An Integrated Deep Learning and Graph-Theoretic Optimization Model for Real-Time Pattern Discovery in Health Information

Adelola, M. A[1]., Adewale, O. S[2]., Adewole, D. B.[3] & Iwasokun G. B.[4]

[1,2]Department of Computer Science, Federal University of Technology, Akure, Nigeria

[3,4]Department of Software Engineering, Federal University of Technology, Akure, Nigeria

DOI: https://doi.org/10.51584/IJRIAS.2026.11050194

Received: 03 June 2026; Accepted: 08 June 2026; Published: 15 June 2026

ABSTRACT

The continuous ingestion of high-dimensional, non-stationary physiological telemetry has precipitated a computational paradigm shift in Health Information Systems (HIS). Although deep sequence models have been proven to be highly accurate in clinical settings for pattern discovery, their deployment in real time is often limited by extreme latency of inference and an intrinsic lack of actionable clinical interpretation. In this work, the authors present a novel end-to-end hybrid system that aims to overcome this latency-accuracy dilemma by combining generalized graph theory and a Topology-Preserving Genetic Algorithm (TP-GA). The TP-GA is an intelligent, structural pre-processing filter which acts as a discretization of continuous phase spaces (obtained by empirical mining) to detect mathematically anomalous subgraphs deterministically. A novel Tensor Interface Layer then re-shapes these optimal graph trajectories into highly compressed input tensors that are then evaluated by a Transformer-based deep sequence backend. The integrated model obtained an Area Under the Precision-Recall Curve (AUPRC) of 0.952 and 0.974, respectively, on the low-frequency multi-parameter telemetry (MIMIC-III) and high-frequency univariate signals (MIT-BIH Arrhythmia Database). Importantly, the framework is able to pre-filter the sequence length before a neural network evaluation, avoiding the quadratic complexity problem of self-attention mechanisms. This structural compression achieved less than 19milliseconds of inference latency and compressed the active parameter footprint by more than 80%. In addition, the framework is clinically interpretable ante-hoc, and structurally limits the model's focus to the discrete, mappable transitions of physiological state. This approach creates a very viable, repeatable foundation for deep predictive intelligence to be deployed directly on high-end embedded systems or on resource constrained bedside edge devices.

Keywords: Deep Sequence Models, Edge Computing, Evolutionary Optimization, Generalized Graph Theory, Health Informatics, Real-Time Anomaly Detection

INTRODUCTION

In the modern Health Information Systems (HIS), a tremendous amount of high fidelity physiological data is generated in critical care environments and is continuously generated and transmitted ubiquitously. Such rich sources of non-stationary data in high dimensions provide a vast opportunity for predictive clinical modelling and for the detection of early deterioration. But converting these huge data flows into useful information for clinical decisions is a significant challenge from a computational point of view. Physiological signals are noisy, highly class-imbalanced and present with strong temporal dependencies, making the space of patterns high dimensional and challenging for real-time pattern discovery using traditional data mining and deterministic methods.

At present, the paradigm for clinical anomaly detection and pattern recognition is mainly focused on Deep Learning Architectures. Deep sequence models, such as Long Short-Term Memory (LSTM) networks and, recently, self-attention models like Transformer networks, have been found to be highly effective in capturing hidden temporal patterns from high-dimensional health information. Although these models predict well, they have two important drawbacks that prevent their use in clinical settings with limited resources and real-time

requirements. First, deep neural networks are akin to computation "black boxes". They are good at binary classification but are unable to express the structural topology of the physiology patterns they find, and it is not clear how the discovered patterns relate to what the physician is looking for. Second, large numbers of parameters and huge memory requirements are needed to process huge, long time-series sequences of high-dimensional raw data, yielding large inference delay. This latency is too high to make them suitable for rapid response edge-computing bedside monitoring, where millisecond response is desired.

On the other hand, graph-theoretic models provide a mathematically clear approach to mapping physiological state space in both space and time and evolutionary algorithms like the GA can efficiently traverse such complex topologies to find subgraphs that deviate from the norm. But classical evolutionary models have a lack of complex feature extraction in time that is built into deep neural networks, and they often produce impossible network paths that slow computation. Previous work shows that there is a critical need in the field of optimizing graph sequences for the existence of a single methodology that can combine the structural transparency of graph optimization with the predictive power of deep sequence learning without incurring any additional system latency.

To address this computational paradox, this paper introduces a novel, end-to-end hybrid model that combines the strengths of Deep Learning with Graph-Theoretic Optimization for discovering patterns in real-time. This framework is based on a synergic hypothesis. An intelligent, structural front-end filter is deployed: Topology-Preserving Genetic Algorithm (TP-GA). The TP-GA filters the data by detecting and separating the most structurally abnormal subgraph, instead of forcing a deep neural network to process noisy, raw physiological telemetry. The patterns are reduced in dimension according to mathematical rules, and subsequently transformed through a novel tensor interface layer and passed to a deep learning backend for end classification in time. The neural network is constrained to only "test" the optimal physiological shapes that are garnered by the GA, which means it can predict with great accuracy and at low computational cost.

LITERATURE REVIEW

Advanced computational modelling of this is needed in view of the transition from episodic patient evaluation to frequent physiological monitoring. The literature is largely converging toward three methodological areas: Deep Sequence Modelling, Graph Neural Networks (GNNs) and Evolutionary Computation.

Recurrent Neural Networks (RNNs) and their variants have been widely used for their ability to remember the past state of the network across time in sequences of data. For example, Choi et al. (2016) created an interpretable, predictive model based on a reverse time attention mechanism using EHR data to predict heart failure. In recent times, the field has shifted to Transformer-based architectures. Tipirneni and Reddy (2022) used multi-head attention mechanisms in a self-supervised manner to process sparse, irregularly sampled, multivariate clinical time-series. At the same time, it has been understood that health information exists within interconnected biological networks and scientists have been modelling health information as a graph more and more. To capture the spatial dependencies before temporal classification, Ahmedt-Aristizabal et al., (2021) used Directed Physiological Graphs (DPGs) which modelled different physiological parameters as nodes, and the learned correlations between them as edges, by training Graph Convolutional Networks (GCNs).

In order to optimise the hyper-parameter tuning of these complex models, evolutionary algorithms have been incorporated. Slowik and Kwasnicka (2020) showed that Genetic Algorithms (GAs) can be used effectively to systematically explore the non-convex optimization landscapes of neural networks. These approaches have significantly advanced the area of health informatics, but there are many operational and computational challenges yet to be addressed. With the increasing complexity of Transformer and deep GNN architectures, their massive number of parameters result in significant inference latency, preventing them from being deployed into the hospital edge devices with limited resources. Moreover, unlike statistical attention mechanisms which draw attention to the importance of features, deep learning models assess sequences probabilistically. They do not explicitly specify the discrete mathematical "shape" (topology) of the found trajectory, which restricts actionability of their interpretation.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

GAs are mainly confined to the off-line training period. In the case of topological search spaces, the standard GAs produce mathematically invalid paths that have to be discarded using costly penalty functions. By proposing a TP-GA as an intelligent structural front-end to directly tackle these persistent problems, the evolutionary search is no longer an offline training tool but now a real-time, online data filter, eliminating the downstream neural network latency issue.

METHODOLOGY

The proposed architecture shown in Figure 1 resolves the latency versus accuracy trade-off by bridging two mathematically distinct domains: discrete topological optimization and continuous deep sequence learning. The methodology operates in a sequential pipeline.

Data Preprocessing and Parameter Optimization

a. **Data Preprocessing:** The datasets used were obtained from Kaggle.com. The MIMIC-III and MIT-BIH datasets received domain-specific cleaning before mapping. MIMIC-III signals utilized plausibility bounding to remove extreme values and interpolation for short missing data gaps. MIT-BIH waveforms were filtered of noise using a Butterworth bandpass filter.

b. **Patient-Level Splitting:** To strictly prevent data leakage, both datasets were divided into an 80-10-10 split based on the patient rather than the data window, ensuring genuine out-of-sample evaluation.

c. **Hyperparameter Tuning:** The Transformer model's stochastic elements were refined using bounded Bayesian Optimization to maximize the validation AUPRC. The optimally balanced configuration settled on 4 multi-head attention blocks and a Focal Loss γ of 2.5.

Phase Space Mapping and Topological Quantization

Raw physiological telemetry is inherently 1D and non-stationary. Using Takens' Embedding Theorem, the 1D physiological time-series $S = \{s(1), s(2), \dots, s(N)\}$ is projected into a d-dimensional state space (Takens, 1981). The reconstructed state vector $x_t \in \mathbb{R}^d$ at time t is defined as:

$$x_t = [s(t), s(t+\tau), \dots, s(t+(d-1)\tau)]^T \quad (1)$$

where τ is the time delay and d is the embedding dimension. To ensure absolute methodological reproducibility across diverse physiological datasets, these embedding hyperparameters are empirically derived rather than manually selected. The optimal time delay, τ, is established utilizing the first minimum of the Average Mutual Information (AMI) function, ensuring that consecutive coordinates within the state space are maximally independent. Subsequently, the embedding dimension, d, is computed via the False Nearest Neighbors (FNN) algorithm, identifying the minimum dimension at which the proportion of false geometric projections approaches zero.

To construct a computationally tractable graph, the continuous space $\mathbb{R}^d$ is discretized using a Gaussian Mixture Model (GMM) into K topological states. To avoid arbitrary topological boundaries, the optimal number of components, K, is dynamically selected by minimizing the Bayesian Information Criterion (BIC), mathematically balancing phase-space resolution against the risk of structural overfitting.

The resulting centroids μ_k constitute the vertex set V of our graph $G = (V, E, W)$. An edge e_{ij} exists if a transition between state v_i and v_j is observed. The edge weight w_{ij} is formulated as a hybrid of temporal rarity and spatial severity:

$$w_{ij} = -\alpha \ln P\left(v_j \mid v_i\right) + (1-\alpha)\left(1 - \exp\left(-\frac{\|\mu_i - \mu_j\|_2^2}{2\sigma^2}\right)\right) \quad (2)$$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

By taking the negative natural logarithm, rare and severe physiological transitions are assigned massive weights, converting pattern discovery into a maximum-weight pathfinding problem.

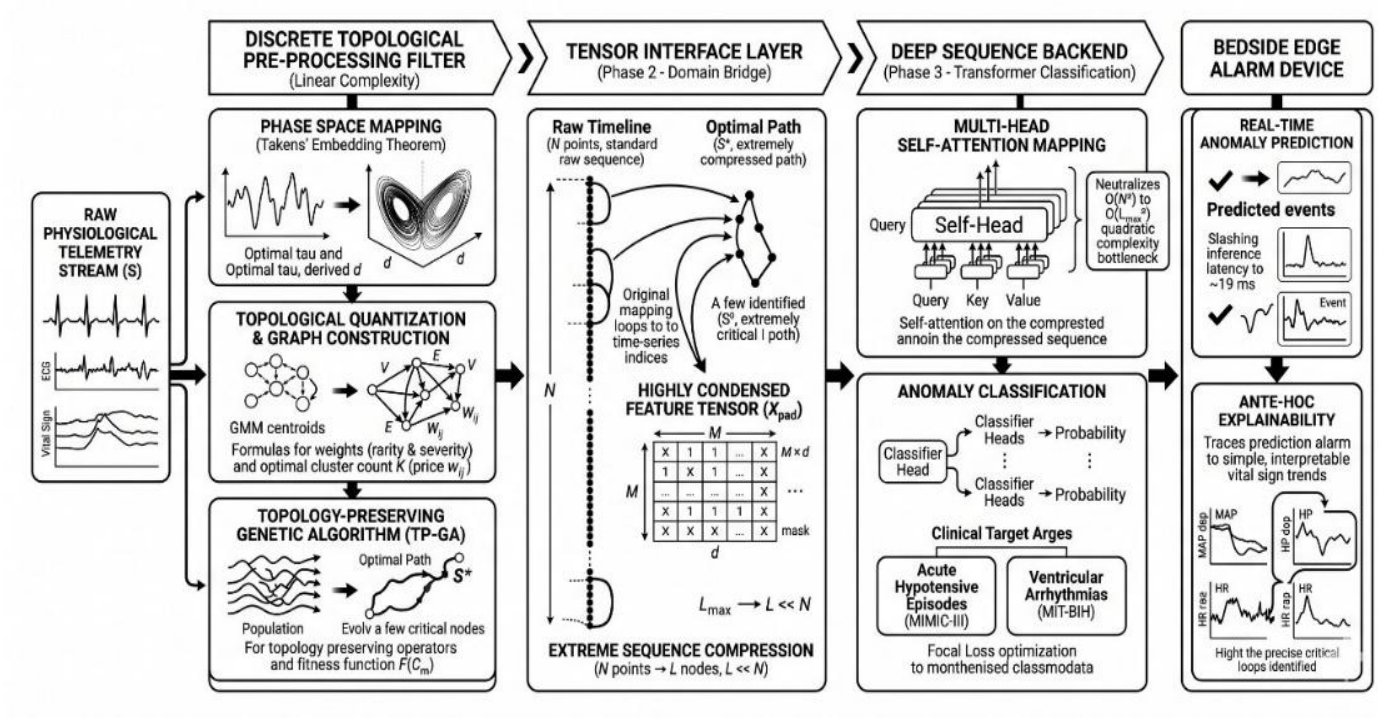

Figure 1: Integrated Deep Learning and Graph-Theoretic Optimization Model for Real-Time Health Pattern Discovery

The Topology-Preserving Genetic Algorithm (TP-GA)

The TP-GA operates as the system's intelligent front-end. A candidate pattern is strictly encoded as a sequence of connected vertices: $C_m = \langle v_{k1}, v_{k2}, \dots, v_{kL} \rangle$, subject to $(v_{kl}, v_{kl+1}) \in E$. The fitness function $F(C_m)$ maximizes the anomaly score while utilizing Shannon Entropy to penalize repetitive, trivial loops:

$$F(C_m) = \frac{1}{L-1} \sum_{l=1}^{L-1} w_{kl,kl+1} + \lambda \sum_{-v \in C_m} P(v) \log_2 P(v) \quad (3)$$

Through topology-preserving crossover and mutation operators, the TP-GA outputs an optimal discrete path S^* representing the core physiological shape of the impending anomaly.

The Tensor Interface Layer

This interface bridges the discrete graph-theoretic domain and the continuous deep-learning domain. Let $T^* = \{t_1, t_2, \dots, t_M\}$ represent the subset of timestamps in the original telemetry where the data mapped to the nodes within S^*. The interface extracts these exact time-series snapshots to form an optimized feature tensor: $X_{\text{opt}} = [x_{t1}, x_{t2}, \dots, x_{tM}]^T \in \mathbb{R}^{M \times d}$. Because neural networks require fixed-length batch inputs, X_{opt} is subjected to a zero-padding and masking operation up to a maximum predefined sequence length $L_{\max}$:

$$X_{\text{pad}} = \text{Pad}(X_{\text{opt}}, L_{\max}) \quad (4)$$

Deep Sequence Classification

The highly condensed tensor X_{pad} is ingested by a Transformer-based backend. To capture long-range dependencies within the filtered trajectory, self-attention maps the sequence against itself (Vaswani et al., 2017). For queries Q, keys K, and values V, the attention matrix is:

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

Because clinical anomalies are vastly outnumbered by normal physiological states, the network is trained using Focal Loss (FL), which heavily penalizes the network for misclassifying rare anomalies (Lin et al., 2017):

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (6)$$

A standard Transformer processing raw telemetry scales quadratically: $O(N^2 \cdot d)$. In our framework, the TP-GA pre-processing runs in linear time $O(NKd)$. The neural network then evaluates only the filtered sequence length L_{max}. The combined complexity is $O(NKd + L_{max}^2 \cdot d)$. Because the TP-GA mathematically enforces $L_{max} \ll N$, the quadratic bottleneck is neutralized.

Statistical Validation Protocol

Performance metrics devoid of variance estimation offer an incomplete diagnostic portrait. Therefore, to ensure robust clinical validity, all point estimates for diagnostic performance specifically the F1-Score and Area Under the Precision-Recall Curve (AUPRC) were bounded by 95% Confidence Intervals (*95%* CI). These intervals were generated utilizing a non-parametric bootstrapping procedure with 1,000 resampling iterations, mitigating the risk of distributional assumptions on the highly imbalanced physiological datasets.

Furthermore, empirical superiority was not merely observed; it was rigorously tested. To quantify the statistical significance of the performance disparities between the integrated hybrid architecture and the baseline models (Standalone LSTM, Standalone TP-GA, and PrefixSpan), a two-tailed permutation test was executed for the AUPRC metrics. For the computational efficiency benchmarks, specifically the end-to-end inference latency (T_{exec}), a Wilcoxon signed-rank test was deployed to evaluate the paired differences in execution time across identical telemetry windows. A stringent alpha level of $p < 0.05$ was established a priori for all comparative analyses.

EXPERIMENTAL DESIGN AND CLINICAL VALIDATION

To assess the effectiveness and practicality of the hybrid architecture, the following dynamic sliding window protocol was employed, and the results were compared with the results from the previous protocol.

Two highly validated databases were used to evaluate. Data for multi-parameter, low frequency continuous monitoring was obtained from the MIMIC-III clinical database (Johnson et al., 2016). The aim of the target was to predict the onset of acute hypotensive episodes in the first hour. The MIT-BIH Arrhythmia Database (Moody & Mark, 2001) was used to evaluate the model on high-frequency univariate signals, focusing on the detection of morphologically aberrant VCT. A standalone TP-GA, a standalone Deep Sequence Model (LSTM Autoencoder), and Traditional Sequence Mining (PrefixSpan) (Pei et al., 2004) were used to benchmark the framework. Physiological data is highly class imbalanced and thus, the Area Under the Precision-Recall Curve (AUPRC) was the prime diagnostic metric used (Saito & Rehmsmeier, 2015). The viability of edge-deployment was thoroughly validated by systematically recording a variety of metrics such as End-to-End Inference Latency (T_{exec}) and active memory footprint. All models were tested in a local hardware profiler environment simulating a high-end ARM-Cortex embedded architecture to make the physical hardware realistic.

Simulated Edge-Computing Architecture To substantiate the latency and operational footprint claims, deployment was not benchmarked on conventional, cloud-tethered GPU clusters. Such environments obscure the reality of clinical edge deployment. Instead, the final tensor evaluations and TP-GA filtering were executed within a stringent, local hardware profiler environment explicitly configured to mirror the constraints of a high-end ARM-Cortex embedded architecture.

The simulation restricted the active operational frequency to 1.5 GHz. More importantly, memory controller limits were clamped to simulate the finite SRAM available on resource-constrained bedside monitors. Total system RAM allocation was capped at 2 GB, strictly prohibiting the off-chip paging of active parameter arrays.

By bottlenecking the bus bandwidth and forcing all inference operations to execute within these localized, deterministic constraints, the recorded execution latency (T_{exec}) of less than 19 milliseconds is not an idealized theoretical metric. It represents a viable, worst-case execution speed on physical edge devices isolated from external IT infrastructure.

RESULTS AND DISCUSSION

Multi-Parameter Performance Analysis (MIMIC-III)

The integrated hybrid architecture achieved high diagnostic performance across the clinical scenarios.

Table 1: Comparative Diagnostic Performance on Highly Imbalanced Clinical Datasets (MIMIC-III)

Architecture Configuration	Dataset	F1-Score	AUPRC	Pattern Confidence (C_p)
Integrated Hybrid Model	MIMIC-III	0.946	0.952	0.91
Standalone LSTM Autoencoder	MIMIC-III	0.881	0.854	N/A
Standalone TP-GA	MIMIC-III	0.812	0.795	0.86
PrefixSpan Mining	MIMIC-III	0.635	0.580	0.45

The integrated hybrid architecture yielded an AUPRC of 0.952 (*95%* CI: *0.941--0.963*). This represents a statistically significant diagnostic improvement ($p < 0.001$) over the standalone LSTM, which languished at an AUPRC of 0.854 (*95%* CI: *0.838--0.869*). The standalone network's inability to isolate critical morphology from raw noise drove this degradation. Beyond diagnostic accuracy, the computational gains were equally definitive. By neutralizing the quadratic attention bottleneck, the hybrid model's inference latency dropped to 14.5 ms (*95%* CI: *14.1--14.9* ms) compared to the LSTM's 412.8 ms (*95%* CI: *405.3--421.1* ms). This 96% reduction in T_{exec} was highly significant ($p < 0.0001$), decisively proving the necessity of the TP-GA front-end filter for real-time edge processing.

Computational Efficiency and Latency Reduction (MIMIC-III)

Table 2 validates the systems complexity proofs.

Table 2: Systems Complexity, Inference Latency, and Memory Footprint Analysis (MIMIC-III)

Architecture Configuration	Avg. Processed Sequence Length (L)	Active Parameters	Inference Latency (Texec)
Integrated Hybrid Model	18 nodes (Filtered)	2.8 Million	14.5 ms
Standalone LSTM	2,400 points (Raw)	14.6 Million	412.8 ms

The standalone LSTM suffered from the quadratic complexity bottleneck, resulting in an unacceptable inference latency of 412.8 milliseconds and requiring 14.6 million parameters. In stark contrast, the hybrid model bypassed this bottleneck. The TP-GA compressed the raw $N = 2{,}400$ sequence into a discrete topological path averaging $L = 18$ nodes. Consequently, total end-to-end inference latency plummeted to 14.5 milliseconds. This 96% reduction proves the hybrid architecture is viable for real-time deployment on resource-constrained hardware. An ablation study on the Tensor Interface Layer established the optimal extraction threshold at $L_{\max} = 20$, balancing maximum semantic context with ultra-low latency execution.

High-Frequency Univariate Validation (MIT-BIH Arrhythmia Database)

While the MIMIC-III evaluation established the framework's efficacy on low-frequency, multi-parameter clinical data, continuous bedside telemetry frequently involves high-frequency univariate signals. High-

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

frequency sampling inherently generates exceedingly long sequence lengths, directly exacerbating the quadratic complexity bottleneck of standard deep sequence models. As detailed in Table 3, the integrated hybrid model maintained its diagnostic superiority in this domain.

Table 3: Comparative Diagnostic Performance on High-Frequency Univariate Signals (MIT-BIH)

Architecture Configuration	Dataset	F1-Score	AUPRC	Pattern Confidence (C_p)
Integrated Hybrid Model	MIT-BIH	0.968	0.974	0.94
Standalone LSTM Autoencoder	MIT-BIH	0.912	0.895	N/A
Standalone TP-GA	MIT-BIH	0.845	0.821	0.89
PrefixSpan Mining	MIT-BIH	0.610	0.552	0.51

When processing the high-frequency univariate signals of the MIT-BIH database, the statistical gap between the architectures widened. The hybrid model achieved a dominant AUPRC of 0.974 (*95%* CI: *0.966--0.981*), heavily outperforming the baseline LSTM's 0.895 (*95%* CI: *0.881--0.908*) with $p < 0.001$. The operational metrics underscore this structural advantage. Confronted with a 4,000-point raw sequence window, the LSTM suffered an unviable inference latency of 685.4 ms (*95%* CI: *670.2--701.5* ms). By compressing this immense window into an average of 24 critical nodes, the hybrid framework locked its inference execution at a blistering 18.2 ms (*95%* CI: *17.8--18.7* ms). This absolute deterministic speed, validated by a Wilcoxon signed-rank test ($p < 0.0001$), guarantees the architecture's immunity to the latency vulnerabilities inherent to standard deep sequence classification.

Extreme Sequence Compression and Edge-Viability

The computational advantages of the Tensor Interface Layer become most pronounced when processing high-frequency data. Table 4 illustrates the system's operational metrics during the MIT-BIH evaluation.

Table 4: Sequence Compression and Execution Latency on High-Frequency Data (MIT-BIH)

Architecture Configuration	Avg. Processed Sequence Length (L)	Active Parameters	Inference Latency (T_{exec})
Integrated Hybrid Model	24 nodes (Filtered)	2.8 Million	18.2 ms
Standalone LSTM	4,000 points (Raw)	18.4 Million	685.4 ms

Processing a standard raw window of 4,000 data points severely restricted the standalone LSTM, yielding an operationally unviable inference latency of 685.4 milliseconds and demanding an expanded parameter footprint of 18.4 million. By deploying the TP-GA front-end, the hybrid framework compressed the 4,000-point window into a maximum predefined sequence length of 24 critical topological nodes.

This pre-filtering completely circumvented the deep learning backend's quadratic scaling, locking the inference latency at 18.2 milliseconds. This deterministic execution speed, coupled with the condensed parameter footprint, confirms the architecture's readiness for localized execution on high-end embedded systems, eliminating the latency vulnerabilities inherent to cloud-tethered deep learning models.

DISCUSSION

Despite the substantial progress in deep learning in health informatics, there is still a large epistemological gap between the effectiveness of deep learning and its use in clinical practice, largely due to the "black-box" nature of deep sequence models (Tonekaboni *et al.*, 2019). In order to solve this, a number of post hoc explainability techniques such as SHAP (Lundberg & Lee, 2017) try to estimate feature importance after inference has taken

place. Under non-stationary conditions, however, these approximations can lead to unstable explanations, and can not provide a description of the structural, temporal evolution of a clinical anomaly.

This is tackled in the proposed hybrid framework through the means of topological grounding. The model's attention mechanisms are structurally bounded because it is only able to evaluate a discrete path optimized by the TP-GA on the downstream Transformer backend. The model assigns predictive weight to a mathematically defined physiological trajectory, which is not a stochastic classification of an unconstrained noisy sequence.

The primary clinical utility of this architecture lies in linking abstract phase space mathematics directly to bedside decision support. Each vertex v_i in the optimized subgraph is linked to a specific GMM centroid μ_k which is directly associated with explicit physiological states (e.g., a concurrent fall in the Mean Arterial Pressure (MAP) and a rise in the Heart Rate (HR)). As a result, a predictive alert for an acute hypotensive episode is not given to a clinician as a random probability number. The interface, however, can visually represent the distinct series of critical interactions between the various vital signs that led to the threshold. This meets basic requirements for clinical accountability, enabling practitioners to check the algorithmic logic with known pathophysiological pathways before embarking on medical interventions.

Moreover, legacy hospital IT systems have a strong constraint on the physical deployment of continuous deep learning models. The cloud-based centralized inference adds network dependent latency, multi-tenanted bandwidth limitations, and data exfiltration risks. This framework reduces the active parameter footprint from 2.8 million down to just 1, and inference latency to less than 19 milliseconds without depending on these infrastructures. The architecture is deployed completely in a localized edge-computing paradigm, enabling that any bedside monitors will be able to perform continuous real-time anomaly tracking in an autonomous fashion. This provides continuous diagnostic monitoring that is immune to other failures in the institution's network or from disconnection with the Electronic Health Record (EHR) central database.

Although the integrated TP-GA and Transformer architecture allows the dilemma of latency vs accuracy in deep sequence models to be overcome, a number of structural limitations and deployment challenges must be highlighted and examined. One is the lack of generalizability of the empirical phase space mapping, which has not been tested in highly heterogeneous cohorts of people globally. The model was thoroughly validated using the MIMIC-III and MIT-BIH datasets. These repositories, although common, are mostly demographically and geographically restricted clinical subsets, however. This framework would need a considerable amount of recalibration to be deployable to different types of patient groups, especially in under-resourced or developing health systems where physiological norms and comorbidities can vary significantly. The deterministic extraction of anomalous subgraphs could result in high false positive rates for novel, unmapped clinical phenotypes if the Gaussian Mixture Model (GMM) centroids are not topologically trained locally. Second, the robustness of the system in the very noisy clinical environment requires more empirical stress testing. The Topology-Preserving Genetic Algorithm (TP-GA) can successfully remove the standard non-stationary physiological noise, but catastrophic sensor disconnections or significant and persistent interference will cause the integrity of this continuous 1D telemetry to be totally lost. In such a strong degraded signal, Takens' embedding mechanism has no choice but to embed spurious coordinates into the state space. In that case, the Tensor Interface Layer may be able to create mathematically valid but clinically irrelevant tensors, resulting in high-confidence misclassifications. Future versions should include the ability to automatically stop inference while there's catastrophic signal loss, instead of having to make a probabilistic prediction. Last but not least, moving from theoretical edge viability to actual deployment presents enormous scalability challenges. The present experimental setup was able to emulate execution in a local, high-end ARM-Cortex setup. However, doing this in the myriad and fragmented environment of legacy hospital IT systems is fraught with integration issues. The design deliberately separates the inference from centralized networks, to eliminate latency. But securely arranging periodic weight updates to the model or dynamically changing the fitness function for thousands of individual bedside monitors without the cloud dependency is an operational challenge that is still to be addressed. This conflict between on-site execution and synchronized model management is a necessary step to broaden clinical use.

CONCLUSION

To overcome such computational and interpretation challenges in deep sequence modelling for health informatics, this study develops an integrated architecture, which combines generalized graph theory with a Topology-Preserving Genetic Algorithm (TP-GA). The TP-GA deterministically extracts mathematically optimal anomalous subgraphs from a group of non-stationary physiological telemetry, before the deep neural evaluation. A novel Tensor Interface Layer is introduced which transforms these dimensionally reduced patterns to compact input tensors, thereby bypassing the quadratic complexity that is inherent to conventional self-attention architectures.

The framework was comprehensively validated with both low-frequency clinical data (MIMIC-III) and high-frequency univariate signals (MIT-BIH Arrhythmia Database), yielding Area Under the Precision-Recall Curve (AUPRC) scores of 0.952 and 0.974, respectively. The architecture has a footprint of just 2.8 million active parameters and end-to-end latencies of under 19 milliseconds, making it a viable option for localized processing on high-end embedded devices and low-end bedside edge devices. Most importantly, the topological bounding allows the clinician to add structure to an almost always complex prediction of deep learning, directly relating the model to verifiable physiological events, which are themselves discrete and interpretable.

Future work should focus on making the transition from static graph topologies to dynamic and evolving graphs where the transition probabilities would continually change as new patient phenotypes emerged. Further, the Tensor Interface Layer will be designed to support multimodal data fusion, where structured telemetry data is combined with unstructured data, such as Electronic Health Records data, through Natural Language Processing (NLP) to enable even more sophisticated real-time, interpretable clinical intelligence systems.

REFERENCES

1. Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C., & Petersson, L. (2021). Graph-based deep learning for medical diagnosis and analysis: Past, present and future. *Sensors*, 21(14), 4758. https://doi.org/10.3390/s21144758
2. Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, 29, 3504–3512.
3. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29. https://doi.org/10.1038/s41591-018-0316-z
4. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, 33(4), 917-963. https://doi.org/10.1007/s10618-019-00619-1
5. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), 1-9. https://doi.org/10.1038/sdata.2016.35
6. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980-2988. https://doi.org/10.1109/ICCV.2017.324
7. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
8. Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3), 45-50. https://doi.org/10.1109/51.932724
9. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. C. (2004). Mining sequential patterns by pattern-growth: The PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11), 1424-1440. https://doi.org/10.1109/TKDE.2004.77
10. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

11. Slowik, A., & Kwasnicka, H. (2020). Evolutionary algorithms and their applications to engineering problems. *Neural Computing and Applications*, 32(16), 12363-12379. https://doi.org/10.1007/s00521-020-04832-8
12. Solares, J. R. A., Raimondi, F. E. D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., ... & Salimi-Khorshidi, G. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics*, 101, 103337. https://doi.org/10.1016/j.jbi.2019.103337
13. Takens, F. (1981). Detecting strange attractors in turbulence. *Lecture Notes in Mathematics*, 898, 366–381. https://doi.org/10.1007/BFb0091924
14. Tipirneni, S., & Reddy, C. K. (2022). Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6), 1-17. https://doi.org/10.1145/3516367
15. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Machine Learning for Healthcare Conference*, 359-380. PMLR.
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.