Agente de inteligência artificial para verificar a validade de sinais auto-coletados por pacientes de telemedicina

Francisco Paulo Maraschin[1], **Luiz Eduardo S. Spalding**[2], **Marcelo Trindade Rebonatto**[1]

[1]Programa de Pós-graduação em Computação Aplicada - Universidade de Passo Fundo
BR 285 Km 292,7 — Campus I, Bairro São José – Passo Fundo - RS, 99052-900

[2]Elomed Indústria e Comércio de Equipamentos Eletrônicos Ltda
Av. Sete de Setembro, 99, sala 4, sobreloja, Centro – Passo Fundo - RS, 99010-120

66128@upf.br, spading@elomed.com.br, rebonatto@upf.br

Abstract. *This paper presents a proposal for the development of a solution that utilizes generative artificial intelligence to evaluate the self-collection of physiological data performed by the patient itself, identifying its validity and aiding the patient to perform quality acquisitions.*

Resumo. *Este artigo apresenta uma proposta para o desenvolvimento de uma solução que utilize inteligência artificial generativa para avaliar a coleta de dados fisiológicos realizados pelo próprio paciente, identificando sua validade e auxiliando o paciente a realizar aquisições com qualidade.*

1. Introdução

A telemedicina permite que pacientes localizados em áreas remotas conectem-se aos grandes centros de medicina. Essa modalidade contribui para um melhor controle clínico dos pacientes e da redução da taxa de faltas em consultas [SHAO et al. 2025, ANAWADE et al. 2024, BARBOSA et al. 2021].

Por meio de equipamentos portáteis, os pacientes podem coletar dados fisiológicos e compartilhá-los com médicos que se encontram em localidades distantes, eliminando a necessidade de deslocamento de um paciente até uma unidade de saúde [SIAM et al. 2023, VEGESNA et al. 2017].

Entretanto, a coleta de dados, quando realizada por pessoas não treinadas, pode gerar dados pouco confiáveis. Isso acontece pois o paciente não possui o mesmo treinamento que o especialista possui e não toma o devido cuidado ao coletar esses dados pela falta de conhecimento de suas implicações [SHIN et al. 2023].

A literatura médica dispõe de regras para a avaliação das medidas indicando quando elas devem ser desconsideradas. Essas regras encontram-se na forma textual, sendo que é possível processar esses textos e extrair essas informações por meio do uso de processamento de linguagem natural. Modelos de linguagem podem ser utilizados para fazer esse processamento, em especial, os modelos de linguagem de grande escala (*Large Language Models* - LLM) [HE et al. 2024, CHEN et al. 2023, DELGADO-CHAVES et al. 2025].

O objetivo deste trabalho é desenvolver um agente de inteligência artificial para auxiliar o usuário a realizar automedições de dados clínicos com uso de equipamentos

médicos portáteis, buscando a coleta de dados válidos em telessaúde. Esse agente será embarcado no equipamento médico, comunicando-se diretamente com o equipamento.

2. Revisão de literatura

Os modelos de linguagem de grande escala são modelos generativos que possuem bilhões de parâmetros e são treinados com bases massivas [YILDIRIM and CHENAGHLU 2024, JURAFSKY and MARTIN 2025]. Na área da saúde, os LLMs mostram um potencial em várias aplicações, como a educação médica, através da geração de diagnósticos diferenciais e respondendo a questões em estilo de teste [Lucas et al. 2024].

Um *prompt* é a entrada que é fornecida ao LLM da qual se espera a obtenção de uma resposta. Falhas de comunicação em interações humanas são comuns - acontecem a toda hora. Esse tipo de problema é inerente da comunicação do ser humano e quando *prompts* são aplicados a um LLM, eles também podem ocorrer. Algumas técnicas podem fazer com que os *prompts* sejam mais assertivos, como *zero-shot*, *few-shot* e *chain-of-thoughts* [BERRYMAN and ZIEGLER 2025].

Em um LLM, o raciocínio não ocorre da mesma forma que nos seres humanos. A resposta produzida é um produto direto da *prompt* fornecido. O LLM simplesmente transforma a entrada de acordo com as probabilidades estatísticas adquiridas em seu treinamento. Produzir uma conversação nas interações com o LLM melhoram a precisão de sua saída. Esse tipo de interação pode ser obtido com maior eficácia por meio do uso de programas construídos para realizar esse processo: são os Agentes de Inteligência Artificial (IA) [BERRYMAN and ZIEGLER 2025].

Agentes de IA são sistemas de software que fazem o uso de LLMs para realizar tarefas de forma (semi)autônoma. Eles são empregados para realizar a tomada de decisões complexas, execução de tarefas autônomas e de tarefas adaptativas [MAZUMDER 2025].

Um agente de IA é composto por três partes fundamentais: Percepção, raciocínio e ação. O uso integrado destes componentes permite ao agente interagir com o ambiente e adaptar-se às tarefas [MAZUMDER 2025].

3. Metodologia

A Figura 1 contém o diagrama de caso de uso para o agente proposto. Nela é possível observar os principais componentes da solução e como será sua interação com o usuário.

Conforme a Figura 1, o usuário realiza alguma medição disponível no equipamento medidor. Assim que a medição é concluída, o agente invoca o modelo de IA generativa, enviando um *prompt* com os dados coletados e com o conjunto de restrições a serem atendidas para a validação dos dados coletados. A IA deverá analisar os dados coletados e emitir um veredito.

No *prompt* inicial é descrita a tarefa que o modelo de IA deverá executar e quais as ferramentas estarão a sua disposição para poder auxiliá-lo neste processo. Essas ferramentas poderão ser acionadas pelo modelo de IA durante seu processo de raciocínio, caso seja necessário complementá-lo.

O agente conterá algumas possibilidades de ferramentas que executam funções específicas que não fazem parte das capacidades de raciocínio do modelo de IA generativa consultado ou possuem uma qualidade de resultado superior.

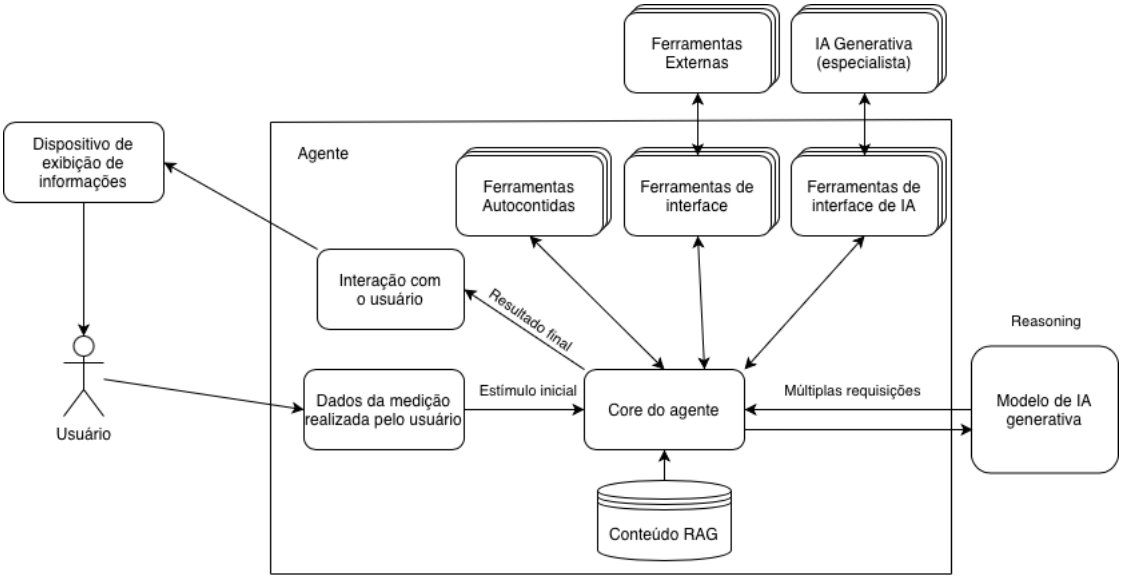

Figura 1. Caso de uso de construção do agente.

O módulo de interação com o usuário permite ao agente enviar ao usuário, por meio das interfaces disponíveis (*display*, áudio, etc.), informações sobre o resultado final, que pode ser uma mensagem de sucesso ou uma mensagem com instruções para o usuário corrigir os problemas ocorridos na medição e repetir o processo.

O agente será executado em um ambiente computacional que simulará as características de um equipamento medidor, a ser desenvolvido pela empresa parceira do projeto, para a devida coleta de dados. Esse equipamento medidor possuirá um hardware com capacidade de realizar as medições de ECG, pressão arterial, saturação de oxigênio, temperatura corporal, fluxo respiratório e concentração de oxigênio, além da capacidade de executar o agente em um hardware embarcado. Para os protótipos, uma simulação da aquisição desses dados será utilizada.

A validação será realizada com sinais gerados por simuladores, entre eles o ProSim 4 [1] da Fluke Biomedical e o IP300-6 [2] da Elomed. Além disso, durante o desenvolvimento serão utilizados sinais sintéticos simulando situações de validação e invalidação.

4. Implementação

O agente inicial foi construído com o uso do framework LangChain (https://www.langchain.com) por ser o mais difundido em conjunto com um modelo de IA generativa comercial hospedada na nuvem. Logo, o agente necessitará de acesso à internet para sua operação. Serão aproveitadas as interfaces do LangChain para criar um padrão de implementação para essas ferramentas [PATTHI 2025, HUANG 2025].

O agente final será construído com o objetivo de ser executado em um Sistema-em-Módulo (*System-on-Module* - SoM). Esse ambiente simplifica a integração do agente no ambiente de operação do equipamento final, facilitando a aquisição dos sinais eletrônicos das medições realizadas pelo paciente. O sistema operacional utilizado será

[1] https://www.flukebiomedical.com/sites/default/files/resources/Prosim4_POR_A_W.PDF
[2] https://www.elomed.com.br/produtos/calibrador-de-pressao-ip300-6/

o Linux. Para a execução no SoM, será construída uma distribuição específica do Linux com o uso do projeto Yocto (http://www.yoctoproject.org), que é uma ferramenta que permite compilar os pacotes que compõem o sistema Linux de forma automatizada.

O agente encontra-se em fase de desenvolvimento, com resultados iniciais promissores. O LangChain trata da maior parte dos detalhes de construção do agente. Muitas IAs comerciais permitem uso da API para comunicação apenas em modalidades pagas, entre elas, OpenAI GPT (https://openai.com/), Anthropic Claude (https://claude.com) e Deepseek (https://www.deepseek.com/en/). Inicialmente, o agente está sendo testado com o Google Gemini (https://gemini.google.com/app) que permite o uso da API de forma gratuita com limitações na quantidade de dados trafegados. Entretanto, pretende-se testar o agente utilizando 3 modelos de cada uma das IAs comerciais citadas: um modelo menor (GPT-5.4-nano, gemini-2.5-flash-lite, deepseek-v4-flash e claude-haiku-4.5), o modelo padrão (GPT-5.5, gemini-2.5-flash e claude-sonnet-4.6) e um modelo avançado (GPT-5.5-pro, gemini-2.5-pro, deepseek-v4-pro e claude-opus-4.7).

Os testes realizados foram feitos com dados de medições de pressão arterial. São disponibilizados para a IA os dados brutos da medição, formados por 12.000 amostras da pressão do manguito e 12.000 amostras da pressão oscilométrica, adquiridas a cada 5 ms, e o resultado do equipamento medidor para que a IA analise e emita seu veredito. Observou-se a necessidade de padronizar a nomenclatura utilizada para referir-se aos sinais no *prompt*, informar a base de tempo da aquisição, bem como um escalonamento dos valores para uma unidade conhecida, para permitir que a IA consiga interpretar corretamente as grandezas, descartando artefatos do sinal irrelevantes.

O envio do conjunto de dados da pressão arterial com as mensagens de *prompt* e a resposta devolvida pela IA em todas as interações do agente consome cerca de 200.000 *tokens*, conforme medição realizada pela própria API da IA. Observa-se que essa quantidade é bastante significativa, podendo gerar um custo considerável para a operação. Formas de reduzir esse consumo são estudadas, como compactar as informações no *prompt* e trabalhar com o uso de ferramentas que possam analisar os dados brutos para a IA.

5. Conclusão

Informações técnicas sobre coleta de dados clínicos encontram-se na literatura na forma textual, em linguagem natural. O processamento dessas informações e a sua utilização pode ser realizado por programas de processamento de linguagem natural para extrair essas regras e utilizá-las para analisar dados obtidos de medições fisiológicas.

O uso de um agente de IA, um programa construído para utilizar a capacidade de raciocínio da IA generativa de forma eficiente, possibilita a extração de respostas da IA, além de permitir a combinação com ferramentas que ampliam a sua capacidade de atuação.

A implementação deste trabalho atualmente encontra-se no estágio inicial, onde a programação do agente de teste foi realizada e foi feita a análise do comportamento deste agente. Os próximos passos incluem o aperfeiçoamento da programação do agente, principalmente no ajuste dos *prompts* enviados à IA, além do teste de desempenho com modelos pagos. Outra atividade considerada é o uso de ferramentas para tentar eliminar a necessidade do envio de todos os dados brutos à IA.

Agradecimentos

Este trabalho conta com recursos do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), através de bolsa de mestrado do programa MAI/DAI, edital nº 05/2025/ PROACAD MAI/DAI - CNPQ/UPF, projeto "Análise de dados de parâmetros cardiorrespiratórios para uso como apoio à telessaúde".

Este trabalho é integrante do projeto de pesquisa CIARS (60116.671.27164.23122021), aprovado no Edital FAPERGS 06/2021 Programa de Redes Inovadores e Tecnologias Estratégicas do Rio Grande do Sul - RITEs-RS.

Referências

ANAWADE, P. A., SHARMA, D., and GAHANE, S. (2024). A comprehensive review on exploring the impact of telemedicine on healthcare accessibility. *Cureus*, 16(3). https://doi.org/10.7759/cureus.55996.

BARBOSA, W., ZHOU, K., WADDELL, E., MYERS, T., and DORSEY, E. R. (2021). Improving access to care: telemedicine across medical domains. *Annual Review of Public Health*, 42(1):463–481. http://doi.org/10.1146/annurev-publhealth-090519-093711.

BERRYMAN, J. and ZIEGLER, A. (2025). *Prompt engineering for LLMs: The art and science of building large language model-based applications*. O'Really Media Inc., Sebastopol, CA, USA.

CHEN, Q., DU, J., HU, Y., KELOTH, V., PENG, X., RAJA, K., ZHANG, R., LU, Z., and XU, H. (2023). Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1). https://doi.org/10.1038/s41467-025-56989-2.

DELGADO-CHAVES, F., JENNINGS, M. J., ATALAIA, A., WOLFF, J., HORVÁTH, R., MAMDOUH, Z. M., BAUMBACH, J., and BAUMBACH, L. (2025). Transforming literature screening: the emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences of the United States of America*, 122(2). https://doi.org/10.1073/pnas.2411962122.

HE, Y., TANG, B., and WANG, X. (2024). Generative models for automatic medical decision rule extraction from text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 7034–7048, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.399.

HUANG, K. (2025). *Agentic AI: theories and practices*. Progress in IS. Springer Nature Switzerland, Cham, Suiça. https://doi.org/10.1007/978-3-031-90026-6.

JURAFSKY, D. and MARTIN, J. H. (2025). *Speech and language processing: an introduction to natural language processing, Computational Linguistics, and Speech Recognition with Language Models*. 3 edition. Disponível em https://web.stanford.edu/~jurafsky/slp3/old_jan25/ed3book_Jan25.pdf. Acesso em 11 set. 2025.

Lucas, H. C., Upperman, J. S., and Robinson, J. R. (2024). A systematic review of large language models and their implications in medical education. *Medical education*, 58(11):1276–1285.

MAZUMDER, E. R. (2025). *AI agent for all: a creative guide to understanding and building AI agents*. via tolino media. Disponível em https://books.google.com.br/books?id=9bFlEQAAQBAJ. Acesso em 10 set. 2025.

PATTHI, T. (2025). *Agentic AI:a practical guide to build agent-based AI systems that think and act: build intelligent agents for real-world AI automation, Machine Learning, and Autonomous Systems with Practical Code and Tools*. Disponível em https://books.google.com.br/books?id=11xpEQAAQBAJ. Acesso em 13 set. 2025.

SHAO, C. C., KATTA, M. H., SMITH, B. P., JONES, B. A., GLEASON, L. T., ABBAS, A., WADHWANI, N., WALLACE, E. L., MUGAVERO, M. J., and CHU, D. I. (2025). Reducing no-show visits and disparities in access: the impact of telemedicine. *Journal of Telemedicine and Telecare*, 31(7):1041–1049. https://doi.org/10.1177/1357633X241241357.

SHIN, D. A., KIM, J., CHOI, S.-W., and LEE, J. C. (2023). DNN based reliability evaluation for telemedicine data. *Biomedical Engineering Letters*, 13(1):11–19. https://doi.org/10.1007/s13534-022-00248-6.

SIAM, A. I., EL-AFFENDI, M. A., ELAZM, A. A., EL-BANBY, G. M., EL-BAHNASAWY, N. A., EL-SAMIE, F. E. A., and EL-LATIF, A. A. A. (2023). Portable and real-time iot-based healthcare monitoring system for daily medical applications. *IEEE Transactions on Computational Social Systems*, 10(4):1629–1641. https://doi.org/10.1109/TCSS.2022.3207562.

VEGESNA, A., TRAN, M., ANGELACCIO, M., and ARCONA, S. (2017). Remote patient monitoring via non-invasive digital technologies: a systematic review. *Telemedicine and e-Health*, 23(1):3–17. https://doi.org/10.1089/tmj.2016.0051.

YILDIRIM, S. and CHENAGHLU, M. (2024). *Mastering transformers: the journey from BERT to large language models and stable diffusion*. Packt Publishing Ltd., Birmingham, Reino Unido.