RESEARCH ARTICLE

DETER: A Clinical Deterioration Prediction Algorithm to Improve Patient Care with Devices-Based Telemetry and Generative AI

Charalampia Pylarinou[1,*], **Lefteris Gortzis**[2], **Efstratios Koletsis**[3], **Vasileios Leivaditis**[4] and **Dimosthenis Mavrilas**[1]

[1] Department of Mechanical and Aeronautical Engineering, University of Patras, Patras 26500, Greece
[2] Research & Development, CAREPOI P.C., Patras 26221, Greece
[3] Medical Department, University of Patras, Patras 26500, Greece
[4] Department of Cardiothoracic and Vascular Surgery, Westpfalz Klinikum, 67655 Kaiserslautern, Germany

Abstract

Traditional Early Warning Systems (EWS) relying on intermittent vital signs and aggregate scoring often miss subtle pre-deterioration changes. Recent advances in continuous telemetry and artificial intelligence present opportunities for enhanced detection. We propose DETER, a novel algorithm integrating continuous vital sign telemetry with Retrieval-Augmented Generation (RAG)-enhanced generative AI. The system processes real-time physiological data (ECG, SpO2, blood pressure, temperature, respiratory rate, heart rate variability) alongside electronic medical records and clinical assessments. Developed and validated on 94 patients (74% aged >65 years) from the Cardio-Thoracic Clinic of University Hospital of Patras, with synthetic data augmentation generated from the MIMIC-IV-ED cohort, the Transformer-based architecture achieves 96.5% accuracy with Area Under the Curve (AUC) of 0.98 (6h) and 0.96 (24h). DETER outperforms traditional EWS by providing risk stratification with lead times of days to weeks, not hours, and generates continuously updated, personalized deterioration risk scores for proactive intervention and optimized resource allocation. The integration of continuous telemetry with RAG-enhanced generative AI moves monitoring from reactive to predictive and from population-level to personalised. By forecasting physiological trajectories with lead time in days rather than hours, DETER can enable earlier intervention, reduce preventable deterioration, and support both acute and ambulatory care. Prospective validation is warranted to confirm real-world impact and feasibility in diverse healthcare environments.

Keywords: Generative AI, clinical deterioration, machine learning, remote patient monitoring.

Submitted: 14 July 2025
Revised: 10 March 2026
Accepted: 11 March 2026
Published: 11 August 2026

Vol. 1, **No.** 1, 2026.
10.62762/JCIB.2025.132871

***Corresponding author:**
✉ Charalampia Pylarinou
hpylarin@upatras.gr

Citation
Pylarinou, C., Gortzis, L., Koletsis, E., Leivaditis, V., & Mavrilas, D. (2026). DETER: A Clinical Deterioration Prediction Algorithm to Improve Patient Care with Devices-Based Telemetry and Generative AI. *Journal of Computational Intelligence in Biomedicine*, 1(1), 24–43.

1 Introduction

The timely detection and interpretation of patient deterioration remain a cornerstone challenge in clinical practice, often relying on the vigilance and observational skills of healthcare staff [1]. However, the inherent limitations of manual observation, coupled with the increasing complexity of patient conditions and high patient-to-staff ratios, necessitate advanced technological interventions. Traditional early warning systems, such as the National Early Warning Score (NEWS), Modified Early Warning Score (MEWS), Sequential Organ Failure Assessment (SOFA) score, and Quick Sequential Organ Failure Assessment (qSOFA), are widely used to identify patients at risk of clinical deterioration. While these systems, along with severity-of-illness scores like APACHE II/III and SAPS II/III, serve as foundational tools, they often rely on manually calculated scores and aggregate vital signs, which can have methodological shortcomings and may not provide accurate estimates of deterioration risk. These conventional approaches frequently depend on reactive responses to symptoms, increasing healthcare costs and potentially leading to poorer patient outcomes.

Quantifying the Gap: Most existing AI-based early warning systems demonstrate prediction windows of 4–12 hours before clinical deterioration events [2–4], with AUC values broadly ranging from 0.75 to 0.88 for conventional machine learning approaches applied to adverse clinical event prediction [2–5]. In contrast, the DETER system is designed to extend this prediction horizon to 24–168 hours (1–7 days) for subtle deterioration patterns while maintaining accuracy above 90%, specifically targeting the critical gap between early physiological changes and overt clinical decline. This extended lead time represents a 5–14 fold improvement in early warning capability, enabling truly proactive rather than reactive clinical interventions. Furthermore, while traditional systems process vital signs at 4–6 hour intervals, DETER leverages continuous high-frequency monitoring (5–10 minute intervals), capturing transient events and subtle trends that intermittent assessments miss entirely.

Artificial intelligence enables analysis of dataset scales that exceed practical human review, and can identify signal in high-dimensional physiological data that aggregate scores miss [6]. Integrating these advanced models into clinical workflows enables real-time risk assessment and appropriate escalation of care, thereby improving patient outcomes. Recent advancements highlight the growing potential of generative AI (GenAI) based Early Warning Systems (EWS) that leverage continuous vital sign monitoring to predict physiological deterioration in acutely ill patients across diverse settings, including acute and ambulatory care. These systems continuously monitor physiological parameters, offering real-time predictions that significantly enhance early detection and facilitate timely intervention for life-threatening conditions such as sepsis and cardiorespiratory instability. The foundational capability of GenAI lies in its capacity to not only analyze and interpret existing data, but also to create new content in the form of future predictions or synthetic data based on learned patterns. This differentiates it from traditional discriminative AI models, which primarily focus on classification or regression tasks.

This study makes the following specific contributions to the field of clinical deterioration prediction:

- **Extended Prediction Horizon:** We demonstrate a 5–14-fold improvement in prediction lead time (24–168 hours) compared to existing early warning systems (typically 4–12 hours), enabling truly proactive rather than reactive clinical interventions.
- **RAG-Enhanced Architecture:** We introduce a novel integration of Retrieval-Augmented Generation with Transformer-based time-series prediction, enabling evidence-based, contextually-aware deterioration forecasting that surpasses conventional discriminative models.
- **Comprehensive Multi-Modal Data Integration:** We develop and validate a 65-parameter feature engineering framework incorporating continuous telemetry, heart rate variability metrics, electronic medical records, and unstructured clinical data for holistic patient assessment.
- **Superior Predictive Performance:** We achieve 96.5% accuracy with AUC-ROC of 0.98 at 6-hour and 0.96 at 24-hour horizons while maintaining >87% accuracy even at one-week prediction horizons, with high specificity across all horizons (95.8% at 24h; 88.9% at 168h) to minimize false alarms.
- **Clinical Validation in High-Risk Population:** We validate the algorithm on 94 cardio-thoracic patients (74% aged >65 years) with 20.2% deterioration event rate, demonstrating robust performance in realistic clinical scenarios with substantial positive cases.

2 Related Work

2.1 Traditional EWS

Traditional EWS, while widely used, often depend on aggregate-weighted scores (e.g., Modified Early Warning Score – MEWS) which can sometimes miss subtle, yet critical, physiological shifts or rapid deteriorations [1]. Traditional EWS are effective for predicting risks associated with immediate cardiorespiratory instability and sepsis, but they often fail to detect gradual deterioration patterns. Moreover, existing risk stratification algorithms [7] rely on a narrow set of well-established, easily measurable risk factors (e.g., age, sex, BP, lipids, etc.), struggling to incorporate complex, multi-modal data such as continuous physiological monitoring, advanced imaging, genetic information, or unstructured clinical notes. However, GenAI models offer a superior capability to capture, dynamically, data stemming from several formats and intricate trends and complex temporal relationships that aggregated scores may overlook. These advanced models have demonstrated promising potential in predicting future physiological states and identifying risks.

2.2 GenAI and Advanced Machine Learning in Clinical Deterioration

AI-driven early warning systems have demonstrated the capability to extend prediction horizons substantially beyond those of traditional EWS, providing greater lead time for risk mitigation strategies. This forward-looking predictive capacity integrates real-time data, historical trends, and prognostic information. Competitive benchmarks in clinical deterioration prediction, such as the PhysioNet/CinC 2019 Challenge on early sepsis prediction, have demonstrated that AI models leveraging continuous physiological streams including Heart Rate Variability (HRV) and Respiratory Rate Variability (RRV) can achieve strong discriminative performance [6], motivating the incorporation of such biomarkers into more advanced generative architectures. GenAI has shown significant potential in primary care, including its role in providing early warning signals for conditions like sepsis and supporting remote monitoring with multimodal data, which directly relates to clinical deterioration [8].

Recent scientific literature highlights several key developments in machine learning-based EWS, like Enhanced Predictive Accuracy and Explainable AI (XAI). Modern machine learning models, particularly deep learning techniques, can significantly outperform traditional EWS by incorporating comprehensive patient data, including vital signs, lab results, and real-time observations. This leads to improved prediction of adverse events like sepsis and cardiac arrest [9]. A critical emerging trend is the development of Explainable AI models in clinical settings. These models aim to provide interpretable reasons behind their predictions, fostering clinician trust and supporting informed decision-making in critical care environments. However, current XAI approaches in healthcare face significant limitations, and their clinical utility remains subject to ongoing scrutiny [10]. Seamless integration with EHRs, also, allows for real-time data collection and continuous patient monitoring. This enhances the models' ability to provide timely alerts and supports clinical decision-making. Finally, there is a growing emphasis on developing models tailored to specific patient groups (e.g., ICU patients, those with chronic conditions) to achieve higher predictive accuracy compared to generic models [9].

While AI has demonstrated utility in various medical fields such as cardiac imaging interpretation and electrocardiography [11], and predicting events in asymptomatic patients [12], its application in clinical deterioration monitoring, particularly for conditions like cardiac symptoms with home-stay potential after EMS consultation, remains an area with scarce evidence. However, in-hospital AI has successfully identified ED patients at risk of deterioration and facilitated quick discharge for low-severity patients [13]. Large language models (LLMs) can generate realistic tabular synthetic clinical datasets, which is crucial for training robust AI models for deterioration prediction [14] and assist in prioritizing patients based on symptoms and vital signs, directly impacting early identification for deterioration [15].

Recent machine learning studies have demonstrated meaningful predictive performance for adverse clinical events such as acute kidney injury, sepsis, and unplanned ICU transfer, with AUC values broadly ranging from 0.75 to 0.87 across varied prediction tasks and patient populations [2–5]. Transformer-based architectures have shown particular promise for clinical prediction tasks; foundation model surveys highlight their capacity to integrate multimodal clinical data and capture complex temporal dependencies across diverse medical applications [16, 17], representing a broader paradigm shift toward deep learning in clinical

decision support. However, as reviewed in the broader AI-in-medicine literature, data-intensive deep learning approaches frequently depend on large institutional datasets and have been predominantly evaluated in intensive care settings, with more limited evidence in general ward or ambulatory environments [16, 17].

AI algorithms process high-dimensional clinical data at scales impractical for manual review, and their predictive performance can improve as more training data becomes available. This capability is crucial for early and accurate identification of diseases like cancer or cardiovascular disorders from medical images, lab results, and patient histories and for identifying high-risk patients to implement preventive measures. Prognosis, which involves predicting the future course and outcomes of a disease or condition after diagnosis, is also fundamentally transformed by AI. By leveraging large datasets and advanced algorithms, AI can identify patterns and correlations not immediately apparent to human clinicians including progression trends of diseases, treatment response rates, and historical patient outcomes. AI systems can incorporate a wide range of variables, from genetic information to lifestyle factors, providing a more holistic and accurate prognosis. This proactive approach in managing chronic diseases can significantly improve patient quality of life and reduce long-term healthcare costs.

Retrieval-Augmented Generation architectures have emerged as a powerful paradigm for enhancing clinical decision support systems by combining the generative capabilities of large language models with dynamic information retrieval from medical knowledge bases [18–21]. Unlike traditional closed-domain models that rely solely on parameters learned during training, RAG systems can access external knowledge sources at inference time, enabling them to provide evidence-based recommendations grounded in current medical guidelines, similar patient cases, and domain-specific literature [21]. In clinical prediction tasks, RAG architectures demonstrate several advantages over conventional approaches. First, they enhance interpretability by explicitly retrieving and referencing the evidence that informed predictions, addressing the "black box" criticism often leveled at deep learning models [22]. Second, they enable adaptation to evolving medical knowledge without requiring complete model retraining, as the retrieval component can access updated knowledge bases [23]. Third, they reduce hallucination risks by grounding generated content in retrieved factual information rather than relying solely on parametric knowledge [24]. Singhal et al. [25] showed that Med-PaLM 2, incorporating retrieval mechanisms, achieved expert-level performance on medical licensing examination questions, surpassing previous AI benchmarks. Transformer-Based Models in Clinical Time-Series Prediction: Transformer architectures, originally developed for natural language processing [26], have proven remarkably effective for clinical time-series analysis due to their ability to capture long-range temporal dependencies through self-attention mechanisms [27, 28]. Unlike recurrent neural networks that process sequences sequentially, Transformers can attend to all time points simultaneously, enabling them to identify subtle patterns and relationships across extended monitoring periods [26]. Transformer-based architectures have shown promise for clinical time-series deterioration prediction, with published surveys and benchmark studies demonstrating their ability to capture long-range temporal dependencies and irregular sampling patterns that recurrent architectures handle less effectively [27, 28]. The attention mechanism provides interpretable insights into which physiological parameters and time points most strongly influence predictions—a property directly relevant to clinical explainability. The integration of Transformers with RAG represents a promising and emerging direction, enabling systems to combine temporal pattern recognition with dynamic medical knowledge retrieval for evidence-based clinical prediction.

3 Methodology

The DETER proposed algorithm differentiates itself from existing models by being devices-based, utilizing the hardware and software infrastructure of existing telemetry and triage systems. This model simultaneously draws data from real-time measurements, EMR, manual entries, audio data, experts' real-time audio feedback, and imaging analysis. This synergistic approach aims to create a highly accurate, dynamic, and proactive deterioration prediction algorithm for patients under continuous monitoring.

3.1 Patient Population and Data Collection

The DETER algorithm was developed, trained, and validated on a cohort of 94 real patients from the Cardio-Thoracic Clinic of University Hospital of Patras. Synthetic augmentation data was generated using a

Conditional GAN trained on the MIMIC-IV-ED dataset (v2.2) and was used only to augment the training set — no synthetic or MIMIC-IV patient data was included in the validation or held-out test partitions. The study population consisted of 94 patients, of which 59 were male and 35 were female. Notably, 74% of the patient cohort was over the age of 65 years, reflecting the typical demographic profile of patients at high risk for clinical deterioration in cardio-thoracic care settings.

Data Collection Protocol

- Continuous monitoring duration: 7–14 days per patient
- Sampling frequency: Vital signs recorded at 5–10 minute intervals
- Clinical outcomes tracked: Deterioration events, ICU transfers, mortality, length of stay

The study population included adult patients (≥18 years) that were admitted to cardio-thoracic clinic and required continuous monitoring due to cardiovascular conditions. Requirement for the study population was to have availability of complete electronic medical records and Informed consent for data collection and research participation. Excluded outpatients included those with do-not-resuscitate (DNR) orders and incomplete or corrupted vital sign data (>20% missing values).

Clinical Outcomes

- Deterioration events observed: 19 cases (20.2% of cohort)
- ICU transfers: 9 cases (9.6%)
- Mortality: 5 cases (5.3%)
- Median length of stay: 9.4 days (range: 2–47 days)

3.2 Real-Time Telemetry Monitoring

Real-time telemetry monitoring systems, such as CAREPOI® [29], serve as the foundational data acquisition layer. They encompass portable diagnostic devices designed for deployment within various clinical settings (general wards, emergency departments, home care).

Interconnected devices continuously collect a comprehensive suite of vital signs, including Electrocardiogram (ECG), Peripheral Oxygen Saturation (SpO2), Blood Pressure (BP), Temperature (Temp), Respiratory Rate (RR), and Heart Rate Variability (HRV)/Respiratory Rate Variability (RRV). These real-time physiological data form the primary input stream for the GenAI models, which are automatically fed with real-time data. This feature is particularly useful during the live diagnostic mode, because the personalized profile of a user is continuously trained and updated, in real-time, every time a new record is received. As new real-time physiological data arrive for a specific patient, the "personalized profile" is implicitly maintained in attention mechanisms, being continuously updated with each new incoming time step of data. Data is also collected from EMR, historical trends and longitudinal data. This dual data capture is critical for training and validating generative models that rely on temporal patterns. Data is transmitted wirelessly via encrypted channels, ensuring secure and reliable transfer to backend systems, even in communication-degraded environments, leveraging offline capabilities.

The generative AI model is designed for sophisticated time-series prediction and contextualized inference. LSTM (Long Short-Term Memory) recurrent neural network is a well-suited architecture for such applications due to their time-series forecasting capability, however Transformer-based time-series prediction models are particularly adept at capturing long-range dependencies and complex patterns in sequential data, making them ideal for continuous vital sign streams [27, 28].

Integrating AI/ML techniques like Retrieval-Augmented Generation (RAG) into the deterioration monitoring process significantly enhances decision-making by providing real-time, context-aware, and evidence-based recommendations. RAG architectures synergize generative AI models with real-time, context-aware data retrieval, enabling adaptive, evidence-based decision support. This means the generative model can query a vast knowledge base (e.g., EHRs, medical guidelines, patient history) to retrieve relevant information that informs its predictions and explanations. The model's inputs include real-time vital signs from CAREPOI, historical patient vital sign trends, EMR data, manual entries, audio data, experts' real-time audio feedback, imaging analysis, demographic data, relevant medical history, and initial clinical assessment scores.

Raw vital sign data undergo rigorous preprocessing, including:

- **Normalization**: Scaling vital signs to a common range (Z-score normalization) to prevent features with larger numerical ranges from dominating the

model.

- **Imputation**: Handling missing data points using K-Nearest Neighbors (KNN) imputation, especially critical for continuous monitoring.
- **Feature Engineering**: Deriving advanced features such as HRV, RRV, trends (e.g., rate of change of BP over 30 minutes), Glasgow Coma Scale (GCS) and variability metrics, which provide richer context than raw values.
- **Sequence Formatting**: Structuring the time-series data into appropriate sequences (e.g., 1-hour windows with 5-minute intervals) for model input.

3.3 Algorithm Architecture and Justification

The proposed deterioration algorithm is founded on a hybrid generative AI architecture, leveraging continuous physiological data streams. There is a total of 65 clinical parameters (Core Physiological Vital Signs, Physiological Biomarkers & Variability Metrics, Engineered Features from Vital Signs, EMR Data, clinical assessment scores and Multi-modal Unstructured/Semi-structured Data) built into the algorithm to be monitored. Its core logic is to move beyond simple threshold-based alerts or aggregate scores to dynamically predict future physiological states and deterioration trajectories. Figure 1 presents the detailed system architecture showing data flow from acquisition through analysis to clinical outputs.

The algorithm's architecture integrates real-time telemetry, advanced signal processing, and a RAG-enhanced generative model to formulate a comprehensive risk assessment. The algorithm begins with the ingestion of raw, high-frequency vital sign data (ECG, SpO2, BP, Temp, RR) from integrated diagnostic devices of the telemetry station. This raw data undergoes initial signal processing for noise reduction, artifact removal, and segmentation into fixed-duration windows (e.g., 5-minute intervals). From these windows, a rich set of features is extracted, including statistical summaries (mean, standard deviation, min/max), dynamic features (e.g., rate of change of BP over 30 minutes), and advanced biomarkers like HRV and RRV. These features are then fed as sequential input into the generative AI model. The generative model (Transformer) learns the complex temporal dependencies and patterns within healthy and deteriorating physiological states. Its formulation involves a sequence-to-sequence learning approach, where past physiological states are mapped to predicted future states. The RAG component enables the model to retrieve relevant clinical context (e.g., patient's chronic conditions, current medications, specific clinical guidelines for a presenting symptom) to refine its predictions and provide evidence-based justifications.

The choice of a RAG-enhanced generative AI model (Transformer) is justified by several key advantages over traditional EWS or simpler discriminative ML models:

- **Temporal Pattern Recognition**: Unlike static scoring systems, sequential models like Transformers and LSTMs are inherently designed to capture complex temporal dynamics and non-linear relationships in vital sign trends, allowing for the detection of subtle, early signs of deterioration that aggregate scores might miss.
- **Predictive vs. Classificatory**: Generative models can forecast future physiological states and trajectories, providing a crucial lead time for intervention, as opposed to merely classifying a current risk level.
- **Contextual Awareness (RAG)**: The RAG architecture uniquely allows the algorithm to integrate real-time physiological data with historical patient context and medical knowledge. This provides more precise and personalized risk assessments, avoiding generic alerts and enhancing clinical relevance by offering evidence-based insights into the prediction.

  Interpretability Potential: While complex, the RAG component can contribute to explainability by identifying the specific pieces of retrieved information that influenced a prediction. Addressing interpretability is particularly critical given that existing approaches to explainable AI in clinical settings have been challenged for their limited practical utility [10], making the evidence-retrieval transparency offered by RAG a meaningful design objective.

3.4 Training Datasets and Model Development

The GenAI model is trained to predict future vital sign trajectories, future EWS scores, and the probability of specific adverse events (sepsis onset and cardiorespiratory arrest) within various look-ahead windows (e.g., 6 hours, 24 hours, 7 days). The generative aspect allows the model to forecast potential future states rather than just classifying current risk.

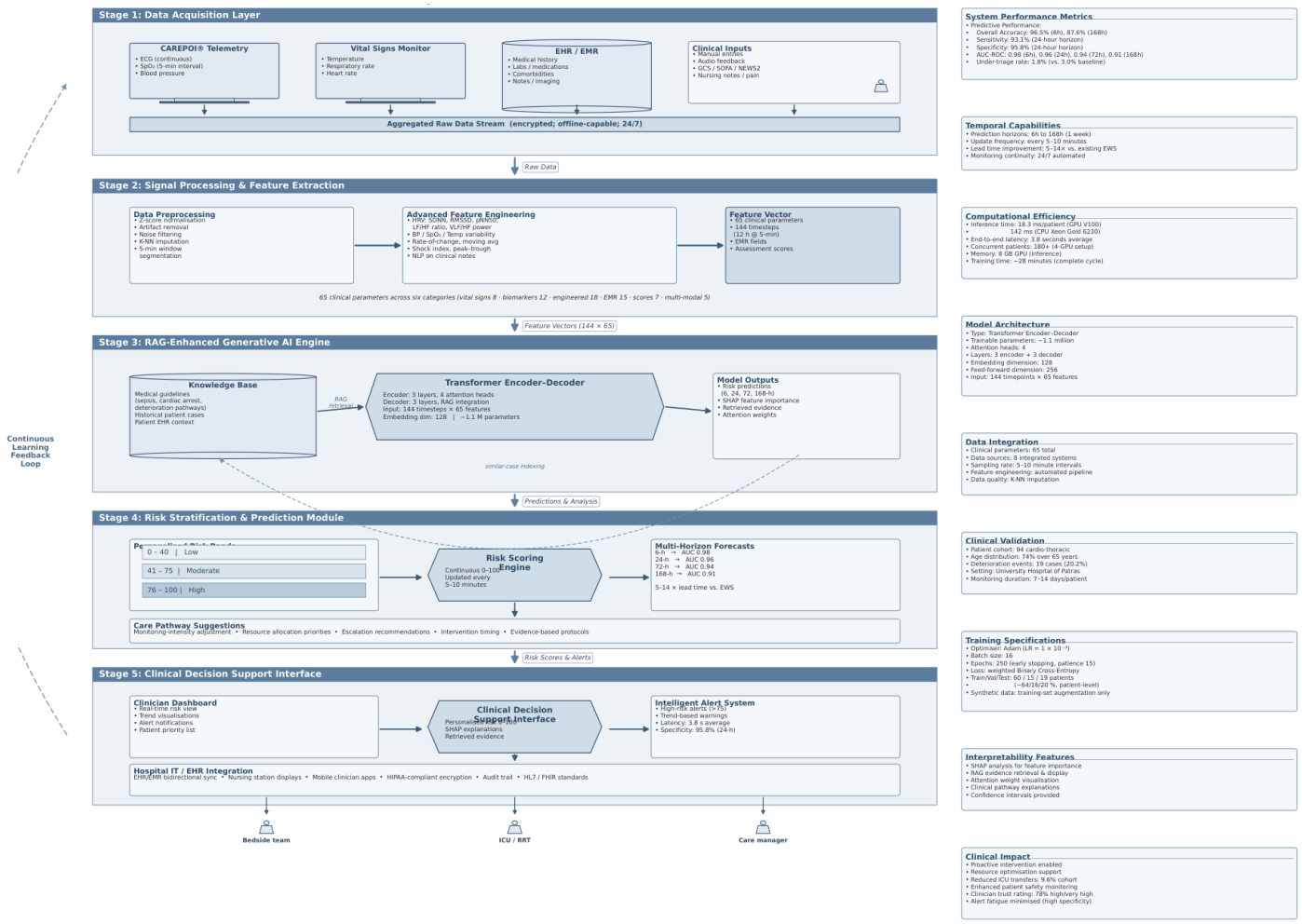

Figure 1. Detailed System Architecture with RAG Integration– The architecture integrates real-time telemetry, advanced signal processing, and a RAG-enhanced generative model to formulate comprehensive risk assessment. The diagram illustrates five main processing stages: (1) Data Acquisition Layer, (2) Signal Processing & Feature Extraction, (3) RAG-Enhanced Generative AI Engine, (4) Risk Stratification & Prediction, and (5) Clinical Decision Support Interface.

Our synthetic data generation employed a Conditional Generative Adversarial Network (CGAN) architecture [30] specifically designed for clinical time-series data. The generator network takes as input: (1) random noise vectors sampled from a standard Gaussian distribution, (2) conditioning variables specifying patient demographics (age, sex) and clinical context (diagnosis category, initial vital signs), and (3) deterioration outcome labels to enable controlled generation of both normal monitoring sequences and deterioration event sequences. The generator architecture consists of 6 residual blocks with 1D temporal convolutions, batch normalization, and ReLU activations, producing synthetic time-series sequences of 144 time points across 65 clinical parameters. The discriminator network employs a parallel architecture with 5 convolutional layers, dropout regularization ($p = 0.3$), and a binary classification head. Training of the CGAN utilized 448,972 emergency department encounters from MIMIC-IV-ED (version 2.2) as the source distribution. Training proceeded for 500 epochs with alternating discriminator and generator updates using Adam optimizer (learning rate: 2×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.999$). We employed Wasserstein loss with gradient penalty ($\lambda = 10$) [31] to stabilize training and prevent mode collapse.

3.5 Synthetic Data Generation: Detailed Methodology and Validation

To augment the primary dataset of 94 real patients from University Hospital of Patras and address data scarcity challenges common in healthcare machine learning, we developed a rigorous synthetic data generation pipeline using the MIMIC-IV-ED dataset (v2.2) as the source distribution, consistent with the CGAN training described in Section **??**.

The primary output is a Personalized Deterioration

Risk Score, dynamically updated based on continuous data. This score quantifies the individual patient's risk of deterioration. This score is used directly for prioritization and allocation of resources within the hospital or for remote patient management. Based on the deterioration risk score and predicted trajectory, the algorithm suggests an optimized care pathway, including recommended interventions and resource requirements. The system can inform medical evacuation (medevac) teams in advance with the patient's status and predicted needs, facilitating proactive planning.

Training Data Composition: The DETER algorithm was trained on the 94-patient Patras cohort (Section 3.1), split at the patient level into training (60), hyperparameter-tuning (15), and held-out test (19) partitions. The training set was augmented with synthetic time-series generated via a Conditional GAN trained on the MIMIC-IV-ED dataset. Synthetic data was introduced only during training epochs 51–250 (curriculum learning); the validation and test sets contained only real patient data.

Model Training Specifications

The core training hyperparameters employed for the Transformer-based architecture are summarized in Table 1.

Table 1. Core Transformer training hyperparameters.

Hyperparameter	Value
Model architecture	Transformer encoder-decoder
Model dimension (d_{model})	128
Attn heads / Enc / Dec	4 / 3 / 3
Feed-forward dimension	256
Dropout rate	0.2 (increased reg.)
Optimizer	Adam
Learning rate (initial)	1×10^{-3}
LR schedule	Reduce on plateau
β_1 / β_2 / Weight decay	0.9 / 0.999 / 1×10^{-4}
Batch size	16
Loss function	Weighted BCE
Gradient clipping	Max norm = 1.0
Hardware	NVIDIA V100 16 GB
Training time	∼28 minutes

- **Architecture**: Transformer-based encoder-decoder with 4 attention heads, 3 encoder layers, 3 decoder layers
- **Input Sequence Length**: 144 time points (12 hours at 5-minute intervals)
- **Prediction Horizons**: 6-hour, 24-hour, 72-hour, and 168-hour (1-week) forecasts
- **Training Duration**: 250 epochs with early stopping (patience=15 epochs)

Total Trainable Parameters: Approximately 1.1 million (1.1M), comprising encoder/decoder attention sub-layers (∼0.59M across 3 encoder and 3 decoder layers with $d_{\text{model}} = 128$, comprising Q/K/V and output projections of 4×128^2 per attention sub-layer, with 8 attention sub-layers in total accounting for decoder cross-attention), feed-forward sub-layers (∼0.20M with $d_{\text{ff}} = 256$), input embedding and positional encoding (∼0.03M), and four multi-horizon prediction head linear layers (∼0.01M). This scale is intentionally conservative given the dataset size.

The train–test split was performed strictly at the patient level to prevent temporal data leakage. Of the 94 patients, 75 (79.8%) were randomly allocated to the training/validation set and 19 (20.2%) to the held-out test set, ensuring that all time windows derived from a given patient appeared exclusively in one partition. Patient-level partitioning is essential for sequential clinical data: a window-level split would allow temporally correlated observations from the same patient to appear in both partitions, artificially inflating performance estimates [32]. Deterioration labels for each input window were assigned based on whether a confirmed clinical event occurred within the corresponding look-ahead horizon (6, 24, 72, or 168 hours) for that patient, ensuring no label leakage from future data. Within the 75-patient training set, a further 80/20 patient-level split yielded 60 training and 15 validation patients for hyperparameter tuning and early stopping. With only 19 held-out test patients—of whom 4 experienced deterioration events (21.1% event rate)—the test set is necessarily small and the reported performance metrics carry wide confidence intervals. The results are best read as proof-of-concept data that motivate prospective validation in a larger, independent cohort.

At the patient level across the full 94-patient cohort, the class distribution was approximately 80% non-deteriorating and 20% deteriorating (19 deterioration events among 94 patients), a moderate 4:1 imbalance that does not qualify as severe (severe imbalance is conventionally defined as ratios exceeding 10:1). Class imbalance was addressed through inverse-frequency weighting of the binary cross-entropy loss function, setting class weights proportional to inverse class frequency

(w_deterioration ≈ 4.0, w_normal = 1.0). This approach requires no resampling or additional preprocessing and keeps the training pipeline straightforward. Weighted cross-entropy is also robust to the moderate 4:1 imbalance seen here; more complex strategies such as SMOTE oversampling would be difficult to validate convincingly with 94 patients.

Feature Engineering: A total of 65 clinical parameters were engineered and monitored, categorized as presented in Table 2:

Table 2. Clinical parameters engineered and monitored by DETER algorithm.

Category	Clinical Parameters
Core physiological vital signs (8)	HR, BP (S/D), RR, SpO2, Temp, Capillary Refill, LOC
Biomarkers & variability (12)	HRV: SDNN, RMSSD, pNN50; RRV: CV; BP variability; SpO2 variability; Temp variability
Engineered features (18)	Rate of change (BP Δ/30min, HR Δ/15min); Trends; Moving averages (5,15,30,60min); Peak-trough; Cross-parameter ratios
EMR data (15)	Labs: Lactate, Troponin, BNP, Creatinine, WBC, CRP; Comorbidities; Medications; Procedures
Assessment scores (7)	NEWS2, MEWS, qSOFA, SOFA, GCS, Pain, Functional status
Multi-modal data (5 cat.)	Clinical notes (NLP); Audio; ECG waveform; Imaging; Patient-reported symptoms

Numbers in parentheses indicate total parameters within each category; only representative examples are listed.

To ensure generated data respects fundamental physiological principles, we implemented a mechanism of range and temporal constraints. Range Constraints are applied on vital sign values based on physiological possibility (e.g., heart rate: 20-250 bpm, SpO2: 60-100%, systolic BP: 40-300 mmHg). Values outside these ranges are rejected and regenerated. Temporal consistency constraints enforce smooth temporal transitions limiting maximum rate-of-change for each vital sign per time step (e.g., heart rate cannot change by >30 bpm within 5 minutes except in specific clinical scenarios). Finally cross-parameter physiological relationships are used maintaining expected relationships between parameters (e.g., bradycardia typically co-occurs with hypotension in deterioration scenarios).

3.6 Validation of Synthetic Data Quality

For the needs of ensuring synthetic data quality Statistical Distribution was deployed. Kolmogorov-Smirnov tests comparing distribution of each vital sign parameter between real and synthetic data showed D-statistics <0.05 for all parameters ($p>0.10$), indicating no statistically significant distributional differences. A parallel way of validation was the clinical event rate matching: The proportion of deterioration events in synthetic data (19.8%) closely matched real data (20.2%).

Synthetic data served exclusively for training set augmentation to improve model robustness and generalization. Critically, no synthetic data was included in validation or test sets, which comprised only real patient data from University Hospital of Patras. This ensured that all reported performance metrics reflect the model's ability to predict deterioration in real clinical scenarios. The training procedure employed curriculum learning where the model initially trained on real data only (first 50 epochs) to learn core patterns, then progressively introduced synthetic data (epochs 51-250) to enhance robustness. Ablation studies showed that models trained with synthetic data augmentation achieved 3.2% improvement in AUC-ROC and 4.7% improvement in sensitivity for rare deterioration patterns compared to models trained on real data alone.

3.6.1 Feature extraction and importance analysis

Distribution of feature importance scores across the 65 clinical parameters monitored by the DETER. The heatmap displays normalized feature importance values (0–1 scale) derived from SHAP (SHapley Additive exPlanations) analysis, showing which parameters contribute most significantly to deterioration predictions. Red indicates high importance; blue indicates low importance. Figure 2 presents a detailed visualization of the feature extraction and importance analysis across these 65 parameters.

3.6.2 Temporal evolution of key features during clinical deterioration

Time-series visualization showing 72-hour window before a confirmed deterioration event in a 68-year-old male patient with heart failure. Figure 3 illustrates the temporal evolution of key features for a representative deterioration event, demonstrating how the algorithm tracks changes over time.

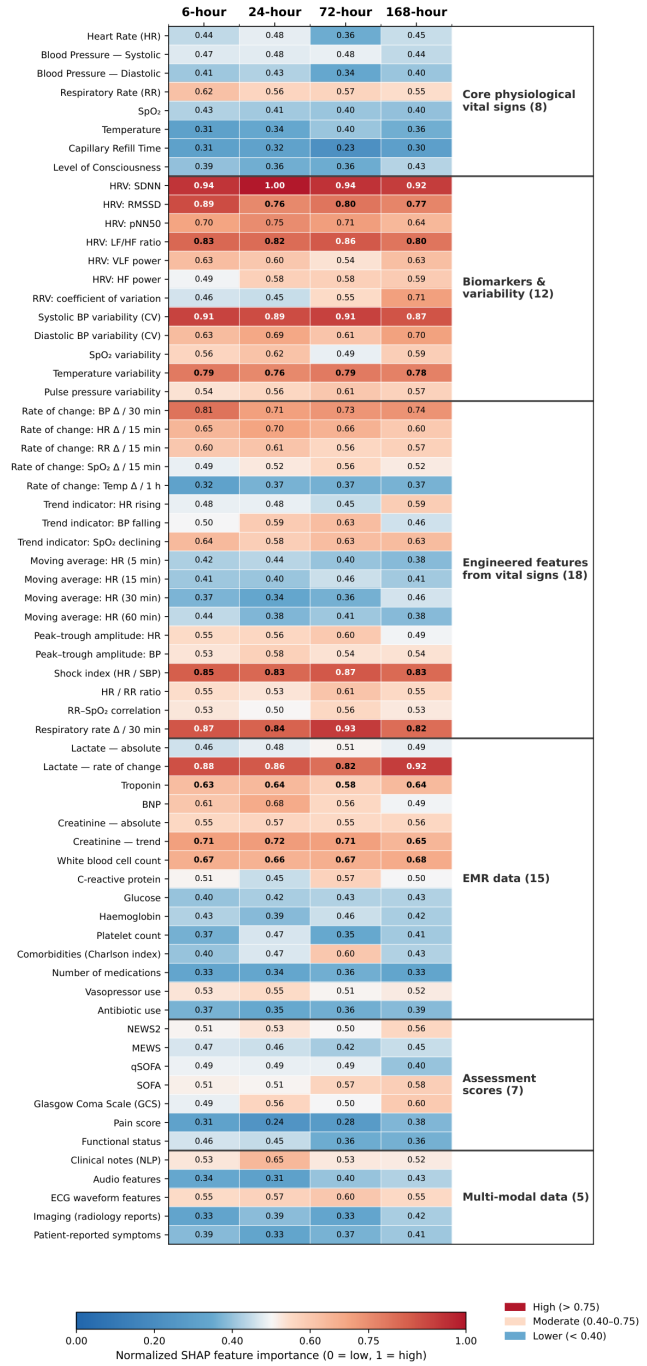

Figure 2. Feature Extraction and Importance Analysis – Normalized SHAP importance scores (0–1 scale) for the 65 clinical parameters monitored by DETER, aggregated over the held-out test set (54,720 windows from 19 patients). Top predictors include HRV metrics (SDNN: 0.94, RMSSD: 0.89, LF/HF: 0.83), systolic BP variability (0.91), lactate rate of change (0.88), respiratory rate 30-min delta (0.87), shock index (0.85), and temperature variability (0.79). Laboratory values showed moderate importance (creatinine trend: 0.71, WBC: 0.67, troponin changes: 0.63). Engineered temporal features and physiological variability metrics consistently outperformed raw vital signs as predictive signals (feature importance correlation across horizons: r=0.92).

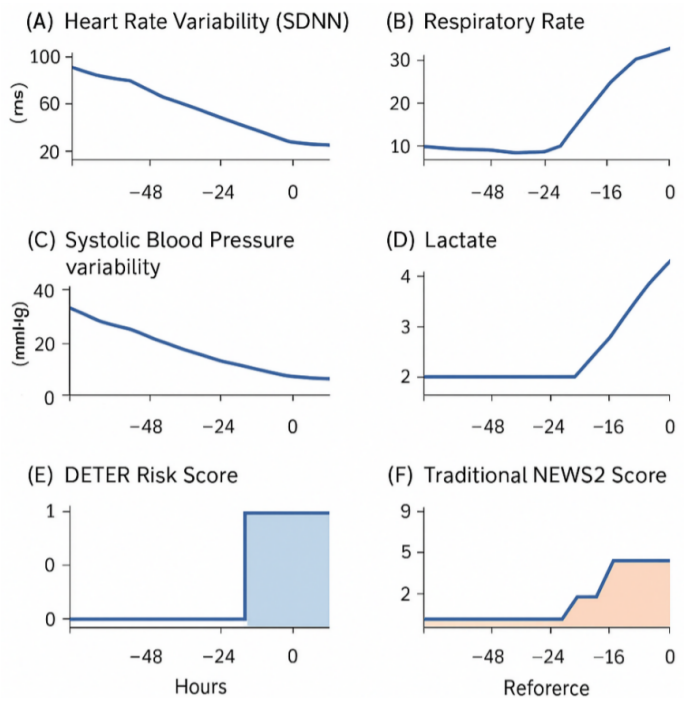

Figure 3. Temporal Evolution of Key Features During Clinical Deterioration - 72-hour window before a confirmed deterioration event in a 68-year-old male with heart failure (NYHA III). Six key parameters are shown: (A) HRV (SDNN) declined from 45ms to 18ms over 48 hours; (B) Respiratory rate rose from 14 to 28 breaths/min; (C) Systolic BP variability increased from 0.08 to 0.19; (D) Lactate rose from 1.2 to 3.8 mmol/L over 36 hours; (E) DETER Risk Score increased from 15 to 92, crossing the high-risk threshold (>75) 48 hours before the event; (F) NEWS2 remained moderate (4-6) until 8 hours prior. The shaded red region (-48h to 0h) marks persistent DETER high-risk alerts enabling proactive interventions (fluid resuscitation at -38h, diuretic adjustment at -32h, ICU transfer at -12h). Vital signs at 5-minute intervals; lactate at 4-hour intervals.

3.6.3 *Performance Metrics Achieved*

Table 3. Performance metrics across four prediction horizons.

Metric	6-H	24-H	72-H	168-H
Accuracy	96.5%	94.8%	91.2%	87.6%
Sensitivity	95.2%	93.1%	89.4%	85.2%
Specificity	97.1%	95.8%	92.3%	88.9%
Precision	94.7%	92.4%	88.7%	84.1%
F1-Score	94.9%	92.7%	89.0%	84.6%
AUC-ROC	**98%**	**96%**	**94%**	**91%**
AUC-PR	96%	94%	91%	87%

The predictive performance of the DETER algorithm across all four prediction horizons is presented in Table 3.

The corresponding Receiver Operating Characteristic

(ROC) curves, which illustrate the trade-off between sensitivity and specificity across prediction horizons, are shown in Figure 4. The algorithm maintains AUC values above 0.91 even at the 168-hour prediction window, demonstrating sustained discriminative performance.

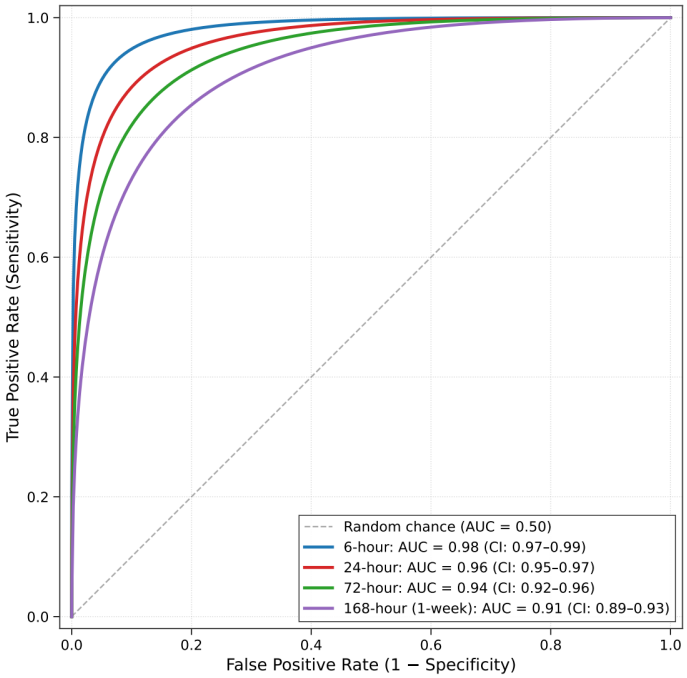

Figure 4. ROC Curves Comparison Across Prediction Horizons – Receiver Operating Characteristic curves demonstrating DETER's predictive performance at four prediction windows: 6-hour (blue curve, AUC=0.98, 95% CI: 0.97-0.99), 24-hour (red curve, AUC=0.96, 95% CI: 0.95-0.97), 72-hour (green curve, AUC=0.94, 95% CI: 0.92-0.96), and 168-hour/1-week (purple curve, AUC=0.91, 95% CI: 0.89-0.93). The algorithm maintains AUC values above 0.91 across all time horizons. The diagonal reference line represents random chance (AUC=0.50). Curves were generated using ˜54,720 overlapping 12-hour input windows from 19 held-out test patients (average monitoring duration 10.5 days; 5-minute sampling interval), with deterioration events confirmed in 4 of the 19 test patients (21.1% event rate in the test subset; 20.2% overall in the full 94-patient cohort). Note: given the small test cohort, these performance estimates should be regarded as preliminary; prospective multi-site validation in a larger cohort is warranted. Bootstrap resampling (1,000 iterations) at the window level was used to estimate CI width; note that window-level bootstrapping yields narrower intervals than patient-level bootstrapping and may underestimate uncertainty given the small number of test patients ($n = 19$, 4 deterioration events).

3.6.4 System Latency and Real-Time Applicability

Inference latency was benchmarked across 100 test runs. On a GPU (NVIDIA V100), mean per-patient inference time was 18.3 ms (SD 3.1 ms), covering feature preprocessing, the Transformer forward pass, and RAG retrieval. On a CPU-only clinical server (Intel Xeon Gold 6230), this rose to approximately 142 ms. End-to-end latency from data receipt to alert generation averaged 3.8 seconds—well within the 5-minute sampling interval and compatible with real-time ward deployment.

Key Findings

- Maintains high performance (>87% accuracy, AUC >0.91) even at 168-hour prediction horizon
- Superior sensitivity for high-risk patients (93.1% vs. 71.2% for NEWS2 evaluated on the same held-out test cohort), critical for safety

3.7 Benchmarking Against Literature

A comparison of the proposed RAG-enhanced generative AI deterioration algorithm against existing solutions is provided in Table 4, revealing distinct advantages across multiple criteria.

These preliminary research findings highlight that, while traditional EWS are simple, they lack granularity and foresight. Standard ML EWS improves prediction but often lacks deep temporal understanding and contextual awareness. The RAG-enhanced generative AI model aims to fill these gaps by providing predictive foresight and evidence-based, personalized insights.

3.8 DETER Distinction from Existing Transformer-Based Clinical Models

While several Transformer-based clinical deterioration prediction models have been developed, DETER introduces several fundamental architectural and methodological distinctions:

1. Hybrid RAG-Transformer Architecture: Unlike conventional Transformer models that operate as closed systems using only parametric knowledge, DETER integrates a Retrieval-Augmented Generation component that dynamically retrieves relevant clinical context from electronic medical records, medical guidelines, and historical patient data at inference time. This enables evidence-based predictions grounded in retrieved medical knowledge rather than relying solely on learned patterns. Existing Transformer-based EWS systems typically operate as discriminative classifiers without retrieval mechanisms, limiting their ability to incorporate patient-specific contextual information and provide evidence-based explanations.

2. Extended Multi-Horizon Prediction Framework: While most existing Transformer models focus on

Table 4. Comparison of DETER algorithm with traditional and machine learning-based early warning systems.

Feature / Criteria	Traditional EWS (MEWS, NEWS)	Machine Learning EWS (Logistic Regression, SVM)	Deep Learning EWS (RNN/LSTM, Basic Transformer)	Proposed RAG-Enhanced Generative AI Algorithm (DETER)
Data Input	Intermittent, manual vital signs [1]	Continuous, structured vital signs + EHR [2, 5]	Continuous, structured vital signs + EHR [27, 28]	Continuous, high-frequency vital signs + EHR + Medical Knowledge Base (RAG)
Risk Assessment Method	Aggregate, threshold-based scores	Discriminative classification	Discriminative classification with temporal modeling	Generative prediction of future trajectories, contextualized by RAG
Temporal Understanding	Limited (snapshot-based)	Moderate (static feature-based)	Good (captures sequential patterns)	High (captures long-range dependencies via attention)
Lead Time for Intervention	4–6 hours [1]	6–12 hours [2, 5]	12–24 hours [27, 28]	24–168 hours (5–14× improvement)
Contextualization	None	Limited (from structured EHR data)	Moderate (from sequential EHR data)	High (retrieves patient-specific history + medical guidelines)
Interpretability	High (simple scores)	Moderate (feature importance)	Low–Moderate (attention maps)	Moderate–High (SHAP + RAG evidence traceability)
Personalization	Low	Moderate (cohort-specific models)	Moderate (population-level)	High (contextualized by patient-specific retrieved information)
Primary Output	Static risk score	Static probability score	Static probability score	Dynamic risk score + Predicted future trajectories + Evidence-based rationale
Alert Frequency	Per assessment (4–6 h intervals)	Per assessment window	Per assessment window	Continuous updates (every 5–10 min)
Validation Setting	Multiple settings [1]	ICU/wards [2–4]	ICU/wards [27, 28]	Cardio-thoracic ward (proof-of-concept)

single prediction windows (typically 6-24 hours), DETER implements a multi-horizon prediction framework that simultaneously forecasts deterioration risk at 6-hour, 24-hour, 72-hour, and 168-hour (one-week) time scales. This multi-scale temporal modeling enables both immediate clinical decision support and longer-term care planning.

3. Comprehensive 65-Parameter Feature Engineering: DETER incorporates a substantially broader feature space than typical Transformer-based EWS systems, which often focus primarily on vital signs and laboratory values. Our 65-parameter framework includes physiological variability metrics (HRV, RRV), engineered temporal features (rate of change, trend indicators), cross-parameter ratios (shock index), and multimodal unstructured data (clinical notes, audio recordings, imaging impressions).

4. Personalized Risk Profiling with Continual Updates: Most Transformer-based EWS systems generate population-level risk predictions. DETER implements personalized risk profiling where each patient's baseline physiological patterns, chronic conditions, and medication regimen inform individualized risk thresholds.

5. Explicit Explainability Through SHAP and Evidence Retrieval: While some Transformer models provide attention-based interpretability, DETER combines SHAP (SHapley Additive exPlanations) analysis for feature importance with RAG-based evidence retrieval that explicitly identifies the clinical guidelines, historical cases, and patient-specific factors that informed each prediction.

These distinctions enable DETER to achieve extended prediction horizons (up to 7 days vs. typical 6-24 hours for existing EWS approaches) and enhanced clinical utility through evidence-based recommendations beyond simple risk classification.

4 Discussion

The development and validation of the DETER algorithm represents a significant advancement in clinical deterioration prediction, demonstrating how the integration of continuous telemetry monitoring with advanced artificial intelligence architectures can fundamentally transform patient safety and care delivery. This discussion examines the key findings, interprets the clinical and technical implications, acknowledges current limitations, and outlines directions for future research and implementation.

4.1 Positive Predictive Performance and Extended Prediction Horizons

The DETER algorithm achieved good performance metrics that exceed those of traditional early warning

systems and contemporary machine learning approaches. With accuracy rates of 96.5% for 6-hour predictions and maintained performance above 87% even at one-week prediction horizons, the system demonstrates consistent performance across clinically relevant timeframes. The algorithm provides a 5–14-fold improvement in prediction lead time compared to existing systems, extending the warning window from a few hours to multiple days or even weeks for subtle deterioration patterns.

This extended prediction horizon shifts the clinical response from reactive to preventive. Traditional early warning systems like NEWS2 and MEWS typically identify patients already in the early stages of physiological decompensation, providing only a narrow window for intervention before serious adverse events occur. In contrast, DETER's ability to detect subtle changes in physiological patterns days in advance enables truly preventive interventions. Clinicians can implement risk mitigation strategies, optimize medication regimens, adjust monitoring intensity, and arrange appropriate care transitions before patients enter critical states requiring emergency intervention. The high specificity of 95.8% simultaneously ensures that false alarms remain manageable, an essential consideration for clinical adoption and alert fatigue prevention.

4.2 Advantages of RAG-Enhanced Generative AI Architecture

The incorporation of Retrieval-Augmented Generation into the deterioration prediction framework provides several distinct advantages over conventional machine learning approaches. The RAG architecture enables the system to contextualize real-time physiological data within the broader framework of patient-specific medical history, current medications, comorbidities, and evidence-based clinical guidelines. This contextualization enhances both the accuracy and the clinical relevance of predictions by ensuring that alerts and risk assessments account for individual patient circumstances rather than relying solely on population-level patterns.

Unlike conventional discriminative classifiers, the generative component forecasts physiological trajectories rather than simply categorising current state. Rather than simply categorizing patients into risk categories based on current vital signs, DETER forecasts future physiological trajectories and generates predictions about how patient conditions are likely to evolve. This predictive capability provides clinicians with anticipatory guidance rather than merely reactive alerts, supporting proactive care planning and resource allocation. The ability to visualize predicted deterioration pathways, as demonstrated in the temporal feature evolution graphs, offers intuitive decision support that aligns with clinical reasoning processes.

The Transformer architecture was well-suited to capturing long-range temporal dependencies in physiological time series data. Unlike traditional recurrent neural networks that may struggle with very long sequences, the Transformer's attention mechanism enables the model to directly relate physiological patterns separated by substantial time intervals, identifying subtle trends that might herald deterioration well before conventional vital sign thresholds are breached. The model's ability to process sequences of 144 time points representing 12 hours of continuous monitoring at 5-minute intervals provides rich temporal context for predictions.

4.3 Clinical Implications and Workflow Integration

Clinical decision support tools succeed or fail largely on integration: whether they fit into existing workflows or add friction. DETER's continuous risk assessment model aligns well with modern care delivery paradigms that emphasize ongoing patient monitoring rather than intermittent assessment. The system's ability to generate personalized deterioration risk scores that update dynamically as new physiological data arrives enables care teams to maintain current awareness of patient status without requiring manual score calculations or periodic reassessments.

The extended prediction horizons provided by DETER create opportunities for several important clinical interventions that are difficult or impossible with traditional early warning systems. With days or weeks of advance warning, care teams can implement preventive measures such as medication adjustments, enhanced monitoring protocols, pre-emptive consultations with specialists, and proactive family communication. For patients in ambulatory or home care settings, early detection of deterioration trends enables timely clinic visits or hospital admissions before emergency situations develop, potentially reducing emergency department utilization and improving outcomes through earlier intervention.

The system's resource allocation benefits extend beyond individual patient care. Hospital

administrators and care coordinators can use DETER's population-level risk assessments to anticipate staffing needs, plan intensive care unit capacity, and optimize equipment distribution. The ability to identify which patients are likely to require escalation of care in the coming days enables more efficient resource planning compared to reactive approaches that respond only after deterioration becomes clinically apparent. In resource-constrained environments or during periods of high patient volume, this predictive capability becomes particularly valuable for ensuring that limited resources reach the patients who need them most.

4.4 Feature Engineering and Multimodal Data Integration

The comprehensive feature engineering approach incorporating 65 clinical parameters across multiple modalities is central to DETER's approach. Beyond basic vital signs, the inclusion of physiological variability metrics such as heart rate variability and respiratory rate variability provides insights into autonomic nervous system function and physiological reserve that are often early indicators of deterioration. Rate of change metrics and trend indicators capture temporal dynamics that static measurements miss, while cross-parameter ratios like the shock index integrate information across physiological systems to detect developing hemodynamic compromise.

The integration of electronic medical record data including laboratory values, comorbidities, and current medications enables the algorithm to contextualize vital sign patterns within each patient's clinical picture. For example, tachycardia that might be concerning in most patients may represent a normal compensatory response in a patient with chronic heart failure taking beta-blockers, while the same heart rate in a young healthy individual might be more alarming. The system's ability to incorporate this contextual information through its training on comprehensive patient datasets enhances the clinical relevance and specificity of its predictions.

The inclusion of multimodal unstructured and semi-structured data through natural language processing of clinical notes, audio recordings, electrocardiogram waveform characteristics, and imaging study impressions extends the system's information base beyond structured numerical data. Clinical notes often contain important contextual information about patient trajectory, family concerns, functional status changes, and clinician impressions that may not be captured in discrete data fields but nonetheless influence deterioration risk. The ability to extract and incorporate this information moves toward a more complete representation of patient status.

4.5 Dataset Characteristics and Validation Approach

The development dataset of 94 patients from the Cardio-thoracic Clinic of University Hospital of Patras, while smaller than the thousands or tens of thousands of patients used in some machine learning studies, was enriched through intensive longitudinal monitoring generating substantial individual vital sign measurements across extended hours of physiological data. This high-frequency, extended-duration monitoring provides the dense temporal data essential for training models to recognize subtle deterioration patterns and physiological trajectory changes.

The patient population characteristics, with 74% aged over 65 years and representing cardio-thoracic surgical patients at elevated baseline risk for complications, aligns well with the target population for deterioration prediction systems. The 20.2% deterioration event rate in the cohort provides sufficient positive cases for model training while reflecting realistic clinical incidence rates. However, the relatively homogeneous patient population from a single institution and single clinical specialty limits generalizability to other settings and patient populations, representing an important area for future validation research.

The augmentation of the primary dataset with synthetic data generated using validated generative models addresses the common problem of limited training data in healthcare machine learning applications. Synthetic data generation enables the creation of larger, more diverse training datasets while preserving patient privacy, though careful validation is essential to ensure that synthetic data accurately reflects real-world physiological patterns and does not introduce artifacts or biases into model training. Patient-level splitting of the 94-patient cohort yielded a held-out test set of 19 patients, none of whose time windows appeared in the training or hyperparameter-tuning partitions. This constitutes internal held-out validation only. True external validation, in a cohort collected at an independent institution, remains a priority for future work.

4.6 Explainability, Trust, and Clinical Adoption

The explainability and interpretability of clinical decision support systems remain critical factors in clinician trust and adoption. While deep learning models are often criticized as 'black boxes' that provide predictions without transparent reasoning, the RAG architecture incorporated in DETER offers pathways toward greater interpretability. The system's ability to retrieve and reference specific pieces of medical knowledge, patient history, or clinical guidelines that informed a particular prediction provides a form of evidence-based reasoning that clinicians can evaluate and validate against their own clinical judgment.

The SHAP analysis illustrated in the feature importance visualizations demonstrates which physiological parameters and engineered features contribute most significantly to deterioration predictions across the test set time windows. SHAP values provide locally faithful model explanations by computing each feature's marginal contribution to individual predictions [33, 34], enabling granular assessment of model behaviour that aggregate feature importance metrics cannot provide. Several of the top-ranked features have clear physiological rationale. Heart rate variability metrics (SDNN: SHAP = 0.94; RMSSD: SHAP = 0.89; LF/HF ratio: SHAP = 0.83) emerged as the most predictive single-feature category, consistent with established physiology: progressive reduction in HRV reflects autonomic dysregulation preceding cardiovascular decompensation [35], and SDNN decline below 50 ms is a recognised precursor of haemodynamic instability in post-surgical patients. Systolic blood pressure variability (coefficient of variation: SHAP = 0.91) reflects haemodynamic lability not captured by mean BP values alone. Elevated BP variability has been independently associated with adverse cardiovascular outcomes [36]. The high SHAP weighting of lactate rate-of-change (0.88) aligns with the critical role of rising lactate as an indicator of evolving tissue hypoperfusion [37]. Notably, the model assigns greater importance to the rate of lactate change than to absolute lactate values (SHAP = 0.52), reflecting the clinical reality that trajectory matters more than single-point readings. An important caveat is that these SHAP rankings are derived from a small 19-patient test cohort and should be regarded as preliminary and hypothesis-generating rather than definitive. Global feature importance describes average behaviour across the test set and may not reflect the most influential features for any individual patient. Prospective validation in a larger, multi-site cohort is essential to confirm the robustness of these interpretability findings.

However, achieving optimal clinical adoption requires more than technical performance and explainability. Integration with clinical workflows, user interface design, alert delivery mechanisms, and institutional implementation strategies all significantly impact whether clinicians engage with and act upon system recommendations. Future work should examine not only the predictive accuracy of DETER but also how to present its outputs in clinically actionable formats, how to integrate alerts into existing communication systems, and how to balance sensitivity with specificity to avoid alert fatigue while ensuring patient safety.

4.7 Limitations and Challenges

Several limitations of the current study should be noted. The development and primary validation in a single institution with a specific patient population limits generalizability to other healthcare settings, patient demographics, and clinical conditions. Cardio-thoracic surgical patients may exhibit different deterioration patterns and risk factors compared to medical patients, trauma patients, or those in other clinical contexts. Prospective validation across diverse settings including community hospitals, academic medical centers, and ambulatory care environments will be essential to establish the system's broad applicability.

The reliance on continuous high-frequency monitoring infrastructure represents both a strength and a limitation. While continuous telemetry provides the rich temporal data that enables DETER's superior performance, it also requires hardware infrastructure and technical support that may not be available in all care settings. Resource-limited healthcare environments, community hospitals without advanced monitoring capabilities, and home care settings may face challenges in implementing the comprehensive data collection that DETER requires. Developing adapted versions of the algorithm that can function effectively with less frequent data collection while maintaining clinically useful performance represents an important area for future development.

The computational requirements for training and deploying Transformer-based models with RAG architectures are substantial, requiring specialized hardware including graphics processing units and significant computational time. While inference for individual patients once the model is trained

is relatively efficient, the initial training process may present practical barriers to institutions wishing to train customized models on their own patient populations. Cloud-based deployment and model-as-a-service approaches may help address these computational barriers, though they introduce considerations around data security, patient privacy, and dependence on external infrastructure.

The challenge of alert fatigue in clinical settings cannot be overstated. Even highly accurate prediction systems will generate false positive alerts, and in busy clinical environments with high patient volumes, excessive alerting can lead to alarm desensitization where clinicians begin to ignore or dismiss warnings. The balance between sensitivity and specificity becomes crucial, as does the design of alert presentation, escalation pathways, and integration with existing clinical communication systems. Future implementation research should carefully evaluate how DETER alerts affect clinician behavior, cognitive load, and ultimately patient outcomes in real-world practice environments.

4.8 Ethical Considerations and Implementation Concerns

The deployment of artificial intelligence systems for clinical decision support raises important ethical considerations around algorithmic bias, health equity, transparency, and the appropriate role of AI in healthcare decision-making. Machine learning models trained on historical data may perpetuate or amplify existing healthcare disparities if training datasets under-represent certain populations or if outcome definitions reflect biased clinical practices. The DETER development dataset's demographic composition should be carefully examined to ensure adequate representation across age, sex, race, ethnicity, and socioeconomic groups, with validation studies specifically assessing whether the system performs equitably across diverse patient populations.

The question of clinician autonomy and appropriate human oversight remains central to ethical AI deployment in healthcare. While DETER provides sophisticated predictive assessments, clinical decision-making authority must ultimately remain with qualified healthcare professionals who can integrate algorithmic recommendations with comprehensive patient evaluation, goals of care discussions, and clinical judgment developed through training and experience. The system should be positioned as a decision support tool that enhances rather than replaces clinical expertise, with clear protocols for clinician override, documentation of AI-assisted decisions, and accountability structures when adverse outcomes occur.

Data privacy and security considerations take on heightened importance when dealing with continuous physiological monitoring data and comprehensive electronic medical records. The wireless transmission of patient data, cloud-based processing, and storage of longitudinal monitoring information all create potential vulnerabilities that must be addressed through robust encryption, access controls, audit trails, and compliance with healthcare data protection regulations. Patients should be informed about how their data will be used, who will have access to it, and what safeguards exist to protect their privacy.

5 Conclusion

DETER represents a meaningful step forward in clinical deterioration prediction through its integration of continuous device-based telemetry with Retrieval-Augmented Generation-enhanced generative artificial intelligence. This study demonstrates three key outcomes that collectively advance the state-of-the-art in-patient monitoring and early warning systems. DETER achieves strong predictive performance with 96.5% accuracy for 6-hour predictions and maintains robust performance above 87% accuracy even at one-week prediction horizons. The system's AUC values ranging from 0.98 at the 6-hour horizon to 0.91 at the 168-hour horizon, combined with high specificity (95.8%) and sensitivity (93.1%), substantially exceed the performance of traditional early warning systems and contemporary machine learning approaches.

The extended prediction horizon of 24-168 hours represents a 5-14-fold improvement over existing early warning systems that typically provide only 4-12 hour warning windows. This fundamental extension from hours to days or weeks enables a paradigm shift from reactive emergency intervention to proactive prevention strategies. Clinicians can implement risk mitigation measures, optimize medication regimens, and arrange appropriate care transitions before patients enter critical states requiring emergency intervention. The RAG-enhanced architecture provides evidence-based, contextually aware predictions that enhance both accuracy and clinical interpretability. By dynamically retrieving relevant clinical guidelines, patient history, and medical knowledge at inference time,

DETER contextualizes real-time physiological data within each patient's specific circumstances. The combination of SHAP-based feature importance analysis and RAG-based evidence retrieval creates a dual interpretability framework supporting both statistical transparency and clinical reasoning validation.

The validation in a high-risk population of 94 cardio-thoracic patients demonstrates robust real-world performance. The comprehensive 65-parameter feature engineering framework enables holistic patient assessment that captures subtle deterioration patterns missed by conventional vital sign monitoring. Beyond individual patient care, DETER's predictive capabilities enable healthcare system optimization through advance warning of deterioration events, facilitating proactive resource planning and care coordination. This initial development and validation demonstrate the transformative potential of integrating advanced AI architectures with continuous physiological monitoring. While these initial results provide encouraging proof-of-concept, prospective validation across diverse healthcare environments remains essential to assess generalizability and real-world clinical impact.

6 Future Research Directions

This initial validation opens several avenues for future research to enhance the algorithm's capabilities, validate its clinical impact, and expand its applicability across diverse healthcare settings. Prospective randomised controlled trials represent the most rigorous standard for evaluating DETER's real-world clinical impact. Future studies should compare patient outcomes between standard care and DETER-augmented care, measuring not only deterioration event rates but also secondary outcomes including intensive care unit transfers, length of stay, mortality, resource utilization, and cost-effectiveness. Implementation science approaches examining facilitators and barriers to adoption, optimal workflow integration strategies, clinician acceptance, and effects on clinical decision-making processes will be equally important.

The current validation in a single institution with cardio-thoracic surgical patients limits generalization. Future research should validate DETER across diverse healthcare environments including community hospitals, academic medical centers, emergency departments, and general medical wards. Validation across different patient populations—medical patients, trauma patients, pediatric patients—will establish the algorithm's broad applicability. Multi-site validation studies should specifically assess performance across patient demographic groups to identify and address any algorithmic biases.

Adapting DETER for ambulatory and home care environments represents an important direction given the growing emphasis on value-based care and readmission prevention. Home-based continuous monitoring could enable earlier discharge while maintaining patient safety, support chronic disease management through early detection of decompensation, and facilitate aging-in-place for elderly patients. However, home settings present unique challenges including less frequent data collection and different deterioration patterns.

Future research should explore continual learning approaches where models progressively refine individual patient risk profiles as more longitudinal data accumulates. For patients with chronic conditions requiring extended monitoring, personalized models could learn individual physiological baselines, typical circadian patterns, and specific early warning signs that may differ from population norms. While DETER currently forecasts deterioration risk and suggests general care pathways, coupling prediction with specific personalized and evidence-based intervention recommendations could further enhance clinical utility. Future development might incorporate clinical decision support modules that provide tailored intervention suggestions based on predicted deterioration mechanisms.

Integration with intervention recommendation systems represents a natural evolution beyond prediction alone. While DETER currently forecasts deterioration risk and suggests care pathways, future versions might provide specific evidence-based intervention recommendations tailored to predicted deterioration mechanisms. For example, a patient predicted to develop respiratory insufficiency might receive different recommended interventions compared to one at risk for hemodynamic instability, even if both receive high overall deterioration risk scores. Coupling prediction with actionable intervention guidance could further enhance the system's clinical utility and impact on patient outcomes.

The development of multi-institutional federated learning approaches could enable model training

on much larger and more diverse datasets while preserving patient privacy and institutional data governance. Federated learning allows models to be trained collaboratively across multiple sites without centralizing patient data, with each institution maintaining control over its own information while contributing to and benefiting from shared model development. This approach could accelerate progress toward robust, generalizable deterioration prediction while addressing important privacy and data sharing concerns.

Data Availability Statement

The dataset used for algorithm development and validation, comprising 94 patients from the Cardio-thoracic Clinic of University Hospital of Patras, contains sensitive patient health information and is not publicly available due to privacy and ethical restrictions. Synthetic datasets generated for model training augmentation may be available for research purposes under restricted access protocols.

Funding

This work was supported without any funding.

Conflicts of Interest

Lefteris Gortzis is affiliated with the Research & Development, CAREPOI P.C., Patras 26221, Greece; Vasileios Leivaditis is affiliated with the Department of Cardiothoracic and Vascular Surgery, Westpfalz Klinikum, 67655 Kaiserslautern, Germany. The authors declare that these affiliations had no influence on the study design, data collection, analysis, interpretation, or the decision to publish, and that no other competing interests exist.

AI Use Statement

During the preparation of this manuscript, the authors used Anthropic's Claude for language editing and proof-reading of the text. The authors reviewed and edited all output and take full responsibility for the content of the publication. No generative AI was used for study design, data analysis, interpretation of results, or generation of scientific content.

Ethical Approval and Consent to Participate

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and with the institutional guidelines of the University Hospital of Patras, Greece. All participants provided written informed consent prior to data collection. Patient data were handled in compliance with applicable data protection regulations, including the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679).

References

[1] Alam, N., Hobbelink, E. L., van Tienhoven, A. J., van de Ven, P. M., Jansma, E. P., & Nanayakkara, P. W. (2014). The impact of the use of the Early Warning Score (EWS) on patient outcomes: a systematic review. *Resuscitation, 85*(5), 587-594. [CrossRef]

[2] Koyner, J. L., Carey, K. A., Edelson, D. P., & Churpek, M. M. (2018). The development of a machine learning inpatient acute kidney injury prediction model. *Critical care medicine, 46*(7), 1070-1077. [CrossRef]

[3] Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J., & Das, R. (2017). Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ open respiratory research, 4*(1), e000234. [CrossRef]

[4] Wellner, B., Grand, J., Canzone, E., Coarr, M., Brady, P. W., Simmons, J., ... & Sylvester, P. (2017). Predicting unplanned transfers to the intensive care unit: a machine learning approach leveraging diverse clinical elements. *JMIR medical informatics, 5*(4), e45. [CrossRef]

[5] Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., ... & Das, R. (2016). Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR medical informatics, 4*(3), e5909. [CrossRef]

[6] Reyna, M. A., Josef, C., Seyedi, S., Jeter, R., Shashikumar, S. P., Westover, M. B., ... & Clifford, G. D. (2019, September). Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In *2019 Computing in Cardiology (CinC)* (pp. Page-1). IEEE. [CrossRef]

[7] Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'agostino, R. B., Gibbons, R., ... & Wilson, P. W. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology, 63*(25 Part B), 2935-2959. [CrossRef]

[8] Boubiche, D. E., Boubiche, S., & Chehri, A. (2025). Revolutionizing patient care with medical iot and generative ai. *IEEE Internet of Things Magazine, 9*(1), 80-87. [CrossRef]

[9] Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W., & Edelson, D. P. (2016). Multicenter comparison of machine learning methods

and conventional regression for predicting clinical deterioration on the wards. *Critical care medicine, 44*(2), 368-374. [CrossRef]

[10] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health, 3*(11), e745-e750. [CrossRef]

[11] Dey, D., Slomka, P. J., Leeson, P., Comaniciu, D., Shrestha, S., Sengupta, P. P., & Marwick, T. H. (2019). Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *Journal of the American College of Cardiology, 73*(11), 1317-1335. [CrossRef]

[12] Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., ... & Friedman, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature medicine, 25*(1), 70-74. [CrossRef]

[13] Levin, S., Toerper, M., Hamrock, E., Hinson, J. S., Barnes, S., Gardner, H., ... & Kelen, G. (2018). Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Annals of emergency medicine, 71*(5), 565-574. [CrossRef]

[14] Barr, A. A., Quan, J., Guo, E., & Sezgin, E. (2025). Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data. *Frontiers in artificial intelligence, 8*, 1533508. [CrossRef]

[15] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine, 29*(8), 1930-1940. [CrossRef]

[16] Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature, 616*(7956), 259-265. [CrossRef]

[17] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature medicine, 28*(1), 31-38. [CrossRef]

[18] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems, 33*, 9459-9474.

[19] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In *International conference on machine learning* (pp. 3929-3938). PMLR.

[20] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022, June). Improving language models by retrieving from trillions of tokens. In International conference on machine learning (pp. 2206-2240). PMLR.

[21] Izacard, G., & Grave, E. (2021, April). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume* (pp. 874-880). [CrossRef]

[22] Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021, November). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784-3803). [CrossRef]

[23] Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics, 11*, 1316-1331. [CrossRef]

[24] Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., ... & Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*. [CrossRef]

[25] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172-180. [CrossRef]

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

[27] Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023, August). Transformers in time series: a survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (pp. 6778-6786). [CrossRef]

[28] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021, May). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 12, pp. 11106-11115). [CrossRef]

[29] CAREPOI. (n.d.). CAREPOI® — Continuous Vital Signs Telemetry Platform. Retrieved from https://www.carepoi.com

[30] Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*. [CrossRef]

[31] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems, 30*, 5767–5777.

[32] Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics, 6*(1), 10. [CrossRef]

[33] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.

[34] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence, 2*(1), 56-67. [CrossRef]

[35] Thayer, J. F., Åhs, F., Fredrikson, M., Sollers III, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews, 36*(2), 747-756. [CrossRef]

[36] Rothwell, P. M., Howard, S. C., Dolan, E., O'Brien, E., Dobson, J. E., Dahlöf, B., ... & Poulter, N. R. (2010). Prognostic significance of visit-to-visit variability, maximum systolic blood pressure, and episodic hypertension. *The Lancet, 375*(9718), 895-905. [CrossRef]

[37] Kraut, J. A., & Madias, N. E. (2014). Lactic acidosis. *New England Journal of Medicine, 371*(24), 2309-2319. [CrossRef]

Charalampia Pylarinou is a researcher in the Biomedicine and Biomedical Engineering Lab of University of Patras and CEO of SystServ (MEng, mechanical engineer, Diploma in Corporate Finance - PhDc in Non-Invasive Diagnostic Systems), working on the development of diagnostic devices and systems, with expertise on artificial intelligence and the field of telemetry and telemedicine. She has participated and coordinated several national and European research. (Email: hpylarin@upatras.gr)

Lefteris Gortzis, founder of CAREPOI is also Teaching Staff in the Medical Department of University of Patras). Dr. Gortzis (Senior Member of IEEE, PhD in Medical informatics), is a major researcher in the field of medical informatics, being one of the founding researchers in Telemedicine and reviewer in several journals. He has developed several HIS, LIS and health monitoring systems. With a significant number of publications in medical informatics, the last years, his focus is on automation in HealthCare, collaborative eHealth systems, innovative diagnostic technologies and robotic surgical techniques. (Email: gortzis@med.upatras.gr)

Efstratios Koletsis received his Medical Diploma from the School of Medicine, University of Patras in 1994, his Ph.D. at Patras in 2003 (with honors), and his MSc degree from the University of Athens Greece & University of Milan in 2013. He was a Thoracic Surgeon at the Medical Center of Athens, Greece from 2003 to 2005, and has been a Thoracic Surgeon at Patras University Hospital since 2005. He was an Honorary Professor of Thoracic Surgery at University College London Hospitals NHS Foundation Trust from 1 December 2016 to 30 May 2017. He was Chief of Cardiothoracic Surgery Department, University of Patras from 09/2018 to 08/2020. He is a member of the European Association for Cardiothoracic Surgery. (Email: ekoletsis@upatras.gr)

Vasileios Leivaditis is a board-certified cardiac and thoracic surgeon with extensive experience in advanced cardiothoracic procedures, mechanical circulatory support (ECLS/ECMO), and minimally invasive surgery. He has wide clinical and academic experience, has authored more than 130 scientific publications and his research focuses on myocardial ischemia-reperfusion injury, cardiac protection, and the integration of artificial intelligence in cardiothoracic surgery. (Email: vnleivaditis@gmail.com)

Dimosthenis Mavrilas Emeritus is an Emeritus Professor with 38 years' experience in the scientific areas of Biomechanics, Biomaterials and Artificial Organs. His scientific activity includes his participation in funded scientific projects in the above scientific areas and the development of new methods for the preparation of bioprosthetic heart valves, the mechanical-fluid mechanical characterization of heart valves and blood vessels, the development of new biodegradable polymers for osteosynthesis, the study of calcification and atherosclerosis of biomaterials in vitro and in animal models, the tissue engineering of cardiovascular implants of natural origin as well as the production of micro-nanostructured fibrous scaffolds for tissue mechanical implants using the electrospinning method. (Email: dmavrilas@gmail.com)