

Inteligência Artificial Generativa com Dados Limitados: Superando Desafios na Avaliação Automática da Dor Neonatal

Lucas F. Buzuti¹, Carlos E. Thomaz¹

¹Departamento de Engenharia Elétrica – Laboratório de Processamento de Imagens
Centro Universitário FEI – São Bernardo do Campo – SP – Brasil

lucas.buzuti@outlook.com, cet@fei.edu.br

Abstract. *This study investigates the potential of Generative Artificial Intelligence (GenAI) in neonatology, focusing on overcoming data scarcity for the automatic assessment of neonatal pain. The research led to the creation of SynUnifesp, a dataset composed entirely of synthetic images of newborn faces, used to expand limited datasets. To improve the evaluation of these generative models, the Fréchet AutoEncoder Distance (FAED) metric was proposed, which utilizes the latent space to ensure greater robustness against perturbations and better generalization than the traditional FID. The effectiveness of SynUnifesp was validated in classification tasks across three Artificial Intelligence paradigms: supervised learning, few-shot learning, and self-supervised learning. The results demonstrate that the inclusion of synthetic data provides significant performance gains in the automatic assessment of neonatal pain, consolidating GenAI as a viable tool for clinical advancements in Neonatal Intensive Care Units (NICU).*

Resumo. *Este trabalho investiga o potencial da Inteligência Artificial Generativa (IAGen) na neonatologia, focando na superação da escassez de dados para a avaliação automática da dor neonatal. A pesquisa resultou na criação do SynUnifesp, um conjunto de dados composto inteiramente por imagens sintéticas de faces de recém-nascidos, utilizado para expandir bases de dados limitadas. Para aprimorar a avaliação desses modelos generativos, foi proposta a métrica Fréchet AutoEncoder Distance (FAED), que utiliza o espaço latente para garantir maior robustez contra perturbações e melhor generalização que o FID tradicional. A eficácia do SynUnifesp foi validada em tarefas de classificação sob três paradigmas de Inteligência Artificial: aprendizado supervisionado, de poucas amostras (few-shot) e auto-supervisionado. Os resultados demonstram que a inclusão de dados sintéticos proporciona ganhos significativos no desempenho da avaliação automática da dor neonatal, consolidando a IAGen como uma ferramenta viável para avanços clínicos em Unidades de Terapia Intensiva Neonatal.*

1. Introdução

O trabalho explora a aplicação da Inteligência Artificial Generativa (IAGen) na neonatologia, destacando seu potencial para superar a escassez de dados na avaliação automática da dor neonatal. Enquanto modelos discriminativos tradicionais exigem grandes volumes de dados [Hinton et al. 2012], a IAGen surge como uma alternativa para

gerar dados sintéticos no intuito de ampliar base de dados com poucas amostras. Essa abordagem é crucial em ambientes de Unidade de Terapia Intensiva Neonatal (UTIN) [Beam et al. 2024], onde a coleta de dados reais é limitada por questões éticas, burocráticas e pela priorização do atendimento imediato ao recém-nascido.

Apesar dos avanços, a IAGen enfrenta desafios significativos, como a dificuldade de treinamento com conjuntos de dados restritos e a ausência de métricas de avaliação padronizadas [Borji 2022], ao contrário dos modelos discriminativos de aprendizado profundo. Na neonatologia, a falta de amostras suficientes tem exigido o uso de técnicas como o aprendizado por transferência e o aumento de dados [Zamzmi et al. 2019]. Nesse cenário, a geração de imagens sintéticas destaca-se como uma solução promissora para ampliar bases de dados insuficientes, tornando o processo de desenvolvimento de modelos de Inteligência Artificial (IA) menos invasivo e mais eficiente.

O objetivo central do estudo é utilizar a IAGen para aprimorar a avaliação da dor neonatal através da criação do conjunto de dados sintético SynUnifesp. A pesquisa busca avaliar a viabilidade prática dessas imagens geradas artificialmente, comparando o desempenho de diferentes paradigmas de aprendizado de máquina, incluindo o aprendizado supervisionado, o de poucas amostras (few-shot) e o auto-supervisionado. Dessa forma, o trabalho pretende validar como a ampliação diversificada de dados pode impactar positivamente a precisão diagnóstica em contextos clínicos reais de baixa disponibilidade de dados. Além disso, este trabalho propôs e implementou uma nova métrica para avaliação de imagens sintéticas denominada Fréchet AutoEncoder Distance (FAED) [Buzuti and Thomaz 2023]. Por limitação de página não foi possível descrever o FAED, para mais detalhes ler o artigo [Buzuti and Thomaz 2023].

2. Metodologia

2.1. Conjunto de Dados Unifesp

O conjunto de dados Unifesp [Heiderich et al. 2015] (Figura 1a) compreende 3.900 imagens faciais de 30 recém-nascidos brasileiros de etnia mista, das quais 360 imagens frontais foram selecionados para análise técnica. A amostra incluiu neonatos com idade gestacional média de 37 ± 1 semanas, abrangendo casos pré-termo tardio e a termo. A rotulagem clínica para detecção de dor foi conduzida por seis especialistas em neonatologia baseando-se no Sistema de Codificação Facial Neonatal (NFCS) [Rve 1987]. O estudo foi realizado no Hospital São Paulo sob aprovação dos Comitês de Ética em Pesquisa da Unifesp (protocolos n. 1299/09 e 3.116.146), garantindo o rigor ético e a validade clínica das imagens utilizadas para a avaliação automática.

2.2. Modelo Generativo Baseado em Auto-codificador Variacional

Os Auto-codificadores Variacionais Hierárquicos (HVAEs) [Klushyn et al. 2019] descrevem a distribuição conjunta $p(\mathbf{x}, \mathbf{z})$ via maximização do limite inferior variacional (ELBO) no treinamento [Klushyn et al. 2019]. Para aumentar a expressividade, as variáveis latentes $\mathbf{z}$ são organizadas em grupos disjuntos, com as probabilidades a priori e a posteriori aproximadas, estruturadas condicionalmente e representadas por distribuições Gaussianas [Klushyn et al. 2019] [Vahdat and Kautz 2020]. A implementação utiliza redes neurais *bottom-up* e *top-down* para inferência e geração. Emprega-se o truque de reparametrização [Kingma and Welling 2013] para viabilizar o treinamento via retropropagação.

O modelo *Nouveau Variational AutoEncoder* (NVAE) [Vahdat and Kautz 2020], variante avançada dos HVAE, foi implementado para gerar imagens sintéticas a partir do conjunto de dados da Unifesp, utilizando redes residuais com camadas de *squeeze and excitation* (SE). O pré-processamento com RetinaFace [Deng et al. 2020] resultou em 3.290 faces de recém-nascidos redimensionadas para 128×128 pixels, servindo de base para o treinamento. A qualidade das imagens geradas (Figura 1b) foi validada qualitativa e quantitativamente pela métrica FAED, comparando 50.000 amostras sintéticas com as originais para garantir a fidelidade perceptiva dos atributos faciais. Com um valor de FAED de 21,60, as imagens apresentaram um equilíbrio ideal entre fidelidade e diversidade.

2.3. Conjunto de Dados SynUnifesp

O conjunto de dados SynUnifesp [Buzuti et al. 2024], Figura 1b, foi construído por meio da seleção aleatória de um número de imagens sintéticas geradas pelo NVAE treinado sobre o conjunto de dados da Unifesp, com o objetivo de dobrar a quantidade original de imagens pré-processadas e rotuladas disponíveis originalmente. Essas imagens foram então avaliadas por três profissionais de saúde especialistas em neonatologia. Cada profissional avaliou as expressões faciais dos recém-nascidos nas imagens, determinando se estavam com dor ou não. Essa avaliação envolveu a pontuação de cinco regiões faciais: fronte saliente, fenda palpebral estreitada, sulco nasolabial aprofundado, boca aberta e boca estirada (horizontal ou verticalmente), também de acordo com o NFCS [Rve 1987]. Cada neonato fora considerado em estado de dor se três ou mais regiões faciais recebessem pontuação positiva. Utilizou-se o voto de maioria entre todos os profissionais para determinar o estado final.

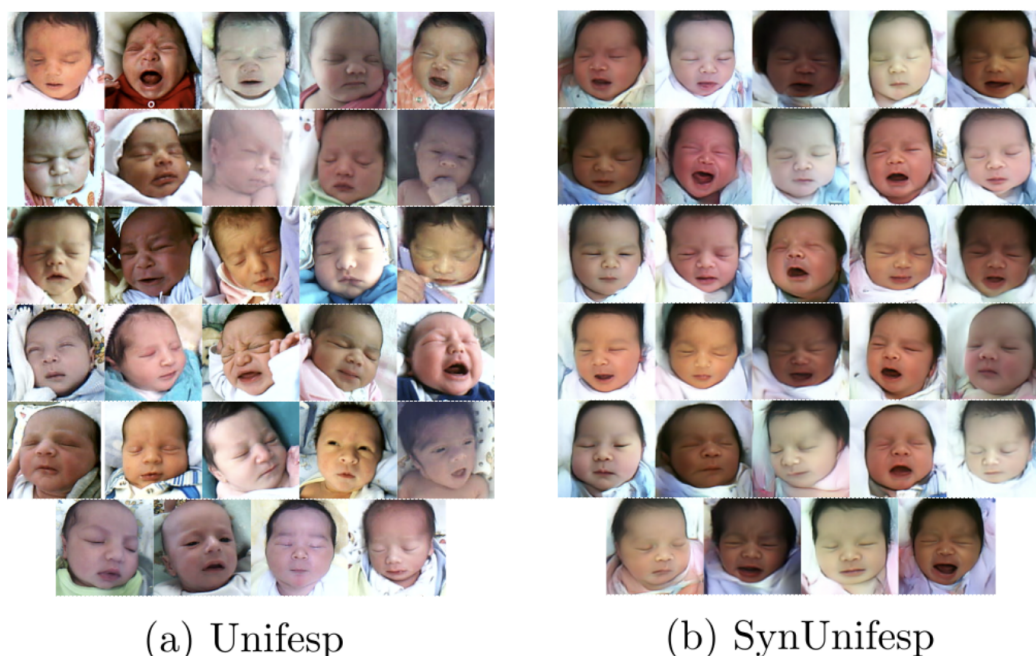


Figura 1. Amostras do conjunto de dados Unifesp (a) e SynUnifesp (b).

2.4. Paradigmas da Inteligência Artificial

Nesta seção apresenta-se os seguintes paradigmas da IA para avaliar o desempenho da IAGen no contexto da avaliação automática da dor neonatal: Aprendizado Supervisio-

nado (SL), Aprendizado de Poucos Amostras (FSL) e Aprendizado Auto-Supervisionado (SSL). Foram analisados dois cenários experimentais: o primeiro utilizando exclusivamente o conjunto de dados Unifesp (356 imagens pré-processadas via RetinaFace [Deng et al. 2020]) e o segundo combinando-o ao conjunto sintético SynUnifesp (712 imagens no total).

No paradigma SL, foram empregados os modelos VGG19 [Karen 2014], ResNet50 [He et al. 2016] e DeiT-Base-16 [Touvron et al. 2021], utilizando técnicas de transferência de aprendizado e aumento de dados padrão [Zamzmi et al. 2019]. Já no FSL, focado em cenários com escassez de dados, aplicou-se as arquiteturas Matching Networks [Vinyals et al. 2016] e Prototypical Networks [Snell et al. 2017] em configurações 2-way com 1-shot e 5-shot para a classificação binária da dor. No paradigma SSL utilizou-se o modelo DINOv2 [Oquab et al. 2023], pré-treinado com 50.000 imagens sintéticas geradas por NVAE para extração de características especializadas. Para a tarefa de classificação final, o codificador foi congelado e acoplado a classificadores como k -Vizinhos Mais Próximos (kNN), Regressão Logística e uma Rede Neural, garantindo uma avaliação abrangente da eficácia das representações aprendidas.

3. Resultados

Tabela 1. Resultados comparativos de acurácia entre os paradigmas SL, FSL e SSL utilizando os conjuntos Unifesp (356 imagens) e Unifesp + SynUnifesp (712 imagens).

Paradigma / Modelo	Conjunto	Acurácia (%)	Notas / Shot
<i>Aprendizado Supervisionado (SL)</i>			
VGG19	Unifesp	80.38 ± 7.32	–
ResNet50	Unifesp	86.04 ± 3.89	–
DeiT-Base-16	Unifesp	78.49 ± 3.70	–
VGG19	Unifesp + Syn	85.98 ± 3.07	–
ResNet50	Unifesp + Syn	87.10 ± 0.92	–
DeiT-Base-16	Unifesp + Syn	83.93 ± 4.36	–
<i>Aprendizado de Poucos Exemplos (FSL)</i>			
Matching Networks	Unifesp	$58.48 \pm 3.14 / 64.99 \pm 4.40$	1-s / 5-s
Prototypical Nets	Unifesp	$59.13 \pm 2.16 / 69.08 \pm 3.22$	1-s / 5-s
Matching Networks	Unifesp + Syn	$70.45 \pm 2.95 / 77.41 \pm 3.30$	1-s / 5-s
Prototypical Nets	Unifesp + Syn	$68.50 \pm 2.51 / 78.15 \pm 2.67$	1-s / 5-s
<i>Aprendizado Auto-Supervisionado (SSL - DINOv2)</i>			
kNN (k=3)	Unifesp	76.45 ± 1.61	–
Regressão Logística	Unifesp	80.36 ± 4.13	–
Rede Neural	Unifesp	87.41 ± 3.19	–
kNN (k=3)	Unifesp + Syn	77.48 ± 1.81	–
Regressão Logística	Unifesp + Syn	85.78 ± 2.01	–
Rede Neural	Unifesp + Syn	88.60 ± 2.39	–

A análise comparativa entre os cenários revelou que a inclusão do conjunto SynUnifesp elevou o desempenho em todos os paradigmas testados (Tabela 1). No SL e

FSL, houve um aumento significativo na acurácia média e redução do desvio padrão nas maiorias das comparações. No SSL, o DINOv2 atuou como um codificador especializado, otimizando a capacidade de generalização e da precisão dos classificadores kNN, Regressão Logística e Redes Neurais, com a expansão do conjunto de dados original adicionando imagens sintéticas.

4. Conclusão

O modelo NVAE, embora não tenha sido originalmente projetado para gerar novas amostras a partir de conjuntos com poucas amostras, demonstrou resultados promissores neste cenário. Nas análises realizadas, observou-se sua capacidade de gerar novas amostras com notável diversidade, evidenciando seu potencial para aumentar conjuntos de imagens destinados à automatização da avaliação da dor neonatal. Este desempenho pode ser atribuído à abordagem hierárquica do NVAE, que utiliza uma distribuição Multi-Gaussiana em vez de uma única distribuição Gaussiana, permitindo uma navegação mais refinada do espaço latente e maior controle sobre a distribuição dos dados. Com base nos resultados obtidos com o conjunto de dados Unifesp, observou-se que o modelo NVAE é adequado para a geração de imagens sintéticas que podem efetivamente contribuir para a automatização da avaliação da dor neonatal.

Questões Éticas

O trabalho alinha-se ao princípio fundamental defendido pelos comitês de ética. Entende-se que a priorização deve ser no tratamento imediato do recém-nascido em detrimento de procedimentos de construção de um conjunto de dados extensivo. Ao utilizar dados sintéticos, a metodologia desenvolvida respeita integralmente este imperativo ético, ao mesmo tempo em que proporciona avanços na capacidade de reconhecimento automático, sem comprometer a qualidade do atendimento ou submeter os pacientes neonatais a procedimentos desnecessários para fins de coleta de dados.

Agradecimentos

Ao LNCC (Laboratório Nacional de Computação Científica) pelo acesso ao supercomputador Santos Dumont, ao grupo de neonatologia da Universidade Federal de São Paulo (UNIFESP) pela colaboração e aos profissionais de saúde por avaliar as imagens de fezes neonatal sintéticas geradas por este trabalho. O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) processo nº 2018/13076-9 e do Centro Universitário FEI.

Referências

- Beam, K., Sharma, P., Levy, P., and Beam, A. L. (2024). Artificial intelligence in the neonatal intensive care unit: the time is now. *Journal of Perinatology*, 44(1):131–135.
- Borji, A. (2022). Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329.
- Buzuti, L., Giraldi, G., Heiderich, T., Barros, M., Guinsburg, R., and Thomaz, C. (2024). Generative ai for neonatal pain assessment: a sound approach to improve data-driven automatic recognition in intensive care unit. *Available at SSRN 4994076*.

- Buzuti, L. F. and Thomaz, C. E. (2023). Fréchet autoencoder distance: A new approach for evaluation of generative adversarial networks. *Computer Vision and Image Understanding*, page 103768.
- Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Heiderich, T. M., Leslie, A. T. F. S., and Guinsburg, R. (2015). Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. *Acta Paediatrica*, 104(2):e63–e69.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.
- Karen, S. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556*.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Klushyn, A., Chen, N., Kurle, R., Cseke, B., and van der Smagt, P. (2019). Learning hierarchical priors in vaes. *Advances in neural information processing systems*, 32.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2023). DINOv2: Learning robust visual features without supervision.
- Rve, G. (1987). Pain expression in neonates: facial action and cry. *Pain*, 28:395–410.
- Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357.
- Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679.
- Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *Advances in neural information processing systems*, 29.
- Zamzmi, G., Paul, R., Goldgof, D., Kasturi, R., and Sun, Y. (2019). Pain assessment from facial expression: Neonatal convolutional neural network (n-cnn). In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.