Quantização Guiada pela Lei de Benford: Compressão Log-Uniforme de Pesos para Modelos de Visão Médica Eficientes

Arthur Negrão[1], Guilherme Silva[1], Matheus Vieira[1], Ederson N. F. G. Júnior[1], Eduardo José da Silva Luz[2], Pedro Silva[2]

[1] Programa de Pós-Graduação em Ciência da Computação
Universidade Federal de Ouro Preto (UFOP) – 35400-000 – Ouro Preto – MG – Brasil
arthur.negrao@aluno.ufop.edu.br

[2]Departamento de Computação – UFOP – 35400-000 – Ouro Preto – MG – Brasil
silvap@ufop.edu.br

Abstract. *Benford-Quant, a post-training quantization method guided by Benford's Law, is evaluated for medical image classification. The study considers ResNet-18, ResNet-50, and ViT-Base on the BloodMNIST, PathMNIST, and OrganCMNIST datasets, with comparisons against RTN and NF4. Results show that BenQ preserves competitive F1-score while achieving up to 7.7× memory reduction, indicating that its benefits extend beyond LLMs. This behavior persists even under challenging conditions such as Gaussian noise and contrast variations, where non-uniform quantization methods tend to exhibit greater robustness. These findings position BenQ as a viable alternative for medical vision models in resource-constrained environments.*

Resumo. *Benford-Quant, um método de quantização pós-treinamento guiado pela Lei de Benford, é avaliado na classificação de imagens médicas. O estudo considera ResNet-18, ResNet-50 e ViT-Base nos conjuntos BloodMNIST, PathMNIST e OrganCMNIST, com comparação contra RTN e NF4. Os resultados mostram que o BenQ preserva desempenho competitivo em F1-Score, com redução de memória de até 7,7x, evidenciando que seus benefícios não se restringem a LLMs, mesmo em cenários desfavoráveis como ruído gaussiano e variações de contraste, onde métodos não uniformes tendem a maior robustez. Esses achados posicionam o BenQ como uma alternativa viável para visão médica em ambientes com recursos limitados.*

1. Introdução

A utilização de Inteligência Artificial (IA) em aplicações médicas tem crescido significativamente na literatura recente. Embora tais sistemas não substituam a *expertise* humana em processos clínicos e levantem importantes discussões éticas, há amplo reconhecimento de seu potencial como ferramentas de apoio à prática médica [Zhai et al. 2022]. Nesse contexto, métodos baseados em aprendizado profundo têm sido aplicados em diversas tarefas, como detecção de doenças, análise de exames laboratoriais e processamento de imagens médicas [Shandhi and Dunn 2022, Manduva 2024]. Entretanto, a crescente sofisticação desses modelos impõe desafios práticos relevantes, especialmente relacionados ao elevado consumo de memória e poder computacional.

Esse problema torna-se particularmente crítico em cenários nos quais o processamento precisa ocorrer localmente, como em dispositivos móveis, sistemas embarcados ou aplicações de *edge computing* utilizadas para coleta e análise de dados médicos [Zhao 2023, Lu et al. 2021]. Modelos modernos frequentemente demandam recursos muito superiores aos disponíveis nesses ambientes. Por exemplo, arquiteturas de visão amplamente utilizadas, como o *Google's ViT-Base-Patch-16-224*[1], demandam aproximadamente 346 *megabytes* (MB) de armazenamento, enquanto plataformas embarcadas típicas operam com ordens de grandeza inferiores de memória, como no caso da STM32 NUCLEO-L432KC[1] e do Arduino Nano 33 BLE[1]. Esse descompasso evidencia a necessidade de técnicas capazes de reduzir o custo computacional de modelos de IA.

Entre as abordagens mais utilizadas para esse fim está a quantização de redes neurais, que consiste na redução da precisão numérica utilizada para representar os parâmetros do modelo [Nagel et al. 2021]. Por exemplo, uma matriz de pesos originalmente representada em precisão de 32 bits pode ser convertida para representações de 4 bits, resultando em modelos significativamente mais compactos. Essa redução pode ser realizada durante o treinamento, no processo conhecido como *Quantization Aware Training* (QAT), ou após o treinamento, por meio de *Post Training Quantization* (PTQ). Considerando a maior simplicidade de aplicação e o menor custo computacional, abordagens baseadas em PTQ tornam-se particularmente atrativas em cenários com restrições de recursos [Marchisio et al. 2024]. No entanto, a eficácia de métodos de quantização pode variar significativamente conforme a arquitetura, a tarefa e o domínio de aplicação [Mekala et al. 2025].

Nesse contexto, este trabalho investiga a aplicabilidade do método Benford-Quant (BenQ) [Negrão et al. 2026] em tarefas de classificação de imagens médicas sob restrições de recursos computacionais com o objetivo de desenvolver um benchmark sistemático para analisar a robustez em imagens médicas de métodos de quantização. Originalmente proposto no contexto de *Large Language Models* (LLMs), o método utiliza propriedades estatísticas derivadas da Lei de Benford [Benford 1938] para caracterizar a distribuição dos pesos de redes neurais. A partir dessa propriedade, o método adota duas estratégias principais: **(1)** quantização seletiva de camadas que apresentam aderência à Lei de Benford e **(2)** utilização de um *grid* de quantização log-uniforme, motivado pela relação entre benford-aderência e distribuições logarítmicas da mantissa [Hill 1995]. Este trabalho investiga se tais propriedades também se mantêm eficazes no contexto de modelos de visão aplicados à análise de imagens médicas.

No intuito de nortear o presente trabalho, estabelecem-se as seguintes Perguntas de Pesquisa (PP): **PP1.** Benford-Quant é capaz de gerar modelos quantizados eficientes na tarefa de classificação de imagens médicas, sendo a eficiência mensurada através de métricas como acurácia ou F1-*Score*? Ou sua eficiência está restrita ao seu contexto original de LLMs?; **PP2.** Como o *grid* log-uniforme se desempenha em relação ao *grid* uniforme e a outros *grids* não-uniformes (*ex. grid* normal) em tarefas de classificação de imagens médicas?; **PP3.** Benford-Quant é capaz de gerar modelos quantizados mais resistentes a ruídos e variações típicas do contexto médico, mais especificamente ruído gaussiano e variações do contraste, que outros métodos de quantização?

As principais contribuições deste trabalho são três: **(1)** uma avaliação empírica

[1]Dados sobre o modelo foram extraídos da plataforma *Hugging Face*. Dados sobre as placas embarcadas foram extraídos das documentações fornecidas pelos fabricantes.

do método Benford-Quant em tarefas de classificação de imagens médicas utilizando arquiteturas amplamente empregadas em visão computacional (ResNet-18, ResNet-50 e *Google's ViT Base Patch16*); **(2)** uma análise comparativa entre o *grid* log-uniforme utilizado pelo método e outros esquemas de quantização, incluindo *grid* uniforme e *grid* normal; e **(3)** uma investigação da robustez de modelos quantizados frente a perturbações comuns em imagens médicas, como ruído gaussiano e variações de contraste. Para essa avaliação, utilizamos o *dataset* MedMNIST e comparamos o Benford-Quant com dois *baselines*: *NormalFloat* 4-bits (NF4) e *Round-to-Nearest* (RTN). Os resultados indicam que o método apresenta desempenho competitivo em relação aos baselines, inclusive na presença de ruído, embora nenhum método tenha se mostrado consistentemente superior em todos os cenários avaliados.

2. Trabalhos Relacionados

A Lei de Benford (LB) [Benford 1938] descreve a distribuição do Primeiro Dígito Significativo (PDS) (o primeiro dígito diferente de zero em um número quando lido da esquerda para a direita) em diversos conjuntos numéricos. Diferentemente de uma distribuição uniforme, a ocorrência dos dígitos segue uma distribuição logarítmica. Formalmente, para um dígito $d \in \{1, \ldots, 9\}$, a probabilidade de ocorrência do PDS é dada por $P(d) = \log_{10}\left(1 + \frac{1}{d}\right)$. Por exemplo, a probabilidade de $d = 1$ ocorrer como primeiro dígito é aproximadamente 30,1%.

No contexto de Inteligência Artificial, a LB tem sido explorada com diferentes objetivos. Em [Sahu et al. 2021], os autores investigam sua relação com a capacidade de generalização de modelos, propondo métricas baseadas na aderência à distribuição de Benford como indicador de generalização e como critério de parada durante o treinamento. De forma complementar, [Ott et al. 2025] exploram a LB como mecanismo de regularização, demonstrando que sua incorporação pode melhorar o desempenho de CNNs em tarefas de classificação de imagens.

Mais recentemente, [Negrão et al. 2026] propõem o método BenQ, uma abordagem de quantização baseada na LB. A ideia central consiste em explorar propriedades estatísticas da distribuição do primeiro dígito significativo para orientar o processo de quantização, buscando reduzir a perda de informação associada à compressão numérica dos parâmetros do modelo. Detalhes são apresentados na Seção 3.4.

A quantização é amplamente empregada para reduzir o custo computacional de modelos de *deep learning*, permitindo sua execução em ambientes com recursos limitados. No domínio da saúde digital, essa técnica é particularmente relevante para aplicações em dispositivos embarcados e sistemas de monitoramento contínuo. Nesse contexto, [Xi et al. 2025] destacam o papel da quantização de CNNs na redução de consumo de memória, latência e gasto energético em aplicações médicas baseadas em computação de borda.

Especificamente no processamento de imagens médicas, estudos têm demonstrado que a quantização pode melhorar a eficiência computacional com impacto limitado no desempenho. Em [Xu et al. 2018], a quantização aplicada a CNNs para segmentação de imagens médicas resultou em reduções de até 6,4 vezes no uso de memória, acompanhadas de melhorias modestas de desempenho. De forma semelhante, [Abid et al. 2021] investigam a quantização na classificação de radiografias de tórax, reportando reduções de

até 57% no tempo de inferência em arquiteturas ARM e diminuições de 2 a 4 vezes no consumo de memória, com impacto marginal na métrica AUC-ROC.

Apesar desses avanços, abordagens de quantização baseadas em propriedades estatísticas como a Lei de Benford ainda permanecem pouco exploradas no contexto de imagens médicas. Até onde é de nosso conhecimento, métodos de quantização baseados na Lei de Benford, como o BenQ, ainda não foram avaliados sistematicamente em tarefas de classificação de imagens médicas. Nesse sentido, este trabalho investiga empiricamente a aplicação do método BenQ em tarefas de classificação de imagens médicas bidimensionais, avaliando seu impacto em desempenho e eficiência computacional.

3. Materiais e Métodos

3.1. Base de Dados

Para avaliar os métodos de quantização em tarefas de classificação de imagens médicas, utilizou-se a base MedMNIST [Yang et al. 2023], considerando três *subsets* 2D: BloodMNIST, PathMNIST e OrganCMNIST. Todas as imagens foram utilizadas com resolução de 224×224 *pixels*. A Figura 1 apresenta exemplos dos conjuntos selecionados.

Figura 1. Conjuntos de dados escolhidos da MedMNIST. Para cada um é apresentada a quantidade de instâncias dos conjuntos de *treino* / *validação* / *teste* e a quantidade de (*classes*) do mesmo. Fonte: adaptado de [Yang et al. 2023].

Esses conjuntos foram escolhidos por representarem diferentes modalidades e escalas de informação visual em imagens médicas. BloodMNIST e PathMNIST são compostos por imagens microscópicas (de células sanguíneas e tecidos histopatológicos do cólon, respectivamente) nas quais a discriminação entre classes depende principalmente de padrões celulares e texturas finas. Em contraste, OrganCMNIST contém imagens de tomografia computadorizada da região abdominal, caracterizadas por estruturas anatômicas de maior escala e padrões espaciais mais amplos. Essa diversidade permite avaliar o comportamento dos métodos investigados em cenários com diferentes níveis de granularidade visual, o que é particularmente relevante para analisar o impacto da quantização em tarefas de classificação médica. Os conjuntos de dados foram utilizados tanto para o ajuste fino dos modelos quanto para a avaliação dos modelos quantizados.

Além disso, em relação à PP3, foram consideradas duas perturbações nas imagens: ruído gaussiano e variações de contraste. Essas transformações simulam artefatos frequentemente observados em dados de *medical imaging* [Kusk and Lysdahlgaard 2023] e permitem investigar a robustez dos métodos de quantização frente a variações nas condições de aquisição e qualidade das imagens. Exemplos dessas perturbações são apresentados na Figura 2.

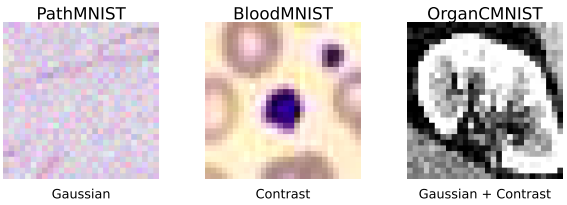

Figura 2. *Subsets* com aplicação de ruído gaussiano, variação do contraste e ambos, respectivamente. Fonte: adaptado de [Yang et al. 2023].

3.2. Modelos Utilizados

Para este trabalho, foram selecionadas arquiteturas de duas naturezas diferentes: Redes Neurais Convolucionais (ResNet-18/RN18, a ResNet-50/RN50 [He et al. 2015]) e *Vision Transformers* (Google's ViT Base Patch16-224/ViT [Wu et al. 2020]). O objetivo é observar os efeitos de BenQ sobre cada uma delas, à medida que o trabalho original tem enfoque sobre arquiteturas de LLMs. Os modelos são ilustrados na Figura 3.

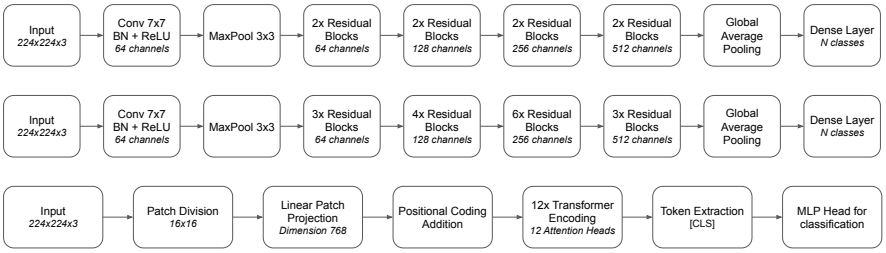

Figura 3. Arquiteturas dos modelos utilizados no estudo. A primeira é a ResNet-18, a segunda, ResNet-50 e a última, Google's ViT Base Patch16-224. Fonte: autoria própria (2026).

Antes da quantização, os modelos passaram por um processo de *finetuning* em cada um dos conjuntos de dados de treinamento. Para o *finetuning* utilizou-se 5 épocas, *learning rate* de 3e-4, e um *early stopping* com 2 épocas de paciência considerando a *loss* de validação. Todos os métodos de quantização partiram do mesmo modelo treinado. Logo, todas as diferenças observadas nos resultados são integralmente oriundas do algoritmo quantizador.

3.3. Processo de Avaliação

A avaliação dos modelos quantizados foi conduzida com o conjunto de teste de cada subconjunto de dados. A métrica escolhida para a análise foi o F1-Score. Tendo em vista que, TP representa verdadeiros positivos, TN representa verdadeiros negativos, FP representa falsos positivos e FN representa falsos negativos, tem-se: a **precisão** mede a proporção de verdadeiros positivos entre todas as predições positivas, sendo definida como $PRE = \frac{TP}{TP+FP}$; o ***recall*** mede a proporção de verdadeiros positivos entre todos os positivos reais, sendo definido como $REC = \frac{TP}{TP+FN}$; e, por fim, o **F1-Score** é a média harmônica entre precisão e *recall*, isto é $F1 = 2 \cdot \frac{PRE \cdot REC}{PRE+REC}$. O tempo gasto para a execução do algoritmo de quantização é utilizado como análise de tempo computacional e o espaço gasto (em MB) antes e depois da quantização como análise de memória.

Para a análise da PP2, estabeleceram-se dois *baselines* comparativos: o *Uniform Round-to-Nearest* (RTN) e o 4-bit *NormalFloat* (NF4) [Dettmers et al. 2023]. O primeiro apresenta *grid* de quantização uniforme, enquanto o segundo apresenta *grid* não-uniforme, mais especificamente *grid* normal. Ambos *baselines* adotaram a mesma quantização seletiva (melhor detalhada na Seção 3.4) presente em BenQ, assegurando que as diferenças observadas fundamentem-se essencialmente na diferença entre os *grids*.

3.4. Benford-Quant

O BenQ [Negrão et al. 2026] analisa a relação entre pesos de redes neurais e a Lei de Benford (LB). Construções teóricas e experimentos práticos apresentados no artigo revelaram que, no contexto de redes neurais, camadas baseadas em transformações lineares (por exemplo, camadas de atenção e MLP) tendem a seguir distribuições benfordianas. Em contraste, camadas de normalização e *embeddings* não exibem esse comportamento.

Essa observação fundamenta duas políticas práticas que compõem o BenQ. A primeira consiste na quantização seletiva de camadas: apenas camadas cuja distribuição de pesos apresenta aderência à Lei de Benford são quantizadas utilizando o esquema proposto, enquanto camadas de normalização e *embeddings* permanecem na precisão original. Essa decisão é motivada por três fatores principais: (i) essas camadas não apresentam distribuição benfordiana, (ii) representam uma fração relativamente pequena da memória total do modelo e (iii) desempenham um papel crítico na estabilidade das ativações. Assim, mantê-las em ponto flutuante tende a preservar a estabilidade do modelo sem impacto significativo no consumo de memória.

A segunda política deriva de uma propriedade fundamental da LB. Uma variável aleatória X segue a LB se, e somente se, a mantissa de X possui distribuição logarítmica [Hill 1995]. Considerando ainda que pesos de redes neurais tipicamente apresentam forte concentração em torno de zero, essa propriedade sugere um *proxy* prático para o projeto do quantizador: utilizar níveis de quantização log-uniformes. Intuitivamente, esse tipo de discretização tende a alocar maior resolução para valores de menor magnitude, potencialmente reduzindo o erro de quantização.

Formalmente, para uma quantização com largura de B bits, o BenQ define um conjunto de níveis composto por 2^{B-1} valores positivos, $2^{B-1}-1$ valores negativos e um nível exatamente igual a zero. O *grid* de quantização, de caráter log-uniforme, pode então ser descrito como: $\mathcal{L} = (-\mathcal{L}_*^+) \cup \{0\} \cup \mathcal{L}^+$, onde $\mathcal{L}^+ = \left\{ \exp\Big(\log(\epsilon) + i \cdot \frac{\log(1)-\log(\epsilon)}{(2^{B-1}-1)-1} \Big) \;\middle|\; i = 0, 1, \ldots, (2^{B-1}-1)-1 \right\}$, $\mathcal{L}_*^+$ corresponde a $\mathcal{L}^+$ sem o seu nível cujo valor é o mais próximo de 0, haja vista que ele será substituído por tal, e ϵ é uma pequena constante de estabilidade numérica.

Dada uma matriz de pesos W, o BenQ aplica quantização *group-wise*. A matriz é dividida em grupos w_g de tamanho G, e para cada grupo é calculada uma escala $s_g = max(|w_g|)$, utilizado para normalizar os valores do grupo. Após a normalização, cada elemento é mapeado para o índice mais próximo no *grid* $\mathcal{L}$.

O processo de dequantização realiza a operação inversa. Dada a matriz quantizada W_q e o conjunto de escalas S, a matriz reconstruída $\tilde{W}$ é composta por blocos $\tilde{w}_g$ definidos por $\tilde{w}_g = \mathcal{L}[\mathbf{i}_g] \cdot s_g$. Nesta expressão, $\mathcal{L}[\mathbf{i}_g]$ representa a operação de busca elemento a elemento no *grid* $\mathcal{L}$ utilizando os índices quantizados do grupo g, enquanto s_g corresponde ao fator de escala associado a esse grupo.

Salienta-se que o BenQ é utilizado como um método de quantização pós-treinamento, avaliando se as propriedades observadas no trabalho original também se mantêm no contexto de modelos de classificação de imagens médicas. Por fim, a Figura 4 ilustra o fluxo completo do método BenQ. O código fonte do método encontra-se disponível na plataforma Github[2].

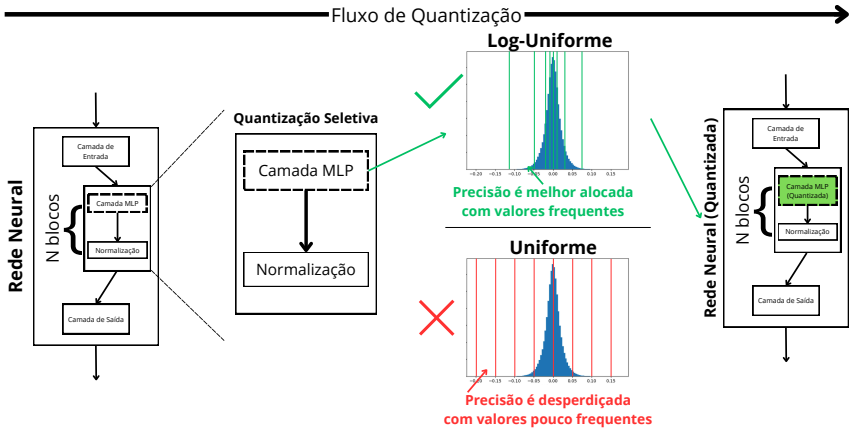

Figura 4. BenQ seleciona as camadas baseadas em transformações lineares para a quantização, mantendo inalteradas as outras. Fonte: autoria própria (2026).

4. Resultados e Discussão

Os modelos avaliados foram quantizados em regime de 4 bits, com *group size* igual a 128. A Figura 5 apresenta os resultados obtidos, evidenciando a diferença (em termos de F1-Score) entre o modelo quantizado e a versão de precisão original (FP16).

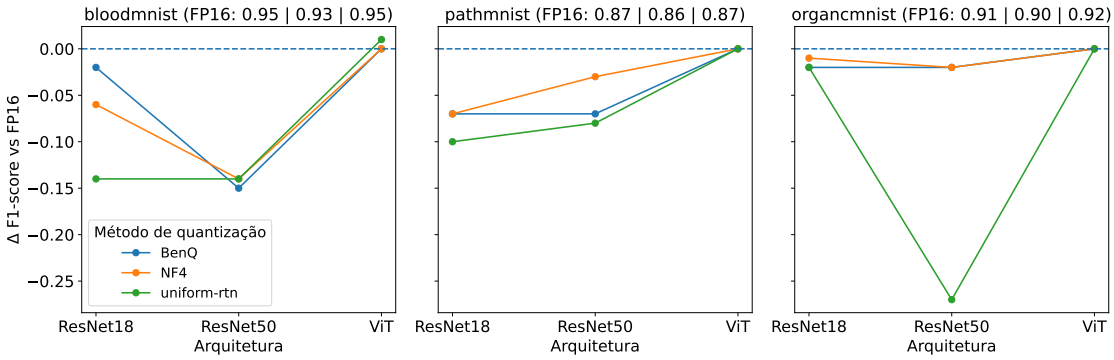

Figura 5. Variação do F1-score (△) dos métodos de quantização em relação aos diferentes modelos em precisão original (FP16). Os valores obtidos em FP16 são apresentados no título de cada subgráfico, respectivamente para RN18, RN50 e ViT. Fonte: autoria própria (2026).

Considerando inicialmente a PP1, os resultados indicam que o método BenQ não está restrito ao contexto de LLMs, apresentando desempenho competitivo na tarefa de classificação de imagens médicas. Como mostrado na Figura 5, em alguns cenários, BenQ

[2]https://github.com/arthurnfmc/med-sbcas

superou os métodos de referência. No *subset* BloodMNIST com a arquitetura RN18, por exemplo, BenQ superou NF4 e RTN em aproximadamente 4 e 12 pontos percentuais de F1-Score, respectivamente.

Entretanto, os experimentos também mostram que BenQ não é consistentemente superior em todos os cenários. Em BloodMNIST com RN50, RTN superou BenQ em cerca de 1%, enquanto em PathMNIST com RN50 o método NF4 apresentou vantagem de aproximadamente 4%. Esses resultados reforçam observações da literatura de que a escolha do método de quantização depende de fatores como a arquitetura do modelo, a tarefa considerada e as restrições computacionais [Mekala et al. 2025]. Assim, a seleção do quantizador deve ser realizada com base na análise empírica do cenário de aplicação, pois pode gerar resultados como o RTN aplicado a RN50 na OrganCMNIST.

Ao analisar especificamente a arquitetura baseada em *Vision Transformers*, observa-se que as reduções de desempenho em relação ao modelo em FP16 foram, em geral, menores quando comparadas às CNNs. Uma possível explicação está na forma como mecanismos de *self-attention* agregam informações globalmente, tornando as representações menos sensíveis ao ruído introduzido pela quantização.

Adicionalmente, observa-se que no *subset* BloodMNIST o método RTN superou o F1-Score do modelo em FP16. Esse comportamento está alinhado com observações anteriores de que o ruído introduzido pela quantização pode atuar como um mecanismo de regularização, potencialmente melhorando a generalização do modelo [Askari Hemmat et al. 2022].

Em relação à PP2, os resultados indicam que nenhum dos *grids* de quantização não-uniformes avaliados se mostrou consistentemente superior ao outro. Existem cenários nos quais BenQ apresenta melhor desempenho que NF4 e outros nos quais ocorre o inverso, reforçando a necessidade de avaliação empírica na escolha do esquema de quantização.

No que se refere ao consumo de memória, todos os métodos apresentaram reduções semelhantes devido à política de quantização seletiva adotada. Especificamente, a arquitetura RN18 apresentou redução de 7,31 vezes (42,61MB→5,82MB), a RN50 reduziu 7,75 vezes (89,54MB→11,54MB) e o modelo ViT reduziu 7,11 vezes (326,26MB→45,88MB). A Tabela 1 apresenta o tempo necessário somente para o processo de quantização, onde são desconsiderados o tempo gasto para carregar os modelos e o tempo gasto na inferência em momento de teste.

Método	RN18	RN50	ViT
BenQ	69ms	157ms	250ms
NF4	69ms	154ms	249ms
RTN	58ms	113ms	161ms

Tabela 1. Tempo gasto na quantização, em milissegundos. Fonte: autoria própria (2026).

Os resultados indicam que RTN apresentou o menor tempo de quantização, enquanto BenQ e NF4 apresentaram tempos semelhantes e ligeiramente superiores. Observa-se também uma tendência de aumento aproximadamente linear do tempo de quantização em função do número de parâmetros do modelo.

Essa característica torna os métodos avaliados particularmente adequados para cenários com restrições computacionais. Para efeito de comparação, o método GPTQ (amplamente utilizado na quantização de LLMs) apresenta tempos significativamente maiores [Frantar et al. 2022], onde os autores reportam aproximadamente 2,9 minutos para quantizar um modelo com 1,7 bilhões de parâmetros. Considerando que o ViT utilizado neste trabalho possui 86,6 milhões de parâmetros, uma extrapolação proporcional indicaria cerca de 4,9 segundos para BenQ. Assim, métodos de baixa complexidade, como os avaliados neste estudo, mostram-se particularmente atrativos para cenários de *edge computing*, por exemplo.

4.1. Imagens Ruidosas

Para investigar a PP3, foi conduzido um estudo de ablação avaliando o desempenho dos métodos de quantização sob diferentes perturbações na entrada: ruído gaussiano, variação de contraste e a combinação de ambos. A Figura 6 apresenta a variação do F1-Score em relação ao desempenho obtido pelos modelos em FP16 sem ruído.

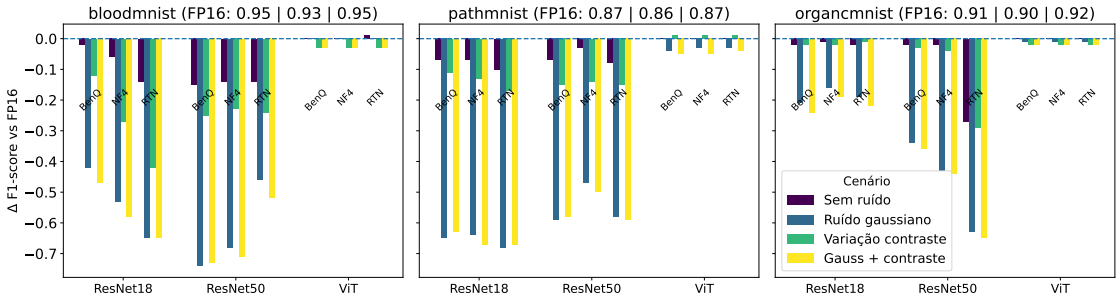

Figura 6. Variação do F1-Score dos métodos de quantização em relação ao modelo FP16 sem ruído para diferentes arquiteturas, *subsets* e tipos de perturbação. Fonte: autoria própria (2026).

Em alguns cenários, BenQ apresentou maior robustez à presença de ruído, como na RN18 em BloodMNIST e na RN50 em OrganCMNIST, onde superou NF4 e RTN em todas as condições avaliadas. Entretanto, de forma consistente com as análises das PP1 e PP2, nenhum quantizador demonstrou superioridade sistemática em todos os cenários.

Ao restringir a análise às CNNs, observa-se que métodos de quantização não-uniforme (*i.e.* BenQ e NF4) tendem a apresentar maior resistência a perturbações. Por exemplo, na RN50 com RTN em OrganCMNIST, a combinação de ruído gaussiano e variação de contraste resultou em uma redução de aproximadamente 65% no F1-Score em relação ao modelo FP16 sem ruído. Para BenQ e NF4, essas reduções foram de 36% e 44%, respectivamente. Esse comportamento pode ser explicado pelo fato de que esquemas não-uniformes alocam maior precisão nas regiões mais densas da distribuição dos pesos, reduzindo o erro médio de quantização.

Para analisar o impacto dos ruídos independentemente da quantização, a Tabela 2 apresenta o desempenho dos modelos em FP16 sob cada condição de perturbação.

Observa-se que a variação de contraste provoca degradações menores do que o ruído gaussiano nas CNNs, pois preserva a estrutura espacial da imagem enquanto altera apenas a escala das intensidades. Em contraste, o ruído gaussiano introduz perturbações

Tipo de ruído	BloodMNIST			PathMNIST			OrganCMNIST		
	RN18	**RN50**	**ViT**	**RN18**	**RN50**	**ViT**	**RN18**	**RN50**	**ViT**
Sem Ruído	0,95	0,93	0,95	0,87	0,86	0,87	0,91	0,90	0,92
Contraste	0,84	0,84	0,92	0,78	0,76	0,88	0,91	0,88	0,91
Ruído Gaussiano	0,54	0,39	0,95	0,21	0,33	0,84	0,76	0,51	0,91
Gauss. + Contraste	0,47	0,32	0,92	0,15	0,32	0,83	0,74	0,49	0,91

Tabela 2. F1-Score (FP16) para os *subsets* Blood, Path e OrganCMNIST sob diferentes condições de ruído. Fonte: autoria própria (2026).

aleatórias em nível de *pixel*, afetando diretamente padrões locais capturados pelas camadas convolucionais. Já o modelo baseado em *Vision Transformers* apresentou menor sensibilidade às perturbações avaliadas. Esse comportamento pode estar relacionado ao processamento da imagem em *patches* e à agregação de informações por meio de mecanismos de *self-attention*, que reduzem a influência de ruídos locais.

Por fim, observa-se que as perturbações tiveram menor impacto no *subset* OrganCMNIST quando comparado aos *subsets* BloodMNIST e PathMNIST. Isso provavelmente ocorre porque BloodMNIST e PathMNIST dependem fortemente de padrões texturais microscópicos, enquanto OrganCMNIST apresenta estruturas mais amplas e semanticamente distintas.

4.2. Limitações

Algumas limitações devem ser reconhecidas sobre este trabalho. Dentre elas, a não utilização de *kernels* especializados para dequantização (*ex.* Triton e ExLlamaV2) impediu a análise dos ganhos (ou perdas) de cada um dos métodos em termos de tempo de inferência para cenários reais. A não utilização de plataformas IoT para os experimentos também pode ter omitido a descoberta de dinâmicas relevantes, cuja análise enriqueceria a literatura científica. Ademais, a presente pesquisa utilizou apenas métodos de PTQ, o que impossibilitou a comparação de abordagens QAT [Tong et al. 2026].

5. Considerações Finais

Este trabalho apresentou uma avaliação empírica do método de quantização BenQ no contexto da classificação de imagens médicas. Os experimentos foram conduzidos em três subconjuntos do *benchmark* MedMNIST (BloodMNIST, PathMNIST e OrganCMNIST), selecionados de modo a representar diferentes características visuais e níveis de granularidade espacial. Os resultados indicam que o BenQ não está restrito ao contexto de grandes modelos de linguagem (LLMs), para o qual foi originalmente proposto. Em vez disso, o método também pode ser considerado uma alternativa viável de quantização para pipelines de IA médica que operam sob restrições de tempo computacional e memória. Entretanto, comparações com os métodos NF4 e RTN revelaram que nenhum método supera consistentemente os demais em todos os cenários avaliados, indicando que a escolha de uma estratégia de quantização deve considerar fatores como a arquitetura do modelo e as características da tarefa.

Diferenças significativas foram observadas entre as arquiteturas avaliadas. De modo geral, os *Vision Transformers* (ViT) demonstraram maior resiliência à quantização, com degradações de desempenho, em termos de F1-score, inferiores a 1%. Em contraste, redes neurais convolucionais (CNNs) apresentaram quedas de desempenho substancialmente maiores, ultrapassando 20% em determinadas configurações. Uma possível explicação é

que os mecanismos de agregação baseados em *self-attention* podem atenuar perturbações locais introduzidas pela quantização, enquanto filtros convolucionais podem ser mais sensíveis a essas perturbações, potencialmente amplificando o ruído de quantização ao longo das camadas da rede.

Por fim, os experimentos também investigaram a robustez dos métodos de quantização avaliados sob perturbações de entrada, especificamente ruído gaussiano e variações de contraste. Embora nenhum método tenha se mostrado consistentemente superior em todos os cenários, abordagens de quantização não-uniforme tenderam a apresentar maior robustez em diversos casos, possivelmente devido à sua capacidade de melhor se adaptar à distribuição estatística dos pesos da rede. Ao alocar níveis de quantização de forma mais densa em regiões com maior concentração de valores, esses métodos podem reduzir o erro médio de quantização. Esses achados reforçam a necessidade de uma seleção cuidadosa das estratégias de quantização em modelos de imagens médicas. Trabalhos futuros podem investigar mais profundamente a interação entre métodos de quantização, arquiteturas de modelos e características dos dados, a fim de compreender melhor como projetar sistemas de IA médica mais eficientes e robustos.

Agradecimentos

Os autores agradecem ao apoio da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, projeto APQ-01768-24), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), à Universidade Federal de Ouro Preto (PROPPI/UFOP) e a seu Programa de Pós-Graduação em Ciência da Computação (PPGCC/UFOP).

Referências

Abid, A., Sinha, P., Harpale, A., Gichoya, J., and Purkayastha, S. (2021). Optimizing medical image classification models for edge devices. In *International Symposium on Distributed Computing and Artificial Intelligence*, pages 77–87. Springer.

Askari Hemmat, M. H., Hemmat, R. A., Hoffman, A., Lazarevich, I., Saboori, E., Mastropietro, O., Sah, S., Savaria, Y., and David, J.-P. (2022). Qreg: On regularization effects of quantization. *arXiv preprint arXiv:2206.12372*.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American philosophical society*, pages 551–572.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2022). Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Hill, T. P. (1995). The significant-digit phenomenon. *The American Mathematical Monthly*, 102(4):322–327.

Kusk, M. W. and Lysdahlgaard, S. (2023). The effect of gaussian noise on pneumonia detection on chest radiographs, using convolutional neural networks. *Radiography*, 29(1):38–43.

Lu, Z.-x., Qian, P., Bi, D., Ye, Z.-w., He, X., Zhao, Y.-h., Su, L., Li, S.-l., and Zhu, Z.-l. (2021). Application of ai and iot in clinical medicine: summary and challenges. *Current medical science*, 41(6):1134–1150.

Manduva, V. C. (2024). Advancing ai in edge computing with graph neural networks for predictive analytics. *The Metascience*, 2(2):75–102.

Marchisio, K., Dash, S., Chen, H., Aumiller, D., Üstün, A., Hooker, S., and Ruder, S. (2024). How does quantization affect multilingual llms? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947.

Mekala, A., Atmakuru, A., Song, Y., Karpinska, M., and Iyyer, M. (2025). Does quantization affect models' performance on long-context tasks? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9433–9481.

Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., Van Baalen, M., and Blankevoort, T. (2021). A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*.

Negrão, A., Silva, P., Freitas, V. L., Moreira, G., and Luz, E. (2026). Benford's law as a distributional prior for post-training quantization of large language models. *arXiv preprint arXiv:2602.00165*.

Ott, J., Sun, H., Rinaldi, E., Mauro, G., Servadei, L., and Wille, R. (2025). Exploiting benford's law for weight regularization of deep neural networks. *Transactions on Machine Learning Research*.

Sahu, S. K., Java, A., and Shaikh, A. (2021). On the connection of benford's law and neural networks. *CoRR*.

Shandhi, M. M. H. and Dunn, J. P. (2022). Ai in medicine: Where are we now and where are we going? *Cell Reports Medicine*, 3(12).

Tong, Y., Yuan, J., and Hu, C. (2026). Enhancing quantization-aware training on edge devices via relative entropy coreset selection and cascaded layer correction. *IEEE Transactions on Mobile Computing*.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., and Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision.

Xi, L., Li, C., Anari, M. S., and Rezaee, K. (2025). Integrating wearable health devices with ai and edge computing for personalized rehabilitation. *Journal of Cloud Computing*, 14(1):64.

Xu, X., Lu, Q., Yang, L., Hu, S., Chen, D., Hu, Y., and Shi, Y. (2018). Quantization of fully convolutional networks for accurate biomedical image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8300–8308.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. (2023). Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific data*, 10(1):41.

Zhai, S., Wang, H., Sun, L., Zhang, B., Huo, F., Qiu, S., Wu, X., Ma, J., Wu, Y., and Duan, J. (2022). Artificial intelligence (ai) versus expert: A comparison of left ventricular outflow tract velocity time integral (lvot-vti) assessment between icu doctors and an ai tool. *Journal of applied clinical medical physics*, 23(8):e13724.

Zhao, H. (2023). Applications of embedded systems in medicine: Challenges and future trends. *Highlights Sci. Eng. Technol*, 62:31–35.