# ViSPAC: Priorização e compressão adaptativa guiadas por NEWS2 em um ciclo Edge–Fog–Cloud avaliado na AWS

**Mateus Roveda**[1], **Rodrigo da Rosa Righi**[1]

[1]Universidade do Vale do Rio dos Sinos (Unisinos)
Av. Unisinos – 93.022-750 – Sáo Leopoldo – RS – Brazil

mroveda@edu.unisinos.br, RRRIGHI@unisinos.br

***Abstract.*** *Remote monitoring of vital signs generates continuous data streams that frequently overload low-cost edge devices and networks. This paper presents ViSPAC, a closed-loop Edge–Fog–Cloud model that uses the National Early Warning Score 2 (NEWS2) to dynamically adapt sampling and compression parameters at the edge based on patient risk. Evaluated in a distributed Amazon Web Services (AWS) environment, ViSPAC reduced transmissions by 96.7%, achieved an 81.6% average compression rate, preserved clinical fidelity (global Percent Root-mean-square Difference – PRD 1.16%), and kept the feedback-loop latency around 1.05 s.*

***Resumo.*** *O monitoramento remoto de sinais vitais gera fluxos contínuos de dados que frequentemente sobrecarregam redes e dispositivos de borda. Este artigo apresenta o ViSPAC, um modelo em ciclo fechado Edge–Fog–Cloud que utiliza o National Early Warning Score 2 (NEWS2) para adaptar dinamicamente a amostragem e a compressão na borda, com base no risco clínico do paciente. Avaliado em um ambiente distribuído na Amazon Web Services (AWS), o ViSPAC demonstrou uma redução de 96,7% nas transmissões e alcançou 81,6% de compressão média. O modelo preservou a fidelidade clínica, mantendo o Percent Root-mean-square Difference (PRD) global em 1,16%, com uma latência no ciclo de controle de aproximadamente 1,05 s.*

## 1. Introdução

O avanço dos dispositivos vestíveis e da Internet das Coisas Médicas (IoMT) popularizou o monitoramento contínuo de sinais vitais fora do ambiente hospitalar. Essa prática, contudo, esbarra em desafios significativos: o tráfego massivo de dados, as restrições de processamento e energia nos dispositivos de borda (Edge) e a urgência de respostas rápidas frente à deterioração clínica do paciente [Damera et al. 2025, Ali et al. 2025]. Para equilibrar escalabilidade e tempo de resposta, arquiteturas distribuídas Edge–Fog–Cloud despontam como uma solução natural, transferindo a análise e a coordenação para camadas intermediárias (Fog) e serviços centrais (Cloud) [Bonomi et al. 2012].

Apesar desses avanços, a literatura recente aborda a compressão e a priorização de forma isolada. Trabalhos como o de Chang e Sobelman [Chang and Sobelman 2024] validam a eficácia da compressão híbrida em cenários médicos. Simultaneamente, abordagens recentes exploram a compressão diretamente na computação em névoa [Andrade et al. 2025]. Contudo, é raro encontrar sistemas em ciclo fechado. Falta uma

arquitetura onde a inteligência das camadas superiores reconfigure ativamente a coleta na borda, ajustando a amostragem e a compressão em tempo real.

Para suprir essa lacuna, propomos o **ViSPAC** (*Vital Sign Prioritization and Adaptive Compression*). Trata-se de um modelo Edge–Fog–Cloud operando em **ciclo fechado** e guiado por risco clínico. Nele, a camada Fog calcula a gravidade do quadro do paciente via NEWS2 [Royal College of Physicians 2017] e envia comandos de ajuste de volta à borda. Isso permite reduzir o tráfego de forma agressiva durante períodos de estabilidade clínica, garantindo ao mesmo tempo alta fidelidade assim que o paciente entra em estado crítico. A estratégia adota uma compressão sensível ao contexto, combinando métodos com e sem perdas em alinhamento com abordagens híbridas recentes [Andrade et al. 2025, Chang and Sobelman 2024].

As contribuições deste artigo englobam:

- a modelagem de um ciclo adaptativo de priorização e compressão guiado por NEWS2 com realimentação Fog/Cloud→Edge;
- a avaliação quantitativa desse sistema em um ambiente de nuvem multirregião (AWS), com foco em métricas de eficiência, fidelidade clínica e latência do ciclo de controle.

## 2. Fundamentação e Trabalhos Relacionados

### 2.1. Edge–Fog–Cloud e sistemas sensíveis a latência

A computação em névoa (*Fog computing*) é frequentemente adotada em sistemas de Internet das Coisas (IoT) para diminuir a latência e elevar a qualidade do serviço (QoS), aproximando o processamento da origem dos dados [Bonomi et al. 2012]. Na área da saúde, o grande obstáculo é balancear responsividade e consumo de recursos. Transmitir sinais vitais de forma ininterrupta consome muita banda e bateria; em contrapartida, coletas muito espaçadas podem atrasar a identificação de um quadro clínico grave [Damera et al. 2025, Ali et al. 2025].

### 2.2. NEWS2 como escore clínico

O *National Early Warning Score 2 (NEWS2)* é um protocolo amplamente utilizado para padronizar a avaliação de doenças agudas, definindo a intensidade da vigilância necessária [Royal College of Physicians 2017]. No ViSPAC, o escore funciona como o gatilho principal: riscos elevados exigem alta precisão e comunicação quase instantânea, enquanto quadros estáveis autorizam o relaxamento dos parâmetros.

### 2.3. Compressão para transmissores biomédicos e métricas de distorção

Métricas como a taxa de compressão e o PRD (ou MSE/PSNR) costumam guiar a avaliação de eficiência na IoT médica [Chang and Sobelman 2024, Hassan and Mohsen 2024]. Soluções mais recentes têm explorado modelos híbridos (mesclando compressão com e sem perdas) para diminuir o volume sem sacrificar a utilidade médica em trechos críticos [Andrade et al. 2025]. O diferencial do ViSPAC reside em integrar essa compressão híbrida a um ajuste dinâmico de coleta dentro de um ciclo guiado por risco.

## 3. Modelo ViSPAC

A arquitetura do sistema organiza-se em três frentes de atuação. Na **Edge**, um *gateway* captura os sinais vitais, comprime os dados e despacha os lotes. Na **Fog**, os pacotes são abertos para o cálculo do NEWS2, permitindo que um módulo de controle devolva instruções de ajuste à borda. Já na **Cloud**, ocorre a persistência dos dados para consultas históricas, análises e emissão de alertas a longo prazo.

Conforme as Figuras 1 e 2 ilustram, o sistema opera em um laço fechado contínuo: as decisões sobre priorização e nível de detalhamento dos dados são ajustadas em tempo de execução pela Fog e repassadas à borda.

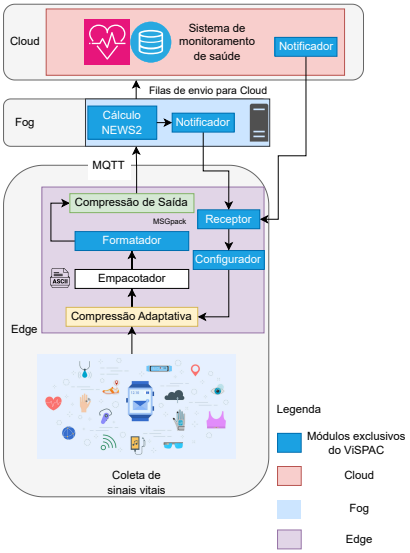


**Figura 1. Arquitetura macro do modelo ViSPAC.**

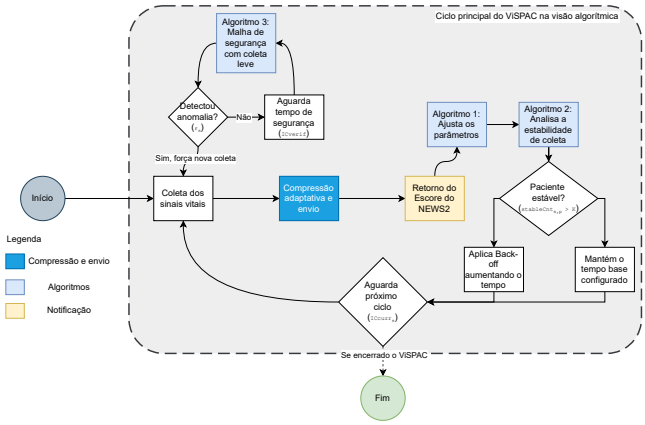


**Figura 2. Pipeline do ViSPAC em ciclo fechado em nível algorítmico.**

### 3.1. Pipeline de compressão híbrida

O processamento ocorre em dois passos na borda. Primeiro, aplica-se uma compressão *lossy* via *Swinging Door Trending (SDT)* [Bristol 1990], selecionada por sua complexidade linear que se adequa bem a hardwares restritos. Em seguida, entra em cena uma

compressão *lossless* (Huffman/LZW) para enxugar o pacote durante períodos sem risco clínico. Para garantir velocidade máxima em emergências, o sistema aciona uma regra de *bypass*: se o risco for ALTO, a etapa *lossless* é ignorada. Dispositivos de borda restritos adicionam latência ao executar algoritmos de entropia. Em situações críticas, o tempo gasto compactando o pacote excede o tempo de transmissão bruta na rede. Assim, ignorar a etapa de compressão minimiza o esforço computacional e reduz o tempo de resposta ponta a ponta.

### 3.2. Serialização binária e protocolo

Para enfrentar possíveis instabilidades de rede e oscilações de banda, o JSON tradicional foi substituído pelo MessagePack nas mensagens Edge→Fog [Friesel and Spinczyk 2021], enxugando o *payload* logo na origem. A comunicação principal utiliza o protocolo MQTT [Banks and Gupta 2014], que foi escolhido justamente por sua resiliência contra perdas de pacotes. Por fim, o tráfego em direção à Cloud segue via HTTP, o que facilita a integração com bancos de dados relacionais.

### 3.3. Ciclo de controle: parâmetros por risco e malha de segurança

A malha de realimentação depende de três mecanismos integrados: (i) reconfiguração imediata frente a mudanças de risco; (ii) *back-off* exponencial durante estabilidade prolongada; e (iii) vigilância contínua (*keep-alive*) para prevenir silêncios prolongados na rede.

**Intervalos base por risco.** Com base nas diretrizes do NEWS2 [Royal College of Physicians 2017], o risco ALTO exige um monitoramento quase contínuo (ex: 15s para Frequência Cardíaca), enquanto riscos menores permitem espaçar as medições (2, 5 ou 10 minutos).

**Back-off condicionado ao risco.** Se o sistema registrar $K$ leituras estáveis consecutivas, o intervalo dobra até atingir um teto $IC_{max}$ (30min para MODERADO, 2h para BAIXO e 6h para MÍNIMO). O parâmetro $T_{SDT}$ ajusta-se proporcionalmente para manter a coerência da redução de dados:

$$T_{SDT,s}^{curr} = \max\left(T_{SDT,s}^{base}, 2 \times IC_s^{curr}\right). \tag{1}$$

**Vigilância contínua (keep-alive).** Para evitar que uma piora súbita passe despercebida nos intervalos mais longos do *back-off*, executam-se checagens leves periódicas [Braden 1989]. Se houver uma alteração acima do limiar $\varepsilon_s$, uma coleta completa é disparada imediatamente.

## 4. Ambiente de Avaliação na AWS

A distribuição geográfica e as restrições de hardware afetam diretamente o desempenho de sistemas IoT em larga escala. Para refletir essa realidade e simular latências representativas de um ambiente de produção, implementamos uma topologia multirregião na AWS (Figura 3).

A arquitetura separou a Edge e a Fog na região us-east-1. A Cloud foi instanciada na região us-west-1. As camadas comunicam-se via *VPC Peering*. Isso mantém o tráfego protegido no *backbone* da AWS. Ao mesmo tempo, o modelo é submetido a latências inter-regionais naturais, superiores a 200,ms de atraso *round-trip*. Essa latência injetada é indispensável para testar o comportamento e a responsividade do ciclo de controle em condições realistas.

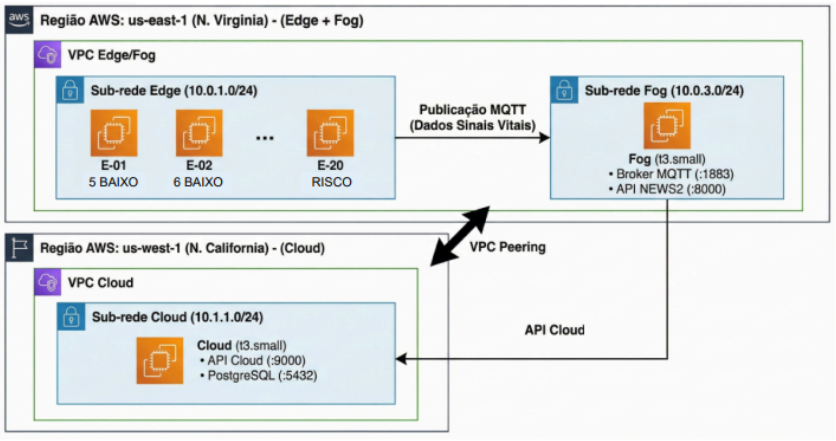


**Figura 3. Topologia experimental distribuída na AWS.**

### 4.1. Componentes, restrições e reprodutibilidade

Na borda, foram utilizadas 20 instâncias EC2 t2.small rigidamente limitadas via *cgroups* a um máximo de 256,MB de RAM e 50% de uso de uma vCPU. Essa restrição foi aplicada propositalmente para emular com alta fidelidade os gargalos físicos de *Single Board Computers* (SBC) de baixo custo, como o *Raspberry Pi Zero*, comprovando assim a viabilidade do *pipeline* de compressão mesmo em dispositivos embarcados sujeitos a severas limitações de hardware. Por sua vez, Fog e a Cloud rodaram em instâncias t3.small dedicadas. A Fog e a Cloud rodaram em instâncias t3.small dedicadas (2 vCPU, 2GB RAM), lidando respectivamente com o broker MQTT/cálculo do NEWS2 e com a persistência em PostgreSQL.

Toda a infraestrutura foi automatizada como código (Terraform), assegurando um provisionamento reprodutível dos 22 nós, incluindo a injeção determinística de sinais fisiológicos **reais** para cada paciente. Foram utilizados registros públicos validados do *BIDMC PPG and Respiration Dataset* e do *Biosensor Student Health Fitness Data*. O *payload* serializado via MessagePack priorizou as variações dinâmicas da Frequência Cardíaca (FC) e da Saturação de Oxigênio (SpO2), enquanto os demais parâmetros do escore NEWS2 foram mantidos em valores de referência estáveis para isolar o impacto da compressão. A distribuição dos dados ocorreu via *seeding* atrelado ao EDGE_ID.

## 5. Metodologia Experimental

### 5.1. Cenários de teste e janela temporal

Três cenários de teste foram definidos, executados por 12 horas cada, para capturar a variabilidade natural dos sinais sob diferentes cargas: **(1) Baseline**: transmissão contínua

e bruta com intervalo fixo de $IC = 1$s; **(2) Estático (VSAC)**: aplicação das compressões SDT e Huffman/LZW com parâmetros fixos e encaminhamento síncrono; **(3) ViSPAC**: adaptação dinâmica via NEWS2, encaminhamento assíncrono e filas por nível de risco na Fog.

### 5.2. Métricas

A taxa de compressão média (TC) comparou o volume inicial ($S_i$) e o final trafegado ($S_f$):

$$TC = \left(1 - \frac{S_f}{S_i}\right) \times 100\%. \tag{2}$$

A latência avaliou o atraso de resposta do laço de controle:

$$T_{loop} = t_{ajuste} - t_{coleta}. \tag{3}$$

A distorção do sinal foi calculada pelo PRD, no qual valores próximos de zero representam alta fidelidade do sinal reconstruído ($Y$) em relação ao original ($X$) [Chang and Sobelman 2024, Hassan and Mohsen 2024]:

$$PRD = \sqrt{\frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{\sum_{i=1}^{N} X_i^2}} \times 100. \tag{4}$$

## 6. Resultados e Discussão

A Tabela 1[1] resume o comportamento do sistema para 20 nós de borda em 12 horas. Ao reduzir o número absoluto de transmissões em 96,7%, o ViSPAC proporciona indiretamente uma drástica economia de bateria nos dispositivos de borda, visto que o uso contínuo do rádio-transmissor costuma ser o principal dreno energético em redes IoT médicas. No que tange à escalabilidade, a eficiência computacional do modelo assíncrono mostrou-se notável: a camada Fog consumiu menos de 0,1% de CPU para gerenciar os 20 nós simultâneos, o que permite estimar, por meio de projeções lineares, que um único nó Fog suportaria mais de 10.000 instâncias na borda sem apresentar gargalos de processamento. Cabe notar que o volume bruto registrado nos três cenários é praticamente o mesmo (dados estruturados em MessagePack), evidenciando que a economia gerada provém estritamente das decisões inteligentes de corte e compressão da arquitetura.

### 6.1. Fidelidade clínica por nível de risco

Conforme a Tabela 2, o PRD atinge valores mais altos nas faixas de baixo risco (8,25% a 9,86%) em decorrência da compressão mais agressiva aplicada durante longos platôs fisiológicos. Do ponto de vista clínico, as pequenas variações descartadas nesse período representam ruídos que não alteram o diagnóstico de um paciente estável. Assim que o risco evolui para MODERADO ou ALTO, a distorção cai drasticamente (1,27% a 0,56%), mantendo-se em patamares reconhecidamente seguros [Hassan and Mohsen 2024]. É importante destacar que esse comportamento foi validado qualitativamente por uma especialista com doutorado em Enfermagem, confirmando que o descarte agressivo em fases estáveis não impõe riscos ao paciente, desde que os mecanismos de *keep-alive* garantam uma resposta imediata a qualquer deterioração súbita.

[1] `https://roveda.dev/dissertacao/results/dashboard.html`

**Tabela 1. Métricas globais (20 nós edge, 12 horas).**

| Métrica | Baseline | Estático | ViSPAC |
|---|---|---|---|
| Transmissões totais (pacotes) | 6.642.048 | 564.237 | 220.403 |
| Volume de dados brutos (MB) | 245,11 | 246,96 | 245,01 |
| Volume trafegado final (MB) | 245,11 | 61,65 | 45,21 |
| Taxa de compressão média (%) | 0,00 | $75{,}0 \pm 10{,}6$ | $81{,}6 \pm 9{,}7$ |
| PRD médio global (%) | 0,00 | $3{,}62 \pm 4{,}94$ | $1{,}16 \pm 2{,}67$ |
| Latência média (ms) | $1.358{,}3 \pm 249{,}5$ | $1.271{,}2 \pm 149{,}0$ | $1.048{,}9 \pm 3{,}8$ |

**Tabela 2. PRD por nível de risco clínico.**

| Risco | Estático | | ViSPAC | |
|---|---|---|---|---|
| | PRD (%) | Amostras | PRD (%) | Amostras |
| Alto | 0,50 | 207.742 | 0,56 | 200.503 |
| Moderado | 0,70 | 40.396 | 1,27 | 5.429 |
| Baixo | 5,59 | 85.310 | 8,25 | 5.401 |
| Mínimo | 6,21 | 230.789 | 9,86 | 9.070 |
| Média global | 3,62 | 564.237 | 1,16 | 220.403 |

### 6.2. Latência e efeito do encaminhamento assíncrono

A decomposição do tempo de resposta (Tabela 3) revela o grande trunfo do ViSPAC: o uso de filas assíncronas. Ao desacoplar a resposta imediata da Fog em relação ao tempo de confirmação do banco de dados na Cloud, o sistema retira a latência inter-regional do caminho crítico, assegurando reconfigurações rápidas na borda.

**Tabela 3. Decomposição da latência média por componente e cenário.**

| Componente | Baseline | Estático | ViSPAC |
|---|---|---|---|
| Processamento NEWS2 (Fog) | 2,31 ms | 1,65 ms | 0,88 ms |
| Encaminhamento Fog–Cloud | 230,06 ms | 207,73 ms | assíncrono |
| Rede + processamento Edge | 1.125,9 ms | 1.061,8 ms | 1.048,0 ms |
| Latência total percebida (Edge) | 1.358,3 ms | 1.271,2 ms | 1.048,9 ms |

## 7. Conclusão

Este artigo demonstrou que atrelar escores de risco clínico às políticas de coleta em camadas de borda e névoa é um caminho altamente eficaz para otimizar recursos em IoT médica. O ViSPAC, atuando em ciclo fechado, entregou compressão severa (81,6%), redução massiva no uso da rede (96,7%) e tempos de adaptação curtos ($\sim$1,05 s), tudo isso mantendo a distorção do sinal em níveis seguros (PRD de 1,16%).

Diferente das abordagens isoladas, o ViSPAC enxerga o paciente e a infraestrutura de forma integrada, suportado por uma topologia multirregião na AWS. Passos futuros incluem a validação em hardwares embarcados físicos rodando sob redes com perda de

pacotes, a inclusão de novos sinais vitais e a avaliação clínica formal dos sinais reconstruídos por médicos especialistas.

## 8. Declaração sobre uso de Inteligência Artificial

Em conformidade com o Código de Conduta para autores da SBC, os autores declaram o uso de ferramentas de Inteligência Artificial (IA) generativa como apoio à escrita. Em particular, modelos de IA generativa foram utilizados para (i) revisão ortográfica e sugestões de melhoria de fluidez/clareza no manuscrito; e (ii) apoio na organização e formatação do texto estrutural em LaTeX. As ferramentas não foram utilizadas para gerar dados experimentais, executar simulações, criar figuras ou substituir a verificação de referências. Os autores revisaram criticamente todo o conteúdo e assumem total responsabilidade pelo artigo.

## Referências

Ali, A., Montanaro, T., Sergi, I., Carrisi, S., Galli, D., Distante, C., and Patrono, L. (2025). An innovative iot and edge intelligence framework for monitoring elderly people using anomaly detection on data from non-wearable sensors. *Sensors*, 25(6).

Andrade, A., da Costa, C. A., Roehrs, A., Muchaluat-Saade, D., and da Rosa Righi, R. (2025). Blending lossy and lossless data compression methods to support health data streaming in smart cities. *Future Generation Computer Systems*, 167:107748.

Banks, A. and Gupta, R. (2014). Mqtt version 3.1.1. Technical report, OASIS. OASIS Standard.

Bonomi, F., Milito, R., Zhu, J., and Addepalli, S. (2012). Fog computing and its role in the internet of things. In *First Edition of the MCC Workshop on Mobile Cloud Computing*, pages 13–16. Association for Computing Machinery.

Braden, R. T. (1989). Requirements for internet hosts - communication layers. Technical Report 1122, RFC Editor. Request for Comments.

Bristol, E. H. (1990). Swinging door trending: adaptive trend recording. In *ISA National Conference*, pages 749–753.

Chang, Y. and Sobelman, G. E. (2024). Lightweight lossy/lossless ecg compression for medical iot systems. *IEEE Internet of Things Journal*, 11(7):12450–12458.

Damera, V. K., Cheripelli, R., Putta, N., Sirisha, G., and Kalavala, D. (2025). Enhancing remote patient monitoring with ai-driven iomt and cloud computing technologies. *Scientific Reports*, 15(1):24088.

Friesel, D. and Spinczyk, O. (2021). Data serialization formats for the internet of things. *Electronic Communications of the EASST*, 80.

Hassan, A. M. A. and Mohsen, S. (2024). Compression of electrocardiogram signals using compressive sensing technique based on curvelet transform toward medical applications. *Multimedia Tools and Applications*, 84(12):11203–11219.

Royal College of Physicians (2017). *National Early Warning Score (NEWS) 2: standardising the assessment of acute-illness severity in the NHS*. Royal College of Physicians, London. Updated report of a working party.