Metodologia para Diagnóstico de Sepse Pediátrica via LLM, RAG e Fine-Tuning sob Escassez de Dados Reais

Adriano Lages dos Santos[1], Isabela Torres[2], Melissa Oliveira[2], Mylena Maria Guedes de Almeida[2] , Lilian Martins Oliveira Diniz[2], Cristiane Dos Santos Dias[2], Zilma Reis[2], Eduardo Araujo de Oliveira[2]

[1] Instituto Federal de Minas Gerais (IFMG) - Belo Horizonte - MG - Brasil
[2] Faculdade de Medicina, Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte - MG - Brasil

{adrianolagesdossantos, isabelatorrese, melissaoliveirar3, mylenamgalmeida, lilianmodiniz, profacristianedias, zilma.medicina, eduolive812}@gmail.com

Resumo. *A sepse pediátrica é uma emergência médica de alta complexidade, exigindo detecção precoce para reduzir a morbimortalidade. Este trabalho propõe uma metodologia para o desenvolvimento de um sistema de suporte à decisão clínica fundamentado em Grandes Modelos de Linguagem (LLM), utilizando Geração Aumentada por Recuperação (RAG) e Fine-tuning. A escassez de dados reais é mitigada pelo framework MedSyn para geração de dados sintéticos. O sistema integra os Critérios Phoenix, assegurando conformidade com consensos internacionais. A arquitetura proposta combina dados estruturados e notas clínicas, provendo recomendações fomentadas por evidências clínicas no domínio da informática em saúde.*

Abstract. *Pediatric sepsis is a highly complex medical emergency, requiring early detection to reduce morbidity and mortality. This work proposes a methodology for developing a clinical decision support system based on Large Language Models (LLM), using Retrieval Augmented Generation (RAG) and fine-tuning. The scarcity of real data is mitigated by the MedSyn framework for generating synthetic data. The system integrates the Phoenix Criteria, ensuring compliance with international consensus. The proposed architecture combines structured data and clinical notes, providing recommendations supported by clinical evidence in the field of health informatics.*

1. Introdução

A sepse pediátrica é definida como uma disfunção orgânica potencialmente fatal, decorrente de uma resposta desregulada do hospedeiro a uma infecção, especificamente quantificada pelo *Phoenix Sepsis Score* superior a dois pontos em crianças com suspeita de infecção [Schlapbach et al. 2024]. A transição dos critérios de Síndrome de Resposta Inflamatória Sistêmica (SIRS) para critérios baseados em disfunção orgânica reflete a necessidade de maior especificidade diagnóstica em ambientes de cuidados críticos [Weiss et al. 2020].

Esta proposta metodológica diferencia-se ao adotar os critérios Phoenix (2024), o mais recente consenso internacional para sepse pediátrica, integrando-os a técnicas de *Retrieval-Augmented Generation* (RAG) [Fan 2024] e fine-tunning [Anisuzzaman 2024, Savage 2025]. A abordagem foca na viabilidade técnica de um sistema capaz de processar simultaneamente sinais vitais estruturados e o contexto clínico contido em notas de prontuário por meio de modelos de linguagem especializados. Este artigo detalha a fundamentação técnica e a arquitetura sistêmica, estabelecendo uma base

metodológica para futuras validações clínicas prospectivas. Outra contribuição do nosso trabalho é a utilização do framework MedSyn [Kumichev 2024] para geração de dados sintéticos para treinamento do LLM e RAG. Isso não impede que durante a duração do projeto dados reais possam ser integrados ao sistema.

A contribuição central deste trabalho reside na proposição de uma arquitetura de suporte à decisão clínica que transcende modelos de triagem univariados, integrando o raciocínio clínico semântico de *Large Language Models* (LLMs) à precisão algorítmica de modelos preditivos tradicionais. Esta abordagem aborda a lacuna de interpretabilidade em sistemas de IA na saúde ao ancorar a geração de linguagem em diretrizes clínicas baseadas em evidências via Geração Aumentada por Recuperação (RAG) [Lewis et al. 2020] [Nagori, 2025]. Este trabalho detalha a organização metodológica do sistema, estruturado da seguinte forma: a Seção 2 apresenta os critérios Phoenix; a Seção 3 descreve a arquitetura proposta para o problema; a Seção 4 detalha a geração de dados sintéticos; a Seção 5 aborda a camada RAG; a Seção 6 especifica o fine-tuning no modelo escolhido; a Seção 7 discute aspectos regulatórios e a Seção 8 conclui o trabalho apresentando as futuras etapas do desenvolvimento do projeto.

2. Critérios Clínicos e Fundamentação

A fundamentação teórica deste framework repousa sobre os critérios do *Phoenix Sepsis Score*, desenvolvidos pelo *Society of Critical Care Medicine's Pediatric Sepsis Definition Task Force*. A validade desses critérios foi estabelecida através de uma análise retrospectiva de mais de 3 milhões de encontros pediátricos, demonstrando uma área sob a curva (AUC) superior aos critérios de SIRS para a predição de mortalidade intra-hospitalar [Schlapbach et al. 2024]. O sistema proposto codifica as quatro dimensões de disfunção (cardiovascular, respiratória, neurológica e de coagulação) como variáveis de estado para o mecanismo de inferência do LLM.

3. Metodologia e Arquitetura do Sistema

O desenvolvimento de um sistema de suporte à decisão clínica de alto impacto requer uma arquitetura modular que sintetize dados estruturados (sinais vitais e exames laboratoriais) e dados não estruturados (notas do Prontuário Eletrônico do Paciente - PEP). Tal integração é essencial para mimetizar o raciocínio holístico do especialista intensivista. Optamos pelo modelo Llama 3.2 14B (14 bilhões de parâmetros) como motor de inferência devido ao seu equilíbrio otimizado entre latência de processamento em ambientes de produção e capacidade de raciocínio clínico especializado. A orquestração de tecnologias é realizada por um backend híbrido em Node.js e Python, garantindo interoperabilidade via FHIR R4 [HL7 2019] e alto desempenho na disponibilidade do modelo.

A Figura 1 detalha a arquitetura proposta para o sistema de diagnóstico de sepse pediátrica. O fluxo operacional do sistema inicia-se com a ingestão multidimensional de dados do paciente, abrangendo tanto dados estruturados (sinais vitais, exames laboratoriais, dados demográficos, comorbidades e medicações) quanto dados não

estruturados provenientes das notas clínicas contidas no Prontuário Eletrônico do Paciente (PEP). Em uma primeira etapa de triagem, os dados estruturados são processados por um Modelo de Machine Learning/Deep Learning (ML/DL) encarregado de calcular uma pontuação de risco contínua. Na etapa de triagem inicial, os dados estruturados são processados pelo classificador *Extreme Gradient Boosting* XGBoost, selecionado devido à sua eficácia comprovada no processamento de dados tabulares médicos e na gestão de valores omissos em prontuários eletrônicos [Chen & Guestrin 2016]. Este modelo atua como um filtro de alta latência, operando em paralelo ao cálculo determinístico do *Phoenix Score* para assegurar redundância diagnóstica.

O sistema opera sob uma lógica de decisão baseada em limiares críticos (θ): caso a pontuação de risco supere o limiar de alta confiança (θ_1), um Alerta de Sepse é gerado imediatamente, fundamentado na análise individual dos critérios do Score de Phoenix (SP). Quando a pontuação situa-se em uma zona de incerteza clínica (θ_2 <risco<θ_1), o caso é automaticamente escalonado para a camada de LLM Suporte para auxílio no diagnóstico diferencial.

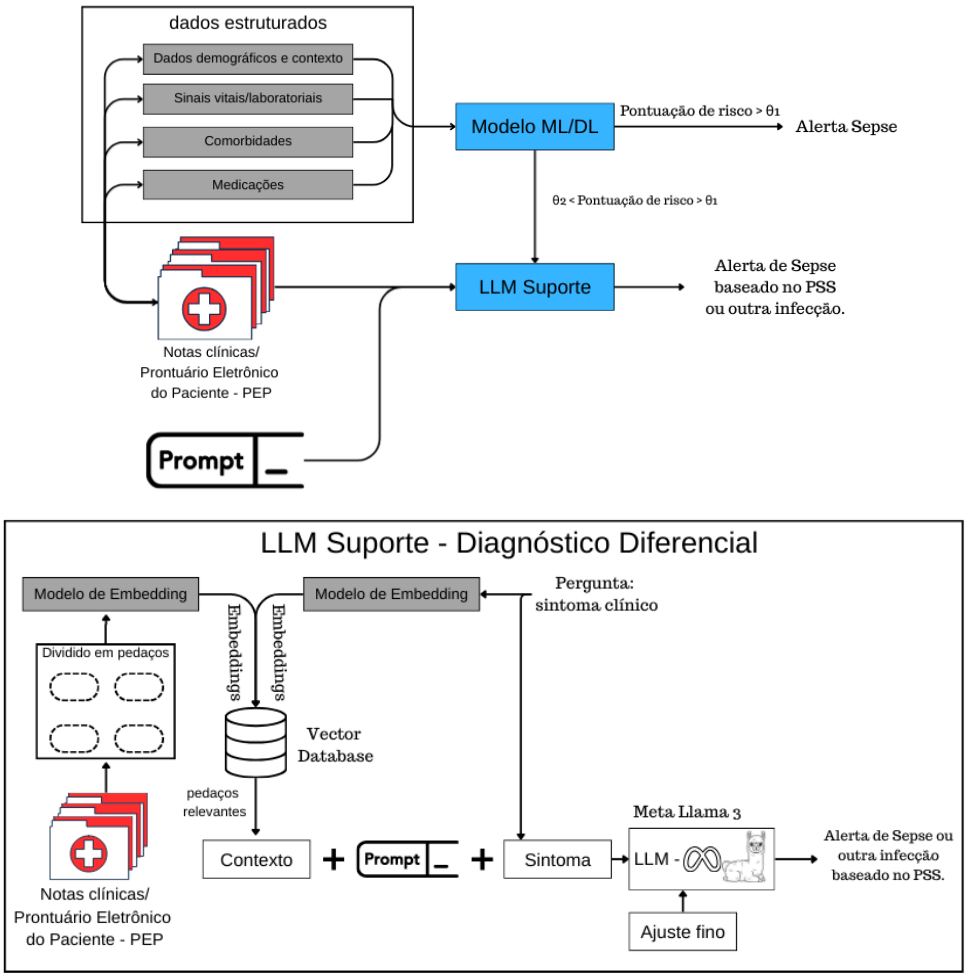

Figura 1. Arquitetura da solução híbrida para suporte à decisão clínica em sepse pediátrica. O diagrama detalha o fluxo de ingestão de dados estruturados e notas clínicas (PEP), a triagem inicial via modelos de aprendizado de máquina com base em limiares de risco (θ) e a camada de suporte

para diagnóstico diferencial utilizando RAG (Retrieval-Augmented Generation) com o modelo Meta Llama 3, fundamentado nos critérios Score de Phoenix.

4. Metodologia de Geração de Dados Sintéticos

A geração de dados sintéticos via MedSyn é estratégica para prover robustez estatística ao modelo, assegurando a privacidade conforme a Lei Geral de Proteção de Dados (LGPD). O volume alvo estabelecido é de 3.000 a 5.000 prontuários sintéticos. Os dados gerados pelo framework se mostraram robustos e muito próximos de dados reais. Para dados reais base para geração vamos utilizar algumas bases de dados reais disponibilizadas gratuitamente PIC (Paediatric Intensive Care Database): Banco de dados de UTI pediátrica, MIMIC-IV (Medical Information Mart for Intensive Care IV): Banco de dados que inclui um subconjunto pediátrico, eICU Collaborative Research Database: Contém mais de 200.000 admissões em UTI de múltiplos centros 2024 Pediatric Sepsis Challenge Dataset [Schlapbach 2024].

A qualidade dos dados sintéticos gerados pelo framework MedSyn é avaliada através da comparação de distribuições marginais e correlações multivariadas entre os dados reais (MIMIC-IV, eICU) e as amostras simuladas. Utiliza-se a Distância de Wasserstein para quantificar a divergência entre as distribuições, assegurando que o modelo preserve as propriedades estatísticas essenciais para o treinamento de algoritmos de detecção de sepse sem comprometer a privacidade diferencial [Kumichev et al. 2024].

4.1. Estratégia de Validação dos Dados

A fidelidade dos dados gerados é mensurada por métricas estatísticas, incluindo o teste de Kolmogorov-Smirnov ($p > 0.05$) e a correlação de Pearson ($r > 0.85$). Clinicamente, os dados devem apresentar 100% de conformidade com os critérios Phoenix para casos graves. A distribuição etária é controlada para refletir a realidade clínica: 30% neonatal, 40% infantil (lactentes) e 30% para crianças acima de um ano. A validação clínica seguirá um protocolo Delphi modificado com 4 pediatras, sendo dois intensivistas. O objetivo é atingir sensibilidade > 90% e especificidade > 85%, conforme métricas de utilidade clínica e calibração de Brier Score.

5. Pipeline RAG e vetorização dos dados

A base de conhecimento clínico integra ontologias como SNOMED CT e LOINC, além das diretrizes da Surviving Sepsis Campaign. Isso é importante porque toda linguagem de resposta do sistema deve aderir estritamente a linguagem médica, evitando assim qualquer sobrecarga ao profissional de saúde para interpretar termos que não sejam referentes a área de saúde. O processo de vetorização utiliza o modelo ClinicalBERT [Huang 2019] para gerar embeddings especializados para a área da medicina. Os documentos são fragmentados em blocos de 512 tokens (sobreposição de 50 tokens), permitindo que o recuperador identifique contextos específicos de disfunção orgânica neonatal para compor o prompt enviado ao modelo de linguagem**.**

6. Fine-Tunning via QLoRA

O modelo Llama 3.2-14B-Instruct é submetido a um ajuste fino utilizando a técnica Quantized Low-Rank Adaptation (QLoRA). Esta escolha técnica permite o treinamento em hardware com restrição de memória (16 até 24GB VRAM) através de quantização de 4 bits, preservando o conhecimento prévio do modelo base. O ajuste foca na adaptação à terminologia médica em língua portuguesa e à lógica de pontuação do Score de Phoenix, elevando a utilidade clínica do sistema. A parametrização técnica define um Rank de 16 e Alpha de 32, direcionando os adaptadores para os módulos q_proj, k_proj, v_proj e o_proj. O treinamento utiliza uma sequência máxima de 2048 tokens, processando instruções que correlacionam o contexto recuperado pelo RAG (diretrizes Phoenix de 2024) aos dados reais e sintéticos do paciente.

7. Discussão Metodológica e Aspectos Regulatórios

Embora a validação em ambiente clínico prospectivo esteja prevista para etapas futuras, realizaram-se experimentos preliminares de 'provas de conceito' utilizando o subconjunto de dados sintéticos validados. Estes testes iniciais indicam a estabilidade da arquitetura de duplo limiar, embora a eficácia diagnóstica definitiva dependa da integração de fluxos de dados reais em tempo real [Ali et al. 2023]. Sistemas de suporte diagnóstico são classificados como Software as a Medical Device (SaMD). No cenário brasileiro, a RDC Nº 657/2022, ANVISA (2022) que enquadra tais tecnologias na Classe de Risco II (risco moderado). Esta classificação exige que o ciclo de vida do software seja documentado com rigor, abrangendo desde a proveniência dos dados sintéticos até a validação de performance por subgrupos etários.

A explicabilidade (Explainable AI - XAI) [Ali 2023] é tratada como requisito central para sistemas de saúde. O uso da arquitetura RAG auxilia na transparência ao permitir que o sistema aponte as evidências clínicas e diretrizes específicas que fundamentaram a recomendação. Adicionalmente, a segurança é reforçada pelo cumprimento da LGPD, utilizando privacidade diferencial (ϵ<1.0) para assegurar que os dados gerados não permitam a reidentificação de pacientes em ataques de inferência ao LLM ou qualquer tipo de ataque malicioso a base de dados do sistema.

8. Conclusão e próximas etapas do projeto

O sucesso da implementação será monitorado através de métricas de desempenho técnico (Precisão, Sensibilidade e F1-Score) e métricas de impacto clínico (tempo para administração de antibióticos e tempo de permanência em UTI). A etapa atual compreende o ajuste fino do modelo Llama 3.2 em língua portuguesa, seguida pela fase de integração via API REST em um ambiente de *sandbox* hospitalar. O acesso aos dados reais será mediado por protocolos de anonimização estritos, em conformidade com o Comitê de Ética em Pesquisa (CEP) da instituição colaboradora.

Agradecimentos

Este estudo foi financiado com recursos do Centro de Inovação e Inteligência Artificial para Saúde (CI-IA Saúde), em parte com recursos da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) Processo nº 2020/09866-4, da Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) Processo nº PPE-00030-21, UNIMED Belo Horizonte e Instituto Federal de Minas Gerais.

References

Ali, S., Abuhmed, et al. (2023). "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence". Information Fusion, 99(101805), 101805. sciencedirect. https://doi.org/10.1016/j.inffus.2023.101805

Anisuzzaman, D. M et al. (2024). "Fine-Tuning LLMs for Specialized Use Cases". Mayo Clinic Proceedings: Digital Health, 3(1). https://doi.org/10.1016/j.mcpdig.2024.11.005

ANVISA (2022) "Resolução da Diretoria Colegiada - RDC nº 657, de 24 de março de 2022", Diário Oficial da União.

Baracat, E. C. E. (2018). Validação Dos Sistemas De Triagem Em Emergência Pediátrica. Revista Paulista De Pediatria, 36(4), 386–387. https://doi.org/10.1590/1984-0462/;2018;36;4;00018

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

Bender, D., & Sartipi, K. (2013). HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems.

Fan, W. et al. (2024). "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models". In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)

Health Level Seven International. (2019). Fast Healthcare Interoperability Resources (FHIR v4.0.1) [Webpage]. https://hl7.org/fhir/R4/index.html

Kumichev, G. et al. (2024) "MedSyn: LLM-based Synthetic Medical Text Generation Framework", In Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024.

Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems.

Huang, K., Altosaar, J., & Ranganath, R. (2019). "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission". ArXiv, abs/1904.05342.

Nagori, A., Gautam, A., Wiens, M. O., Nguyen, V., Mugisha, N. K., Kabakyenga, J., Kissoon, N., Ansermino, J. M., & Kamaleswaran, R. (2025). "Contextual Phenotyping of Pediatric Sepsis Cohort Using Large Language Models". AMIA. Annual Symposium proceedings. AMIA Symposium, 2024,

Savage, T. et al. (2025) "Fine-Tuning Methods for Large Language Models in Clinical Medicine", Journal of Medical Internet Research, Vol. 27, e76048.

Schlapbach, L. J. et al. (2024) "International Consensus Criteria for Pediatric Sepsis and Septic Shock: The Phoenix Sepsis Score", JAMA, Vol. 331, No. 8.

Weiss, S. L., et al. (2020). Surviving Sepsis Campaign International Guidelines for the Management of Septic Shock and Sepsis-Associated Organ Dysfunction in Children. Pediatric Critical Care Medicine.