v7.n12.2026

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

TRIAGEM POR VOZ COM MODELO DE LINGUAGEM E ÁRVORE DE DECISÃO PARA CLASSIFICAÇÃO DE RISCO: INTEGRAÇÃO COM MONITOR DE PRESSÃO ARTERIAL E RELÓGIO INTELIGENTE

VOICE TRIAGE WITH A LANGUAGE MODEL AND A DECISION TREE FOR RISK CLASSIFICATION: INTEGRATION WITH A BLOOD PRESSURE MONITOR AND A SMARTWATCH

TRIAJE POR VOZ CON MODELO DE LENGUAJE Y ÁRBOL DE DECISIÓN PARA LA CLASIFICACIÓN DE RIESGO: INTEGRACIÓN CON TENSIÓMETRO Y RELOJ INTELIGENTE

Giovana Nascimento Ambroselli[1], Vinicius Marçal[2], Gabriel Camargo Felix[3], Juan David Fontecha, Hannah Daga Theodoro

e71210065
https://doi.org/10.47820/recima21.v7i12.10065

PUBLICADO: 12/2026

RESUMO

Em serviços de urgência, a classificação de risco por cores define quem é atendido primeiro, mas depende de um profissional disponível para cada triagem. Este trabalho descreve um protótipo de triagem por voz, concebido como pré-triagem que apoia, e não substitui, o enfermeiro, e que devolve uma das cinco cores de um protocolo de classificação de risco derivado do Sistema Manchester. Um modelo de linguagem de grande escala (LLM) conversa com o usuário e converte o relato em chaves de um catálogo fechado de 102 itens; a cor, porém, é decidida por uma árvore de decisão com regras fixas, que combina os sintomas com sinais vitais medidos por um monitor de pressão arterial e por um relógio inteligente, que mede saturação de oxigênio e temperatura da pele, e com o perfil do paciente. Em 142 casos fictícios, a árvore reproduziu a cor esperada em todos. Um técnico de enfermagem, avaliador único, classificou 101 desses casos e escolheu a mesma cor que a árvore em 98 (97,0%; kappa de 0,96); nos três restantes, a árvore deu uma cor mais grave, sem subtriagem. A extração de sintomas foi testada com 118 relatos escritos, com 76,6% de execuções exatas e cor mais leve que a esperada em 8,8%, o que indica risco de subtriagem; em 10 medições, a pressão coincidiu com o visor do monitor e a SpO2 ficou a até 1 ponto de um oxímetro, mas a temperatura da pele ficou abaixo da referência e não pode ser usada como temperatura clínica sem correção. Separar a conversa da decisão mostrou-se viável e torna cada classificação rastreável até a regra que a gerou, mas a validação clínica ainda é preliminar, por se basear em casos fictícios e em um avaliador único.

PALAVRAS-CHAVE: Triagem. Classificação de risco. Modelo de linguagem. Protocolo de Manchester. Bluetooth Low Energy.

ABSTRACT

In emergency services, color-coded risk classification defines who is seen first, but it requires a professional to be available for each triage. This paper describes a voice triage prototype, conceived as a pre-triage that supports, and does not replace, the nurse, and that returns one of the five colors of a risk classification protocol derived from the Manchester Triage System. A large

[1] Graduanda em Engenharia da Computação na Universidade São Judas Tadeu (USJT), São Paulo, SP, Brasil.
[2] Graduando em Engenharia da Computação na Universidade São Judas Tadeu (USJT), São Paulo, SP, Brasil.Tecnólogo em Automação Industrial.
[3] Graduando em Engenharia da Computação na Universidade São Judas Tadeu (USJT), São Paulo, SP, Brasil.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

language model (LLM) talks to the user and converts the report into keys of a closed catalog of 102 items; the color, however, is decided by a decision tree with fixed rules, which combines the symptoms with vital signs measured by a blood pressure monitor and a smartwatch, which measures oxygen saturation and skin temperature, and with the patient profile. In 142 fictitious cases, the tree reproduced the expected color in all of them. A nursing technician, acting as the only evaluator, classified 101 of these cases and chose the same color as the tree in 98 (97.0%; kappa of 0.96); in the remaining three, the tree gave a more severe color, with no undertriage. Symptom extraction was tested on 118 written reports, with 76.6% exact runs and a milder color than expected in 8.8%, which indicates a risk of undertriage; in 10 measurements, blood pressure matched the monitor display and SpO2 was within 1 point of an oximeter, but skin temperature was below the reference and cannot be used as clinical temperature without correction. Separating the conversation from the decision proved feasible and makes each classification traceable to the rule that generated it, but clinical validation is still preliminary, since it relies on fictitious cases and a single evaluator.

KEYWORDS: *Triage. Risk classification. Language model. Manchester Triage System. Bluetooth Low Energy.*

RESUMEN

En los servicios de urgencias, la clasificación de riesgo por colores define quién es atendido primero, pero depende de un profesional disponible para cada triaje. Este trabajo describe un prototipo de triaje por voz, concebido como pretriaje que apoya, y no sustituye, al enfermero, y que devuelve uno de los cinco colores de un protocolo de clasificación de riesgo derivado del Sistema Manchester. Un modelo de lenguaje de gran escala (LLM) conversa con el usuario y convierte el relato en claves de un catálogo cerrado de 102 elementos; el color, sin embargo, lo decide un árbol de decisión con reglas fijas, que combina los síntomas con signos vitales medidos por un monitor de presión arterial y por un reloj inteligente, que mide la saturación de oxígeno y la temperatura cutánea, y con el perfil del paciente. En 142 casos ficticios, el árbol reprodujo el color esperado en todos. Un técnico de enfermería, único evaluador, clasificó 101 de esos casos y eligió el mismo color que el árbol en 98 (97,0%; kappa de 0,96); en los tres restantes, el árbol dio un color más grave, sin subtriaje. La extracción de síntomas se probó con 118 relatos escritos, con 76,6% de ejecuciones exactas y color más leve que el esperado en 8,8%, lo que indica riesgo de subtriaje; en 10 mediciones, la presión coincidió con la pantalla del tensiómetro y la SpO2 quedó a 1 punto de un oxímetro, pero la temperatura de la piel quedó por debajo de la referencia y no puede usarse como temperatura clínica sin corrección. Separar la conversación de la decisión resultó viable y hace que cada clasificación sea rastreable hasta la regla que la generó, pero la validación clínica aún es preliminar, porque se basa en casos ficticios y en un único evaluador.

PALABRAS CLAVE: *Triaje. Clasificación de riesgo. Modelo de lenguaje. Protocolo de Manchester. Bluetooth Low Energy.*

1. INTRODUÇÃO

Em um pronto-socorro cheio, atender por ordem de chegada pode deixar uma pessoa com dor no peito atrás de dez com dor de garganta. Os protocolos de classificação de risco existem para evitar isso: um profissional treinado avalia cada paciente na chegada e define, por cor, quanto tempo ele pode esperar com segurança (BRASIL, 2009). O Sistema Manchester, um dos

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

mais usados, é aplicado em serviços de urgência de vários países, inclusive no Brasil (MACKWAY-JONES; MARSDEN; WINDLE, 2014; SOUZA *et al.*, 2011).

O gargalo está na própria triagem, que consome tempo de enfermagem e se acumula nos horários de pico, quando o serviço está superlotado (BITTENCOURT; HORTALE, 2009). Automatizar parte dela, sobretudo a coleta de sinais vitais e a primeira conversa com o paciente, é uma ideia atraente. Modelos de linguagem de grande escala (LLMs) tornam a conversa possível, já que entendem relatos em linguagem comum e respondem por voz. Entregar a eles a decisão de gravidade, no entanto, traz um problema: o mesmo relato pode gerar respostas diferentes, e não há como apontar com precisão por que uma cor foi escolhida. A avaliação feita por pessoas também varia entre avaliadores (SOUZA *et al.*, 2011; PARENTI *et al.*, 2014), o que reforça a vantagem de regras explícitas e rastreáveis. Estudos recentes já testaram LLMs justamente na decisão de gravidade: um deles relatou desempenho comparável ao de um médico revisor ao comparar pares de pacientes (WILLIAMS *et al.*, 2024), e outro, concordância elevada com um padrão de referência em um pronto-socorro (PASLI *et al.*, 2024). A proposta deste trabalho é diferente, porque a LLM não decide, e esse cenário justifica investigar formas de automatizar parte da triagem sem perder a rastreabilidade da decisão.

O problema que orienta o trabalho é como usar um modelo de linguagem na conversa de triagem sem entregar a ele a decisão de gravidade, e como combinar a fala do paciente com sinais vitais medidos por sensores para sugerir uma cor de forma explicável.

Este trabalho propõe uma divisão de papéis. A LLM conversa, faz perguntas e traduz o que ouviu para um conjunto fechado de sintomas. A cor é definida por uma árvore de decisão com regras fixas, que qualquer pessoa pode ler, testar e questionar, e que também recebe sinais vitais medidos por sensores. O sistema é um protótipo acadêmico que segue um protocolo municipal derivado do Manchester, e não uma implementação oficial dele. Seu resultado é uma sugestão de prioridade a ser confirmada pelo enfermeiro, e não uma classificação de risco oficial (Seção 4.9).

O objetivo geral é descrever e verificar um protótipo de triagem por voz, concebido como pré-triagem, em que a LLM conversa e extrai sintomas e uma árvore de decisão com regras fixas define a cor.

Os objetivos específicos são: (a) construir uma árvore de decisão com catálogo de 102 chaves (94 sintomas, 7 fatores de risco e a duração da queixa) e 10 regras de combinação, baseada no protocolo de classificação de risco de Aparecida de Goiânia (APARECIDA DE GOIÂNIA, 2025), que registre o motivo de cada classificação; (b) restringir a LLM à extração de sintomas e à explicação do resultado; (c) integrar um monitor de pressão Omron HEM-7156T,

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

lido por Bluetooth Low Energy, e um Galaxy Watch 6, que mede SpO2 e temperatura da pele; e (d) fazer uma verificação inicial, com 142 casos fictícios para a árvore, a avaliação das cores de 101 desses casos por um técnico de enfermagem, 118 relatos para a extração de sintomas e uma reunião de avaliação com uma profissional da área da saúde.

2. REFERENCIAL TEÓRICO

2.1. Sistema Manchester de Classificação de Risco

O Sistema Manchester surgiu na década de 1990, no Reino Unido, e não busca diagnóstico. A pergunta que ele responde é quanto tempo o paciente pode aguardar até ser visto por um médico (MACKWAY-JONES; MARSDEN; WINDLE, 2014). A resposta vem em cinco cores, cada uma com um tempo-alvo: vermelho, atendimento imediato; laranja, 10 minutos; amarelo, 60 minutos; verde, 120 minutos; azul, 240 minutos. Revisões da literatura mostram concordância entre avaliadores que vai de moderada a muito boa, mas também subtriagem e supertriagem (PARENTI *et al.*, 2014; SOUZA; ARAÚJO; CHIANCA, 2015), o que justifica medir esses dois desfechos em qualquer sistema que classifique por cores.

O profissional escolhe um fluxograma a partir da queixa principal, entre cerca de cinquenta. Cada fluxograma é uma lista de perguntas, os discriminadores, ordenadas da mais grave para a mais leve. A primeira resposta afirmativa encerra a triagem e define a cor; se nenhuma for afirmativa, o paciente é azul.

O protótipo segue o Protocolo de Classificação de Risco (baseado no protocolo de Manchester) da Secretaria Municipal de Saúde de Aparecida de Goiânia (APARECIDA DE GOIÂNIA, 2025), documento público que adapta o Manchester e traz 24 fluxogramas de adultos, a escala de dor e as tabelas de sinais vitais. Não é uma implementação oficial do Manchester, cujo manual é licenciado. Quando o protocolo dá duas cores para a mesma faixa, o protótipo adota a mais grave.

2.2. National Early Warning Score 2 (NEWS2)

O NEWS2 é um escore do Royal College of Physicians (2017) para identificar pacientes em piora clínica. Cada sinal vital recebe de 0 a 3 pontos conforme se afasta da faixa normal, e a soma indica o risco. O escore original usa sete parâmetros; neste trabalho foram usados quatro (frequência cardíaca, pressão sistólica, saturação de oxigênio e temperatura), que são os que o sistema consegue medir, e por isso a versão é chamada de adaptada. Em urgência e no atendimento pré-hospitalar, ainda não está claro qual escore de alerta prevê melhor a piora do

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

paciente (GUAN *et al.*, 2022), e uma revisão sistemática apontou fragilidades metodológicas em muitos escores (GERRY *et al.*, 2020). Por isso, neste trabalho o escore continua sendo calculado e registrado, mas apenas como informação: a cor não depende dele.

2.3. Crise hipertensiva

As Diretrizes Brasileiras de Hipertensão Arterial (BARROSO *et al.*, 2021) definem crise hipertensiva como pressão sistólica (PAS) a partir de 180 mmHg e/ou diastólica (PAD) a partir de 120 mmHg e a dividem em dois tipos. Na emergência hipertensiva há sinais de lesão de órgão-alvo, como dor no peito, falta de ar ou cefaleia intensa. Na urgência, a pressão está igualmente elevada, mas sem esses sinais. No protótipo, a cor da pressão não depende dessa distinção, e sim das faixas do protocolo de referência: pressão sistólica de 160 a 199 mmHg ou diastólica a partir de 100 mmHg resulta em amarelo, e sistólica a partir de 200 mmHg, em laranja. Os cortes de 160 e 100 mmHg coincidem com o início do estágio 2 das diretrizes.

2.4. Chamada de função em modelos de linguagem

A chamada de função (*function calling*) permite que o modelo devolva uma estrutura com campos definidos, conforme um esquema fornecido pelo desenvolvedor (OPENAI, 2026a). Quando o esquema restringe um campo a uma lista fechada de valores, o modelo só pode escolher entre eles. O sistema usa esse recurso para limitar a saída da LLM às chaves do catálogo de sintomas.

3. METODOLOGIA

3.1. Visão geral

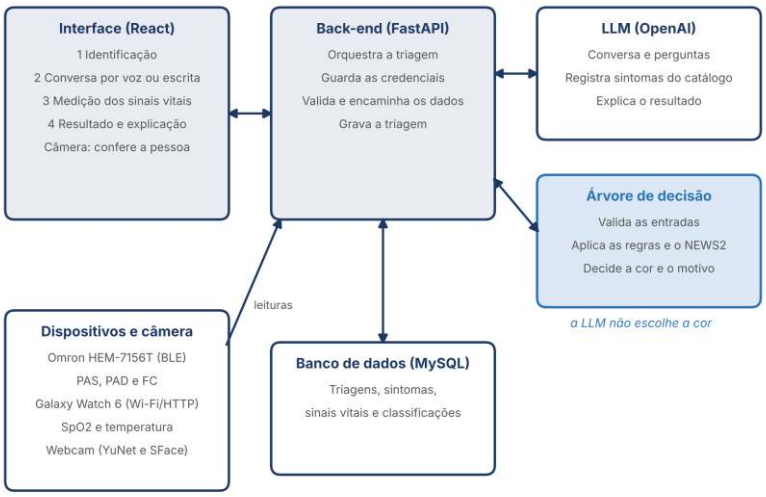
A interface em React capta o áudio, reproduz a voz da assistente e exibe o resultado. O servidor em Python, implementado com FastAPI, guarda as credenciais, comunica-se com a LLM, recebe as medições e chama a árvore de decisão (Figura 1). Antes da identificação, a interface apresenta o termo de consentimento. Após o aceite, o usuário se identifica, responde à entrevista por voz ou por escrito e realiza a medição dos sinais vitais. A conclusão apresenta a classificação, a duração da triagem e o prazo recomendado de atendimento. A assistente explica o resultado por voz. Quando a gravação está ativada, o sistema registra a triagem em MySQL ou MariaDB. Quem informa uma deficiência que impede o uso do protótipo sem ajuda é encaminhado à enfermagem.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

3.2. Catálogo de sintomas

O catálogo tem 102 chaves: 94 sintomas, 7 fatores de risco e a chave queixa_antiga, que registra a duração da queixa. Os sintomas são agrupados por parte do corpo ou sistema: neurológico, coração e circulação, respiratório, digestivo, urinário e reprodutivo, alergia e pele, trauma, olhos, ouvidos, nariz e boca, músculos e articulações, saúde mental e sintomas gerais. Cada sintoma carrega uma cor-base, a prioridade que ele justifica sozinho, igual à do discriminador correspondente no fluxograma da queixa do protocolo de referência (APARECIDA DE GOIÂNIA, 2025). A dor entra pela escala do protocolo, com as chaves dor_leve (nota 1 a 3, verde), dor_moderada (4 a 6, amarela) e dor_intensa (7 a 10, laranja). Se uma queixa que daria verde começou há mais de sete dias, a cor passa a azul. A Tabela 1 mostra a distribuição.

Figura 1. Arquitetura do sistema: a LLM conversa e extrai sintomas, e só a árvore de decisão define a cor



Fonte: Elaborado pelos autores (2026).

Tabela 1. Distribuição dos 94 sintomas do catálogo por cor-base

Cor-base	Sintomas	Exemplos
Vermelho	8	não responde, convulsão ativa, engasgo, lábios roxos
Laranja	19	déficit neurológico, dor no peito intensa, vômito com sangue, intoxicação
Amarelo	18	dor no peito, falta de ar, dor moderada, desmaiou
Verde	43	febre, tosse, vômito, diarreia, dor leve

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Cor-base	Sintomas	Exemplos
Azul	6	tremores, insônia, fraqueza, falta de apetite

Fonte: Elaborado pelos autores (2026).

Há ainda sete fatores de risco, que não têm cor própria e não mudam a cor da classificação: gestante, idoso, diabético, cardiopata, asma ou DPOC, imunidade baixa e uso de anticoagulante. A chave outro_sintoma, que faz parte dos 94 sintomas e tem cor-base verde, recebe os relatos que não têm correspondência no catálogo, o que evita que a LLM ignore o que foi dito ou force um sintoma parecido.

3.3. Árvore de decisão

A árvore recebe pressão sistólica (PAS) e diastólica (PAD), frequência cardíaca (FC), saturação de oxigênio (SpO2), temperatura, o aviso de batimento irregular do monitor Omron, a lista de sintomas (que inclui os fatores de risco e a duração da queixa) e o perfil do paciente. Antes de classificar, ela valida as entradas: valores fora da faixa fisiológica, PAD maior que PAS ou chaves inexistentes geram erro, e nenhuma cor é devolvida. A escolha foi deliberada. Uma leitura errada do sensor ou um sintoma inventado pela LLM não deve virar uma classificação aparentemente válida.

As regras estão organizadas em seis níveis (Tabela 2). A árvore avalia todas, registra as que se aplicam e fica com a de maior gravidade; em empate de cor, vale a primeira na ordem da tabela. As demais regras acionadas voltam junto com o resultado, como achados secundários.

Tabela 2. Níveis de decisão da árvore

Nível	Critério	Cor
N1	Risco de morte: sintoma de alarme (via aérea, cianose, convulsão ativa, não responde, hemorragia incontrolável), choque ou choque anafilático	Vermelho
N2	FC < 40 bpm, SpO2 < 90% ou temperatura ≤ 34 °C	Vermelho
N2	FC de 40 a 59 ou ≥ 150 bpm, ritmo irregular, PAS < 80 ou PAD < 50 mmHg, PAS ≥ 200 mmHg ou temperatura ≥ 41 °C	Laranja
N2	FC de 101 a 149 bpm, SpO2 de 90 a 94%, PAS de 160 a 199 ou PAD ≥ 100 mmHg, ou temperatura de 38,5 a 40,9 °C	Amarelo
N2	Temperatura de 37,5 a 38,4 °C ou de 34,1 a 36 °C	Verde

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Nível	Critério	Cor
N3	Gestante com PAS ≥ 160 ou PAD ≥ 110 mmHg [exceção obstétrica]	Vermelho
N3	Gestante com suspeita de pré-eclâmpsia ou com dor de barriga [exceção obstétrica]	Laranja
N4	Sintoma muito urgente; dor precordial; falta de ar com SpO2 < 92%; sangramento vaginal na gestação	Laranja
N4	Sintoma urgente; ardência ao urinar ou dor de garganta com febre	Amarelo
N4	Sintoma pouco urgente (evento recente)	Verde
N4	Sintoma não urgente	Azul
N5	Queixa há mais de 7 dias que daria verde	Azul
N6	Sinais vitais normais e nenhum discriminador de urgência	Azul

Fonte: Elaborado pelos autores (2026). Nota: FC: frequência cardíaca; PAS e PAD: pressão arterial sistólica e diastólica; SpO2: saturação de oxigênio.

O NEWS2 adaptado pontua os quatro parâmetros disponíveis (frequência cardíaca, pressão sistólica, saturação e temperatura) de 0 a 3 pontos cada, a partir das faixas do escore original. Na versão atual, o escore é devolvido pelo *back-end* e gravado no banco, mas é apenas informativo: não decide a cor, que vem das faixas do protocolo (Tabela 2).

A árvore tem ainda 10 regras de combinação, necessárias porque um sintoma isolado nem sempre reflete o risco. Todas vêm do protocolo de referência, exceto as três da gestação, marcadas no código como exceção obstétrica. As demais são duas de alarme (choque e choque anafilático), três muito urgentes (dor precordial, falta de ar com SpO2 abaixo de 92% e sangramento vaginal na gestação) e duas urgentes (ardência ao urinar e dor de garganta, ambas com febre). Dor no peito sozinha é amarela; com suor frio, enjoo ou vômito, passa a laranja, por suspeita de infarto. Alergia na pele com falta de ar, chiado ou sensação de desmaio vai a vermelho, por suspeita de anafilaxia. Para essas regras, temperatura a partir de 37,8 °C conta como febre. Cada regra ocupa uma linha no código, com o motivo escrito por extenso, o que permite a um profissional de saúde revisar a lista sem ler o programa.

A árvore também usa o perfil do paciente, informado na identificação e não na conversa. Com 65 anos ou mais, o fator de risco “idoso” entra sozinho, sem mudar a cor; abaixo de 14 anos, o resultado avisa que o protocolo usa fluxogramas e limites pediátricos e que a enfermagem deve confirmar a classificação, já que a árvore aplica os limites de adulto. O sexo

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

faz a árvore ignorar chaves que só fazem sentido no outro sexo, como gestante para o sexo masculino. Deficiência, 60 anos ou mais e gestação não mudam a cor, mas indicam atendimento prioritário dentro dela, conforme a Lei nº 10.048/2000 (BRASIL, 2000).

3.4. Medição da pressão arterial

O servidor lê a pressão sistólica (PAS), a pressão diastólica (PAD) e a frequência cardíaca (FC) do monitor Omron HEM-7156T (OMRON HEALTHCARE ASIA PACIFIC, 2026) por Bluetooth Low Energy, com a biblioteca Bleak. A comunicação usa o protocolo proprietário BLESmart, documentado por engenharia reversa no projeto omblepy (OMBLEPY, 2026), consultado na revisão a201ebd e usado como referência documental, e não como dependência; o protocolo não é documentado pelo fabricante. O *software* estabelece uma conexão GATT, desbloqueia o aparelho com uma chave na camada de aplicação e decodifica os registros com base no mapa de memória do modelo HEM-7155T.

Na etapa de medição, o servidor consulta o aparelho à procura de um novo registro. O prazo padrão é de 120 s. Se a leitura automática falhar ou o monitor estiver desativado, a interface permite inserir os valores manualmente. O banco registra a origem da leitura e distingue as medições reais das simuladas.

3.5. Oximetria e temperatura pelo relógio

O Galaxy Watch 6 (SM-R940, Wear OS) mede a saturação de oxigênio (SpO2) e a temperatura da pele do pulso. O aplicativo usa o Samsung Health Sensor SDK 1.4.1 e os exemplos do Samsung Code Lab (SAMSUNG ELECTRONICS, 2026a, 2026b, 2026c). Após o comando do usuário, o aplicativo consulta primeiro o sensor de temperatura da pele e libera esse sensor antes de iniciar a oximetria sob demanda.

Ao concluir a oximetria, o relógio envia pela rede Wi-Fi local um JSON com a SpO2, a temperatura da pele e os horários de medição, que a rota POST /spo2 do FastAPI valida e repassa à interface.

A temperatura da pele preenche o campo de temperatura da triagem e entra no cálculo com o valor recebido; o banco registra a origem cutânea e o horário. Sem essa leitura, a temperatura é digitada. A temperatura da pele do pulso (temperatura cutânea), porém, não equivale à temperatura corporal, e as faixas do protocolo foram definidas para a temperatura clínica; por isso, os resultados que dependem desse valor exigem cautela (Seções 4.5 e 4.8).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

3.6. Papel da LLM

A LLM tem um papel deliberadamente estreito. Ela conversa, faz perguntas de esclarecimento e traduz o relato para chaves do catálogo. Não escolhe a cor, não dá diagnóstico nem indica medicamento, e o *prompt* a proíbe de dizer qual é a classificação antes de a árvore responder.

O *prompt* é gerado automaticamente a partir do catálogo, de modo que os dois não ficam desencontrados. Ele pede respostas de uma ou duas frases, por serem faladas, e uma pergunta por vez. Para a queixa principal, a assistente deve esclarecer intensidade (com a nota da dor de 0 a 10), início súbito, duração e sintomas associados; na dor no peito, por exemplo, pergunta sobre suor frio, enjoo e falta de ar, porque essas respostas mudam a cor. Pergunta uma única vez por condições especiais, como gravidez, diabetes ou uso de anticoagulante. Diante de um relato muito grave, encerra as perguntas e registra os sintomas na hora. O *prompt* recebe ainda a idade, o sexo e as deficiências da identificação, para não perguntar o que já se sabe e adaptar a conversa.

Os sintomas chegam à árvore pela ferramenta registrar_sintomas, que recebe a lista de chaves e um resumo de uma frase. A lista é restrita às 102 chaves do catálogo, e a árvore rejeita qualquer chave desconhecida, o que cria uma segunda barreira. Depois da classificação, a LLM recebe cor, prioridade, motivo e orientação e os explica em duas ou três frases, sem alterar o resultado. Se a chamada à LLM falhar nesse momento, o *back-end* monta a frase a partir do resultado da árvore.

Há dois modos de conversa. No clássico, o sistema transcreve a fala, envia o texto ao modelo de *chat* e sintetiza a resposta. O código configura gpt-4o-mini por padrão e usa temperatura de amostragem 0,3 nas chamadas de *chat*. No modo em tempo real, o navegador troca áudio com a API Realtime da OpenAI por WebRTC (OPENAI, 2026b). O modelo padrão é gpt-realtime-mini e a detecção de fim de fala usa 600 ms de silêncio. O servidor fornece uma credencial temporária para a sessão e mantém a credencial principal no servidor. Se a conexão em tempo real falhar, a interface passa ao modo clássico.

3.7. Protocolo de verificação

A árvore foi testada com 142 pacientes fictícios: 80 de sintomas e sinais vitais (P01 a P80), 38 de limites exatos dos sinais vitais (S01 a S38), 14 de dor, consciência, combinações e duração da queixa (D01 a D14) e 10 de perfil, com idade, sexo e deficiências (Q01 a Q10). Cada caso tem sinais vitais, sintomas e a cor esperada, definida pelos autores a partir do protocolo de

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

referência, e não da saída do programa. O teste da interface de ferramenta usada pela LLM foi feito à parte e não integra esses 142 casos.

A extração de sintomas foi avaliada com 118 relatos escritos para este estudo, em seis categorias: linguagem comum (51), gíria (15), vários sintomas na mesma fala (16), negação (14), erros de transcrição (12) e tentativas de induzir erro (10). Os autores definiram as chaves esperadas de cada relato; cinco relatos não têm chave esperada. Em 16/11/2026, cada relato foi enviado em texto, três vezes, à LLM (gpt-4o-mini, temperatura 0,3) pelo código do *back-end*, sem dados de identificação (354 execuções). Quando a LLM fazia uma pergunta, respondia-se “Só isso mesmo, não tenho mais nada.”, com limite de cinco rodadas. A execução foi exata (chaves iguais às esperadas), parcial (alguma chave em comum) ou errada (nenhuma em comum). Calcularam-se ainda a precisão e a cobertura por chave e a cor que a árvore devolveria para as chaves extraídas, comparada à das esperadas, com sinais vitais normais.

Além dos testes automáticos, o protótipo foi apresentado a uma profissional da área da saúde, Mayara Santos Machado, do Hospital Santa Rita (São Paulo, SP), que o usou com medições do monitor de pressão e do relógio e também com valores digitados à mão, e comentou as tabelas de sintomas, cores e pontuação do NEWS2. A reunião foi qualitativa e informal, sem protocolo de medida de concordância.

Os casos foram avaliados ainda por Gabriel Mariano, técnico de enfermagem da Fundação ABC, com dois anos de experiência em classificação de risco e experiência prévia com o protocolo de Manchester. Em 09/11/2026, o avaliador recebeu uma planilha com 101 dos 142 casos (P01 a P80, D01 a D14 e sete casos de perfil), com os sinais vitais e as queixas de cada paciente, e escolheu a cor que daria; a resposta da árvore ficava em outra aba, aberta depois. Ficaram de fora os 38 casos de limites exatos (S01 a S38) e três de perfil (Q04, Q06 e Q07). Ele indicou também se concordava com cada uma das 20 regras de sinais vitais e com a cor de cada uma das 102 chaves. Calcularam-se a concordância e o kappa de Cohen, simples e ponderado, com intervalo de confiança de 95% (*bootstrap* para o kappa; Wilson para a concordância). Subtriagem foi a cor da árvore menos grave que a do avaliador, e supertriagem, a mais grave.

4. RESULTADOS E DISCUSSÃO

4.1. Verificação da árvore de decisão

Os 142 casos fictícios foram executados, e a árvore devolveu a cor e o nível esperados em todos (142 de 142). Entre os casos de fronteira, SpO2 de 89% resultou em vermelho e de 90%, em amarelo; pressão de 199/110 mmHg, em amarelo, e de 200/110 mmHg, em laranja;

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

falta de ar com SpO2 de 91%, em laranja, e com 92%, em amarelo. O limite conhecido da versão anterior, em que 179/119 mmHg sem sintomas resultava em azul, deixou de existir, porque a faixa de 160 a 199 mmHg passou a ser amarela (caso P14). Esses testes mostram que o código segue as regras escritas, não que elas estejam clinicamente corretas (Seção 4.6).

Variando a SpO2 e a temperatura isoladamente, com os demais valores normais, a cor mudou nestes pontos. Sem sintomas, a SpO2 resultou em azul com 95% ou mais, amarelo de 94% a 90% e vermelho com 89% ou menos; com falta de ar, em amarelo com 92% ou mais, laranja com 91% e 90%, e vermelho com 89% ou menos. Na temperatura, sem sintomas, houve vermelho até 34,0 °C, verde de 34,1 a 36,0, azul de 36,1 a 37,4, verde de 37,5 a 38,4, amarelo de 38,5 a 40,9 e laranja a partir de 41,0 °C; com falta de ar, vermelho até 34,0, amarelo de 34,1 a 40,9 e laranja a partir de 41,0 °C.

A explicação da LLM foi testada em 12 cenários com SpO2 e temperatura alterados (36 explicações). Ela recebe da árvore só o resumo do resultado (cor, prioridade, motivo e orientação). Nas 36, a fala citou a cor da árvore e nenhuma outra, não trouxe diagnóstico nem dose e, nos resultados laranja e vermelho, orientou atendimento imediato.

4.2. Extração de sintomas pela LLM

Em 271 das 354 execuções (76,6%; intervalo de confiança de 95% de 71,9% a 80,7%), as chaves extraídas foram exatamente as esperadas; 49 (13,8%) foram parciais e 34 (9,6%), erradas. Por chave, a precisão foi de 88,1% e a cobertura, de 89,0%; nas chaves de cor-base vermelha, a cobertura foi de 93,3% (28 de 30) e nas de cor-base laranja, de 90,8% (79 de 87), com as falhas concentradas em trauma_grave (0 de 3), picada_animal_peconhento (1 de 3), intoxicacao (3 de 6) e inconsciente (4 de 6). Os relatos com vários sintomas na mesma fala (62,5% de execuções exatas) e com negação (57,1%) tiveram o pior desempenho (Tabela 3). Em 107 dos 118 relatos, as três repetições deram o mesmo resultado. A cor da árvore para as chaves extraídas foi igual à esperada em 308 execuções (87,0%), mais leve em 31 (8,8%) e mais grave em 15 (4,2%). Em dez execuções de quatro relatos, a cor esperada era laranja ou vermelha e a obtida foi mais leve: acidente de moto sem trauma_grave (R015), picada de escorpião sem picada_animal_peconhento (R018), ingestão de medicamento com ideia de suicídio sem intoxicacao (R077) e a frase “ele disacordo naum rispondi”, sem inconsciente (R101). Esse é o erro de maior consequência clínica, porque corresponde a uma subtriagem potencial: o paciente seria orientado a esperar mais do que deveria. Como a árvore não subtriou nenhum dos 101 casos avaliados (Seção 4.6), a principal fonte de subtriagem do sistema está na extração, quando a LLM deixa de registrar um discriminador grave.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Tabela 3. Extração de sintomas pela LLM em 118 relatos (354 execuções, três por relato), por categoria

Categoria (relatos)	Execuções	Exato	Parcial	Errado
Linguagem comum (51)	153	120 (78,4%)	18 (11,8%)	15 (9,8%)
Gíria (15)	45	40 (88,9%)	2 (4,4%)	3 (6,7%)
Vários sintomas (16)	48	30 (62,5%)	18 (37,5%)	0
Negação (14)	42	24 (57,1%)	6 (14,3%)	12 (28,6%)
Erro de transcrição (12)	36	32 (88,9%)	2 (5,6%)	2 (5,6%)
Tentativa de induzir erro (10)	30	25 (83,3%)	3 (10,0%)	2 (6,7%)
Total (118)	354	271 (76,6%)	49 (13,8%)	34 (9,6%)

Fonte: Elaborado pelos autores (2026). Nota: exato: chaves iguais às esperadas; parcial: alguma chave em comum; errado: nenhuma chave em comum.

4.3. Tempo de resposta

Nos 118 relatos da Seção 4.2 (354 execuções e 615 chamadas à LLM), cada rodada levou em média 1,20 s (desvio-padrão de 0,37 s; mediana de 1,10 s; percentil 95 de 1,79 s), e o registro dos sintomas, 2,04 s por execução, com 1,74 rodada em média. A explicação do resultado levou 1,32 s (desvio-padrão de 0,24 s; percentil 95 de 1,60 s; 36 chamadas). Os testes rodaram em um computador com Windows 11, processador Intel Core i5-10400F e 16 GB de RAM, ligado à internet por cabo, e os tempos incluem a ida e a volta à API da OpenAI. Em medições anteriores do modo clássico, a voz demorou 1,80 s até o primeiro trecho de áudio, e a criação da sessão no modo tempo real, 0,38 s.

4.4. Integração com o monitor de pressão

A primeira tentativa de conexão caía durante o pareamento. Um registro HCI (btsnoop) indicou que o aparelho usa o protocolo proprietário BLESmart, e não o perfil padrão de pressão arterial, e que a causa provável era a chamada a client.pair(), do pareamento nativo do WinRT, com falhas documentadas (BLEAK, 2024). Sem essa chamada, a conexão deixou de cair e o código leu e decodificou os 49 registros guardados no aparelho, de 28/08 a 10/10/2026.

Os *scripts* leem a memória do monitor e decodificam PAS, PAD, FC, data e hora, com sinalizadores de movimento e batimento irregular. Em 10 medições, a PAS, a PAD e a FC decodificadas pelo *software* foram idênticas às exibidas no visor do monitor.

4.5. Oximetria e temperatura pelo relógio

Em 10 medições, o relógio indicou SpO2 de 97% a 98%, com diferença de até 1 ponto percentual em relação a um oxímetro, e temperatura da pele do pulso de 34 a 35 °C, enquanto uma câmera termográfica FLIR indicou de 36,0 a 36,4 °C. A câmera termográfica também não é um termômetro clínico de referência, de modo que a diferença de 1 a 2 °C indica a ordem de

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

grandeza do desvio, e não um fator de correção. Como a temperatura da pele entra na árvore sem correção, leituras como 34,8 °C resultariam em verde por hipotermia leve em uma pessoa sem hipotermia, e uma febre real poderia não ser reconhecida.

4.6. Avaliação com profissionais da área da saúde e teste com usuários

O protótipo foi apresentado a uma profissional da área da saúde, Mayara Santos Machado, do Hospital Santa Rita (São Paulo, SP), que avaliou as tabelas de sintomas e cores e a pontuação do NEWS2 adaptado em uma reunião qualitativa e informal, sem revisão sistemática nem medida de concordância. Ela também usou o sistema, com medições dos aparelhos e valores digitados, e concordou com os resultados de dois relatos simulados por voz.

Para um caso de exemplo (febre e tosse, 38 °C, SpO2 de 95%, frequência cardíaca de 92 bpm e idade avançada), a árvore indicou amarelo; a profissional considerou-o adequado, atribuiu o peso principal à idade e observou que, dependendo do hospital, poderia ser verde. Lembrou que comorbidades, como diabetes, podem elevar a prioridade, o que reforça a pergunta sobre fatores de risco feita pela assistente, e discutiu-se definir limites por sintoma e sinal vital que acionem o chamado de um profissional.

Na avaliação estruturada, em 98 dos 101 casos (97,0%; intervalo de confiança de 95% de 91,6% a 99,0%) a cor escolhida por Gabriel Mariano foi igual à da árvore. O kappa de Cohen foi de 0,96 (IC 95% de 0,91 a 1,00) e o kappa ponderado, de 0,98 (0,93 a 1,00). As três divergências foram de supertriagem, isto é, a árvore deu uma cor mais grave que a dele: dor de garganta com febre (D09), amarelo contra azul; pancada na cabeça (P61), amarelo contra verde; e coriza (P77), verde contra azul. Não houve subtriagem. A Tabela 4 mostra a matriz entre as cores.

Tabela 4. Cores do avaliador (linhas) e da árvore (colunas) nos 101 casos avaliados

Avaliador / Árvore	Vermelho	Laranja	Amarelo	Verde	Azul	Total
Vermelho	11	0	0	0	0	11
Laranja	0	19	0	0	0	19
Amarelo	0	0	37	0	0	37
Verde	0	0	1	24	0	25
Azul	0	0	1	1	7	9
Total	11	19	39	25	7	101

Fonte: Elaborado pelos autores (2026).

Nas regras, o avaliador concordou com 98 das 102 chaves do catálogo e com 16 das 20 regras de sinais vitais. Discordou de quatro faixas: frequência cardíaca a partir de 150 bpm (acima de 180 bpm seria vermelho); saturação abaixo de 90% (vermelho só abaixo de 80% e laranja de 80 a 92%); falta de ar com saturação abaixo de 92% (vermelho, e não laranja); e

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

pressão sistólica abaixo de 80 ou diastólica abaixo de 50 mmHg (vermelho, e não laranja). Entre as chaves, discordou de gestante e idoso, com o comentário “Prioridade” (o protótipo já os sinaliza como prioritários dentro da mesma cor), e, sem comentário, de imunidade baixa e da duração da queixa de mais de sete dias.

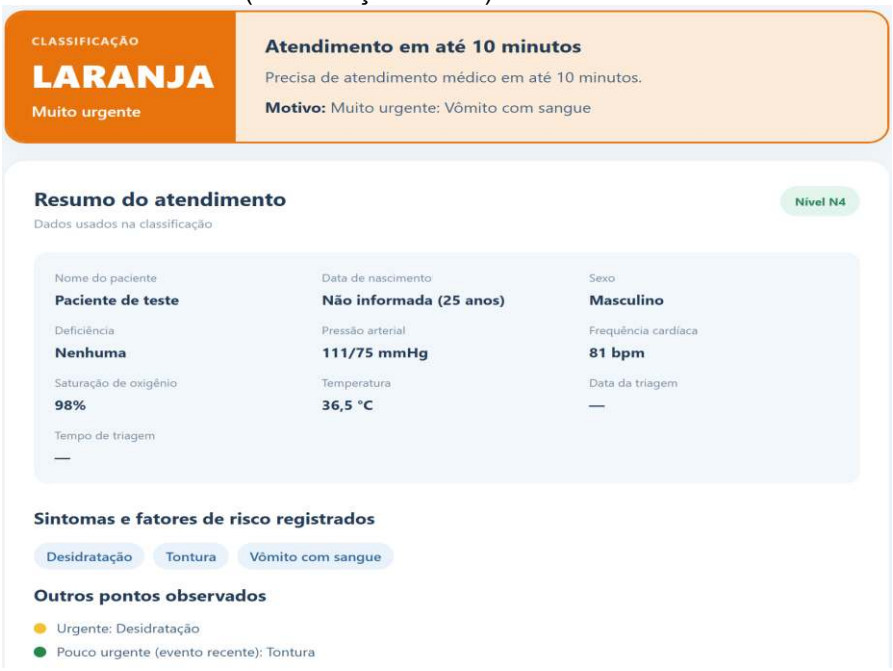
Nos casos, os comentários do avaliador pediram atenção a possíveis causas graves (meningite em P33; AVC ou glaucoma em P35; hipertensão em P45) e à glicemia em um diabético com tremores (P71); em todos eles a cor dele foi igual à da árvore. Para D07 (falta de ar com SpO2 de 91%), ele escolheu laranja, mas nas regras sugeriu vermelho para saturação abaixo de 92% com falta de ar. As sugestões por regra não foram implementadas. Em casos equivalentes ao exemplo da reunião (P68 e P70), atribuiu verde, a cor atual da árvore.

Esse resultado tem limites. Houve um único avaliador, técnico de enfermagem e não enfermeiro, e a classificação de risco é privativa do enfermeiro (Seção 4.9). Os casos foram escritos pela equipe, o que pode favorecer a concordância. O kappa mede a concordância entre ele e a árvore, e não entre avaliadores, e os intervalos são largos para as cores com poucos casos. A concordância mostra coerência com o julgamento de uma pessoa, e não segurança clínica.

Em 16/11/2026, 14 pessoas de 17 a 58 anos, três delas autoras deste trabalho, usaram o protótipo por voz em uma instituição de ensino superior, em ambiente não silencioso, após aceitar o termo de consentimento. Houve 15 triagens (um autor repetiu a triagem): 14 concluídas e uma cancelada pela câmera, após 10 s sem rosto. A conferência de presença, ajustável na área de configuração, também encerrou a triagem quando outra pessoa entrou. A pressão veio do monitor Omron; SpO2 (98%) e temperatura (36,5 °C; 37,9 °C em um caso) foram valores padrão digitados. As 14 concluídas resultaram em azul (2), verde (6), laranja (4) e vermelho (2); a Figura 2 mostra a tela de conclusão para os dados de uma delas.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 2. Tela de conclusão gerada pelo protótipo para os sintomas e sinais vitais de uma triagem do teste com usuários (identificação fictícia)



Fonte: Elaborado pelos autores (2026).

Os autores encontraram problemas de extração em 7 das 14 triagens concluídas: omissão de falta_de_ar ou cefaleia súbita (três), sintomas inventados após respostas ininteligíveis (dois, um deles levando a laranja) e local da dor errado (dois). A voz em ambiente ruidoso piorou a extração em relação ao texto limpo (Seção 4.2). Nas três omissões, a cor registrada pode ter sido mais leve que a adequada, o que também caracteriza risco de subtriagem. Essa classificação dos erros não foi revisada por profissional de saúde. Repetir os sintomas entendidos e pedir confirmação reduziria o risco, mas não foi implementado.

4.7. Discussão

O ganho principal da arquitetura é poder explicar cada resultado. Para qualquer triagem, dá para dizer quais sintomas foram registrados, quais regras se aplicaram e qual delas definiu a cor. Isso permite testar o sistema com casos escritos de antemão, algo difícil quando a LLM decide sozinha; os 142 casos garantem que a árvore devolva o que as regras determinam, o que não equivale a acerto clínico. A avaliação por um técnico de enfermagem mostrou concordância alta, com divergências só de supertriagem, o desfecho menos perigoso em uma pré-triagem, mas com um único avaliador e casos escritos pela equipe (Seção 4.6).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

O risco passa para outro ponto: a extração. Se a LLM omite um sintoma ou registra outro, a árvore classifica corretamente o que recebe, mas a cor será coerente com uma situação que não é a real. Os estudos citados na Introdução usaram a LLM para decidir a gravidade (WILLIAMS *et al.*, 2024; PASLI *et al.*, 2024); aqui ela só extrai sintomas, e a extração passou a ser o ponto mais frágil. Nos 118 relatos, as chaves exatas ficaram em 76,6% das execuções, com pior resultado em negações e em falas com vários sintomas, e a cor obtida foi mais leve que a esperada em 8,8% das execuções, incluindo dez de relatos graves; no teste falado, houve três omissões de sintomas que elevariam a cor. Em uma pré-triagem, esse é o risco que mais importa, porque a subtriagem atrasa o atendimento de quem precisa dele primeiro, enquanto a supertriagem apenas antecipa um atendimento.

Três medidas podem reduzir esse risco e ficam como trabalho futuro: fazer a assistente repetir os sintomas entendidos e pedir confirmação antes da medição; acrescentar regras de alerta independentes da LLM para termos graves, como desmaio, sangramento, envenenamento ou falta de ar; e manter a confirmação do resultado pelo enfermeiro, que é a condição de uso do sistema como pré-triagem (Seção 4.9). A validação contra um padrão definido por profissionais de saúde, com relatos reais e falados, ainda falta.

4.8. Limitações

Validação clínica incompleta. As cores dos sintomas e as faixas dos sinais vitais foram transcritas de um protocolo municipal, e as regras da gestação são dos autores. Os 142 casos e os 101 avaliados foram escritos pela mesma equipe que escreveu as regras, a partir do mesmo protocolo, o que tende a aumentar a concordância: são descrições completas e sem ambiguidade, sem a informação incompleta, a comunicação não verbal e a evolução de um paciente real. Por isso, os 142 de 142 da árvore mostram que o código segue as regras, e não que elas estejam corretas, e o kappa de 0,96 deve ser lido como um limite superior. A avaliação estruturada (Seção 4.6) teve um único avaliador, técnico de enfermagem, e não enfermeiro: não há medida de concordância entre avaliadores, o julgamento de uma pessoa pode refletir hábitos do seu serviço, e as cores com poucos casos têm intervalos de confiança largos. A reunião com uma profissional da área da saúde foi informal, sem medida de concordância. Faltam a avaliação por mais de um enfermeiro de triagem, com casos escritos por terceiros ou reais, e a revisão das faixas de sinais vitais graves apontadas pelo avaliador.

Risco de subtriagem na extração. A árvore não subtriou nenhum caso avaliado, mas a extração resultou em cor mais leve que a esperada em 8,8% das execuções, inclusive em relatos

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

de trauma grave, intoxicação, picada de animal peçonhento e perda de consciência, e omitiu sintomas no teste falado. O sistema ainda não tem mecanismo que detecte essas omissões.

Validação dos sensores e temperatura cutânea. As 10 medições de comparação não incluem análise de concordância (Bland-Altman) nem um termômetro clínico. A pressão decodificada coincidiu com o visor e a SpO2 ficou a até 1 ponto do oxímetro, mas a temperatura da pele ficou de 1 a 2 °C abaixo da câmera termográfica. A temperatura cutânea do pulso varia com o ambiente e a circulação periférica e não equivale à temperatura corporal; na versão atual, ela entra na árvore como se fosse temperatura clínica, o que pode gerar hipotermia aparente ou esconder febre. Até a troca por um termômetro clínico, ela não deve ser interpretada pelas faixas clínicas do protocolo, e os resultados que dependem da temperatura não devem ser considerados confiáveis.

NEWS2 apenas informativo. O escore usa quatro dos sete parâmetros originais e não decide a cor, de modo que a redução não afeta a classificação; ele é calculado e mostrado a título de informação.

Faixas de sinais vitais graves. O avaliador propôs revisar as faixas de frequência cardíaca, saturação, falta de ar associada à baixa saturação e hipotensão. A Seção 4.6 descreve as sugestões e as diferenças entre as respostas por caso e por regra.

Testes de voz e de usuários limitados. Os 118 relatos foram escritos pelos autores, enviados em texto e testados em um único dia. O teste falado teve 14 pessoas, em um dia, com SpO2 e temperatura digitadas e erros classificados pelos próprios autores, sem apreciação prévia por comitê de ética em pesquisa.

Custo. O modo tempo real é mais rápido e mais caro por minuto que o clássico. O custo por triagem não foi medido: depende da duração da conversa e do volume de áudio e de texto enviado à API, e precisa ser estimado antes de qualquer uso contínuo.

Pediatria. Abaixo de 14 anos, o protocolo usa fluxogramas e limites pediátricos que o protótipo não implementa; a árvore aplica os limites de adulto e pede a confirmação da enfermagem.

O sistema é um protótipo acadêmico e não substitui a triagem feita por profissional qualificado.

4.9. Privacidade, regulação e uso clínico

O protótipo apresenta um termo antes da identificação, que informa os dados usados, o envio da conversa à OpenAI e o uso da câmera, e indica a equipe como contato para pedidos

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

sobre os dados. A data, a hora do aceite e a versão do termo são gravadas com cada triagem, concluída, encaminhada ou cancelada.

A aplicação envia nome completo, CPF e data de nascimento ao banco do sistema. A entrevista usa o relato, o áudio, a idade, o sexo e as deficiências. No modo clássico, o primeiro nome integra a saudação sintetizada. Quando a câmera está habilitada, YuNet detecta o rosto (WU; PENG; YU, 2023) e SFace gera a representação facial usada para verificar a continuidade da pessoa diante do equipamento (ZHONG *et al.*, 2021). O processamento ocorre localmente e a representação permanece em memória durante a sessão. Dados de saúde e dados biométricos vinculados a uma pessoa são dados pessoais sensíveis (BRASIL, 2018). A documentação do provedor descreve as condições de uso e retenção de dados da API (OPENAI, 2026c). A base legal considerada é o consentimento do titular, dado de forma específica e destacada pelo termo (art. 11, inciso I, da LGPD). A reidentificação facial foi demonstrada, com encerramento da triagem quando outra pessoa entrou, mas não teve os erros medidos (falso aceite e falso bloqueio). Para uso institucional, ainda é preciso estabelecer prazos de guarda, avaliar a transferência internacional de dados e revisar a base legal com a assessoria jurídica da instituição.

Quanto à regulação, a classificação de risco e a priorização da assistência em serviços de urgência são privativas do enfermeiro (CONSELHO FEDERAL DE ENFERMAGEM, 2021), e um *software* com finalidade médica pode ser enquadrado pela Anvisa como *software* como dispositivo médico (BRASIL, 2022). Por isso, o sistema deve ser entendido como uma pré-triagem, cujo resultado é uma sugestão a ser confirmada pelo enfermeiro; o enquadramento do protótipo nessas normas ainda precisa ser verificado.

O Galaxy Watch é um equipamento de uso pessoal, e não hospitalar, e serve apenas ao protótipo; um produto real usaria equipamentos aprovados para uso hospitalar, de preferência ligados por cabo, porque o Bluetooth pode falhar e atrasar a transmissão. O protocolo do monitor Omron veio de engenharia reversa (omblepy), e não do fabricante.

5. CONSIDERAÇÕES FINAIS

O protótipo reúne entrevista por voz e classificação por regras com justificativa rastreável. A avaliação estruturada de Gabriel Mariano apresentou concordância em 98 de 101 casos (97,0%, kappa de Cohen de 0,96), com três casos de supertriagem e nenhum de subtriagem. A extração de sintomas foi testada com 118 relatos escritos, com as chaves esperadas em 76,6% das execuções e cor mais leve que a esperada em 8,8%, o que mostra que o risco de subtriagem

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

está na extração, e não na árvore. Os ensaios no macOS confirmaram a integração do Omron e do Galaxy Watch com o banco de dados; em 10 medições, a pressão coincidiu com o visor do monitor e a SpO2 ficou a até 1 ponto do oxímetro, mas a temperatura da pele ficou abaixo da medida pela câmera termográfica. No teste com 14 pessoas, a fala em ambiente ruidoso piorou a extração.

A avaliação seguinte deve substituir a temperatura da pele por um termômetro clínico, ampliar a revisão das regras e dos casos com enfermeiros de triagem e testar a extração com o catálogo atual em entrevistas variadas. A confirmação dos sintomas pelo paciente e a confirmação do resultado pelo enfermeiro dentro da aplicação permanecem como trabalhos futuros.

O código está disponível no GitHub sob licença MIT (TRIAGEM INTELIGENTE, 2026), com os *prompts* completos e a ferramenta da LLM (docs/prompts_llm.md), as instruções para repetir os testes (docs/REPRODUCAO.md) e os 142 casos da árvore com a cor esperada e a obtida (docs/casos_teste_arvore_142.csv). Usaram-se os nomes de modelo da API, sem versão datada, em testes de 16/11/2026; versões posteriores podem dar resultados diferentes. Os 101 casos avaliados, sem identificação do avaliador, e os 118 relatos com as 354 execuções constam do material suplementar (Casos_avaliados_101.xlsx e Relatos_extracao_LLM.xlsx).

DECLARAÇÕES

Uso de inteligência artificial. O sistema usa um modelo de linguagem da OpenAI para conduzir a entrevista e extrair os sintomas (Seção 3.6). Na elaboração do trabalho, os autores usaram o assistente Claude (Anthropic) para revisar o código em busca de erros, apoiar a organização do catálogo de sintomas da árvore de decisão e revisar a redação do artigo; todo o conteúdo foi conferido pelos autores, que são responsáveis por ele.

Contribuição dos autores. Vinicius Marçal: concepção, metodologia, *software* (árvore de decisão, monitor de pressão, LLM e integração), validação, análise e redação. Gabriel Camargo Felix: *software* (relógio e sensores) e validação. Juan David Fontecha: *software* (*front-end*). Hannah Daga Theodoro: banco de dados e redação (referencial e documentação). Giovana Nascimento Ambroselli: banco de dados, redação (documentação e monografia) e correção de falhas. Todos participaram dos testes e aprovaram a versão final.

Conflito de interesses. Os autores declaram não haver conflito de interesses.

Financiamento. Esta pesquisa não recebeu financiamento de agências de fomento, empresas ou instituições.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Agradecimentos. Ao Prof. Nelson Aguiar, pela orientação; a Mayara Santos Machado e Gabriel Mariano, pela avaliação do protótipo; e aos colegas que participaram do teste com usuários.

Ética e consentimento. A avaliação estruturada usou casos fictícios elaborados pelos autores, sem participação de pacientes. O teste de 16/11/2026 foi uma avaliação de usabilidade do protótipo com 14 voluntários, que aceitaram o termo de consentimento do sistema; nenhum nome ou documento dos participantes é apresentado neste artigo, e uma das participantes tinha 17 anos. O teste não foi submetido previamente a um Comitê de Ética em Pesquisa, como prevê a Resolução CNS nº 510/2016 para trabalhos de conclusão de curso (BRASIL, 2016), o que se registra como limitação (Seção 4.8). Os dois profissionais de saúde que avaliaram o protótipo, Mayara Santos Machado (Hospital Santa Rita) e Gabriel Mariano (Fundação ABC), autorizaram a citação de seus nomes e locais de trabalho.

REFERÊNCIAS

APARECIDA DE GOIÂNIA. Secretaria Municipal de Saúde. **Protocolo de classificação de risco**: baseado no protocolo de Manchester. Aparecida de Goiânia: Secretaria Municipal de Saúde, 2025.

BARROSO, W. K. S. *et al.* Diretrizes Brasileiras de Hipertensão Arterial – 2020. **Arquivos Brasileiros de Cardiologia**, v. 116, n. 3, p. 516-658, 2021. DOI: 10.36660/abc.20201238.

BITTENCOURT, R. J.; HORTALE, V. A. Intervenções para solucionar a superlotação nos serviços de emergência hospitalar: uma revisão sistemática. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 25, n. 7, p. 1439-1454, 2009. DOI: 10.1590/S0102-311X2009000700002.

BLEAK: issue n. 1528: extremely low timeout for bluetooth pairing on win11. [S. l.]: GitHub, 2024. Disponível em: https://github.com/hbldh/bleak/issues/1528. Acesso em: 22 set. 2026.

BRASIL. Lei nº 10.048, de 8 de novembro de 2000. Dá prioridade de atendimento às pessoas que especifica, e dá outras providências. **Diário Oficial da União**: seção 1, Brasília, DF, 9 nov. 2000.

BRASIL. Ministério da Saúde. **Acolhimento e classificação de risco nos serviços de urgência**. Brasília, DF: Ministério da Saúde, 2009.

BRASIL. Conselho Nacional de Saúde. Resolução nº 510, de 7 de abril de 2016. Dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais. **Diário Oficial da União**: seção 1, Brasília, DF, 24 maio 2016.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União**: seção 1, Brasília, DF, 15 ago. 2018.

BRASIL. Agência Nacional de Vigilância Sanitária. Resolução de Diretoria Colegiada – RDC nº 657, de 24 de março de 2022. Dispõe sobre a regularização de software como dispositivo médico

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

(Software as a Medical Device – SaMD). **Diário Oficial da União**: seção 1, Brasília, DF, 30 mar. 2022.

CONSELHO FEDERAL DE ENFERMAGEM. **Resolução Cofen nº 661, de 9 de março de 2021**. Atualiza e normatiza, no âmbito do Sistema Cofen/Conselhos Regionais de Enfermagem, a participação da Equipe de Enfermagem na atividade de Classificação de Risco. Brasília, DF: Cofen, 2021. Disponível em: https://cofen.gov.br/?p=85839. Acesso em: 3 nov. 2026.

GERRY, S. *et al.* Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. **BMJ**, v. 369, m1501, 2020. DOI: 10.1136/bmj.m1501.

GUAN, G. *et al.* The use of early warning system scores in prehospital and emergency department settings to predict clinical deterioration: a systematic review and meta-analysis. **PLoS ONE**, v. 17, n. 3, e0265559, 2022. DOI: 10.1371/journal.pone.0265559.

MACKWAY-JONES, K.; MARSDEN, J.; WINDLE, J. (ed.). **Emergency triage**: Manchester Triage Group. 3. ed. Chichester: Wiley Blackwell, 2014.

OMBLEPY: CLI tool to read records from Omron Bluetooth-LE measurement instruments. [S. l.]: GitHub, 2026. Revisão a201ebd, de 8 jul. 2026. Disponível em: https://github.com/userx14/omblepy. Acesso em: 22 set. 2026.

OMRON HEALTHCARE ASIA PACIFIC. **HEM-7156T**: automatic blood pressure monitor. [S. l.]: Omron Healthcare, 2026. Disponível em: https://www.omronhealthcare-ap.com/ap/product/1645-hem-7156t. Acesso em: 22 set. 2026.

OPENAI. **Function calling**. [S. l.]: OpenAI, 2026a. Disponível em: https://platform.openai.com/docs/guides/function-calling. Acesso em: 5 out. 2026.

OPENAI. **Realtime API**. [S. l.]: OpenAI, 2026b. Disponível em: https://platform.openai.com/docs/guides/realtime. Acesso em: 5 out. 2026.

OPENAI. **Your data**. [S. l.]: OpenAI, 2026c. Disponível em: https://developers.openai.com/api/docs/guides/your-data. Acesso em: 5 nov. 2026.

PARENTI, N. *et al.* A systematic review on the validity and reliability of an emergency department triage scale, the Manchester Triage System. **International Journal of Nursing Studies**, v. 51, n. 7, p. 1062-1069, 2014. DOI: 10.1016/j.ijnurstu.2014.01.013.

PASLI, S. *et al.* Assessing the precision of artificial intelligence in emergency department triage decisions: insights from a study with ChatGPT. **American Journal of Emergency Medicine**, v. 78, p. 170-175, 2024. DOI: 10.1016/j.ajem.2024.01.037.

ROYAL COLLEGE OF PHYSICIANS. **National Early Warning Score (NEWS) 2**: standardising the assessment of acute-illness severity in the NHS. London: RCP, 2017.

SAMSUNG ELECTRONICS. **Measure blood oxygen on Galaxy Watch**. [S. l.]: Samsung Developer, 2026a. Disponível em: https://developer.samsung.com/codelab/health/blood-oxygen.html. Acesso em: 21 set. 2026.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

SAMSUNG ELECTRONICS. **Measure skin temperature on Galaxy Watch**. [S. l.]: Samsung Developer, 2026b. Disponível em: https://developer.samsung.com/codelab/health/skin-temperature.html. Acesso em: 8 out. 2026.

SAMSUNG ELECTRONICS. **Samsung Health Sensor SDK**. [S. l.]: Samsung Developer, 2026c. Disponível em: https://developer.samsung.com/health/sensor/overview.html. Acesso em: 21 set. 2026.

SOUZA, C. C. *et al.* Classificação de risco em pronto-socorro: concordância entre um protocolo institucional brasileiro e Manchester. **Revista Latino-Americana de Enfermagem**, v. 19, n. 1, p. 26-33, 2011. DOI: 10.1590/S0104-11692011000100005.

SOUZA, C. C. de; ARAÚJO, F. A.; CHIANCA, T. C. M. Scientific literature on the reliability and validity of the Manchester Triage System (MTS) protocol: an integrative literature review. **Revista da Escola de Enfermagem da USP**, v. 49, n. 1, p. 144-151, 2015. DOI: 10.1590/S0080-623420150000100019.

TRIAGEM INTELIGENTE: triagem de sintomas com classificação de risco. [S. l.]: GitHub, 2026. Revisão fda2f22. Disponível em: https://github.com/Marcaltw20/triagem-inteligente. Acesso em: 17 nov. 2026.

WILLIAMS, C. Y. K. *et al.* Use of a large language model to assess clinical acuity of adults in the emergency department. **JAMA Network Open**, v. 7, n. 5, e248895, 2024. DOI: 10.1001/jamanetworkopen.2024.8895.

WU, W.; PENG, H.; YU, S. YuNet: a tiny millisecond-level face detector. **Machine Intelligence Research**, v. 20, n. 5, p. 656-665, 2023. DOI: 10.1007/s11633-023-1423-y.

ZHONG, Y. *et al.* SFace: sigmoid-constrained hypersphere loss for robust face recognition. **IEEE Transactions on Image Processing**, v. 30, p. 2587-2598, 2021. DOI: 10.1109/TIP.2020.3048632.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.