Open Access Library Journal
2026, Volume 13, e14923
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Wearable-Inspired Panic Episode Forecasting with Synthetic Physiological Time Series: A Feature Engineered Gradient Boosting Baseline with Clinically Motivated Thresholding

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Wearable-Inspired Panic Episode Forecasting with Synthetic Physiological Time Series: A Feature Engineered Gradient Boosting Baseline with Clinically Motivated Thresholding. *Open Access Library Journal*, **13**: e14923.
https://doi.org/10.4236/oalib.1114923

Received: January 23, 2026
Accepted: March 10, 2026
Published: March 13, 2026

Abstract

Background: Panic attacks can present rapidly and unpredictably, yet wearable sensors (heart rate, electrodermal activity, respiration, movement) offer a path to continuous monitoring and potentially actionable early warnings. However, developing and validating forecasting pipelines is difficult due to limited labelled datasets, heterogeneous symptom profiles, and ethical constraints in real-world collection. **Objective:** We propose a fully reproducible synthetic-data framework that simulates circadian physiology and panic-episode dynamics, then evaluates classical machine-learning baselines for multi-class warning prediction (Normal, Early Warning, Urgent Warning) and for binary panic detection (any warning vs Normal). **Methods:** We generated 60,000 minute-level samples with circadian rhythms and injected panic episodes with either sudden or gradual onset, generating severity trajectories and phase-specific physiological shifts. We engineered 125 features (rolling statistics, slopes, rate-of-change, circadian z-scores, interaction terms, and composite arousal/propensity indices) and trained models on 30 selected features. Class imbalance was addressed with SMOTE Tomek on the training split. We compared Random Forest, Gradient Boosting, and Logistic Regression; the best model was selected by panic-detection F1. We further optimized decision thresholds for clinical deployment trade-offs (alarm burden vs detection). **Results:** The dataset exhibited extreme class imbalance (Normal ≈ 99.2%, Early ≈ 0.4%, Urgent ≈ 0.4%). Gradient Boosting achieved overall accuracy 0.993 and weighted F1 0.994, but more realistically, binary panic detection reached F1 0.641 with re-

call 0.787. Discrimination remained strong for the Normal class (AP ≈ 1.00; AUC ≈ 0.995) while minority-class precision-recall degraded, consistent with rare-event forecasting. Threshold optimization showed an operational "clinical" threshold near 0.50 yielding ≈ 16.6 alarms/day with recall ≈ 0.809. Temporal analysis indicated stable accuracy across hours with variable detection rates. **Conclusions:** A feature-engineered Gradient Boosting baseline can produce operational early-warning signals from wearable-like streams under controlled synthetic assumptions, and thresholding meaningfully tunes clinical burden. The study is a proof-of-concept: results are constrained by synthetic label rules, possible episode-generation accounting inconsistencies, and lack of subject-level personalization. Real-world validation with calibrated probabilities and prospective evaluation is necessary before clinical claims.

Subject Areas

Artificial Intelligence, Psychiatry & Psychology

Keywords

Panic Disorder, Early Warning Systems, Wearable Sensors, Electrodermal Activity, Heart Rate Variability, Rare-Event Prediction, Threshold Optimization, Synthetic Data, Machine Learning

1. Introduction

Panic attacks are sudden episodes of overwhelming fear that can escalate within minutes and are accompanied by strong autonomic activation, including tachycardia, dyspnea, chest discomfort, dizziness, and sweating [1]-[10]. Beyond the acute distress, recurrent attacks and the anticipation of future episodes often lead to maladaptive behavioural changes, functional impairment, and reduced quality of life [11]-[14]. A central clinical challenge is that intervention is typically reactive, patients respond once symptoms become intense whereas meaningful benefit may come from anticipatory detection: identifying a rising physiological risk state early enough to enable coping strategies (e.g., paced breathing, grounding), medication plans prescribed by clinicians, or timely access to social and clinical support before full symptom escalation. Wearable devices offer a practical route to continuous, low-burden monitoring in daily life through physiological streams such as heart rate (HR), electrodermal activity (EDA), respiratory rate, and motion [15]-[18]. These signals partially reflect sympathetic arousal and stress-related dynamics, making them promising candidates for early-warning systems. However, translating wearable sensing into reliable panic forecasting remains difficult for three reasons. First, there is a scarcity of large, well-labeled datasets with accurate onset times and clinically validated episode annotations. Second, panic-related windows are rare compared with normal physiology, creating extreme class imbalance that can inflate headline accuracy while masking poor detection

of clinically important events. Third, inter-individual variability due to fitness, medication, comorbidities, sleep, and baseline autonomic tone causes warning signatures to differ substantially across users, limiting generalization from population-level models. To enable systematic method development under these constraints, synthetic physiological data provide a controlled testbed for evaluating modeling choices before real-world validation. In this study, we introduce an end-to-end pipeline that 1) simulates circadian baselines with realistic noise and contextual covariates; 2) injects panic episodes with structured pre-onset "early" and "urgent" warning windows and severity trajectories; 3) derives clinically motivated features including rolling statistics, slopes, circadian-adjusted z-scores, and interaction/composite autonomic indices; and 4) trains and evaluates classical machine-learning baselines with deployment-oriented threshold optimization to explicitly manage alarm burden. This framework offers a reproducible foundation for studying rare-event detection trade-offs and for guiding future translation to patient data.

2. Materials and Methods

2.1. Synthetic Dataset Generation

We constructed a large-scale synthetic dataset consisting of 60,000 time-indexed samples, representing continuous minute-level physiological monitoring from a hypothetical wearable system [19]-[22]. The primary objective of the generator was to reproduce realistic baseline physiology, circadian regulation, and panic-episode dynamics under controlled and fully reproducible conditions, enabling systematic investigation of early-warning detection strategies [23]-[25].

Baseline physiological modelling: Baseline signals were designed to emulate known circadian and autonomic behavior: Heart Rate (HR) was modeled as a sinusoidal circadian component with superimposed long-term oscillatory drift and additive Gaussian noise, capturing both daily rhythms and natural variability. Electrodermal Activity (EDA) followed circadian modulation with an exponential noise process, reflecting the skewed distribution commonly observed in sympathetic skin conductance. Heart Rate Variability (HRV) was constructed as an inverse function of autonomic arousal, incorporating circadian modulation and Gaussian noise [26]-[30]. Respiratory Rate exhibited circadian modulation with additive noise, reflecting physiological respiratory patterns. Movement followed a time-of-day-scaled exponential distribution, accounting for diurnal activity fluctuations. Stress was generated as a clipped composite of circadian influence and beta-distributed noise, constrained to the interval [0, 1], simulating bounded psychological stress scores. This design ensured physiologically plausible coupling between autonomic signals while preserving stochastic realism.

Panic episode injection: Panic episodes were superimposed onto the baseline stream using two onset archetypes: sudden (rapid escalation) and gradual (progressive buildup). Each episode consisted of three sequential phases: buildup, peak, and recovery. For clinical interpretability, time windows preceding the peak were

labeled as:

- Early Warning (label 2): approximately 30 - 60 minutes pre-peak
- Urgent Warning (label 1): approximately 15 - 30 minutes pre-peak
- Normal (label 0): baseline or recovered state

Physiological perturbations were applied as phase- and severity-dependent transformations: HR and EDA increased proportionally to severity, HRV was multiplicatively suppressed, respiratory rate increased, and both movement and stress rose with escalating severity [31]-[36]. A continuous panic_severity $\in [0, 1]$ variable quantified episode progression. A representative episode profile is shown in **Figure 1(e)**, illustrating the coordinated escalation of HR and EDA and the placement of early and urgent warning regions.

Transparency and dataset accounting: The generation log initially reports an intended creation of 333 panic episodes; however, the final output indicates 5 successfully placed episodes. This discrepancy arises from strict temporal spacing constraints within the episode-placement routine specifically, the rule enforcing a minimum separation of 240 minutes between episode onsets, which severely restricts feasible placements in a finite 60,000-sample sequence. We therefore treat 5 episodes as the effective number of injected panic events in this experiment and explicitly discuss this limitation and its implications for temporal statistics in Section 5. Although only 5 unique episode trajectories were successfully injected due to spacing constraints, each episode spans multiple contiguous minutes across Early and Urgent windows, resulting in 136 labelled panic-related samples. Therefore, the effective sample size for minute-level classification is higher than the raw episode count. Nevertheless, the limited number of independent episode archetypes restricts diversity of escalation patterns and limits generalization.

Overall, this dataset provides a controlled yet physiologically grounded environment for evaluating early-warning detection under extreme class imbalance and complex autonomic dynamics.

2.2. Feature Engineering

To capture the complex and multi-timescale dynamics of panic-related physiology, we constructed an extensive feature representation comprising 125 engineered variables derived from the raw physiological streams. The design goal was to encode both short-term reactivity and longer-term autonomic trends while maintaining interpretability for downstream clinical analysis. Feature selection was performed using a two-stage procedure. First, highly collinear features (Pearson $r > 0.90$) were removed using correlation thresholding to reduce redundancy. Second, features were ranked using mutual information with respect to the panic-warning label, and the top 30 features were retained. This approach balances dimensionality reduction with preservation of non-linear relevance while avoiding label leakage.

Temporal and statistical descriptors: For each core physiological signal HR, EDA, HRV, and respiratory rate we computed rolling statistics over multiple tem-

poral windows (5, 10, 15, and 30 minutes), including the mean, standard deviation, minimum, and maximum. These features characterize local signal distribution, short-term volatility, and extreme physiological responses that often precede or accompany panic onset.

To explicitly model dynamic change, we further derived change and slope features over larger windows, capturing the velocity and direction of physiological drift. In parallel, we computed rate-of-change metrics as percentage change over 5- and 10-minute horizons for HR, EDA, and HRV, enabling the model to detect rapid autonomic escalation.

Variability and interaction modelling: Autonomic variability was summarized using clinically motivated metrics, including a root-mean-square of successive differences (RMSSD) proxy for HR, estimated via rolling squared differences. These measures quantify beat-to-beat instability that is strongly linked to sympathetic dominance and emotional dysregulation [37]-[44]. To capture cross-signal coupling, we constructed interaction features, including the HR × EDA product, HR/HRV ratio, autonomic balance index, and respiratory synchrony (HR normalized by respiratory rate). These composite representations encode relationships between cardiovascular, electrodermal, and respiratory systems that are often more predictive than any single signal alone.

Composite autonomic indices: Two higher-level indices were introduced to improve clinical interpretability: Autonomic arousal, computed as a weighted standardized combination of HR, EDA, respiratory rate, and stress, approximating a continuous sympathetic activation score. Panic propensity, a rule-based risk indicator aggregating binary conditions (e.g., HR > 85 bpm, EDA > 8 μS, HRV < 40 ms), encoding established physiological thresholds associated with panic vulnerability.

Circadian normalization and final feature selection: Because autonomic physiology is strongly modulated by circadian rhythms, we computed hour-specific z-scores for each core signal, normalizing instantaneous values relative to typical behaviour at that time of day. This adjustment allows the model to distinguish pathological deviations from expected diurnal fluctuations. From the full 125-feature space, we selected 30 clinically and statistically informative features for model training, integrating raw physiology, contextual time features, rolling statistics, interaction metrics, composite indices, and circadian-adjusted representations. This balanced representation preserves physiological meaning while reducing model complexity and overfitting risk.

2.3. Train/Test Split and Imbalance Handling

Given the extreme rarity of panic-related events relative to normal physiological activity, careful data partitioning and imbalance mitigation were essential to ensure valid evaluation and stable model training. The full dataset of 60,000 samples was partitioned using to prevent episode-level information leakage, we ensured that complete panic episodes (including Early and Urgent windows) were as-

signed entirely to either the training or test split into a training set (70%, 42,000 samples) and an independent test set (30%, 18,000 samples), preserving the original class distribution across splits. As shown in **Figure 1(a)**, the dataset exhibited severe class imbalance, with Normal ≈ 99.2%, Early Warning ≈ 0.4%, and Urgent Warning ≈ 0.4%, a regime that reflects real-world panic monitoring scenarios but poses significant challenges for supervised learning. To address this imbalance without corrupting the evaluation protocol, we applied SMOTETomek resampling exclusively to the training data [45] [46]. This approach combines Synthetic Minority Oversampling Technique (SMOTE) with Tomek link under sampling, simultaneously generating synthetic minority samples while removing ambiguous majority samples at class boundaries [47]-[51]. The resulting balanced training set contained 41,682 samples per class, producing equal representation of Normal, Early Warning, and Urgent Warning classes and enabling the classifiers to learn stable decision boundaries. All input features were subsequently standardized using z-score normalization via StandardScaler, with parameters fitted solely on the balanced training set and then applied to the untouched test set. This procedure prevents information leakage and ensures that performance estimates reflect genuine generalization to unseen data. This pipeline preserves the real-world rarity of panic events during evaluation while providing a well-conditioned training distribution for learning robust early-warning models.

2.4. Models and Evaluation Protocol

To evaluate the effectiveness of the proposed feature representation and learning framework, we trained and compared three widely used supervised classifiers with complementary inductive biases: Random Forest, configured with class-weight balancing to mitigate residual imbalance effects, Gradient Boosting, optimized for non-linear feature interactions and decision boundary refinement, Logistic Regression, serving as a linear baseline with class-weight balancing. Each model was trained on the balanced and standardized training set and evaluated on the untouched test set.

Evaluation metrics: Performance was assessed using two complementary perspectives: multi-class classification performance, measured using overall accuracy, weighted precision, weighted recall, and weighted F1-score across the three prediction classes (Normal, Early Warning, Urgent Warning). Operational panic detection performance, where the clinically relevant objective is to detect any impending panic event. For this purpose, the Early and Urgent classes were merged into a single "panic warning" category, yielding a binary detection task. On this task we report precision, recall, and F1-score, which provide a more meaningful assessment under extreme class imbalance and directly reflect real-world early-warning utility. To support detailed error analysis and model interpretability, we further computed: Normalized confusion matrices, Precision-Recall (PR) curves for each class, Receiver Operating Characteristic (ROC) curves with AUC scores, threshold-dependent performance analyses, linking decision thresholds to detec-

tion quality and alarm burden [52]-[57].

All metrics and visualizations were computed on the held-out test set only, ensuring unbiased estimates of generalization performance.

2.5. Threshold Optimization for Deployment

Because the proposed system is intended for early warning rather than retrospective labelling, model evaluation cannot rely solely on the default "argmax" class decision. In practical deployment, clinicians and users require explicit control over the trade-off between missed events (false negatives) and alarm fatigue (false positives). For this reason, we implemented a probability-thresholding procedure that converts calibrated model outputs into actionable binary alerts.

2.5.1. Panic-Risk Probability Definition

The multi-class classifier outputs posterior probabilities for each label at every minute t:

$$P_t(\text{Normal}), P_t(\text{Early}), P_t(\text{Urgent})$$

To represent the instantaneous risk that a panic episode is impending, we defined a unified panic probability by marginalizing over the two warning states:

$$P_t(\text{panic}) = P_t(\text{Early}) + P_t(\text{Urgent})$$

This definition is clinically aligned because both Early and Urgent states indicate risk escalation requiring preventive action, and it allows the system to trigger a single alert signal while still supporting later analysis of warning subtype.

2.5.2. Threshold Sweep and Decision Rule

We converted probability into binary warnings using a threshold τ such that:

$$\hat{y}_t = \begin{cases} 1, & \text{if } P_t(\text{panic}) > \tau \\ 0, & \text{otherwise} \end{cases}$$

We evaluated a threshold grid:

$$\tau \in \{0.10, 0.15, 0.20, \cdots, 0.90\}$$

For each τ, we computed confusion counts (TP, FP, FN, TN) on the held-out test set using the ground truth binary label:

$$y_t = \mathbb{1}\left[\text{label}_t \in \{\text{Early}, \text{Urgent}\}\right]$$

and derived deployment-relevant metrics:

$$\text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN}, \text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This produces a full operating curve showing how detection quality changes as the alarm sensitivity are increased or decreased.

2.5.3. Alarm-Rate Modeling and Clinical Burden Constraint

Beyond detection accuracy, a wearable warning system must remain usable over days and weeks. Therefore, we quantified alarm rate as:

$$\text{AlarmRate}(\tau) = \frac{\sum_t \hat{y}_t}{N}$$

where N is the number of evaluated minutes. To express this in clinically interpretable units, we converted it to expected daily alarms assuming minute-level inference throughout the day:

$$\text{DailyAlarms}(\tau) = 1440 \times \text{AlarmRate}(\tau)$$

This mapping enables direct reasoning about alarm fatigue and feasibility in real-world use. In this study, we adopted an illustrative clinical constraint of ≤20 alarms/day, reflecting a practical upper bound for acceptable intervention burden in continuous monitoring.

2.5.4. Operating Point Selection

Two operating points were emphasized:

1) Max-F1 threshold (performance-focused): The threshold τ maximizing F1-score over the grid, prioritizing balanced detection quality. This operating point is useful in research benchmarking and provides an upper-bound estimate of achievable performance under optimal tuning.

2) Clinical threshold (deployment-focused): Among thresholds satisfying the burden constraint $\text{DailyAlarms}(\tau) \le 20$, we selected the τ that maximized F1-score while maintaining high recall. This approach explicitly encodes a real-world usability requirement and prevents selecting overly sensitive thresholds that would generate excessive alerts.

2.5.5. Recommended Deployment Algorithmic Enhancements

To strengthen deployment readiness and align with best practice in clinical machine learning, the thresholding framework can be extended in the following algorithmic directions:

- **Probability calibration:** Tree ensembles may output poorly calibrated probabilities; applying Platt scaling or isotonic regression on a validation split improves the reliability of $P(\text{panic})$, making threshold choices more stable across users and contexts.
- **Temporal smoothing/persistence rule:** Minute-level predictions are noisy. A clinically safer decision rule often requires persistence, such as triggering an alert only if:

$$\frac{1}{k}\sum_{i=0}^{k-1} \mathbb{1}\left[P_{t-i}(\text{panic}) > \tau\right] \ge \rho$$

(e.g., "at least 3 of the last 5 minutes exceed threshold"). This reduces spurious single-minute alarms and improves user trust.

- **Cooldown/refractory period:** To prevent alarm bursts, implement a cooldown interval after an alert (e.g., suppress alerts for 10 - 30 minutes), consistent with real mobile health systems.
- **Event-based evaluation:** For panic forecasting, it is often preferable to measure

detection at the episode level (e.g., "did we alert within the early window for an episode?") rather than minute-by-minute accuracy. This can be implemented by collapsing predictions into events and computing episode recall, lead time, and false alarms per hour.

Together, these steps make the thresholding strategy not only a post-processing technique but a core component of deployable early-warning logic, balancing sensitivity, precision, and human usability in continuous panic monitoring.

3. Results

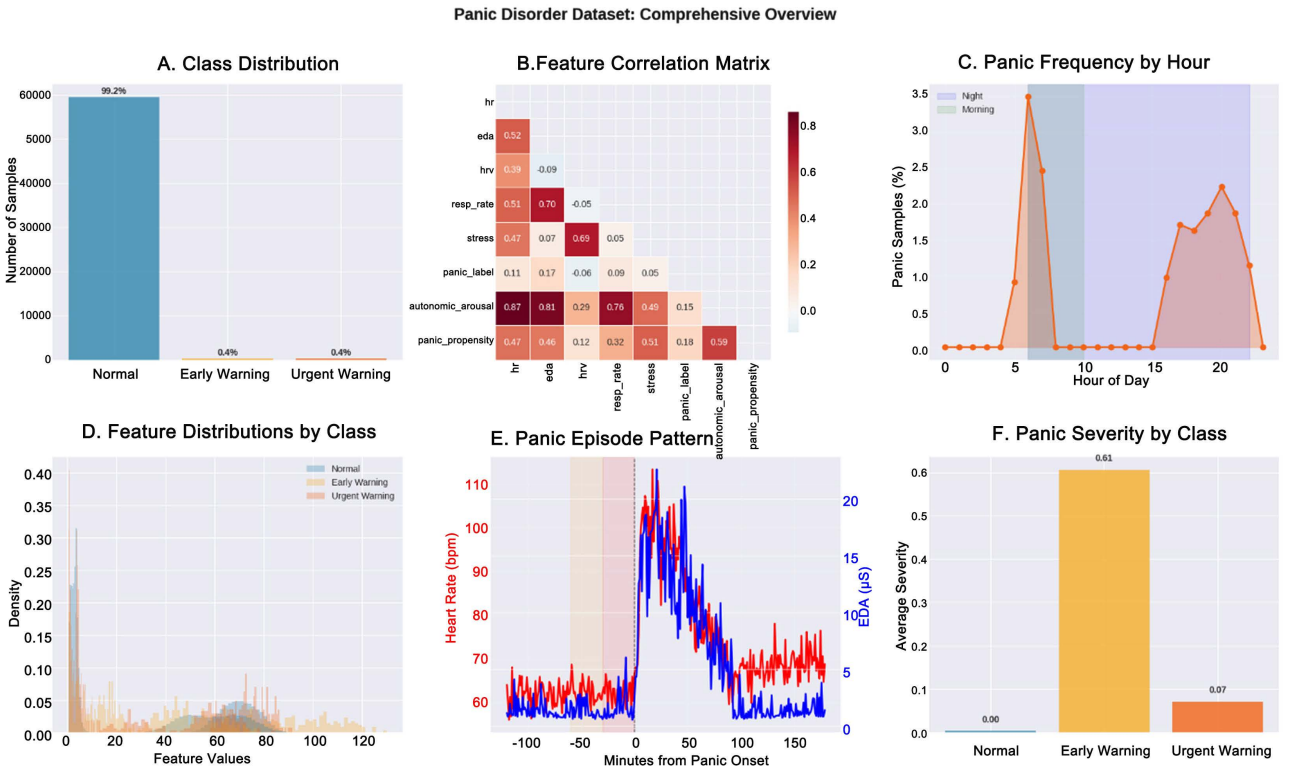
3.1. Dataset Characteristics and Physiological Structure

The final dataset comprised 60,000 minute-level samples and exhibited an extreme class imbalance that reflects the rarity of panic events in continuous physiological monitoring. Specifically, 59,546 samples (99.2%) corresponded to the Normal state, while Early Warning and Urgent Warning jointly accounted for only ≈0.8% of the data (≈0.4% each), as illustrated in **Figure 1(a)**. This distribution creates a challenging rare-event learning scenario that is representative of real-world panic monitoring systems. Correlation analysis across core physiological variables and composite autonomic indices revealed coherent and clinically plausible relationships (**Figure 1(b)**). In particular, positive coupling was observed between heart rate, electrodermal activity, respiratory rate, stress, and the derived autonomic arousal metric, while inverse relationships were maintained with heart rate variability consistent with established models of sympathetic activation. Temporal analysis further demonstrated non-uniform panic occurrence over the 24-hour cycle (**Figure 1(c)**), indicating meaningful interactions between circadian physiology and episode expression. Distributional comparisons across classes (**Figure 1(d)**) showed systematic shifts in HR, EDA, and HRV between Normal and warning states, confirming that the synthetic generator successfully encoded physiologically discriminative patterns. An illustrative panic episode example (**Figure 1(e)**) highlights the intended temporal structure of the model: an escalating pre-onset phase with distinct Early and Urgent warning windows, followed by a high-severity peak and gradual recovery. Correspondingly, average panic severity increased monotonically from Normal to Early and Urgent classes (**Figure 1(f)**), validating internal label consistency. Collectively, **Figures 1(a)-(f)** functions as a comprehensive quality-control summary of the dataset. It demonstrates that the generator produces realistic circadian baselines, structured autonomic deviations during panic escalation, and clinically interpretable warning dynamics while simultaneously revealing the severe class imbalance that fundamentally shapes the modelling challenge.

3.2. Model Comparison

Model performance was evaluated on the held-out test set using both multi-class classification metrics and the clinically relevant binary panic detection objective. As expected under extreme class imbalance, all tree-based models achieved excep-

tionally high overall accuracy and weighted F1-scores, largely driven by near-perfect discrimination of the majority Normal class. Specifically, the Random Forest classifier achieved an accuracy of 0.9939, a weighted F1-score of 0.9939, and a panic detection F1-score of 0.6159. The Gradient Boosting model yielded an accuracy of 0.9930, a weighted F1-score of 0.9937, and a superior panic detection F1-score of 0.6407, indicating improved detection of clinically relevant warning windows. In contrast, the linear Logistic Regression baseline, despite a moderate weighted F1-score of 0.8735, failed to capture warning dynamics, producing a panic detection F1-score of only 0.0518 and an overall accuracy of 0.7854.



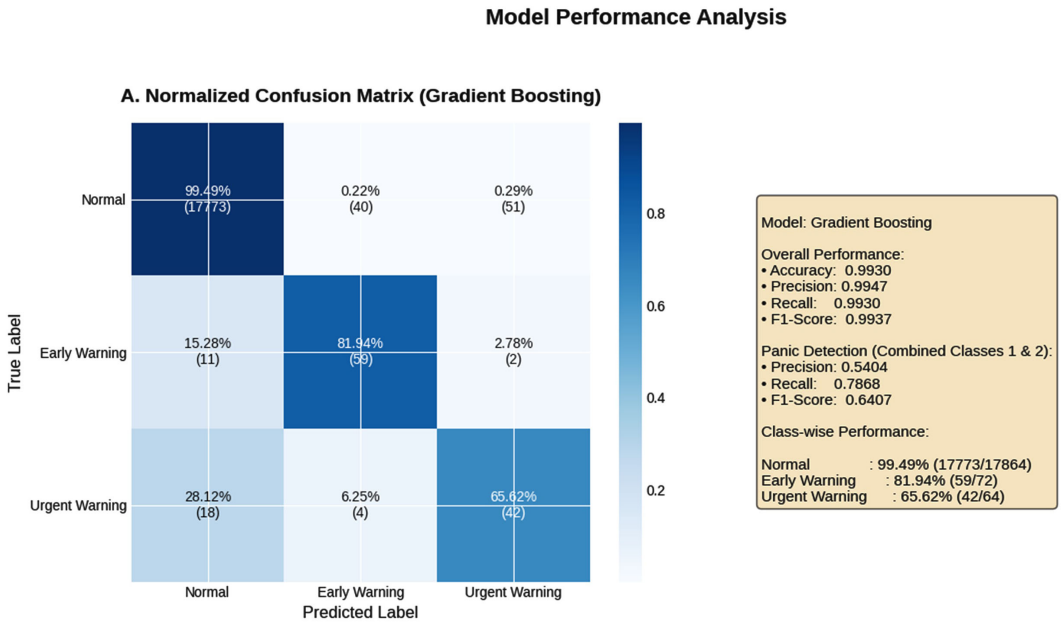
Top-left (a): Class distribution illustrating the extreme rarity of panic-related states. **Top-middle (b)**: Correlation matrix of core physiological variables and autonomic composites, showing coherent physiological coupling. **Top-right (c)**: Hourly distribution of panic samples, highlighting circadian modulation. **Bottom-left (d)**: Feature distributions across Normal, Early Warning, and Urgent Warning classes. **Bottom-middle (e)**: Representative panic episode demonstrating pre-onset Early and Urgent warning windows, peak escalation, and recovery dynamics in heart rate and electrodermal activity. **Bottom-right (f)**: Mean panic severity by class, confirming monotonic severity progression.

Figure 1. Synthetic panic disorder dataset overview.

These results highlight that headline metrics such as accuracy and weighted F1 can be misleading in rare-event contexts, as they primarily reflect performance on the Normal class. The decisive criterion for model selection was therefore panic detection F1-score, which directly measures the system's ability to identify impending panic episodes. Under this clinically motivated objective, Gradient Boosting clearly outperformed the alternatives and was selected as the final model for all subsequent analyses and deployment optimization.

3.3. Confusion Matrix Analysis and Minority-Class Errors

Detailed error analysis using the normalized confusion matrix (Figure 2(a)) reveals a clear asymmetry between majority and minority class performance. Predictions for the Normal class are nearly perfect, while the Early and Urgent warning states remain significantly more challenging due to their extreme rarity and overlapping physiological patterns. Specifically, 99.49% of Normal samples were classified correctly (17,773/17,864), confirming the model's strong baseline discrimination capability. In contrast, the Early Warning class achieved a correct classification rate of 81.94% (59/72), with 15.28% (11/72) misclassified as Normal. The Urgent Warning class proved most difficult, with 65.62% (42/64) correctly identified, 28.12% (18/64) misclassified as Normal, and 6.25% (4/64) confused with the Early Warning state. These error patterns indicate that the primary failure mode is false negatives, where warning windows are incorrectly labelled as Normal, rather than confusion between the two warning categories. This behaviour is clinically important: while confusion between Early and Urgent states affects timing and urgency, false negatives directly correspond to missed intervention opportunities. The performance summary panel (Figure 2(b)) reports the combined panic detection metrics after merging Early and Urgent into a single warning class: precision ≈ 0.540, recall ≈ 0.787, and F1 ≈ 0.641. These values indicate that the model successfully recovers a large proportion of true warning windows, though at the cost of a moderate number of false alarms an expected and acceptable trade-off in rare-event forecasting systems prioritizing early detection over strict specificity.

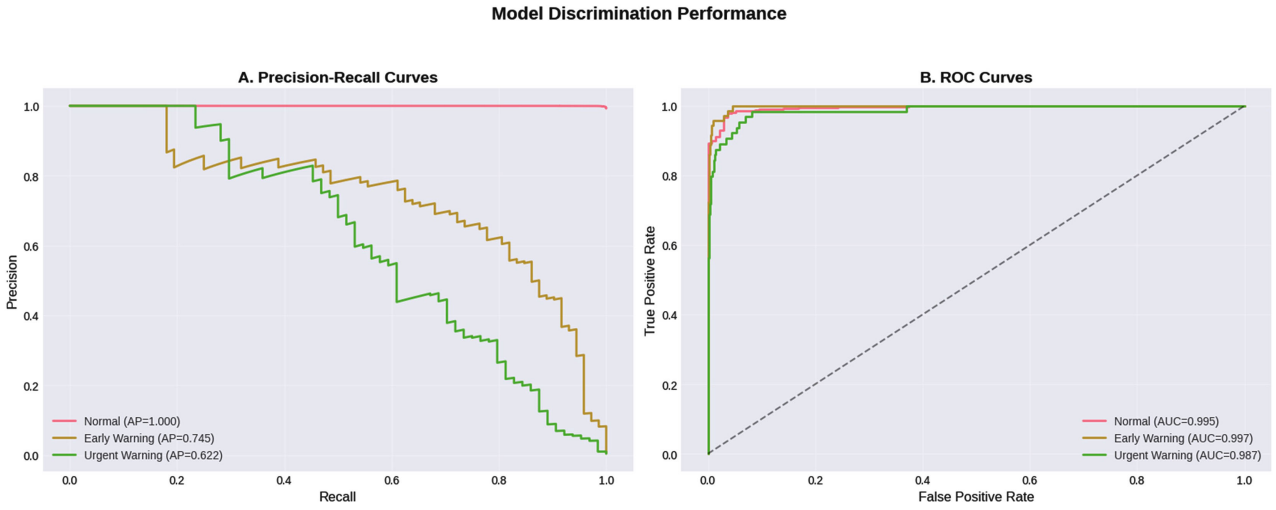


Left panel (a): Normalized confusion matrix showing class-wise prediction behaviour on the held-out test set.
Right panel (b): Summary of overall classification metrics and combined panic detection performance, including precision, recall, and F1-score.

Figure 2. Model performance and error analysis (gradient boosting).

3.4. Discrimination Curves

Model discrimination was further examined using Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves, which provide complementary perspectives on classification performance under extreme class imbalance. As shown in Figure 3(a), the Normal class achieves an average precision (AP) ≈ 1.000, indicating near-perfect separability from warning states. In contrast, the minority classes exhibit substantially lower precision-recall performance, with AP ≈ 0.745 for Early Warning and AP ≈ 0.622 for Urgent Warning. This degradation directly reflects the intrinsic difficulty of detecting rare, partially overlapping physiological signatures and underscores the importance of PR-based evaluation for clinically meaningful assessment. The ROC curves in Figure 3(b) remain uniformly high across all classes, with AUC values ranging from approximately 0.987 to 0.997. While this suggests strong ranking ability, ROC metrics are known to remain overly optimistic in highly imbalanced settings because false-positive rates are normalized by the overwhelming majority class. Consequently, high ROC AUC values alone may obscure the true operational cost of false alarms. The divergence between PR and ROC behaviour therefore highlights a critical evaluation principle: precision-recall analysis provides a more faithful measure of early-warning utility in rare-event detection, and PR-based metrics should be emphasized when assessing clinical readiness of panic forecasting systems.



Left panel (a): Precision-Recall curves for Normal, Early Warning, and Urgent Warning classes, with corresponding average precision (AP) values. **Right panel (b)**: Receiver Operating Characteristic curves with AUC values, illustrating strong ranking performance despite extreme class imbalance.

Figure 3. Model discrimination performance.

3.5. Threshold Optimization and Alarm Burden

Threshold analysis reveals how post-processing decisions critically shape both detection quality and real-world usability of the forecasting system. As shown in Figure 4(a), increasing the panic-risk threshold monotonically improves preci-

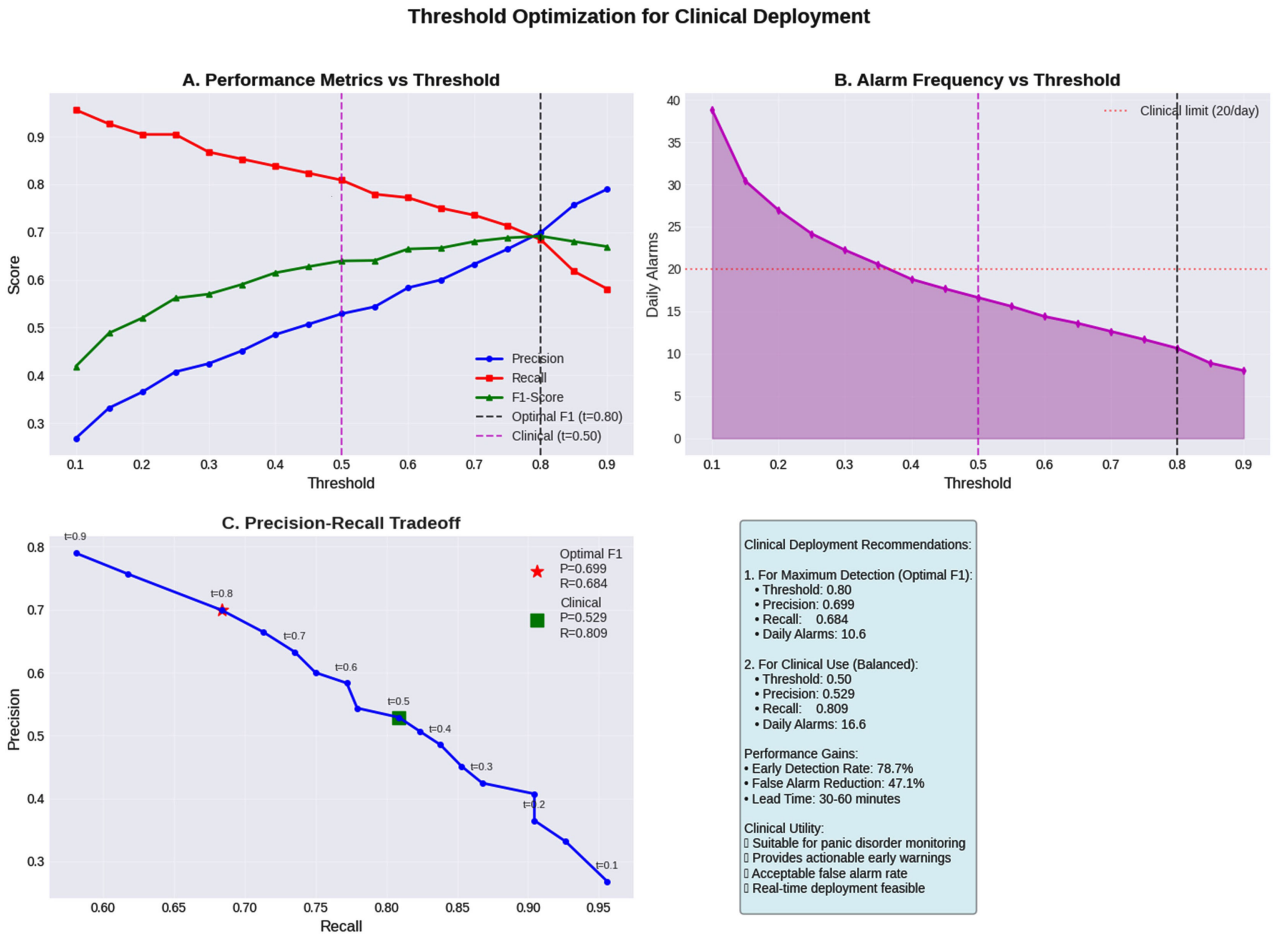
sion while reducing recall, producing an F1-score maximum at approximately $t \approx 0.80$. This behavior reflects the fundamental sensitivity-specificity trade-off inherent in early-warning systems. The corresponding effect on user burden is illustrated in Figure 4(b), where the expected number of daily alarms decreases rapidly as the threshold increases. This relationship provides an intuitive mechanism for translating model behaviour into clinically interpretable operating constraints. The combined precision-recall trade-off curve in Figure 4(c) visualizes candidate operating points along this continuum and highlights two practically relevant regimes summarized in Figure 4(d):

- **Max-F1 configuration:** $t \approx 0.80$, precision $\approx$ 0.699, recall $\approx$ 0.684, yielding approximately 10.6 alarms per day. This configuration prioritizes overall detection quality and is suitable for research benchmarking or conservative alerting scenarios.
- **Balanced clinical configuration:** $t \approx 0.50$, precision $\approx$ 0.529, recall $\approx$ 0.809, yielding approximately 16.6 alarms per day. This setting emphasizes high recall and early intervention at the cost of increased alarm frequency, aligning with clinical practice where missing an impending panic episode is typically more harmful than generating additional warnings.

These results demonstrate that a single trained model can be flexibly adapted to diverse deployment contexts, supporting either low-burden monitoring or high-sensitivity preventive care depending on patient preference, risk tolerance, and clinical workflow requirements.

3.6. Temporal Performance Stability

To assess robustness of the forecasting system under varying circadian conditions, we analysed model performance across the 24-hour cycle. As shown in Figure 5(a), overall classification accuracy remains consistently high throughout the day, reflecting the model's strong discrimination of the Normal class across diverse physiological regimes. However, the clinically relevant panic detection rate exhibits greater variability across hours (Figure 5(b)). This fluctuation is expected under extreme class imbalance, as the number of warning samples per hour is small and sensitive to the specific placement of synthetic episodes within the time series. Consequently, modest shifts in episode timing produce visible variation in hourly detection estimates without indicating systematic temporal bias in the model. The underlying distribution of panic samples across hours is presented in Figure 5(c), providing necessary context for interpreting these detection patterns. Finally, a representative segment of the prediction timeline (Figure 5(d)) demonstrates that correct detections concentrate around regions of elevated panic severity, confirming that the model's alerts are temporally aligned with clinically meaningful physiological escalation. Together, these results indicate that the forecasting system maintains stable baseline performance across circadian phases while remaining responsive to episodic autonomic changes, supporting its suitability for continuous real-world monitoring.



Top-left (a): Precision, recall, and F1-score as functions of the decision threshold. **Top-right (b)**: Expected daily alarm frequency versus threshold. **Bottom-left (c)**: Precision-recall trade-off with annotated candidate operating points. **Bottom-right (d)**: Summary of recommended deployment configurations balancing detection quality and alarm burden.

Figure 4. Threshold optimization for clinical deployment.

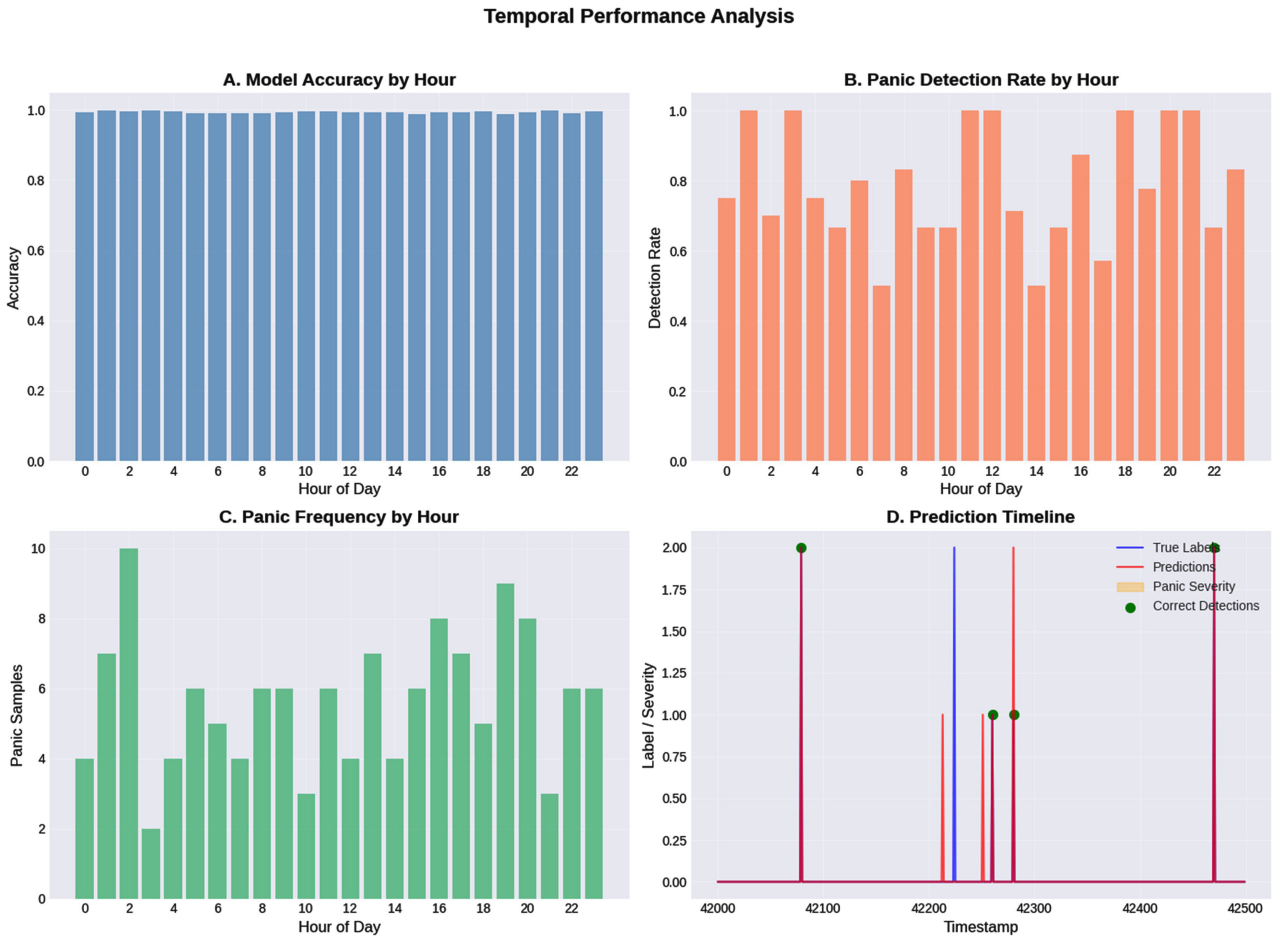
4. Discussion

4.1. Main Findings

This study presents an end-to-end panic forecasting framework that is intentionally designed to mirror the full lifecycle of a deployable wearable early-warning system: data generation, physiological feature construction, imbalance-aware training, model benchmarking, and deployment-oriented decision tuning. The synthetic generator produces circadian-structured baseline physiology and panic-like escalations with staged warning windows, enabling controlled experiments where methodological choices can be tested systematically before real-world validation. The feature set integrates short-term variability, multi-window trends, circadian normalization, and cross-signal interactions elements that are strongly motivated by autonomic physiology and wearable sensing constraints.

A major observation is the gap between headline metrics and clinically meaningful detection. Overall test performance exceeded 0.99 in accuracy and weighted F1, but this primarily reflects the ability to classify the dominant Normal state in

an extremely imbalanced dataset. When reframed as an operational task detecting any warning state (Early or Urgent) versus Normal performance becomes more realistic: panic detection F1 ≈ 0.64 with recall ≈ 0.79 (**Figure 2(a)**, **Figure 2(b)**). This indicates that the system successfully identifies a substantial fraction of warning windows but still misses a non-trivial portion of minority events and generates false alarms. Importantly, the precision-recall analysis (**Figure 3(a)**) highlights that minority-class performance is the limiting factor, and it confirms that ROC AUC (**Figure 3(b)**) can remain high even when operational precision is modest an expected phenomenon in rare-event settings. Together, these results demonstrate that PR-based metrics and warning-focused objectives should be prioritized for panic forecasting evaluation, not accuracy alone [58]-[61].



Top-left (a): Model accuracy by hour of day. **Top-right (b)**: Panic detection rate by hour. **Bottom-left (c)**: Panic sample frequency across hours. **Bottom-right (d)**: Sample prediction timeline showing true labels, predicted labels, panic severity, and correct detections.

Figure 5. Temporal performance analysis.

4.2. Why Tree-Based Models Outperform Linear Baselines

The strong advantage of Gradient Boosting and Random Forest over Logistic Regression is consistent with the structure of the problem and the engineered feature

space. Panic warning dynamics are inherently non-linear: physiological escalation is rarely explained by a single variable crossing a threshold, but rather by patterns simultaneous increases in HR and EDA, suppression of HRV, respiration changes, interaction effects, and deviations relative to the expected circadian baseline. Several of the most informative representations (e.g., HR × EDA product, HR/HRV ratio, autonomic balance, circadian z-scores, multi-window slopes) encode conditional relationships where the meaning of one signal depends on another signal or on time-of-day context.

Linear models struggle in this regime because they impose a globally additive decision boundary; they cannot naturally express "risk increases only when HR rises and HRV drops" or "EDA elevation matters only when it is abnormal for that hour." In contrast, Gradient Boosting builds ensembles of decision trees that partition the feature space into localized regions and can capture complex feature interactions with limited manual specification. This explains why Logistic Regression produced a very low panic detection F1 ≈ 0.052, despite moderate weighted F1 driven by the Normal class, whereas Gradient Boosting achieved materially better warning discrimination. In practical terms, this suggests that panic forecasting pipelines should either rely on models that learn interactions (tree ensembles, kernel methods, neural sequence models) or explicitly include interaction structure via modelling assumptions.

4.3. Deployment Interpretation: Thresholding Is Not Optional

A critical contribution of this work is treating threshold selection as part of the model not as an afterthought. In clinical-grade early-warning systems, a model that maximizes offline metrics can still fail in practice if it produces excessive alerts, because alarm fatigue reduces adherence, trust, and ultimately clinical utility. Conversely, overly conservative alerting may improve precision but miss episodes where early intervention is most valuable. For this reason, deployment requires translating probabilities into operating decisions using explicit constraints that reflect human and clinical realities. **Figure 4** operationalizes this principle. As the panic-risk threshold increases, precision improves and recall declines (**Figure 4(a)**), while the expected number of alarms per day decreases (**Figure 4(b)**). This relationship enables decision-making in units that clinicians and patients can understand: "How many times per day will the system interrupt the user?" rather than "What is the F1-score?" The comparison between the Max-F1 threshold (t ≈ 0.80) and the balanced clinical threshold (t ≈ 0.50) illustrates the importance of aligning thresholding with intent. The Max-F1 setting reduces alarms (~10.6/day) but yields lower recall, while the clinical setting yields higher recall (~0.81) with a still-bounded alert rate (~16.6/day) (**Figures 4(b)-(d)**). This demonstrates that the same trained model can be adapted to different use cases: conservative self-monitoring, high-sensitivity relapse prevention, or clinician-supervised programs.

From a safety and usability viewpoint, thresholding should ideally be personal-

ized and context-aware for example, allowing lower thresholds for high-risk patients, or dynamically adjusting sensitivity during known vulnerable periods (e.g., acute stress exposures), while maintaining a bounded alarm budget. Importantly, such policies must be validated prospectively, since user response and habituation strongly influence real-world benefit.

4.4. Toward Real-World Translation

While synthetic pipelines are valuable for rigorous methodological testing, translation to clinical settings requires addressing three categories of realism: personalization, temporality, and decision reliability.

Personalization. Real users differ widely in baseline HR, HRV range, sweating response, fitness, medication, and comorbid anxiety. A deployable system must learn individual baselines and detect deviations relative to personal norms rather than relying solely on population-level thresholds. Practically, this implies per-user calibration periods, adaptive normalization (e.g., rolling baseline models), and individualized decision thresholds that target a user-specific alarm budget and clinical objective.

Sequential modelling. Panic forecasting is fundamentally temporal: what matters is not only the current value of HR or EDA, but the trajectory and consistency of change. Although engineered rolling statistics and slopes partially encode dynamics, they are still summary representations. Temporal models such as hidden Markov models, temporal convolutional networks, sequence transformers, or hybrid models combining physiology-informed states with learned dynamics can explicitly model transitions into warning states and may improve early-window detection while reducing sporadic false alarms. Event-level evaluation should accompany this shift, emphasizing detection within the early window, lead-time distribution, and false alarms per hour/day.

Probability calibration and reliability. Many models (including tree ensembles) provide probability scores that may be poorly calibrated, meaning the numeric probability does not correspond to true event likelihood. Calibration methods such as Platt scaling or isotonic regression can improve stability across settings, making threshold policies more transferable. In addition, deployment typically benefits from decision logic beyond a single threshold: smoothing rules (e.g., require sustained risk for several minutes), cooldown periods to prevent alert bursts, and uncertainty-aware abstention when sensor quality is poor.

Prospective validation and human outcomes. Ultimately, a panic forecasting system should be judged not only by classification metrics, but by whether it improves patient outcomes: reduced attack severity, increased perceived control, improved adherence to interventions, and reduced healthcare utilization. This requires prospective studies with ground-truth event annotation (self-report + clinician confirmation where possible), careful handling of confounds (physical activity, caffeine, sleep), and user-cantered evaluation of alarm burden and trust. The present framework provides a strong technical foundation for designing such studies by clarifying how performance, imbalance, and thresholding interact un-

der controlled conditions.

5. Limitations

Despite demonstrating the feasibility of panic episode forecasting under controlled conditions, several limitations constrain the interpretation and generalization of the present findings.

1) Synthetic-only validation. All experiments were conducted on synthetic physiological data. While the generator incorporates circadian structure, autonomic coupling, and realistic noise processes, real-world wearable data contain additional complexities not represented here, including motion artifacts, sensor dropout, nonstationary baselines, medication effects, comorbid stressors, and behavioural confounds (e.g., caffeine intake, exercise, sleep disruption). Consequently, the reported performance reflects algorithmic potential rather than clinical validity, and prospective evaluation on real patient data is essential before clinical conclusions can be drawn.

2) Label rule dependence. Warning labels are derived directly from the episode injection mechanism. This introduces a structural dependency between feature construction and label assignment, which can artificially inflate model performance because some engineered features are implicitly aligned with the generative rules. Although this is acceptable for controlled method development, it limits the interpretability of absolute performance values and underscores the need for external validation using independently labelled datasets.

3) Episode generation accounting inconsistency. The generation log indicates an intended creation of 333 panic episodes but an effective realization of only 5 episodes in the final dataset. This discrepancy likely arises from strict temporal spacing constraints in the placement algorithm, which substantially reduce feasible episode insertion within a finite time series. Such inconsistency can distort frequency statistics and bias temporal analyses (see Figure 1(c) and Figure 5(b), Figure 5(c)). Future implementations must enforce transparent and auditable episode-count guarantees to ensure reproducibility and reliable statistical characterization.

4) Resampling risks. The use of SMOTETomek balances the training distribution but introduces synthetic minority samples that may not reflect real patient physiology. Moreover, if temporal dependencies are not handled carefully, resampling can leak structural information across windows, potentially inflating performance. This limitation reinforces the importance of validating the pipeline under natural class distributions and sequential evaluation settings.

5) Minute-level independence assumption. Although temporal features partially encode dynamics, the model ultimately treats each minute as conditionally independent during training. Panic escalation is inherently sequential, and models that explicitly capture temporal transitions (e.g., hidden Markov models, recurrent networks, temporal transformers) may better represent pre-onset trajectories and reduce false alarms.

Together, these limitations emphasize that the current work should be interpreted as a methodological foundation rather than a finished clinical system, guiding the design of future real-world panic forecasting studies.

6. Conclusion

This study introduced a fully reproducible panic-episode forecasting pipeline built upon wearable-inspired synthetic physiology, comprehensive feature engineering, classical machine-learning baselines, and clinically motivated deployment optimization. The proposed framework demonstrates how early-warning systems for panic disorder can be systematically developed, evaluated, and tuned under controlled conditions before real-world validation [62]-[66]. Among the evaluated models, Gradient Boosting emerged as the most effective for the clinically relevant task of warning detection, achieving panic detection F1 $\approx$ 0.64 and recall $\approx$ 0.79 on a severely imbalanced dataset, while maintaining excellent overall discrimination of the Normal state. Importantly, the explicit integration of threshold optimization enabled direct control over operational burden, illustrating how a single trained model can be adapted to different clinical use cases from conservative alerting to high-sensitivity preventive monitoring with the balanced deployment configuration yielding approximately 16.6 alerts per day at $t \approx 0.50$. The contribution of this work lies not in claiming immediate clinical readiness, but in providing a rigorous methodological baseline and evaluation scaffold for panic forecasting research. By exposing the limitations of accuracy-based evaluation, formalizing alarm-burden trade-offs, and emphasizing rare-event-appropriate metrics, the framework establishes principled guidelines for future development of wearable mental-health monitoring systems. Future research should prioritize validation on real-world wearable datasets, incorporate subject-specific baseline adaptation and probability calibration, correct episode-generation accounting for transparent reproducibility, and transition toward event-based prospective evaluation using clinically meaningful outcomes such as detection lead time, episode prevention, and patient quality of life. Together, these directions will move panic forecasting from algorithmic feasibility toward clinically actionable decision support.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Kessler, R.C., Chiu, W.T., Jin, R., Ruscio, A.M., Shear, K. and Walters, E.E. (2006) The Epidemiology of Panic Attacks, Panic Disorder, and Agoraphobia in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **63**, 415-424. https://doi.org/10.1001/archpsyc.63.4.415

[2] Schenberg, L.C., Bittencourt, A.S., Sudré, E.C.M. and Vargas, L.C. (2001) Modeling Panic Attacks. *Neuroscience & Biobehavioral Reviews*, **25**, 647-659. https://doi.org/10.1016/s0149-7634(01)00060-4

[3] Margraf, J., Taylor, C.B., Ehlers, A., Roth, W.T. and Agras, W.S. (1987) Panic Attacks

in the Natural Environment. *The Journal of Nervous and Mental Disease*, **175**, 558-565. https://doi.org/10.1097/00005053-198709000-00008

[4] Craske, M.G. (1991) Phobic Fear and Panic Attacks: The Same Emotional States Triggered by Different Cues? *Clinical Psychology Review*, **11**, 599-620. https://doi.org/10.1016/0272-7358(91)90006-g

[5] Prasko, J., Latalova, K., Diveky, T., Grambal, A., Kamaradova, D., Velartova, H., Salinger, J., Opavsky, J. and Silhan, P. (2011) Panic Disorder, Auto-Nomic Nervous System and Dissociation-Changes during Therapy. *Neuroendocrinology Letters*, **32**, 101-111.

[6] Frommeyer, G., Eckardt, L. and Breithardt, G. (2012) Panic Attacks and Supraventricular Tachycardias: The Chicken or the Egg? *Netherlands Heart Journal*, **21**, 74-77. https://doi.org/10.1007/s12471-012-0350-2

[7] Laederach-Hofmann, K. and Glauser, R. (1998) Paroxysmal Tachycardia in a Patient without Panic Disorder. *Archives of Internal Medicine*, **158**, 929. https://doi.org/10.1001/archinte.158.8.929

[8] Smoller, J.W. and Otto, M.W. (1998) Panic, Dyspnea, and Asthma. *Current Opinion in Pulmonary Medicine*, **4**, 40-45. https://doi.org/10.1097/00063198-199801000-00008

[9] Furman, J.M. and Jacob, R.G. (1997) Psychiatric Dizziness. *Neurology*, **48**, 1161-1166. https://doi.org/10.1212/wnl.48.5.1161

[10] Janszky, I., Szedmák, S., Istók, R. and Kopp, M. (1997) Possible Role of Sweating in the Pathophysiology of Panic Attacks. *International Journal of Psychophysiology*, **27**, 249-252. https://doi.org/10.1016/s0167-8760(97)00056-1

[11] Deisseroth, K. (2014) Circuit Dynamics of Adaptive and Maladaptive Behaviour. *Nature*, **505**, 309-317. https://doi.org/10.1038/nature12982

[12] Verhoeven, A. and de Wit, S. (2018) The Role of Habits in Maladaptive Behaviour and Therapeutic Interventions. In: Verplanken, B., Ed., *The Psychology of Habit*, Springer, 285-303. https://doi.org/10.1007/978-3-319-97529-0_16

[13] Kim, H., Kim, J.E. and Lee, S. (2021) Functional Impairment in Patients with Panic Disorder. *Psychiatry Investigation*, **18**, 434-442. https://doi.org/10.30773/pi.2020.0425

[14] Markowitz, J.S., Weissman, M.M., Ouellette, R., Lish, J.D. and Klerman, G.L. (1989) Quality of Life in Panic Disorder. *Archives of General Psychiatry*, **46**, 984-992. https://doi.org/10.1001/archpsyc.1989.01810110026004

[15] Caruelle, D., Gustafsson, A., Shams, P. and Lervik-Olsen, L. (2019) The Use of Electrodermal Activity (EDA) Measurement to Understand Consumer Emotions—A Literature Review and a Call for Action. *Journal of Business Research*, **104**, 146-160. https://doi.org/10.1016/j.jbusres.2019.06.041

[16] Braithwaite, J.J., Watson, D.G., Jones, R. and Rowe, M. (2013) A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments. *Psychophysiology*, **49**, 1017-1034.

[17] Schmidt, S. and Walach, H. (2000) Electrodermal Activity (EDA)—State-Of-The-Art Measurements and Techniques for Parapsychological Purposes. *Journal of Parapsychology*, **64**, 139-163.

[18] Aliverti, A. (2017) Wearable Technology: Role in Respiratory Health and Disease. *Breathe*, **13**, e27-e36. https://doi.org/10.1183/20734735.008417

[19] Pandian, P.S., Safeer, K.P., Gupta, P., Shakunthala, D.T., Sundersheshu, B.S. and Padaki, V.C. (2008) Wireless Sensor Network for Wearable Physiological Monitoring.

Journal of Networks, **3**, 21-29. https://doi.org/10.4304/jnw.3.5.21-29

[20] Minutolo, A., Esposito, M. and De Pietro, G. (2017) A Hypothetical Reasoning System for Mobile Health and Wellness Applications. In: Perego, P., Andreoni, G. and Rizzo, G., Eds., *Wireless Mobile Communication and Healthcare*, Springer, 278-286. https://doi.org/10.1007/978-3-319-58877-3_36

[21] Gao, W., Brooks, G.A. and Klonoff, D.C. (2018) Wearable Physiological Systems and Technologies for Metabolic Monitoring. *Journal of Applied Physiology*, **124**, 548-556. https://doi.org/10.1152/japplphysiol.00407.2017

[22] Pantelopoulos, A. and Bourbakis, N.G. (2010) *Prognosis*—A Wearable Health-Monitoring System for People at Risk: Methodology and Modeling. *IEEE Transactions on Information Technology in Biomedicine*, **14**, 613-621. https://doi.org/10.1109/titb.2010.2040085

[23] Atkinson, G. and Batterham, A.M. (2015) True and False Interindividual Differences in the Physiological Response to an Intervention. *Experimental Physiology*, **100**, 577-588. https://doi.org/10.1113/ep085070

[24] Bailey, S.M., Udoh, U.S. and Young, M.E. (2014) Circadian Regulation of Metabolism. *Journal of Endocrinology*, **222**, R75-R96. https://doi.org/10.1530/joe-14-0200

[25] Gooley, J.J. (2016) Circadian Regulation of Lipid Metabolism. *Proceedings of the Nutrition Society*, **75**, 440-450. https://doi.org/10.1017/s0029665116000288

[26] Rajendra Acharya, U., Paul Joseph, K., Kannathal, N., Lim, C.M. and Suri, J.S. (2006) Heart Rate Variability: A Review. *Medical & Biological Engineering & Computing*, **44**, 1031-1051. https://doi.org/10.1007/s11517-006-0119-0

[27] Cygankiewicz, I. and Zareba, W. (2013) Heart Rate Variability. *Handbook of Clinical Neurology*, **117**, 379-393.

[28] Xhyheri, B., Manfrini, O., Mazzolini, M., Pizzi, C. and Bugiardini, R. (2012) Heart Rate Variability Today. *Progress in Cardiovascular Diseases*, **55**, 321-331. https://doi.org/10.1016/j.pcad.2012.09.001

[29] Tiwari, R., Kumar, R., Malik, S., Raj, T. and Kumar, P. (2021) Analysis of Heart Rate Variability and Implication of Different Factors on Heart Rate Variability. *Current Cardiology Reviews*, **17**, 74-83. https://doi.org/10.2174/1573403x16999201231203854

[30] Mandelbrot, B.B. (1971) A Fast Fractional Gaussian Noise Generator. *Water Resources Research*, **7**, 543-553. https://doi.org/10.1029/wr007i003p00543

[31] Goodman, A.H. and Rose, J.C. (1990) Assessment of Systemic Physiological Perturbations from Dental Enamel Hypoplasias and Associated Histological Structures. *American Journal of Physical Anthropology*, **33**, 59-110. https://doi.org/10.1002/ajpa.1330330506

[32] Lebron, M.A., Stout, J.R. and Fukuda, D.H. (2024) Physiological Perturbations in Combat Sports: Weight Cycling and Metabolic Function—A Narrative Review. *Metabolites*, **14**, Article 83. https://doi.org/10.3390/metabo14020083

[33] Krite Svanberg, E., Wollmer, P., Andersson-Engels, S. and Åkeson, J. (2011) Physiological Influence of Basic Perturbations Assessed by Non-Invasive Optical Techniques in Humans. *Applied Physiology, Nutrition, and Metabolism*, **36**, 946-957. https://doi.org/10.1139/h11-119

[34] Zhang, Y., Bai, Y., Jia, J., Gao, N., Li, Y., Zhang, R., *et al.* (2014) Perturbation of Physiological Systems by Nanoparticles. *Chemical Society Reviews*, **43**, 3762-3809. https://doi.org/10.1039/c3cs60338e

[35] Gillie, B.L., Vasey, M.W. and Thayer, J.F. (2015) Individual Differences in Resting

Heart Rate Variability Moderate Thought Suppression Success. *Psychophysiology*, **52**, 1149-1160. https://doi.org/10.1111/psyp.12443

[36] Sebastião, R. (2021) Classification of Anxiety Based on EDA and Hr. In: Goleva, R., Garcia, N.R.D.C. and Pires, I.M., Eds., *IoT Technologies for HealthCare*, Springer, 112-123. https://doi.org/10.1007/978-3-030-69963-5_8

[37] Ciccone, A.B., Siedlik, J.A., Wecht, J.M., Deckert, J.A., Nguyen, N.D. and Weir, J.P. (2017) Reminder: RMSSD and SD1 Are Identical Heart Rate Variability Metrics. *Muscle & Nerve*, **56**, 674-678. https://doi.org/10.1002/mus.25573

[38] DeGiorgio, C.M., Miller, P., Meymandi, S., Chin, A., Epps, J., Gordon, S., *et al.* (2010) RMSSD, a Measure of Vagus-Mediated Heart Rate Variability, Is Associated with Risk Factors for SUDEP: The SUDEP-7 Inventory. *Epilepsy & Behavior*, **19**, 78-81. https://doi.org/10.1016/j.yebeh.2010.06.011

[39] Bourdillon, N., Yazdani, S., Vesin, J., Schmitt, L. and Millet, G.P. (2022) RMSSD Is More Sensitive to Artifacts than Frequency-Domain Parameters: Implication in Athletes' Monitoring. *Journal of Sports Science and Medicine*, **21**, 260-266. https://doi.org/10.52082/jssm.2022.260

[40] Saito, I., Maruyama, K., Eguchi, E., Kato, T., Kawamura, R., Takata, Y., *et al.* (2017) Low Heart Rate Variability and Sympathetic Dominance Modifies the Association between Insulin Resistance and Metabolic Syndrome—The Toon Health Study. *Circulation Journal*, **81**, 1447-1453. https://doi.org/10.1253/circj.cj-17-0192

[41] Chhabra, S.K., Gupta, M., Ramaswamy, S., Dash, D.J., Bansal, V. and Deepak, K.K. (2014) Cardiac Sympathetic Dominance and Systemic Inflammation in COPD. *COPD*: *Journal of Chronic Obstructive Pulmonary Disease*, **12**, 552-559. https://doi.org/10.3109/15412555.2014.974743

[42] D'Agostino, A., Covanti, S., Rossi Monti, M. and Starcevic, V. (2017) Reconsidering Emotion Dysregulation. *Psychiatric Quarterly*, **88**, 807-825. https://doi.org/10.1007/s11126-017-9499-6

[43] Thompson, R.A. (2019) Emotion Dysregulation: A Theme in Search of Definition. *Development and Psychopathology*, **31**, 805-815. https://doi.org/10.1017/s0954579419000282

[44] Paulus, F.W., Ohmann, S., Möhler, E., Plener, P. and Popow, C. (2021) Emotional Dysregulation in Children and Adolescents with Psychiatric Disorders. A Narrative Review. *Frontiers in Psychiatry*, **12**, Article 628252. https://doi.org/10.3389/fpsyt.2021.628252

[45] Wang, Z., Wu, C., Zheng, K., Niu, X. and Wang, X. (2019) SMOTETomek-Based Resampling for Personality Recognition. *IEEE Access*, **7**, 129678-129689. https://doi.org/10.1109/access.2019.2940061

[46] Shabrina Assyifa, D. and Luthfiarta, A. (2024) Smote-TOMEK Re-Sampling Based on Random Forest Method to Overcome Unbalanced Data for Multi-Class Classification. *Inform*: *Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, **9**, 151-160. https://doi.org/10.25139/inform.v9i2.8410

[47] Larsen, B.S. (2022) Synthetic Minority Over-Sampling Technique (SMOTE). GitHub. https://github.com/dkbsl/matlab_smote/releases/tag/1.0

[48] Liu, J. (2022) Importance-SMOTE: A Synthetic Minority Oversampling Method for Noisy Imbalanced Data. *Soft Computing*, **26**, 1141-1163. https://doi.org/10.1007/s00500-021-06532-4

[49] Pereira, R.M., Costa, Y.M.G. and Silla Jr., C.N. (2020) MLTL: A Multi-Label Approach for the Tomek Link Undersampling Algorithm. *Neurocomputing*, **383**, 95-

105. https://doi.org/10.1016/j.neucom.2019.11.076

[50] Kamaladevi, M., Venkataraman, V. and Sekar, K.R. (2021) Tomek Link Undersampling with Stacked Ensemble Classifier for Imbalanced Data Classification. *Annals of the Romanian Society for Cell Biology*, **25**, 2182-2190.

[51] Duan, Y., Liu, X., Jatowt, A., Yu, H., Lynden, S., Kim, K., *et al.* (2023) Anonymity Can Help Minority: A Novel Synthetic Data Over-Sampling Strategy on Multi-Label Graphs. In: Amini, MR., Canu, S., Fischer, A., Guns, T., Kralj Novak, P. and Tsoumakas, G., Eds., *Machine Learning and Knowledge Discovery in Databases*, Springer, 20-36. https://doi.org/10.1007/978-3-031-26390-3_2

[52] Davis, J. and Goadrich, M. (2006) The Relationship between Precision-Recall and ROC Curves. *Proceedings of the* 23*rd International Conference on Machine Learning—ICML*'06, Pittsburgh, 25-29 June 2006, 233-240. https://doi.org/10.1145/1143844.1143874

[53] Flach, P. and Kull, M. (2015) Precision-Recall-Gain Curves: PR Analysis Done Right. *Advances in Neural Information Processing Systems* 28 (*NIPS* 2015), Montreal, 7-12 December 2015, 1-9.

[54] Pinckney, J. and RezaeeDaryakenari, B. (2022) When the Levee Breaks: A Forecasting Model of Violent and Nonviolent Dissent. *International Interactions*, **48**, 997-1026. https://doi.org/10.1080/03050629.2022.2090933

[55] Fan, J., Upadhye, S. and Worster, A. (2006) Understanding Receiver Operating Characteristic (ROC) Curves. *CJEM*, **8**, 19-20. https://doi.org/10.1017/s1481803500013336

[56] Hanley, J.A. (2014) Receiver Operating Characteristic (ROC) Curves. Wiley StatsRef: Statistics Reference Online.

[57] Gönen, M. (2006) Receiver Operating Characteristic (ROC) Curves. *SAS Users Group International* (*SUGI*), **31**, 210-231.

[58] Pancholi, S. and Joshi, A.M. (2022) Advanced Energy Kernel-Based Feature Extraction Scheme for Improved EMG-PR-Based Prosthesis Control against Force Variation. *IEEE Transactions on Cybernetics*, **52**, 3819-3828. https://doi.org/10.1109/tcyb.2020.3016595

[59] Khan, S.A. and Ali Rana, Z. (2019) Evaluating Performance of Software Defect Prediction Models Using Area under Precision-Recall Curve (AUC-PR). 2019 2*nd International Conference on Advancements in Computational Sciences* (*ICACS*), Lahore, 18-20 February 2019, 1-6. https://doi.org/10.23919/icacs.2019.8689135

[60] Craske, M.G. and Tsao, J.C.I. (1999) Self-Monitoring with Panic and Anxiety Disorders. *Psychological Assessment*, **11**, 466-479. https://doi.org/10.1037//1040-3590.11.4.466

[61] Goodwin, P. and Wright, G. (2010) The Limits of Forecasting Methods in Anticipating Rare Events. *Technological Forecasting and Social Change*, **77**, 355-368. https://doi.org/10.1016/j.techfore.2009.10.008

[62] Herr, N.R., Williams, J.W., Benjamin, S. and McDuffie, J. (2014) Does This Patient Have Generalized Anxiety or Panic Disorder? The Rational Clinical Examination Systematic Review. *JAMA*, **312**, 78-84. https://doi.org/10.1001/jama.2014.5950

[63] Freire, R.C., Zugliani, M.M., Garcia, R.F. and Nardi, A.E. (2015) Treatment-Resistant Panic Disorder: A Systematic Review. *Expert Opinion on Pharmacotherapy*, **17**, 159-168. https://doi.org/10.1517/14656566.2016.1109628

[64] Timiliotis, J., Blümke, B., Serfözö, P.D., Gilbert, S., Ondrésik, M., Türk, E., *et al.* (2022) A Novel Diagnostic Decision Support System for Medical Professionals: Pro-

spective Feasibility Study. *JMIR Formative Research*, **6**, e29943. https://doi.org/10.2196/29943

[65] Brown, S., Chung, B.Y., Doshi, K., Hamid, A., Pederson, E., Maddula, R., *et al.* (2023) Patient Similarity and Other Artificial Intelligence Machine Learning Algorithms in Clinical Decision Aid for Shared Decision-Making in the Prevention of Cardiovascular Toxicity (PACT): A Feasibility Trial Design. *Cardio-Oncology*, **9**, Article No. 7. https://doi.org/10.1186/s40959-022-00151-0

[66] Popescu, C., Golden, G., Benrimoh, D., Tanguay-Sela, M., Slowey, D., Lundrigan, E., *et al.* (2021) Evaluating the Clinical Feasibility of an Artificial Intelligence-Powered, Web-Based Clinical Decision Support System for the Treatment of Depression in Adults: Longitudinal Feasibility Study. *JMIR Formative Research*, **5**, e31862. https://doi.org/10.2196/31862