Human vs Machine Towards Neonatal Pain Assessment: A Comparison of the Facial Features Extracted by Adults and Convolutional Neural Networks

Lucas Pereira Carlini[1], **Carlos Eduardo Thomaz**[1]

[1]FEI, Depto. de Engenharia Elétrica, São Bernardo do Campo, SP, Brazil

lucaspcarlini10@gmail.com, cet@fei.edu.br

Resumo. *Recém-nascidos (RNs) em estado crítico ou prematuros passam diariamente por inúmeros procedimentos dolorosos, necessitando avaliação contínua da dor. No entanto, resultados recentes evidenciaram a subjetividade dos métodos aplicados por profissionais de saúde. Neste contexto, este trabalho aplica modelos computacionais para compreender de maneira específica a dor expressa por RNs, comparando as características faciais relevantes para a máquina com as regiões observadas por Médicos e Pais de RN durante avaliação da dor. Os resultados obtidos mostraram que as regiões relevantes para os modelos computacionais são clinicamente importantes e, em partes, concordam com a percepção facial humana.*

Abstract. *Critically ill or premature neonates experience numerous painful procedures daily, requiring continuous pain assessment. However, recent results have shown the subjectivity of the tools applied by health professionals. In this context, this work applies computational models to specifically understand the pain expressed by neonates, comparing machine-relevant facial features with the regions observed by Physicians and Parents of neonates during pain assessment. The results showed that the regions relevant to computational models are clinically important and partially agree with the human facial perception.*

1. Introdução

Recém-nascidos (RNs) são a prioridade para a continuidade dos seres humanos [Darmstadt et al. 2015]. No entanto, RNs em estado crítico ou prematuros passam por inúmeros procedimentos dolorosos durante permanência em Unidades de Terapia Intensiva Neonatal [Cruz et al. 2016], gerando efeitos degradantes no decorrer de suas vidas [Grunau 2020]. Consequentemente, várias escalas clínicas foram desenvolvidas para avaliar a dor neonatal, analisando principalmente a expressão facial do RN [Guinsburg 1999]. Embora utilizadas em situações clínicas, essas escalas são subjetivas, produzindo resultados variados baseados nas características do RN, seu cuidador e do ambiente onde se encontram [Barros et al. 2021]. Dessa forma, a avaliação precisa e objetiva da dor neonatal continua sendo um grande desafio para a comunidade científica.

Neste contexto, métodos baseados em Inteligência Artificial (IA) mostram-se como uma alternativa não-invasiva e específica para a avaliação contínua da dor neonatal [Carlini et al. 2021, Gkikas and Tsiknakis 2023]. Além disso, técnicas de Interpretação de IA (IIA) revelam o processo de tomada de decisão do modelo computacional, permitindo a identificação das características faciais relevantes para a tarefa

[Carlini et al. 2021]. Portanto, este trabalho investiga as características faciais relacionadas à avaliação da dor neonatal por humanos e máquinas, comparando as regiões faciais observadas por Médicos e Pais de RN com as características mais relevantes de um modelo de IA treinado especificamente para a avaliação da dor neonatal. As principais questões da atual pesquisa são: (1) Quais são as regiões faciais mais relevantes para um modelo de IA?, (2) Usando o mesmo modelo de IA, métodos de IIA distintos concordam entre si?, (3) Um modelo de IA concorda com a percepção facial de Médicos e Pais de RN?, (4) Existe alguma diferença na concordância entre grupos ao comparar a percepção em relação às imagens de "dor" e "sem dor"?.

2. Materiais e Métodos

Esta seção descreve os bancos de imagens, o arcabouço de rastreamento de olhar, os modelos computacionais e os métodos de IIA utilizados e implementados.

2.1. Bancos de Imagens de Faces

Para os presentes métodos, foram utilizados os seguintes bancos de imagens de face: UNIFESP [Heiderich et al. 2015] e *infant Classification of Pain Expression* (iCOPE) [Brahnam et al. 2005]. O Banco UNIFESP possui 164 imagens classificadas como "dor" e 196 amostras "sem dor", com dimensões de 320x233, de 30 RNs com até 168 horas de vida. Já o Banco iCOPE, com resolução de 3008x2000, possui 63 imagens categorizadas como "sem dor" e 60 amostras de "dor" de 26 RNs com até 72 horas de vida.

2.2. Arcabouço de Dados de Rastreamento de Olhar

Para analisar a percepção facial humana, foram utilizados dados oculares obtidos em pioneiro experimento anterior [Carlini et al. 2020]. Utilizando equipamento de captura de olhar, 44 Médicos e 29 Pais de RN avaliaram 10 imagens faciais de RN classificadas como "dor" e 10 imagens como "sem dor" do Banco UNIFESP. Cada foto foi selecionada por especialistas, considerando posição do RN e visibilidade da face. Após observarem cada imagem por 7s[1], o voluntário verbalizou uma nota de 0 (sem dor) a 10 (dor extrema). Os experimentos tiveram aprovação do Comitê de Ética em Pesquisa da Universidade Federal de São Paulo (1299/09, 3.116.151, 3.116.146 e 3.201.307).

2.3. Modelos Computacionais de Classificação

Foram utilizados os seguintes modelos computacionais: (1) VGG-Face, no qual as camadas convolucionais originais, treinadas para reconhecimento facial, foram conectadas a duas camadas densas sequenciais treinadas especificamente para a avaliação de dor neonatal [Parkhi et al. 2015]; (2) Neonatal Convolutional Neural Network (N-CNN), a qual sua arquitetura foi recentemente proposta especificamente para a avaliação de dor neonatal [Zamzmi et al. 2019]. Utilizando a técnica *leave-sample-subjects-out*, ambos os modelos foram treinados com a junção dos Bancos UNIFESP e iCOPE, porém, apenas a N-CNN foi treinada ponta-a-ponta [Coutrin et al. 2022].

[1]Tempo sugerido por testes piloto realizados previamente.

2.4. Métodos de Interpretação de Inteligência Artificial e Pós-Processamento

Para compreender o processo de tomada de decisão dos modelos computacionais, obtendo os escores de atribuição da imagem de entrada, foram aplicadas duas técnicas de IIA: (1) Integrated Gradients (IG), que calcula a contribuição individual de cada *pixel* da imagem de entrada para a resposta obtida pelo modelo [Sundararajan et al. 2017]; (2) Gradient-weighted Class Activation Mapping (GradCAM), que calcula a relevância, em relação à resposta do modelo, de cada mapa de características [Selvaraju et al. 2017].

Posteriormente, os escores de atribuição foram processados conforme os seguintes passos: (1) mapeamento para escala RGB, (2) agrupamento dos escores utilizando k-Means, (3) suavização usando filtro Gaussiano, e (4) filtro dos escores menos relevantes usando canal alpha. A partir dos escores processados, foi criada uma máscara de contorno, que revela quais regiões da imagem foram observadas por cada modelo e grupo de voluntário [Schiller et al. 2020]. Por fim, todas as máscaras foram comparadas entre si utilizando a Similaridade do Cosseno, Correlação de Spearman e Overlap (do inglês, sobreposição) [Schiller et al. 2020]. A última métrica revela a porcentagem total de *pixels* da imagem que ambos do par de comparação observaram ou ignoraram.

3. Resultados Experimentais

Esta seção apresenta as métricas de desempenho dos modelos computacionais e detalha as análises de comparação das regiões observadas por modelos humanos e máquinas.

3.1. Desempenho dos Modelos Computacionais

A VGG-Face obteve o melhor desempenho médio, resultando em elevada acurácia $(86,2\%)$, escore F1 $(87,7\%)$ e AUC $(85,4\%)$. O modelo também apresentou grande capacidade de generalização, evidenciado por sua precisão (85.9%) e sensibilidade (90.3%). Já o modelo N-CNN apresentou desempenho inferior, com comparada baixa acurácia $(77,1\%)$, escore F1 $(80,8\%)$ e AUC $(76,0\%)$. Mesmo com elevada sensibilidade $(89,0\%)$, o escore F1 foi negativamente impactado pela baixa precisão obtida $(74,6\%)$.

3.2. Análise Qualitativa das Regiões Faciais Observadas

Como exibido na Figura 1, o modelo VGG-Face, quando aplicado o método GradCAM, apresentou uma extração agrupada de características em regiões específicas da face. Para imagens de "dor", o método destacou as regiões da fronte, olhos, nariz e boca. Enquanto em imagens classificadas como "sem dor", foram observadas características faciais centrais, tais como a região entre sobrancelhas, nariz e boca. Já o método IG resultou em extração granular de características. Em imagens de "dor", foram destacadas as regiões dos olhos, nariz, sulco nasolabial e boca. Para imagens categorizadas como "sem dor", regiões centrais da face tiveram maior relevância.

Novamente, quando aplicado na N-CNN, o GradCAM apresentou uma extração agrupada de características, enquanto o IG teve uma extração granular. Interessantemente, ambos os métodos de IIA focaram na boca em quase todas as imagens de "dor", demonstrando a relevância da região. Esse destaque também foi obtido pelo método IG em imagens "sem dor". No entanto, para essa mesma categoria, o método GradCAM não apresentou nenhum padrão de destaque de região.

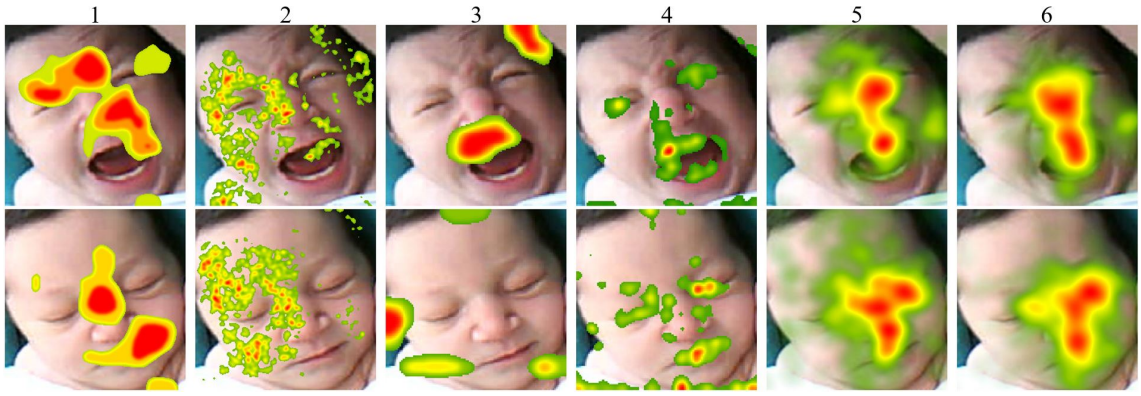

Figura 1. Exemplos de atribuições do GradCAM (1) e IG (2) aplicados na VGG-Face, GradCAM (3) e IG (4) aplicados na N-CNN, Pais (5) e Médicos (6).

Como esperado, Médicos e Pais de RN apresentaram uma percepção facial similar e holística baseada nas regiões dos olhos, nariz e boca, sem evidentes diferenças entre si e independente da categoria da imagem de RN. No entanto, Médicos aparentemente possuem uma percepção mais focada do que Pais. Comparando humanos com máquina, a extração agrupada de características do GradCAM na VGG-Face é a mais similar a percepção facial holística humana.

3.3. Análise Quantitativa da Comparação das Regiões Faciais Observadas

Para imagens de "dor"[2], a Similaridade do Cosseno indicou moderada correlação ($0,5 < \mu_{cosseno} < 0,7$) entre o GradCAM aplicado na VGG-Face com Médicos (0,5812) e Pais (0,5618), e forte correlação ($\mu_{cosseno} \geq 0,7$) entre humanos (0,7505). Já a Correlação de Spearman demonstrou moderada concordância ($0,4 < \mu_{Spearman} < 0,6$) entre o GradCAM aplicado na VGG-Face e Médicos (0,4272), e forte correlação ($\mu_{Spearman} \geq 0,6$) entre Médicos e Pais (0,6161). Para a métrica Overlap, todos os pares de comparação tiveram média ($0,5 < \mu_{Overlap} < 0,7$) ou forte concordância ($\mu_{Overlap} \geq 0,7$). Os principais destaques foram as concordâncias: entre humanos (0,8167), entre ambos os métodos de IIA aplicados na N-CNN (0,8016), e entre o GradCAM aplicado na VGG-Face com Médicos (0,7336) e Pais (0,6974).

Para imagens de "sem dor", novamente, a Similaridade do Cosseno indicou moderada correlação entre o GradCAM aplicado na VGG-Face com Médicos (0,5784) e Pais (0,5490), e forte concordância entre Médicos e Pais (0,7600). Além disso, foi observada também moderada Similaridade de Cosseno entre ambos os métodos de IIA aplicados na VGG-Face (0,5022). Em seguida, analisando a Correlação de Spearman, foram obtidas apenas moderadas concordâncias entre o GradCAM aplicado na VGG-Face com Médicos (0,4346) e entre humanos (0,5888). Em relação à métrica Overlap, exceto pela concordância entre GradCAM aplicado na N-CNN e Pais, todos os pares indicaram moderada ou forte correlação. Os resultados mais relevantes foram as concordâncias: entre humanos (0,7714), entre ambos os métodos de IIA aplicados na VGG-Face (0,7605), e entre o GradCAM aplicado na VGG-Face com Médicos (0,7033) e Pais (0,6452).

[2]Devido ao número limite de páginas, consultar dados completos nas páginas 61 e 62 da dissertação, disponível em `https://fei.edu.br/~cet/dissertacao_LucasPCarlini_2023.pdf`

Estatisticamente, investigando se há diferença da correlação dos pares entre imagens de "dor" e "sem dor", o GradCAM aplicado na N-CNN apresentou uma maior concordância com humanos quando analisando imagens de "dor", como evidenciado por todas as métricas. Além disso, a métrica Overlap indicou uma maior concordância entre o GradCAM aplicado na VGG-Face e Pais quando analisando imagens de "dor".

4. Discussão

Os achados deste trabalho evidenciam que técnicas de IIA se mostram adequadas para melhor compreender a tomada de decisão de modelos computacionais quando avaliando a dor neonatal. Especificamente, os resultados mostraram que regiões clinicamente importantes da face do RN são relevantes para estes modelos e que os mesmos concordam parcialmente com a percepção facial de Médicos e Pais de RN quando realizando essa mesma tarefa. Além disso, estes modelos apresentaram desempenho comparável a outras arquiteturas quando submetidas ao mesmo protocolo de treinamento [Coutrin et al. 2022].

Para a VGG-Face, as características faciais mais destacadas foram a fronte, regiões entre sobrancelhas, olhos, nariz e boca. Já para a N-CNN, houve clara relevância da região específica da boca. Ao comparar os dois métodos de IIA, foram observadas relevâncias distintas para um mesmo par de imagem e modelo. O GradCAM apresentou uma extração agrupada de características e o IG extraiu informação granular da face do RN. Estes resultados geram preocupações sobre o uso e interpretação destes métodos e, principalmente, sobre quais regiões da imagem são relevantes para a tomada de decisão da máquina. Analisando a concordância entre humanos e máquina, o método GradCAM aplicado na VGG-Face apresentou a maior similaridade com Médicos e Pais de RN. No entanto, a relevância da boca para a N-CNN é notável, visto que a região está associada a maior chance de humanos identificarem corretamente a presença de dor [Barros et al. 2021].

Como limitações desse trabalho, a baixa quantidade de imagens para treinamento dos modelos computacionais pode ter impactado no desempenho da N-CNN em particular e na qualidade das extrações de ambas arquiteturas. Em relação ao arcabouço de rastreamento ocular, a amostragem de voluntários foi por conveniência e o número de imagens de RN para teste reduzido, fatos que sugerem parcimônia nas comparações realizadas.

Agradecimentos

Ao apoio da FAPESP (2018/13076-9), CAPES (001), FEI e UNIFESP.

Referências

Barros, M. C. d. M., Thomaz, C. E., da Silva, G. V. T., do Carmo Azevedo Soares, J., Carlini, L. P., Heiderich, T. M., Orsi, R. N., Balda, R. d. C. X., Silva, P. A. S. O., Sanudo, A., et al. (2021). Identification of pain in neonates: the adults' visual perception of neonatal facial features. *Journal of Perinatology*, 41(9):2304–2308.

Brahnam, S., Chuang, C.-F., Shih, F. Y., and Slack, M. R. (2005). Svm classification of neonatal facial images of pain. *International Workshop on Fuzzy Logic and Applications*, pages 121–128.

Carlini, L. P., Ferreira, L. A., Coutrin, G. A., Varoto, V. V., Heiderich, T. M., Balda, R. C., Barros, M. C., Guinsburg, R., and Thomaz, C. E. (2021). A convolutional neural

network-based mobile application to bedside neonatal pain assessment. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 394–401. IEEE Computer Society.

Carlini, L. P., Soares, J. C. A., Silva, G. V. T., Heideirich, T. M., Balda, R. C. X., Barros, M. C. M., Guinsburg, R., and Thomaz, C. E. (2020). A visual perception framework to analyse neonatal pain in face images. In Campilho, A., Karray, F., and Wang, Z., editors, *Image Analysis and Recognition*, volume 12131 of *Lecture Notes in Computer Science*, pages 233–243, Cham. Springer International Publishing.

Coutrin, G. A., Carlini, L. P., Ferreira, L. A., Heiderich, T. M., Balda, R. C., Barros, M. C., Guinsburg, R., and Thomaz, C. E. (2022). Convolutional neural networks for newborn pain assessment using face images: A quantitative and qualitative comparison. In *Proceedings of the 3rd International Conference on Medical Imaging and Computer-Aided Diagnosis, MICAD 2022*. Springer LNEE.

Cruz, M., Fernandes, A., and Oliveira, C. (2016). Epidemiology of painful procedures performed in neonates: a systematic review of observational studies. *European Journal of Pain*, 20(4):489–498.

Darmstadt, G. L., Shiffman, J., and Lawn, J. E. (2015). Advancing the newborn and stillbirth global agenda: priorities for the next decade. *Archives of disease in childhood*, 100(Suppl 1):S13–S18.

Gkikas, S. and Tsiknakis, M. (2023). Automatic assessment of pain based on deep learning methods: A systematic review. *Computer Methods and Programs in Biomedicine*, 231:107365.

Grunau, R. E. (2020). Personal perspectives: Infant pain—a multidisciplinary journey. *Paediatric and Neonatal Pain*, 2(2):50–57.

Guinsburg, R. (1999). Avaliação e tratamento da dor no recém-nascido. *J Pediatr (Rio J)*, 75(3):149–60.

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.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.

Schiller, D., Huber, T., Dietz, M., and André, E. (2020). Relevance-based data masking: a model-agnostic transfer learning approach for facial expression recognition. *Frontiers in Computer Science*, 2:6.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

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.