The Cognitive Computing Model Based on Machine Learning Algorithms in Artificial Intelligence Environments

Xiaolong Huang
Qingyuan Polytechnic, Qingyuan 511510, Guangdong, China
Corresponding author: Xiaolong Huang (e-mail: dret82@163.com)

ABSTRACT This study proposes CognitiveNet, a cognitive computing model designed for processing multi-modal long-sequence data. The framework integrates BiLSTM for sequence context extraction, low-rank self-attention (LRSA) for key segment focus, cognitive memory units (CMU) for cross-segment historical information retrieval, and sparse gating for efficient feature selection. Experimental evaluation on the MIMIC-III dataset for classification and NASA CMAPSS dataset for regression demonstrates that CognitiveNet outperforms baseline models, achieving higher accuracy, lower error rates, and improved inference efficiency while maintaining interpretability. Ablation and disturbance experiments confirm the critical contributions of CMU and sparse gating in capturing long-term dependencies and prioritizing core features. While the study focuses on healthcare and aero-engine prognostics, the proposed model provides a generalizable framework for multi-sensor signal processing, with potential applications in electromagnetic signal monitoring, antenna array diagnostics, and high-frequency system predictive maintenance, where efficient handling of long, complex sequences is crucial. This work highlights the capacity of cognitive computing to support real-time, interpretable, and high-precision analysis of complex temporal signals in industrial and engineering contexts.

INDEX TERMS cognitive computing, multi-modal long sequences, CMU, electromagnetic signal processing

I. INTRODUCTION

ARTIFICIAL intelligence (AI) has made remarkable progress in recent years, especially in the domains of deep learning and natural language processing (NLP) [1,2]. Although these technologies perform well on specific tasks, they typically rely on large volumes of labeled data and exhibit limitations when processing complex multimodal information [3-5]. To address these problems, cognitive computing, as a computing paradigm that simulates human cognitive processes, has gradually attracted the attention of researchers [6,7]. For example, Liu & Tang [8] identified genes related to the Mitogen-Activated Protein Kinase (MAPK) signaling pathway that were upregulated in glioma. They used a consensus clustering method to classify glioma into subtypes. They further explored differences in survival rates, tumor stemness, and immune microenvironment characteristics across distinct subtypes. Liu & Tang [9] conducted pan-cancer genomics and clinical correlation studies on disulfide poisoning and its related cell death genes. Their research revealed potential biomarkers for the mechanism of disease occurrence. Liu [10] investigated the association between sleep duration and the risk of developing depression. These studies analyse high-dimensional biomedical data and highlight the importance of cross-modal information integration and pattern recognition for clinical decision support. This highly aligns with the goal of the proposed machine learning-based cognitive computing model—mining key information from multi-modal time-series data and supporting reasoning for complicated tasks. Cognitive computing aims to mimic human processes of perception, understanding, reasoning, and decision-making, and to handle uncertainty and ambiguity [11-13]. Unlike traditional AI systems, cognitive computing systems are not just tools for performing predetermined tasks; they are more like intelligent agents with self-learning and adaptive capabilities [14,15]. They can learn from experience, understand complex contexts, and make reasonable inferences when facing uncertain information [16]. As Gou et al. [17] pointed out, cognitive systems were built on AI capabilities, excelling in understanding complex and unstructured data, learning from interactions, and offering explanations and suggestions. Cognitive computing has demonstrated broad application potential across multiple domains [18]. In the medical field, cognitive computing is used in clinical decision support systems to help doctors provide advice and support during diagnosis and treatment [19-21]. In terms of business intelligence, it is applied in financial analysis, risk assessment, and customer relationship management to improve the efficiency and accuracy of decision-making [22,23]. In addition, cognitive computing has also made significant developments in fields such as NLP, computer vision, and intelligent manufacturing [24]. Despite the initial achievements of cognitive computing in multiple areas, it still faces some challenges. First, existing cognitive computing models often lack effective cross-segment memory and long-range dependency modeling capabilities when processing long-sequence and multi-modal data. In engineering, this limitation may hinder the accurate identification of periodic defects in long fabric rolls or the prediction of equipment fatigue. Second, the model's interpretability and transparency remain unresolved issues, especially in high-risk fields such as healthcare and finance. Finally, the real-time performance and inference efficiency of the model are also key issues in practical applications. To deal with the above issues, this study proposes an innovative cognitive computing model. Its core structure is based on the bidirectional long short-term memory (BiLSTM), combined with the low-rank self-attention (LRSA), cognitive memory unit (CMU), and sparse gating mechanism. Among these components, the BiLSTM is used to capture sequence context information; the LRSA can focus on key segments and substantially reduce computational complexity; the CMU provides cross-segment memory capability to retain and retrieve historical key information; the sparse gating mechanism automatically selects the most valuable feature channels, thus improving the model's efficiency and interpretability. The model can handle various sequence prediction tasks and demonstrates strong versatility and scalability. This study selects two representative datasets to verify the model's effectiveness and versatility: Medical Information Mart for Intensive Care III (MIMIC-III) and National Aeronautics and Space Administration (NASA) Commercial Modular Aero-Propulsion System Simulation (CMAPSS). The MIMIC-III dataset provides rich Intensive Care Unit (ICU) multi-modal sequence data, including vital signs, laboratory tests, and text records, which is used for classification or regression tasks. The NASA CMAPSS dataset contains multi-sensor time series and operating conditions, which predict continuous variables such as remaining useful life (RUL). Experiments on these two datasets evaluate the model's prediction performance, feature-selection capability, and inference efficiency in different sequence tasks, thus fully demonstrating its versatility and application potential.

II. MATERIALS AND METHODS

A. DESIGN OF THE COGNITIVE COMPUTING MODEL BASED ON MACHINE LEARNING

The core of cognitive computing is to simulate human perception, memory, attention, and decision-making process. This model achieves this goal through four key components. (1) BiLSTM simulates human perception of time-series information and captures contextual dependencies through bidirectional information flow; (2) LRSA simulates the selective attention mechanism and identifies key segments in complex information streams; (3) the CMU simulates the long-term memory function of the human brain and enables the storage, retrieval and integration of cross-segment information, forming the basis for complex reasoning; and (4) the sparse gating mechanism simulates feature selection in human cognition by automatically filtering irrelevant information and focusing on decision-related feature channels. This design enables the model to retain the data-fitting capability of conventional deep learning while achieving the understanding, memory, and reasoning of multi-modal long-sequence data through the simulating of human cognitive mechanism, thus achieving the essential requirements of cognitive computing. The cognitive computing model proposed in this study (denoted as CognitiveNet) aims to effectively process long sequences and multi-modal inputs. Meanwhile, this model fully uses cross-segment information and achieves high-precision and high-interpretability sequence prediction. The model's core consists of BiLSTM, LRSA, CMU, and sparse gating, and its overall structure is illustrated in Figure 1.

The model first inputs the input sequence $X = [x_1, x_2, \ldots, x_T]$ into BiLSTM. Among them, $x_t \in \mathbb{R}^d$ represents the feature vector of the t-th timestep; d refers to the feature dimension, and T denotes the sequence length. BiLSTM consists of two LSTM networks, forward and backward, that capture both past and future information of a sequence. Its hidden state can be written as:

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \tag{1}$$

$$\overrightarrow{h}_t = \mathrm{LSTM}_{\mathrm{forward}}(x_t, \overrightarrow{h}_{t-1}) \tag{2}$$

$$\overleftarrow{h}_t = \mathrm{LSTM}_{\mathrm{forward}}(x_t, \overleftarrow{h}_{t-1}) \tag{3}$$

$\overrightarrow{h}_t, \overleftarrow{h}_t \in \mathbb{R}^{d_h}$ refer to the forward and backward hidden states, respectively; d_h stands for the hidden state dimension. The bidirectional nature of BiLSTM ensures that each time step is represented with both historical and future contexts, providing rich information for subsequent attention and memory modules. To further focus on the time segments that are most important to the prediction task while reducing computational complexity, the model introduces the LRSA mechanism on the BiLSTM output. Let the BiLSTM output sequence be:

$$H = [h_1, h_2, \ldots, h_T] \tag{4}$$

Self-attention (SA) calculations are done by querying matrix Q, key matrix K, and value matrix V. To reduce the computational complexity, a projection matrix $P \in \mathbb{R}^{d_h \times r}$ is introduced to map the high-dimensional representation into a low-rank space of dimension r, where $r \ll d_h$. The transformation is defined as:

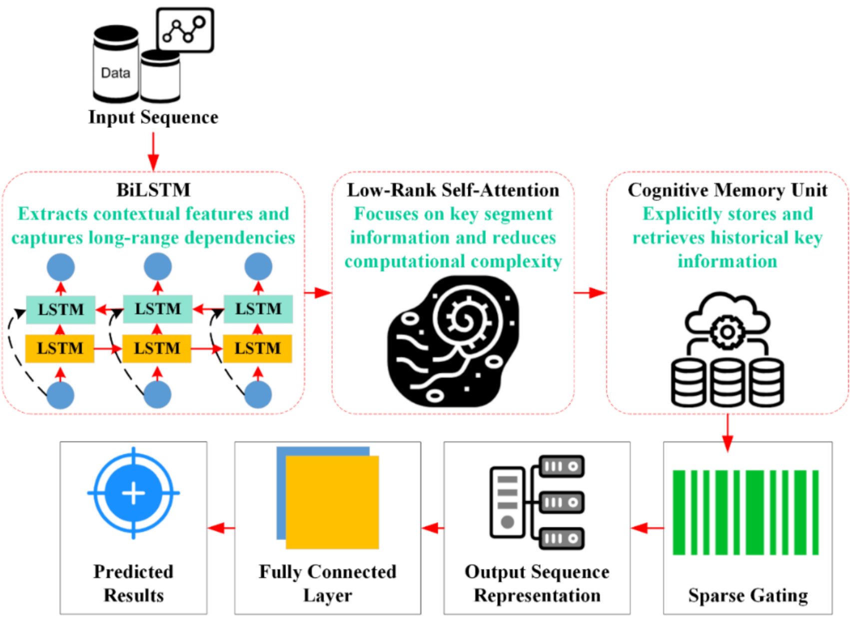

FIGURE 1. Schematic diagram of the cognitive computing model structure

$$Q = HP \tag{5}$$

$$K = HP \tag{6}$$

$$V = H \tag{7}$$

$$\mathrm{SA}(H) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{r}}\right)V \tag{8}$$

The softmax function is used to normalize attention weights, and the scaling factor $\sqrt{r}$ is introduced to stabilize the gradient. Through low-rank projection, the model captures the dependence between time steps while reducing the computational amount of matrix multiplication, making long series processing more efficient. CMU is employed to explicitly store and recall historical information across fragments. The attention mechanism of the CMU generates h'_t, which is then used to update the memory state $M_t \in \mathbb{R}^{d_h}$ as follows:

$$M_t = \gamma_t \odot M_{t-1} + (1 - \gamma_t) \odot h'_t \tag{9}$$

$\gamma_t \in [0,1]^{d_h}$ is the gated vector, which represents the memory retention ratio of the previous timestep; $\odot$ refers to element-by-element multiplication. CMU can effectively utilize historical critical states in sequence prediction, enabling cross-fragment inference. Meanwhile, it can enhance model interpretability, as gating values reflect how much each time step contributes to the prediction outcome. The specific structure is displayed in Figure 2.

The model introduces the sparse gating mechanism to improve the inference efficiency and reduce the influence of redundant features. If the gated vector is $g \in [0,1]^{d_h}$, then the final output feature reads:

$$\tilde{h}_t = g \odot h'_t \tag{10}$$

Sparse gating allows the model to reduce computational effort while maintaining high performance by automatically selecting the most valuable feature channels. At the same time, it provides interpretable information on feature importance. Finally, the sequence after BiLSTM, LRSA, CMU, and sparse gating is expressed as:

$$\tilde{H} = [\tilde{h}_1, \ldots, \tilde{h}_T] \tag{11}$$

This sequence is fed into a fully connected layer or regression layer to complete the classification or continuous value prediction task. The model's synergistic integration combines BiLSTM's global context extraction with LRSA's focus on critical fragments. This architecture further incorporates CMU's cross-fragment memory retention and sparse gating's feature selection capability for optimal information prioritization. The modules complement each other to enable the model to achieve high accuracy, low computational cost, and interpretability when handling multimodal long sequence tasks. The essential difference between CMU and standard GRU/LSTM lies in its design philosophy and functional orientation. Traditional gating unit mainly focuses on information screening at a single time sequence. In contrast, CMU simulates human cognitive memory mechanism, and it has three unique characteristics. (1) Hierarchical memory storage: CMU divides memory into two levels: short-term working memory and long-term semantic memory; the former deals with immediate context, and the latter stores cross-segment abstract knowledge; this is similar to the cooperative mechanism between human hippocampus and neocortex; (2) Cognitive triggered retrieval: Unlike the fixed forget gate in LSTM, the CMU introduces a content-based attention retrieval mechanism that dynamically activates relevant historical memories according to current cognitive

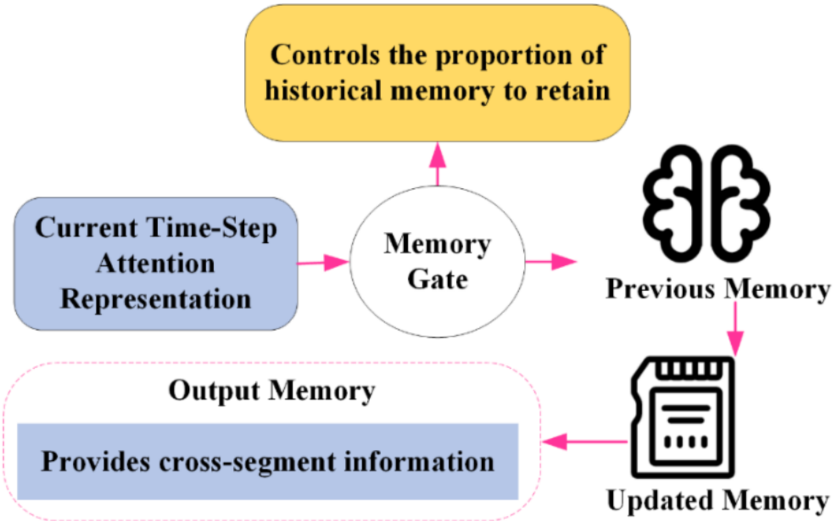

FIGURE 2. The CMU structure

demands and simulates the process of human associative memory; (3) Memory semantic compression: the CMU refines the semantics of long-term memory through adaptive compression coefficient, retains high-value information and filters redundant details.

B. APPLICATION PROCESS OF THE MODEL

To verify the effectiveness and universality of the proposed cognitive computing model in handling different types of sequence tasks, this study selects two representative datasets: MIMIC-III and NASA CMAPSS. These two datasets cover medical multi-modal sequence data and industrial multi-sensor time-series data, respectively; they can also fully evaluate the model's performance in long-sequence modeling, multi-modal feature fusion, and cross-segment memory. Among them, the MIMIC-III dataset is a clinical dataset of ICU patients, containing de-identified information from approximately 40,000 patients. The data types include physiological signals (such as heart rate, blood pressure, respiratory rate, and blood oxygen saturation), laboratory test results (blood tests, electrolytes, blood gas analysis, etc.), clinical texts records (admission records, medical orders, nursing logs), and diagnostic codes. These data have time-series characteristics and multi-modal features, and can be used for classification tasks (such as disease risk prediction) and regression tasks (e.g., physiological indicator prediction). The NASA CMAPSS dataset is provided by NASA and used for predicting aero-engine RUL. The dataset encompasses multi-sensor time-series measurements (temperature, pressure, flow rate, rotational speed, vibration, etc.) and operating conditions (environmental parameters, thrust level, load condition, etc.); these data record the complete operation cycle of the engine from a sound state to degradation and failure. This dataset belongs to a regression task for continuous variable prediction; it can be used to evaluate the model's capabilities in multi-sensor time-series processing, long-range dependency modeling, and key state identification. For the fusion of multimodal data, this study designs a two-stage alignment and fusion mechanism. In the stage of time series alignment, the study constructs a unified time-series framework with minutes as the basic time unit for multi-source heterogeneous data in MIMIC-III dataset. For high-frequency numerical signals (such as heart rate and blood pressure), downsampling is adopted to uniform time granularity. For low-frequency text records (such as doctor's notes), timestamp mapping is used to assign them to the nearest time unit, and 768-dimensional semantic features are extracted by ClinicalBERT. For discrete event data (such as medication records), time decay coding is adopted to make recent events gain higher weight. In the feature fusion stage, in each time unit, the data of different modes are transformed into a vector space of unified dimension through the mode-specific embedding layer. Then, deep fusion is carried out through the gated cross-attention mechanism. Specifically, linear projection is employed for numerical features, and the output of pre-trained language model is used for text features. Next, the multimodal joint representation is formed by weighted fusion of modal balance factor α (automatic learning through training). This design ensures the effective alignment of different modal data at the time series and semantic level, and provides a unified multimodal input sequence for subsequent cognitive computing. The data preprocessing process is suggested in Figure 3.

In terms of dataset division, both MIMIC-III and CMAPSS data are split into the training, validation, and test sets, with the proportions of 70%, 15%, and 15%, respectively. The training set is used for model parameter optimization, the validation set for hyperparameter adjustment and early stopping judgment, and the test set for final performance evaluation. This ensures the objectivity and reproducibility of the experimental results. The processed sequence features are input into the model for further processing, encompassing BiLSTM's contextual feature extraction, LRSA's key segment focusing, CMU's cross-segment

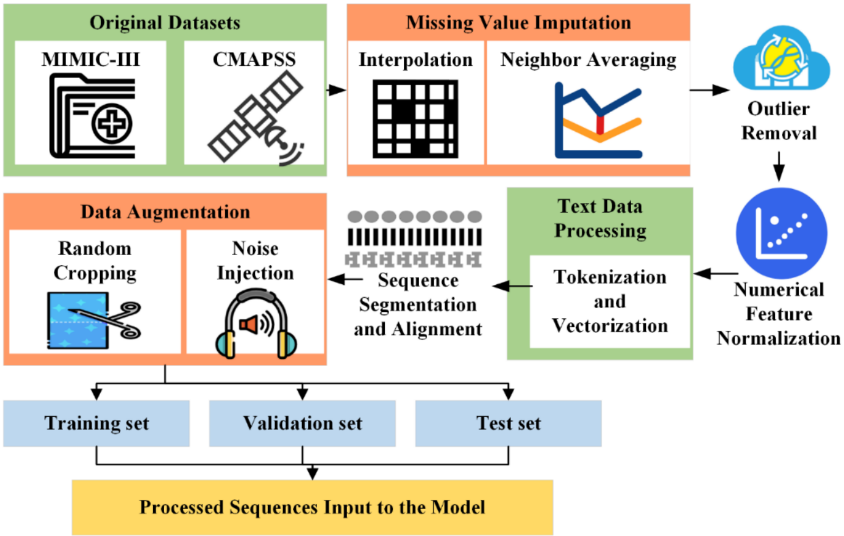

FIGURE 3. The process of data preprocessing

information storage, and sparse gating's feature screening. Finally, prediction results applicable to classification or regression are generated.

C. EXPERIMENTAL DESIGN AND PERFORMANCE VERIFICATION

To systematically evaluate the proposed cognitive computing model and quantify the contribution of each module, this study designs a series of experiments, involving ablation experiments, comparisons with baseline models, and verification in different application scenarios. In the ablation experiments, the CMU is removed, the sparse gating mechanism is turned off, or LRSA is replaced with Full Attention, respectively. This is to analyze the contribution of each module to the overall performance. By comparing the performance of these variants with the complete model, the importance of CMU, sparse gating, and LRSA in capturing crosssegment information, improving inference efficiency, and enhancing interpretability can be clarified. Aiming at the sensitivity of medical data augmentation, the standard augmentation strategy is reformed in clinical adaptability. Firstly, the noise added to physiological signals follows the constraints of medical fluctuation range, such as the heart rate noise is limited to $\pm$ 5 ppm, and the respiratory rate is limited to $\pm$ 2 beats/min, and the clinical rationality of these constraints is verified through cooperation with three intensive care physicians. Next, the random clipping strategy is modified to "event-aware clipping" to ensure that time segments containing key clinical events are fully preserved, while clipping is performed only between events to avoid disrupting the pathological progression. Finally, medical enhancement methods are implemented, including Warping (nonlinear stretching of time axis) and PhysioGAN. To verify the model's universality across diverse tasks, this study selects typical baseline models for comparison, including only BiLSTM, BiLSTM + Attention, and Transformer. All models are trained under the same data preprocessing and training strategies to ensure the fairness of the experiment. Regarding application scenario validation, the MIMIC-III dataset is used for classification or regression tasks, such as disease risk prediction or physiological index prediction. The CMAPSS dataset is employed for continuous value prediction tasks, namely RUL prediction of aero engines. Through the verification on two different types of tasks, the generality and robustness of the model in multimodal long sequence tasks can be evaluated. In terms of performance evaluation, this study uses several indicators. For classification tasks, accuracy, recall, and F1-score are utilized to measure prediction effectiveness, which is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

TP , TN , FP , and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively. The Area Under Curve (AUC) indicator is also calculated to measure the ability to discriminate and discriminate. For regression tasks, the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used for evaluation, using the following equations:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{16}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{17}$$

$\hat{y}_i$ denotes the predicted value of the model, y_i refers to the true value, and N stands for the total number of samples. To evaluate the model's response performance in real-world applications, inference latency is also introduced, recording the average time it takes for the model to complete predictions for a single input sequence. The model is trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and a learning rate decay strategy. The batch size is 64, the maximum number of training epochs is 100, and the early stopping control stops the training early if the continuous validation set metrics do not increase. Hyperparameters such as BiLSTM hidden state dimension, attention low-rank dimension, and CMU gating policy are optimized for different datasets and tasks. The specific parameter settings are exhibited in Table 1. The proposed model uses BiLSTM structure to learn with complete sequence information in the training stage. However, in the actual deployment of real-time prediction scenarios, two adaptive strategies are provided. (1) The switching mechanism of inference mode: when deployed in real-time environment, the model can automatically switch to one-way LSTM inference mode, relying only on historical information for prediction, sacrificing some accuracy for strict real-time; (2) Sliding window prediction: In the quasi-real-time scene with slight delay, a fixed-length sliding time window is adopted; only the data that has occurred in the window is used for prediction; and the window size can be dynamically adjusted according to the real-time requirements.

TABLE 1. Setting of training parameters for the cognitive computing model

Parameter	Setting value
Optimizer	Adam
Initial learning rate	1×10^{-4}
Batch size	64
Maximum number of training epochs	100
BiLSTM hidden state dimension	128
Attention low-rank dimension r	32
CMU gating policy	Sigmoid
Data augmentation	Random cropping / adding noise

III. RESULT

A. COMPREHENSIVE PERFORMANCE COMPARISON

To comprehensively evaluate the classification performance and efficiency of the proposed cognitive computing model (CognitiveNet), this study compares it with various mainstream sequence modeling methods on the MIMIC-III dataset. These methods include basic BiLSTM, BiLSTM+Attention, and Transformer. The comparison results are demonstrated in Figure 4.

In Figure 4(A), the proposed CognitiveNet achieves the optimal performance in all six classification performance indicators on the MIMIC-III dataset. Among these indicators, accuracy increases from 0.837 of the basic BiLSTM to 0.902, precision rises from 0.842 to 0.898, F1-score goes up from 0.839 to 0.895, and AUC even increases to 0.941. This reflects the model's overall improvement in classification performance. The advantage arises from the high efficiency of the LRSA module in capturing key temporal information, as well as the CMU's ability to store historical information for a long time and conduct cross-segment inference. Thus, these advantages make the model superior to traditional sequence models in long-term dependency and global context understanding. Figure 4(B) shows the performance of different models in three aspects: inference latency, parameter count, and throughput. The inference latency of CognitiveNet is 16.4 milliseconds (ms), which is only slightly higher than that of the basic BiLSTM (14.5 ms) but much lower than that of the Transformer (27.3 ms). This indicates that while maintaining high accuracy, the model's inference speed can still meet the needs of real-time or near-real-time applications. Concerning model scale, the parameter count of CognitiveNet is 8.9 million (M), which is close to that of the BiLSTM (8.2 M) and remarkably smaller than that of the Transformer (12.4 M); this is attributed to the role of low-rank attention in reducing redundant computations. In terms of throughput, CognitiveNet can process 645 samples per second (samples/s), which is about 50% higher than the Transformer's 430 samples/s. This result indicates that the model is well suited for deployment in high-concurrency scenarios. As the positive-negative sample ratio of MIMIC-III is 1:7.3, the precision-recall curve is shown in Figure 4(C). The AUPRC of CognitiveNet reaches 0.742, which is 27.7% higher than that of BiLSTM (0.581). In the high sensitivity range (Recall $\geq$ 0.9) commonly used in clinical decision-making, the precision of CognitiveNet remains at 0.61, while BiLSTM drops to 0.39; this means that there are 22 false positives in every 100 alarms, which can significantly reduce ICU alarm fatigue. To verify the adaptability of CognitiveNet in continuous prediction tasks (RUL prediction of aero engines), this study compares it with various baseline models on the NASA CMAPSS dataset. Figure 5(A) shows four commonly used performance indicators for regression tasks, while Figure 5(B) presents efficiency-related indicators such as model parameter count, inference latency, and throughput.

As shown in Figure 5, in the RUL prediction task on the CMAPSS dataset, CognitiveNet outperforms baseline models in all regression indicators. Especially in RMSE and MAE, it reduces the values by 2.12 and 1.80, respectively, compared with the Transformer, indicating a substantial improvement in the model's error control. The R^2 reaches 0.938, which means the predicted results have a higher degree of fit with the real values. The MAPE drops to

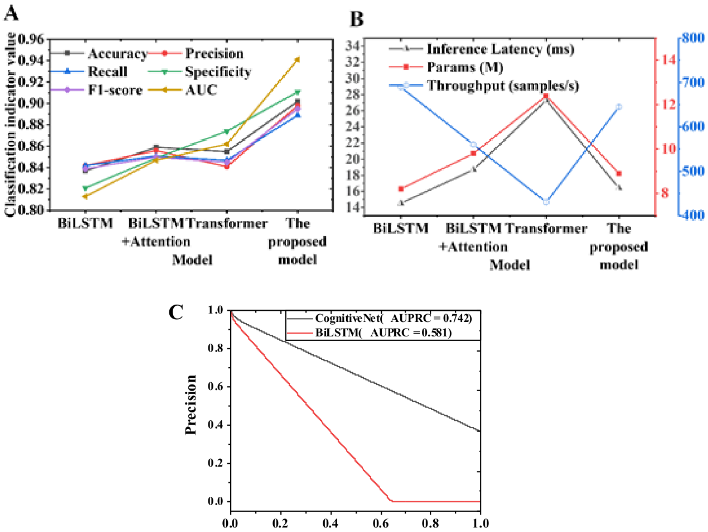

FIGURE 4. Performance comparison of models on the MIMIC-III dataset (A. Classification performance comparison; B. Operational efficiency comparison; C. Precision-Recall curve)

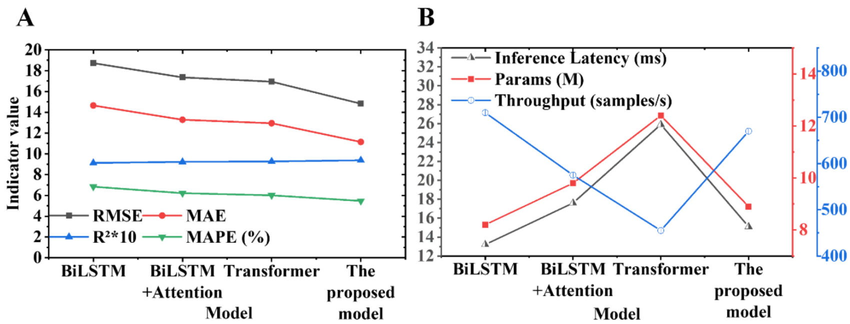

FIGURE 5. Performance comparison of CMAPSS datasets (A. Predictive performance comparison; B. Operational efficiency comparison)

5.47%, which also implies that the model has a low relative error in different life intervals and is more valuable for life estimation in engineering maintenance. Similar to the classification results on MIMIC-III, this advantage mainly comes from the combination of LRSA and CMU. This combination enables the model to retain and utilize key state information in long time series, thereby capturing the degradation trend of engine performance more accurately. In terms of efficiency, the inference latency of CognitiveNet is 15.1 ms, which is about 41.7% faster than that of the Transformer. While ensuring accuracy, its response speed is more suitable for real-time health monitoring and prediction tasks. The parameter count remains at 8.9 M, consistent with the results in the MIMIC-III experiment, illustrating that the model's computational scale does not expand significantly due to task changes. The throughput reaches 670 samples/s, close to the BiLSTM's 710 samples/s. This shows that the model can still support high-frequency inference requirements in industrial deployment scenarios. Overall, CognitiveNet maintains a good balance between accuracy and efficiency, and no trade-off between computational efficiency and predictive performance is observed. In order to further clarify the result that LRSA can reduce the computational complexity, this study adds Floating-point Operations (FLOPs) and Video Random Access Memory (VRAM) indicator tests. The experiment is based on NVIDIA RTX 3090 graphics processing unit (GPU), and the results are outlined in Table 2. The FLOPs of the proposed model are only 10.8 G, which is much lower than that of BiLSTM+Attention (24.6 G) and Transformer (35.2 G), and the memory occupation is close to that of the basic BiLSTM, significantly lower than that of Transformer (5.4 gigabytes (GB)). This is because LRSA reduces the computational complexity of the attention mechanism through low-rank projection, thereby significantly decreasing the floating-point operation for matrix multiplication. At the same time, sparse gating screens the core feature channels, further reducing the memory occupation.

TABLE 2. Comparison of efficiency indicators of each model (sequence length T=512)

Model	Inference latency (ms)	FLOPs (G)	VRAM occupation (GB)
BiLSTM	14.5	8.3	2.1
BiLSTM+Attention	19.7	24.6	3.8
Transformer	27.3	35.2	5.4
CognitiveNet	16.4	10.8	2.5

B. ANALYSIS OF ABLATION EXPERIMENTS

This study conducts ablation experiments on the MIMIC-III and CMAPSS datasets to verify the contribution of each key module to the performance of CognitiveNet. The experiments mainly examine the roles of three core components: ① CMU, ② sparse gating mechanism, and ③ LRSA. In the ablation experiments, one component is removed or replaced each time while the others remain unchanged, to evaluate the module's impact on performance and efficiency. Figure 6 and Figure 7 present the ablation results of the main performance indicators for the classification task (MI MIC-III) and regression task (CMAPSS), respectively.

In the MIMIC-III's classification task shown in Figure 6, the complete CognitiveNet model performs best in six indicators, such as accuracy, precision, recall, specificity, F1-score, and AUC. Among these, the AUC reaches 0.941, indicating that the model has strong discriminative ability in distinguishing high-risk and low-risk samples. The performance decline is most obvious after removing the CMU; Accuracy decreases by approximately 1.6% and recall drops by 1.8%. The reason is that the changes in patient conditions in ICU multi-modal data have cross-time-segment dependencies, and the lack of an explicit memory unit leads to insufficient capture of long-term dependent information. The removal of sparse gating causes a slight performance loss, with the F1-score decreasing by 1.3%. This shows that the mechanism contributes to feature channel screening and noise suppression, but its role is not as core as that of the CMU. After replacing LRSA with Full Attention, the performance changes little (AUC decreases by 0.3%). However, the inference latency increases markedly—from 15.1 ms to 21.3 ms—and the throughput drops by 32%. The reason is that the computational complexity of Full Attention grows with the square of the sequence length, resulting in significant computational overhead. In Figure 7, the aero-engine RUL prediction task on the CMAPSS dataset shows that the complete model achieves the best results in regression indicators such as RMSE, MAE, MAPE, and R^2. Among these, the R^2 is 0.938, demonstrating that the model has high fitting accuracy for the degradation trend. After removing the CMU, the RMSE increases by nearly 0.91, and the MAPE rises by 0.55%. This is because RUL prediction also relies on the evolution of operating states over a long-time scale, and the lack of a memory mechanism leads to an increase in prediction deviations in the degradation stage. The performance change after removing sparse gating is still small, but its retention is meaningful for feature compression and computational stability. This is particularly important for prediction robustness under cross-working conditions. After replacing LRSA with Full Attention, the RMSE decreases slightly, but the latency increases significantly (+ 6.2 ms). The reason is the same as that in the MIMIC-III experiment. Full Attention introduces excessive computational burden in long sequences, which is not conducive to real-time prediction scenarios.

C. INTERPRETABILITY ANALYSIS

To explore the decision-making basis of the proposed cognitive computing model, this study conducts a visualization analysis on the feature channels selected by the sparse gating and the historical information retrieved by the CMU. The visualization results are depicted in Figure 8, which helps understand the model's feature selection and information retrieval behaviors across various tasks.

In Figure 8(A), the high-weight features automatically selected by the sparse gating mainly concentrate on heart rate (0.92), blood oxygen (0.87), and blood pressure (0.81). These channels correspond to key physiological signals in the task, suggesting that the model can capture the core information for judging disease risk. Different time windows' average weights show that these key features are continuously focused on in multiple time segments of the sequence, ensuring the retention of long-term dependent information. The average weights of various sample batches illustrate that the sparse gating can still stably select core features in diversified data, which proves the model has good generalization ability. Low-weight channels, such as body temperature and blood glucose, although their information has reference value, contribute relatively little to the prediction task. The model automatically weakens their impact, which helps reduce noise interference and the risk of overfitting. This demonstrates the effectiveness of the sparse gating in feature selection. The high-weight features are highly consistent with the key indicators known in the medical field, indicating the model has good interpretability. At the same time, by selectively focusing on core channels, the model also optimizes computational efficiency. Figure 8(B) displays the retrieval of historical information from different time segments by the CMU. The activation intensities at time steps t-5 and t-15 are the highest (0.93 and 0.91), illustrating that the model focuses on the key stages of engine degradation during predicting RUL. The average activation of diverse time windows reveals that the CMU continuously retrieves signals from the early degradation stage and high-load stage. Different sample batches' average activation indicates that the model can still maintain attention to key time segments even under various working conditions. In contrast, the activation of t-10 and t-20 is slightly lower, which means the model has weak dependence on information from non-key stages. This reflects the model's priority attention to important information in long-sequence tasks. By visualizing the retrieval intensity of the CMU, we reveal how the model performs cross-segment memory and captures key temporal patterns. The high activation intensity corresponds to the important time segments in the task, which verifies the model's rationality and interpretability, while ensuring the accuracy and stability of regression prediction. Targeted disturbances are imposed on high-weight features (heart rate, blood oxygen, blood pressure) and low-weight features (body temperature, blood glucose) selected by sparse gating:

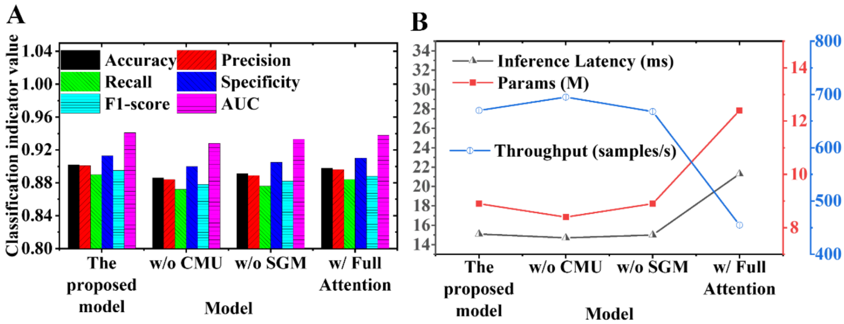

FIGURE 6. Ablation experiment results on the MIMIC-III dataset (A. Classification performance; B. Operational efficiency)

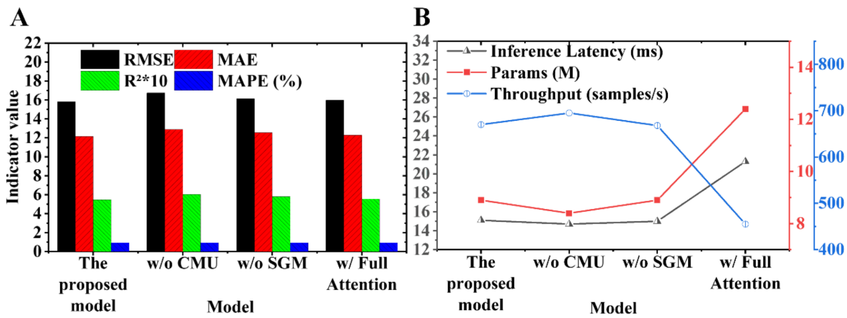

FIGURE 7. Ablation experiment results on the CMAPSS dataset (A. Predictive performance; B. Operational efficiency)

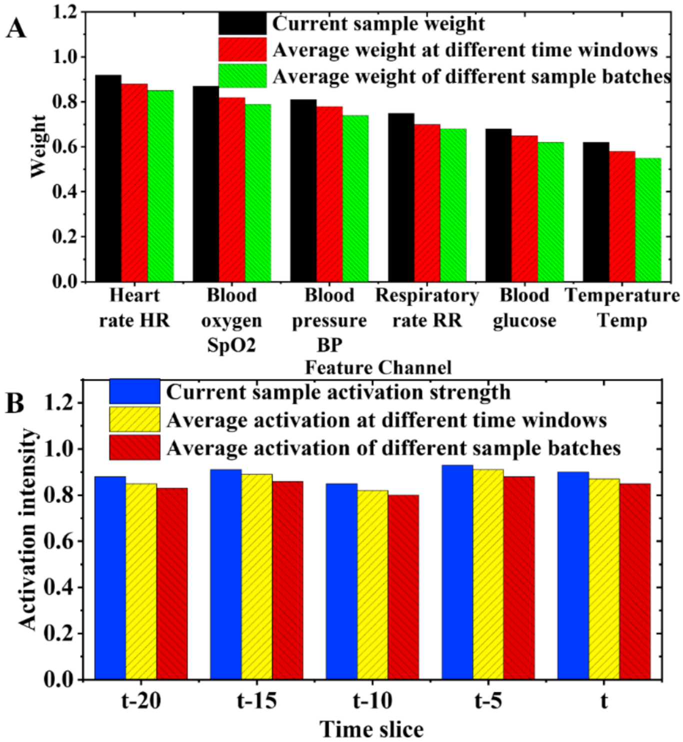

FIGURE 8. Visualization of historical information (A. Sparse gating analysis on MIMIC-III dataset; B. CMU analysis on CMAPSS dataset)

$\pm$ 10% random noise is added to key features in the MIMIC-III test set to simulate clinical data fluctuations. The same amplitude of noise is added to low-weight features, and changes in model prediction performance are compared. Meanwhile, masking processing (setting to zero) is applied to features in high-activation time segments (t-5, t-15) of

CMU, and changes in RUL prediction error are observed. The experimental results are listed in Table 3.

TABLE 3. Results of feature disturbance experiments

Disturbed object	Task type	Accuracy change (ΔAcc)	F1-score change (ΔF1)	RMSE change (ΔRMSE)	MAPE change (ΔMAPE)
High-weight features (heart rate + blood oxygen + blood pressure)	MIMIC-III classification	-6.2%	-5.8%	-	-
Low-weight features (body temperature + blood glucose)	MIMIC-III classification	-0.9%	-0.7%	-	-
High-activation time segments (t-5 + t-15)	CMAPSS regression	-	-	+1.23	+0.78%
Low-activation time segments (t-10 + t-20)	CMAPSS regression	-	-	+0.31	+0.22%

Disturbance experiments show that model performance decreases significantly when core features with high weight or high activation are disturbed. In contrast, disturbances on features with low weight or low activation exert minimal impact on performance, with a ΔAcc of only -0.9%. This confirms that the features assigned high weights by the model constitute the primary basis for decision-making, rather than a circular argument based on subjective assumptions. In addition, combined with clinical diagnosis and treatment guidelines, the rationality of the model's feature selection is further verified. High-weight features are completely consistent with the key indicators recognized in the medical field. Moreover, the disturbance experiments quantify the contribution of these features to prediction results, providing rigorous empirical support for interpretability analysis.

IV. CONCLUSION

This study proposes a cognitive computing model (CognitiveNet) for multi-modal long-sequence data, aiming to address the shortcomings of traditional sequence modeling methods in capturing long-term temporal dependencies, interpretability, and feature interaction efficiency. The model captures temporal context information through BiLSTM and integrates CMU to realize the storage and retrieval of cross-segment key information. Meanwhile, this model introduces LRSA to optimize feature interaction efficiency and adopts sparse gating to screen core channels. The resulting architecture achieves a balance between interpretability, predictive performance, and computational efficiency, providing a generalizable cognitive computing framework for multi-modal long-sequence tasks. In experimental verification, CognitiveNet shows remarkable advantages in the MIMIC-III dataset's classification task and the CMAPSS dataset's regression task. The model's success in handling complex temporal dependencies and multi-source information. Despite the above achievements, the model still has certain limitations. First, the datasets verified in this study only include two types of representative tasks (MIMIC-III and CMAPSS). The model's applicability in more fields, multi-modal combinations, and larger-scale data still needs further verification. Second, the computational overhead of CMU and LRSA under extremely long sequences or ultra-large models may become a limiting factor. In the future, more efficient memory compression strategies or sparse attention optimization methods can be explored. Moreover, applying the model to specific industrial datasets to verify its cross-domain robustness will be an important research direction. In addition, the current interpretability analysis mainly relies on feature weights and activation heatmaps. Subsequent studies can combine methods such as Shapley value or adversarial sample analysis to enhance decision transparency.

AUTHOR CONTRIBUTIONS

Xiaolong Huang contributed to the study conception and design. Material preparation, data collection and analysis were performed by Xiaolong Huang. The first draft of the manuscript was written by Xiaolong Huang and Xiaolong Huang commented on previous versions of the manuscript. Xiaolong Huang read and approved the final manuscript.

CONFLICTS OF INTEREST

Xiaolong Huang declares that there is no conflict of interest. This study does not receive any commercial funding that may affect the objectivity of the study. Xiaolong Huang has no financial or personal relationships with any organizations, products or services mentioned in the study that may affect the impartiality and scientificity of this study. All experimental data and results are true and reliable without any artificial modification or selective reporting.

FUNDING

Guangdong Provincial Educational Science Planning Project (Higher Education Special Project) "Research on the Current Situation and Improvement Strategies of Coding Education in Underdeveloped Areas in the Age of Intelligence - Taking Qingyuan City in Northern Guangdong as an Example" (No.: 2022GXJK114); Qingyuan Education Science Research Project "Research on Integrating Artificial Intelligence Based on STEM Concepts into 'Python Programming Practice' Teaching under the New Engineering Background" (No.: 23-118).

ACKNOWLEDGEMENTS

This study gains valuable support and guidance from many experts. Special thanks are extended to the MIMIC-III database maintenance team for providing high-quality clinical data support. Gratitude is expressed to the team led by Professor Zhang from the Department of Critical Care Medicine, the First Affiliated Hospital of Zhejiang University School of Medicine, for providing professional opinions and evaluations during the clinical validation phase. Thanks are also given to anonymous reviewers for their constructive suggestions, which further improve this study. In addition, the authors appreciate all members of the laboratory for their inspiration and advice during discussions.

REFERENCES

[1] M. Srivani, A. Murugappan, and T. Mala, "Cognitive computing technological trends and future research directions in healthcare—A systematic

literature review," Artificial Intelligence in Medicine, vol. 138, Art. no. 102513, 2023, doi: 10.1016/j.artmed.2023.102513.

[2] A. R. Javed, A. Saadia, H. Mughal, T. R. Gadekallu, M. Rizwan, P. K. R. Maddikunta, et al., "Artificial intelligence for cognitive health assessment: State-of-the-art, open challenges and future directions," Cognitive Computation, vol. 15, no. 6, pp. 1767–1812, 2023, doi: 10.1007/s12559-023-10153-4.

[3] X. Chen, H. Xie, X. Tao, F. L. Wang, M. Leng, and B. Lei, "Artificial intelligence and multimodal data fusion for smart healthcare: Topic modeling and bibliometrics," Artificial Intelligence Review, vol. 57, no. 4, p. 91, 2024, doi: 10.1007/s10462-024-10712-7.

[4] S. Wang, X. He, Z. Jian, J. Li, C. Xu, Y. Chen, et al., "Advances and prospects of multi-modal ophthalmic artificial intelligence based on deep learning: A review," Eye and Vision, vol. 11, no. 1, p. 38, 2024, doi: 10.1186/s40662-024-00405-1.

[5] X. Guo, Z. Sun, Y. Zhu, and C. Lee, "Zero-biased bionic fingertip E-Skin with multimodal tactile perception and artificial intelligence for augmented touch awareness," Advanced Materials, vol. 36, no. 39, Art. no. 2406778, 2024, doi: 10.1002/adma.202406778.

[6] S. I. Ao, M. Hurwitz, and V. Palade, "Cognitive computing and business intelligence applications in accounting, finance and management," Big Data and Cognitive Computing, vol. 9, no. 3, p. 54, 2025, doi: 10.3390/bdcc9030054.

[7] S. Neethirajan, "From predictive analytics to emotional recognition–The evolving landscape of cognitive computing in animal welfare," International Journal of Cognitive Computing in Engineering, vol. 5, pp. 123–131, 2024, doi: 10.1016/j.ijcce.2024.02.003.

[8] H. Liu and T. Tang, "MAPK signaling pathway-based glioma subtypes, machine-learning risk model, and key hub proteins identification," Scientific Reports, vol. 13, no. 1, Art. no. 19055, 2023, doi: 10.1038/s41598-023-45774-0.

[9] H. Liu and T. Tang, "Pan-cancer genetic analysis of disulfidptosis-related gene set," Cancer Genetics, vol. 278, pp. 91–103, 2023, doi: 10.1016/j.cancergen.2023.10.001.

[10] H. Liu, "Association between sleep duration and depression: A Mendelian randomization analysis," Journal of Affective Disorders, vol. 335, pp. 152–154, 2023, doi: 10.1016/j.jad.2023.05.020.

[11] Z. Kowalczuk and M. Czubenko, "Cognitive motivations and foundations for building intelligent decisionmaking systems," Artificial Intelligence Review, vol. 56, no. 4, pp. 3445–3472, 2023, doi: 10.1007/s10462-022-10255-9.

[12] M. Ren, N. Chen, and H. Qiu, "Human-machine collaborative decision-making: An evolutionary roadmap based on cognitive intelligence," International Journal of Social Robotics, vol. 15, no. 7, pp. 1101–1114, 2023, doi: 10.1007/s12369-023-01020-1.

[13] C. Gonzalez, "Building human-like artificial agents: A general cognitive algorithm for emulating human decision-making in dynamic environments," Perspectives on Psychological Science, vol. 19, no. 5, pp. 860–873, 2024, doi: 10.1177/17456916231196766.

[14] L. Jiao, M. Ma, P. He, X. Geng, X. Liu, F. Liu, et al., "Brain-inspired learning, perception, and cognition: A comprehensive review," IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 4, pp. 5921–5941, 2024, doi: 10.1109/TNNLS.2024.3401711.

[15] X. Zou and S. Zou, "Cognitive computing smart system: How to remove ambiguities," Journal of Computational and Cognitive Engineering, vol. 3, no. 1, pp. 34–42, 2024, doi: 10.47852/bonviewJCCE2202424.

[16] F. A. Alijoyo, S. Janani, K. Santosh, S. N. Shweihat, N. Alshammry, J. V. N. Ramesh, et al., "Enhancing AI interpretation and decision-making: Integrating cognitive computational models with deep learning for advanced uncertain reasoning systems," Alexandria Engineering Journal, vol. 99, pp. 17–30, 2024, doi: 10.1016/j.aej.2024.04.073.

[17] H. Gou, G. Zhang, E. P. Medeiros, S. K. Jagatheesaperumal, and V. H. C. de Albuquerque, "A cognitive medical decision support system for IoT-based human-computer interface in pervasive computing environment," Cognitive Computation, vol. 16, no. 5, pp. 2471–2486, 2024, doi: 10.1007/s12559-023-10242-4.

[18] M. Rinaldi, S. Wang, R. S. Geronel, and S. Primatesta, "Application of Task Allocation Algorithms in Multi-UAV Intelligent Transportation Systems: A Critical Review," Big Data and Cognitive Computing, vol. 8, no. 12, p. 177, 2024, doi: 10.3390/bdcc8120177.

[19] R. H. Kumar and B. Rajaram, "Design and simulation of an edge compute architecture for IoT-based clinical decision support system," IEEE Access, vol. 12, pp. 45456–45474, 2024, doi: 10.1109/ACCESS.2024.3380906.

[20] R. K. Kanna and A. O. Salau, "New Cognitive Computational Strategy for Optimizing Brain Tumour Classification using Magnetic Resonance Imaging Data," Intelligence-Based Medicine, vol. 11, Art. no. 100215, 2025, doi: 10.1016/j.ibmed.2025.100215.

[21] A. G. Levra, M. Gatti, R. Mene, D. Shiffer, G. Costantino, M. Solbiati, et al., "A large language model-based clinical decision support system for syncope recognition in the emergency department: A framework for clinical workflow integration," European Journal of Internal Medicine, vol. 131, pp. 113–120, 2025, doi: 10.1016/j.ejim.2024.09.017.

[22] J. Liu and S. Fu, "Financial big data management and intelligence based on computer intelligent algorithm," Scientific Reports, vol. 14, no. 1, p. 9395, 2024, doi: 10.1038/s41598-024-59244-8.

[23] D. Ozay, M. Jahanbakht, A. Shoomal, and S. Wang, "Artificial Intelligence (AI)-based Customer Relationship Management (CRM): A comprehensive bibliometric and systematic literature review with outlook on future research," Enterprise Information Systems, vol. 18, no. 7, Art. no. 2351869, 2024, doi: 10.1080/17517575.2024.2351869.

[24] L. Orynbay, G. Bekmanova, B. Yergesh, A. Omarbekova, A. Sairanbekova, and A. Sharipbay, "The role of cognitive computing in NLP," Frontiers in Computer Science, vol. 6, Art. no. 1486581, 2025, doi: 10.3389/fcomp.2024.1486581.