

REDUÇÃO DO EXCESSO DE CONFIANÇA EM MODELO DE DIAGNÓSTICO DE CÂNCER DE PELE

Marco Antonio Capucho Ribeiro¹, Kamilly Victória dos Santos Batista², Alan Afif Helal³,
Gledson Melotti⁴
Engenharias

Introdução e Objetivo

A inteligência artificial (IA) tem ganhado destaque em diversas áreas do conhecimento, impulsionada pelo avanço de algoritmos capazes de reconhecer padrões em bancos de dados de diferentes naturezas — sejam eles tabulares, visuais ou audiovisuais (GOODFELO; BENGIO; COURVILLE, 2016; BISHOP; BISHOP, 2023). Na busca por soluções mais práticas, a IA tem demonstrado habilidade para realizar tarefas tradicionalmente atribuídas a especialistas humanos, como o diagnóstico de doenças a partir da análise automatizada de imagens médicas (SUN, et al., 2024; ALSHMRANI, et al., 2023).

Na área da saúde, o uso da IA oferece benefícios relevantes. Ela possibilita a redução de custos operacionais, racionalização de exames, aumento da capacidade de atendimento e aprimoramento de estratégias de prevenção e acompanhamento de doenças. Um aspecto crítico é a otimização do tempo de atendimento: segundo dados da CNN, o tempo médio de espera em hospitais brasileiros varia de 90 a 180 minutos, e cerca de 24% da população já apresentou agravamento de saúde em função da demora em consultas e exames pelo Sistema Único de Saúde (SUS). Situações como essas poderiam ser mitigadas por meio de sistemas inteligentes capazes de priorizar casos urgentes, reduzir filas e otimizar recursos (SUN, et al., 2024; ALSHMRANI, et al., 2023).

Esses sistemas baseiam-se na análise de dados clínicos — como febre, dores, náuseas ou desmaios — para identificar padrões característicos de diferentes patologias. Com isso, torna-se viável prever ou confirmar diagnósticos de doenças como câncer de mama, pneumonia, tuberculose, Parkinson, doenças de pele e cardiovasculares, entre outras, oferecendo suporte rápido e embasado à tomada de decisão médica. Entretanto, um desafio

¹ Graduando em Engenharia Elétrica, Instituto Federal do Espírito Santo, marcocapucho@gmail.com

² Graduando em Engenharia Elétrica, Instituto Federal do Espírito Santo,
kamilly.batista@estudante.ifes.edu.br

³ Mestre em Engenharia Informática, Instituto Federal do Espírito Santo, alan.helal@ifes.edu.br

⁴ Doutor em Engenharia Eletrotécnica e de Computadores, Instituto Federal do Espírito Santo,
gledson@ifes.edu.br

recorrente nos modelos de aprendizado de máquina é o excesso de confiança em previsões incorretas, o que em contextos clínicos pode comprometer a segurança do paciente. A calibração probabilística surge, portanto, como etapa essencial para que as previsões reflitam adequadamente a incerteza envolvida (GUO et al., 2017; KULL, FILHO, FLACH, 2017).

Diante disso, o objetivo desta pesquisa é avaliar e comparar diferentes técnicas de calibração probabilística — Temperature Scaling, Isotonic Regression, Platt Scaling, Histogram Binning e Beta Calibration — aplicadas a modelos de IA para diagnóstico médico. Busca-se verificar quais métodos conseguem ajustar os escores previstos, tornando-as compatíveis com a frequência empírica dos eventos e, assim, garantir previsões mais seguras e clinicamente confiáveis (SUN et al., 2024).

Método e Metodologia

Esta pesquisa caracteriza-se como aplicada, de abordagem quantitativa e de natureza experimental, utilizando os conjuntos de dados histológicos NCT-CRC-HE-100K e o CRC-VAL-HE-7K (KATHER; HALAMAE, 2018).

O NCT-CRC-HE-100K (dados de treino, 70% dos dados, e validação, 30% dos dados) contém 100.000 imagens histológicas (224×224 pixels; 0,5 μm por pixel), coradas com hematoxilina e eosina (H&E) e normalizadas em cor. As imagens foram extraídas de 86 lâminas provenientes de amostras coletadas do *National Center for Tumor Diseases* (Centro Nacional de Doenças Tumoriais), *Heidelberg*, Alemanha e na *University Medical Center Mannheim* (Centro Médico Universitário de Mannheim), *Mannheim*, Alemanha. As classes de tecido incluem: adiposo (*Adipose* - ADI), fundo (BACK), detritos (*debris* - DEB), linfócitos (*lymphocytes* - LYM), muco (*mucus* - MUC), músculo liso (*smooth muscle* - MUS), mucosa normal do cólon (*normal colon mucosa* - NORM), estroma associado ao câncer (*cancer-associated stroma* - STR) e epitélio do adenocarcinoma colorretal (*colorectal adenocarcinoma epithelium* - TUM). Já o CRC-VAL-HE-7K (dados de teste) reúne 7.180 imagens de 50 pacientes com adenocarcinoma colorretal, sem sobreposição de indivíduos com o conjunto anterior, mantendo as mesmas especificações técnicas.

Para a classificação das imagens, foram avaliados diferentes modelos de redes neurais convolucionais (CNNs) consagrados na literatura: ConvNeXtLarge, ConvNeXtBase, EfficientNetB3, EfficientNetB0, InceptionV3, IResNet50, com ou sem o *Global Average*

Pooling-GAP (Agregação Média Global) após a última camada de convolução. Todas as redes foram treinadas a partir do zero, tendo o F1-score como métrica para fornecer o resultado de classificação. O melhor resultado foi com a rede EfficientNetB3_GAP.

Sobre esses modelos, foram aplicadas cinco técnicas de calibração probabilística: Temperature Scaling, Histogram Binning, Platt Scaling, Beta Calibration e Isotonic Regression. A avaliação da eficácia das calibrações foi conduzida por meio das métricas Erro de Calibração Esperado (*Expected Calibration Error* - ECE), Erro de Calibração Máximo (*Maximum Calibration Error* - MCE), Log-Verossimilhança Negativa (*Negative Log-Likelihood* - NLL) e apresentadas pelos Diagramas de Confiabilidade (GUO et al., 2017). Essas medidas permitiram verificar se as probabilidades previstas refletiam de forma adequada a incerteza real dos eventos, requisito essencial para aplicações clínicas seguras.

Discussão e Resultados

A figura a seguir exibe a calibração de cada modelo, medida por ECE, MCE e NLL (com $M = 15$ bins), antes e depois das aplicações das técnicas de calibração.

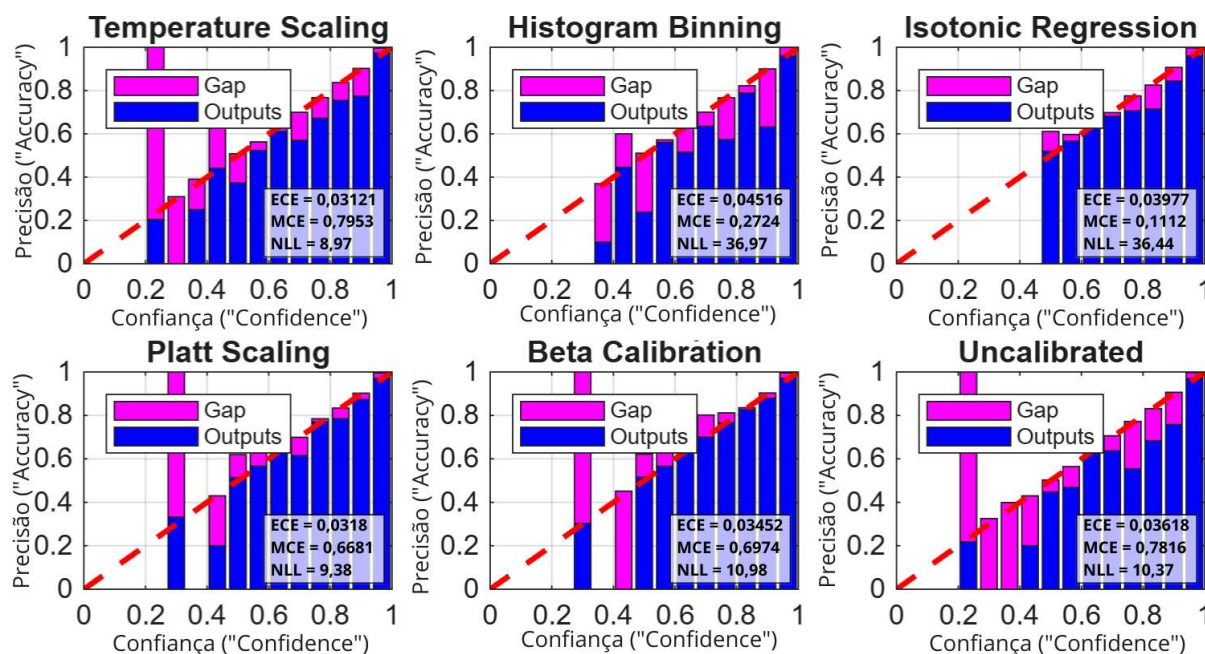


Figura 1 - Diagramas de Confiabilidade da rede EfficientNetB3_GAP com suas respectivas calibrações

Fonte: Autoria Própria (2025).

Os resultados evidenciam diferenças significativas entre os métodos de calibração aplicados ao modelo de classificação. No ECE, destacaram-se o Platt Scaling e o Temperature Scaling, que alcançaram os menores valores, conforme Figura 1. Em contrapartida, o Histogram Binning apresentou o pior resultado, sugerindo limitações para uso nos dados histológicos deste trabalho.

No MCE, o Isotonic Regression demonstrou maior robustez, reduzindo os piores casos de descalibração, enquanto o Temperature Scaling apresentou desempenho inconsistente, com variações que comprometem sua aplicabilidade clínica.

Nas métricas Negative Log-Likelihood (NLL), Temperature Scaling e Platt Scaling mostraram melhores qualidades probabilísticas, enquanto Histogram Binning e Isotonic Regression obtiveram valores elevados, indicando distorções nas estimativas de probabilidade.

Esses achados confirmam resultados de estudos anteriores quanto à eficácia do Platt Scaling.

Do ponto de vista prático, os resultados indicam que Platt Scaling é um método confiável para aumentar a segurança de sistemas de IA a partir dos dados histológicos apresentados neste trabalho, particularmente em diagnósticos críticos como câncer colorretal, onde reduzir erros graves é essencial. Já técnicas como Isotonic Regression e Histogram Binning devem ser utilizadas com cautela devido à instabilidade observada.

Considerações Finais

Este estudo teve como objetivo avaliar métodos de calibração de modelos de aprendizado de máquina aplicados à classificação médica. Os resultados demonstraram que Platt Scaling se destaca como técnica mais eficaz, atendendo ao objetivo de identificar um método capaz de reduzir erros de confiança e aumentar a segurança preditiva.

O trabalho dá conta de mostrar que esse método é uma alternativa viável para aplicação clínica, sobretudo em cenários críticos como o diagnóstico de câncer colorretal, em que decisões dependem de estimativas probabilísticas confiáveis.

Entre as potências do estudo, destacam-se a comparação sistemática de diferentes métricas (ECE, MCE, NLL) e a validação em arquiteturas variadas, o que fortalece a generalização dos

achados. Como dificuldades, ressalta-se a instabilidade observada em técnicas como Isotonic Regression e Histogram Binning, além da limitação de recursos computacionais para explorar um número maior de modelos e bancos de dados.

Foram respondidas as questões relativas à eficácia comparativa dos métodos e à adequação clínica de cada abordagem, confirmando achados da literatura. Contudo, não foi possível esgotar análises sobre o impacto da calibração em cenários de desbalanceamento severo de classes ou em dados longitudinais, o que permanece como questão em aberto.

Como perspectiva futura, sugere-se ampliar os experimentos para outros conjuntos de dados médicos, investigar estratégias híbridas de calibração e explorar a integração desses métodos em sistemas de apoio à decisão clínica em tempo real, avaliando sua aplicabilidade prática no fluxo hospitalar.

Referências

I. Goodfellow, Y. Bengio, and A. Courville. **Deep Learning**. The MIT Press, 2016.

C. M. Bishop, H. Bishop. **Deep Learning: Foundations and Concepts**. Springer, 2023.

J. Sun, D. Wei, L. Wang, and Y. Zheng. **Hybrid unsupervised representation learning and pseudo-label supervised self-distillation for rare disease imaging phenotype classification with dispersion-aware imbalance correction**, Medical Image Analysis, vol. 93, 2024.

G. M. M. Alshmrani, Q. Ni, R. Jiang, H. Pervaiz, and N. M. Elshennawy. **A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images**. Alexandria Engineering Journal, vol. 64, 923-935, 2023.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. **On calibration of modern neural networks**. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17). JMLR.org, 1321–1330, 2017.

Sun, S., Zhang, Y., Chen, M., Xiao, F., Wu, W., Li, X., Zheng, M., Luo, X., & Wang, D. **Reducing overconfident errors in molecular property classification using Posterior Network**. Patterns (New York, N.Y.), 5(6), 2024.

J. N. Kather, N. Halamae A. Marx. **100,000 histological images of human colorectal cancer and healthy tissue**. Zenodo, abr. 07, 2018.

Kull, M., Filho, T. S., Flach, P. **Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers**. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, In Proceedings of Machine Learning Research 54:623-631, 2017.