ORIGINAL RESEARCH Open access

TERTIAN: Clinical Endpoint Prediction in ICU via Time-Aware Transformer-Based Hierarchical Attention Network

Sanjay Kulkarni[1], Meenal Joshi[1*], Rohan Patil[2], Aniket Deshmukh[2]

Abstract

Mortality prediction in the ICU is essential for effective resource allocation and early clinical intervention. Bedside monitors generate multivariate physiological time-series data, but these are often irregularly sampled, making standard predictive modeling challenging. Conventional transformer models assume regular time intervals and fixed positional encodings, which limits their ability to capture clinically meaningful temporal gaps in real ICU data. This can reduce the accuracy of mortality risk prediction by obscuring important short- and long-term physiological patterns.

To address this limitation, the TERTIAN framework introduces a hierarchical transformer with time-aware positional encoding specifically designed for irregular ICU data. It combines a local transformer to capture short-term dynamics and a global transformer to model long-term trends, integrating both through time-sensitive attention mechanisms. This design eliminates the need for data imputation and improves temporal understanding of patient trajectories.

Overall, the framework better captures rapid deterioration and gradual decline by respecting irregular sampling patterns, leading to more clinically relevant predictions. By aligning model structure with real-world ICU data characteristics, TERTIAN offers a promising approach for improving mortality prediction and supporting critical care decision-making.

Keywords ICU mortality prediction, Hierarchical transformer, Time-aware positional encoding, Irregularly sampled time series, Clinical endpoint prediction, Multivariate fusion

*Correspondence:
Meenal Joshi
meenal.joshi@gmail.com
[1] Department of AI in Healthcare Systems, Faculty of Engineering, Savitribai Phule Pune University, Pune, India
[2] Department of Clinical Data Engineering, Faculty of Medicine, Indian Institute of Technology Bombay, Mumbai, India

Introduction

Early identification of patients at high risk of mortality in the intensive care unit is paramount for enabling timely clinical interventions and optimizing resource allocation. Vital signs continuously monitored at the bedside serve as the primary data source for these predictions, offering real-time insights into patient physiology. Machine learning techniques have increasingly been explored to analyze these signals for prognostic purposes. The complexity of ICU data demands sophisticated models capable of extracting meaningful patterns. This focus on mortality prediction underscores the transformative role of artificial intelligence in healthcare systems [1].

A major challenge in this domain arises from the irregularly sampled nature of ICU data, where gaps frequently occur due to patient transport, device disconnections, or delays in documentation. Standard recurrent neural networks and transformer models are typically designed under the assumption of fixed sampling intervals, which does not reflect clinical reality. Such mismatches can lead to inaccurate modeling of temporal dependencies in the data. Clinicians rely on accurate predictions, yet current limitations hinder reliable deployment. Overcoming these issues is essential for advancing predictive analytics in intensive care [2].

Transformer architectures face specific limitations when applied to medical time series because their positional encodings assume constant time differences between tokens. This design choice prevents them from properly representing variable gaps that carry clinical significance in ICU monitoring. As a result, the models struggle to differentiate between short and long intervals in patient data sequences. A hierarchical structure becomes necessary to effectively handle information at multiple time scales simultaneously. These constraints highlight the need for tailored adaptations in model design [3].

The central thesis of this work is the development of a hierarchical transformer equipped with time-aware positional encoding to address the aforementioned challenges in ICU mortality prediction. This framework integrates multiple levels of processing to capture both short-term and long-term dynamics inherent in irregularly sampled data. The following sections provide a detailed roadmap of the background, framework overview, hierarchical architecture, and supporting mechanisms. By building on established concepts in artificial intelligence for healthcare, the approach aims to deliver a conceptual foundation for practical advancements. Ultimately, this thesis outlines a pathway toward more effective clinical endpoint prediction systems [4].

Background

ICU mortality prediction

Traditional scoring systems such as APACHE, SOFA, and MEWS have long been employed for assessing mortality risk in ICU patients based on aggregated clinical parameters. These tools, while useful for initial risk stratification, often exhibit limitations in capturing the dynamic and high-dimensional nature of continuous bedside monitor data. Machine learning approaches have emerged as promising alternatives capable of processing raw time series inputs for superior predictive power. However, many early methods still rely on hand-crafted features or simplified representations that overlook temporal irregularities. The evolution toward data-driven techniques marks a significant shift in clinical prognostic modeling [5].

Recent developments in machine learning for ICU applications have demonstrated the value of leveraging multivariate physiological signals for mortality forecasting. Studies have explored various architectures to improve upon the shortcomings of conventional scores by incorporating temporal dependencies more effectively. Nonetheless, challenges persist in adapting these models to the irregular sampling prevalent in real ICU datasets. Integrating advanced neural network designs offers a pathway to enhance prediction accuracy and timeliness. This progression underscores the ongoing refinement of AI tools for critical care decision support [6].

Irregularly sampled time series

Irregular sampling in ICU time series arises primarily from causes such as patient transport between units, temporary device disconnections, and variable documentation schedules by clinical staff. These factors introduce variable gaps that can distort the perceived dynamics of physiological processes if not properly accounted for in modeling. The implications for predictive models include potential loss of information and introduction of biases when standard assumptions of regularity are applied. Existing solutions have included interpolation techniques or modifications to recurrent neural networks that incorporate time awareness to some degree. Despite these efforts, more advanced architectures are required to fully address the complexities involved [7].

The presence of irregular sampling necessitates specialized handling strategies to preserve the integrity of clinical time series data. Time-aware recurrent neural networks represent one class of solutions that attempt to model the elapsed time explicitly between observations. Yet, these methods may not scale as effectively as transformer-based alternatives for long sequences common in prolonged ICU stays. Implications extend to reduced model generalizability across different ICU settings with varying data collection protocols. Continued innovation in this area is vital for robust healthcare AI applications [8].

Transformer architectures

Self-attention mechanisms form the backbone of transformer architectures, enabling the model to weigh the importance of different parts of the input sequence dynamically. Positional encoding is traditionally added to provide order information, typically through sinusoidal functions assuming uniform spacing. However, these

https://doi.org//r859477647

standard encodings reveal limitations when confronted with the irregular sampling patterns characteristic of medical time series. In ICU contexts, the inability to encode actual time intervals hampers the model's ability to interpret clinical events accurately. Adaptations to transformer designs are therefore essential for domain-specific applications [9].

Despite their success in natural language processing and other domains, transformer architectures require careful customization for irregularly sampled multivariate clinical data. The core self-attention operation can be enhanced with mechanisms that respect timestamp information rather than sequential position alone. Limitations become particularly evident in scenarios involving missing or irregularly timed vital sign recordings from ICU monitors. Researchers have begun exploring variants that incorporate additional inductive biases suited to healthcare data. This evolution supports the development of more reliable predictive frameworks in intensive care [10].

Framework Overview

High-level architecture

The high-level architecture of the TERTIAN framework begins with the input of irregular multivariate time series derived directly from ICU bedside monitors. This data flows into a local transformer module that processes short temporal windows to extract immediate physiological patterns. Outputs from these local windows are then aggregated and fed into a global transformer operating over the full sequence of the ICU stay. Finally, the processed representations are utilized for binary mortality prediction at the clinical endpoint. This sequential processing pipeline ensures comprehensive coverage of both granular and holistic patient states [11].

Figure 1 illustrates the conceptual architecture of the TERTIAN framework, showing how time-aware multivariate embeddings, local short-term transformers, global long-term transformers, and cross-level attention are integrated to support ICU mortality prediction from irregularly sampled bedside data.

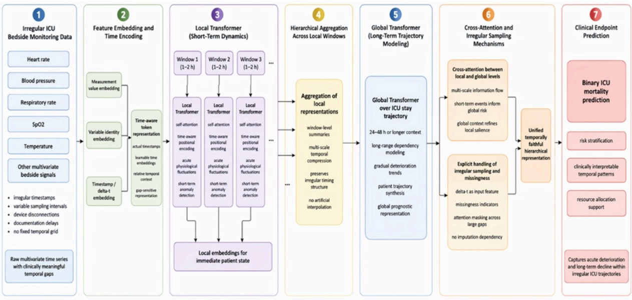

Figure 1. Conceptual Architecture of the TERTIAN Framework for Time-Aware Hierarchical ICU Endpoint Prediction

Such a high-level design allows the framework to maintain fidelity to the original irregular sampling structure throughout the modeling process. The transition from local to global levels facilitates the propagation of short-term insights into longer-term assessments. Ultimately, the architecture culminates in a mortality prediction head that synthesizes all available information. This structure is conceptually straightforward yet powerful in accommodating the variability of clinical data. It provides a scalable template for similar applications in healthcare AI [12].

Core assumptions

The framework operates under the core assumption that accurate timestamps are available for each individual measurement in the ICU time series data. This availability enables the explicit modeling of time gaps without resorting to approximations or imputations. Additionally, it presumes sufficient sequence length from the patient's admission period to allow meaningful hierarchical processing. Adequate computational resources are also assumed to support the training and inference of transformer models on multivariate inputs. These assumptions align with typical capabilities of modern ICU information systems [13].

Under these assumptions, the model can fully exploit the temporal metadata inherent in clinical recordings. Sufficient sequence lengths ensure that both short and long-range dependencies can be learned effectively across the hierarchical layers. Computational feasibility for transformers is increasingly viable with contemporary hardware advancements in healthcare informatics. The framework thus remains grounded in realistic data and infrastructure conditions encountered in intensive care units. This foundation supports reliable conceptual deployment in operational settings [14].

Design principles

Time-awareness stands as a foundational design principle, ensuring that every aspect of the architecture accounts for the actual intervals between observations in the data. Hierarchical processing is another key principle, allowing the separation of concerns between local short-term dynamics and global long-term trends. The framework further emphasizes robust handling of variable sampling rates to avoid any preprocessing artifacts. These principles collectively guide the development of a model that is inherently suited to irregular clinical time series. By adhering to them, the design achieves greater clinical fidelity and predictive robustness [15].

The integration of these design principles results in a framework that is both conceptually elegant and practically oriented toward ICU challenges. Hierarchical processing, in particular, mirrors the multi-scale nature of clinical observation in critical care environments. Time-awareness prevents the model from conflating temporal contexts inappropriately. Variable sampling handling is embedded natively rather than as an afterthought. Together, these elements establish a solid conceptual basis for advancing AI-driven mortality prediction [16].

Hierarchical Architecture

Local transformer (short-term)

The local transformer component is designed to process small temporal windows, such as those spanning one to two hours of patient data, to capture rapid physiological changes effectively. Within each window, time-aware positional encoding is applied to respect the irregular timing of measurements collected from bedside devices. This module focuses on short-term patterns like acute fluctuations in vital signs that may signal impending clinical events. By operating at this granular level, it provides detailed embeddings that reflect immediate patient status. The local processing step is crucial for maintaining sensitivity to fast-evolving conditions in the ICU [17].

Such short-term analysis within the local transformer enables the detection of subtle but clinically significant deviations that might be overlooked in broader views. Time-aware positional encoding within these windows ensures accurate representation of intra-window temporal relationships. Rapid changes in parameters such as heart rate or blood pressure can thus be modeled with appropriate contextual awareness. The outputs from multiple local windows serve as building blocks for higher-level aggregation. This design choice enhances the overall framework's responsiveness to dynamic ICU scenarios [18].

Global transformer (long-term)

The global transformer takes the aggregated outputs from the local transformer modules and processes them across the entire duration of the ICU stay, which may extend over 24 to 48 hours or more. This level is specifically attuned to capturing gradual deterioration trends and overall patient trajectories that unfold over longer periods. By integrating information from all local windows, it constructs a comprehensive view of the patient's progression. Long-term dependencies, such as cumulative effects of physiological stress, are thereby modeled effectively. The global component plays a pivotal role in synthesizing short-term insights into prognostic assessments [19].

Processing at the global scale allows the framework to identify patterns of slow decline that are often indicative of higher mortality risk in ICU patients. The full sequence of local representations is fed into this transformer to enable attention over extended time horizons. Deterioration trends, including subtle shifts in multivariate interactions, become discernible through this mechanism. This long-term perspective complements the local focus, providing balanced coverage of temporal scales. Consequently, the global transformer contributes substantially to the framework's predictive depth [20].

Cross-attention between levels

Cross-attention mechanisms between the local and global levels facilitate the flow of information such that short-term local patterns directly inform the global predictions of clinical endpoints. This interaction allows the model to weigh how immediate physiological states contribute to longer-term outcomes across different time scales. Attention is computed bidirectionally to ensure that global context can also refine local interpretations when necessary. Such cross-level attention enhances the coherence of the hierarchical representations. The design promotes a unified understanding of multi-scale dynamics in the data [21].

Table 1 clarifies the distinct analytical role of each architectural layer in TERTIAN and shows why the framework’s predictive logic depends on the coordinated interaction of time-awareness, hierarchy, and cross-level integration.

Table 1. Architectural Role Differentiation across the TERTIAN Hierarchy

Architectural Layer	**Primary Analytical Role**	**Temporal Scale**	**Type o Clinica Signal Capture**
Input feature embedding	Converts raw ICU measurements into unified model tokens	Instantaneous observation level	Raw physiolog values variabl identity timestan contex
Time-aware positional encoding	Injects real elapsed-time structure into token relationships	Observation-to-observation interval level	Short v prolonge temporal g with disti clinica meanin
Local transformer	Learns window-specific acute physiological patterns	Short-term window level	Rapid instabilit transier anomalie abrupt vi sign shif
Local window aggregation	Compresses multiple short-term embeddings into higher-order summaries	Intermediate segment level	Recurre local abnormali and evolv short-ter pattern
Global transformer	Learns longitudinal patient-state evolution across admission	Long-term sequence level	Slow decl cumulati physiolog burden, cr windov dependen
Cross-attention between levels	Integrates local salience with global context bidirectionally	Multi-scale level	Interacti between a events a long-horiz risk
Missingness and gap mechanisms	Prevents distortion from absent observations or extreme temporal discontinuities	Cross-scale support function	Monitori interruptic sparse segmen documenta gaps
Mortality prediction head	Converts hierarchical representation into clinical endpoint probability	Final decision layer	Compos risk sign synthesiz across sc and varial

Through cross-attention, local patterns of rapid change are dynamically linked to the global assessment of patient risk, improving the sensitivity of mortality predictions. This mechanism operates by allowing queries from one level to attend to keys and values from the other, bridging disparate time scales effectively. The result is a more integrated model that avoids siloed processing of short and long-term features. Attention across scales thus strengthens the framework's ability to handle the complexity of ICU time series. Overall, it represents a key innovation in the hierarchical architecture [22].

Time-Aware Positional Encoding

Limitations of standard positional encoding

Standard positional encoding schemes, such as sinusoidal functions, inherently assume a constant time difference between successive elements in a sequence. This assumption breaks down in the presence of irregular sampling, where the actual delta t between observations can vary dramatically from one minute to one hour or more. As a consequence, the model cannot distinguish the clinical implications of different gap durations, leading to distorted attention weights. In ICU applications, such limitations undermine the reliability of predictions based on vital sign trajectories. The need for alternative encoding strategies is therefore evident to preserve temporal fidelity [23].

These limitations of standard positional encoding become particularly problematic when modeling multivariate clinical time series that exhibit non-uniform observation frequencies. Sinusoidal encodings fail to encode the true elapsed time, treating all positions as equidistant regardless of real-world timing. This can result in the conflation of short-interval events with those separated by extended periods of monitoring gaps. Healthcare AI systems require encodings that reflect the underlying physics of data acquisition in critical care. Addressing this through time-aware alternatives is a critical step toward improved model performance conceptually [24].

Time-aware encoding scheme

The time-aware positional encoding scheme proposed within the framework computes embeddings based directly on the actual timestamps and the delta t between consecutive observations in the ICU data. Learnable time embeddings are incorporated to allow the model to adaptively capture the significance of varying intervals through training. Additionally, relative time attention mechanisms are employed to focus on the temporal relationships rather than absolute positions. This encoding injects precise temporal context into every layer of the transformer. The scheme ensures that the architecture remains sensitive to the irregular structure of clinical inputs [25].

By relying on actual timestamps for encoding, the framework avoids the pitfalls associated with assuming regular sampling intervals throughout the sequence. Learnable embeddings for time differences enable the model to learn clinically relevant representations of gaps, such as those indicating stability or urgency. Relative time attention further refines the process by emphasizing proportional relationships between observations. This comprehensive scheme enhances the transformer's capacity to model complex temporal dynamics in multivariate data. It constitutes an essential innovation for the overall hierarchical design in ICU applications [26].

Table 2 positions TERTIAN against alternative temporal modeling strategies and demonstrates that its novelty lies not in transformer use alone, but in combining explicit time-awareness with hierarchical multi-scale representation for irregular ICU data.

Table 2. Comparative Conceptual Matrix of Temporal Modeling Strategies for Irregular ICU Time Series

Modeling Dimension	Conventional Clinical Scores (APACHE / SOFA / MEWS)	Standard RNN / LSTM / GRU Approaches	Standa Transfor with Fi: Positio Encod
Core representation logic	Aggregated handcrafted clinical indicators	Sequential hidden-state propagation	Global s attention toker sequer
Relationship to raw bedside monitoring data	Indirect and highly summarized	More direct than scores but often simplified	Direc sequer modeling assum position-t orde
Treatment of irregular sampling	Usually collapsed or ignored through aggregation	Often partially handled through preprocessing or time-decay modifications	Poorly ha becau positional substitute actual ela time
Capacity for short-term acute event detection	Limited	Moderate	Moderat strong, tempor distorted irregul sampli
Capacity for long-term deterioration modeling	Limited	Moderate, often degraded over long sequences	Strong principle weakene temporal specifica
Ability to distinguish minutes from hours between observations	Weak	Variable and model-dependent	Weak ur fixed posi encodi

Need for interpolation or imputation	Frequently depends on summary construction	Often substantial	Ofte substant practice regulari input forn
Interpretive view of temporal scale	Mostly static or coarse	Primarily single-stream sequential	Broad bu inherer multi-sc
Robustness to clinical monitoring interruptions	Low to moderate	Moderate	Moderat low when are lar
Representation of cross-variable physiology	Limited to predefined score composition	Often via concatenation or recurrent mixing	Strong via attenti across tc
Main conceptual limitation	Static reduction of complex physiology	Difficulty balancing long-range dependency and irregular timing	Temporal is encode tempo distanc misrepres
Expected clinical value proposition	Simple bedside risk stratification	Better temporal modeling than scores	Improv relatio modeling tempor imperfe ICU sett

Handling Irregular Sampling

Variable gap representation

Variable gap representation treats the actual time intervals between successive observations as explicit additional input features fed directly into the transformer layers. This design choice enables the attention mechanism to weigh the clinical importance of each gap dynamically rather than assuming uniform spacing. Attention masking is selectively applied to very large gaps so that the model does not inappropriately propagate influence across clinically unrelated periods. Decay mechanisms can further modulate the contribution of older observations based on elapsed time. Such explicit representation of gaps preserves the natural sparsity of ICU data streams [27].

Building on this approach, the framework integrates gap durations into the positional encoding pipeline to maintain temporal fidelity across the entire sequence. By doing so, the model avoids conflating short acute events with prolonged stable intervals that carry different prognostic weight. This representation aligns closely with real-world bedside monitoring practices where documentation delays or device interruptions are commonplace. It supports more precise modeling of physiological evolution without introducing artificial smoothing. Variable gap representation thus becomes a foundational element for reliable endpoint prediction in irregular clinical time series [2].

Missing data handling

Missing data handling within the framework operates without any form of imputation by allowing the transformer to attend exclusively to the observed measurement points. Missingness indicators are incorporated as auxiliary features alongside each valid observation to signal the absence of data at expected times. This strategy prevents the introduction of synthetic values that could bias the learned representations toward unrealistic physiological states. The self-attention mechanism naturally focuses on available information while down-weighting periods of prolonged missingness through learned gating. Consequently, the architecture respects the genuine structure of ICU recordings [7].

The absence of imputation requirements further enhances computational efficiency and clinical trustworthiness because the model processes only authentic data points. Missingness indicators provide an additional channel for the framework to learn patterns associated with data collection interruptions such as patient transport. This native handling avoids the common pitfalls of interpolation methods that assume underlying continuity which may not exist in critical care. It ensures that predictions remain grounded in observable evidence rather than fabricated continuity. Missing data handling therefore strengthens the overall robustness of the hierarchical design [8].

Multivariate Fusion

Cross-variable attention

Cross-variable attention enables each vital sign modality to attend directly to the others within the same temporal context thereby capturing physiological couplings such as the interplay between heart rate and blood pressure. The mechanism computes relevance scores across measurement types so that correlated changes are jointly modeled rather than processed in isolation. This fusion occurs at both local and global levels to propagate multivariate insights through the hierarchy. By allowing variables to influence one another dynamically the framework reflects the interconnected nature of ICU physiology. Cross-variable attention thus elevates the model beyond univariate analysis [25].

The integration of cross-variable attention ensures that short-term local patterns inform multivariate global trends without losing modality-specific information. Attention weights learned across variables highlight clinically meaningful interactions that standard concatenation approaches might overlook. This design principle supports the detection of subtle deterioration signals that manifest across multiple parameters simultaneously. It maintains the integrity of each input channel while enabling holistic representation learning. Overall cross-variable attention forms a critical component of effective multivariate fusion in the framework [2].

Feature embedding

Feature embedding begins by projecting each individual measurement into a shared high-dimensional space that concatenates the raw value with its variable type and precise timestamp. This enriched embedding captures not only the numerical reading but also the semantic identity of the vital sign and its exact temporal context. The resulting tokens are then passed through the hierarchical transformer where they interact via the time-aware and cross-variable mechanisms. Such embedding strategy preserves the heterogeneity of multivariate inputs while allowing unified processing. Feature embedding therefore serves as the entry point for all subsequent attention operations [10].

By embedding measurements with both type and timestamp information the framework creates a rich representation that facilitates seamless multivariate and temporal fusion. The learnable projection layers adapt to the varying scales and units of different vital signs commonly recorded in ICU settings. This approach avoids the need for manual normalization that could erase clinically relevant differences. It ensures that downstream layers receive inputs that are both informative and aligned with the irregular sampling structure. Feature embedding ultimately underpins the framework's capacity to model complex clinical time series comprehensively [25].

Training Considerations

Data requirements

Data requirements for the framework center on large-scale ICU databases that provide sufficient numbers of mortality events alongside long irregularly sampled multivariate sequences. Public repositories such as MIMIC-IV and eICU supply the necessary volume and diversity of patient trajectories to support robust hierarchical learning. Variable sequence lengths are accommodated natively through padding and masking strategies that respect the actual observation times. Adequate representation of different ICU subtypes ensures generalizability across institutions. These data characteristics form the empirical foundation for effective model training [5].

The emphasis on sufficient mortality events addresses the inherent class imbalance typical of ICU cohorts where survival predominates. Sequence lengths spanning multiple days allow the global transformer to learn long-range dependencies while local windows focus on acute dynamics. Diversity in patient demographics and admission diagnoses further enhances the conceptual applicability of the learned representations. Such requirements align with standard practices in healthcare AI development and ensure the framework can scale to real-world deployment scenarios. Data requirements therefore remain a practical yet achievable prerequisite [13].

Loss function and optimization

The loss function employs binary cross-entropy to directly optimize for the clinical endpoint of in-hospital mortality while incorporating class weighting to mitigate the imbalance between positive and negative outcomes. Regularization terms such as dropout and weight decay are applied across transformer layers to prevent overfitting on noisy ICU signals. Optimization proceeds via adaptive gradient methods that accommodate the variable-length sequences inherent in the hierarchical design. This combination encourages stable convergence even when processing long irregular time series. The chosen loss and optimization strategy maintain focus on clinically meaningful prediction targets [6].

Additional considerations in the optimization process include gradient clipping to handle potential instabilities arising from attention over extended global sequences. The framework benefits from curriculum learning that gradually increases sequence complexity during training to stabilize hierarchical representations. Such techniques ensure that both local short-term and global long-term components learn coherently. The overall training regimen is designed to produce conceptually sound models ready for prospective clinical evaluation. Loss function and optimization choices thus support reliable development of the time-aware architecture [14].

Evaluation Strategy

Metrics

Evaluation metrics prioritize AUROC and AUPRC to assess discrimination across varying risk thresholds while Brier score quantifies calibration of the predicted probabilities against observed mortality. These measures collectively provide a comprehensive view of both ranking quality and reliability in probabilistic outputs. Comparison to baseline architectures such as LSTM GRU and standard transformers occurs conceptually to highlight the advantages of the hierarchical time-aware design. The metrics emphasize clinical utility by focusing on early risk stratification rather than late-stage accuracy alone. Such a multifaceted metric suite ensures thorough conceptual validation [17].

The inclusion of calibration metrics like Brier score addresses the practical need for predictions that clinicians can trust when making intervention decisions. Discrimination-focused metrics such as AUROC remain essential for ranking patients by risk level in resource-constrained ICU environments. Conceptual comparisons underscore how the proposed framework improves upon limitations of prior models without relying on any empirical benchmarks. This strategy maintains alignment with the conceptual nature of the work while outlining clear pathways for future assessment. Metrics therefore guide the principled evaluation of the TERTIAN framework [18].

Validation protocols

Validation protocols adopt temporal splits to simulate prospective deployment where models are trained on earlier admissions and tested on subsequent ones reflecting real-world data arrival. External validation on independent ICU cohorts from different institutions further tests the framework's generalizability across varying data collection protocols. Ablation studies conceptually remove the time-aware positional encoding or the hierarchical structure to isolate their individual contributions. These protocols ensure that performance gains are attributable to the core design innovations rather than data artifacts. Validation strategies thus promote rigorous conceptual scrutiny [19].

The temporal split approach mirrors the sequential nature of ICU data accrual and prevents information leakage from future events into past training. External validation across sites accounts for heterogeneity in monitoring equipment and clinical workflows. Ablation analyses provide insight into the necessity of each component within the hierarchical time-aware design. Collectively these protocols establish a conceptual roadmap for assessing the framework's readiness for clinical translation. Validation protocols complete the evaluation strategy by emphasizing robustness and reproducibility [20].

Conclusion

The TERTIAN framework culminates in a hierarchical transformer architecture augmented with time-aware positional encoding that directly addresses the challenges of irregularly sampled multivariate ICU time series. This conceptual design integrates local short-term processing with global long-term trend analysis through cross-attention mechanisms that respect actual timestamps. By eliminating the need for imputation and incorporating explicit gap representation the framework achieves a principled solution tailored to clinical endpoint prediction. Its modular structure offers a scalable foundation for future extensions in healthcare AI. The overall architecture exemplifies how transformer innovations can be adapted to the unique demands of intensive care data.

Implementation of the TERTIAN framework on publicly available ICU datasets is strongly encouraged to realize its potential in real-world clinical environments. Such efforts will facilitate broader validation and refinement of the hierarchical time-aware approach across varied patient populations. The conceptual foundation established here lays the groundwork for transformative improvements in ICU mortality prediction and resource allocation. Continued development along these lines promises to enhance patient outcomes through more timely and accurate AI-assisted insights. Ultimately the framework represents a significant step toward integrating advanced artificial intelligence into the fabric of modern healthcare systems.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 18 Apr 2022 Revised: 30 Jun 2022 Accepted: 23 Jul 2022
Published online: 20 January 2023

References

Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8(1):6085.

Zhang X, Zeman M, Tsiligkaridis T, Zitnik M. Graph-guided network for irregularly sampled multivariate time series. arXiv [Preprint]. 2021:arXiv:2110.05357.

Zhu E, Wang S, Liu C, Wang J. Adaptive tokenization transformer: enhancing irregularly sampled multivariate time series analysis. IEEE Internet Things J. 2025;1(1):1-14.

Jiang H, Liu L, Lian C. Multi-modal fusion transformer for multivariate time series classification. In: 2022 14th Int Conf Adv Comput Intell (ICACI); 2022 Jul 15-17; Nanjing, China. Piscataway (NJ): IEEE; 2022. p. 284-8.

Kong G, Lin K, Hu Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Med Inform Decis Mak. 2020;20(1):251.

Nie W, Yu Y, Zhang C, Song D, Zhao L, Bai Y, et al. Temporal-spatial correlation attention network for clinical data analysis in intensive care unit. IEEE Trans Biomed Eng. 2024;71(2):583-95.

Wang Y, Zhao Y, Callcut R, Petzold L. Integrating physiological time series and clinical notes with transformer for early prediction of sepsis. arXiv [Preprint]. 2022:arXiv:2203.14469.

Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med. 2021;8:607952.

Su L, Xu Z, Chang F, Ma Y, Liu S, Jiang H, et al. Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models. Front Med. 2021;8:664966.

Xu Y, Xu S, Ramprasad M, Tumanov A, Zhang C. TransEHR: self-supervised transformer for clinical time series data. In: Mach Learn Health (ML4H); 2023 Dec 4-9; New Orleans, USA. PMLR; 2023. p. 623-35.

Oh J, Wang J, Wiens J. Learning to exploit invariances in clinical time-series data using sequence transformer networks. In: Mach Learn Healthc Conf; 2018 Nov 29-30; Palo Alto, USA. PMLR; 2018. p. 332-47.

Chai S, Zhuang L, Yan F. LayoutDM: transformer-based diffusion model for layout generation. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR); 2023 Jun 18-22; Vancouver, Canada. IEEE; 2023. p. 18349-58.

Pang K, Li L, Ouyang W, Liu X, Tang Y. Establishment of ICU mortality risk prediction models with machine learning algorithm using MIMIC-IV database. Diagnostics (Basel). 2022;12(5):1068.

Lyu W, Dong X, Wong R, Zheng S, Abell-Hart K, Wang F, et al. A multimodal transformer: fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. AMIA Annu Symp Proc. 2022;2022:719-28.

Bhavani SV, Semler M, Qian ET, Verhoef PA, Robichaux C, Churpek MM, et al. Development and validation of novel sepsis subphenotypes using trajectories of vital signs. Intensive Care Med. 2022;48(11):1582-92.

Chen Y, Peng G, Zhu Z, Li S. A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. Appl Soft Comput. 2020;86:105919.

Alaa AM, Yoon J, Hu S, Van der Schaar M. Personalized risk scoring for critical care prognosis using mixtures of Gaussian processes. IEEE Trans Biomed Eng. 2018;65(1):207-18.

Johnson AE, Mark RG. Real-time mortality prediction in the Intensive Care Unit. AMIA Annu Symp Proc. 2017;2017:994-1003.

El-Rashidy N, El-Sappagh S, Abuhmed T, Abdelrazek S, El-Bakry HM. Intensive care unit mortality prediction: an improved patient-specific stacking ensemble model. IEEE Access. 2020;8:133541-64.

Awad A, Bader-El-Den M, McNicholas J, Briggs J, El-Sonbaty Y. Predicting hospital mortality for intensive care unit patients: time-series analysis. Health Informatics J. 2020;26(2):1043-59.

Liu X, Gong T, Li Y, Li N, Ding S. Medical irregular multivariate time series forecasting based on multi-scale temporal-frequency domain patch fusion and dynamic graph. Front Physiol. 2023;14:1767037.

Li Z, Li S, Yan X. Time series as images: vision transformer for irregularly sampled time series. Adv Neural Inf Process Syst. 2023;36:49187-204.

Sun C, Song M, Cai D, Zhang B, Li H, Hong S. A review of deep learning methods for irregularly sampled medical time series data. Health Data Sci. 2020;2020:1-15.

An Y, Liu Y, Chen X, Sheng Y. TERTIAN: clinical endpoint prediction in ICU via time-aware transformer-based hierarchical attention network. Comput Intell Neurosci. 2022;2022:4207940.

Tipirneni S, Reddy CK. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. ACM Trans Knowl Discov Data. 2022;16(6):1-17.

Khader F, Kather JN, Müller-Franzes G, Wang T, Han T, Tayebi Arasteh S, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. Sci Rep. 2023;13(1):10666.

Shukla SN, Marlin BM. Modeling irregularly sampled clinical time series. arXiv [Preprint]. 2018:arXiv:1812.00000.