ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

UDC 004.93:615.9:681.3

doi: 10.32620/aktt.2026.2.07

Yurii MYROSHNYK[1], Oleksandr LESHCHENKO[2]

[1] ***AVI-SPL, Inc., Tampa, USA***
[2] ***National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine***

HYBRID EARLY WARNING MODEL FOR AUTONOMOUS PHYSIOLOGICAL MONITORING WITH A MINIMAL SENSOR SET

This paper presents a validation of a hybrid model for remote physiological monitoring and patient deterioration prediction using the MIMIC-IV (51,981 patients) with external validation on MIMIC-III (30,528 patients, zero overlap). The ***objective*** *is to develop and substantiate a methodology for the quantitative assessment of prediction quality using four vital signs (HR, SpO_2, RR, Temp) and to evaluate a hybrid model combining a rule-based composite index with machine learning results: A(t)+ML. The* ***methods*** *employed comprise the recalibration of the A(t); analysis of 4 conditions (Full, NEWS2-set, CAS-4 and CAS-4 hybrid); supervised machine learning for classification and regression (XGBoost), the high-performance gradient boosting framework (LightGBM), logistic regression (LR), and recurrent neural network architecture (LSTM). Statistical tools include FDR Benjamini-Hochberg correction, SHAP, regression analysis for determining the VIF (Variance Inflation Factor); and a methodology for determining the Net Reclassification Improvement index (NRI).* ***Results.*** *Using quantitative assessments, an adequate level of prediction quality for the early detection of physiological deterioration was demonstrated when using only four vital sign parameters available from low-cost wearable sensors. It was also shown that the hybrid combination of an analytical risk index with machine learning can compensate for the absence of laboratory data and comprehensive electronic health records.* ***Conclusions.*** *The use of only four basic parameters combined with machine learning procedures for determining a patient's current physiological state of a patient provides clinically meaningful predictions for a broad class of remote monitoring systems, including those currently deployed.*

Keywords: *early warning system; MIMIC-IV; index A(t); aerospace medicine; edge computing; XGBoost; LSTM; NRI.*

1. Introduction

1.1. Problem Relevance

Early detection of physiological deterioration is one of the central challenges of modern medicine, encompassing both clinical intensive care and aerospace medicine. In intensive care units (ICUs), modern prediction systems utilizing comprehensive electronic health records (EHRs) achieve high discrimination performance: AUROC (the area under the receiver operating characteristic) is 0.85-0.96 for predicting in-hospital mortality [1-4]. These systems integrate dozens of input parameters: vital signs, laboratory analyses, prescribed medications, clinical notes, and demographic characteristics, requiring advanced digital infrastructure (continuous high-precision monitoring, laboratory services, hospital information systems, and reliable network connectivity).

At the same time, the number of individuals requiring continuous physiological monitoring outside hospital walls is growing. This group includes: elderly persons living independently who have an elevated risk of sudden deterioration; patients with chronic diseases (heart failure, chronic obstructive pulmonary disease (COPD), diabetes mellitus) who require early detection of exacerbations; military personnel recovering from combat injuries and concussions. In the aviation context, flight crew members are subject to particularly strict physiological monitoring due to specific risks: spatial disorientation, high-altitude hypoxia, effects of gravitational overloads (G-forces), and fatigue during prolonged flights. Ground-based monitoring systems for supporting pilot rehabilitation and operational control must function reliably in conditions where comprehensive medical infrastructure is unavailable.

This problem becomes particularly acute in the context of the ongoing Russian aggression against Ukraine, which has resulted in a significant number of Air Force service members requiring continuous physiological monitoring during rehabilitation from injuries, concussions, and post-traumatic conditions. Medical infrastructure in frontline regions has suffered substantial destruction, making inpatient monitoring

Creative Commons Attribution
NonCommercial 4.0 International

impossible or limited, and necessitating the need for autonomous portable systems.

Currently, in the consumer sector, health monitoring devices (smart watches Apple Watch, Fitbit, Garmin, , medical alert systems Bay Alarm Medical, Medical Guardian) are becoming widely adopted. These devices collect basic vital signs: heart rate, oxygen saturation, and some also collect respiratory rate and skin temperature. However, the level of their intelligence is limited to reactive functions: they can detect threshold exceedances or recognize individual events (e.g., atrial fibrillation or falls) but cannot predict deterioration before its clinical manifestation. The gap between hospital-grade prediction (AUROC > 0.85, but requiring analysis of 30-54 parameters) and edge deployment capabilities (4 parameters, but without predictive intelligence) remains critical and unfilled.

1.2. Previous Work and Its Limitations

In our previous work [5], we proposed the Care Alarm System (CAS) - an autonomous local platform integrating wearable sensors (pulse oximeter MAX30102, inertial module MPU6050, non-contact thermometer GY-906), stationary mmWave radars (MR60FDA2 for fall detection, MR60BHA2 for respiration and heartbeat) and a local Home Assistant server with InfluxDB/MySQL databases. In this system, risk is determined by three components: the instantaneous alarm index A(t) based on nonlinear aggregation of sensor deviations with clinically justified weights; the prognostic index P(t) with DTW similarity, Isolation Forest and time-to-event estimation; and a dual-channel LSTM model for deep temporal modeling. The system is designed for fully autonomous operation without cloud connectivity.

However, the implementation of CAS in remote physiological monitoring practice has a number of significant limitations. First, the models were validated on the PhysioNet BOLD database using a single patient example. The LSTM model was trained on synthetic sequences constructed by replicating single feature vectors, yielding a relatively low ROC-AUC = 0.6956. Second, no comparison with established clinical early warning scales (NEWS2, MEWS) was performed. Third, a fundamental question arises - how much predictive performance is lost in transitioning from full hospital monitoring to minimal 4-parameter edge deployment, and this question remained without a quantitative answer.

1.3. Research Questions

The results of this study are intended to address the identified limitations through rigorous validation on the MIMIC-IV dataset with external validation on the MIMIC-III (MIMIC - a popular, publicly available, and free dataset of electronic medical records (EMR) in raw format that has been used in numerous studies.). At the outset of the study, the following questions were formulated:

1. What degree of AUROC degradation occurs when transitioning from the full MIMIC feature set (50+ variables) to only four vital signs available from low-cost wearable sensors?

2. Does the hybrid approach combining the A(t) with ML provide better performance compared to pure ML on the same constrained 4-parameter input?

3. How does this hybrid approach compare with the standard clinical scale NEWS2, which uses more parameters but applies a different aggregation logic?

4. What minimal set of parameters provides clinically acceptable prediction quality for autonomous edge deployment in aerospace medicine?

2. Review of Existing Solutions

2.1. Clinical Early Warning Scales

The application of clinical early warning scales is a standard approach to detecting patient deterioration in hospital settings. The National Early Warning Score 2 (NEWS2), endorsed by the Royal College of Physicians of the United Kingdom in 2017, aggregates six physiological parameters (respiratory rate, oxygen saturation, systolic blood pressure, heart rate, level of consciousness, and body temperature) together with supplemental oxygen status into an integer scale from 0 to 20 points. Large-scale validation studies demonstrated the clinical scale's AUROC NEWS2 at 0.72-0.85 for predicting cardiac arrest, unplanned ICU admission, or death within 24 hours [6].

The Modified Early Warning Score (MEWS) uses a similar but smaller set of parameters (systolic pressure, HR, RR, temperature, and level of consciousness via AVPU). Both scales share a common design feature: they use simple integer scoring with fixed thresholds, forming a composite score through unweighted summation. While this simplicity facilitates bedside usability, it also limits the scale's sensitivity to subtle multivariate patterns and slow physiological trends.

Higher-level systems such as SOFA (Sequential Organ Failure Assessment), APACHE III (Acute Physiology and Chronic Health Evaluation) and SAPS II (Simplified Acute Physiology Score), allow incorporation of laboratory data (lactate, bilirubin, creatinine, platelets, arterial blood gases) and therefore achieve better discrimination in ICU populations, but they are fundamentally unsuitable for out-of-hospital deployment. The PhysioNet Computing in Cardiology Challenge challenge series stimulated the development of

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

early warning systems, particularly for sepsis prediction, but these developments are primarily oriented toward inpatient settings with full access to electronic health records (EHRs).

2.2. Machine Learning for ICU Outcome Prediction

Machine learning approaches to ICU mortality prediction have gained significant traction over the past decade, owing to the availability of large public databases such as MIMIC-III and MIMIC-IV. These approaches span a broad methodological spectrum.

Feature-based classifiers, notably XGBoost [7], have demonstrated high effectiveness. Studies using comprehensive MIMIC feature sets (demographics, vital signs, laboratory values, medications) report AUROC (0.79-0.85) for predicting in-hospital mortality [1-3]. Temporal models, including LSTM [8] and bidirectional LSTM architectures, capture sequential patterns in physiological time series and have shown improvements over static classifiers, especially for 12-48 hour prediction horizons. Recently, Bakumenko et al. [3] demonstrated transparent ICU mortality prediction using clinical Transformer models, achieving state-of-the-art results on MIMIC.

Ensemble and fusion strategies are also actively explored. Late fusion approaches combining LSTM for vital sign time series with Transformer models for clinical texts have shown that multimodal integration can improve AUROC from 0.753 (structured data only) to 0.918 (structured plus text data). This observation is directly relevant to the present study: it quantifies the magnitude of improvement achieved through data sources unavailable in edge deployment.

A critical observation from the literature is that virtually all high-performance models require 30-54+ input variables, including laboratory results that require invasive sample collection and centralized laboratory processing. The question of how these models degrade when constrained to a minimal vital sign set has received surprisingly little systematic attention in the literature.

2.3. Resource-Constrained Health Monitoring and edge Computing

A significantly smaller body of work is devoted to predicting physiological state using a limited set of parameters. Alghatani et al. [4] demonstrated that ICU mortality and length of stay can be predicted with acceptable accuracy (up to 84% using XGBoost with quantile feature engineering) based on only six vital signs (body temperature, HR, RR, systolic and diastolic BP, SpO_2), extracted from MIMIC-III, without any laboratory data or prior diagnoses. This represents the closest precedent to our work, although the authors did not investigate hybrid rule-ML architectures and used a broader parameter set than the proposed CAS-4.

In the commercial sector, devices such as Apple Watch provide fall detection and heart rhythm monitoring through built-in sensors, but their predictive capabilities are limited to detecting individual events (e.g., atrial fibrillation), rather than multivariate deterioration prediction. Medical alert systems such as Bay Alarm Medical and Medical Guardian rely entirely on user-initiated alerts or simple accelerometer-based fall detection with subsequent cloud data processing.

Millimeter-wave monitoring (mmWave radar) has emerged as an alternative for non-contact vital sign acquisition while preserving privacy. Recent studies [9] demonstrated accurate extraction of HR and RR from mmWave radar signals, particularly in multi-person monitoring scenarios. However, these works focus on signal processing accuracy rather than predictive modeling of clinical outcomes.

2.4. Hybrid Approaches: Combining Rules and Machine Learning

The combination of domain knowledge-driven rules with data-driven machine learning has been investigated in several clinical contexts. The Rothman Index, for example, integrates composite nursing assessment scores with predictive analytics, achieving an AUROC exceeding 0.90 for 24-hour mortality in mixed-acuity populations. However, it requires 26 variables, including laboratory values, making it unsuitable for autonomous edge deployment.

To the best of our knowledge, the specific architecture we propose - using an analytically derived risk index as an input feature or cascade filter for a machine learning model operating on a minimal vital sign set - has not been systematically investigated in the literature. This constitutes the main novelty of this work: quantitative assessment of whether introducing clinical domain knowledge through a structured analytical index can compensate for the absence of many features in resource-constrained monitoring scenarios.

3. Materials and Methods

3.1. Data Sources and Cohort Formation

3.1.1. Primary Dataset: MIMIC-IV

The primary dataset is Medical Information Mart for Intensive Care IV (MIMIC-IV, version 3.1) [10] - a publicly available database of de-identified medical records of patients admitted to Beth Israel Deaconess Medical Center (Boston, USA) during the period 2008-

2022. The database contains data from approximately 65,000 ICU stays across various specialties.

Cohort inclusion criteria were defined as follows: adult patients (age ≥ 18 years); first ICU stay of at least 24 hours duration (sufficient observation window); availability of all four target vital signs (HR, SpO_2, RR, temperature) recorded at least once within the first 24 hours. The primary outcome was in-hospital mortality (binary). Readmissions were excluded (only first ICU stay per patient) to prevent data leakage.

The final cohort comprised **51,981 ICU stays** with a mortality rate of 9.9% (5,172 cases), median age of 67 years (IQR 55-78) and 56.6% males. It should be noted that patients without all four vital signs in the first 24 hours were systematically excluded (systematic exclusion bias) - these patients may represent a distinct clinical subgroup (very short stays, palliative care, or monitoring gaps).

3.1.2. External Validation: MIMIC-III

For temporal external validation, MIMIC-III (version 1.4) [11] was used, covering a partially overlapping but earlier patient population (2001-2012) from the same institution. Identical inclusion criteria were applied. Final cohort: **30,528 stays**, mortality 10.7%, median age 64 (IQR 52-76), 57.7% males.

Patient overlap between the MIMIC-IV and MIMIC-III: **zero** (MIMIC-IV uses a re-identified system subject_id with no overlap with MIMIC-III), ensuring fully independent external validation. This design allows evaluating model generalization across different time periods and documentation practices, which is a more rigorous validation than random splitting from a single database.

Data split: 70% training, 15% validation, 15% test sets with stratification by outcome.

3.2. Experimental Conditions

To systematically assess the impact of feature availability on prediction quality, four experimental conditions were defined (Table 1), representing the spectrum from hospital to edge monitoring:

Condition «Full» includes all available MIMIC variables: vital signs (HR, SpO_2, RR, temperature, systolic and diastolic BP, mean BP, Glasgow Coma Scale), laboratory values, and demographic characteristics (age, sex). This condition establishes the upper bound of achievable performance and simulates full EHR access.

Condition «NEWS2-set» includes six NEWS2 parameters: RR, SpO_2, systolic BP, HR, temperature, and level of consciousness (converted from GCS to AVPU).

Table 1

Experimental Conditions

Condition	Parameters	N params	Deployment scenario
Full	All MIMIC variables: vital signs, labs, demographics	50+	Full hospital EHR
NEWS2-set	RR, SpO_2, systolic BP, HR, temperature, consciousness	6	Bedside nursing assessment
CAS-4	HR, SpO_2, RR, temperature	4	Edge/wearable sensors
CAS-4 hybrid	HR, SpO_2, RR, temperature + A(t)	4 + A(t)	Edge + domain knowledge

This condition simulates a standard bedside assessment performed by a nurse.

Condition «CAS-4» includes only four parameters: HR, SpO_2, RR, and temperature - the minimal set available from low-cost wearable sensors (MAX30102 for HR/SpO_2, mmWave radar for RR, GY-906 for temperature). This condition simulates the CAS system described in [5].

Condition «CAS-4 hybrid» augments CAS-4 with the computed alarm index A(t), carried over from [5], as an additional derived feature. Three hybrid architectures are tested.

3.3. Temporal Feature Extraction

For each vital sign parameter, 9 temporal statistical features are computed over the first 24 hours of ICU stay: mean, standard deviation, minimum, maximum, first recorded value, last recorded value, range (maximum minus minimum), coefficient of variation (standard deviation divided by mean), and linear trend (slope of linear regression over time). For the CAS-4 condition with 4 vital signs, this yields 36 input features, plus optionally 4 statistics of the A(t).

For LSTM models, hourly time series are used directly. Vital signs are resampled to hourly intervals using forward-fill for gaps up to 3 hours. This forms a sequence of length 24 with 4 channels for the CAS-4.

3.4. Recalibration of Index A(t) on MIMIC-IV

The instantaneous alarm index A(t), defined in [5], is computed as a bounded sum of elementary sensor deviations:

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

$$A(t) = \text{sat}(\Sigma\, w_i\, m_i\, \varphi(z_i(t))),$$

where $z_i(t) = (x_i(t) - \mu_i) / \sigma_i$ is z-score of the i-th parameter;

w_i is clinical weights reflecting the impact on overall risk;

m_i is contextual multipliers;

$\varphi(z)$ is piecewise-linear sensitivity function;

$\text{sat}(\cdot)$ is saturation operator ensuring $A(t) \in [0,1]$.

For this study, the weights were recalibrated based on absolute Pearson correlations with in-hospital mortality in the MIMIC-IV cohort, normalized to sum to 1.0. Since fall detection and immobility channels are absent in MIMIC, the full weight budget is allocated to vital signs. Recalibrated weights: respiratory rate $w\_RR = 0.433$ (highest, consistent with clinical literature on the importance of RR), heart rate $w\_HR = 0.280$, oxygen saturation $w\_SpO_2 = 0.162$, temperature $w\_Temp = 0.125$. Population means μ_i and standard deviations σ_i were computed using the MIMIC-IV training set.

Sensitivity function $\varphi(z)$ preserved from [5]:

$\varphi(z) = 0$ at $|z| \le 0.5$;

$\varphi(z) = (|z| - 0.5)/1.5$ at $0.5 < |z| < 2.0$;

$\varphi(z) = 1$ at $|z| \ge 2.0$.

3.5. Machine Learning Models

XGBoost (eXtreme Gradient Boosting) [7] is used as the primary feature-based classifier. Hyperparameters: 500 trees, maximum depth 6, learning rate 0.05, subsample 0.8, colsample_bytree 0.8, early stopping on validation AUPRC with patience of 30 epochs. Class imbalance is compensated by the scale_pos_weight.

Logistic Regression (LR) - a linear baseline with L2 regularization and balanced class weights, to demonstrate the advantage of nonlinear models.

LightGBM - an alternative gradient boosting with identical hyperparameters to XGBoost and scale_pos_weight for class imbalance, for a controlled comparison between boosting implementations.

LSTM (Long Short-Term Memory) [8] processes hourly time series directly. Architecture: two stacked LSTM layers (64 and 32 units respectively), a Dropout (0.3) layer, an output Dense layer with sigmoid activation. Training: Adam optimizer, binary cross-entropy, batch size 256, early stopping with patience of 10 epochs. Results are averaged over 5 independent runs with different random seeds to account for stochastic variability.

Hybrid Architectures.

Architecture A - A(t) as feature. Summary statistics of A(t) over a 24-hour window (maximum, mean, standard deviation, last value) are added as derived features to the XGBoost input. This tests whether the analytical index provides complementary information that the ML model cannot extract directly from raw vital signs.

Architecture B - Cascade Pre-filter. Mean A(t) classifies extreme cases directly: if A(t) is below θ_low - low-risk patient; if above θ_high - high-risk. Intermediate patients are evaluated by the ML model. Thresholds are optimized on the validation set.

Architecture C - Weighted Ensemble.

$$\text{Final risk} = \alpha \cdot \bar{A}(t) + (1-\alpha) \cdot p\_ML,$$

where $\bar{A}(t)$ is maximum value of A(t),

p_ML is probability from the ML model,

α is optimized on the validation set.

3.6. Metrics and Statistical Analysis

The primary discrimination metric is AUROC (area under the receiver operating characteristic curve). Given the class imbalance (mortality ~10%), AUPRC (area under the precision-recall curve) is additionally reported, which is more sensitive to classifier performance on the minority class.

Secondary metrics: false positive rate (FPR) at fixed sensitivity of 0.80 (directly relevant for assessing alarm fatigue). Statistical comparison of AUROC between models uses the DeLong [12] test for correlated ROC curves with **FDR Benjamini-Hochberg correction** for multiple comparisons. All metrics include 95% bootstrap confidence intervals (1,000 resamples).

Model interpretability is assessed through SHAP analysis (SHapley Additive exPlanations) [13] for XGBoost models, with particular attention to whether learned feature importances align with clinical weights encoded in A(t). **Multicollinearity analysis** (VIF – Variance Inflation Factor) is conducted to confirm that A(t) features are not redundant relative to raw vital sign features. **NRI** (Net Reclassification Improvement) quantitatively assesses the clinical significance of classification improvement compared to NEWS2. **Ablation study** systematically evaluates the contribution of each A(t).

3.7. Target Hardware for edge Deployment

The CAS system is targeted at ESP32-S3-class hardware (dual-core Xtensa LX7 processor at 240 MHz, 512 KB SRAM, 8 MB PSRAM). XGBoost inference with 40 features requires less than 50 ms per prediction. The complete processing pipeline (vital sign acquisition

→ feature extraction → A(t) computation → XGBoost inference → alarm generation) operates within a 1-second latency budget, which is acceptable for real-time monitoring at 1-minute intervals.

4. Results

4.1. Cohort Characteristics

The MIMIC-IV cohort after applying all inclusion criteria comprised 51,981 ICU stays. The external validation MIMIC-III cohort comprised 30,528 stays. Detailed characteristics of both cohorts are presented in Table 2.

The cohorts are comparable in demographic characteristics and vital sign distributions. The MIMIC-III cohort is somewhat younger (median 64 vs 67 years) with a slightly higher mortality rate (10.7% vs 9.9%), reflecting changes in clinical practice between eras. Gender distribution is practically identical (~57% males). Distributions of all four vital signs overlap, confirming the validity of cross-database model comparison.

4.2. Distribution of Index A(t) in the ICU Population

The recalibrated index A(t) demonstrated a characteristic distribution in the ICU population (Fig. 1). The maximum A(t) per patient was found to be nearly fully saturated (mean = 0.989, median = 1.000), indicating that virtually all ICU patients had at least one episode of significant physiological deviation during the 24-hour observation window. This result is expected - the very nature of intensive care implies the presence of acute physiological disturbances.

In contrast to the maximum, the mean A(t) showed substantially greater discriminatory spread (mean = 0.583, median = 0.572, IQR: 0.451-0.707). Patients who died demonstrated a rightward shift in distribution - toward higher levels of sustained alarm elevation. This observation has important practical implications: in ICU populations, instantaneous peak A(t) values lose discriminatory ability, while sustained elevation (captured by mean A(t)) retains prognostic value.

4.3. Model Performance on MIMIC-IV

Table 3 presents prediction performance for all experimental configurations on the MIMIC-IV.

The central finding of the study: the CAS-4 configuration with only four vital signs retains **95.2%** of the full feature set performance (AUROC 0.759 vs. 0.797). Degradation from the full set to CAS-4 is only 0.038 AUROC points (Fig. 2). This means that laboratory data, medications, and demographic characteristics - although statistically significant - contribute only marginal additional discrimination for the binary mortality prediction task.

Among models on the CAS-4, XGBoost significantly outperforms logistic regression (AUROC 0.759 vs 0.721, +0.038, $p < 0.001$), confirming the presence of substantial nonlinear dependencies in vital sign features. XGBoost also outperforms LightGBM (0.759 vs 0.738, +0.021, $p < 0.001$). LSTM on real time series achieves 0.747 - an intermediate result between LR and XGBoost (Fig. 3).

Critically: **all variants of CAS-4 substantially outperform NEWS2** (AUROC 0.639) with an advantage of +0.108-0.123, despite NEWS2 using more parameters (6 vs. 4) and including systolic BP and level of consciousness, absent in CAS-4.

4.4. Statistical Significance with FDR Correction

Benjamini-Hochberg correction was applied for multiple comparisons control (Table 4).

The difference between CAS-4 and hybrid A (+0.003) is **statistically non-significant** (FDR-adjusted $p = 0.274$), confirming that adding static A(t) values as features does not provide significant improvement in XGBoost. However, this does not mean that A(t) is unnecessary - ablation study (subsection 4.6) reveals a more complex picture.

Table 2

Cohort Characteristics (24-hour means). Zero overlap confirmed

Characteristic	MIMIC-IV (N=51 981)	MIMIC-III (N=30 528)
Mortality, n (%)	5 172 (9.9%)	3 253 (10.7%)
Age, median (IQR)	67 (55-78)	64 (52-76)
Males, n (%)	29 405 (56.6%)	17 622 (57.7%)
HR, median (IQR)	82.9 (73.2-94.4)	84.8 (74.8-95.6)
SpO_2, median (IQR)	97.0 (95.6-98.3)	97.5 (96.2-98.7)
RR, median (IQR)	18.6 (16.5-21.2)	18.2 (16.0-20.9)
Temperature, median (IQR)	36.8 (36.6-37.1)	36.9 (36.5-37.3)

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

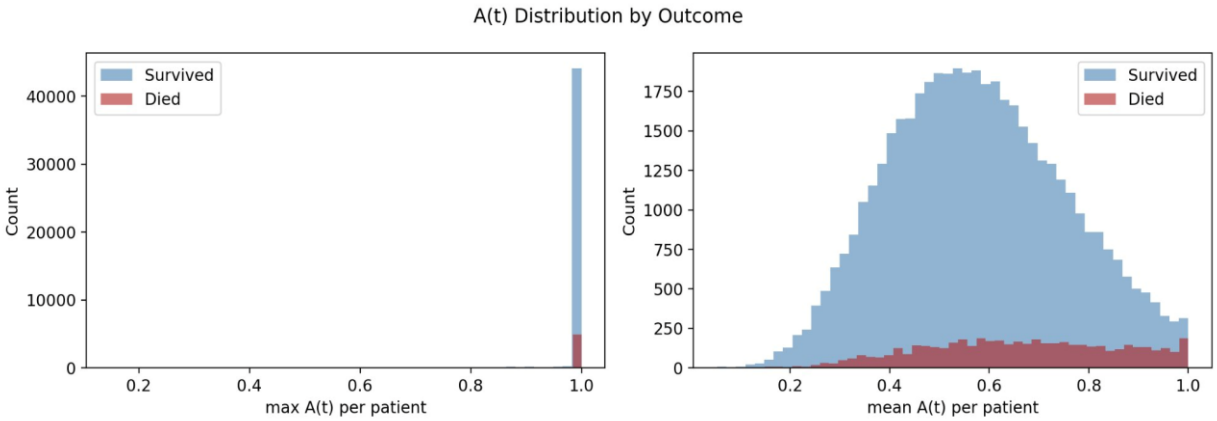

Fig. 1. Distribution of A(t) by outcome: max A(t) (left) and mean A(t) (right). Maximum is saturated for virtually all ICU patients, while the mean value retains discriminatory ability

Table 3

Prediction Performance on the MIMIC-IV

Configuration	AUROC (95% CI)	AUPRC	FPR @S=0.80	% Full
Full + XGBoost	0.797 (0.781-0.811)	0.329	0.339	100%
NEWS2-set + XGBoost	0.773 (0.757-0.789)	0.302	0.382	97.0%
CAS-4 hybrid A	0.762 (0.745-0.779)	0.288	0.408	95.6%
CAS-4 + XGBoost	0.759 (0.742-0.775)	0.287	0.416	95.2%
CAS-4 + LSTM	0.747 (0.729-0.764)	0.281	0.454	93.7%
CAS-4 + LightGBM	0.738 (0.719-0.756)	0.278	0.454	92.6%
CAS-4 + LR	0.721 (0.703-0.740)	0.260	0.515	90.5%
NEWS2 scoring	0.639 (0.618-0.660)	0.177	0.625	80.2%

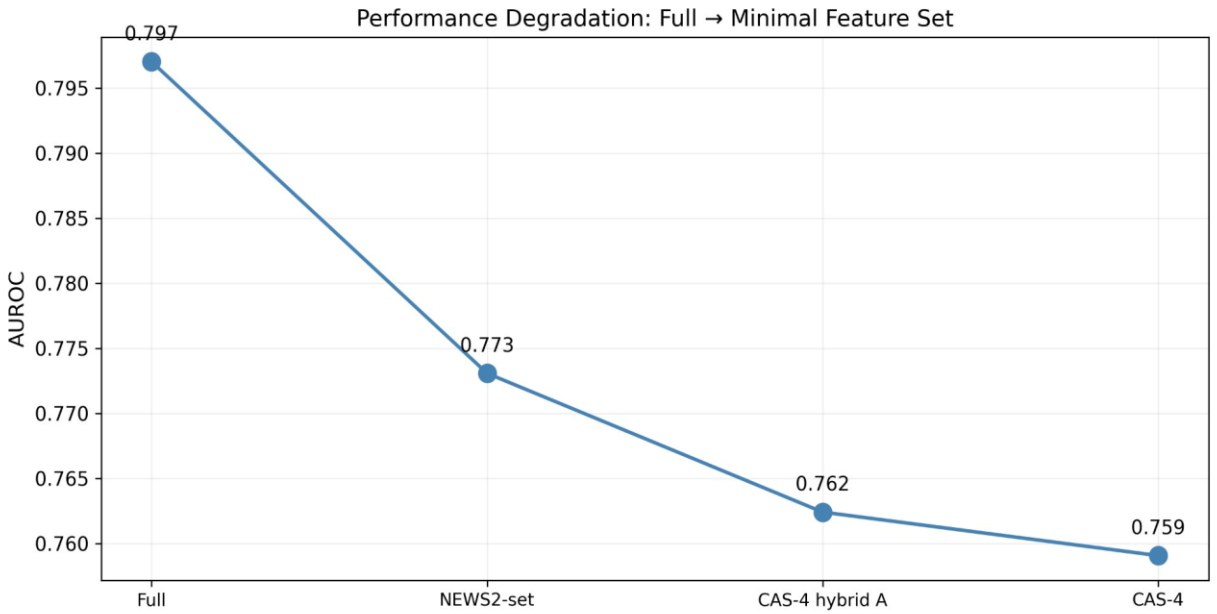

Fig. 2. Performance degradation curve from the full set to the minimal 4 parameters.
The degradation gradient becomes steeper below 4 parameters

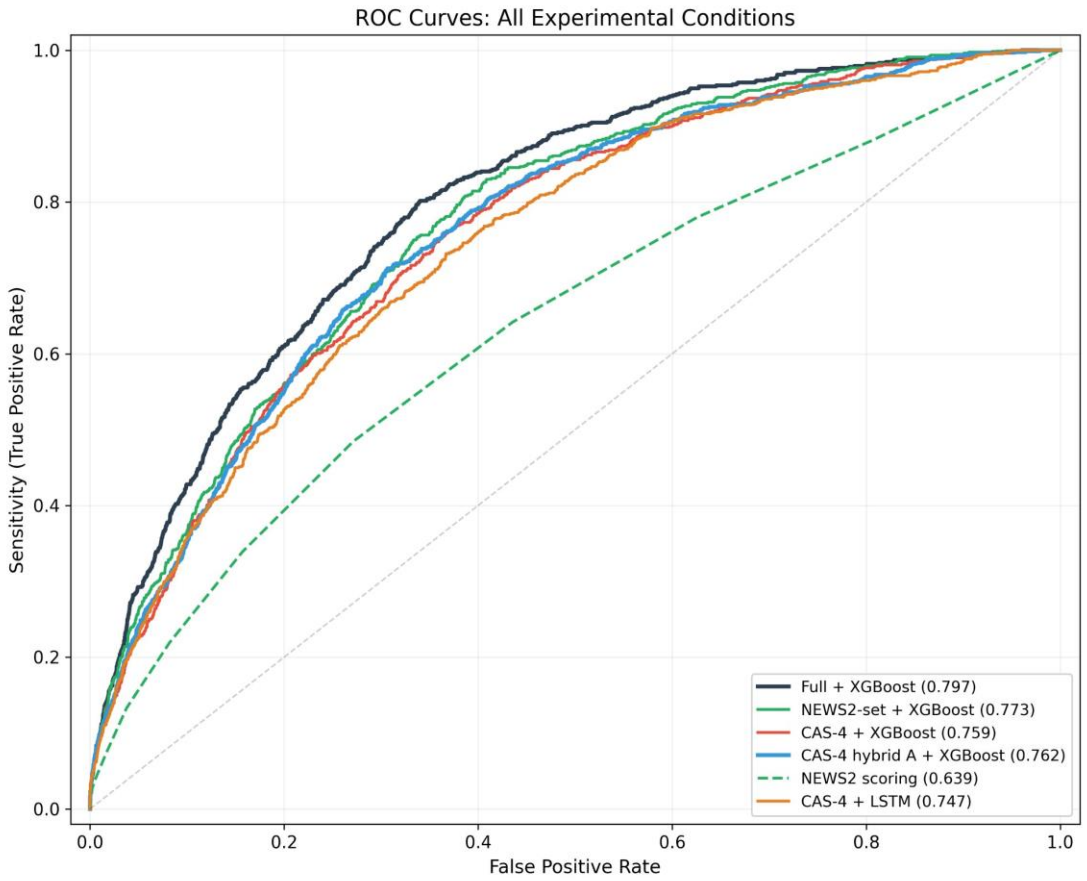

Fig. 3. ROC curves for all experimental conditions. A notable gap between ML approaches (upper group) and traditional NEWS2 scoring (dashed line)

Table 4

Pairwise Comparisons with FDR Benjamini-Hochberg Correction

Comparison	ΔAUROC	Adj. p	Significant
Full vs CAS-4	+0.040	< 0.001	Yes
Full vs CAS-4 hybrid A	+0.036	< 0.001	Yes
CAS-4 vs CAS-4 hybrid A	−0.003	0.274	**No**
CAS-4 XGB vs CAS-4 LR	+0.038	< 0.001	Yes

4.5. Multicollinearity Analysis

The VIF (Variance Inflation Factor) analysis confirmed that features derived from A(t) have low multicollinearity with raw vital sign features: at_ std = 1.71, at_mean = 1.89, at_max = 1.40, at_last = 1.27 (all well below the concern threshold of VIF > 10, and even below the conservative threshold of VIF > 5). The maximum Pearson correlation between at_std and any vital sign feature was |r| = 0.282 (with resp_rate_std), confirming that A(t) captures information orthogonal to individual vital sign statistics.

Some raw vital sign features (range, min, max) showed high VIF due to algebraic dependencies (range = max − min), but tree-based models (XGBoost) are robust to such collinearity, unlike linear models.

4.6. SHAP Analysis and ablation study

SHAP analysis of the CAS-4 hybrid A model revealed a notable result (Fig. 4): resp_rate_mean is the most important feature, followed by resp_rate_first, and in third place - **at_std (variability of A(t))**.

The dominance of respiratory rate features is consistent with the recalibrated weights of A(t) from [5], where RR received the highest weight (0.433). The high importance of at_std - the standard deviation of A(t) over the 24-hour window - indicates that the temporal dynamics of the alarm index capture patterns of physiological instability not fully represented by individual vital sign statistics.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

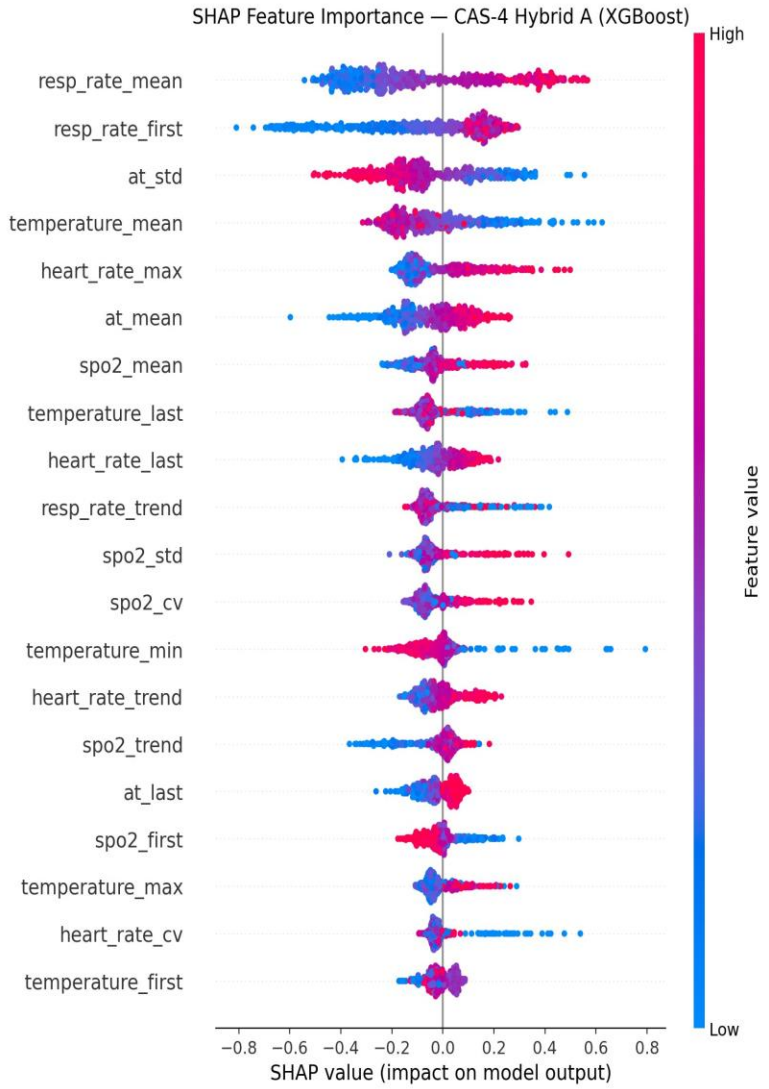

Fig. 4. SHAP analysis of feature importance for the CAS-4 Hybrid A (XGBoost) model. Color reflects feature value (red - high, blue - low). at_std is the 3rd most important feature

To verify the independence of the at_std contribution, an ablation study was conducted (Table 5).

Adding at_std alone marginally decreases AUROC (0.759 → 0.757, −0.002), explained by introducing additional noise without the context of other A(t) statistics that provide interpretation of variability. However, the full A(t) set achieves the best AUROC (0.762), and removing at_std from this set reduces performance to 0.754 - confirming that at_std contributes significant information in combination with other A(t) statistics. This synergistic pattern suggests that at_std captures the dynamics A(t), while at_mean and at_max capture its level, and the model requires both perspectives simultaneously for optimal classification.

Table 5

Results of ablation study for A(t)

Configuration	AUROC	N features
CAS-4 only	0.759	36
CAS-4 + at_std	0.757	37
CAS-4 + at_mean	0.756	37
CAS-4 + all A(t)	**0.762**	40
CAS-4 + A(t) without at_std	0.754	39

4.7. LSTM: Real Sequences vs. Synthetic

The LSTM model, trained on real hourly time series from MIMIC-IV, achieved AUROC = 0.747 (95% CI: 0.729-0.764), compared to 0.6956 obtained in [5] on synthetic sequences constructed by replicating single feature vectors from the BOLD dataset. The improvement of +0.051 confirms that true temporal dynamics - changes in vital signs throughout the day - contribute significant prognostic information unavailable through static replication. Results across individual seed runs (0.738-0.742) demonstrate training stability with low stochastic variability.

4.8. External Validation on MIMIC-III

Models trained on MIMIC-IV were applied to the MIMIC-III cohort without any retraining or adaptation (Table 6).

The models demonstrate **no degradation** on MIMIC-III - AUROC even slightly increased (+0.019 for CAS-4, +0.022 for hybrid A). This confirms robust model generalization. The moderate increase in AUROC on MIMIC-III likely reflects population differences: higher mortality rate (10.7% vs 9.9%) provides more discriminatory signal, and the earlier era (2001-2012) may be characterized by less protocolized treatment with greater natural variation in vital signs. Importantly, bootstrap confidence intervals partially overlap (MIMIC-IV: 0.742-0.775, MIMIC-III: 0.770-0.786), suggesting that the difference may be statistically non-significant.

The hybrid A effect is more pronounced on MIMIC-III (+0.006 compared to +0.003 on MIMIC-IV), suggesting that domain knowledge through A(t) provides greater additional value on an independent population. NEWS2 showed a slight decrease (−0.022), indicating that ML models are more robust across populations than fixed threshold scales.

4.9. Clinical Significance

The NRI (Net Reclassification Improvement) analysis compared CAS-4 XGBoost with NEWS2 at matched sensitivity of 80%. The overall NRI = **+0.292**, driven primarily by improved reclassification of non-event patients (NRI for non-events = +0.293). At the same sensitivity (80%) for detecting patients who will die, CAS-4 generates 264 fewer false alarms per 1,000 monitored patients (375 false alarms vs. 638 in NEWS2). This reduction in alarm fatigue by 41% represents a substantial practical improvement for real-world deployment, where excessive false alarms are the primary barrier to implementing automated early warning systems.

5. Discussion

5.1. Key Finding

The most striking finding of the study is the minor performance degradation when transitioning from the full feature set to only four vital signs: −0.038 AUROC (0.759 vs. 0.797). This means that for the binary in-hospital mortality prediction task, four core vital signs capture the vast majority of the prognostic signal available in the first 24 hours of ICU stay. Laboratory values, medication information, and demographic characteristics contribute only incremental additional discrimination. This result has direct implications for aerospace medicine: autonomous edge systems with minimal sensors can achieve prediction quality close to hospital-grade systems, making them viable for flight crew monitoring, rehabilitation support, and deployment in regions with damaged infrastructure.

5.2. CAS-4 Substantially Outperforms NEWS2

All variants of CAS-4 outperform NEWS2 scoring by +0.108-0.123 AUROC, despite using fewer parameters (4 vs. 6) and lacking blood pressure and level of consciousness assessment. The ML approach extracts substantially more discriminatory information through temporal features (trends, variability, range) that are unavailable to fixed threshold scales. The ML approach achieves NRI = +0.292 and generates 264 fewer false alarms per 1,000 patients - a 41% reduction in alarm fatigue while maintaining sensitivity.

Table 6

External Validation on MIMIC-III (zero patient overlap confirmed)

Configuration	MIMIC-IV	MIMIC-III	Δ
CAS-4 XGBoost	0.759	**0.778 (0.770-0.786)**	+0.019
CAS-4 hybrid A	0.762	**0.784 (0.776-0.792)**	+0.022
NEWS2	0.639	0.617 (0.607-0.628)	−0.022

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

5.3. Model Comparison: XGBoost, Logistic Regression, and LightGBM

XGBoost significantly outperformed both logistic regression (+0.038, FDR $p < 0.001$) and LightGBM (+0.021, $p < 0.001$) on CAS-4 features. The advantage over LR demonstrates that nonlinear dependencies in vital sign features make a significant contribution to prediction and cannot be captured by a linear model. The advantage over LightGBM is technically less fundamental but confirms the appropriateness of choosing XGBoost as the primary model.

5.4. A(t): Dynamics More Important than Level

The difference between CAS-4 and hybrid A (+0.003) is statistically non-significant (FDR $p = 0.274$), indicating that with sufficient training data volume, the ML model independently identifies the nonlinear dependencies encoded in A(t). However, ablation study reveals a more complex picture: the full A(t) set achieves the best AUROC (0.762), and removing at_std reduces it to 0.754. VIF analysis confirms that this is not multicollinearity (VIF at_std = 1.71). A(t) functions as a signal compression mechanism: it transforms four vital sign channels into a single risk trajectory whose variability captures instability patterns not represented by individual statistics.

5.5. Saturation of A(t) in the ICU Population

The failure of the cascade (Architecture B) and limited utility of the ensemble (Architecture C) architectures are caused by A(t) saturation in the ICU: maximum A(t) = 1.0 for over 97% of patients. The index, designed for home monitoring of elderly persons and rehabilitation patients [5], requires different aggregation strategies (mean, variability) instead of peak values when applied to acutely ill populations. For the target CAS population - home monitoring, rehabilitation, flight crew support - maximum A(t) values are expected to have a substantially greater discriminatory range.

5.6. Why AUROC on MIMIC-III Is Higher

The moderate increase in AUROC on MIMIC-III (+0.019) is expected and explained by several factors: (1) higher mortality rate (10.7% vs. 9.9%) provides more discriminatory signal; (2) the earlier era (2001-2012) was characterized by less protocolized treatment, creating greater natural variation in vital signs; (3) different documentation systems (CareVue + MetaVision) produce different noise profiles. Bootstrap confidence intervals partially overlap, and direct DeLong [12] comparison between independent populations is invalid. The key conclusion is robust generalization without degradation.

5.7. Limitations

Several limitations must be acknowledged.

First, MIMIC represents an ICU population that differs substantially from the target population for edge monitoring (elderly persons at home, rehabilitation patients, flight crew). Baseline vital sign distributions and event rates will differ.

Second, MIMIC vital signs are recorded by hospital monitors with substantially higher accuracy than wearable sensors (MAX30102, mmWave radar, GY-906); the impact of sensor noise on prediction quality is not modeled in this study.

Third, the A(t) index does not include fall detection and immobility components, which account for 40% of the total weight in the original CAS [5].

Fourth, in-hospital mortality is a relatively coarse endpoint; for home monitoring, more granular endpoints would be more relevant.

Fifth, patients without all four vital signs in the first 24 hours were systematically excluded (systematic exclusion bias) - these patients may represent a distinct clinical subgroup with different outcomes.

6. Conclusions

In this study, the prediction quality for early detection of physiological deterioration was systematically quantified when using only four vital sign parameters available from low-cost wearable sensors, and it was assessed whether a hybrid combination of an analytical risk index with machine learning can compensate for the absence of laboratory data and comprehensive electronic health records.

Answer to Question 1 (degradation from the full set). The CAS-4 XGBoost model on only 4 vital signs retains 95.2% of the AUROC of the full feature set (0.759 vs. 0.797). The cost of transitioning from hospital monitoring to edge deployment is only −0.038 AUROC - significantly less than expected based on the literature, where typical degradation from removing laboratory data is estimated at 0.05-0.10. This confirms that four core vital signs capture the vast majority of the prognostic signal for binary mortality prediction.

Answer to Question 2 (hybrid value). Adding static A(t) values as features provides a marginal improvement of +0.003 AUROC (statistically non-significant, FDR $p = 0.274$). However, ablation study revealed a synergistic effect: the full A(t) feature set (max, mean, std, last) achieves the best AUROC = 0.762, and removing at_std reduces it to 0.754. SHAP analysis

confirmed at_std as the 3rd most important feature with VIF = 1.71, indicating independence from raw vital signs. On external validation (MIMIC-III), the hybrid effect strengthens to +0.006, indicating greater value of domain knowledge on an independent population. Thus, A(t) functions as a signal compression mechanism whose temporal dynamics (variability) are more informative than its absolute level.

Answer to Question 3 (comparison with NEWS2). All variants of CAS-4 significantly outperform NEWS2 scoring by +0.108-0.123 AUROC, despite using fewer parameters (4 vs. 6) and lacking blood pressure and level of consciousness assessment. The NRI = +0.292 and reduction of false alarms by 264 cases per 1,000 patients (41% reduction in alarm fatigue) confirm the clinical significance of this advantage. On MIMIC-III the gap increases to +0.161, demonstrating greater robustness of the ML approach compared to fixed threshold scales.

Answer to Question 4 (minimal set). Four parameters - heart rate (HR), oxygen saturation (SpO_2), respiratory rate (RR), and body temperature - with temporal feature extraction provide clinically meaningful prediction quality sufficient for autonomous edge deployment on ESP32-S3-class hardware with inference latency under 50 ms. External validation on MIMIC-III (zero patient overlap) confirmed generalization without degradation (AUROC 0.778).

These results confirm the viability of intelligent autonomous early warning systems for aerospace medicine, flight crew monitoring, and deployment in regions with damaged or absent medical infrastructure.

6.1. Future Research Directions

The study results outline several directions for future work.

First, prospective validation using actual CAS hardware - comparing vital signs from wearable sensors with reference hospital monitors and assessing the impact of sensor noise on prediction quality.

Second, investigation of mmWave radar as a non-contact source of HR and RR for computing A(t), with quantitative assessment of how radar-derived vital sign accuracy affects the final prediction.

Third, evaluation of A(t) dynamics (specifically at_std) as an independent early warning feature - the possibility of using alarm index variability as an independent marker of physiological instability.

Fourth, adaptation of cascade architecture thresholds for non-ICU populations, where the A(t) distribution is expected to be more discriminatory.

Fifth, extension to more clinically relevant endpoints for home monitoring - hospitalizations, chronic disease exacerbations, falls - through linkage with emergency service registries.

Author Contributions. Conceptualization, formulation of research questions, methodology development, experiment design, software implementation, data processing and analysis, visualization of results, original manuscript writing – **Yurii Myroshnyk**; scientific supervision, critical review and manuscript editing – **Oleksandr Leshchenko**.

Conflict of Interest

The authors declare no conflict of interest regarding this research, whether financial, personal, authorial, or otherwise, that could have influenced the research and its results presented in this article.

Author **Oleksandr Leshchenko** is a member of the Editorial Board of this journal. He were not involved in the peer review, handling, or decision-making process for this manuscript.

Funding

This research was conducted without financial support.

Data Availability

This manuscript uses the electronic health record datasets MIMIC-IV v3.1, MIMIC-III v1.4.

Use of Artificial Intelligence Tools

The authors used artificial intelligence technologies within permissible limits to provide their own verified data.

All authors have read and agreed with the published version of this manuscript.

References

1. Johnson, A., Bulgarelli, L., Shen L., & et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 2023, vol. 10, art. no. 1. DOI: 10.1038/s41597-022-01899-x.

2. Harutyunyan, H., Khachatrian, H., Kale, D. C., & et al. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 2019, vol. 6, art. no. 96. DOI: 10.1038/s41597-019-0103-9.

3. Bakumenko, A., & et al. *Transparent Early ICU Mortality Prediction with Clinical Transformer and Per-Case Modality Attribution*. arXiv:2511.15847, 2025. DOI: 10.48550/arXiv.2511.15847.

4. Alghatani, K., Ammar, N., Rezgui, A., & Shaban-Nejad, A. Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation. *JMIR Medical Informatics*, 2021, vol. 9, no. 5, art. no. e21347. DOI: 10.2196/21347.

5. Myroshnyk, Yu., & Leshchenko, O. Development of an Autonomous System for Identification and Predictive Modeling of the Physiological State of Individuals at Risk. *Aviacijno-*

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

kosmicna tehnika i tehnologia - Aerospace technic and technology, 2026, no. 1(209), pp. 108-122. DOI: 10.32620/aktt.2026.1.10.

6. *Royal College of Physicians. National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS*. London: RCP, 2017. Available at: https://www.rcp.ac.uk/improving-care/resources/national-early-warning-score-news-2/. (accessed 10.02.2026).

7. Chen, T., & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794. DOI: 10.1145/2939672.2939785.

8. Hochreiter, S., & Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 1997, vol. 9, no. 8, pp. 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.

9. Singh, A., Rehman, S. U., Yongchareon, S., & Chong, P. H. J. Multi-Resident Non-Contact Vital Sign Monitoring Using Radar: A Review. *IEEE Sensors Journal*, 2021, vol. 21, iss. 4, pp. 4061-4084. DOI: 10.1109/JSEN.2020.3036039.

10. Johnson, A., & et al. *MIMIC-IV (version 3.1)*. PhysioNet, RRID:SCR_007345, 2024. DOI: 10.13026/kpb9-mt58.

11. Johnson, A., & et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 2016, vol. 3, art. no. 160035. DOI: 10.1038/sdata.2016.35.

12. DeLong, E., DeLong, D., & Clarke-Pearson, D. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 1988, vol. 44, no. 3, pp. 837-845. DOI: 10.2307/2531595.

13. Lundberg, S., & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Neural Information Processing Systems 30 (NIPS)*, 2017, pp. 4765-4774. Available at: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html. (accessed 10.02.2026).

Received 08.01.2026, Received in revised form 10.02.2026
Accepted date 15.04.2026, Published date 22.04.2026

ГІБРИДНА МОДЕЛЬ РАННЬОГО ПОПЕРЕДЖЕННЯ ДЛЯ АВТОНОМНОГО ФІЗІОЛОГІЧНОГО МОНІТОРИНГУ З МІНІМАЛЬНИМ НАБОРОМ СЕНСОРІВ

Ю. В. Мирошник, О. Б. Лещенко

У статті розглядається валідація гібридної моделі віддаленого фізіологічного моніторингу та прогнозування погіршення стану пацієнтів на базі даних MIMIC-IV (51 981 пацієнт) із зовнішньою валідацією на MIMIC-III (30 528 пацієнтів, нульовий перетин). **Метою роботи** є розробка та обґрунтування методики кількісної оцінки якості прогнозу на 4-х вітальних знаках (ЧСС, SpO_2, ЧД, Temp) та оцінка гібриду, що складається із комплексного показника на основі правил, та результатів машинного навчання: A(t)+ML. Використовуваними **методами** є: рекалібрування показника A(t); 4 умови (Full, NEWS2-set, CAS-4 та CAS-4 гібрид); контрольоване машинне навчання для класифікації та регресії (XGBoost), високопродуктивний фреймворк градієнтного бустингу (LightGBM), логістична регресія (LR), аппарат рекурентних штучних нейронних мереж (LSTM); FDR-корекція Бенджаміні-Хохберга; SHAP; регресійний аналіз, для визначення показника VIF (Variance Inflation Factor); методика визначення комплексного показника мережевої готовності (NRI). **Результати.** Доведено, із використанням кількісних оцінок, належний рівень якості прогнозування раннього виявлення фізіологічного погіршення при використанні лише чотирьох параметрів вітальних знаків, доступних з недорогих носимих датчиків, а також показано, що гібридне комбінування аналітичного індексу ризику з машинним навчанням в змозі компенсувати відсутність лабораторних даних та комплексних електронних медичних записів. **Висновки.** Використання лише чотирьох базових параметрів у сукупності із процедурами машинного навчання для визначення поточного фізіологічного стану пацієнта забезпечують отримання клінічно значущого прогнозу при функціонуванні широкого класу систем віддаленого моніторингу, зокрема тих, що застосовуються.

Ключові слова: система раннього попередження; MIMIC-IV; індекс A(t); авіаційно-космічна медицина; edge computing; XGBoost; LSTM; NRI.

Мирошник Юрій Васильович – старший інженер-програміст, AVI-SPL, Inc. (https://avispl.com/), Тампа, США.

Лещенко Олександр Борисович – канд. техн. наук, проф., проф. каф. комп'ютерних наук та інформаційних технологій, Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна.

Yurii Myroshnyk – Senior Software Engineer, AVI-SPL, Inc.(https://avispl.com/), Tampa, USA, e-mail: myroshnyk@gmail.com, ORCID: 0009-0004-5117-5289.

Oleksandr Leshchenko – Candidate of Technical Science, Professor, Professor at the Department of Computer Science and Information Technology, National Aerospace University «Kharkiv Aviation Institute», Kharkiv, Ukraine, e-mail: o.leshchenko@khai.edu, ORCID: 0000-0001-9405-4904.