Temporal Concept Tracing: Making Deep Learning Predictions Interpretable and Actionable for ICU Acute Kidney Injury Prevention

S M Saiful Islam Badhon[1], Serdar Bozdag[2], Mohammad Adibuzzaman[3], Ana D. Cleveland[1], Junhua Ding[4], K S M Tozammel Hossain[4]

[1]Department of Information Science, University of North Texas, Denton, TX
[2]Department of Computer Science and Engineering, University of North Texas, Denton, TX
[3]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR
[4]Anuradha and Vikas Sinha Department of Data Science, University of North Texas, Denton, TX
smsaifulislambadhon@my.unt.edu, Serdar.Bozdag@unt.edu, adibuzza@ohsu.edu, ana.cleveland@unt.edu, Junhua.Ding@unt.edu, Tozammel.Hossain@unt.edu

Abstract

Deep learning models have demonstrated impressive accuracy in predicting acute kidney injury (AKI), a condition affecting up to 20% of ICU patients, yet their black-box nature prevents clinical adoption in high-stakes critical care settings. While existing interpretability methods like SHAP, LIME, and attention mechanisms can identify important features, they fail to capture the temporal dynamics essential for clinical decision-making, and are unable to communicate when specific risk factors become critical in a patient's trajectory. This limitation is particularly problematic in the ICU, where the timing of interventions can significantly impact patient outcomes. We present a novel interpretable framework that brings temporal awareness to deep learning predictions for AKI. Our approach introduces three key innovations: (1) a latent convolutional concept bottleneck that learns clinically meaningful patterns from ICU time-series without requiring manual concept annotation, leveraging Conv1D layers to capture localized temporal patterns like sudden physiological changes; (2) Temporal Concept Tracing (TCT), a gradient-based method that identifies not only which risk factors matter but precisely when they become critical addressing the fundamental question of temporal relevance missing from current XAI techniques; and (3) integration with MedAlpaca to generate structured, time-aware clinical explanations that translate model insights into actionable bedside guidance. We evaluate our framework on MIMIC-IV data, demonstrating that our approach performs better than existing explainability frameworks, Occlusion and LIME, in terms of the comprehensiveness score, sufficiency score, and processing time. The proposed method also better captures risk factors inflection points for patients timelines compared to conventional concept bottleneck methods, including dense layer and attention mechanism. This work represents the first comprehensive solution for interpretable temporal deep learning in critical care that addresses both the what and when of clinical risk factors. By making AKI predictions transparent and temporally contextualized, our framework bridges the gap between model accuracy and clinical utility, offering a path toward trustworthy AI deployment in time-sensitive healthcare settings.

Introduction

The advancement of deep-learning-based predictive methods, particularly Transformer-based Models(Vaswani et al. 2017; Li et al. 2020; Choi et al. 2019), has inspired researchers to apply them to different domains, including healthcare. However, deep learning's black-box nature remains a key barrier to clinical adoption, with concerns about explainability, bias, fairness, and patient safety limiting its real-world use (Tonekaboni et al. 2019). These concerns are especially pronounced in critical care settings, where clinicians must trust that model predictions align with clinical reasoning and reflect genuine risk factors rather than statistical artefacts. The challenge of interpretability becomes even more acute for clinical time-series models, where understanding when specific risk factors become critical is as important as identifying what those factors are.

The AKI syndrome is characterized by a rapid loss in kidney function, with or without kidney damage, that occurs over a few hours or days. While recent deep learning models have achieved impressive accuracy in AKI prediction, their opaque decision-making processes limit adoption in clinical settings where high-stakes decisions demand transparency. Existing approaches to model interpretability, including SHAP(Lundberg and Lee 2017a), LIME(Ribeiro, Singh, and Guestrin 2016), and attention mechanisms(Bahdanau, Cho, and Bengio 2015), can identify important features but fail to capture the temporal dynamics crucial for clinical decision-making. These methods typically produce static saliency maps or feature importance rankings that lack temporal context: they cannot communicate, for instance, that a sudden drop in urine output at hour 12 post-admission carries different clinical significance than the same drop at hour 48. Even when feature importance is available, the outputs often lack the narrative context and domain alignment necessary for bedside decision-making (Lundberg and Lee 2017b). The ideal interpretability framework for AKI prediction in the ICU should therefore:

1. Identify clinically meaningful risk factors without requiring expensive manual annotation.
2. Pinpoint when these factors become critical in the patient's trajectory.

3. Translate these insights into actionable clinical language that supports time-sensitive interventions.

Current methods fail to achieve all three objectives simultaneously, leaving a gap between model predictions and clinical utility. To address these challenges, we propose a novel interpretability framework that brings temporal awareness to deep learning predictions in critical care. Fig. 1 illustrates the complete pipeline, showing how our interpretability framework integrates with black-box temporal models to produce clinically actionable insights.

Our approach centers on a Concept Block(a convolutional neural sub-network) that can be integrated into various temporal predictive architectures during training. Unlike traditional Concept Bottleneck Models that require annotated Concept Labels, our framework learns latent concepts directly from ICU time-series data. We chose convolutional layers over dense networks because they naturally capture localised temporal patterns critical in clinical events (such as sudden creatinine spikes preceding AKI), while maintaining parameter efficiency and feature-channel independence that enables clearer attribution of which physiological variables changed when. We extend this architecture with TCT, a gradient-based technique that not only identifies which risk factors influence predictions but also determines when they become important in the patient's timeline. Finally, we leverage MedAlpaca (Han et al. 2023), a medical language model, to transform TCT outputs into structured, time-aware clinical explanations tailored to each patient's trajectory. Together, these innovations provide a comprehensive solution that makes deep learning predictions both interpretable and actionable at the bedside. The key contributions of this paper are as follows:

- We develop a novel convolutional concept bottleneck framework for explaining temporal deep learning predictions in critical care. Though integration requires training with our Concept Block included, the design is architecture-agnostic, enabling researchers to add interpretability to LSTMs, Transformers, or other temporal models without fundamental architectural constraints.
- We introduce TCT, a gradient-based method that identifies both which risk factors influence AKI risk and when they become critical, going beyond static feature importance to provide temporal insights essential for clinical decision-making.
- We demonstrate the application of large language models for generating time-aware clinical explanations, using MedAlpaca to translate complex model outputs into natural language that aligns with clinical reasoning patterns.

Together, this provides a first-of-its-kind approach for interpretable AKI prediction from ICU time-series, combining model transparency with clinical reasoning grounded in patient data.

Related Work

Concept Bottleneck Models (CBMs) aim to induce intermediate concept predictions, often human-labeled, before the model makes its final decision, enhancing interpretability and enabling concept-level interventions. Koh et al. (Koh et al. 2020) presented the foundational CBM framework and demonstrated that classifiers predicting from concept representations can achieve competitive performance while allowing post-hoc intervention on concepts. Margeloiu et al. (Margeloiu et al. 2021) critically evaluated whether CBMs consistently learn meaningful concepts and highlighted limitations in fidelity and interpretability. Extensions like Post-hoc CBMs have allowed concept-space explanations even without concept annotations during training (Yuksekgonul, Wang, and Zou 2022), and recent works explore robustness to mislabeled concept data (Park et al. 2025).

Hybrid architectures combining Transformers and LSTMs seek to leverage both long-range attention and sequential modeling for healthcare time-series. Shukla and Marlin introduced the Multi-Time Attention Network (MTAN) to handle sparsity and irregular sampling in physiological data by combining time embeddings and attention mechanisms (Shukla and Marlin 2021). Their earlier work also developed interpolation–prediction networks for irregular EHR sequences (Shukla and Marlin 2018). Additionally, fully hierarchical or dual-stream Transformer-RNN designs have been proposed for ICU tasks like mortality and sepsis prediction, demonstrating improved robustness (Xu and Staniek 2025).

Explainability in sequential models is an active research area. Methods like MTAN provide implicit interpretability for sparse ICU sequences (Shukla and Marlin 2021). RETAIN proposed reverse-time attention for healthcare predictions, enabling human-interpretable attention weights (Choi et al. 2016). AttentionViz visualizes model attention to aid clinicians in interpreting time-series predictions (Yeh et al. 2024). TIMEX provides benchmarks for evaluating explainability methods across time-series domains (Liu et al. 2024).

Recent work also explores large language models for clinical reasoning and explainability (Singhal et al. 2023; Li et al. 2025; Mullenbach et al. 2018). On the EHR modeling front, Li et al. (Li et al. 2023) proposed *Hi-BEHRT*, a hierarchical Transformer designed to process very long patient histories.Similarly, Rasmy et al.'s *Med-BERT* (Rasmy et al. 2021) adapted the BERT framework to structured EHR data.

Methodology

We propose an explainability framework tailored for high-dimensional, multivariate time-series models in critical care. Our approach enables temporal reasoning, concept-level summarisation, and natural language clinical justification for predictions made by a black-box AKI prediction model. While the predictive model (based on Transformer and LSTM layers) is developed separately, our focus here is on enhancing model transparency, temporal attribution, and clinician trust in AI-driven decision support systems.

Our proposed model integrates a hybrid Transformer–LSTM framework augmented with a concept-level explainability mechanism. The architecture is designed to effectively capture both global temporal dependencies and localised sequential patterns in multivariate clinical time-

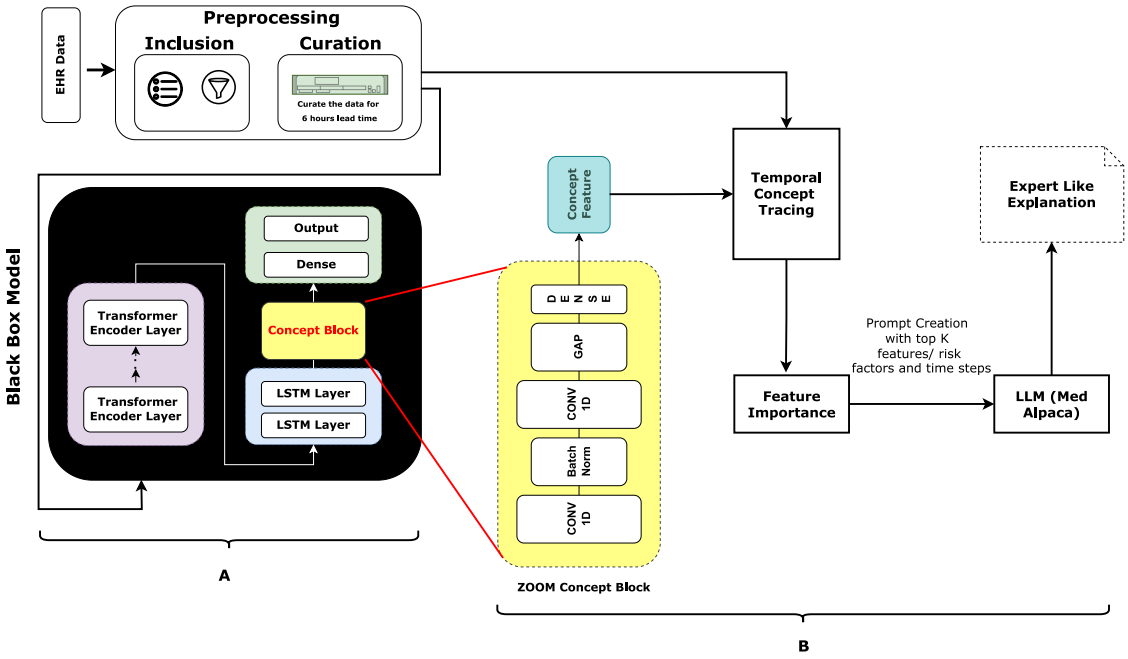

Figure 1: **The proposed framework. A**) Black Box Predictive model for AKI where convolutional concept block can be integrated during training. **B**) Temporal Concept Tracin (TCT), a gradient-based method and the integration of LLM for expart like explanation

series data, with an emphasis on enhancing interpretability through a concept-level representation.

Problem Formulation

We developed a temporal model to anticipate AKI at least six hours before onset, using the most recent 48 hours of patient history. The model exploits sequential clinical measurements and demographic (age,gender) data to predict AKI risk. Each patient's history is represented as a sequential time series of observations of physiological and clinical risk factors collected over the last 48 hours. Each time step includes measurements of several risk factors, including creatinine levels, blood urea nitrogen (BUN), temperature, pH levels, potassium, plateau pressure, hemoglobin, and others. We cast the AKI risk anticipation as a binary classification problem, where the outcome "yes" indicates that the patient is likely to develop AKI after a predefined lead time (e.g., 6, 12, or 24 hours) and the outcome "no" indicates that AKI is unlikely to occur after the lead time.

Let,

$$X = \{X_1, X_2, \ldots, X_T\}$$

be the sequence of a patient's data over T time steps (e.g., 48 hours with each hour as a time step). Here, X_t R^d represents the vector of risk factors at time t, where d is the number of risk factors. Let y_{T+k} element of $\{0, 1\}$ be the target variable, capturing the AKI label after a specified lead time window k. The objective is to learn a function:

$$f : R^{T \times d} \rightarrow \{0, 1\}$$

that maps the input time series X to the binary target y, maximizing predictive accuracy while balancing sensitivity (recall) and specificity (precision). To improve interpretability, we decompose the model into two functions:

Concept Mapping:

$$g : R^{T \times d} \rightarrow R^c$$

where g is a function (neural network block) that transforms the raw input into a latent concept vector. The concept space R^c has size c, where each dimension can represent a high-level clinical pattern, such as "Creatinine rising", "Urine output low", or "stable blood pressure".

Concept-based Classifier:

$$f_c : R^c \rightarrow \{0, 1\}$$

where f_c is a prediction function that uses the concept vector to predict whether AKI will occur. Putting both together:

$$f(X) = f_c(g(X))$$

AKI Predictive Model (Black Box)

The proposed model comprises two stacked Transformer encoder blocks (Vaswani et al. 2017). Each encoder consists of a multi-head self-attention layer, followed by dropout, residual connections, and layer normalization. Figure 1(A) illustrates the architecture of the black-box model. This configuration enables the model to learn attention-based representations across all time steps and features, facilitating long-range dependency modeling. The second Transformer layer

incorporates a skip connection from the first, allowing richer hierarchical temporal abstraction.

The output of the Transformer layers is passed into a stacked Long Short-Term Memory (LSTM) network, consisting of two layers. These LSTMs capture sequential dynamics and preserve the temporal order of clinical events, which is critical for modeling disease progression.

Concept Block

Following the LSTM layers in the black-box model, we introduce a novel Concept Block (See Figure 1B) to project the high-dimensional LSTM outputs into a lower-dimensional, interpretable concept space. The Concept Block consists of two submodules: a. Concept Extraction Module, b. Concept Bottleneck Layer.

a) Concept Extraction Module To extract structured, interpretable patterns from temporal clinical data, we introduce a convolution-based Concept Extraction Module prior to the bottleneck layer. Given an input,

$$X \in R^{T \times F},$$

where T = 48 represents the time steps and F = 52 denotes the number of clinical features, the Concept Block applies temporal convolutions across features to capture local temporal dynamics:

$$Z_1 = \text{ReLU}(\text{Conv1D}_{64}(X)),$$
$$Z_2 = \text{ReLU}(\text{Conv1D}_{32}(Z_1)).$$

We choose convolutional networks to extract localized temporal dynamics such as rising creatinine or transient hypotension common in AKI onset. Conv1D is parameter-efficient, preserves sequence locality, and allows disentangled per-feature abstractions. Unlike attention-based models, its inductive bias supports clearer gradient-based attribution, making it well-suited for integration with TCT.

b) Concept Bottleneck Layer The output Z_2 is passed through batch normalization and globally averaged across the time dimension, yielding a concept vector:

$$C = \text{GAP}(\text{BatchNorm}(Z_2)),$$

where C is element of R^{16}. This vector acts as the concept bottleneck, providing an intermediate, compressed representation of the temporal input before the final binary prediction layer.The Concept Block supports post-hoc explainability via the technique TCT. It acts as a bridge between deep representations and human-understandable features, enabling interpretability without compromising predictive performance.

Temporal Concept Tracing (TCT)

To achieve temporal explainability, we introduce TCT, a gradient-based method that highlights the contribution of specific features at specific time steps toward the model's prediction.

Given a test instance X_{test}, we compute the gradient of the model's output with respect to each input feature at each time step, denoted as:

$$\nabla_{X_{\text{test}}} \hat{y} \in R^{T \times F}$$

We normalize these gradients to obtain an importance heatmap that can be visualized across time (x-axis) and features (y-axis). This allows tracing which features become influential, and when, during the lead-up to AKI onset.

To focus attention on the most influential moments and features, we apply a Top-k filtering strategy:

1. Compute the average absolute gradient across features for each time step.
2. Select the top 5 time steps with highest overall importance.
3. Within each selected time step, identify the top 3 features based on their individual gradient magnitudes.
4. Collect these time-feature pairs for final interpretation.

This process yields a compact set of temporally localized risk factors, which are then associated with actual clinical values to ground the explanation.

Explanation Generation

To generate human-understandable rationales for predictions, we integrate the temporally filtered TCT output with a large language model (LLM) fine-tuned for clinical instruction-following, MedAlpaca-7B(Han et al. 2023).

Prompt Construction Logic We construct a natural language prompt in the following format:

> The model predicted the patient's AKI risk increased starting around hour {t_start}.
> Feature importance over time:
> Hour {t1}: FeatureA (value1), FeatureB (value2), ...
> Hour {t2}: FeatureX (valueX), FeatureY (valueY), ...
> Based on the above, explain why AKI is likely.

This prompt captures temporal dynamics, feature salience, and clinical values in a form interpretable by both humans and LLMs. For low-risk predictions, the prompt template adjusts to state that AKI is not likely and omits specific hours.

Integration with MedAlpaca We use the HuggingFace pipeline API to load `medalpaca/medalpaca-7b` in 4-bit quantization, ensuring feasible inference even on limited GPU memory. The pipeline processes the prompt and returns an explanation written in fluent clinical language.

To enhance credibility, we prepend a context instruction:

> Context: You are a kidney expert. Do not say the model thinks. Instead, explain like a human nephrologist would.

Results

We evaluate the efficacy of the proposed approach by addressing the following research questions:

- **RQ1:** Does the proposed explainability method truly make a difference, and why does it matter?
- **RQ2:** Does the method identify critical time points in a patient's history? How meaningful are the Concept Block representations?
- **RQ3:** Can LLM be used to generate explanation reports?

Datasets

We utilize the publicly available MIMIC-IV v1.0 dataset, developed by MIT and Beth Israel Deaconess Medical Center (Johnson et al. 2023). MIMIC-IV contains de-identified electronic health records (EHR) of ICU and non-ICU patients from 2008 to 2019. As the data is de-identified and collected under HIPAA guidelines, this study qualifies as nonhuman subject research and does not require additional IRB approval (Johnson et al. 2023). For this study, we focus on ICU stays, including 61,735 admissions with high-resolution clinical measurements.

Inclusion Criteria:

- Adult patients (18 years old)
- ICU stays with at least 48 hours of observation
- Availability of core clinical variables (e.g., creatinine, BUN, MAP, urine output)

Exclusion Criteria:

- Patients with end-stage renal disease (ESRD) on admission
- Missing data for key prediction features
- Multiple ICU stays (only the first admission considered)

And after inclusion and exclusion, finally we worked with 18131 patients.

Experimental Setup

We vary different parameters, such as learning rate and batch size, of the black-box model to ensure robustness (see Table **??**). The model maintains high performance for learning rates of 0.001 and 0.0001, with F1-scores of 0.86 and 0.84, respectively. However, a significantly lower F1-score (0.76) is observed at 0.00001, likely due to underfitting or slow convergence. Across batch sizes of 32, 64, and 128, the model performs consistently well, with F1-scores ranging from 0.83 to 0.85. Notably, the batch size of 128 yields the highest Recall (0.86), which is crucial for early AKI detection. These results indicate that the model is relatively stable and robust across a reasonable range of hyperparameters, especially at learning rates 0.0001. This consistency strengthens the model's reliability and suitability for deployment in real-world clinical settings.

Learning Rate	**P**	**R**	**F1**	**Batch Size**	**P**	**R**	**F1**
0.001	0.87	0.85	0.86	**32**	0.86	0.80	0.83
0.0001	0.87	0.82	0.84	**64**	0.88	0.81	0.84
0.00001	0.70	0.82	0.76	**128**	0.84	0.86	0.85

Table 1: **Performance metrics (P: Precision, R: Recall, F1: F1-score) of the black-box model.** The model shows stable performance across different learning rates and batch size, supporting its reliability.

RQ1: Does the proposed explainability method truly make a difference, and why does it matter?

To answer this question, we compared our explainability method, TCT, with two popular techniques: Occlusion and LIME. The objective was to evaluate both interpretability and computational efficiency on randomly selected 1000 test case (see Table **??**).

Metric	**Occlusion**	**LIME**	**TCT**
Avg. Comprehensiveness Score	0.68	0.48	**0.75**
Avg. Sufficiency Score	0.005	0.0018	**0.0017**
Processing Time (Seconds) for one sample	334.8	0.30	**0.19**

Table 2: **Comparison of TCT with other explainability methods.** TCT shows higher interpretability and faster runtime.

We used two primary metrics for evaluation:

- **Comprehensiveness Score**: Measures the drop in prediction confidence when the most important features are removed. A higher score indicates that the identified features are crucial for the prediction.
- **Sufficiency Score**: Measures how well the top features alone can explain the prediction. A lower score is preferable, indicating that only a small subset of features is sufficient to maintain the prediction.

Our method achieved the highest comprehensiveness score (0.75) and the lowest sufficiency score (0.0017), indicating that TCT effectively identifies the most meaningful and predictive features. Furthermore, TCT was the fastest, generating explanations in approximately 0.19 seconds per case, compared to over 300 seconds for Occlusion and 0.30 seconds for LIME. This efficiency makes TCT feasible for real-time deployment in clinical environments.

Most importantly, unlike existing methods, TCT not only identifies *what* features are important but also reveals *when* they become significant in a patient's history. This temporal insight is critical for tracking disease progression and enabling timely interventions, particularly in acute conditions such as AKI.

RQ2: Does the method identify critical time points in a patient's history? How meaningful are the Concept Block representations?

The proposed TCT method effectively identifies critical time points when influential features contribute most to the model's prediction of acute kidney injury. In the provided heatmap (See Figure 2) for Patient #1049 (AKI Risk: 0.99, True Label: 1) from the test set, we observe that the majority of time steps exhibit minimal feature importance (blue regions). However, a sharp increase in importance is observed between Hour 27 and Hour 40, particularly around Hour 28 and Hour 39, where features such as Creatinine, Iron Binding Capacity (Total), and Dialysis patients show strong activation. We also observe that the proposed method capture similar critical time points for other patients (not shown for space constraints). This temporal concentration of importance not only highlights what features matter but crucially reveals when they become relevant for risk prediction. Such insights are particularly valuable in clinical settings where early intervention can prevent deterioration.

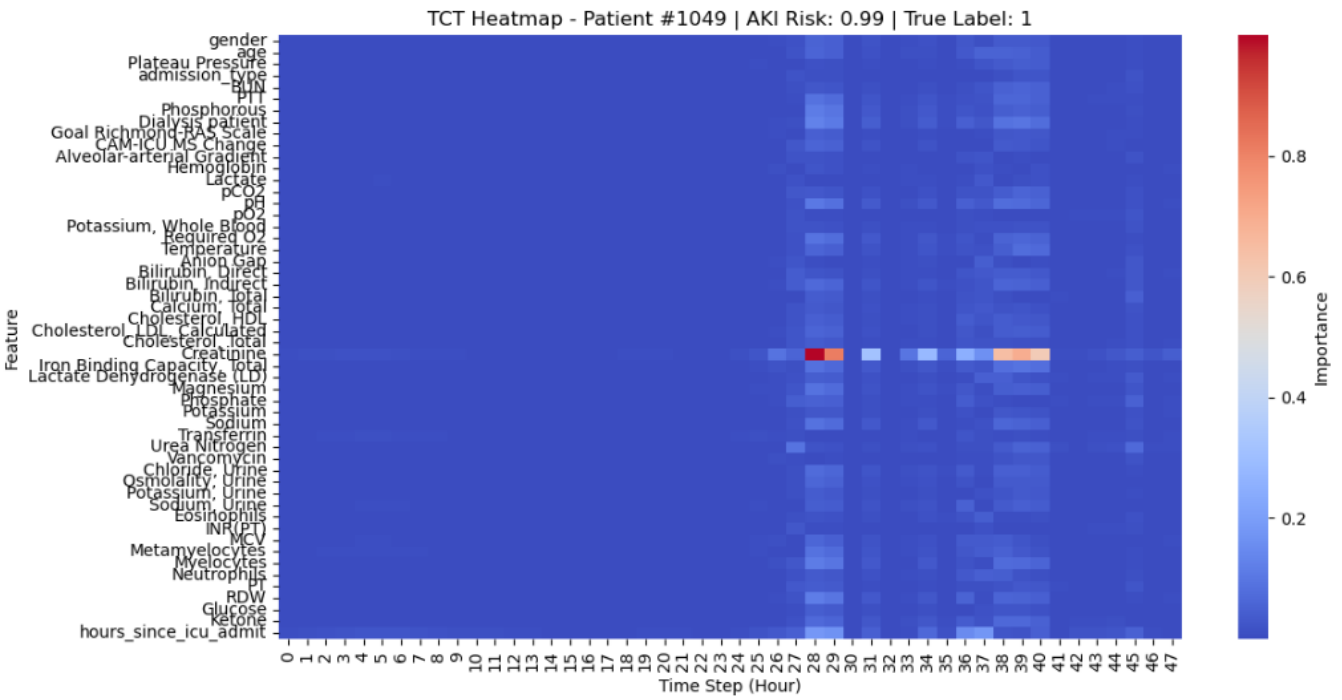

Figure 2: **A Temporal Concept Tracing (TCT) heatmap for Patient #1049.** The model assigns high importance to specific features during critical time steps (around hours 28–40), demonstrating its ability to identify when key physiological changes contribute to AKI risk prediction

To evaluate whether the Concept Block representations in our model encode meaningful information, we experiment with three different concept bottleneck designs:

1. A simple dense layer,
2. A transformer-based block, and
3. The proposed convolutional concept bottleneck model.

All three variants were integrated into the same LSTM-Transformer encoder backbone. We analyzed both the AKI prediction scores and the temporal feature attributions for a held-out ICU patient (#1049, True AKI label: 1).

Concept Type	AKI Risk Predict	Time Activation	Comment
Dense Layer	82.7%	Hour 46–47	Late and compressed output
Transformer Based	40%	Hours 0–2, 36, 45	Wrong prediction, scattered temporal activations
Proposed	**99%**	Hours 28–40	**Accurate and timely prediction; highlights relevant features**

Table 3: **A case study for comparing proposed concept block design with alternative bottleneck designs: a) Dense Layer and b) Transformer.** The result is shown for Patient #1049 with true AKI label.

RQ3: Can LLMs be leveraged to generate explanation reports?

To assess whether large language models (LLMs) can generate clinically meaningful explanations, we integrated MedAlpaca(Han et al. 2023), a domain-adapted LLM built on LLaMA-7B, into our interpretability pipeline. MedAlpaca is fine-tuned on a combination of publicly available medical datasets, including Medical QA, Clinical Knowledge, Instruction Tuning, and Multi-Source Corpus to improve its medical reasoning and instruction-following capabilities.

In our approach, we constructed structured prompts from the TCT output, summarising key features, their temporal activations, and associated patient values. These prompts were passed to MedAlpaca, which generated patient-specific natural language explanations contextualising the model's prediction. For instance:

The generated outputs closely aligned with clinical reasoning and avoided hallucinations due to the structured input. Unlike attention heatmaps or feature importance scores alone, these textual summaries provide narrative, time-aware justifications making the model's decisions more transparent and interpretable to clinicians (See Figure 3).

The dense-layer model predicted 82.7% AKI risk but only activated at the final hour (46–47), offering poor temporal insight. The transformer-based model performed worse, predicting 40% risk with scattered importance across time, including irrelevant early hours. In contrast, the CNN-based

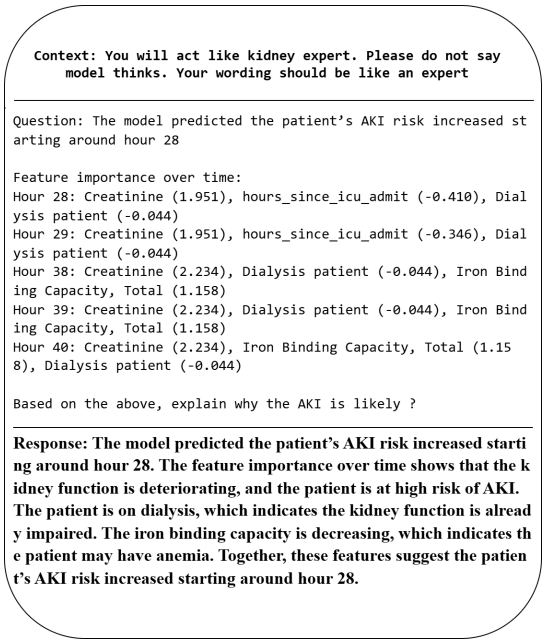

Figure 3: **A sample of LLM-generated explanation report using MedAlpaca.** The report translates temporal risk factor importance into clinically coherent justification for AKI Risk.

model predicted 99% risk and consistently highlighted clinically relevant features (e.g., Creatinine, Iron Binding Capacity) from Hours 28–40, aligning well with AKI progression. This indicates that our CNN-based concept bottleneck effectively captures meaningful temporal patterns, offering a strong balance between abstraction and interpretability.

Limitations and Future Directions

While our results are promising, several limitations warrant discussion. First, our evaluation is limited to the MIMIC-IV dataset, and validation across different healthcare systems and populations is needed to establish generalizability. The reliance on retrospective data, while appropriate for development, should be complemented by prospective clinical validation to assess real-world utility and clinician acceptance.

The concept learning approach, while effective at identifying meaningful temporal patterns, still operates as a latent representation without explicit clinical concept labels. Future work could explore hybrid approaches that combine our learned concepts with clinician-annotated concepts to enhance interpretability further. Additionally, while MedAlpaca successfully generates coherent explanations, evaluation with clinical experts is needed to assess the clinical accuracy and utility of these generated reports.

The framework's architecture-agnostic design, while advantageous for adoption, requires integration during training rather than post-hoc application. This limitation may affect deployment in settings with existing trained models, suggesting the need for transfer learning approaches that could retrofit interpretability into pre-trained systems.

Conclusion

We have presented a novel interpretability framework that successfully addresses the critical challenge of temporal explainability in deep learning models for AKI anticipation. Our approach makes three key contributions that collectively advance the state of interpretable AI in healthcare: a convolutional concept bottleneck that learns meaningful temporal patterns without manual annotation, TCT that identifies both what features matter and when they become critical, and integration with large language models to generate clinical explanations.

The experimental results demonstrate clear advantages over existing interpretability methods, with superior comprehensiveness and sufficiency scores, significantly faster processing times, and the unique ability to provide temporal context essential for clinical decision-making. The CNN-based concept bottleneck architecture proves particularly effective at capturing localized temporal patterns in physiological data, while the LLM integration successfully translates complex model outputs into actionable clinical insights.

This framework represents a significant step toward bridging the gap between model accuracy and clinical utility in healthcare AI. By making AKI predictions both transparent and temporally contextualized, our approach addresses fundamental barriers to deep learning adoption in critical care settings. The architecture-agnostic design enables integration with various temporal models, potentially facilitating broader adoption across different healthcare AI applications.

Acknowledgments

This research was supported by a Research Seed Grant from the College of Information at the University of North Texas (2023).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2019. Graph-Based Attention Model for Healthcare Representation Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 715–723. ACM.

Choi, E.; et al. 2016. RETAIN: Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Han, T.; Adams, L. C.; Papaioannou, J.; Grundmann, P.; Oberhauser, T.; Löser, A.; Truhn, D.; and Bressem, K. K. 2023. MedAlpaca: An Open-Source Collection of Medical Conversational AI Models and Training Data. *arXiv preprint arXiv:2304.08247*.

Johnson, A. E. W.; Bulgarelli, L.; Shen, L.; et al. 2023. MIMIC-IV, a Freely Accessible Electronic Health Record Dataset. *Scientific Data*, 10(1): 1–11.

Koh, P. W.; et al. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Li, J.; Zhou, Z.; Lyu, H.; and Wang, Z. 2025. Large Language Models-Powered Clinical Decision Support: Enhancing or Replacing Human Expertise? *Intelligent Medicine*, 5(1): 1–4.

Li, Y.; Mamouei, M.; Salimi-Khorshidi, G.; Rao, S.; Hassaine, A.; Canoy, D.; Lukasiewicz, T.; and Rahimi, K. 2023. Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records. *IEEE Journal of Biomedical and Health Informatics*, 27(4): 1106–1117.

Li, Y.; Sun, K.; Wang, Y.; and Zhang, F. 2020. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(7155).

Liu, Z.; Wang, T.; Shi, J.; Zheng, X.; Chen, Z.; Song, L.; Dong, W.; Obeysekera, J.; Shirani, F.; and Luo, D. 2024. TIMEX++: Learning Time-Series Explanations with Information Bottleneck. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Lundberg, S.; and Lee, S.-I. 2017a. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 4765–4774. NeurIPS.

Lundberg, S. M.; and Lee, S. I. 2017b. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Margeloiu, A.; et al. 2021. Do Concept Bottleneck Models Learn as Intended? *arXiv preprint arXiv:2105.04289*.

Mullenbach, J.; et al. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Park, S.; et al. 2025. An Analysis of Concept Bottleneck Models: Measuring, Understanding... *arXiv preprint arXiv:2505.16705*.

Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; and Zhi, D. 2021. Med-BERT: Pretrained Contextualized Embeddings on Large-Scale Structured Electronic Health Records for Disease Prediction. *NPJ Digital Medicine*, 4: 13.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Shukla, S. N.; and Marlin, B. 2018. Modeling Irregularly Sampled Clinical Time Series. In *ML4H Workshop at NeurIPS*.

Shukla, S. N.; and Marlin, B. 2021. Multi-Time Attention Networks for Irregularly Sampled Time Series. In *International Conference on Learning Representations (ICLR)*.

Singhal, K.; et al. 2023. Large Language Models Encode Clinical Knowledge. *Nature*, 620: 172–180.

Tonekaboni, S.; Joshi, M.; McCradden, J.; and Goldenberg, A. 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *NPJ Digital Medicine*, 2(1): 1–7.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 5998–6008. NeurIPS.

Xu, J.; and Staniek, M. 2025. Multimodal Transformers for Clinical Time Series Forecasting and Early Sepsis Prediction. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, 100–108. Albuquerque, New Mexico: Association for Computational Linguistics.

Yeh, C.; Chen, Y.; Wu, A.; Chen, C.; Viegas, F.; and Wattenberg, M. 2024. AttentionViz: A Global View of Transformer Attention. *IEEE Transactions on Visualization and Computer Graphics*, 30(1): 262–272.

Yuksekgonul, M.; Wang, M.; and Zou, J. 2022. Post-hoc Concept Bottleneck Models. *arXiv preprint arXiv:2205.15480*.