At the global level, SHAP aggregates attributions across people to characterize the model's typical behavior. Summary and dependence plots move beyond simple bar charts by showing effect direction, variability, and interaction structure. In psychiatric applications, such views have surfaced clinically meaningful nonlinearities (e.g., thresholds and plateaus) and interactions (e.g., when the impact of stress depends on sleep or social connection), while also reminding us that features with low average importance can be decisive for specific individuals. This global picture enables alignment checks against domain knowledge and helps identify spurious proxies before deployment.

At the local level, SHAP decomposes a single prediction into supportive and opposing feature contributions for that person. Force or waterfall plots make the "composition" of risk legible: which elements of mood, sleep, interpersonal strain, prior self-harm, or functional impairment are currently pushing risk up or down, and by how much. In practice, these individualized explanations have revealed non-monotonic cut-points (e.g., risk rising only beyond a symptom intensity threshold), protective plateaus (e.g., very high functioning dampening risk), and interaction-driven flips (e.g., the combined effect of health status and socioeconomic adversity). Such patterns often align with clinical intuition and give clinicians and patients a common language for proportionate action—tightening sleep hygiene, scheduling social contact, or triggering just-in-time support—without over-interpreting a single scalar score.

Taken together, SHAP's theory-backed guarantees and dual global/local views make black-box models behave more like accountable instruments. When paired with calibrated probabilities and person-specific baselines, these explanations turn predictions into actionable, dialogue-ready intelligence that supports early intervention and individualized care in mental health contexts.

2.4. Gaps in the current landscape

Current shortcomings in data capture, modeling, and explainability are not isolated gaps but form an interconnected systemic failure chain; the present framework is explicitly designed to close that chain end-to-end. Despite steady advances in data capture, modeling, and explainability, most projects still cover only fragments of the full pipeline. Data collection initiatives, standalone

DOI: 10.54254/2755-2721/2025.LD28426

modeling efforts, and isolated dashboards remain common, whereas continuous, ethically grounded loops that connect observation, interpretation, and action are rare. This framework targets four missing links.

Feature engineering should be designed around within-person dynamics and short-horizon instability, not only long-window aggregates. Because suicidal ideation and its antecedents can shift over hours to days, features need to encode deviations from personal baselines, volatility, intraday rhythms, end-of-day summaries, and cumulative effects in which chronic distress is amplified by acute events. Time-aware transforms, sliding windows, and change-point or autocorrelation signals preserve temporal information, while digital phenotyping helps distinguish highly labile from stable individuals to support precise short-term prediction and stratified intervention.

Modeling protocols must elevate probability calibration and fairness auditing to first-class requirements. Risk estimates should be probabilistically trustworthy across subgroups and over time, using subject-level validation (e.g., grouped or leave-one-subject-out cross-validation), nested or repeated cross-validation for small or imbalanced datasets, and external testing where feasible to detect dataset shift. Pre-deployment guardrails—youth-appropriate, GDPR-compatible data agreements, clinician training, and transparent reporting of EMA adherence, attrition, and adverse events—are necessary to evaluate reactivity and safety with credibility [9].

An explanation layer should translate quantitative outputs into shared, clinically usable meaning rather than stopping at raw feature rankings. Tools such as SHAP can provide global views that surface dominant patterns, nonlinear thresholds, protective plateaus, and risky conjunctions, alongside local decompositions that show which factors are currently pushing an individual's risk up or down and by how much. Standardized, patient-friendly summaries that align with domain knowledge facilitate clinician–patient dialogue and connect explanations to concrete decisions.

Finally, predictions must be linked to proportionate, personalized action pathways co-designed with users and clinicians. Systems should detect risk lability and trigger low-latency, context-aware support (JITAIs); combine survival or treatment-effect modeling (e.g., DeepSurv-style approaches) with user-centered product design; and leverage digital behavioral health platforms for continuous, situated guidance. Designing for adherence—human-factors principles, minimal burden, optional micro-pauses, personalized feedback, and graded escalation—helps convert explained predictions into sustained, ethically grounded clinical benefit.

3. The proposed framework: design principles and a five-layer framework

The framework is anchored in four design principles that steer both the technical stack and the ethics of use. They reflect a simple stance: near-term risk is about change, explanations should serve people rather than impress them, predictions must lead to doable steps, and trust is earned by doing the hard methodological work.

3.1. Design principles

Time is the principal dimension. Near-term risk tends to live not in a static level but in patterns of change—recent trajectories, volatility, and departures from one's own baseline. The research therefore treats time as first-class: features and models capture intraday rhythms, end-of-day summaries, and cumulative effects in which chronic strain is amplified by acute stressors. The emphasis is within-person signal rather than cross-sectional averages, so that short-horizon predictions track how someone is moving, not merely where they are.

DOI: 10.54254/2755-2721/2025.LD28426

Interpretability—and respect for narrative—is a precondition, not an afterthought. In high-stakes care, transparency and dignity are non-negotiable. Explanations function as a scaffold for dialogue: they translate model inferences into everyday language, surface thresholds and interactions that clinicians recognize, and make space for the person's own account of what matters. Model-agnostic attribution (e.g., SHAP-style methods) is used deliberately as part of documentation, informed consent, and proportionate response, rather than as decorative analytics.

Closed-loop actionability is required for clinical usefulness, a risk score must be coupled to a feasible next step. This framework links risk estimates to proportionate, low-latency supports (e.g., just-in-time adaptive interventions) that are co-designed with users to be safe, realistic, and actionable even in non-ideal conditions.

Trustworthiness comes through rigor. Probabilities must mean what they say; systems must not amplify inequities; and participation must remain safe and livable. Concretely, the research uses subject-level validation with external testing where feasible, routine probability calibration across subgroups and time, and active bias monitoring and mitigation. The paper reports adherence, attrition, and adverse events transparently, and the research treat user-centric governance (e.g., youth-appropriate, GDPR-compatible agreements) and clinician training as part of the workflow itself, not as post-hoc patches.

3.2. The five-layer architecture

The framework operates as a continuous loop that supports real-time inference while learning from its own use. The five-layer architecture operates under cross-cutting guardrails—revisitable youth consent, data minimization and least-privilege access, fairness and calibration audits with drift monitoring, and versioned governance (model cards, shadow evaluation, rollback)—that ensure ethical, fair, and secure operation across the loop; detailed instantiation appears in Section 3.2.6. Each layer has a clear purpose, inputs, and outputs so the pipeline stays auditable and clinically usable end to end. The integrated architecture is summarized below.

3.2.1. Layer 1: high-frequency data acquisition

This layer specifies sampling cadence, instrument selection, and privacy-preserving preprocessing rather than restating inputs/outputs. This research assembles a time-stamped "film" of each person's recent life over a focused monitoring window (typically 2–4 weeks). Brief EMA prompts (about 3–4 per day) sample mood (valence, arousal), stress, hopelessness, loneliness, perceived burdensomeness, rumination, self-efficacy, sleep quality, and salient interpersonal events; where appropriate and with safety protocols, SI is queried with clear distinctions between passive ("wish I were dead") and active ("plan to kill myself") forms. With consent, passive signals contribute objective context: sleep efficiency/duration/fragmentation from wearables; activity rhythms and mobility from accelerometers and GPS; and content-free communication metadata (e.g., call/text frequency and reciprocity) as a proxy for social contact. Cadence is tuned to balance sensitivity with burden, with adaptive sampling available (e.g., temporarily increasing prompts when passive sensors detect instability). Within this framework, EMA prompts are delivered 3–4 times per day; raw inputs are preprocessed on-device under least-privilege encryption, and consent/privacy are managed through an age-appropriate, revisit-enabled protocol.

DOI: 10.54254/2755-2721/2025.LD28426

3.2.2. Layer 2: feature engineering for dynamic risk

This layer formalizes time-aware transforms and windowing choices and justify interaction terms with clinical hypotheses. Core families include trend (e.g., exponentially weighted moving averages; short rolling slopes over 3–7 days), volatility (rolling variance/SD for mood or stress), persistence (lag-1 autocorrelation as a marker of "critical slowing down"), threshold exceedance (counts of moments above clinically meaningful cut-points, indexing cumulative allostatic load), baseline deviations (current vs. personal median/mean, such as last night's sleep vs. typical), and clinically motivated interactions (e.g., stress × sleep-shortfall). Transformations are explicitly time-aware (sliding windows, event-triggered features) to avoid treating time-sensitive signals as time-insensitive aggregates. An EWMA ($\alpha \approx 0.3$) over a 3–7-day window and a 3–5-day rolling standard deviation quantify trend and volatility, while threshold exceedance is defined dynamically from each participant's empirical percentiles.

3.2.3. Layer 3: short-horizon risk modeling and calibration

This layer details subject-level validation, probability calibration, and hysteresis policies that make short-horizon risks decision-ready. When transparency is paramount, EBM/GAM with learned pairwise terms are used; for performance at scale, gradient-boosted trees are adopted with post-hoc SHAP. Evaluation respects the person as the unit of generalization via grouped k-fold or leave-one-subject-out cross-validation to prevent leakage. Discrimination is summarized with AUC and, given class imbalance, AUPRC; calibration is treated as non-negotiable, with isotonic or Platt scaling fit on held-out subjects and monitored using calibration plots, Brier score, and expected calibration error. Operational metrics—daily alert precision, lead time (latency), and EMA adherence—are tracked for deployment readiness. To prevent alert "thrash," probabilities can be binned into risk states with hysteresis (separate on/off thresholds). Random Forest and XGBoost are trained under leave-one-subject-out cross-validation, probabilities are calibrated on an independent hold-out set via Platt scaling, performance is tracked with AUROC/AUPRC, Brier score, and ECE, and state transitions use hysteresis with on/off cutoffs at 0.10/0.15.

3.2.4. Layer 4: explanation and phenomenological translation

This layer focuses on explanation stability (over time and subgroups), ICE/ALE cross-checks under multicollinearity, and clinician-facing language templates. A probability is not a reason. Each day's prediction is decomposed into risk-increasing and risk-reducing contributions (e.g., SHAP values) and translates the quantitative pattern into ordinary language that is clinically resonant and culturally grounded. Explanations emphasize both drivers and buffers—e.g., "Recent sleep shortfalls appear to be magnifying today's dip in mood; steady contact with your cousin seems to be buffering some of that strain." The goal is dialogue, not verdicts: explanations are framed as negotiable accounts that clinicians and individuals can refine together, and global summaries (e.g., stability of feature effects, interaction hotspots) inform model governance and design. Local and global attributions are produced with SHAP (TreeExplainer), explanation stability is stress-tested with ICE/ALE diagnostics, and attributions are rendered into clinician-facing narrative templates.

3.2.5. Layer 5: from explanation to action—and back again

This layer specifies triggering logic, escalation pathways aligned to safety plans, and logging for learning and recalibration. Insights are coupled to proportionate, personalized steps. For low-to-

DOI: 10.54254/2755-2721/2025.LD28426

moderate risk, just-in-time adaptive interventions provide in-the-moment skills matched to that day's drivers (e.g., short grounding for rumination spikes). Social prompts nudge contact with pre-identified supports. Clinicians receive succinct one-page briefs highlighting current drivers/buffers and suggested focus areas to ease workflow integration. For persistent or high-risk states, predefined escalation pathways follow the individual's safety plan to connect with higher levels of care. When supportive others are unavailable, "just-right" scaffolds—micro-routines, micro-practice cards, contextual reminders—reduce friction and keep momentum without overreach. All actions, together with adherence and subsequent EMA observations—are captured in a single structured schema that supports continuous policy learning and periodic recalibration.

3.2.6. Design guardrails across layers

Least-privilege access, on-device preprocessing where feasible, and revisitable, youth-appropriate consent keep participation livable; fairness and calibration audits (with transparent reporting of adherence, attrition, and adverse events) are treated as workflow, not post-hoc patches. The result is a loop that can act in real time while improving with use—without sacrificing clarity, dignity, or safety. Moreover, explanatory stability is routinely audited across time windows and subgroups and cross-checked with ICE/ALE to guard against spurious attributions under multicollinearity or input drift; out-of-domain inputs are gated. Governance includes versioned model cards, shadow evaluation before promotion, an auditable change log, and rollback procedures.

4. Hypothetical demonstration and integrated evaluation

This section shows the framework in motion and outlines how it should be tested. The demonstration begins with a conversational risk trajectory that threads together time-aware features, calibrated prediction, and explanation. A staged evaluation protocol is specified that maps cleanly onto the framework's layers and design principles.

4.1. Demonstration case: 28-day risk trajectory

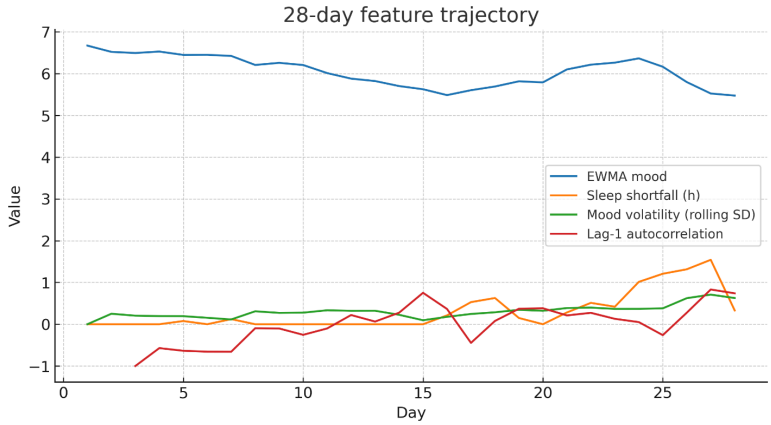

Figure 1. 28-day within-person feature trajectories. EWMA mood (3–7-day window), sleep shortfall, rolling volatility, and lag-1 autocorrelation over a 28-day window. (see Section 3.2.2)

Over a 28-day monitoring window, Layer 1 assembles brief EMA and passive signals; Layer 2 converts them into time-aware, within-person features (e.g., exponentially weighted trends, short rolling slopes, volatility, lag-1 autocorrelation as persistence, threshold exceedance, and clinically

DOI: 10.54254/2755-2721/2025.LD28426

motivated interactions such as stress × sleep-shortfall). During Days 1–21 the participant's mood and stress remain close to their personal baseline; trend and volatility are muted, persistence is low, and the Layer-3 model assigns a calibrated short-horizon risk below five percent. Figure 1 summarizes these within-person dynamics: mood exhibits a gentle downward EWMA drift, while volatility and lag-1 autocorrelation rise in the final week—consistent with reduced recovery speed after perturbations ("critical slowing down"). Sleep shortfall emerges mid-window and persists. Together, these time-aware features operationalize how the person is moving, not merely where they are, which is the signal that matters for short-horizon risk.

By Days 22–24, actigraphy shows fragmentation and shorter sleep; EWMA sleep dips, mood begins a downward slope, and calibrated risk edges toward eight percent. Calibrated next-day SI risk (Figure 2) remains low early on, then rises sharply as sleep shortfall and mood slope intensify. To avoid alert oscillation, this framework applies hysteresis bands (illustrative on/off thresholds at 0.10/0.15): actions trigger when risk enters the band and de-trigger only after risk returns below the lower bound. This preserves sensitivity while respecting clinical workload and user experience. A SHAP decomposition attributes most of this shift to sleep shortfall and the negative mood trajectory, with protective contributions from stable social contact and intact self-efficacy partially offsetting the rise.

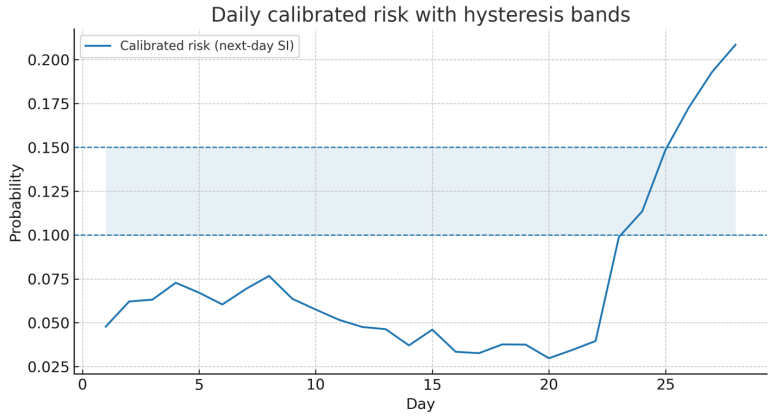

Figure 2. Calibrated next-day SI risk with hysteresis bands. Risk after isotonic/Platt calibration; dashed lines show example on/off thresholds (0.10/0.15). (Methods refs: Section 3.2.3/Section 3.2.6)

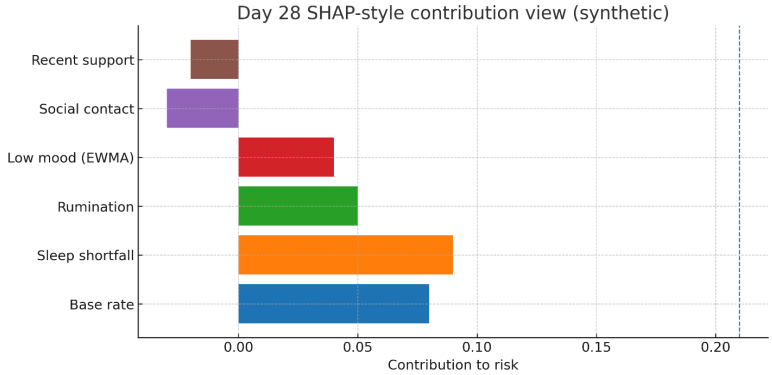

Figure 3. Local SHAP contribution view (Day 28). Risk-increasing vs. protective contributions relative to the calibrated base rate. (Methods refs: Section 3.2.4/Section 3.2.6)

DOI: 10.54254/2755-2721/2025.LD28426

An interpersonal conflict on Days 25–26 produces a marked stress spike and elevates the stress × sleep interaction feature, which becomes the leading positive SHAP term. Volatility in affect increases, lag-1 autocorrelation rises (a "critical slowing down" signal suggestive of slower recovery from perturbation), and risk climbs toward fifteen percent. On Day 27, risk peaks at approximately twenty-two percent. A local SHAP view (Figure 3) renders reasons in ordinary language: hopelessness (+0.06), the stress–sleep synergy (+0.05), and the ongoing mood decline (+0.04) dominate risk-increasing contributions, while recent check-ins with a cousin and an evening walk contribute small protective effects. For the participant, the message is that recent sleep disruption appears to amplify stress, and mood has slipped below the usual personal line; for the clinician, the system proposes proportionate next steps (stabilize sleep tonight; a brief grounding exercise now) without over-interpreting a single scalar score. With these steps in place, Day 28 shows partial rebound—sleep stabilizes, short-window slopes ease, volatility falls—and the calibrated risk decelerates toward twelve percent. Throughout, probabilities are monitored for calibration against held-out subjects; alerts employ hysteresis (separate on/off thresholds) to avoid oscillation; and global SHAP summaries slowly accumulate a phenotypic "map" that informs governance and product iteration in Layer 4. The example illustrates the core stance: time is first-class, within-person change is the signal, explanation scaffolds dialogue, and prediction is valuable because it connects to feasible action.

4.2. A staged protocol for evaluation

Validating a closed-loop, dynamic system requires an evaluation protocol that is itself longitudinal and multi-faceted. This framework therefore proposes a staged evaluation protocol, analogous to the phased clinical trial model, as a gold standard for assessing such digital mental health interventions. This protocol is designed to systematically build a body of evidence, moving from initial feasibility. Stage 1 establishes feasibility, acceptability, and adherence in a focused, four-week pilot (approximately 20–30 adolescents) using mixed methods. Quantitatively, the framework tracks EMA response rates with a target above eighty percent, average prompt completion time, and standardized burden/usability scales; qualitatively, semi-structured interviews and focus groups with young people and clinicians probe clarity, perceived intrusiveness, and whether explanations feel usable rather than judgmental. These data also stress-test consent language, safety protocols, and the balance between sampling sensitivity and participant burden.

Stage 2 examines predictive validity and calibration in a prospective cohort (approximately 150–200 participants followed for three to six months). Person-level generalization is enforced via grouped k-fold or leave-one-subject-out cross-validation to prevent leakage. Discrimination is summarized with AUROC and, given class imbalance, AUPRC; calibration is treated as non-negotiable using Brier score, reliability plots, and post-hoc isotonic or Platt scaling fit on held-out subjects. Clinical utility is quantified with decision-curve analysis across plausible thresholds to test whether using the model would improve decisions relative to treat-all or treat-none policies. Analysis plans are preregistered to reduce researcher degrees of freedom, and all results are stratified by key subgroups to surface fairness issues early.

Stage 3 tests actionability and proximal clinical impact using a micro-randomized trial. When the Layer 3 risk estimate crosses a personalized threshold (with hysteresis), eligible moments are randomized to deliver a just-in-time adaptive intervention versus a minimal self-monitoring prompt. The primary proximal endpoint is next-day suicidal-ideation severity or frequency; secondary endpoints include targeted drivers (e.g., sleep quality, affect regulation, rumination, self-efficacy). By repeatedly randomizing across time within individuals, the design estimates context-dependent

DOI: 10.54254/2755-2721/2025.LD28426

causal effects and identifies which interventions help whom, and under what conditions. Layer-5 logging ensures that actions, adherence, and outcomes can be linked to refine policies.

Stage 4 assesses transportability and monitors drift during multi-site implementation over longer horizons. Site-wise calibration audits verify reliability in new settings; subgroup analyses check for equitable performance; and statistical tests track dataset drift (input shifts) and concept drift (changes in input–outcome relations). Protocols for periodic recalibration or retraining are pre-specified, and governance remains youth-appropriate and GDPR-compatible. Adherence, attrition, and adverse events are reported transparently across stages, and clinician training is treated as part of the workflow rather than a post-hoc patch. The result is an evaluation path that matches the framework's commitments: probabilities mean what they say, explanations serve agency, and predictions lead to proportionate, feasible action.

5. Discussion

5.1. Principal contributions and clinical implications

This work positions the system not as an oracle but as a translator and gap-filler. Its core contribution is to convert shifting, real-world contexts into dignity-preserving supports that adolescents can actually use, especially when conditions are less than ideal. By pairing calibrated probabilities with explanations that read as reasons (rather than verdicts) and with proportionate, feasible options, the framework strengthens clinicians' judgment while cultivating young people's sense of agency. In practice, it enables a shift from reactive crisis management to proactive prevention: brief encounters become more targeted because the system highlights which drivers are active today, which buffers remain intact, and what next steps are realistic. For adolescents, everyday struggles are seen and met in time, in language that demystifies patterns and invites collaboration. In this sense, the technology is a scaffold: it reduces friction, reframes constraints as affordances, and keeps agency at the center of care.

5.2. Clinical integration and implementation

Successful adoption depends on fitting the grain of clinical work. This entails delivering concise, secure summaries integrated into existing EHR workflows; training clinicians to interpret and communicate probabilistic, explainable outputs carefully [16]; personalizing thresholds and notifications to manage alert fatigue; and co-designing recommendations with clinical teams to ensure actionable escalation pathways while balancing adherence, burden, and sensitivity.

5.3. Ethical considerations: fairness, privacy, and autonomy

Fairness is treated as a running obligation rather than a one-time check. Training data are diversified where possible; performance and calibration are audited across subgroups; and mitigation tactics (e.g., reweighting, group-specific calibration, threshold adjustments) are applied when disparities appear. Privacy and security rely on least-privilege access, encryption end-to-end, auditable role-based controls, and—where feasible—on-device feature computation to minimize exposure. Consent is voluntary, transparent, youth-appropriate, and revisitable; participants can set notification windows, pause participation, or request deletion. Above all, autonomy is preserved: explanations are framed as negotiable accounts that support shared decision-making, not as edicts. Safety protocols (e.g., predefined escalation consistent with a personal safety plan) are present but proportionate, so support remains available without becoming surveillance.

DOI: 10.54254/2755-2721/2025.LD28426

5.4. Limitations

This paper presents a blueprint rather than a deployed, validated system. EMA adherence can wane over time, and missing-data patterns may bias estimates; while engineered features capture useful dynamics, they remain simplifications that may miss context. Models will drift as life circumstances change, requiring continuous monitoring, recalibration, and external validation across settings. Finally, this paper focus is near-term ideation; the transition from ideation to attempt involves additional mechanisms—capability for suicide, access to means, impulsivity—that demand dedicated modeling and distinct safety considerations. Future work should therefore (i) test adherence-preserving designs (short prompts, adaptive sampling, user-set windows), (ii) embed subject-level validation, routine calibration, and fairness audits into operations, (iii) use micro-randomized trials to estimate which JITAI(s) help whom and when, and (iv) extend beyond ideation to model attempt-related mechanisms in partnership with clinical services. The goal is a loop that remains accurate, equitable, and livable as it learns.

6. Conclusion

Suicide risk is a moving target. This paper advances a dynamic, interpretable, closed-loop framework that treats time as first-class, centers within-person variation, and links calibrated short-horizon predictions to explanations that read as reasons and to proportionate next steps. The aim is not to replace judgment or relationships, but to equip both with timely, legible information at moments when trajectories can still change. Clinical usefulness follows from a pipeline that is accurate and livable (time-aware features, subject-level validation, routine calibration and fairness auditing, plain-language explanations, feasible actions integrated into EHR workflows). The path forward is empirical and pragmatic: staged evaluations, micro-randomized trials of JITAI(s), and multi-site transportability with drift monitoring to determine whether this approach improves proximal outcomes and, ultimately, reduces harm.

References

[1] Kivelä, L., Van Der Does, W. A. J., Riese, H., et al. (2022) Don’t miss the moment: a systematic review of ecological momentary assessment in suicide research. Frontiers in Digital Health, 4: 876595.
[2] Sedano-Capdevila, A., Porras-Segovia, A., Bello, H. J., et al. (2021) Use of ecological momentary assessment to study suicidal thoughts and behavior: a systematic review. Current Psychiatry Reports, 23(7): 41.
[3] Czyz, E. K., King, C. A., Al-Dajani, N., et al. (2023) Ecological momentary assessments and passive sensing in the prediction of short-term suicidal ideation in young adults. JAMA Network Open, 6(8): e2328005.
[4] Gowin, J. L., Stoddard, J., Doykos, T. K., et al. (2024) Sleep disturbance and subsequent suicidal behaviors in preadolescence. JAMA Network Open, 7(9): e2433734.
[5] Kampa, S., Raposa, M., Purcell, M., et al. (2025) Predicting risk of suicidal ideation with digital ecological momentary assessment of sleep, stress, mood, and pain: implications for tech-enabled suicide prevention. Journal of Technology in Behavioral Science.
[6] Baryshnikov, I., Rosenström, T., Isometsä, E. (2024) Predicting a short-term change of suicidal ideation in inpatients with depression: an ecological momentary assessment. Journal of Affective Disorders, 350: 1-6.
[7] Ballester, P. L., Cardoso, T. D. A., Moreira, F. P., et al. (2021) 5-year incidence of suicide-risk in youth: a gradient tree boosting and SHAP study. Journal of Affective Disorders, 295: 1049-1056.
[8] Lundberg, S. M., Erion, G., Chen, H., et al. (2020) From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1): 56-67.
[9] Niu, B., Wan, M., Zhou, Y. (2025) Development of an explainable machine learning model for predicting depression in adolescent girls with non-suicidal self-injury: a cross-sectional multicenter study. Journal of Affective Disorders, 379: 690-702.

DOI: 10.54254/2755-2721/2025.LD28426

[10] Katzman, J. L., Shaham, U., Cloninger, A., et al. (2018) DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Medical Research Methodology, 18(1): 24.
[11] Oquendo, M. A., Galfalvy, H. C., Choo, T. H., et al. (2024) Occurrence and characteristics of suicidal ideation in psychiatrically healthy individuals based on ecological momentary assessment. Molecular Psychiatry, 29(11): 3356-3363.
[12] Kivelä, L. M. M., Fried, E. I., Van Der Does, W., et al. (2024) Examining contemporaneous and temporal associations of real-time suicidal ideation using network analysis. Psychological Medicine, 54(12): 3357-3365.
[13] Kivelä, L. M. M., Fiß, F., Van Der Does, W., et al. (2024) Examination of acceptability, feasibility, and iatrogenic effects of ecological momentary assessment (EMA) of suicidal ideation. Assessment, 31(6): 1292-1308.
[14] Tang, H., Miri Rekavandi, A., Rooprai, D., et al. (2024) Analysis and evaluation of explainable artificial intelligence on suicide risk assessment. Scientific Reports, 14(1): 6163.
[15] Tougui, I., Jilbab, A., Mhamdi, J. E. (2021) Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. Healthcare Informatics Research, 27(3): 189-199.
[16] Bradshaw, T. J., Huemann, Z., Hu, J., et al. (2023) A guide to cross-validation for artificial intelligence in medical imaging. Radiology: Artificial Intelligence, 5(4): e220232.