GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Article

Construction And Optimization Of AI-Based Real-Time Clinical Decision Support System

Danyating Shen [1,*]

[1] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
* Correspondence: Danyating Shen, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Abstract: With the rapid development of AI technology, its role in clinical decision support systems (CDSS) has become increasingly prominent. This paper focuses on the construction ideas of real-time clinical decision support systems, elaborates in detail the key elements of multi-source data fusion, high-precision prediction and real-time reasoning, and proposes improvement paths to accelerate system response, strengthen model iteration and optimize interaction structure, aiming to provide reliable, efficient and sustainable technical support for smart medical practice.

Keywords: AI; clinical decision support system; real-time reasoning; data integration; model optimization

Received: 13 November 2025
Revised: 24 November 2025
Accepted: 02 December 2025
Published: 09 December 2025

1. Introduction

As an important tool for improving the quality and efficiency of medical services, the clinical decision support system is undergoing a transformation from the traditional model to the real-time clinical decision support system and then towards the "AI clinical decision support system". With the development of AI technology, the system has been continuously optimized in terms of data processing capabilities, reasoning depth, architecture design and response mechanisms, providing more solid support for clinical practice. This article focuses on the real-time clinical decision support system based on AI technology, introduces its architecture principle and key optimization strategies, aiming to explore the optimization path for its efficient implementation in medical scenarios.

2. The Core Advantages of AI Technology Empowering Clinical Decision Support Systems

2.1. Integration and Feature Extraction of Multi-Source Heterogeneous Data

There are a large number of data types with diverse sources and complex structures in the clinical environment, including electronic medical records (EMR), laboratory test results (LIS), imaging data (PACS), vital sign monitoring information, and drug use records, etc. These data often present characteristics such as the coexistence of structured and unstructured data, inconsistent time distribution, and inconsistent format standards. Based on this, AI technology introduces feature engineering and data fusion algorithms. Through embedded coding, multimodal learning and deep representation models, it achieves unified representation and correlation modeling of multi-source data, effectively breaking down information silos [1].

To support real-time clinical decision-making, the system needs to build a patient-centered data integration platform, which has the ability of dynamic access and synchronous processing of heterogeneous data. Through a unified feature extraction process, the system can automatically generate high-quality input variables, providing reliable support for subsequent reasoning, judgment and recommended output. Figure 1 below shows the core processing flowchart of AI in integrating multi-source heterogeneous clinical data:

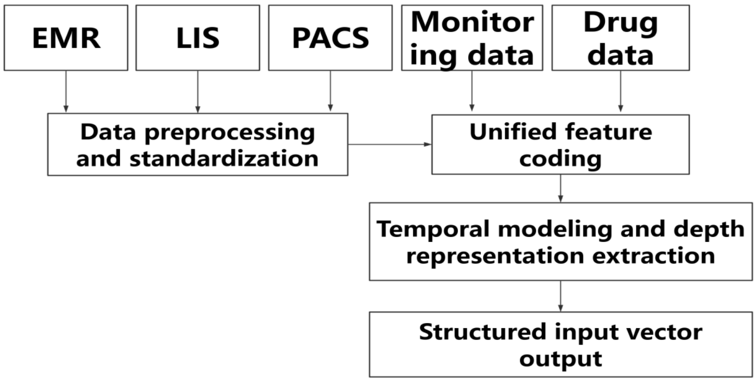

Figure 1. The processing flowchart of AI in the integration of multi-source heterogeneous clinical data.

2.2. High-Precision Prediction and Individualized Decision Recommendation

The key decisions in traditional CDSS are usually based on fixed processes or static databases, which makes it difficult to meet the requirements of the complex and dynamic medical environment. With the introduction of AI technology, especially the predictive models established by deep learning, the system can analyze a large amount of patient data and more accurately identify the risk conditions and disease trends of diseases. By modeling and analyzing complex and extensive data such as the patient's previous medical history, physiological indicators, and medication status, an individual's risk assessment can be modeled, thereby predicting possible development trends and potential risks [2].

In terms of personalized recommendation, the function of AI-CDSS has evolved from a single rule-based suggestion to an intervention strategy that is dynamically adjusted based on the current medical environment. It can generate more targeted examination suggestions, treatment plans and medication strategies based on the classification results of patients' conditions and similar cases, improving accuracy and practicality [3].

2.3. Real-Time Reasoning and Automated Auxiliary Diagnosis and Treatment Efficiency Improvement

Real-time performance is one of the key characteristics that distinguish AI clinical decision support systems from traditional systems. In the actual operating environment, many diagnostic tasks are extremely time-sensitive, such as emergency response in risk prediction, auxiliary decision-making during surgical procedures, and real-time diagnosis in intensive care. Relying on the real-time reasoning ability of AI technology, by using stream data analysis (such as real-time databases and edge computing) and lightweight model deployment architectures (such as TensorRT and ONNX Runtime), the system can complete data processing and inference result output within 1 microsecond, significantly shortening the time of information feedback [4].

Based on the preset digital body model of the patient and key medical information parameters, the system can provide diagnostic suggestions through real-time reasoning. When the diagnostic suggestions are updated due to changes in medical records, the model will immediately regenerate a set of suggestions and actively push corresponding prompts to doctors to ensure that the diagnostic information can be presented in real time on the clinical front line. Figure 2 below shows the workflow of the AI-CDSS system in real-time reasoning and assisted diagnosis and treatment:

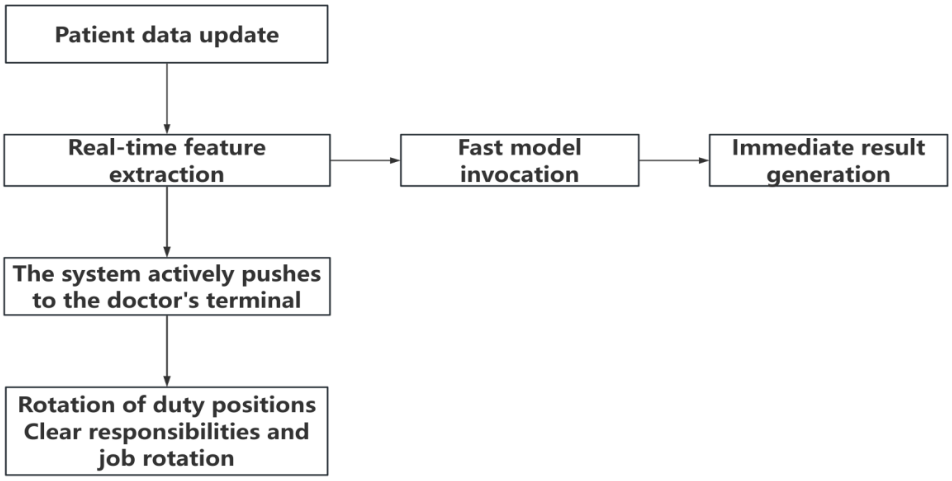

Figure 2. Flowchart of real-time reasoning and assisted diagnosis and treatment of the AI-CDSS system.

3. The Core Construction Mechanism of the AI Real-Time Clinical Decision Support System

3.1. Real-Time Response Mechanism Optimization and System Reasoning Acceleration Technology

In clinical scenarios, due to the delay in the judgment process, it sometimes seriously affects the therapeutic effect and medical safety of patients. To enable the AI-enabled real-time clinical decision support system to respond to requests promptly, it is necessary to comprehensively enhance the optimization capabilities from multiple aspects such as the implementation method of the system, the logical reasoning process of the model, and data exchange [5]. Besides the research on model algorithms, choosing an appropriate reasoning framework, scientifically allocating computing resources, and rationally configuring the system operating environment are also key links to improve the system performance. By reducing the model size, ignoring invalid variables, and applying parallel processing, etc., the generation time of visiting suggestions can be effectively shortened.

To evaluate the practical effectiveness of the above-mentioned several optimization techniques, the time consumption and resource consumption of the conventional model inference acceleration algorithms are compared. The following Table 1 shows the effects of three optimized configurations running on the same medical dataset:

Table 1. Comparative Analysis of Response Efficiency of Different Inference Acceleration Techniques.

Optimize the configuration method	Average response time (ms)	Memory usage (MB)	Reasoning success rate (%)
Deployment of standard deep models	850	720	94.2
Model pruning + ONNX deployment	410	430	93.6
INT8 quantization + TensorRT optimization	230	310	92.8

Judging from the results in Table 1, combining the model compression technology with the hardware acceleration technology can reduce the response time to a quarter or even more of the original, while reducing memory usage by nearly 60%. Although the success rate of reasoning has slightly declined, it is still in acceptable clinical practice.

3.2. Standardization of Clinical Data and Dynamic Quality Control Mechanism

For the clinical AI real-time decision support system based on high-quality data, if the data format is not uniform and the quality does not meet the standards, it may lead to reasoning errors and suggestion deviations. Due to the diverse sources of data, including both structured laboratory test results and partially structured medical records, as well as a large amount of unstructured image and text information, in the process of achieving multi-system data fusion and ensuring the consistency of model input, a standardized data processing procedure must be established. It covers multiple links such as data cleaning, format specification adjustment, and semantic mapping.

The following Table 2 summarizes the changes in key quality indicators in data processing before and after the deployment of AI-CDSS in tertiary hospitals:

Table 2. Comparison of Key Indicators before and after standardization and dynamic quality control.

Indicator type	Pre-optimization level	Optimized level	Improvement range
Data missing rate (%)	6.8	1.9	↓72.1%
Proportion of outliers (%)	3.4	0.7	↓79.4%
Data repetition rate (%)	5.1	0.6	↓88.2%
Unified format coverage	65%	98%	↑50.8%

It can be seen from the results in Table 2 that through the joint deployment of the standardization module and the quality control system, the loss, anomalies and duplications of data have been significantly reduced, and the overall structural consistency has been significantly improved.

3.3. Output Structure and Interaction Design of the Decision Recommendation Module

The decision recommendation module undertakes the key function of connecting the reasoning results of the model with the actual clinical operation. In the limited interface space, the system needs to present diagnosis and treatment suggestions, risk weights, and traceable evidence information simultaneously, striving to strike a balance between information density and readability. This module adopts a hierarchical display strategy: the core conclusion is placed at the visual center of the page, while the evidence chain and similar case index are designed as foldable structures to reduce the visual burden and avoid information redundancy. The interaction design adheres to the "minimum click" principle. The system can automatically record the doctor's acceptance, modification or neglect of the suggested content, and feed these feedbacks back to the model update channel to achieve dynamic fine-tuning and self-optimization of the reasoning mechanism.

To evaluate the performance of different output structures in clinical scenarios, the system conducted comparative tests on three interface prototypes, and the results are shown in Table 3 below.

Table 3. Comparison of Prototypes of Decision Output Interfaces.

Output structure prototype	The average number of operation steps for doctors	Single viewing time (seconds)	Adoption rate (%)
List of basic text prompts	7	68	54.1
Hierarchical visual card	4	41	72.8
Interactive process guidance window	3	35	81.3

The data in Table 3 indicates that in the process of handling emergency or severe cases, the interactive process guidance window is more efficient, especially in emergency situations. Although the acceptance rate of hierarchical visual cards is the lowest, under the conditions of an easy-to-read and concise information architecture and a good reading experience, they are still relatively suitable for the regular outpatient environment.

3.4. Model Interpretability and Construction of Trusted Recommendation Mechanism

The effective deployment of AI models in clinical practice depends not only on the performance of the algorithms but more crucially on whether their output results can be understood and trusted by medical staff. Due to the "black box" effect of deep learning, the results it recommends are difficult to track and verify. This deficiency will affect doctors' willingness to adopt it. Based on this, combined with the perspective of interpretability, starting from the aspect of model structure design, and simultaneously using the attention heat map display, the Shapley value is introduced in the model process to calculate the influence degree of each variable on the final diagnosis, or the weight of key inputs in the decision-making process is displayed through the attention heat map.

In addition to the explanatory function, the trustworthiness of the recommendation system must also meet standards such as clinical knowledge verification, transparency, and consistency. That is, the platform should be able to link the output result of each recommendation to previous clinical guidelines, similar past records, knowledge databases, and other resources to form a chain. To verify the influence of different interpretation mechanisms on doctors' behaviors, a round of multi-center tests was conducted systematically to compare the clinical feedback data of three mainstream interpretable methods. The results are shown in Table 4 below.

Table 4. The influence of different explainable mechanisms on doctors' adoption behavior.

Explainable mechanism type	Recommended adoption rate (%)	Doctor Trust Score (out of 5 points)	Feedback modification rate (%)
Unexplained output (control group)	49.6	2.7	34.5
Visualize the heat map of feature weights	68.3	4.1	21.2
Shapley value + reference case link	79.5	4.6	14.7

The results in Table 4 show that improving interpretability can significantly enhance the trust and adoption of the system by medical staff. Especially when quantitative interpretation (such as Shap values) is combined with practical application, it will more effectively reduce the number of feedback corrections, indicating that doctors have a more thorough understanding of the recommended results and less intervention.

4. Optimization Strategies of AI Real-time Clinical Decision Support System

4.1. Enhance Reasoning Efficiency and System Response Speed

The key to the use of the AI real-time clinical decision-making system lies in whether it can make judgments quickly in the real-time clinical stage to ensure the timeliness and accuracy of clinical diagnosis and treatment decisions. The reasoning time has a direct impact on the output result time. Therefore, the model needs to be improved from three aspects. Regarding the design of the model, Lightweight architectures such as MobileNet or EfficientNet can be utilized during the design process. At the same time, techniques such as pruning and quantization can also be employed to reduce the model volume. Reduce the computational load of the model. At the same time, the selection of the inference framework is also very important. For instance, choosing to utilize excellent inference frameworks such as TensorRT and ONNXRuntime can achieve command-level parallelism, fully utilize memory, and significantly reduce the model's startup and execution time.

At the implementation level, the functions with more applications are migrated to the boundary devices for implementation, reducing data interaction with the remote server and shortening the response time. Using the caching mechanism to store the results of intermediate attributes and prepare them in advance for subsequent operations can also avoid duplicate calculations. System optimization needs to achieve a response level of "seconds", ensuring that it fully meets the requirements of medical practice and integrates AI into the workflow of doctors rather than becoming an "obstacle" in their work. This concept is crucial for enhancing the usability of the system and the acceptance of medical staff.

4.2. Establish a Dynamic Data Management and Model Update Mechanism

Due to the continuous update and complex form of clinical data, this requires the AI decision support system to have dynamic management capabilities and an adaptive update mechanism for the model. The traditional static learning method is difficult to maintain the stability of the prediction effect during long-term operation, especially in the medical environment where the types of diseases change and the examination methods are constantly innovated. It relies more on the learning strategy based on dynamic data to continuously optimize the model. By setting the dynamic threshold of real-time input data, the incremental learning mechanism can be triggered, thereby efficiently integrating the latest data and enabling it to take effect quickly without weakening the performance of the original model, ensuring that the system always maintains a response capability synchronized with the clinical status quo. Performance monitoring indicators, as the basic conditions for update triggers, need to meet the dual constraints of accuracy and time. Common evaluation indicators such as F1 score and AUC value are combined with the rate of change to calculate and form the evaluation function:

$$\Delta P = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100\% \tag{1}$$

Among them, P_t refers to the current model effect score. If ΔP is greater than the given threshold, an update or replacement is required. In addition to data verification measures, the system also adopts a version rollback mechanism and a sandbox testing environment. It first runs offline for testing and then gradually goes online to update the data path in a closed loop. This not only ensures the changes in the model but also guarantees the reproducibility and consistency of diagnosis and treatment recommendations.

4.3. Optimize the Interactive Interface and the Closed Loop of User Behavior Feedback

In the AI real-time clinical decision support system, whether the output results can be accepted and truly be applied in the medical process depends on the friendliness of the system's interaction interface and the closed-loop design of the feedback mechanism. By using block composition, visual marking and suggestive explanations for tracking, it can effectively reduce the cognitive load of doctors and improve the collaborative efficiency of human-computer interaction. The interaction structure should follow the principle of least action, such as "one-step confirmation" or "drag-and-drop correction", etc., to lower the usage threshold.

Taking the quantitative guidance provided by the feedback information to the model as the design goal, the following weight update function can be obtained.

$$w_i^{(t+1)} = w_i^{(t)} + \alpha \cdot fi \tag{2}$$

Among them, $w_i^{(t+1)}$ represents the i -th feature weight in the current model, α represents the learning rate, and fi represents the correction signal of user behavior feedback.

4.4. Strengthen the Security Guarantee and Stable Operation of System Integration

AI real-time clinical decision support systems are often deployed in the complex information ecosystem of hospitals and need to work in coordination with systems such as EMR, HIS, and LIS. This requires that the AI real-time clinical decision support system have extremely high security and stability to ensure the feasibility of the work. From the perspective of security, it is necessary to establish multiple levels of authorization management, access tracking records, and de-identification processing technologies, etc., to protect patient privacy and comply with regulatory requirements, etc. On the premise of ensuring a stable operation process, it should be achieved through methods such as resource allocation technology, fault recovery and parallel control design to avoid the phenomenon of incorrect response under high load conditions. To comprehensively evaluate the operational stability of the integrated system, the following comprehensive scoring model can be adopted:

$$S = \beta_1 \cdot A + \beta_2 \cdot U + \beta_3 \cdot R \quad (3)$$

Among them, s is the system stability score, α is the system average availability, u is the Mean Time between Failures (MTBF), and r is the abnormal recovery speed. $\beta_1, \beta_2, \beta_3$ are all weight coefficients set as empirical values.

5. Conclusion

Empowered by AI technology, real-time clinical decision support systems are gradually reconfiguring traditional medical service processes. By establishing an efficient data integration mechanism, optimizing the reasoning response structure, and improving the closed-loop path of user feedback, the system not only realizes intelligent auxiliary functions, but also significantly enhances the timeliness and reliability of clinical decision-making. Further strengthening the construction of the interpretability, security and operational stability of the model will lay a solid foundation for the wide application of AI-CDSS in actual clinical practice and provide continuous impetus for the high-quality development of intelligent medical care.

References

1. V. Shanmugarajeshwari, and M. Ilayaraja, "IoT-based multiclass decision support system of chronic kidney disease using optimal DNN," *International Journal of Modeling, Simulation, and Scientific Computing*, vol. 15, no. 04, p. 2441002, 2024. doi: 10.1142/s1793962324410022
2. T. Qiao, "Adaptive optimization of thermal energy and information management in intelligent green manufacturing process based on neural network," *Thermal Science and Engineering Progress*, vol. 55, p. 102953, 2024. doi: 10.1016/j.tsep.2024.102953
3. J. Moxam, H. Sanghvi, A. Danesh, S. Graves, S. Gupta, K. V. Chalam, and A. Pandya, "Evaluation of artificial intelligence based Clinical Decision Support System for detecting Nystagmus," *Investigative Ophthalmology & Visual Science*, vol. 65, no. 7, pp. 1149-1149, 2024.
4. A. C. Rathinakumari, B. B. Channabasamma, and G. S. Kumaran, "Physical and mechanical properties of garlic bulbs and cloves (Allium sativum L.) relevant to development of garlic bulb breaker," *) relevant to development of garlic bulb breaker*, 2015.
5. T. Rambach, P. Gleim, S. Mandelartz, C. Heizmann, C. Kunze, and P. Kellmeyer, "Challenges and facilitation approaches for the participatory design of AI-based clinical decision support systems: protocol for a scoping review," *JMIR Research Protocols*, vol. 13, no. 1, p. e58185, 2024. doi: 10.2196/58185