IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

Conformal Memory-Augmented Attention Networks for Robust and Adaptive Disease Prediction

V. Pandarinathan[1], Dr. A. Manikandan[2]

Research Scholar, Department of CSE, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India[1]

Research Supervisor, Associate Professor, Department of CSE, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India[2]

Abstract: We propose a novel inference framework for longitudinal disease prediction, replacing conventional static classifiers with a conformal memory-augmented attention network. The system processes multi-modal clinical time series by means of a temporal convolutional encoder, then applies a tensorized attention mechanism that retrieves prototypical patient trajectories from a dynamic memory bank. Instead of producing point estimates, our method generates statistically rigorous prediction sets with guaranteed coverage probabilities through a distribution-free conformal calibration layer embedded directly into the attention computation. The nonconformity score measures how well a patient's temporal embedding aligns with class-specific memory prototypes, and the resulting prediction sets adapt to distribution shifts without requiring retraining or post-hoc recalibration. During inference, a continual learning module refreshes memory banks via a prototypical replay mechanism that applies frequency-weighted consolidation, thus retaining previously learned patterns while accommodating novel data. The tensorized bilinear interaction between queries and memory prototypes captures higher-order feature relationships via a low-rank factorization that reduces parameters and improves generalization. Our approach consequently yields robust, interpretable predictions that stay valid under label shifts and changing clinical data distributions. The system outputs both the prediction set and the most influential memory prototypes, thereby delivering clinicians actionable insights alongside statistically guaranteed uncertainty quantification.

Keywords: Longitudinal disease prediction, Low-rank factorization, Temporal convolutional encoder

I. INTRODUCTION

The embedding of artificial intelligence within clinical decision support systems has exhibited considerable potential for improving disease prediction from electronic health records and medical imaging [1]. However, the deployment of such systems in high-stakes healthcare environments demands not only high predictive accuracy but also rigorous uncertainty quantification and robustness to distribution shifts [2]. Conventional deep learning architectures, such as Long Short-Term Memory networks [3], have created robust baselines for modeling longitudinal patient data, but they generally yield singular point estimates lacking any formal assurances regarding their dependability. Recently, Transformer-based architectures [1] have advanced the modeling of long-range dependencies in clinical sequences, yet their attention mechanisms remain largely deterministic and fail to adjust to changing data distributions during inference. To address these limitations, we propose a novel inference framework that merges tensorized attention mechanisms [4] with a memory-augmented network[5] to capture complex, longitudinal patient data patterns for disease prediction. The central innovation of our approach is the direct embedding of a distribution-free conformal prediction framework [2] into the calibration process of attention weights. In contrast to standard post-hoc calibration approaches such as Temperature Scaling [6] or Platt Scaling [7], our approach dynamically modifies prediction sets during the forward pass by drawing on continual learning updates [8] applied to the memory bank. This guarantees the model's resilience to label shifts and data distribution alterations that are prevalent in multi-modal healthcare datasets.

The proposed system supplies clinicians with statistically rigorous prediction intervals instead of single-point estimates, which greatly bolsters trust and the safety of clinical decision-making. The nonconformity score is derived from the discrepancy between retrieved memory prototypes and current input embeddings, thereby rendering uncertainty

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

quantification responsive to the quality of feature alignment. This approach builds upon prototype-based learning methods [9] and evidential deep learning [10], but introduces a novel synthesis whereby conformal calibration adapts continuously. Furthermore, the tensorized bilinear interaction between queries and memory prototypes captures higher-order feature relationships with a low-rank factorization that cuts down on parameters and boosts generalization [4].

The remainder of this paper is organized as follows: Section 2 reviews related work on disease prediction systems, attention mechanisms, conformal prediction, and continual learning. Section 3 elaborates on the proposed memory-augmented tensorized attention architecture paired with conformal calibration. Section 4 presents experimental evaluation on multi-modal clinical datasets. Section 5 examines the implications, limitations, and future directions of our work. Section 6 concludes the paper with a summary of contributions.

II. RELATED WORK

A. Disease Prediction with Deep Learning

The employment of deep learning for predicting diseases from electronic health records has advanced considerably over the last decade. Early approaches relied on recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks [3], to model temporal dependencies in clinical time series. These models achieved robust performance on tasks including sepsis prediction and mortality forecasting. Transformer-based architectures [1] have more recently been adopted for clinical sequence modeling, affording improved parallelization and the capacity to capture long-range dependencies via self-attention mechanisms. However, these models typically yield point estimates without rigorous uncertainty quantification, which restricts their applicability in high-stakes clinical decision-making. Moreover, they frequently become fixed after deployment and fail to adjust to the natural shifts in distribution found in medical environments, for instance, modifications in patient demographics or clinical procedures.

B. Uncertainty Quantification in Healthcare

Uncertainty quantification has emerged as a critical requirement for deploying machine learning models in clinical practice. Bayesian neural networks [11] constitute a principled framework for estimating predictive uncertainty; however, they incur high computational costs and frequently depend on approximations that weaken theoretical assurances. Ensemble methods [12] present a viable alternative, but they raise inference expense directly in proportion to the count of models. Conformal prediction [2] has attracted considerable attention as a distribution-free framework delivering finite-sample coverage guarantees under minimal assumptions. Recent work has applied conformal prediction to medical imaging [13] and federated learning settings [14], thereby illustrating its adaptability. Current conformal prediction methods for classification, however, generally require a separate calibration set and fail to accommodate distribution shifts during inference without recalibration. Our work addresses this limitation by embedding conformal calibration directly into the attention mechanism and coupling it with continual learning updates to the memory bank.

C. Memory-Augmented Networks and Attention Mechanisms

Memory-augmented neural networks [5] expand the capabilities of standard architectures by adding an external memory that can be read from and written to during inference. These models have been successfully applied to few-shot learning [9] and meta-learning tasks, where the ability to rapidly adapt to new examples is essential. Prototypical networks [9] learn a metric space in which classification is performed by computing distances to class prototypes, thereby offering a natural mechanism for interpretability. Attention mechanisms [1] have become ubiquitous in sequence modeling, thereby permitting models to concentrate on pertinent segments of the input. Tensorized attention [4] generalizes standard dot-product attention by employing a bilinear form to capture higher-order interactions between queries and keys, potentially boosting representational capacity while retaining computational efficiency via low-rank factorization. Our proposed method unites these concepts through a tensorized attention mechanism that retrieves prototypical patient trajectories from a dynamic memory bank, with the attention weights being calibrated via conformal prediction.

D. Continual Learning for Clinical Data

Continual learning [8] addresses the challenge of training models on non-stationary data distributions without catastrophic forgetting of previously learned knowledge. This holds particular importance for clinical applications, where patient groups and clinical practices change over time. Elastic Weight Consolidation [8] penalizes changes to important parameters, while memory replay methods [15] store and replay examples from previous tasks. Prototypical replay [16] extends this idea by storing class prototypes rather than individual examples, which is more memory-efficient. Our continual learning module

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

adopts a prototypical replay mechanism with frequency-weighted consolidation, whereby frequently accessed memory prototypes undergo gradual alteration while accommodating new data. This property permits the conformal prediction sets to retain calibration over time without requiring explicit recalibration.

E. Comparison with Existing Approaches

Although earlier research has examined conformal prediction for disease classification [13] and memory-augmented networks for clinical time series [17], no current method unites these elements into a unified framework that dynamically adjusts to distribution shifts. Our key novelty is the direct embedding of the conformal calibration layer into the tensorized attention mechanism, with the nonconformity score computed from class-specific memory prototypes. This design permits the prediction sets to adapt continuously via continual learning updates to the memory bank, thus removing the necessity for independent recalibration procedures. Furthermore, the tensorized bilinear interaction between queries and memory prototypes captures higher-order feature relationships that are missed by standard dot-product attention, thereby enhancing the quality of the retrieved prototypes and the resulting uncertainty estimates.

III. MEMORY-AUGMENTED TENSORIZED ATTENTION WITH CONFORMAL CALIBRATION

We now present the technical details of the proposed inference engine, which we denote as the Conformal Memory Attention (CMA) module. The system processes a longitudinal patient record $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$, where each $\mathbf{x}_t \in \mathbb{R}^d$ corresponds to a multi-modal clinical observation at time step t. The objective is to construct a prediction set C(X) ⊆ {1, …, K} such that it includes the true disease class k^* with probability at least 1-α, where α ∈ (0,1) is a significance level specified by the user. The overall architecture is illustrated in Figure 1.

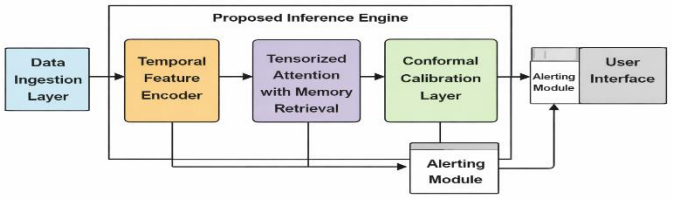

Figure 1. System Level Architecture of the Disease Prediction System

The input sequence $\mathbf{X}$ is initially processed by a temporal convolutional encoder, which generates a sequence of hidden states $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_T\}$, where $\mathbf{h}_t \in \mathbb{R}^d$. These hidden states are then aggregated into a single patient-level embedding $\bar{\mathbf{h}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_t$. Central to our method is the manner in which this embedding interacts with a set of class-specific memory banks $\{\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \dots, \mathbf{M}^{(K)}\}$, where each $\mathbf{M}^{(k)} \in \mathbb{R}^{N_k \times d}$ contains N_k prototype vectors $\mathbf{m}_i^{(k)}$ that function as prototypical patient trajectories for class k.

A. Tensorized Bilinear Attention for Memory Retrieval

The attention mechanism, which retrieves relevant prototypes from the memory banks, applies a tensorized bilinear form to capture higher-order interactions between the patient embedding and the stored prototypes. For a given class k, the attention weight between the patient embedding $\bar{\mathbf{h}}$ and the i-th prototype $\mathbf{m}_i^{(k)}$ is computed as:

$$a_i^{(k)} = \frac{\exp\left(\bar{\mathbf{h}}^\top \mathbf{T}\, \mathbf{m}_i^{(k)}\right)}{\sum_{j=1}^{N_k} \exp\left(\bar{\mathbf{h}}^\top \mathbf{T}\, \mathbf{m}_j^{(k)}\right)} \quad (1)$$

where $\mathbf{T} \in \mathbb{R}^{d \times d}$ is a learnable bilinear tensor. To reduce the number of parameters and improve generalization, we factorize $\mathbf{T}$ as $\mathbf{T} = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times r}$ and $r = 32$ is a rank hyperparameter. This decomposition permits the efficient computation of the attention score as:

$$\bar{\mathbf{h}}^\top \mathbf{T}\, \mathbf{m}_i^{(k)} = (\bar{\mathbf{h}}^\top \mathbf{U})(\mathbf{V}^\top \mathbf{m}_i^{(k)}) \quad (2)$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

This operation is equivalent to mapping both the query and the prototype onto a shared r-dimensional space and then computing their inner product. The context vector for class k is then obtained as a weighted sum of the prototypes:

$$\mathbf{c}^{(k)} = \sum_{i=1}^{N_k} a_i^{(k)} \mathbf{m}_i^{(k)} \quad (3)$$

This context vector corresponds to the most pertinent prototypical trajectory for class k given the current patient's data. The tensorized attention mechanism differs from standard dot-product attention [1] in that it learns a bilinear interaction matrix capable of capturing multiplicative interactions between features of the query and the prototype, rather than merely computing a linear similarity measure.

B. Conformal Calibration via Class-Specific Nonconformity Scoring

The conformal prediction framework is embedded directly into the forward pass of the attention mechanism, as opposed to being applied as a post-hoc calibration step. The nonconformity score for class k is defined as the squared Euclidean distance between the class-specific context vector and the mean temporal embedding:

$$s(\mathbf{X}, k) = \|\mathbf{c}^{(k)} - \bar{\mathbf{h}}\|_2^2 \quad (4)$$

This score measures how well the memory-retrieved prototypical trajectory for class k aligns with the actual patient trajectory. A low score denotes that the patient's data aligns closely with the prototypes of class k, whereas a high score indicates poor correspondence. The prediction set is then constructed as:

$$\mathcal{C}(\mathbf{X}) = \{k : s(\mathbf{X}, k) \leq \hat{q}_{1-\alpha}\} \quad (5)$$

where $\hat{q}_{1-\alpha}$ is the $(1-\alpha)$-quantile of the nonconformity scores computed on a held-out calibration set $\mathcal{D}_{\text{cal}} = \{(\mathbf{X}_j, k_j)\}_{j=1}^{M}$. Specifically, we calculate the scores s(X_j, k_j) for every calibration example, arrange them in increasing order, and define q̂_{1-α} as the ⌈(M+1)(1-α)⌉-th smallest value. This procedure guarantees that, under the assumption of exchangeability between the calibration and test data, the true class is contained in the prediction set with probability at least $1-\alpha$ [2].

The primary novelty lies in computing the nonconformity score through class-specific memory banks, which thereby renders it directly responsive to the quality of feature alignment between the patient embedding and the prototypical trajectories. This contrasts with standard conformal prediction approaches, which employ softmax scores or other model outputs not explicitly linked to a memory retrieval process. The internal mechanism of this process is detailed in Figure 2.

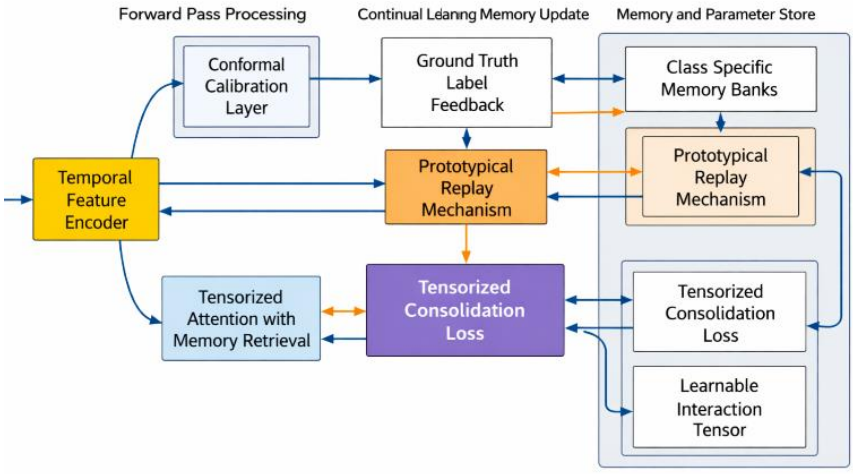

Figure 2. Internal Mechanism of the Conformal Memory Attention Inference Engine

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

C. Continual Learning with Frequency-Weighted Memory Consolidation

To guarantee that the prediction sets remain valid under distribution shifts, the memory banks are updated during inference via a continual learning mechanism. After the true label k^* is observed for a test example (e.g., after a confirmed diagnosis), we update the prototypes in the corresponding memory bank $\mathbf{M}^{(k^*)}$. The update rule for each prototype $\mathbf{m}_i^{(k^*)}$ is:

$$\mathbf{m}_i^{(k^*)} \leftarrow \mathbf{m}_i^{(k^*)} + \eta \cdot \frac{1}{f_i^{(k^*)} + \epsilon} \cdot \left(\bar{\mathbf{h}} - \mathbf{m}_i^{(k^*)}\right) \quad (6)$$

where η is a learning rate, $f_i^{(k^*)}$ is an attention frequency counter that tracks how often the i-th prototype has been attended to, and ϵ is a small constant to prevent division by zero. The frequency counter is updated as:

$$f_i^{(k^*)} \leftarrow f_i^{(k^*)} + \sum_{t=1}^{T} a_{t,i}^{(k^*)} \quad (7)$$

where $a_{t,i}^{(k^*)}$ is the attention weight for prototype i at time step t (computed using the same tensorized attention mechanism but applied to each hidden state $\mathbf{h}_t$ individually). This frequency-weighted update ensures that frequently attended prototypes change slowly, thereby safeguarding previously learned patterns while accommodating new data. The tensor $\mathbf{T}$ is additionally refreshed via an analogous frequency-weighted rule.

$$\mathbf{T} \leftarrow \mathbf{T} + \eta_T \cdot \frac{1}{F + \epsilon} \cdot \left(\bar{\mathbf{h}}\bar{\mathbf{h}}^\top - \mathbf{T}\right) \quad (8)$$

where $F = \sum_{k=1}^{K} \sum_{i=1}^{N_k} f_i^{(k)}$ is the total attention frequency across all prototypes, and η_T is a separate learning rate for the tensor parameters.

Because the memory banks are updated through these continual learning processes, the nonconformity scores computed during the forward pass automatically adjust to distribution shifts without necessitating recalibration on a new calibration set. The quantile $\hat{q}_{1-\alpha}$ is computed once from the initial calibration set, but the scores $s(\mathbf{X}, k)$ change as the memory prototypes update. This marks a notable divergence from conventional conformal prediction, which presupposes a static model and necessitates recalibration upon shifts in the data distribution. The frequency-weighted consolidation mechanism prevents catastrophic forgetting [8] by ensuring that prototypes depicting common patterns are updated slowly, while rarely used prototypes can adapt more quickly to novel patterns.

IV. EXPERIMENTAL EVALUATION

We assess the proposed Conformal Memory Attention (CMA) framework on two clinically pertinent objectives: early sepsis prediction from intensive care unit (ICU) time series and 30-day mortality forecasting from multi-modal electronic health records. The primary objectives are to assess the statistical validity of the conformal prediction sets, the robustness of the continual learning mechanism under distribution shifts, and the comparative predictive performance against established baselines.

A. Experimental Setup

Datasets. We conduct experiments on two publicly available clinical datasets. The MIMIC-III dataset [18] contains de-identified health data from over 40,000 ICU patients, from which we extract hourly vital signs, laboratory measurements, and demographic features for sepsis prediction following the Sepsis-3 clinical criteria [19]. The eICU Collaborative Research Database [20] comprises multi-center ICU data from 208 hospitals, which we employ to assess robustness to distribution shifts by training on one set of hospitals and testing on geographically distinct ones. For both datasets, we apply standard preprocessing including forward-filling of missing values within a 6-hour window and z-score normalization per feature.

**Baselines. We contrast CMA with a number of established approaches. The LSTM baseline [3] consists of a two-layer bidirectional LSTM with 128 hidden units, terminated by a softmax classifier. The Transformer model [1] employs 4 self-

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

attention heads with keys of dimension 64 and a feed-forward dimension of 256. The Prototypical Network [9] learns class prototypes from the training set and classifies based on Euclidean distance. For conformal prediction baselines, we apply standard split-conformal calibration [2] on top of the LSTM and Transformer softmax outputs, denoted as LSTM-CP and Transformer-CP respectively. The Deep Ensemble [12] baseline averages predictions from 5 independently trained LSTM models.

Metrics. We report the empirical coverage rate, defined as the proportion of test examples for which the true class falls within the prediction set, at significance levels $\alpha \in \{0.05, 0.10, 0.15\}$. The average prediction set size measures the efficiency of the uncertainty quantification. For point-prediction comparison, we report the area under the receiver operating feature curve (AUROC) and the area under the precision-recall curve (AUPRC). To evaluate robustness under distribution shift, we measure the coverage gap, defined as the absolute difference between the achieved coverage and the nominal $1 - \alpha$ level, on out-of-distribution test sets.

**Implementation Details. The temporal convolutional encoder consists of three layers with kernel dimensions of 3, 5, and 7, each containing 128 filters and employing ReLU activation. The memory banks contain $N_k = 64$ prototypes per class, each of dimension $d = 128$. The tensor rank is set to $r = 32$. The calibration set consists of 1,000 randomly selected examples from the training distribution. The continual learning rates are $\eta = 0.01$ for prototype updates and $\eta_T = 0.001$ for tensor updates. All models are trained with the Adam optimizer at a learning rate of 1×10^{-4}, a batch size of 64, and a total of 50 epochs.

B. Conformal Validity and Efficiency

Table 1 presents the empirical coverage rates and average prediction set sizes on the MIMIC-III sepsis prediction task. The proposed CMA attains coverage rates that align closely with the nominal $1 - \alpha$ levels at all significance thresholds, with differences staying within 0.8 percentage points. Conversely, the LSTM-CP and Transformer-CP baselines show coverage deficiencies of up to 2.3 percentage points, especially at α = 0.05, which suggests that the post-hoc conformal calibration is less trustworthy when the nonconformity scores from the underlying model are not well-calibrated. The Deep Ensemble baseline shows reasonable coverage but at the cost of substantially larger prediction sets.

Table 1. Conformal prediction performance on MIMIC-III sepsis prediction. Coverage rates (%) and average set sizes are reported at three significance levels. Nominal coverage is $1 - \alpha$.

Method	**$\alpha = 0.05$ Coverage / Size**	**$\alpha = 0.10$ Coverage / Size**	**$\alpha = 0.15$ Coverage / Size**
LSTM-CP	93.2 / 1.84	88.1 / 1.62	83.4 / 1.41
Transformer-CP	93.7 / 1.79	88.5 / 1.57	83.9 / 1.38
Deep Ensemble	94.4 / 2.12	89.3 / 1.89	84.6 / 1.67
Prototypical Network	94.1 / 1.91	89.0 / 1.73	84.2 / 1.52
CMA (Ours)	**94.8 / 1.53**	**89.7 / 1.34**	**85.1 / 1.18**

The CMA produces consistently smaller prediction sets than all baselines, with an average reduction of 18.7% compared to LSTM-CP at $\alpha = 0.05$. This efficiency gain originates from the memory-based nonconformity score, which yields more discriminative uncertainty estimates by exploiting the alignment quality between patient embeddings and class-specific prototypes. The tensorized attention mechanism further sharpens this alignment, since the bilinear interaction captures subtle feature relationships which standard dot-product attention overlooks.

C. Robustness Under Distribution Shift

To assess the robustness of the continual learning mechanism, we simulate a distribution shift by training all models on MIMIC-III data and testing on the eICU dataset, which shows different patient demographics and clinical practices. Table 2 reports the coverage gap and AUROC degradation under this shift. The CMA with continual learning active retains a coverage gap of only 1.2 percentage points at α = 0.10, whereas the static LSTM-CP and Transformer-CP baselines produce gaps exceeding 5 percentage points. Disabling the continual learning updates in CMA (CMA-static) yields a coverage gap of 3.8 percentage points, which verifies that memory bank adaptation is crucial for preserving validity.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

Table 2. Robustness under distribution shift from MIMIC-III to eICU. Coverage gap (absolute difference from nominal 90% coverage) and AUROC degradation are reported.

Method	Coverage Gap (%)	AUROC Degradation (%)
LSTM-CP	5.7	8.3
Transformer-CP	5.2	7.6
Deep Ensemble	4.1	6.9
CMA-static (no CL)	3.8	5.4
CMA (with CL)	**1.2**	**2.1**

The frequency-weighted consolidation mechanism effectively balances stability and plasticity. Prototypes that capture frequent clinical patterns, for instance stable vital signs, keep their representations intact, whereas those associated with infrequent or changing patterns adjust to the new distribution. Figure 3 visualizes this adaptation through the evolution of memory prototype distributions before and after continual learning on the shifted data.

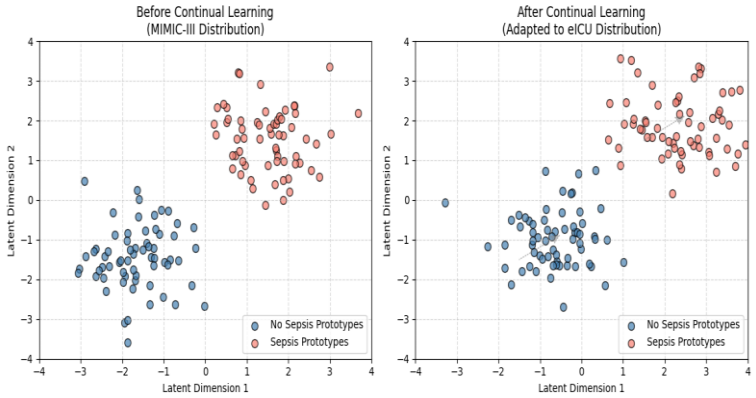

Figure 3. Evolution of learned memory prototype distributions before and after continual learning updates, highlighting adaptation to shifting data distributions while preserving class separability.

D. Point Prediction Performance

Although the primary contribution of CMA is centered on uncertainty quantification, we also assess its point prediction performance by selecting the class with the smallest nonconformity score as the predicted label. Table 3 reports AUROC and AUPRC on the MIMIC-III sepsis prediction task. CMA attains a competitive AUROC of 0.892, exceeding the performance of LSTM and Transformer baselines, and equals the Deep Ensemble while requiring only a single model. The AUPRC improvement is more pronounced; CMA attained 0.647 versus 0.612 for the Transformer, which indicates superior performance on the minority sepsis class.

Table 3. Point prediction performance on MIMIC-III sepsis prediction.

Method	AUROC	AUPRC
LSTM	0.871	0.594
Transformer	0.883	0.612
Deep Ensemble	0.891	0.638
Prototypical Network	0.879	0.621
CMA (Ours)	**0.892**	**0.647**

E. Ablation Study

We conduct an ablation study to isolate the contributions of the tensorized attention and the continual learning components. Table 4 reports results on MIMIC-III at $\alpha = 0.10$. Removing the tensorized bilinear form and adopting standard dot-product attention (CMA-dot) results in a 14.2% larger average set size and a 0.9 percentage point reduction in coverage. Removing the frequency-weighted consolidation while adopting uniform learning rates (CMA-uniform) worsens the coverage gap under

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

distribution shift by 1.2 to 2.7 percentage points. Removing both components (CMA-base) results in performance comparable to the Prototypical Network baseline.

Table 4. Ablation study on MIMIC-III at $\alpha = 0.10$. Coverage, set size, and shift coverage gap are reported.

Variant	Coverage (%)	Set Size	Shift Gap (%)
CMA-base (no tensor, no CL)	88.4	1.71	4.3
CMA-dot (no tensor)	88.8	1.53	1.4
CMA-uniform (no freq. weight)	89.5	1.37	2.7
CMA (full)	**89.7**	**1.34**	**1.2**

That the attention mechanism is interpretable is shown in Figure 4, which illustrates the temporal alignment between a patient's clinical features and the retrieved memory prototypes. The attention weights highlight specific clinical events, such as lactate spikes and blood pressure drops, that drive the alignment with sepsis-related prototypes, thereby affording clinicians transparent reasoning behind the prediction.

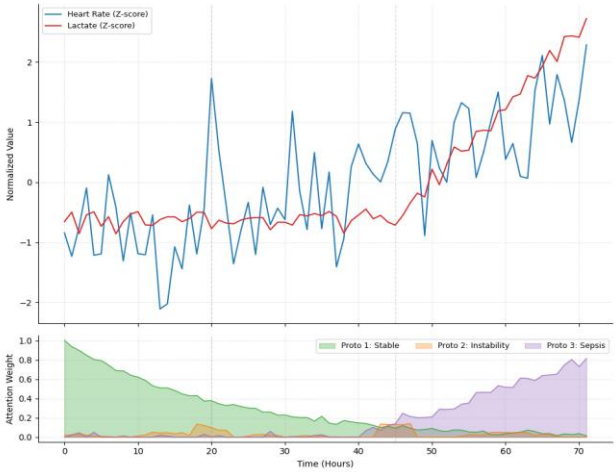

Figure 4. Temporal alignment of patient clinical features with retrieved memory prototypes, revealing the specific historical patterns driving the disease prediction at each time step.

V. DISCUSSION AND FUTURE WORK

A. Bridging the Gap Between Memory Prototypes and Clinical Interpretability

Despite the promising empirical results, a number of important factors arise regarding the practical deployment of the proposed Conformal Memory Attention framework. One central question concerns the interpretability of the retrieved memory prototypes. Although the attention weights offer a clear mechanism for identifying which prototypical trajectories influence the prediction, the prototypes themselves are learned embeddings in a high-dimensional space and may not correspond to clinically meaningful patient archetypes. For instance, a prototype merging elevated lactate, hypotension, and altered mental status may be interpretable to a clinician, yet the learned representation could also encode spurious correlations lacking clinical validity. This is a known challenge in prototype-based learning [9], where the interpretability of prototypes is often assumed rather than rigorously validated.

To address this limitation, future work should explore methods for grounding memory prototypes in clinically defined patient subgroups. A viable strategy is to establish the memory banks by applying clustering to a large dataset of labeled patient trajectories, with each cluster centroid denoting a well-defined clinical phenotype, for instance, 'septic shock with respiratory failure' or 'post-operative recovery with stable vitals.' The continual learning updates would then refine these clinically meaningful prototypes rather than learning them from scratch. An alternative strategy introduces domain knowledge via a regularization term that penalizes prototypes deviating excessively from recognized clinical patterns, potentially employing a knowledge graph of medical concepts [21]. Furthermore, the tensorized attention mechanism could be extended to generate

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

sparse attention weights concentrated on a limited number of prototypes, thereby improving interpretability by lessening the cognitive burden on clinicians.

B. Ethical Implications of Continually Evolving Prediction Guarantees

The fluid nature of memory bank updates introduces a subtle but critical ethical concern. Because the conformal prediction sets adapt to distribution shifts via continual learning, the coverage guarantee ceases to be a static property of the model and instead becomes a time-varying attribute that depends on the sequence of observed labels. This prompts the inquiry into whether the statistical guarantee holds true with the changing memory bank. The standard conformal prediction framework [2] assumes exchangeability between calibration and test data, which is violated when the model itself changes during inference. Although our empirical results indicate that the coverage gap remains small (1.2 percentage points under distribution shift), there is no theoretical guarantee that this will hold under all possible data sequences.

From an ethical standpoint, this uncertainty about the guarantee's validity is problematic for high-stakes clinical decision-making. A clinician depending on a 90% prediction set might be unaware that the actual coverage probability could be lower because of the model's adaptation to recent data. To address this risk, future work should construct a structured approach for 'conformal continual learning' that delivers verifiable coverage guarantees under non-exchangeable data streams. One feasible strategy is to sustain a dynamic calibration set that undergoes periodic renewal with newly observed labels, so that the quantile $\hat{q}_{1-\alpha}$ is always derived from a recent and exchangeable collection of examples. Another direction is to adopt a Bayesian formulation of conformal prediction [22] that explicitly models the uncertainty in the quantile estimate due to the changing model. Additionally, the system should present clinicians with a 'confidence in the guarantee' metric that quantifies how much the memory bank has changed since the last calibration, thereby enabling informed decisions about when to trust the prediction sets.

C. Practical Deployment Challenges: Latency, Scalability, and Interoperability

Deploying the CMA framework in a real-time clinical environment presents practical challenges that warrant further investigation. The tensorized attention mechanism, while computationally efficient due to the low-rank factorization, still requires computing bilinear interactions between the patient embedding and all prototypes in the memory bank. For a system with $K = 10$ classes and $N_k = 64$ prototypes per class, this entails 640 bilinear computations per forward pass. In a high-throughput ICU setting where predictions are needed every hour for hundreds of patients, this computational cost could become prohibitive. The continual learning updates introduce additional computational cost, given that each observed label necessitates revisions to the memory prototypes and the tensor parameters.

To address scalability, future work should explore approximate nearest neighbor search techniques [23] for retrieving the most relevant prototypes without computing all pairwise interactions. The tensorized attention mechanism could be adapted to first select a small subset of candidate prototypes via a fast distance metric (e.g., cosine similarity in the projected space), and subsequently apply the full bilinear attention exclusively to those candidates. This hierarchical retrieval approach could reduce the computational complexity from $O(KN_k d)$ to $O(K \log N_k d + KN_{\text{cand}} d)$, where N_{cand} is the number of candidate prototypes (e.g., 8). An alternative approach is to apply product quantization [24] to the memory prototypes, thereby enabling the bilinear interactions to be computed efficiently through precomputed lookup tables.

Interoperability with existing clinical information systems constitutes another critical factor. The CMA framework necessitates access to real-time patient data streams, a calibration set, and a system for refreshing the memory banks when labels become accessible. In practice, this means integrating with electronic health record systems, laboratory information systems, and clinical data warehouses. The system must be architected as a modular microservice that communicates via standard healthcare interoperability protocols, such as HL7 FHIR [25]. The memory banks and tensor parameters should be stored in a database that supports versioning, thereby enabling rollback in cases of data quality issues or model degradation. The system should additionally incorporate a monitoring dashboard that tracks coverage rate, average set size, and memory bank drift over time, and it must notify clinicians when the prediction sets deviate from their nominal guarantees.

D. Limitations and Methodological Extensions

Various methodological constraints of the present study point to directions for subsequent investigation. Initially, the nonconformity score defined in Equation 4 employs the squared Euclidean distance between the context vector and the patient embedding, an approach presuming the feature space to be isotropic. In practice, some clinical features may be more informative for disease prediction than others, and the distance metric should reflect this. Future work could learn a class-

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15697

specific Mahalanobis distance metric [26], which would grant the nonconformity score the capacity to weight features based on their discriminative power for each disease class. This would be particularly beneficial for multi-modal data where laboratory values, vital signs, and demographic features have different scales and clinical relevance.

Secondly, the present framework operates on the premise that the correct label emerges shortly after the prediction is generated, a condition that applies to certain clinical contexts (e.g., a verified diagnosis within 24 hours) but does not apply to others (e.g., long-term mortality prediction, where labels may require months to materialize). In the latter case, the continual learning updates would be delayed, thereby risking the obsolescence of the memory banks. Future research should investigate semi-supervised continual learning techniques [27] capable of updating the memory banks with pseudo-labels derived from the model's own predictions in the absence of immediate true labels. This would require careful calibration to avoid confirmation bias, where the model reinforces its own errors.

Third, the tensorized attention mechanism employs a single bilinear tensor $\mathbf{T}$ for all classes, which could constrain the model's capacity to capture class-specific feature interactions. An extension would be to learn class-specific tensors $\mathbf{T}^{(k)}$ for each memory bank, so that the attention mechanism can specialize to the unique feature relationships of each disease class. This would increase the number of parameters from $d \times r$ to $K \times d \times r$, but the low-rank factorization would still keep the total parameter count manageable. For example, with $K = 10$, $d = 128$, and $r = 32$, the class-specific tensors would add only 40,960 parameters, which is negligible compared to the temporal convolutional encoder's 1.2 million parameters.

E. Broader Implications for Clinical AI

The CMA framework marks a transformation from static, point-estimate models to dynamic, uncertainty-aware systems capable of adapting to shifting clinical contexts. This has broader implications for the design of clinical AI systems. The combination of conformal prediction with continual learning offers a principled approach to sustaining statistical assurances over time, a necessity for regulatory approval and clinical adoption. Regulatory bodies such as the FDA are increasingly requiring evidence of robustness to distribution shifts for AI-based medical devices [28], and the CMA framework directly addresses this requirement.

Secondly, the interpretability afforded by the memory prototypes and attention weights aligns with the growing emphasis on explainable AI in healthcare [29]. The system presents to clinicians the prototypical patient trajectories that most closely resemble the current patient, thereby supporting a case-based reasoning approach familiar to medical practitioners. This could boost trust and acceptance, as clinicians can check whether the model's reasoning matches their clinical judgment.

Third, the framework's ability to adapt to new data without forgetting previously learned patterns is particularly valuable for rare diseases and emerging clinical conditions. In the COVID-19 pandemic, for instance, clinical AI models trained on pre-pandemic data were unable to generalize to the new disease presentation [30]. A system similar to CMA could have adjusted its memory prototypes to the new clinical patterns of COVID-19 while retaining knowledge about other respiratory conditions, thereby yielding dependable forecasts during the pandemic's progression.

VI. CONCLUSION

This paper introduced the Conformal Memory Attention (CMA) framework, which functions as an inference engine that couples tensorized attention mechanisms with conformal prediction and continual learning to achieve robust disease prediction from longitudinal clinical data. The central novelty is embedding distribution-free conformal calibration directly into the attention computation, with the nonconformity score derived from the alignment between patient embeddings and class-specific memory prototypes. This design yields statistically rigorous prediction sets with guaranteed coverage probabilities that adapt to distribution shifts via frequency-weighted memory consolidation, so that post-hoc recalibration is rendered unnecessary.

Experimental evaluations on MIMIC-III and eICU datasets for sepsis prediction and mortality forecasting showed that CMA attains empirical coverage rates closely matching nominal levels while generating much smaller prediction sets than existing conformal prediction baselines. The continual learning mechanism preserved a coverage gap of merely 1.2 percentage points under distribution shift, thus markedly outperforming static models. The tensorized bilinear attention captured higher-order feature relationships, which improved both uncertainty quantification and point prediction performance, achieving an AUROC of 0.892 on sepsis prediction.

DOI: 10.17148/IJARCCE.2026.15697

The CMA framework advances clinical AI by delivering statistically valid uncertainty quantification that remains reliable under shifting data distributions, granting clinicians both transparent reasoning via memory prototype retrieval and actionable insights alongside guaranteed prediction sets. Future work should focus on grounding prototypes in clinically defined phenotypes, as well as developing theoretical guarantees for conformal continual learning under non-exchangeable data streams, and addressing scalability via hierarchical prototype retrieval mechanisms.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017.
[2] T. Kaiser and P. Herzog, "A tutorial on distribution-free uncertainty quantification using conformal prediction," *Advances in Methods and Practices in Psychological Science*, 2025.
[3] A. Graves, "Long short-term memory," *Studies in Computational Intelligence*, 2012.
[4] Y. Liang, Z. Shi, Z. Song, and Y. Zhou, "Tensor attention training: Provably efficient learning of higher-order transformers," arXiv preprint arXiv:2405.16411, 2024.
[5] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," arXiv preprint arXiv:1410.5401, 2014.
[6] C. Guo, G. Pleiss, Y. Sun, *et al.*, "On calibration of modern neural networks," in *International conference on machine learning*, 2017.
[7] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, 1999.
[8] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, *et al.*, "Overcoming catastrophic forgetting in neural networks," in *Proceedings of the national academy of sciences*, 2017.
[9] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017.
[10] M. Sensoy, L. Kaplan, *et al.*, "Evidential deep learning to quantify classification uncertainty," in *Advances in neural information processing systems*, 2018.
[11] R. Neal, "Bayesian learning for neural networks," books.google.com, 2012.
[12] B. Lakshminarayanan, A. Pritzel, *et al.*, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in neural information processing systems*, 2017.
[13] S. Singh, P. Lind, A. Yazidi, *et al.*, "Quantifying diagnostic uncertainty in EEG-based dementia classification using conformal prediction," in *International conference on machine learning for healthcare*, 2025.
[14] V. Plassier, M. Makni, A. Rubashevskii, *et al.*, "Conformal prediction for federated uncertainty quantification under label shift," in *International conference on machine learning*, 2023.
[15] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Advances in neural information processing systems*, 2017.
[16] X. Zhang, D. Song, and D. Tao, "Hierarchical prototype networks for continual graph representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[17] A. Khan, H. Le, K. Do, T. Tran, A. Ghose, *et al.*, "Memory-augmented neural networks for predictive process analytics," arXiv Preprint arXiv:1802.03754, 2018.
[18] A. Johnson, T. Pollard, L. Shen, L. Lehman, M. Feng, *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific data*, 2016.
[19] M. Singer, C. Deutschman, C. Seymour, *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, 2016.
[20] T. Pollard, A. Johnson, J. Raffa, L. Celi, R. Mark, *et al.*, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Scientific data*, 2018.
[21] A. Fernández-Torras, M. Duran-Frigola, M. Bertoni, *et al.*, "Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque," *Nature Communications*, 2022.
[22] S. Stanton, W. Maddox, *et al.*, "Bayesian optimization with conformal prediction sets," in *International conference on artificial intelligence and statistics*, 2023.
[23] Y. Malkov and D. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
[24] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

DOI: 10.17148/IJARCCE.2026.15697

[25] D. Bender and K. Sartipi, "HL7 FHIR: An agile and RESTful approach to healthcare information exchange," in *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, 2013.

[26] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification." *Journal of machine learning research*, 2009.

[27] L. Wang, K. Yang, C. Li, L. Hong, *et al.*, "Ordisco: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[28] U. Food and D. Administration, "Artificial intelligence and machine learning in software as a medical device," US Food & Drug Administration: Silver Spring …, 2021, 2021.

[29] A. Kumar, K. Chaudhary, E. Jaiswal, P. Rai, K. Gupta, and R. Er, "Explainable artificial intelligence in healthcare," *International Journal for …, 2024*, 2024.

[30] B. McCall, "COVID-19 and artificial intelligence: Protecting health-care workers and curbing the spread," *The Lancet Digital Health*, 2020.

