CENTRO UNIVERSITÁRIO FEI
LEONARDO ANTUNES FERREIRA

SINAL CONTÍNUO DA DOR NEONATAL BASEADO NA EXPRESSÃO FACIAL: uma metodologia de *deep learning* confiável, interpretável e orientada à segurança operacional para a prática clínica

São Bernardo do Campo
2026

LEONARDO ANTUNES FERREIRA

SINAL CONTÍNUO DA DOR NEONATAL BASEADO NA EXPRESSÃO FACIAL: uma metodologia de *deep learning* confiável, interpretável e orientada à segurança operacional para a prática clínica

Tese de doutorado apresentada ao Centro Universitário FEI, como parte dos requisitos necessários para a obtenção do título de Doutor em Engenharia Elétrica. Orientado pelo Prof. Dr. Carlos Eduardo Thomaz.

São Bernardo do Campo
2026

Ferreira, Leonardo Antunes.
Sinal contínuo da dor neonatal baseado na expressão facial: uma metodologia de deep learning confiável, interpretável e orientada à segurança operacional para a prática clínica / Leonardo Antunes Ferreira. São Bernardo do Campo, 2026.
290 p. : il.

Tese - Centro Universitário FEI.
Orientador: Prof. Dr. Carlos Eduardo Thomaz.

1. dor neonatal. 2. sinal contínuo da dor. 3. expressão facial. 4. quantificação da incerteza. 5. inteligência artificial explicável. I. Thomaz, Carlos Eduardo, orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

APRESENTAÇÃO DE TESE
ATA DA BANCA EXAMINADORA

Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica

Doutorado

PGE-10

Aluno(a): Leonardo Antunes Ferreira **Matrícula:** 5241054

Título do Trabalho: SINAL CONTÍNUO DA DOR NEONATAL BASEADO NA EXPRESSÃO FACIAL: uma metodologia de deep learning confiável, interpretável e orientada à segurança operacional para a prática clínica

Área de Concentração: Processamento de Sinais e Imagens

Orientador(a): Prof. Dr. Carlos Eduardo Thomaz

Data da realização da defesa: 06/11/2026

ORIGINAL ASSINADA

Avaliação da Banca Examinadora:

O exame foi realizado em modo virtual e iniciou-se com a apresentação do aluno que durou 40 minutos. A seguir, os examinadores arguiram o aluno individualmente que respondeu de forma adequada todas as questões levantadas. A banca de examinadores avaliou a tese de doutorado como densa, robusta e com contribuição metodológica. A análise global foi de unanimidade para aprovação, com destaque para a excelência acadêmica e relevância para a prática clínica relacionada a dor neonatal.

A Banca Julgadora abaixo-assinada atribuiu ao aluno o seguinte resultado:

APROVADO ☒ REPROVADO ☐

MEMBROS DA BANCA EXAMINADORA

Prof. Dr. Carlos Eduardo Thomaz

Prof. Dr. Marco Antonio Gutierrez

Profª. Drª. Mavilde da Luz Gonçalves Pedreira

Profª. Drª. Maria Fernanda Branco de Almeida

Prof. Dr. Alexei Manso Correa Machado

Aprovação do Coordenador do Programa de Pós-graduação

Prof. Dr. Carlos Eduardo Thomaz

Dedico esta tese àqueles que caminharam comigo e tornaram este trabalho possível.

AGRADECIMENTOS

À minha família, que esteve ao meu lado durante todo este percurso. Em especial, agradeço aos meus pais, Elisabete e Joaquim, pelo apoio e incentivo constantes. Sem vocês, nada disso seria possível. Ao meu irmão, Vitor, agradeço pelos momentos de descontração e diversão. Sou muito grato por tê-los sempre por perto.

Ao meu orientador, Prof. Dr. Carlos Eduardo Thomaz, pela orientação, pelo apoio incansável e pelas contribuições para o desenvolvimento desta tese. Foi uma honra trabalhar com você neste projeto tão especial e impactante, e espero que ainda tenhamos muitas oportunidades de pesquisar juntos.

Aos meus amigos, Lucas Carlini e Gabriel Coutrin, que compartilham comigo uma amizade de longa data, agradeço pela colaboração e pelas discussões produtivas que contribuíram para o desenvolvimento das ideias desta tese. O trabalho não se encerra aqui. Temos um caminho muito longo pela frente, e estou ansioso para percorrê-lo ao lado de vocês.

Agradeço também às Dras. Marina C. M. Barros e Ruth Guinsburg, da Universidade Federal de São Paulo (UNIFESP), pela colaboração de longa data e pelo suporte clínico oferecido a este trabalho. Seus comentários e sugestões enriqueceram a tese e trouxeram uma perspectiva clínica mais ampla.

Ao Centro Universitário FEI, agradeço pelo ambiente de pesquisa, pelos recursos disponibilizados e pelo trabalho dos professores e colaboradores que contribuíram para minha formação acadêmica. Agradeço especialmente à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela bolsa de estudos que viabilizou esta pesquisa, e ao Laboratório Nacional de Computação Científica (LNCC), pelo acesso aos recursos computacionais do SDumont, essenciais para a realização dos experimentos.

Por fim, agradeço ao Prof. Dr. Carlos Francisco Moreno-García e a Craig Pirie por terem me recebido na Robert Gordon University, em Aberdeen, Escócia, durante o período de doutorado sanduíche. Essa experiência foi muito enriquecedora e contribuiu diretamente para o desenvolvimento da tese.

"O conhecido é finito, o desconhecido, infinito; intelectualmente estamos numa ilhota no meio de um oceano ilimitado de inexplicabilidade. Nossa função em cada geração é reivindicar um pouco mais de terra firme."

Huxley (1887)

RESUMO

A avaliação da dor neonatal é dificultada pela impossibilidade de autorrelato e pela natureza intermitente e subjetiva das escalas clínicas. Esta tese teve como objetivo desenvolver e avaliar um sinal contínuo da dor neonatal baseado na expressão facial, mediante uma metodologia formal, reprodutível e orientada à segurança. Nos bancos iCOPE e UNIFESP, os modelos NCNN, VGGFace e ViT-B/32 foram treinados para discriminar dor e ausência de dor, e no iCOPEVid, suas saídas foram avaliadas temporalmente. As arquiteturas apresentaram desempenho global comparável na validação cruzada, evidenciando que métricas discriminativas isoladas não caracterizam a confiabilidade. A incerteza epistêmica associou-se a erros e diferenças entre bases, enquanto a calibração mostrou-se dependente da arquitetura. Explicações agregadas por regiões faciais foram mais estáveis do que mapas em nível de pixel e apresentaram concordância moderada com escalas clínicas. Limiares de classificação, precisão e incerteza permitiram rejeitar predições frágeis, resultando em métricas mais altas no subconjunto aceito, ao custo de menor cobertura. Por fim, probabilidades, incerteza, explicações e limiares foram integrados em uma série temporal, cuja entropia e transições distinguiram padrões estáveis, irregulares, instáveis e indeterminados. O sinal da dor revelou variações de confiabilidade e dinâmica não captadas pelas classificações pontuais comumente utilizadas na literatura, constituindo um arcabouço de IA responsável para apoio ao monitoramento neonatal.

Palavras-chave: dor neonatal; sinal contínuo da dor; expressão facial; *deep learning*; quantificação da incerteza; inteligência artificial explicável.

ABSTRACT

Neonatal pain assessment is hindered by the inability to self-report and by the intermittent and subjective nature of clinical scales. This thesis aimed to develop and evaluate a continuous neonatal pain signal based on facial expressions through a formal, reproducible, and safety-oriented methodology. The NCNN, VGGFace, and ViT-B/32 models were trained on the iCOPE and UNIFESP datasets to discriminate between pain and no pain, and their outputs were evaluated temporally on iCOPEVid. The architectures achieved comparable overall performance in cross-validation, demonstrating that discriminative metrics alone do not characterize reliability. Epistemic uncertainty was associated with errors and differences between datasets, whereas calibration was architecture-dependent. Explanations aggregated by facial region were more stable than pixel-level maps and showed moderate agreement with clinical scales. Classification, precision, and uncertainty thresholds enabled fragile predictions to be rejected, yielding higher performance metrics for the accepted subset at the cost of reduced coverage. Finally, probabilities, uncertainty, explanations, and thresholds were integrated into a time series whose entropy and transitions distinguished stable, irregular, unstable, and indeterminate patterns. The pain signal revealed variations in reliability and temporal dynamics that are not captured by the pointwise classifications commonly used in the literature, constituting a responsible AI framework to support neonatal monitoring.

Keywords: neonatal pain; continuous pain signal; facial expression; *deep learning*; uncertainty quantification; explainable artificial intelligence.

LISTA DE FIGURAS

Figura 1 – Visão geral da metodologia do Capítulo 3. 42
Figura 2 – Exemplos de imagens do banco UNIFESP. 45
Figura 3 – Exemplos de imagens do banco iCOPE. 46
Figura 4 – Detecção facial e pontos fiduciais extraídos pelo SCRFD-10GF. 48
Figura 5 – Arquitetura da NCNN. 49
Figura 6 – Arquitetura da VGGFace. 50
Figura 7 – Arquitetura do ViT-B/32. 51
Figura 8 – Processo de validação cruzada. 53
Figura 9 – Imagens após *Data Augmentation*. 53
Figura 10 – Visão geral da metodologia do Capítulo 4. 57
Figura 11 – Representação gráfica das incertezas aleatória e epistêmica. 58
Figura 12 – Funcionamento do Monte Carlo Dropout. 59
Figura 13 – Distribuição da incerteza epistêmica em função dos parâmetros do MCDP. ... 62
Figura 14 – Distribuição da incerteza epistêmica em função dos parâmetros do Deep Ensembles. 63
Figura 15 – Distribuição da incerteza epistêmica em função das respostas corretas e incorretas. 66
Figura 16 – Distribuição da incerteza epistêmica em função dos bancos de imagens. 68
Figura 17 – Distribuição da incerteza epistêmica em função das predições dos modelos. .. 70
Figura 18 – Faces médias das predições estratificadas por quartis de incerteza epistêmica. 72
Figura 19 – Visão geral da metodologia do Capítulo 5. 75
Figura 20 – Curvas de calibração com excesso e falta de confiança. 81
Figura 21 – Métodos de discretização em intervalos. 82
Figura 22 – Influência do número de intervalos no ECE. 85
Figura 23 – Curvas de calibração da NCNN. 88
Figura 24 – Curvas de calibração da VGGFace. 89
Figura 25 – Curvas de calibração do ViT-B/32. 90
Figura 26 – Faces médias das predições em função da probabilidade predita. 92
Figura 27 – Visão geral da metodologia do Capítulo 6. 94
Figura 28 – Pós-processamento aplicado ao *Integrated Gradients*. 97
Figura 29 – Estratégias de perturbação avaliadas. 99
Figura 30 – Regiões faciais derivadas automaticamente dos pontos fiduciais. 102
Figura 31 – Explicações médias e faces médias para as classes *Dor* e *Sem dor*. 103
Figura 32 – AOPC da estratégia ROAD em relação à perturbação aleatória. 106
Figura 33 – AOPC do consenso global com média simples e ponderada. 108
Figura 34 – Consenso global por média simples e ponderação por AOPC. 109

Figura 35 – Visão geral da metodologia do Capítulo 7. 115
Figura 36 – F1-Score em função da porcentagem de dados retidos por θ_2 e por seleção aleatória. 122
Figura 37 – F1-Score em função da porcentagem de dados retidos por θ_3 e por seleção aleatória. 124
Figura 38 – Visão geral da metodologia do Capítulo 9. 132
Figura 39 – Exemplos de *frames* do banco iCOPEVid. 133
Figura 40 – Matrizes de decisão para sinais da dor nos estados "com dor" e "sem dor". ... 140
Figura 41 – Sinal da dor do tipo estável. 141
Figura 42 – Sinal da dor do tipo irregular. 143
Figura 43 – Sinal da dor do tipo instável. 145
Figura 44 – Sinal da dor do tipo indeterminado do modelo NCNN. 147
Figura 45 – Evolução temporal do sinal da dor e da importância relativa das regiões faciais do modelo VGGFace. 150
Figura 46 – Representação conceitual de um monitor futuro para suporte à avaliação da dor neonatal. 159
Figura 47 – Exemplos de predições com alta incerteza da NCNN. 176
Figura 48 – Exemplos de predições com baixa incerteza da NCNN. 177
Figura 49 – Exemplos de predições com alta incerteza da VGGFace. 178
Figura 50 – Exemplos de predições com baixa incerteza da VGGFace. 179
Figura 51 – Exemplos de predições com alta incerteza do ViT-B/32. 180
Figura 52 – Exemplos de predições com baixa incerteza do ViT-B/32. 181
Figura 53 – Concordância em nível de região entre os explicadores para o modelo NCNN. 187
Figura 54 – Concordância em nível de *pixel* entre os explicadores para o modelo NCNN.. 188
Figura 55 – Concordância em nível de região entre os explicadores para o modelo VGGFace. 189
Figura 56 – Concordância em nível de *pixel* entre os explicadores para o modelo VGGFace. 190
Figura 57 – Concordância em nível de região entre os explicadores para o modelo ViT-B/32. 191
Figura 58 – Concordância em nível de *pixel* entre os explicadores para o modelo ViT-B/32. 192
Figura 59 – Amostras removidas pelo limiar de precisão θ_2. 194
Figura 60 – Amostras removidas pelo limiar de incerteza θ_3. 195
Figura 61 – Amostras removidas pela aplicação conjunta dos limiares θ_2 e θ_3. 196
Figura 62 – Sinal da dor do segmento de dor do recém-nascido S002. 199
Figura 63 – Sinal da dor do segmento sem dor do recém-nascido S003. 200
Figura 64 – Sinal da dor do segmento de dor do recém-nascido S005. 201
Figura 65 – Sinal da dor do segmento sem dor do recém-nascido S006. 202
Figura 66 – Sinal da dor do segmento de dor do recém-nascido S007. 203
Figura 67 – Sinal da dor do segmento sem dor do recém-nascido S008. 204

LISTA DE TABELAS

Tabela 1 – Síntese dos trabalhos relacionados e suas limitações. 38
Tabela 2 – Características dos bancos de dados utilizados. 45
Tabela 3 – Configuração dos ambientes computacionais. 46
Tabela 4 – Matriz de confusão em classificação binária. 50
Tabela 5 – Hiperparâmetros e parâmetros utilizados no treinamento dos modelos. 54
Tabela 6 – Métricas de classificação dos modelos. 55
Tabela 7 – Correlações de Spearman (ρ) e Pearson (r) entre incerteza epistêmica e erro de classificação. 64
Tabela 8 – Métricas de classificação dos modelos após quantificação da incerteza. 65
Tabela 9 – Métricas de classificação e calibração dos modelos após quantificação da incerteza e calibração. 86
Tabela 10 – Presença de regiões faciais em escalas neonatais de dor. 102
Tabela 11 – Explicadores com concordância média em nível de pixel superior a 0,50. 103
Tabela 12 – Explicadores com pelo menos duas regiões concordantes na avaliação regional. 104
Tabela 13 – Ranking das estratégias de perturbação por modelo (AOPC entre parênteses). 105
Tabela 14 – Ranking dos explicadores por modelo para a estratégia ROAD (AOPC entre parênteses). 107
Tabela 15 – Ranking de importância das regiões faciais (em %). 110
Tabela 16 – Concordância global entre o consenso dos explicadores e escalas clínicas de dor. 110
Tabela 17 – Valores dos limiares θ_1, θ_2 e θ_3 para os modelos. 118
Tabela 18 – Métricas de classificação e calibração dos modelos após aplicação do limiar θ_1. 119
Tabela 19 – Métricas de classificação e calibração dos modelos após aplicação do limiar θ_2. 120
Tabela 20 – Métricas de classificação e calibração dos modelos após aplicação do limiar θ_3. 120
Tabela 21 – Métricas de classificação e calibração dos modelos após aplicação do limiar θ_{1+2+3}. 121
Tabela 22 – Estratégias metodológicas aplicadas a cada modelo ao longo da metodologia proposta. 127
Tabela 23 – Evolução das métricas da NCNN ao longo das etapas metodológicas (valores em %). 128
Tabela 24 – Evolução das métricas da VGGFace ao longo das etapas metodológicas (valores em %). 128
Tabela 25 – Evolução das métricas do ViT-B/32 ao longo das etapas metodológicas (valores em %). 129
Tabela 26 – Latência média ± desvio-padrão para cada etapa da metodologia. Valores em milissegundos. 130

Tabela 27 – Características dos bancos de dados utilizados. 134
Tabela 28 – Métricas de classificação e calibração dos modelos no sinal da dor sem aplicação da metodologia proposta (valores em %). 138
Tabela 29 – Métricas de classificação e calibração por tipo de sinal da dor e modelo após a rejeição pelos limiares θ_2 e θ_3..... 139
Tabela 30 – Centroides dos agrupamentos e percentual médio de *frames* rejeitados pelos limiares θ_2 e θ_3, por tipo de sinal da dor..... 140
Tabela 31 – Correlação entre sinais da dor, suas derivadas e concordância categórica κ.... 148
Tabela 32 – Correlações regionais entre a derivada do sinal da dor e as derivadas das séries temporais de XAI por modelo e classe (média ± desvio-padrão). 152
Tabela 33 – Métricas de classificação da NCNN sob diferentes configurações de MCDP e Deep Ensembles. 175
Tabela 34 – Métricas de classificação da VGGFace sob diferentes configurações de MCDP e Deep Ensembles..... 175
Tabela 35 – Métricas de classificação do ViT-B/32 sob diferentes configurações de MCDP e Deep Ensembles. 176
Tabela 36 – Métricas de classificação da NCNN. 183
Tabela 37 – Métricas de calibração da NCNN..... 183
Tabela 38 – Métricas de classificação da VGGFace..... 184
Tabela 39 – Métricas de calibração da VGGFace. 184
Tabela 40 – Métricas de classificação do ViT-B/32. 184
Tabela 41 – Métricas de calibração do ViT-B/32. 185

LISTA DE ABREVIATURAS E SIGLAS

AOPC	*Area Over the Perturbation Curve*
APN-db	*Acute Pain in Neonates database*
APS	*Acute Pain Scale*
AUC	*Area Under the Curve*
AWS	*Amazon Web Services*
BGR	*Blue, Green, Red*
BIIP	*Behavioural Indicators of Infant Pain*
BNN	*Bayesian Neural Network*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CFM	Conselho Federal de Medicina
CNN	*Convolutional Neural Network*
CPU	*Central Processing Unit*
DL	*DeepLIFT*
DLS	*DeepLIFT SHAP*
DP	Desvio-padrão
DSFD	*Dual Shot Face Detector*
ECE	*Expected Calibration Error*
ED	*Edge Detector*
EUA	Estados Unidos da América
EVENDOL	*Evaluation Enfant Douleur*
FLACC	*Face, Legs, Activity, Cry, Consolability Scale*
FLOPs	*Floating-Point Operations*
GPU	*Graphics Processing Unit*
GC	*Grad-CAM*

GS	*Gradient SHAP*
H2AI	*Human-to-Artificial Intelligence*
HITL	*Human-in-the-Loop*
HOG	*Histograms of Oriented Gradients*
IA	Inteligência Artificial
IASP	*International Association for the Study of Pain*
iCOPE	*infant Classification of Pain Expression*
iCOPEVid	*infant Classification of Pain Expression Videos*
IG	*Integrated Gradients*
KNN	*k-Nearest Neighbors*
K–S	Kolmogorov–Smirnov
LBP	*Local Binary Patterns*
LDA	*Linear Discriminant Analysis*
LGPD	Lei Geral de Proteção de Dados
LIDS	*Liverpool Infant Distress Scale*
LIME	*Local Interpretable Model-Agnostic Explanations*
LNCC	Laboratório Nacional de Computação Científica
LSR	*Label Smoothing Regularizer*
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolute Error*
MCDP	*Monte Carlo Dropout*
MCE	*Maximum Calibration Error*
MCGRATH	*McGrath Facial Affective Scale*
MIPS	*Modified Infant Pain Scale*
MLP	*Multilayer Perceptron*
M-NFCS	*Modified Neonatal Facial Coding System*

MoRF *Most Relevant First*

NCNN *Neonatal Convolutional Neural Network*

NFCS *Neonatal Facial Coding System*

NIPS *Neonatal Infant Pain Scale*

NLL *Negative Log-Likelihood*

NNICUPAT *Nepean Neonatal Intensive Care Unit Pain Assessment Tool*

NNSOA *Neural Network Simultaneous Optimization Algorithm*

NPAD *Neonatal Pain Assessment Dataset*

PASPI *Pain Assessment Scale for Preterm Infants*

PCA *Principal Component Analysis*

PDSE Programa de Doutorado Sanduíche no Exterior

PIPP *Premature Infant Pain Profile*

PIPP-R *Premature Infant Pain Profile-Revised*

P-MIPS *Partial Modified Infant Pain Scale*

POPA *Pose-invariant and Occlusion-robust Pain Assessment*

QI Quantificação da Incerteza

RAM *Random Access Memory*

ReLU *Rectified Linear Unit*

RGB *Red, Green, Blue*

RN Recém-Nascido

RNN *Recurrent Neural Network*

ROAD *RemOve And Debias*

ROC *Receiver Operating Characteristic*

SCRFD-10GF *Sample and Computation Redistribution for Efficient Face Detection*

SENet *Squeeze-and-Excitation Network*

SHAP *Shapley Additive Explanations*

SVM	*Support Vector Machine*
UNIFESP	Universidade Federal de São Paulo
USF-MNPAD-I	*University of South Florida Multimodal Neonatal Pain Assessment Dataset I*
UTIN	Unidade de Terapia Intensiva Neonatal
VGG	*Visual Geometry Group*
ViT	*Vision Transformer*
VLM	*Vision-Language Model*
VPN	Valor Preditivo Negativo
VPP	Valor Preditivo Positivo
VRAM	*Video Random Access Memory*
XAI	*Explainable Artificial Intelligence*

LISTA DE SÍMBOLOS

A, B	Parâmetros da função sigmoide de Platt. Em métricas de calibração, B também denota o número de intervalos.
$A = (A_1, \ldots, A_R)$	Vetor de importâncias das regiões faciais.
α	Transparência e canal de transparência do mapa de atribuição.
B_m	m-ésimo intervalo de probabilidade.
$\lvert B_m \rvert$	Número de amostras no intervalo B_m.
C	Escore bruto de concordância com escalas clínicas.
$c(t)$	Classe instantânea do sinal da dor.
D	Distância do teste de Kolmogorov–Smirnov.
$E^{(a)}, E^{(b)}$	Dois explicadores distintos.
ϵ	Parâmetro de suavização no *label smoothing*.
η	Taxas de aprendizado.
$\mathbb{F}(x)$	Saída do modelo para a entrada x.
g_r	Indicador binário da presença da região r na escala clínica.
H, W	Altura e largura do mapa de saliência.
$\mathbb{H}$	Entropia de Shannon do sinal da dor.
$\mathcal{I}$	Função isotônica de calibração.
$\mathbb{I}(\cdot)$	Função indicadora.
k_{ens}	Número de modelos que compõem o *Deep Ensemble*.
κ	Coeficiente kappa de concordância.
L	Número total de etapas de perturbação.
L_1, L_2	Termos de regularização L1 e L2.
$M^{(a)}, M^{(b)}$	Máscara binária de importância e máscaras produzidas por dois explicadores.
N_+, N_-	Quantidades de amostras positivas e negativas na calibração de Platt com *soft-labels*.

N_{bins}	Número de intervalos (*bins*) usados no *Histogram Binning*.
N_{MC}	Número de passagens ou amostragens de Monte Carlo.
N_{trans}	Número total de transições de classe no sinal.
Ω_r	Conjunto de pixels pertencentes à região r.
$P = \sum_r g_r$	Número de regiões clinicamente relevantes na escala.
$\mathcal{P}$	Conjunto de coordenadas dos pixels no mapa de saliência.
$\mathbb{P}(\cdot)$	Operador de probabilidade.
p	Probabilidade predita.
$\hat{p}_m, \hat{p}_i$	Probabilidade calibrada do intervalo m e da amostra i.
$\hat{p}_t, \hat{p}$	Probabilidade predita no instante t e média temporal das probabilidades.
p_{drop}	Taxa de *dropout* empregada no MCDP.
p_{y_+}, p_{y_-}	Probabilidades associadas às classes positiva e negativa.
ϕ	Mapeamento de pixel para região semântica facial.
Q	Estatística do teste de Friedman.
$\mathcal{R} = \{r_1, \ldots, r_R\}$	Conjunto das R regiões faciais.
r	Coeficiente de correlação de Pearson. Em XAI, o símbolo também identifica uma região facial.
ρ	Coeficiente de correlação de Spearman.
S, S_{ij}	Mapa de saliência e valor de saliência no pixel (i, j).
$s_{\text{bruto}}(t), s(t)$	Sinal bruto e sinal suavizado da dor no instante t.
$S_{\text{global}}, S_i, m_i$	Mapa de saliência de consenso, mapa do explicador i e respectivo valor da métrica de fidelidade.
σ	Incerteza epistêmica, calculada como o desvio-padrão das probabilidades preditas.
T	Temperatura no *temperature scaling*.
τ_k	Limiar percentual para selecionar os pixels *top-k*.
$\theta_1, \theta_2, \theta_3$	Limiares operacionais de classificação, precisão e incerteza, respectivamente.

$Top_k^{(a)}, Top_k^{(b)}$	Listas das k regiões mais relevantes de dois explicadores.
$u(t)$	Estado de confiabilidade do sinal no instante t.
w	Tamanho da janela da média móvel do sinal da dor.
$\xi, P(\xi_i)$	Variável aleatória discretizada do sinal e probabilidade de ocorrência de seu i-ésimo intervalo.
y_+, y_-	Evento ou classe positiva e negativa, respectivamente (dor e sem dor).
y, y_i	Classe verdadeira e classe verdadeira da amostra i.
y_c, y'_c	Classe original e classe suavizada da classe c.
z	*Logits* do modelo.
$\langle \cdot \rangle$	Média calculada no conjunto de dados.
$\uparrow, \downarrow$	Melhora e piora de uma métrica em relação ao modelo de referência.
$\uparrow^*, \downarrow^*$	Melhora e piora estatisticamente significativas.
$=$	Resultado inalterado nas tabelas comparativas.

SUMÁRIO

1	**INTRODUÇÃO**	25
1.1	PROBLEMA DE PESQUISA E MOTIVAÇÃO	26
1.2	OBJETIVOS	27
1.3	PERGUNTAS DE PESQUISA	28
1.4	ESTRUTURA DO TEXTO	28
2	**REVISÃO DA LITERATURA**	30
2.1	FENÔMENO DA DOR	30
2.1.1	**Dor neonatal**	31
2.2	HUMAN-IN-THE-LOOP E IA CONFIÁVEL EM SISTEMAS CLÍNICOS	33
2.3	AVALIAÇÃO AUTOMÁTICA DA DOR NEONATAL	34
2.3.1	**Classificação em imagens**	34
2.3.2	**Classificação em vídeos**	35
2.3.3	**Classificação multimodal**	36
2.3.4	**IA Explicável (XAI)**	36
2.4	LACUNAS NA LITERATURA	37
3	**TREINAMENTO DOS MODELOS COMPUTACIONAIS**	42
3.1	INTRODUÇÃO	42
3.1.1	**Convolutional Neural Networks**	43
3.1.2	***Vision transformers***	43
3.2	MATERIAIS	44
3.2.1	**Banco de imagens UNIFESP**	44
3.2.2	**Banco de imagens iCOPE**	45
3.2.3	**Hardware e software**	46
3.3	MÉTODOS	47
3.3.1	**Detecção facial**	47
3.3.2	**Modelos de classificação**	47
3.3.2.1	***NCNN***	48
3.3.2.2	***VGGFace***	48
3.3.2.3	***ViT-B/32***	49
3.3.3	**Métricas de classificação**	49
3.4	EXPERIMENTO	52
3.5	RESULTADOS	54
3.6	DISCUSSÃO	55
3.6.1	**Implicações para a prática clínica**	56

4 QUANTIFICAÇÃO DA INCERTEZA 57
4.1 INTRODUÇÃO 57
4.2 MÉTODOS 59
4.2.1 Monte Carlo Dropout 59
4.2.2 Deep Ensembles 60
4.3 EXPERIMENTO 60
4.4 RESULTADOS 61
4.4.1 Avaliação dos métodos de quantificação da incerteza 61
4.4.2 Fatores associados à incerteza 65
4.4.3 Análise qualitativa da incerteza 71
4.5 DISCUSSÃO 71
4.5.1 Implicações para a prática clínica 73

5 CALIBRAÇÃO 75
5.1 INTRODUÇÃO 75
5.2 MÉTODOS 76
5.2.1 Funções de ativação de saída 77
5.2.2 Calibração de Platt 77
5.2.3 Regressão Isotônica 78
5.2.4 Histogram Binning 78
5.2.5 Temperature Scaling 79
5.2.6 Label Smoothing Regularizer 79
5.2.7 Soft-labels por meio do NFCS 79
5.2.8 Métricas de calibração 80
5.2.8.1 Curva de calibração 80
5.2.8.2 Expected Calibration Error 81
5.2.8.3 Brier Score 82
5.2.8.4 Negative log-likelihood 83
5.3 EXPERIMENTO 83
5.4 RESULTADOS 84
5.4.1 Avaliação dos calibradores 84
5.4.2 Análise qualitativa da calibração 91
5.5 DISCUSSÃO 91
5.5.1 Implicações para a prática clínica 93

6 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL 94
6.1 INTRODUÇÃO 94
6.2 MÉTODOS 95
6.2.1 Explainable artificial intelligence 95
6.2.2 Pós-processamento dos mapas de atribuição 96

6.2.3 **Discordância entre explicadores** 97
6.2.4 **Análise de fidelidade por perturbação de pixels** 98
6.2.5 **Consenso global entre explicadores** 99
6.2.6 **Concordância com escalas clínicas de dor** 100
6.3 EXPERIMENTO 101
6.4 RESULTADOS 102
6.4.1 **Discordância entre explicadores** 102
6.4.2 **Fidelidade por perturbação de pixels** 105
6.4.3 **Consenso global entre explicadores** 107
6.4.4 **Concordância global com escalas clínicas de dor** 110
6.5 DISCUSSÃO 111
6.5.1 **Implicações para a prática clínica** 113

7 **LIMIARES OPERACIONAIS** 115
7.1 INTRODUÇÃO 115
7.2 MÉTODOS 116
7.2.1 **Limiar de classificação θ_1** 116
7.2.2 **Limiar de precisão θ_2** 116
7.2.3 **Limiar de incerteza θ_3** 117
7.3 EXPERIMENTO 117
7.4 RESULTADOS 118
7.5 DISCUSSÃO 125
7.5.1 **Implicações para a prática clínica** 125

8 **MODELOS FINAIS** 127
8.1 INTRODUÇÃO 127
8.2 EXPERIMENTO 127
8.3 RESULTADOS 127
8.4 DISCUSSÃO 130

9 **DESENVOLVIMENTO E AVALIAÇÃO DO SINAL DA DOR** 132
9.1 INTRODUÇÃO 132
9.2 MATERIAIS 133
9.2.1 **Banco de vídeos iCOPEVid** 133
9.3 MÉTODOS 134
9.3.1 **Sinal da dor** 134
9.3.2 **Métricas dinâmicas do sinal da dor** 135
9.3.2.1 ***Média temporal da probabilidade*** 135
9.3.2.2 ***Incerteza temporal do modelo*** 135
9.3.2.3 ***Entropia de Shannon*** 135

9.3.2.4 ***Transições de classe*** 136
9.4 EXPERIMENTO 137
9.5 RESULTADOS 138
9.5.1 Avaliação do sinal da dor e seus tipos 138
9.5.1.1 ***Sinais Estáveis*** 139
9.5.1.2 ***Sinais Irregulares*** 142
9.5.1.3 ***Sinais Instáveis*** 144
9.5.1.4 ***Sinais Indeterminados*** 146
9.5.1.5 ***Correlação entre os sinais da dor*** 148
9.5.2 Associação entre o sinal da dor e as expressões faciais 148
9.6 DISCUSSÃO 151
9.6.1 Implicações para a prática clínica 153

10 CONCLUSÃO 155
10.1 REALIZAÇÃO DO OBJETIVO GERAL E DOS OBJETIVOS ESPECÍFICOS 156
10.1.1 Objetivo específico 1: desenvolver e avaliar modelos de reconhecimento automático da dor neonatal 156
10.1.2 Objetivo específico 2: investigar incerteza e calibração probabilística 156
10.1.3 Objetivo específico 3: avaliar a explicabilidade das predições 156
10.1.4 Objetivo específico 4: definir formalmente os limiares operacionais dos modelos 157
10.1.5 Objetivo específico 5: integrar e avaliar o sinal contínuo da dor 157
10.2 LIMITAÇÕES DA TESE 157
10.3 AGENDA DE PESQUISA 158
10.3.1 Publicações 159

REFERÊNCIAS 161

APÊNDICE A – RESULTADOS AUXILIARES DA QUANTIFICAÇÃO DA INCERTEZA 174
A.1 MÉTRICAS DE CLASSIFICAÇÃO SOB DIFERENTES CONFIGURAÇÕES DE INCERTEZA 175
A.2 EXEMPLOS QUALITATIVOS DE PREDIÇÕES COM ALTA E BAIXA INCERTEZA 176

APÊNDICE B – RESULTADOS AUXILIARES DE CALIBRAÇÃO 182

APÊNDICE C – RESULTADOS AUXILIARES DE XAI 186

APÊNDICE D – RESULTADOS AUXILIARES DOS LIMIARES 193

APÊNDICE E – RESULTADOS AUXILIARES DO SINAL DA DOR ... 197

APÊNDICE F – ARTIGO PUBLICADO NO WORKSHOP *PREDICTIVE INTELLIGENCE IN MEDICINE (PRIME 2023)* 205

APÊNDICE G – ARTIGO PUBLICADO NA REVISTA *COMPUTERS IN BIOLOGY AND MEDICINE* 216

APÊNDICE H – ARTIGO PUBLICADO NA CONFERÊNCIA *WORLD CONFERENCE ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI 2025)* 226

APÊNDICE I – ARTIGO PUBLICADO NA CONFERÊNCIA *WORLD CONFERENCE ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI 2026)* 252

APÊNDICE J – CAPÍTULO PUBLICADO NO LIVRO *GRANDES DESAFIOS DA COMPUTAÇÃO APLICADA À SAÚDE* 277

1 INTRODUÇÃO

A dor é um problema relevante de saúde pública, com efeitos clínicos, sociais e econômicos amplamente documentados (Tracey; Mantyh, 2007; Filho; Silva, 2011; Treede *et al.*, 2015; Souza *et al.*, 2017; Abd-Elsayed, 2019; Gkikas; Tsiknakis, 2023). Na prática clínica, porém, sua avaliação torna-se particularmente difícil quando o paciente não consegue descrevê-la verbalmente. Escalas de autorrelato, embora extensamente utilizadas, não se aplicam a populações pré-verbais ou com limitações de comunicação (Merskey, 1979; Barua *et al.*, 2022; Llerena *et al.*, 2023). Nesses casos, a dor precisa ser inferida por indicadores comportamentais e fisiológicos observados por terceiros.

A definição de dor adotada em 1979 pela *International Association for the Study of Pain* (IASP) vinculava a experiência dolorosa à lesão tecidual real ou potencial ou à possibilidade de descrevê-la em termos dessa lesão (IASP, 1979). Entre as propostas de atualização, Williams e Craig (2016) defenderam a inclusão explícita dos componentes cognitivos e sociais, considerando a influência de fatores como memória, expectativas, crenças, ansiedade e contexto interpessoal. Em 2020, a IASP passou a definir a dor como "uma experiência sensitiva e emocional desagradável associada, ou semelhante àquela associada, a uma lesão tecidual real ou potencial" (Raja *et al.*, 2020). Logo, a literatura reconhece a dor como uma experiência pessoal, influenciada em graus variados por fatores biológicos, psicológicos e sociais, e esclarece que a incapacidade de se comunicar verbalmente não exclui a possibilidade de sentir dor (Raja *et al.*, 2020).

Em recém-nascidos (RNs), a expressão facial ocupa posição central nesse processo, pois está presente desde o nascimento e constitui um dos principais parâmetros observacionais das escalas neonatais de dor (Franck; Miaskowski, 1997; Guinsburg *et al.*, 2000; Schiavenato *et al.*, 2008; Fitzgerald; Walker, 2009). Essa avaliação é particularmente crítica em Unidades de Terapia Intensiva Neonatal (UTIN), nas quais os pacientes são submetidos a um número elevado de procedimentos dolorosos, muitas vezes superior a 300 durante a hospitalização (Perry *et al.*, 2018). Estima-se que aproximadamente dois terços desses pacientes não recebam monitoramento sistemático e contínuo da dor (Anand *et al.*, 2017). A limitação de recursos humanos e a subjetividade das avaliações intermitentes podem, assim, contribuir para intervenções tardias ou inadequadas (Balda *et al.*, 2000; Cruz; Fernandes; Oliveira, 2016; Prestes *et al.*, 2016). A relevância clínica do problema também decorre do reconhecimento, consolidado nas últimas décadas, de que a dor neonatal foi historicamente subestimada e de que a exposição repetida à dor pode produzir consequências de curto e longo prazo no desenvolvimento neurológico e comportamental do RN (Swafford; Allan, 1968; Anand; Hickey *et al.*, 1987; Grunau; Holsti; Peters, 2006; Walker, 2008; Rodkey; Riddell, 2013; Grunau, 2013; Zamzmi *et al.*, 2018a).

Nesse contexto, modelos de Inteligência Artificial (IA) têm sido investigados para automatizar o reconhecimento da dor neonatal a partir da face e de outros indicadores comportamentais e fisiológicos (Zamzmi *et al.*, 2018b; Sario *et al.*, 2023; Gkikas; Tsiknakis, 2023; Heiderich *et al.*, 2023). Modelos de *deep learning* são particularmente promissores e têm se destacado entre

as pesquisas mais recentes (Zamzmi *et al.*, 2018a; Zamzmi *et al.*, 2019; Carlini *et al.*, 2021; Salekin *et al.*, 2021; Coutrin *et al.*, 2023; Zhao *et al.*, 2023; Ferreira *et al.*, 2023; Carlini *et al.*, 2024; Ferreira *et al.*, 2025). Contudo, a maior parte da literatura da área, a exemplo de outras pesquisas sobre IA em saúde, ainda se concentra na melhoria de métricas discriminativas em condições controladas. Questões essenciais ao uso clínico seguro, como quantificação da incerteza e calibração probabilística, ainda recebem avaliação insuficiente ou pouco padronizada (Seoni *et al.*, 2023; Collins *et al.*, 2024). De modo semelhante, explicabilidade, robustez e definição dos limiares operacionais permanecem desafios para a translação tecnológica desses modelos à prática clínica (Lekadir *et al.*, 2025; Kim *et al.*, 2026). Persiste, portanto, uma lacuna entre o bom desempenho experimental e a confiança necessária para apoiar decisões em ambientes reais, tema central do conceito de IA responsável (Arrieta *et al.*, 2020).

Essa lacuna ganhou relevância adicional no contexto regulatório brasileiro recente, pois a Lei n. 15.378/2026, que institui o Estatuto dos Direitos do Paciente, reforça direitos relacionados à segurança, à informação, à confidencialidade, ao consentimento e à participação do paciente ou de seu representante nos cuidados em saúde (Brasil, 2026). De modo complementar, a Resolução do Conselho Federal de Medicina (CFM) n. 2.454/2026 normatiza o uso da IA na medicina e estabelece que esses sistemas devem apoiar a prática médica com transparência, validação científica, governança, proteção de dados, supervisão humana e preservação da autonomia do médico (Conselho Federal de Medicina, 2026). O problema investigado nesta tese, portanto, não se restringe ao desempenho computacional apenas. Ele integra uma demanda clínica e normativa por sistemas capazes de informar seus limites, justificar suas saídas e apoiar a tomada de decisão sem substituir a avaliação profissional. Essa integração é compatível com uma arquitetura *Human-in-the-Loop* (HITL), na qual a IA não atua como agente autônomo de diagnóstico, mas produz evidências que permanecem sujeitas à interpretação, à contestação e à eventual rejeição por uma pessoa responsável (Holzinger, 2016; Shneiderman, 2020).

1.1 PROBLEMA DE PESQUISA E MOTIVAÇÃO

O nascimento seguro e o desenvolvimento saudável na primeira infância constituem fundamentos do capital humano, com repercussões duradouras sobre a saúde, a aprendizagem, o bem-estar e a produtividade ao longo da vida (Richter *et al.*, 2017). Essa constatação torna a qualidade do cuidado neonatal particularmente importante, sobretudo para prematuros internados em UTIN. A exposição repetida à dor e ao estresse decorrentes de procedimentos pode afetar o desenvolvimento cerebral, o neurodesenvolvimento e a reatividade posterior à dor (Grunau; Holsti; Peters, 2006; Grunau, 2013). Avaliar a dor de forma oportuna e consistente não é apenas uma medida de conforto imediato, mas uma condição para proteger o desenvolvimento neurológico e comportamental do RN, com impactos de longo prazo sobre a saúde e a qualidade de vida.

O problema central desta tese consiste em transformar predições pontuais de modelos

de *deep learning* em um sinal contínuo da dor neonatal baseado na expressão facial, que representa tanto a dinâmica temporal das probabilidades estimadas quanto a confiabilidade e a interpretabilidade dessas predições. Para isso, essa proposição deve ser sustentada por uma metodologia formal, reprodutível e orientada à segurança operacional para a prática clínica, que explicite as condições de aceitação, rejeição ou encaminhamento das saídas à avaliação humana sob a perspectiva HITL.

A motivação decorre de duas limitações complementares. A primeira está no processo assistencial, em outras palavras, a avaliação manual da dor neonatal é intermitente, suscetível à variabilidade entre avaliadores e dependente de treinamento especializado (Anand *et al.*, 2017; Zamzmi *et al.*, 2018b; Sario *et al.*, 2023). A segunda está na literatura computacional, em que modelos preditivos em saúde frequentemente reportam desempenho médio, mas carecem de análises sistemáticas sobre a confiabilidade de suas probabilidades, a coerência de suas decisões e as condições em que deixam de ser adequados (Seoni *et al.*, 2023; Collins *et al.*, 2024; Lekadir *et al.*, 2025; Kim *et al.*, 2026).

1.2 OBJETIVOS

O objetivo geral desta tese é desenvolver e avaliar um sinal contínuo da dor neonatal baseado na expressão facial, sustentado por uma metodologia formal, reprodutível e orientada à segurança operacional para a prática clínica, que integre modelos de *deep learning*, quantificação da incerteza, calibração, IA explicável e limiares operacionais, de modo a incorporar mecanismos de confiabilidade, interpretabilidade e segurança à avaliação contínua.

A proposta constitui uma base computacional para uma avaliação contínua da dor neonatal à beira do leito e tem potencial para apoiar a identificação de alterações ao longo do tempo, sem substituir o julgamento dos profissionais de saúde, constituindo uma nova ferramenta para a prática clínica.

Para alcançar esse objetivo geral, foram estabelecidos os seguintes objetivos específicos:

a) desenvolver e avaliar modelos de *deep learning* para estimar, a partir da expressão facial neonatal, probabilidades de presença e ausência de dor;

b) investigar a quantificação da incerteza e a calibração das saídas, avaliando sua contribuição para a confiabilidade do sinal;

c) avaliar métodos de IA explicável e sua associação com regiões faciais reconhecidas em escalas clínicas de dor neonatal;

d) definir limiares operacionais para classificação, confiabilidade e incerteza, estabelecendo condições explícitas de aceitação, rejeição e supervisão humana;

e) integrar e avaliar esses componentes em um sinal contínuo da dor, caracterizando seus padrões temporais, sua confiabilidade, a concordância entre modelos e sua associação com a dinâmica das expressões faciais.

1.3 PERGUNTAS DE PESQUISA

A tese é guiada pela seguinte pergunta central de pesquisa: Como desenvolver e avaliar, para apoio à prática clínica, um sinal contínuo da dor neonatal baseado na expressão facial que integre desempenho preditivo, confiabilidade, explicabilidade e segurança operacional?

Essa pergunta central é desdobrada nas seguintes perguntas específicas:

a) em que medida modelos de *deep learning* conseguem discriminar dor versus não dor a partir da expressão facial neonatal de forma consistente?

b) a quantificação da incerteza e a calibração melhoram a confiabilidade das predições, tornando as saídas mais adequadas para suporte à decisão clínica?

c) as explicações produzidas pelos modelos são coerentes com regiões faciais reconhecidas na literatura e úteis para analisar a plausibilidade clínica das predições?

d) é possível definir limiares operacionais explícitos que determinem quando o sistema pode ser utilizado com segurança e quando sua saída deve ser rejeitada e analisada por especialistas?

e) como a integração das probabilidades, da incerteza, dos limiares operacionais e das explicações em um sinal contínuo permite caracterizar a dinâmica e a confiabilidade temporal das expressões faciais associadas à dor neonatal?

1.4 ESTRUTURA DO TEXTO

Esta tese está organizada em capítulos interdependentes que desenvolvem, de forma progressiva, os pilares metodológicos necessários à construção e à avaliação do sinal contínuo da dor:

a) **capítulo 2**: apresenta a revisão da literatura sobre dor neonatal, métodos clínicos de avaliação e reconhecimento automático da dor em RNs;

b) **capítulo 3**: estabelece a base discriminativa do sinal por meio do desenvolvimento dos modelos de *deep learning*, incluindo bases de dados, pré-processamento, arquiteturas e protocolo de treinamento, teste e validação;

c) **capítulo 4**: investiga métodos de quantificação da incerteza preditiva, que constituem um dos mecanismos de confiabilidade do sinal;

d) **capítulo 5**: analisa técnicas de calibração para conferir significado empírico às probabilidades que compõem o sinal;

e) **capítulo 6**: estabelece o pilar de explicabilidade, examinando a coerência entre as regiões faciais relevantes para os modelos e o conhecimento clínico descrito na literatura;

f) **capítulo 7**: estabelece o pilar de segurança operacional, definindo cenários de falha e condições para aceitação, rejeição ou revisão das predições;

g) **capítulo 8**: consolida os resultados das etapas anteriores e apresenta os modelos finais após a integração das estratégias de confiabilidade adotadas;

h) **capítulo 9**: materializa a contribuição central da tese ao construir e avaliar o sinal contínuo da dor, integrando as saídas probabilísticas, a incerteza, os limiares operacionais e as explicações em uma representação temporal interpretável;

i) **capítulo 10**: sintetiza as contribuições da tese, discute limitações metodológicas e clínicas e aponta direções para trabalhos futuros.

2 REVISÃO DA LITERATURA

Este capítulo apresenta uma síntese do estado da arte da avaliação automática da dor neonatal com base na expressão facial, articulando fundamentos fisiológicos da dor, sistemas de IA confiáveis com HITL, contexto regulatório recente sobre IA na medicina, escalas clínicas tradicionais e abordagens computacionais contemporâneas.

A revisão da literatura foi conduzida com base nos indexadores *Google Scholar* e *Semantic Scholar*, por meio de combinações dos descritores: “neonatal pain”, “neonatal pain assessment”, “facial expression”, “automatic”, “machine learning” e “deep learning”. O recorte temporal priorizou publicações entre 2015 e 2026. Foram incluídos estudos que (I) aplicaram métodos computacionais à detecção ou estimação da dor neonatal e (II) reportaram métricas quantitativas de desempenho. Trabalhos exclusivamente descritivos ou sem validação experimental foram excluídos dessa etapa, porém foram utilizados como base da referência bibliográfica da tese. A busca foi complementada por análise de citações cruzadas, com o objetivo de identificar estudos seminais. Ao final, a Tabela 1 sintetiza os trabalhos relacionados e suas limitações.

2.1 FENÔMENO DA DOR

A dor resulta da interação entre os sistemas nervosos periférico e central (Millan, 1999). O sistema nervoso periférico detecta estímulos nocivos por meio de receptores especializados e os transmite à medula espinhal (Basbaum *et al.*, 2009). O sistema nervoso central, por sua vez, integra essas informações em estruturas como o tronco encefálico e o córtex, onde se estabelece a percepção consciente (Hatfield, 2014). O processamento nociceptivo compreende quatro etapas:

a) **transdução**: ativação de nociceptores que convertem estímulos térmicos, mecânicos ou químicos em potenciais elétricos de ação (Basbaum *et al.*, 2009);

b) **transmissão**: propagação dos sinais elétricos à medula espinhal por meio de fibras A-delta (dor rápida e localizada) e fibras C (dor lenta e difusa), com projeções ascendentes ao tálamo e ao córtex (Millan, 1999);

c) **percepção**: interpretação cortical dos sinais nociceptivos, com integração de componentes sensoriais, emocionais e cognitivos. Trata-se de uma experiência subjetiva modulada por fatores biológicos e contextuais (Tracey; Mantyh, 2007);

d) **modulação**: mecanismos endógenos que amplificam ou inibem a sinalização dolorosa, contribuindo para a variabilidade interindividual na resposta à dor (Fields, 2004).

Tradicionalmente, a dor é classificada como aguda ou crônica. A dor aguda exerce função protetora ao sinalizar lesões reais ou potenciais, tem causa e localização identificáveis e apresenta duração limitada (Carr; Goudas, 1999). Em contraste, a dor crônica é persistente, frequentemente dissociada de uma função adaptativa e associada a alterações neuroplásticas, o

que exige abordagens terapêuticas multidimensionais (Apkarian; Baliki; Geha, 2009; Treede *et al.*, 2015). A complexidade desses mecanismos, da transdução periférica à modulação central, dificulta a adaptação do manejo clínico às necessidades de cada paciente.

2.1.1 Dor neonatal

Neonatos apresentam competência neuroanatômica para a percepção da dor desde fases gestacionais precoces. A partir da 24ª semana de gestação, nociceptores periféricos encontram-se funcionais e, por volta da 26ª semana, prematuros extremos já manifestam respostas comportamentais a estímulos nocivos (Hatfield, 2014).

Embora compartilhem vias de transmissão com adultos, os RNs apresentam imaturidade dos sistemas inibitórios descendentes. A elevada densidade de receptores cutâneos, associada à mielinização incompleta e à modulação inibitória reduzida, torna os neonatos particularmente vulneráveis a experiências dolorosas intensas e difusas (Hatfield, 2014). A exposição repetida à dor nesse período crítico associa-se a alterações duradouras no desenvolvimento neurológico, incluindo redução do limiar doloroso e maior reatividade ao estresse ao longo da vida (Guinsburg, 2011; Hatfield, 2014).

Em razão da incapacidade de verbalização, a avaliação da dor em RNs baseia-se exclusivamente em indicadores indiretos. Sinais comportamentais, como expressão facial, choro e atividade motora, e parâmetros fisiológicos, como frequência cardíaca e saturação de oxigênio, constituem as principais fontes de informação (Guinsburg, 2011). Contudo, os indicadores fisiológicos apresentam especificidade limitada para a detecção da dor (McGrath, 1990; Pereira *et al.*, 1999; Llerena *et al.*, 2023).

Com o objetivo de reduzir a subjetividade inerente à avaliação observacional (Balda *et al.*, 2000; Prkachin *et al.*, 2001; Cruz; Fernandes; Oliveira, 2016), diversas escalas clínicas foram desenvolvidas. Tamanaka *et al.* (2023) identificaram 52 escalas baseadas em expressão facial publicadas até 2020, o que evidencia a ausência de uma escala de dor universalmente adotada.

A expressão facial é considerada o indicador isolado mais sensível e específico da dor neonatal (Schiavenato *et al.*, 2008). Ações como franzir a testa, cerrar os olhos e aprofundar o sulco nasolabial compõem um padrão facial consistente, observado em RNs a termo e pré-termo (Grunau; Craig, 1987; Bucsea *et al.*, 2023). Entre as escalas que descrevem características faciais específicas relacionadas à dor neonatal, destacam-se:

a) **APS (*Acute Pain Scale*)**: avalia dor aguda por meio de indicadores faciais observáveis, incluindo fechamento dos olhos, contração das sobrancelhas e aprofundamento do sulco nasolabial (Carbajal *et al.*, 1997);

b) **BIIP (*Behavioural Indicators of Infant Pain*)**: combina cinco ações faciais, movimentos das mãos e estado comportamental para avaliar a dor em prematuros (Holsti; Grunau, 2007);

c) **EVENDOL (*Evaluation Enfant Douleur*)**: avalia dor por meio de cinco domínios comportamentais: expressão vocal ou verbal, expressão facial, movimentos, postura corporal e interação com o ambiente (Fournier-Charrière *et al.*, 2012);

d) **LIDS (*Liverpool Infant Distress Scale*)**: quantifica desconforto e sofrimento infantil a partir de indicadores comportamentais e expressões faciais, incluindo choro e movimentos corporais (Horgan; Choonara, 1996);

e) **MCGRATH (*McGrath Facial Affective Scale*)**: utiliza representações graduadas de expressões faciais para caracterizar a intensidade do desconforto ou da dor percebida (McGrath, 1985);

f) **MIPS (*Modified Infant Pain Scale*)**: avalia dor por meio da expressão facial, choro e respostas motoras (Buchholz *et al.*, 1998);

g) **M-NFCS (*Modified Neonatal Facial Coding System*)**: versão simplificada da NFCS baseada em ações faciais específicas associadas à dor neonatal, incluindo fronte saliente, olhos cerrados e sulco nasolabial aprofundado (Rushforth; Levene, 1994);

h) **NFCS (*Neonatal Facial Coding System*)**: avalia ações faciais específicas associadas à dor, incluindo fronte saliente, olhos cerrados, sulco nasolabial aprofundado, abertura da boca, estiramento vertical ou horizontal da boca e tremor do queixo (Grunau; Craig, 1987);

i) **NNICUPAT (*Nepean Neonatal Intensive Care Unit Pain Assessment Tool*)**: combina indicadores comportamentais e fisiológicos, incluindo expressão facial, choro, frequência cardíaca e parâmetros respiratórios, para avaliação da dor em neonatos hospitalizados (Marceau, 2003);

j) **PASPI (*Pain Assessment Scale for Preterm Infants*)**: avalia dor em prematuros por meio de indicadores faciais, comportamentais e fisiológicos (Liaw *et al.*, 2012);

k) **PIPP (*Premature Infant Pain Profile*)**: integra idade gestacional, estado comportamental, frequência cardíaca, saturação de oxigênio e três indicadores faciais (fronte saliente, olhos cerrados e sulco nasolabial aprofundado) (Stevens *et al.*, 1996);

l) **PIPP-R (*Premature Infant Pain Profile-Revised*)**: versão revisada da PIPP que mantém os indicadores fisiológicos, contextuais e faciais originais, com ajustes nos critérios de pontuação para diferentes níveis de prematuridade (Stevens *et al.*, 2014);

m) **P-MIPS (*Partial MIPS*)**: versão reduzida da MIPS baseada em um subconjunto de indicadores faciais e motores utilizados na avaliação da dor em lactentes (Buchholz *et al.*, 1998).

2.2 HUMAN-IN-THE-LOOP E IA CONFIÁVEL EM SISTEMAS CLÍNICOS

Em sistemas clínicos de apoio à decisão, HITL descreve uma organização sociotécnica na qual a saída algorítmica integra um processo decisório conduzido por pessoas, em vez de substituir esse processo por automação completa. O papel humano não se limita a confirmar uma recomendação, em outras palavras, o profissional deve interpretar a evidência à luz do contexto clínico, reconhecer situações não representadas nos dados, solicitar novas informações e desconsiderar a saída quando ela for inadequada. Essa complementaridade procura combinar a capacidade computacional de processamento sistemático e contínuo com o conhecimento contextual, a integração multimodal e a responsabilidade clínica humana (Holzinger, 2016; Shneiderman, 2020).

A mera presença de um profissional, contudo, não torna o sistema seguro nem caracteriza supervisão efetiva. A interação depende de informações que permitam avaliar a saída e de mecanismos que indiquem quando intervir (Amershi *et al.*, 2019). Para a avaliação da dor neonatal, uma classe binária isolada oferece pouco suporte a essa análise crítica. Probabilidades calibradas, estimativas de incerteza, explicações semanticamente organizadas, limiares operacionais e estados indeterminados podem fornecer elementos para decidir quando utilizar, questionar ou rejeitar uma predição. Outro risco relevante é o *automation bias*, entendido como a tendência de favorecer uma recomendação automática mesmo quando existem evidências conflitantes. Em sistemas clínicos de apoio à decisão, esse viés pode levar profissionais a abandonar avaliações inicialmente corretas diante de uma saída algorítmica incorreta (Goddard; Roudsari; Wyatt, 2012).

A incorporação crescente da IA à saúde vem sendo acompanhada por marcos regulatórios que tornam essas condições de supervisão parte de uma abordagem mais ampla de gestão de riscos. O Regulamento (UE) 2024/1689, *EU AI Act*, adota uma classificação baseada em risco e estabelece requisitos adicionais para sistemas de alto risco. No campo da saúde, essa qualificação não decorre do simples uso médico, mas pode alcançar sistemas que constituam ou integrem determinados dispositivos médicos, conforme a finalidade prevista e o enquadramento regulatório aplicável. Nesses casos, são enfatizadas a gestão de riscos, a qualidade dos dados, a documentação e a rastreabilidade, a transparência, a supervisão humana, a precisão, a robustez e o monitoramento ao longo do ciclo de vida (European Union, 2024).

No Brasil, a Lei n. 15.378/2026, embora não seja uma regulamentação específica de IA, institui o Estatuto dos Direitos do Paciente e reforça garantias diretamente relevantes ao uso de sistemas algorítmicos, como segurança, informação, confidencialidade, consentimento, autonomia e participação do paciente ou de seu representante no cuidado (Brasil, 2026). A Resolução CFM n. 2.454/2026, por sua vez, normatiza o uso da IA na medicina como apoio, e não substituição, da prática médica, preservando a autonomia e a responsabilidade do médico e enfatizando transparência, explicabilidade, validação científica, governança, proteção de dados, segurança e monitoramento. Ao requerer julgamento crítico das saídas e manter a decisão

final sob responsabilidade humana, a resolução fortalece normativamente o paradigma HITL (Conselho Federal de Medicina, 2026).

2.3 AVALIAÇÃO AUTOMÁTICA DA DOR NEONATAL

A avaliação automática da dor neonatal evoluiu de abordagens baseadas em descritores geométricos definidos manualmente para modelos de *deep learning* capazes de extrair representações diretamente dos dados. Zamzmi *et al.* (2018b) classificaram as técnicas em métodos de classificação e regressão. A maioria das bases de dados concentra-se em expressões faciais, devido à natureza não invasiva e à relativa facilidade de aquisição desse tipo de dados (Gkikas; Tsiknakis, 2023; Heiderich *et al.*, 2023).

Esta seção apresenta métodos de classificação baseados em imagens, vídeos e abordagens multimodais, bem como o papel crescente da explicabilidade nesse contexto.

2.3.1 Classificação em imagens

Os estudos pioneiros empregaram algoritmos clássicos, como análise de componentes principais (PCA), análise discriminante linear (LDA) e *support vector machine* (SVM). No banco *infant Classification of Pain Expression* (iCOPE), Brahnam *et al.* (2006) demonstraram que esses classificadores distinguiam expressões de dor com acurácia de até 88%. Posteriormente, o método *Neural Network Simultaneous Optimization Algorithm* (NNSOA), proposto por Brahnam *et al.* (2007), alcançou 90% de acurácia. Contudo, esses métodos são sensíveis a alterações de brilho e cor, bem como a faces fora do enquadramento. Os trabalhos também reportavam predominantemente a acurácia, métrica sujeita a vieses em bases desbalanceadas.

A análise baseada em pontos fiduciais foi explorada por Heiderich, Leslie e Guinsburg (2015) no banco da Universidade Federal de São Paulo (UNIFESP). Ao mapear 16 pontos faciais com limiares derivados da NFCS, o método alcançou alta sensibilidade e especificidade em ambiente controlado. Entretanto, a necessidade de calibração individual e a sensibilidade ao posicionamento da câmera limitaram sua aplicabilidade clínica.

Com a introdução das Redes Neurais Convolucionais (CNNs), Zamzmi *et al.* (2018a) investigaram arquiteturas da família *Visual Geometry Group* (VGG) como extratoras de características, combinadas a classificadores tradicionais, como *k-nearest neighbors* (KNN) e SVM. A configuração VGGFace + KNN atingiu 90% de acurácia, o que mostrou a relevância do pré-treinamento em faces humanas. Em seguida, Zamzmi *et al.* (2019) propuseram a *Neonatal Convolutional Neural Network* (NCNN), uma arquitetura rasa que obteve 91% de acurácia nos bancos iCOPE e *Neonatal Pain Assessment Dataset* (NPAD). O modelo superou a ResNet50, sugerindo que arquiteturas menos complexas podem ser mais adequadas a bases neonatais de tamanho reduzido.

Métodos de regressão para estimar a intensidade da dor foram explorados por Egede *et al.* (2019) e Salekin *et al.* (2020), que reportaram *Mean Absolute Error* (MAE) de 1,71 e

1,55 na escala *Neonatal Infant Pain Scale* (NIPS), respectivamente. Apesar desses resultados, a escassez de dados rotulados com escores contínuos ainda limita o desenvolvimento dessa linha de investigação.

Modelos mais recentes concentraram-se em estratégias de *fine-tuning* e em comparações arquiteturais. Carlini *et al.* (2021) obtiveram 89% de acurácia ao ajustar a VGGFace para os bancos iCOPE e UNIFESP. Coutrin *et al.* (2023) corroboraram a superioridade de modelos pré-treinados em faces (VGGFace) em relação a arquiteturas generalistas (ResNet, Inception) sob protocolos de validação cruzada mais rigorosos. Para lidar com oclusões e variações de pose, Zhao *et al.* (2023) propuseram o *Pose-invariant and Occlusion-robust Pain Assessment* (POPA), modelo que reconstrói uma vista frontal canônica antes da classificação, atingindo 89% de acurácia.

No que se refere à confiabilidade das predições, Ferreira *et al.* (2023) investigaram a calibração da NCNN. Embora a otimização de hiperparâmetros tenha elevado a acurácia de 78% para 83% nos dados iCOPE e UNIFESP, observou-se degradação da calibração probabilística, com aumento do *Expected Calibration Error* (ECE) de 0,035 para 0,091. Nesse experimento, o ganho de desempenho discriminativo foi acompanhado por maior excesso de confiança, condição que representa risco em aplicações clínicas.

Mais recentemente, Carlini *et al.* (2025) utilizaram o *Vision-Language Model* (VLM) Gemini 1.0 Pro Vision com diferentes estratégias de *prompting* para avaliar a dor neonatal. Os autores exploraram duas abordagens: (I) induzir o modelo a recuperar o conhecimento clínico latente adquirido durante o pré-treinamento e (II) orientá-lo com descrições explícitas de características faciais clinicamente relevantes. O uso do conhecimento clínico implícito produziu um equilíbrio entre precisão (82,3%) e sensibilidade (73,2%). A orientação por atributos faciais específicos alcançou precisão perfeita (100%), mas apresentou menor sensibilidade (40,1%).

Apesar dos avanços, persistem limitações estruturais em relação ao pequeno tamanho e à baixa diversidade das bases de dados, à ausência de padronização na aquisição e na rotulação dos dados (uso de diferentes escalas de dor) e à predominância de análises estáticas (imagens isoladas), que desconsideram a dinâmica temporal e a incerteza associadas às predições.

2.3.2 Classificação em vídeos

A análise temporal permite capturar a dinâmica das expressões faciais, embora sua adoção permaneça limitada pela escassez de bancos de vídeos anotados. O *infant Classification of Pain Expression Videos* (iCOPEVid) (Brahnam *et al.*, 2020) constitui uma das principais fontes disponíveis. Utilizando descritores de textura, como *local binary patterns* (LBP) e *histograms of oriented gradients* (HOG), combinados a SVMs nesse banco, Brahnam *et al.* (2020) reportaram uma *Area Under the Curve* (AUC) de 79%, superior à avaliação humana de 67%.

Em uma estratégia alternativa, Manworren *et al.* (2024) decompuseram vídeos em *frames* classificados individualmente, treinando modelos como *Random Forest* (AUC de 98%). Chen *et al.* (2023) avaliaram a viabilidade clínica do uso de vídeos e encontraram alta concordância

entre avaliações presenciais realizadas por médicos e avaliações remotas realizadas por sistemas inteligentes, apesar de desafios práticos, como oclusões e variações de iluminação em UTINs.

Mais recentemente, Ferreira *et al.* (2025) propuseram a análise das séries temporais de probabilidades geradas por modelos como NCNN, VGGFace e *Vision Transformer* (ViT-B/16), denominada "sinal da dor". Os autores mostraram que a variabilidade temporal da expressão facial contém informação clínica relevante e propuseram classificar as predições em quatro estados: estável, irregular, instável e indeterminado, ampliando a dicotomia estática predominante nas abordagens anteriores (com dor versus sem dor).

2.3.3 Classificação multimodal

A natureza multidimensional da dor favorece abordagens que integrem expressão facial, sinais fisiológicos e comportamento motor. Salekin *et al.* (2021) combinaram CNNs para processar face, corpo e espectrogramas de choro e utilizaram redes *long short-term memory* (LSTM) para a modelagem temporal, alcançando 88% de AUC. Mais recentemente, Yang *et al.* (2025) desenvolveram, em uma plataforma *low-code*, um modelo multimodal que integra informações faciais, fisiológicas e comportamentais. O modelo alcançou 95% de AUC, e os autores discutiram tanto os benefícios da fusão de dados quanto a facilidade de uso da plataforma por profissionais de saúde.

Entretanto, a complexidade da sincronização entre modalidades, a necessidade de múltiplos sensores e o custo computacional impõem barreiras à implementação prática desses sistemas em unidades de terapia intensiva com recursos limitados.

2.3.4 IA Explicável (XAI)

Carlini *et al.* (2021) foram pioneiros na aplicação de *Gradient-weighted Class Activation Mapping* (Grad-CAM) e *Integrated Gradients* a modelos de dor neonatal. As regiões de saliência observadas correspondiam a indicadores clínicos presentes em escalas de dor neonatal. Coutrin *et al.* (2023) ampliaram essa análise e mostraram que a VGGFace produz mapas de saliência mais concentrados em regiões faciais clinicamente relevantes do que arquiteturas treinadas na ImageNet, o que reforça a importância do pré-treinamento em faces.

Carlini *et al.* (2024) compararam mapas de saliência da NCNN e da VGGFace com dados de rastreamento ocular de profissionais de saúde. Embora os modelos utilizem regiões clinicamente relevantes, como olhos e boca, verificou-se divergência entre as áreas de foco máximo humano e computacional. Aprofundando essa análise, Pirie *et al.* (2026) quantificaram a similaridade entre esses dois grupos por meio das métricas *PixelAgreement* e *RegionAgreement*. Os autores observaram maior concordância entre humanos e modelos em instâncias com dor do que em instâncias sem dor, além de proporem uma análise quantificável em nível de regiões semânticas da face associadas a escalas clínicas de dor, que supera a interpretabilidade de abordagens puramente em nível de pixel.

Ferreira *et al.* (2027) ampliaram essa linha de investigação ao avaliar dez métodos de explicação *post-hoc* nas arquiteturas NCNN, VGGFace e ViT-B/32, considerando a concordância em níveis de pixel e de regiões semânticas, a fidelidade por perturbação de pixels e o alinhamento com 13 escalas clínicas de dor neonatal. Os resultados de fidelidade mostraram-se sensíveis à estratégia de perturbação utilizada e aos modelos, não havendo um único explicador universalmente fiel ao comportamento de todos os modelos. Além disso, a concordância com as escalas clínicas variou conforme o modelo. O nariz, porém, foi frequentemente reportado como importante para os modelos, embora não conste nas 13 escalas de dor analisadas.

2.4 LACUNAS NA LITERATURA

À luz do estado da arte e dos requisitos de confiabilidade discutidos, a revisão identifica lacunas que dificultam a translação tecnológica para a prática clínica. A predominância de análises estáticas desconsidera a evolução temporal da dor, aspecto central para o monitoramento contínuo. Mesmo nas abordagens baseadas em vídeo, as saídas tendem a assumir a forma de classificações binárias, insuficientes para representar a complexidade do fenômeno.

No campo da XAI, embora avaliações recentes tenham ampliado a diversidade de explicadores e a análise de sua fidelidade, a utilidade clínica efetiva dessas explicações permanece pouco investigada. De modo semelhante, a calibração probabilística e a quantificação da incerteza permanecem amplamente fora do escopo dos estudos. A maioria das pesquisas prioriza a maximização de métricas discriminativas, sem verificar se os modelos estão aptos a serem utilizados em ambientes críticos, como UTINs, onde erros podem ter consequências graves.

A heterogeneidade metodológica entre bancos de dados também dificulta o estabelecimento de *benchmarks* comparáveis e reprodutíveis. Além disso, permanece pouco explorada a transformação de predições pontuais em um sinal temporal que explicite não apenas a probabilidade de dor, mas também a confiabilidade, a explicabilidade e os limites de validade das saídas.

A metodologia proposta nesta tese aborda essas lacunas ao integrar análise temporal, calibração, quantificação da incerteza, explicabilidade e limiares operacionais na construção de um sinal contínuo da dor, em um arcabouço unificado e agnóstico ao banco de dados. Alguns dos estudos citados nesta seção (Ferreira *et al.*, 2023; Ferreira *et al.*, 2025; Pirie *et al.*, 2026; Ferreira *et al.*, 2027) apresentam resultados diretos desta tese e estão reproduzidos na íntegra nos Apêndices F a J.

Sob a perspectiva HITL, a lacuna não se restringe apenas à qualidade do classificador, deve-se avaliar também como a saída será incorporada ao julgamento humano e como o desempenho conjunto humano–IA será desenvolvido. Assim, a passagem da classificação experimental para o apoio clínico requer o estudo da apresentação da informação, da compreensão dos limites, da carga cognitiva e dos efeitos sobre a decisão profissional, dimensões que permanecem fora do escopo da maior parte dos trabalhos revisados.

Tabela 1 – Síntese dos trabalhos relacionados e suas limitações.

continua

Trabalho	**Base / Tipo**	**Método**	**Desempenho**	**Limitação principal**
Brahnam *et al.* (2006)	iCOPE (Imagem)	SVM com PCA	Acu: 88%	Extração manual de características.
Brahnam *et al.* (2007)	iCOPE (Imagem)	NNSOA	Acu: 90%	Extração manual de características.
Heiderich, Leslie e Guinsburg (2015)	UNIFESP (Imagem)	Pontos fiduciais (NFCS) e dist. euclidiana	Sens: 100% Esp: 100%	Requer calibração manual, altamente sensível à pose e oclusão.
Zamzmi *et al.* (2018a)	Particular (Imagem)	VGGFace + KNN	Acu: 90% AUC: 84%	Uso de características pré-treinadas sem *fine-tuning*.
Zamzmi *et al.* (2019)	iCOPE, NPAD (Imagem)	NCNN	Acu: 91%	Não reporta confiabilidade das predições.
Egede *et al.* (2019)	APN-db (Imagem)	Híbrido (*Handcrafted* + DL)	MAE: 1,71	Base de dados restrita.
Salekin *et al.* (2020)	Particular (Vídeo)	VGG16 + *Recurrent Neural Network* (RNN)	MAE: 1,55	Base proprietária e arquitetura complexa de baixa interpretabilidade.
Brahnam *et al.* (2020)	iCOPEVid (Vídeo)	Textura + SVM	AUC: 79%	Abordagem baseada em vetores de características estáticos, ignorando dinâmica temporal.

Tabela 1 – Síntese dos trabalhos relacionados e suas limitações.

continuação

Trabalho	**Base / Tipo**	**Método**	**Desempenho**	**Limitação principal**
Carlini *et al.* (2021)*	UNIFESP, iCOPE (Imagem)	VGGFace (*Transfer Learning*) + XAI (Grad-CAM e Integrated Gradients)	Acu: 89% AUC: 86%	Carece de métricas quantitativas de fidelidade da explicação.
Salekin *et al.* (2021)	USF-MNPAD-I (Multimodal)	VGG16 + LSTM	Acu: 80% AUC: 88%	Alto custo computacional e complexidade de fusão de dados em ambiente clínico.
Coutrin *et al.* (2023)*	UNIFESP, iCOPE (Imagem)	VGGFace, ResNet, NCNN, SENet + XAI (Grad-CAM)	Acu: 86% AUC: 85%	Foco em métricas de classificação pura, sem considerar a calibração das probabilidades.
Zhao *et al.* (2023)	Particular (Imagem)	Encoder/Decoder (Correção de Pose)	Acu: 89%	Baixa explicabilidade de como a face foi reconstruída.
Ferreira *et al.* (2023)*	iCOPE (Imagem)	NCNN otimizada + análise de calibração + XAI (Grad-CAM e Integrated Gradients)	Acu: 83% ECE: 0,091	Sem avaliação de métodos de calibração.
Chen *et al.* (2023)	Particular (598 RNs) (Vídeo)	Comparação NIPS presencial vs. vídeo + três métodos de IA	Concord.: 96,98% κ: 0,926	Estudo unicêntrico com vídeos limitados pelo ambiente, pela equipe e pelo ângulo de captura.
Carlini *et al.* (2024)*	UNIFESP (Imagem)	Comparação XAI (Grad-CAM e Integrated Gradients) vs. atenção humana	Similaridade parcial em olhos/boca	Uso qualitativo de apenas dois explicadores, sem métricas quantitativas de fidelidade.

Tabela 1 – Síntese dos trabalhos relacionados e suas limitações.

continuação

Trabalho	**Base / Tipo**	**Método**	**Desempenho**	**Limitação principal**
Manworren *et al.* (2024)	iCOPEVid (Vídeo)	Software proprietário + Random Forest	AUC: 98%	Dependência de ferramenta comercial (H2AI).
Ferreira *et al.* (2025)*	iCOPEVid (Vídeo)	VGGFace, NCNN, ViT-B/16 + Série temporal + XAI (Grad-CAM e Integrated Gradients)	Identificação de padrões estáveis/instáveis	Ausência de treinamento temporal.
Yang *et al.* (2025)	Particular (Multimodal)	Modelo multimodal *low-code* (face + fisiologia + comportamento)	AUC: 95%	Complexidade de aquisição multissensorial e baixa reprodutibilidade por plataforma proprietária.
Carlini *et al.* (2025)*	UNIFESP, iCOPE (Imagem)	VLM (Gemini 1.0 Pro Vision) + Prompt detalhado com instruções da NFCS	Prec.: 100,0% Sens.: 40,1%	Uso de apenas um modelo VLM, sem base de comparação com outros modelos do mesmo tipo.
Pirie *et al.* (2026)*	UNIFESP (Imagem)	Comparação XAI (Grad-CAM e Integrated Gradients) vs. atenção humana + Métricas PixelAgreement e RegionAgreement	Maior concordância em instâncias de dor	Estudo limitado à comparação com dois explicadores (Grad-CAM e Integrated Gradients).

Tabela 1 – Síntese dos trabalhos relacionados e suas limitações.

conclusão

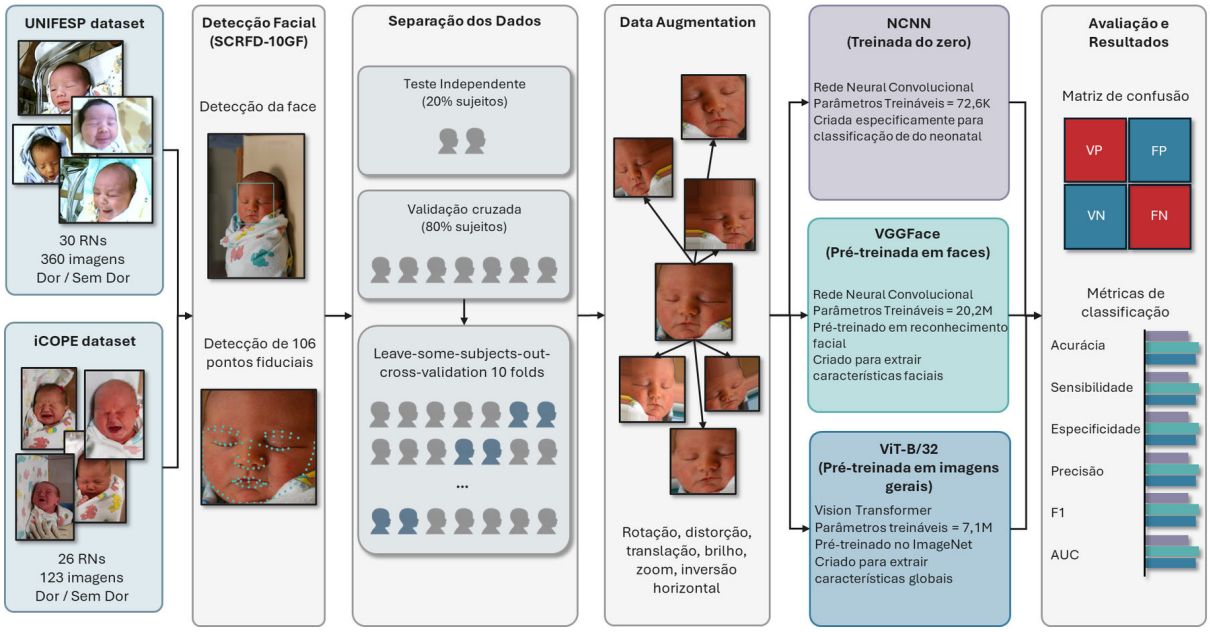
Trabalho	**Base / Tipo**	**Método**	**Desempenho**	**Limitação principal**
Ferreira *et al.* (2027)*	iCOPE, UNIFESP (Imagem)	NCNN, VGGFace e ViT-B/32 + 10 explicadores + concordância, fidelidade (AOPC) e alinhamento com 13 escalas clínicas	Perturbações de pixel realistas contribuíram positivamente para a avaliação da fidelidade dos explicadores. A concordância com as escalas neonatais variou conforme o modelo	Não foi explorada a agregação de explicadores em um consenso global.

Fonte: Autor.

Nota: * Trabalhos com participação do autor desta tese.

3 TREINAMENTO DOS MODELOS COMPUTACIONAIS

Figura 1 – Visão geral da metodologia do Capítulo 3.



Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

3.1 INTRODUÇÃO

Este capítulo apresenta a primeira etapa experimental da tese e estabelece a base discriminativa necessária à construção do sinal da dor (Figura 1). Sobre essa base serão desenvolvidos, nos capítulos seguintes, os pilares de confiabilidade, explicabilidade e segurança operacional que qualificam as probabilidades utilizadas na representação temporal. Como apresentado no Capítulo 2, sistemas de IA procuram executar tarefas tradicionalmente associadas à cognição humana (Russell; Norvig, 2009). Na detecção da dor neonatal, o desafio consiste em desenvolver modelos que identifiquem expressões faciais associadas à dor com precisão e confiança suficientes para apoiar a prática clínica.

O objetivo deste capítulo é desenvolver e avaliar modelos de *deep learning* para classificar a expressão facial neonatal nas categorias *com dor* e *sem dor*. Para isso, são descritos os bancos de dados, o pré-processamento das imagens, as arquiteturas investigadas, o protocolo experimental e as métricas utilizadas. Esses elementos fornecem o ponto de partida para responder à primeira pergunta de pesquisa: em que medida modelos de *deep learning* conseguem distinguir, de forma consistente, a presença e a ausência de dor a partir da expressão facial neonatal?

Entre as abordagens de visão computacional estão as *Convolutional Neural Networks* (CNNs) e os *Vision Transformers* (ViTs). As CNNs são amplamente utilizadas na análise de

imagens médicas e faciais por explorarem a estrutura espacial dos dados por meio de convoluções (Litjens *et al.*, 2017). Os ViTs, por sua vez, empregam mecanismos de atenção global para modelar relações entre diferentes regiões da imagem (Dosovitskiy *et al.*, 2020). Comparar essas famílias arquiteturais permite avaliar se diferentes formas de representação visual produzem desempenho adequado à detecção da dor neonatal.

Neste trabalho, essas arquiteturas são avaliadas em um cenário de aprendizado supervisionado. Os métodos e materiais apresentados neste capítulo apoiam-se em publicações anteriores (Zamzmi *et al.*, 2018a; Carlini *et al.*, 2021; Coutrin *et al.*, 2023).

3.1.1 Convolutional Neural Networks

As *Convolutional Neural Networks* (CNNs) constituem uma classe de redes neurais projetadas para processar dados com estrutura espacial explícita, como imagens biomédicas e faciais. Esses modelos empregam operações de convolução para extrair características relevantes de imagens (LeCun *et al.*, 1998; Goodfellow; Bengio; Courville, 2016) e são amplamente utilizados em tarefas de classificação, segmentação e detecção de objetos ou padrões visuais.

Uma característica das CNNs é a aprendizagem hierárquica de representações. As camadas iniciais tendem a identificar padrões simples, como bordas e texturas, enquanto as mais profundas capturam estruturas complexas, incluindo formas anatômicas e configurações faciais. Essa propriedade favoreceu a ampla aplicação das CNNs na análise de imagens médicas e no reconhecimento facial (Litjens *et al.*, 2017).

Uma arquitetura típica de CNN inclui:

a) **camadas convolucionais**: aplicam filtros (*kernels*) que percorrem a entrada para identificar padrões locais aprendidos durante o treinamento;

b) **camadas de *pooling***: reduzem a dimensionalidade espacial dos mapas de características, preservando informações relevantes e diminuindo a complexidade computacional;

c) **camadas totalmente conectadas**: integram as características extraídas e produzem a saída final, usualmente associada a tarefas de classificação ou regressão;

d) **funções de ativação**: introduzem não linearidade no modelo. Entre as mais utilizadas estão *Rectified Linear Unit* (ReLU), *Sigmoid* e *Softmax*.

3.1.2 *Vision transformers*

Os *Transformers* foram originalmente propostos para tarefas de processamento de linguagem natural, com o objetivo de modelar dependências em sequências de texto por meio de mecanismos de atenção (Vaswani *et al.*, 2017). Posteriormente, essa arquitetura foi adaptada para a visão computacional, dando origem aos *Vision Transformers* (ViTs) (Dosovitskiy *et al.*, 2020). Diferentemente das CNNs, que extraem padrões locais por meio de convoluções, os ViTs

empregam atenção global desde as camadas iniciais, capturando interações complexas entre diferentes regiões da imagem.

Em um ViT, a imagem de entrada é dividida em blocos fixos (*patches*), linearmente projetados em vetores de dimensão fixa e combinados com *embeddings* posicionais, responsáveis por preservar a informação espacial. Esses vetores são processados por múltiplas camadas *Transformer*, nas quais cada *patch* interage com os demais de forma contextual, o que favorece a modelagem de relações espaciais globais.

Uma arquitetura típica de ViT compreende:

a) **divisão em *patches***: segmentação da imagem em blocos de tamanho fixo (por exemplo, 16×16 ou 32×32 pixels);

b) **projeção linear**: transformação de cada *patch* em um vetor de dimensão fixa por meio de uma camada linear (*patch embedding*);

c) ***Positional embeddings***: adicionados para preservar a informação espacial, uma vez que o *Transformer* é invariante à ordem dos blocos;

d) **token de classificação (`[CLS]`)**: vetor especial utilizado em tarefas de classificação;

e) **blocos de *Transformer***: compostos por autoatenção multi-cabeça, normalização em camada, perceptrons multicamadas (MLPs) e conexões residuais;

f) **camada de classificação**: processa a representação associada ao token `[CLS]` por meio de camadas totalmente conectadas e uma função de ativação final (por exemplo, *Softmax* ou *Sigmoid*).

Embora os ViTs demandem grandes volumes de dados e maior capacidade computacional, evidências indicam que, quando pré-treinados em bases extensas e adequadamente ajustados, podem superar CNNs em diversas tarefas visuais (Dosovitskiy *et al.*, 2020; Touvron *et al.*, 2021; Khan *et al.*, 2022).

3.2 MATERIAIS

Foram empregados dois bancos de imagens faciais neonatais: UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006; Brahnam *et al.*, 2007). O banco UNIFESP não é de acesso público, enquanto o iCOPE foi disponibilizado mediante assinatura de termo de uso. A Tabela 2 apresenta as principais características de cada banco.

3.2.1 Banco de imagens UNIFESP

O banco de imagens faciais da Universidade Federal de São Paulo (UNIFESP) foi desenvolvido para apoiar a criação de um *software* de detecção automática de expressões faciais de dor em RNs (Heiderich; Leslie; Guinsburg, 2015).

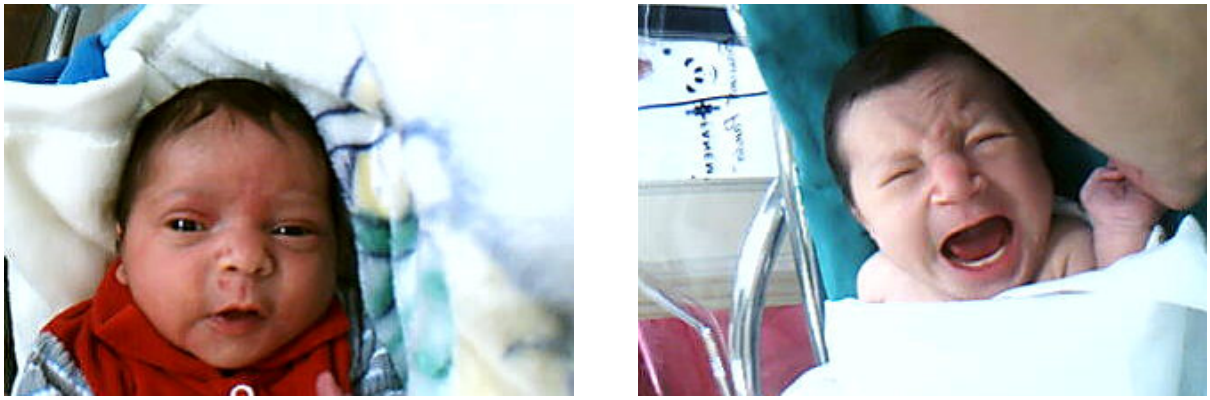
Foram incluídos 30 RNs, com idades entre 24 e 144 horas, classificados como prematuros tardios ou a termo, internados no Hospital São Paulo da Escola Paulista de Medicina/UNIFESP

Tabela 2 – Características dos bancos de dados utilizados.

Banco de dados	**iCOPE**	**UNIFESP**
Tipo de dados	Imagem	Imagem
Nº de RNs	26	30
Masculino	13	15
Feminino	13	15
Idade pós-natal (h)	18–72	24–144
Resolução da câmera	3008×2000	320×233
Total de amostras	123	360
Dor	60	162
Sem dor	63	198

Fonte: Autor.

Figura 2 – Exemplos de imagens do banco UNIFESP.



Fonte: Banco de imagens de face da UNIFESP (Heiderich; Leslie; Guinsburg, 2015).

(Figura 2). Nenhum apresentava histórico de doenças ou malformações congênitas. A coleta foi aprovada pelo Comitê de Ética em Pesquisa da UNIFESP (protocolo nº 1299/09).

Três câmeras Foscam, com resolução de 320×233, registraram vídeos de 10 minutos antes, durante e após procedimentos dolorosos, como injeções intramusculares e punções venosas. A partir desses vídeos, extraiu-se uma imagem a cada 3 segundos, totalizando 360 imagens (138 antes, 30 durante e 192 após os procedimentos). Todas as imagens foram avaliadas por especialistas de acordo com a escala NFCS, resultando em 162 imagens rotuladas como *com dor* e 198 como *sem dor*.

3.2.2 Banco de imagens iCOPE

O banco iCOPE (*infant Classification of Pain Expression*) constitui um dos primeiros esforços voltados ao uso de aprendizado de máquina para o reconhecimento automático da dor neonatal (Brahnam *et al.*, 2006; Brahnam *et al.*, 2007).

Esse banco contém 200 imagens de 26 RNs (Figura 3), com idades entre 18 horas e 3 dias. Todos estavam em bom estado de saúde no momento da coleta, embora seis tenham sido

Figura 3 – Exemplos de imagens do banco iCOPE.




Fonte: Banco de imagens iCOPE (Brahnam *et al.*, 2006; Brahnam *et al.*, 2007).

submetidos à circuncisão no dia anterior. A aquisição seguiu protocolos éticos aprovados pelo St. John's Hospital (EUA). As imagens foram capturadas com uma câmera Nikon D100, com resolução de 3008×2000 pixels, em ambiente controlado.

Cada RN foi exposto a diferentes estímulos: repouso (63 imagens), choro (18), ar (23), fricção (36) e dor (60). Neste trabalho, consideraram-se apenas as classes "repouso" e "dor", de modo a manter alinhamento com o banco UNIFESP.

3.2.3 Hardware e software

Foram utilizados dois ambientes computacionais distintos ao longo desta pesquisa: (I) um ambiente de alto desempenho destinado ao treinamento dos modelos de *deep learning*, representado pelo supercomputador Santos Dumont (SDumont)[1], localizado no Laboratório Nacional de Computação Científica (LNCC), em Petrópolis, Rio de Janeiro; e (II) um computador pessoal empregado na inferência, nos testes de protótipos e na execução de códigos auxiliares. As configurações de cada ambiente são apresentadas na Tabela 3.

Tabela 3 – Configuração dos ambientes computacionais.

Configuração	**SDumont**	**Computador pessoal**
CPU	2x Intel Xeon Skylake 6252 (48 cores)	Intel Core i5-13600K (14 cores)
RAM	384 GB	32 GB
GPU	4x NVIDIA V100 (32 GB VRAM cada)	NVIDIA RTX 4070 Ti (12 GB VRAM)

Fonte: Autor.

Todos os métodos computacionais descritos nesta tese foram implementados em Python 3.10, com a biblioteca PyTorch 2.6 (Paszke *et al.*, 2019) e pacotes complementares para análise e visualização de dados, como NumPy, Pandas, Matplotlib e Captum. Os códigos desenvolvidos estão disponíveis publicamente em

[1] https://sdumont.lncc.br/index.php

https://github.com/darkrubiks/AI-Assessment-of-Pain-in-Newborns, com o objetivo de favorecer a reprodutibilidade dos experimentos.

Declara-se que ferramentas de Inteligência Artificial Generativa foram utilizadas como suporte em etapas específicas desta pesquisa. O modelo ChatGPT-6.1 Sol foi empregado para desenvolver os códigos de programação utilizados na análise dos dados, bem como para dar suporte à revisão gramatical e ao aprimoramento da clareza do texto dos capítulos teóricos.

Ressalta-se que a inteligência artificial atuou estritamente como ferramenta de apoio operacional e computacional. Todo o código gerado foi testado, e a tese passou por rigorosa checagem e validação manual pelo autor.

A concepção metodológica, a interpretação dos resultados e a redação intelectual do trabalho são de autoria estritamente humana. O autor assume integral e exclusiva responsabilidade científica, ética e legal pelos resultados, análises, discussões, conclusões e teor do texto apresentado nesta tese.

3.3 MÉTODOS

3.3.1 Detecção facial

A detecção facial foi realizada com o *Sample and Computation Redistribution for Efficient Face Detection* (SCRFD-10GF) (Guo *et al.*, 2021), um detector de alto desempenho que retorna as coordenadas da face, 106 pontos fiduciais 2D e 68 pontos fiduciais 3D. A implementação está disponível na biblioteca InsightFace[2].

O SCRFD-10GF foi selecionado por combinar elevado desempenho e baixo custo computacional, apresentando tempo médio de inferência de 4,9 ms, em comparação aos aproximadamente 30 ms de detectores como o *Dual Shot Face Detector* (DSFD), RetinaFace, HAMBox e TinaFace (Guo *et al.*, 2021). As imagens foram redimensionadas para 640×640 pixels, conforme o padrão adotado no pré-treinamento do modelo. Após essa etapa, as faces detectadas foram recortadas das imagens originais, sem redimensionamento, em janelas quadradas para evitar distorções, e os 106 pontos fiduciais foram armazenados em arquivos `pkl` (pickle) individuais, para utilização posterior (Figura 4).

3.3.2 Modelos de classificação

Foram empregados três modelos de aprendizado profundo: NCNN (Zamzmi *et al.*, 2019), VGGFace (Parkhi; Vedaldi; Zisserman, 2015) e ViT-B/32 (Dosovitskiy *et al.*, 2020). A VGGFace foi selecionada com base em seu desempenho superior em trabalhos anteriores (Coutrin *et al.*, 2023). A NCNN foi incluída por ter sido desenvolvida especificamente para a detecção de dor em RNs (Zamzmi *et al.*, 2019). O ViT-B/32 foi adotado para avaliar a metodologia em uma arquitetura distinta das CNNs, representativa de modelos recentes baseados em atenção.

[2] https://github.com/deepinsight/insightface/tree/master/python-package

Figura 4 – Detecção facial e pontos fiduciais extraídos pelo SCRFD-10GF.



(a) Caixa delimitadora da face.

(b) Pontos fiduciais detectados.

Fonte: Autor, utilizando banco de imagens iCOPE (Brahnam *et al.*, 2006).

3.3.2.1 NCNN

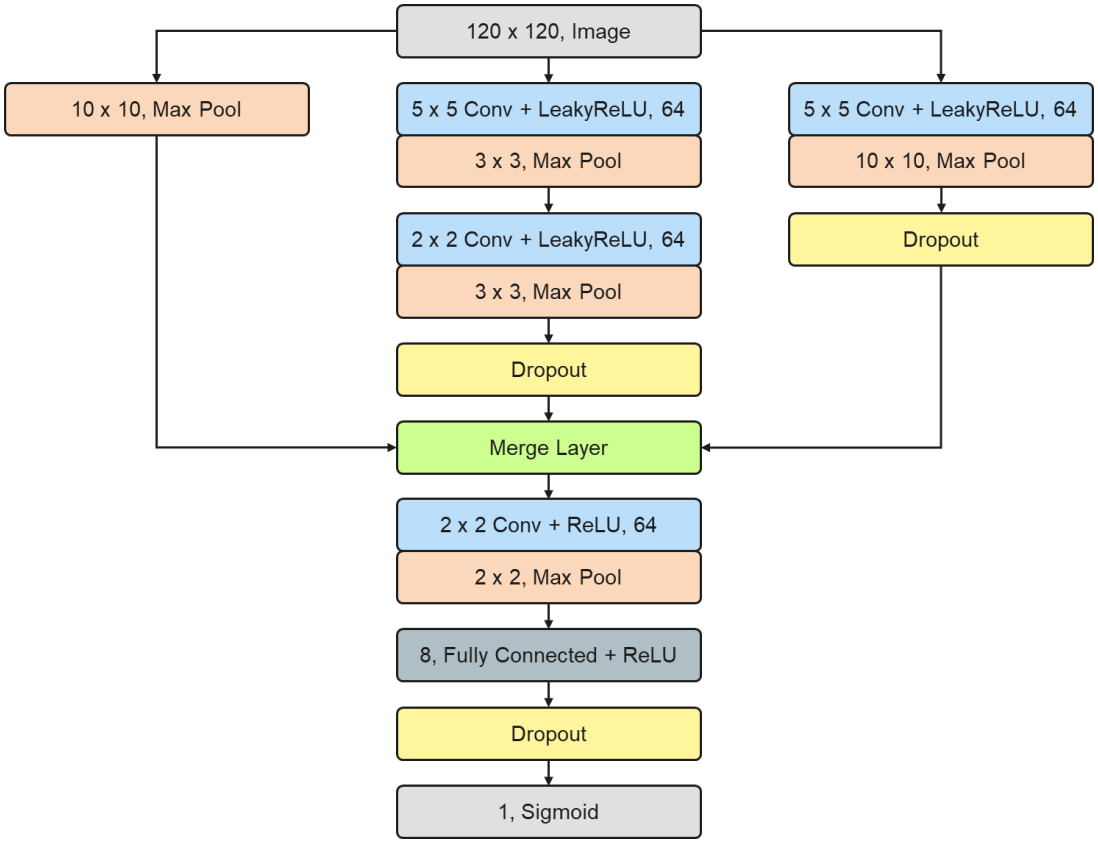
A *Neonatal Convolutional Neural Network* (NCNN), proposta por Zamzmi *et al.* (2019), foi o primeiro modelo de *deep learning* projetado especificamente para a detecção de dor em RNs. Sua arquitetura compreende três ramos convolucionais em cascata, cujas saídas são concatenadas e encaminhadas para um único neurônio com ativação *Sigmoid* (Figura 5). Como o modelo original não foi disponibilizado publicamente, a NCNN foi treinada do zero. No pré-processamento, as imagens foram redimensionadas para 120×120 pixels em formato RGB.

3.3.2.2 VGGFace

A VGGFace é uma CNN derivada da VGG16, treinada em 2,6 milhões de imagens de 2622 indivíduos (Parkhi; Vedaldi; Zisserman, 2015). O modelo consolidou-se como um extrator robusto de atributos faciais em estudos de detecção da dor neonatal (Zamzmi *et al.*, 2018a; Carlini *et al.*, 2021).

Neste trabalho, adotou-se a estratégia descrita por Carlini *et al.* (2021), adicionando duas camadas densas de 512 neurônios com Dropout de 50% às camadas convolucionais pré-treinadas (Figura 6). A saída foi ajustada para um único neurônio com ativação *Sigmoid*. As imagens

Figura 5 – Arquitetura da NCNN.



Fonte: Autor.

foram redimensionadas para 224×224 pixels, convertidas de RGB para BGR e centralizadas pela média dos canais, conforme o procedimento adotado no pré-treinamento original do modelo (Parkhi; Vedaldi; Zisserman, 2015).

3.3.2.3 ViT-B/32

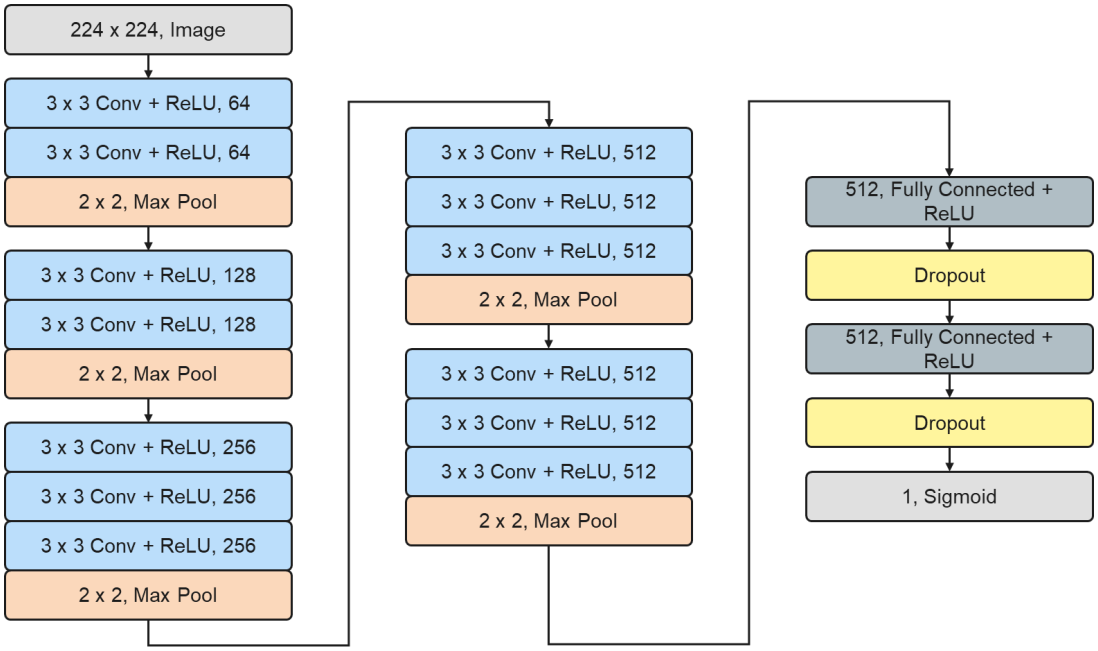
O ViT-B/32 segue a arquitetura *Vision Transformer* proposta por Dosovitskiy *et al.* (2020). Optou-se pela versão Base, com *patches* de 32×32 pixels, buscando a preservação estrutural das regiões faciais (Figura 7). O modelo foi pré-treinado no ImageNet, uma vez que modelos ViT pré-treinados especificamente em faces não estavam disponíveis publicamente. As imagens foram redimensionadas para 224×224 pixels e normalizadas com os valores médios do conjunto de pré-treinamento (Dosovitskiy *et al.*, 2020).

3.3.3 Métricas de classificação

As métricas de classificação quantificam a capacidade do modelo de produzir predições corretas. Em tarefas binárias, os resultados são organizados em uma matriz de confusão, apresentada na Tabela 4.

A partir de VP, VN, FP e FN derivam-se as seguintes métricas:

Figura 6 – Arquitetura da VGGFace.



Fonte: Autor.

Tabela 4 – Matriz de confusão em classificação binária.

	Predito: Positivo	**Predito: Negativo**
Real: Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Real: Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Autor.

a) **acurácia:** proporção de amostras corretamente classificadas.

$$\text{Acurácia} = \frac{VP + VN}{VP + FP + VN + FN} \tag{3.1}$$

b) **sensibilidade:** proporção de casos positivos corretamente identificados.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{3.2}$$

c) **especificidade:** proporção de casos negativos corretamente identificados.

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{3.3}$$

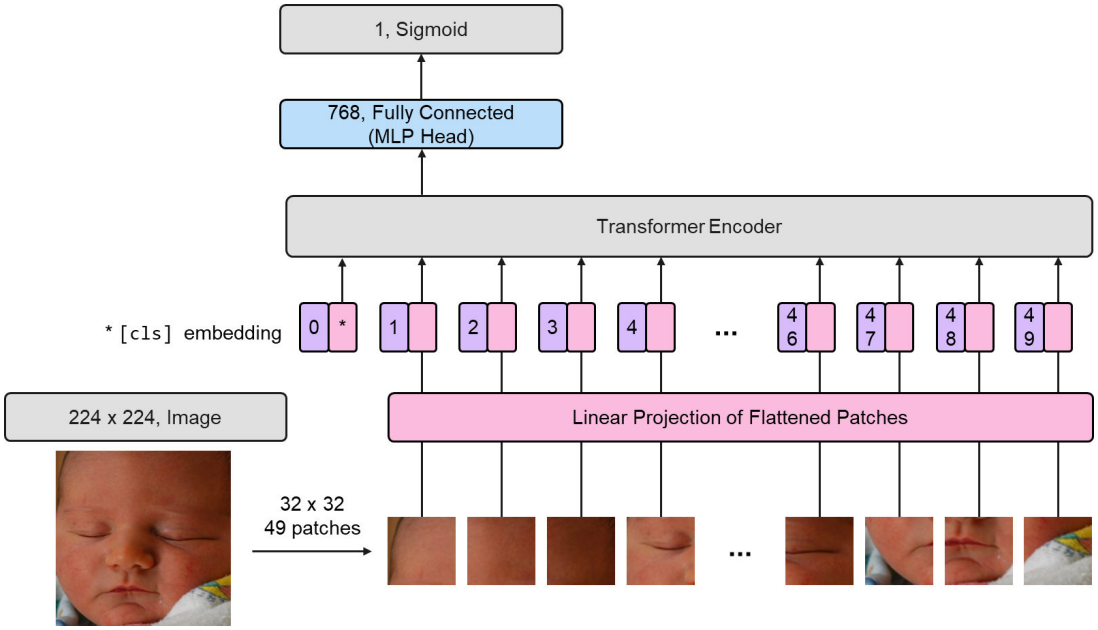
d) **precisão:** proporção de predições positivas corretas.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{3.4}$$

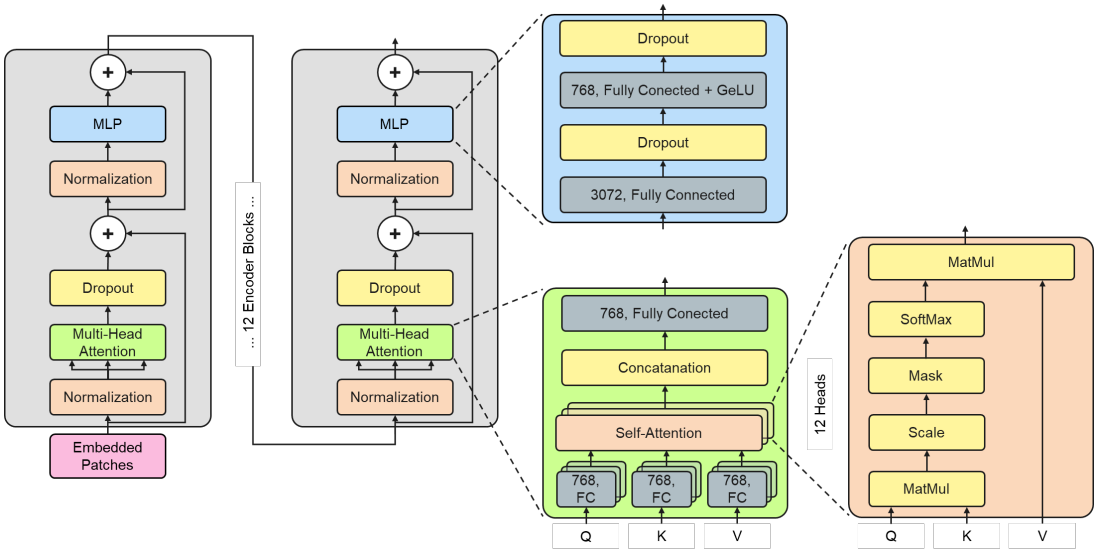
e) **F1-Score:** média harmônica entre precisão e sensibilidade.

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \tag{3.5}$$

Figura 7 – Arquitetura do ViT-B/32.



(a) Visão geral da arquitetura.



(b) Visão detalhada dos blocos *transformer*.

Fonte: Autor, utilizando banco de imagens iCOPE (Brahnam *et al.*, 2006).

f) **AUC:** medida da capacidade discriminativa do modelo ao longo de diferentes limiares de decisão. Na equação, TVP representa a Taxa de Verdadeiros Positivos e TFP, a Taxa de Falsos Positivos.

$$\text{AUC} = \int_0^1 TVP(TFP^{-1}(x))\,dx \tag{3.6}$$

Essas métricas assumem valores entre 0 e 1. A escolha da métrica prioritária depende do contexto de aplicação. Em cenários clínicos, costuma-se priorizar alta sensibilidade, a fim de reduzir a probabilidade de falsos negativos, que resultam em falhas na identificação de dor e, consequentemente, na falta de intervenção adequada.

3.4 EXPERIMENTO

O experimento deste capítulo teve como objetivo quantificar o desempenho discriminativo inicial dos modelos NCNN, VGGFace e ViT-B/32, estabelecendo o *baseline* da tese.

Inicialmente, os bancos de imagens iCOPE e UNIFESP foram combinados, totalizando 483 imagens de 56 sujeitos. Apenas as imagens rotuladas como *com dor* (222) e *sem dor* (261) foram consideradas. Todas passaram pela etapa de detecção facial e, posteriormente, foram organizadas nos conjuntos de treinamento, validação e teste.

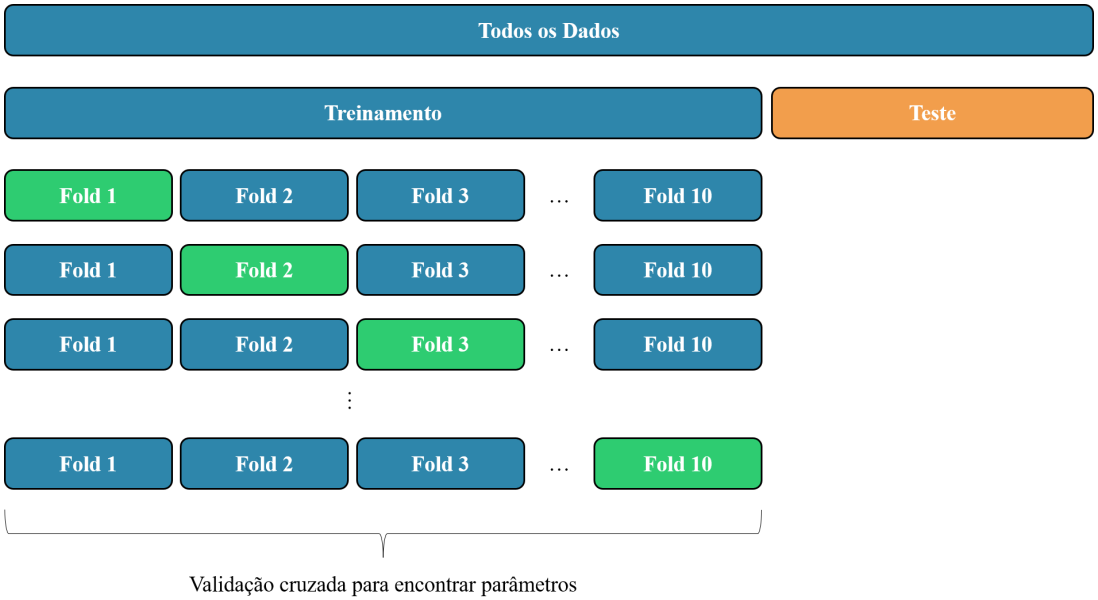
Foram separados 20% dos indivíduos para compor um conjunto de teste independente, enquanto os 80% restantes foram utilizados em validação cruzada segundo a estratégia *leave-some-subjects-out*, conforme Coutrin *et al.* (2023). Essa estratégia assegura que imagens de um mesmo sujeito não sejam utilizadas simultaneamente em treinamento e validação, prevenindo o vazamento de dados (Figura 8). Além disso, foi garantida a estratificação dos bancos iCOPE e UNIFESP nos conjuntos de treinamento, validação e teste.

A etapa de *Data Augmentation*, ilustrada na Figura 9, foi aplicada exclusivamente ao conjunto de treinamento. Para cada amostra, foram geradas 20 imagens adicionais por meio de transformações aleatórias: rotação ($\pm 30°$), distorção (0,15), deslocamento (0,20), ajuste de brilho (0,50–1,10), zoom (0,70–1,50) e inversão horizontal. O detector facial SCRFD-10GF verificou a presença do rosto em todas as imagens geradas. Quando a face não era detectada, novas amostras eram produzidas até que as 20 imagens aumentadas contivessem o rosto do neonato (Coutrin *et al.*, 2023).

O treinamento do modelo utilizou a função de perda *Cross-Entropy* para monitorar a convergência entre treinamento e validação. A cada época em que o erro de validação diminuía, um *checkpoint* do modelo era salvo. Caso não houvesse melhora por cinco épocas consecutivas, o treinamento era interrompido por parada antecipada. Assim, apenas o *checkpoint* com o menor erro de validação foi mantido para as etapas posteriores, reduzindo o risco de sobreajuste.

Para a VGGFace e o ViT-B/32, quando o erro de validação não melhorava por cinco épocas, iniciava-se o *fine-tuning*. Nessa etapa, foram treinadas as duas últimas camadas convolucionais da VGGFace e os dois últimos blocos *transformer* do ViT-B/32. A taxa de aprendizado

Figura 8 – Processo de validação cruzada.



Fonte: Autor.

Figura 9 – Imagens após *Data Augmentation*.



Fonte: Autor, utilizando banco de imagens iCOPE (Brahnam *et al.*, 2006).

foi reduzida de $\eta_{\text{pre-fine-tuning}} = 1 \times 10^{-4}$ para $\eta_{\text{durante-fine-tuning}} = 1 \times 10^{-6}$, a fim de evitar alterações abruptas nos pesos pré-treinados responsáveis pela extração de características de alto nível. Os demais hiperparâmetros e parâmetros são apresentados na Tabela 5.

Tabela 5 – Hiperparâmetros e parâmetros utilizados no treinamento dos modelos.

Configuração	**NCNN**	**VGGFace**	**ViT-B/32**
Tamanho da entrada	120×120 (RGB)	224×224 (BGR)	224×224 (RGB)
FLOPs	326,1 MFLOPS	30,75 GFLOPS	6,37 GFLOPS
Parâmetros treináveis	72,6 K	20,2 M	7,1 M
Pré-treinamento	Nenhum	VGGFace dataset	ImageNet
Função de ativação de saída	Sigmoid	Sigmoid	Sigmoid
Função de perda	Cross-Entropy	Cross-Entropy	Cross-Entropy
Batch size	16	16	16
Taxa de aprendizado inicial	1×10^{-4}	1×10^{-4}	1×10^{-4}
Taxa de aprendizado *fine-tuning*	–	1×10^{-6} (2 últimas conv.)	1×10^{-6} (2 últimos blocos)
Regularização	$L_2 = 0,01$	$L_1 = 5 \times 10^{-4}$	–
Dropout	10%	50%	0%
Parada antecipada	5 épocas sem melhora	5 épocas sem melhora	5 épocas sem melhora

Fonte: Autor.

Para comparar as arquiteturas NCNN, VGGFace e ViT-B/32, empregou-se o teste não paramétrico de Friedman (Friedman, 1937), considerando os mesmos *folds* como blocos pareados. Adotou-se um nível de significância de 5% ($p < 0{,}05$).

3.5 RESULTADOS

A Tabela 6 apresenta a média e o desvio-padrão das métricas de classificação dos três modelos durante a validação cruzada. De modo geral, os modelos apresentaram desempenho semelhante.

A VGGFace obteve as maiores médias de acurácia ($84,2\% \pm 4,8$), F1-Score ($82,0\% \pm 5,9$), sensibilidade ($80,8\% \pm 8,8$), precisão ($84,4\% \pm 9,5$) e AUC ($91,5\% \pm 3,3$). O ViT-B/32 apresentou resultados intermediários e alcançou a maior especificidade ($88,1\% \pm 7,2$). A NCNN apresentou os menores valores, sobretudo de sensibilidade ($72,5\% \pm 16,2$), o que indica maior dificuldade para identificar casos de dor. Ainda assim, sua especificidade ($86,7\% \pm 7,7$) e AUC ($88,7\% \pm 5,6$) foram próximas às dos outros modelos.

A NCNN apresentou os maiores desvios-padrão, especialmente na sensibilidade, e, portanto, menor estabilidade entre as diferentes partições do conjunto de dados. A VGGFace teve a menor variabilidade entre os três modelos ao longo das iterações da validação cruzada.

Apesar das diferenças numéricas observadas entre as arquiteturas, o teste não paramétrico de Friedman não indicou significância estatística em nenhuma das métricas analisadas ($p > 0,05$). Nas condições experimentais adotadas, portanto, o teste não forneceu evidência de diferença no desempenho global dos três modelos.

Tabela 6 – Métricas de classificação dos modelos (média ± DP, em %).

Métrica	**NCNN**	**VGGFace**	**ViT-B/32**	**valor-p**
Acurácia	80,0 ± 9,2	**84,2 ± 4,8**	81,8 ± 6,9	0,8356
F1-Score	76,0 ± 11,8	**82,0 ± 5,9**	78,5 ± 7,7	0,7939
Sensibilidade	72,5 ± 16,2	**80,8 ± 8,8**	74,3 ± 11,2	0,2358
Especificidade	86,7 ± 7,7	87,0 ± 9,6	**88,1 ± 7,2**	0,9178
Precisão	82,2 ± 10,9	**84,4 ± 9,5**	84,3 ± 7,7	0,8357
AUC	88,7 ± 5,6	**91,5 ± 3,3**	90,1 ± 4,7	0,4065

Fonte: Autor.

3.6 DISCUSSÃO

Os resultados mostram que os três modelos discriminam a presença e a ausência de dor com desempenho global sem diferença estatisticamente significativa, embora com perfis distintos. A VGGFace apresentou o comportamento mais consistente entre as arquiteturas avaliadas, o que corrobora estudos que apontam a utilidade de modelos pré-treinados em faces humanas para a detecção da dor neonatal (Zamzmi *et al.*, 2018a; Carlini *et al.*, 2021; Coutrin *et al.*, 2023). O ViT-B/32 apresentou desempenho intermediário, com maior especificidade, enquanto a NCNN obteve resultados mais modestos, possivelmente em razão da ausência de pré-treinamento em larga escala e da menor capacidade de generalização decorrente dessa escolha arquitetural (Coutrin *et al.*, 2023). Ainda assim, como o teste de Friedman não forneceu evidência de diferença entre as arquiteturas, os três modelos foram mantidos nas etapas subsequentes da metodologia.

Em relação à literatura, as métricas observadas aproximam-se do desempenho humano reportado para bases neonatais semelhantes. Dados da base UNIFESP indicam acurácia média de 92,3% ± 7 para profissionais de saúde e de 86,5% ± 9 para pais (Carlini *et al.*, 2024). Os modelos treinados apenas com imagens estáticas alcançam, assim, desempenho discriminativo suficiente para justificar a investigação de outras propriedades, como confiabilidade probabilística.

Os resultados também expõem uma limitação estrutural da tarefa, na qual a classificação baseada em imagens isoladas simplifica um fenômeno inerentemente dinâmico. Manworren *et al.* (2024) mostraram que a concordância entre avaliadores é substancialmente maior em vídeos do que em imagens estáticas, o que indica que parte da informação clinicamente relevante depende do contexto temporal.

Este capítulo mostra que os modelos têm capacidade discriminativa suficiente para sustentar as etapas posteriores, mas também deixa claro que o desempenho classificatório, isoladamente, não resolve o problema central da tese. Torna-se necessário investigar a confiabilidade das predições, a calibração das probabilidades, a plausibilidade clínica das decisões e as condições em que elas podem ser utilizadas com segurança. Essas análises são necessárias para produzir sistemas mais transparentes, auditáveis e operacionalmente seguros (Arrieta *et al.*, 2020).

3.6.1 Implicações para a prática clínica

A principal implicação clínica deste capítulo é que o desempenho discriminativo, embora necessário, é insuficiente para sustentar o uso assistencial. A predição binária isolada não informa quão confiável é a decisão, não explicita quais regiões faciais a sustentam e tampouco define em quais condições o sistema deve se abster de predizer. Em um contexto sensível como a UTIN, essas limitações impedem a transposição direta do modelo para a prática clínica.

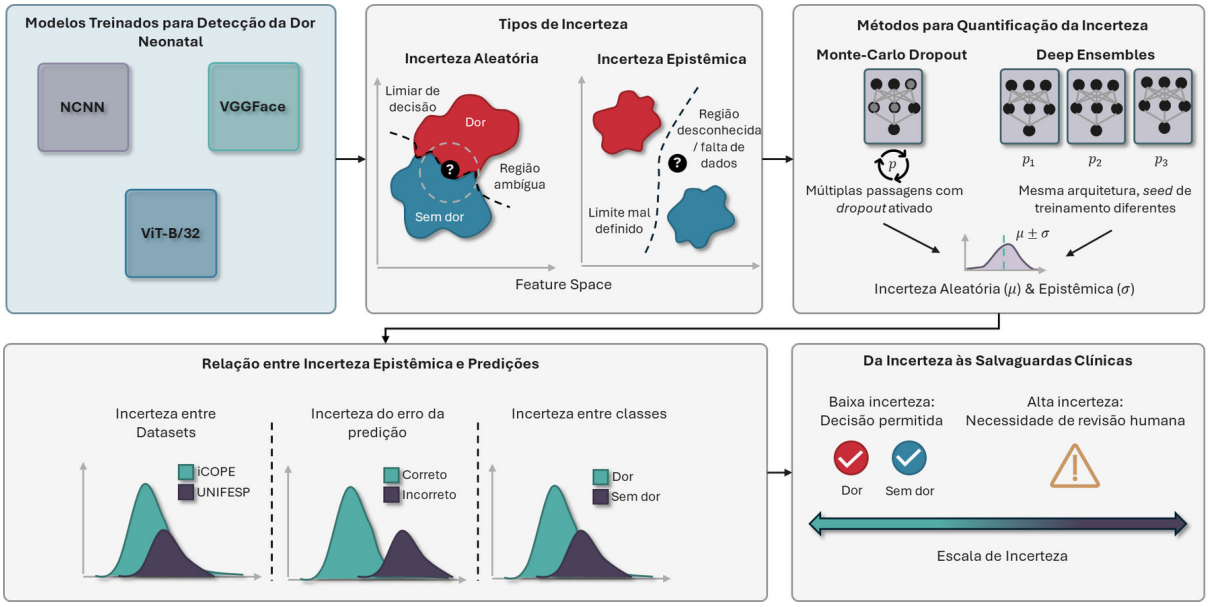
A interpretação das métricas deve considerar o custo clínico de cada tipo de erro. Na triagem neonatal, a sensibilidade é especialmente importante, pois mede a capacidade de identificar episódios de dor e está relacionada à frequência de falsos negativos, isto é, neonatos com dor classificados como “sem dor”. A VGGFace apresentou a maior sensibilidade e pode ser mais adequada a cenários em que a prioridade é não deixar casos de dor passarem despercebidos, ainda que ao custo de avaliações adicionais.

Por outro lado, a especificidade mede a capacidade de reconhecer corretamente os casos sem dor e está associada ao controle de falsos positivos, que podem aumentar a carga de trabalho da equipe e reduzir a confiança no sistema. O ViT-B/32 apresentou maior especificidade, indicando potencial para reduzir alarmes desnecessários. Entretanto, em contextos críticos como a UTIN, maximizar a especificidade à custa da queda da sensibilidade pode ser clinicamente indesejável, pois implica maior risco de não identificar episódios dolorosos.

A NCNN, embora seja uma arquitetura especializada e potencialmente vantajosa em ambientes com recursos computacionais limitados, apresentou menor estabilidade nas condições avaliadas, principalmente pela variabilidade da sensibilidade.

4 QUANTIFICAÇÃO DA INCERTEZA

Figura 10 – Visão geral da metodologia do Capítulo 4.



Fonte: Autor.

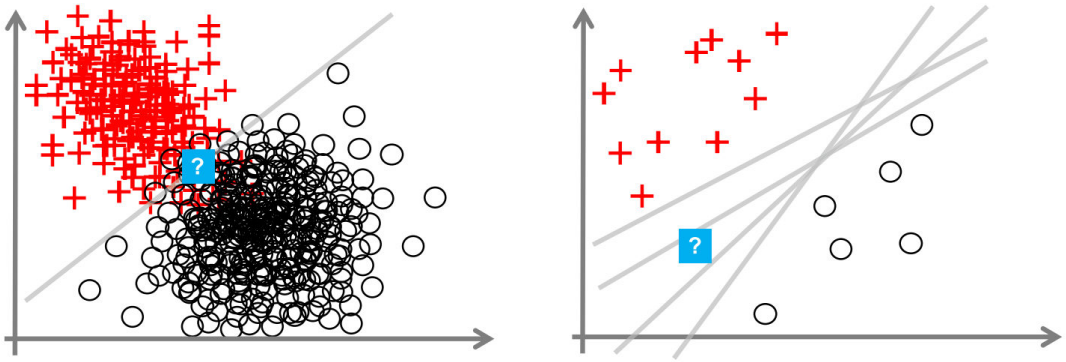
4.1 INTRODUÇÃO

No Capítulo 3, foram desenvolvidos os modelos de classificação da dor neonatal, cujo desempenho discriminativo se mostrou suficiente para sustentar as etapas seguintes da tese. Esse resultado, entretanto, não informa a confiabilidade de cada predição. Este capítulo desenvolve um dos pilares de confiabilidade do sinal da dor, dedicado à quantificação explícita da incerteza preditiva.

O objetivo deste capítulo é investigar se os modelos treinados, além de classificar as expressões nas categorias *com dor* e *sem dor*, conseguem sinalizar situações de dúvida (Figura 10). Essa etapa está diretamente relacionada à segunda pergunta de pesquisa, pois fornece ao sinal um mecanismo adicional de segurança para reduzir erros confiantes e apoiar a definição posterior de limiares operacionais.

Em aplicações médicas, modelos de IA são frequentemente expostos a dados que diferem da distribuição observada durante o treinamento, fenômeno conhecido como *domain shift* (Yoon *et al.*, 2024). Nessas condições, o sistema pode emitir predições incorretas com alta confiança e produzir consequências potencialmente críticas (Abdar *et al.*, 2021; Gawlikowski *et al.*, 2023; Lambert *et al.*, 2024). Quantificar a incerteza permite comunicar ao usuário quando a saída do modelo deve ser interpretada com cautela.

Figura 11 – Representação gráfica das incertezas aleatória e epistêmica.



Fonte: Adaptado de Hüllermeier e Waegeman (2021).

Legenda: À esquerda, mesmo conhecendo as fronteiras de decisão, o ponto de teste (indicado por um ponto de interrogação) apresenta incerteza aleatória, pois as classes se sobrepõem nessa região. À direita, observa-se um caso de incerteza epistêmica, decorrente da ausência de dados naquele segmento do espaço amostral, o que impede a definição precisa das fronteiras de decisão.

A incerteza é tradicionalmente dividida em duas categorias: *aleatória* e *epistêmica*, ilustradas na Figura 11. A incerteza aleatória decorre da variabilidade intrínseca dos dados e manifesta-se mesmo quando o modelo conhece adequadamente o domínio. Por ser uma propriedade da própria distribuição dos dados, é irredutível (Hüllermeier; Waegeman, 2021). A incerteza epistêmica, em contraste, resulta da limitação do conhecimento do modelo, seja pela escassez de dados representativos, seja por restrições em sua capacidade de generalização. Diferentemente da incerteza aleatória, ela é redutível pela incorporação de novos dados ou pelo uso de arquiteturas mais adequadas (Hüllermeier; Waegeman, 2021).

Modelos convencionais de aprendizado profundo, como CNNs e ViTs, não produzem estimativas explícitas de incerteza, pois retornam apenas probabilidades derivadas da camada de saída. Desde a década de 1990, Redes Neurais Bayesianas (BNNs) vêm sendo investigadas com o objetivo de conferir interpretação probabilística a esses modelos, por meio da inferência de distribuições sobre seus pesos (MacKay, 1992). Apesar de sua relevância teórica, a inferência em BNNs permanece desafiadora, tanto do ponto de vista computacional quanto metodológico (Gal; Ghahramani, 2016; Abdar *et al.*, 2021; Gawlikowski *et al.*, 2023).

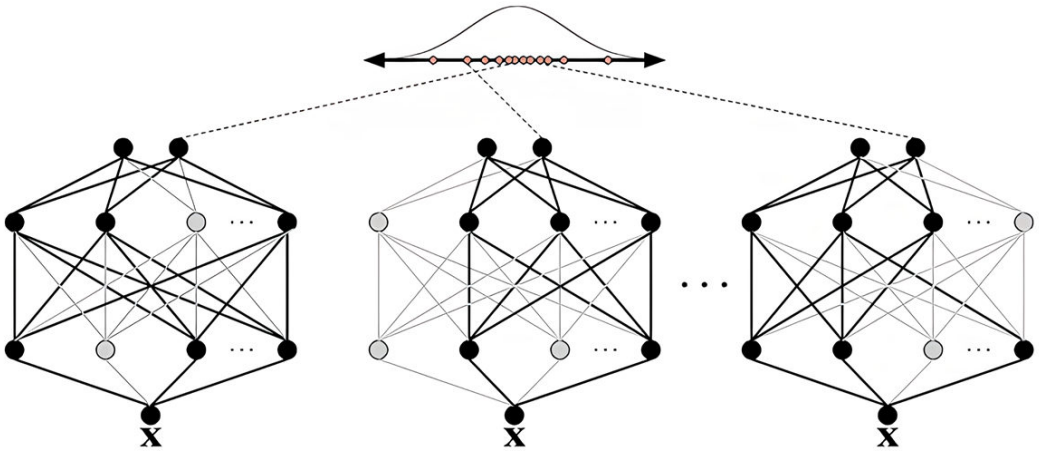
Para contornar essas limitações, foram propostos métodos aproximados, como o Monte Carlo Dropout (Gal; Ghahramani, 2016) e os *Deep Ensembles* (Lakshminarayanan; Pritzel; Blundell, 2017), que permitem estimar distribuições sobre as predições de modelos determinísticos. Neste capítulo, os dois métodos são comparados com duas finalidades: (I) identificar qual deles apresenta a maior associação entre incerteza e erro de classificação; e (II) caracterizar os fatores mais frequentes nas predições incertas.

4.2 MÉTODOS

4.2.1 Monte Carlo Dropout

O método Monte Carlo Dropout (MCDP), proposto por Gal e Ghahramani (2016), aplica-se a diferentes modelos de *deep learning* (Gal; Ghahramani, 2016; Gawlikowski *et al.*, 2023; Lambert *et al.*, 2024). A técnica mantém o *Dropout* ativo durante a inferência. Desse modo, múltiplas passagens de um mesmo dado pelo modelo desativam aleatoriamente diferentes subconjuntos de neurônios. Em vez de uma predição pontual, obtém-se uma distribuição de probabilidades por meio das múltiplas passagens, conforme ilustrado na Figura 12.

Figura 12 – Funcionamento do Monte Carlo Dropout.



Fonte: Adaptado do manual *online* da Amazon Web Services (AWS) (Davis *et al.*, 2020).

Legenda: Neurônios desativados aparecem em cinza e os ativos em preto. No topo, observa-se a distribuição de probabilidades gerada ao aplicar o MCDP.

O *Dropout* foi originalmente concebido como uma técnica de regularização para reduzir sobreajuste, desligando aleatoriamente uma fração dos neurônios a cada época de treinamento. Conforme discutido no Capítulo 3, os modelos empregados já incorporavam essa operação, o que possibilitou a implementação do MCDP sem a adição de novas camadas, em consonância com Gal e Ghahramani (2016).

Com base na distribuição de probabilidades gerada pelo MCDP, estimaram-se as componentes aleatória e epistêmica da incerteza. A média das predições foi utilizada para avaliar a ambiguidade entre as classes, associada à incerteza aleatória, enquanto o desvio-padrão foi adotado como medida de incerteza epistêmica. Valores médios próximos de 50% indicam maior ambiguidade. Desvios-padrão elevados, por sua vez, sinalizam instabilidade entre as amostragens estocásticas e conhecimento limitado do modelo sobre o dado analisado. Nessas situações, decisões semi-automáticas devem ser evitadas, e a avaliação por um profissional qualificado deve ser priorizada.

4.2.2 Deep Ensembles

O método de *Deep Ensembles* consiste no treinamento de múltiplas redes neurais com a mesma arquitetura, porém inicializadas com pesos distintos e expostas a diferentes ordens de apresentação dos dados (Lakshminarayanan; Pritzel; Blundell, 2017). Essa diversidade leva cada rede a convergir para soluções ligeiramente distintas no espaço de parâmetros. Na fase de inferência, as predições individuais são agregadas, produzindo uma distribuição de probabilidades a partir da qual se estima a incerteza epistêmica, tipicamente por meio do desvio-padrão entre modelos, de modo semelhante ao MCDP.

A forma mais direta de promover diversidade entre os modelos é controlar as fontes de aleatoriedade por meio de *seeds* distintas nos geradores pseudoaleatórios, incluindo inicialização de pesos, embaralhamento de *batches* e variações de *data augmentation*. O uso de *seeds* distintas assegura que cada modelo percorra trajetórias de otimização diferentes, produzindo a variabilidade necessária aos *Deep Ensembles*.

Apesar da simplicidade conceitual, os *Deep Ensembles* implicam maior custo computacional, pois requerem o treinamento completo de múltiplos modelos independentes, usualmente entre cinco a dez (Lakshminarayanan; Pritzel; Blundell, 2017). O equilíbrio entre esse custo adicional e o ganho na qualidade da estimativa de incerteza deve ser avaliado à luz das restrições e das exigências da aplicação.

4.3 EXPERIMENTO

Os experimentos avaliaram, para cada arquitetura, qual método de quantificação da incerteza produziu estimativas mais úteis do ponto de vista operacional. Os modelos VGGFace, NCNN e ViT-B/32 foram submetidos à validação cruzada com os mesmos *folds* de treinamento e validação descritos no Capítulo 3. Aplicaram-se MCDP e *Deep Ensembles*, que geraram distribuições de probabilidades a partir das quais se estimou a incerteza epistêmica. Representada por σ, essa incerteza corresponde ao desvio-padrão das distribuições e constitui a variável central do experimento, por representar a dúvida inerente ao modelo.

No MCDP, o número de amostragens estocásticas foi fixado em $N_{\text{MC}} = 50$. A taxa de *Dropout* foi definida empiricamente, considerando-se os valores de 10%, 30% e 50%. Para os *Deep Ensembles*, o número de modelos também foi determinado empiricamente, assumindo 3, 5 e 10 redes. A inicialização aleatória dos pesos foi controlada por diferentes *seeds* definidas aleatoriamente, conforme a seguinte lista:

$$\texttt{seeds} = [1234, 5678, 91011, 121314, 151617, 2023, 31415, 27182, 424242, 8675309],$$

adotando-se a $\texttt{seed} = 1234$ como padrão nos demais treinamentos, enquanto as outras foram aplicadas sequencialmente para a construção de até 10 modelos no *ensemble*.

Após a obtenção dos resultados em ambos os métodos, selecionou-se o método de quantificação da incerteza que apresentou maior correlação entre os altos valores de incerteza

epistêmica e as predições incorretas nos dados de validação. Para essa análise, empregaram-se as correlações de Spearman (ρ) e Pearson (r), que permitem examinar, respectivamente, relações monotônicas e lineares.

Por fim, o método considerado mais adequado para cada modelo foi analisado em maior profundidade, com o objetivo de investigar os fatores associados às respostas classificadas como incertas. Essa análise contemplou características do banco de dados, a distribuição das classes e as variações nas expressões faciais dos RNs. Para todos os casos, foi aplicado o teste de Kolmogorov–Smirnov (K–S) (Kolmogorov, 1933; Smirnov, 1948) para avaliar a separação entre as distribuições de incerteza em função das variáveis analisadas.

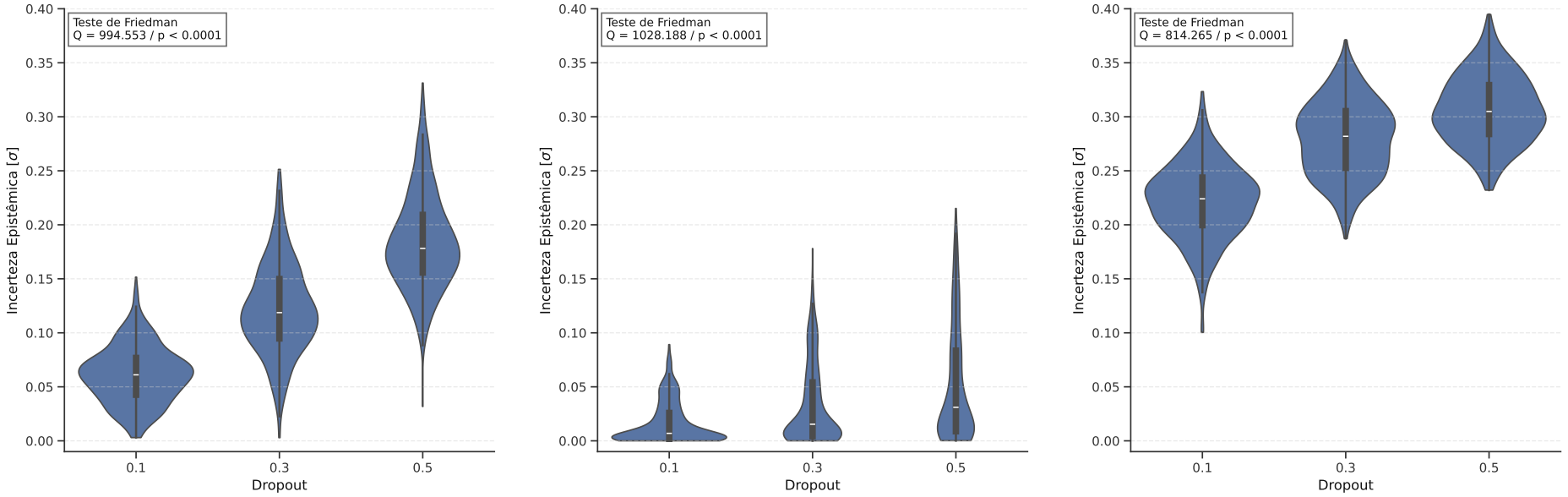
Para as métricas de classificação, as comparações entre NCNN, VGGFace e ViT-B/32 foram realizadas com o teste de Friedman (Friedman, 1937). A comparação pareada de cada arquitetura antes e depois da incorporação do método de quantificação da incerteza foi realizada com o teste de Wilcoxon (Wilcoxon, 1945). Adotou-se $p < 0{,}05$ como critério de significância estatística.

4.4 RESULTADOS

4.4.1 Avaliação dos métodos de quantificação da incerteza

As Figuras 13 e 14 apresentam a distribuição dos valores de incerteza estimados pelos métodos MCDP e *Deep Ensembles* para os modelos NCNN, VGGFace e ViT-B/32. Nessas figuras, também são apresentados a estatística Q e o valor-p do teste de Friedman, pareado e não paramétrico.

Figura 13 – Distribuição da incerteza epistêmica em função dos parâmetros do MCDP.



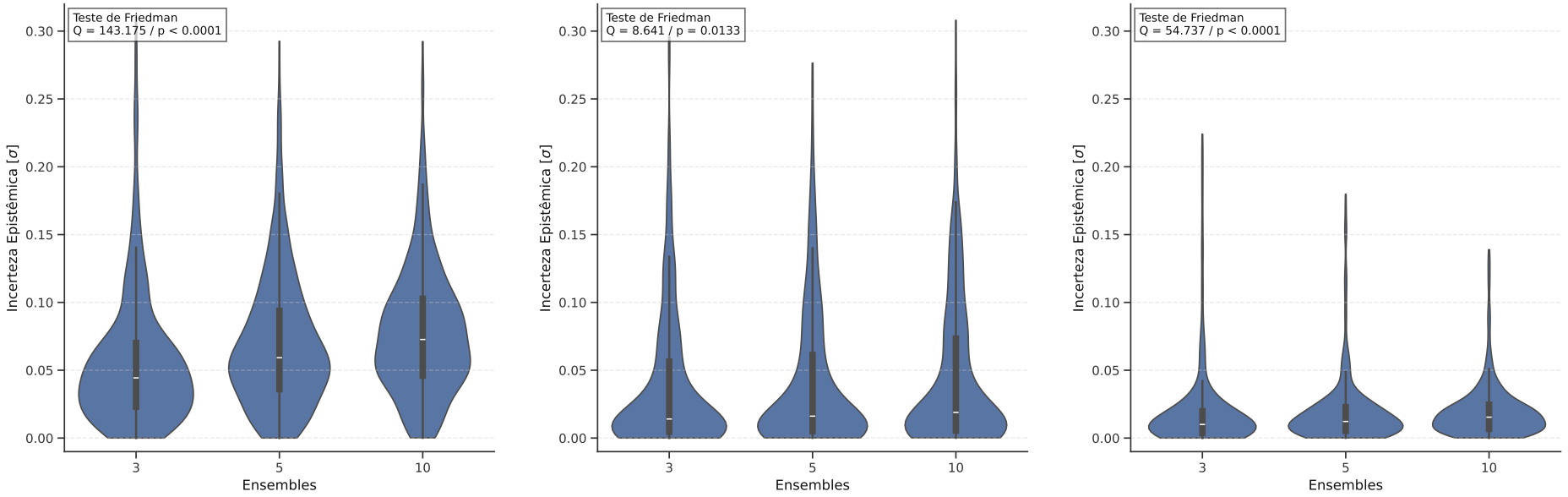
(a) NCNN.

(b) VGGFace.

(c) ViT-B/32.

Fonte: Autor.

Figura 14 – Distribuição da incerteza epistêmica em função dos parâmetros do Deep Ensembles.



(a) NCNN.

(b) VGGFace.

(c) ViT-B/32.

Fonte: Autor.

Em todos os modelos, houve diferença estatisticamente significativa entre as distribuições de incerteza ($p < 0{,}05$). A estatística Q foi maior no MCDP ($Q_{NCNN} = 994$, $Q_{VGGFace} = 1028$, $Q_{ViT-B/32} = 814$) do que nos *Deep Ensembles* ($Q_{NCNN} = 143$, $Q_{VGGFace} = 8$, $Q_{ViT-B/32} = 54$), o que indica maior sensibilidade da incerteza às variações paramétricas no MCDP. Nos *Deep Ensembles*, o aumento do número de redes produziu um efeito mais discreto. Entre as arquiteturas, a VGGFace apresentou distribuições sistematicamente deslocadas para valores menores de incerteza em ambos os métodos, sugerindo maior estabilidade do que a NCNN e o ViT-B/32.

A seleção do método de quantificação não se baseou apenas na incerteza média, mas também na capacidade de identificar predições incorretas. Para isso, utilizaram-se as correlações de Spearman (ρ) e Pearson (r) entre incerteza e erro de classificação. A Tabela 7 apresenta esses resultados para o MCDP (variando a taxa de *dropout*, p_{drop}) e para os *Deep Ensembles* (variando o número de redes, k_{ens}).

Conforme Cohen (1988) e Evans (1996), correlações abaixo de 0,30 são consideradas fracas, entre 0,30 e 0,50 são moderadas e, acima de 0,50, fortes. Correlações positivas indicam que exemplos incorretos tendem a apresentar maiores níveis de incerteza.

Tabela 7 – Correlações de Spearman (ρ) e Pearson (r) entre incerteza epistêmica e erro de classificação (média ± DP).

	MCDP (prob. de dropout p_{drop})			**Deep Ensemble** (nº de redes k_{ens})		
Modelo	p_{drop} = 10%	p_{drop} = 30%	p_{drop} = 50%	k_{ens} = 3	k_{ens} = 5	k_{ens} = 10
NCNN	ρ = 0,25 ± 0,16 r = 0,24 ± 0,17	**ρ = 0,28 ± 0,15** r = 0,27 ± 0,17	ρ = 0,26 ± 0,15 **r = 0,28 ± 0,19**	ρ = 0,13 ± 0,19 r = 0,16 ± 0,19	ρ = 0,26 ± 0,17 r = 0,27 ± 0,17	ρ = 0,24 ± 0,16 r = 0,25 ± 0,14
VGGFace	ρ = 0,34 ± 0,09 r = 0,34 ± 0,14	ρ = 0,34 ± 0,09 r = 0,35 ± 0,13	**ρ = 0,36 ± 0,09** **r = 0,37 ± 0,13**	ρ = 0,32 ± 0,16 r = 0,28 ± 0,23	ρ = 0,29 ± 0,13 r = 0,23 ± 0,20	ρ = 0,33 ± 0,11 r = 0,32 ± 0,19
ViT-B/32	ρ = 0,17 ± 0,12 r = 0,18 ± 0,11	ρ = 0,02 ± 0,14 r = 0,01 ± 0,14	ρ = 0,01 ± 0,20 r = 0,02 ± 0,19	ρ = 0,15 ± 0,18 r = 0,14 ± 0,18	ρ = 0,20 ± 0,20 r = 0,20 ± 0,20	**ρ = 0,22 ± 0,20** **r = 0,22 ± 0,21**

Fonte: Autor.

Na VGGFace, o MCDP com $p_{\text{drop}} = 50\%$ apresentou as maiores correlações ($\rho = 0,36 \pm 0,09$ e $r = 0,37 \pm 0,13$), ligeiramente superiores às demais configurações e aos *Deep Ensembles*.

Na NCNN, MCDP e *Deep Ensembles* apresentaram correlações de magnitude semelhante. O melhor resultado para Spearman ocorreu em $p_{\text{drop}} = 30\%$ ($\rho = 0,28 \pm 0,15$) e, para Pearson, em $p_{\text{drop}} = 50\%$ ($r = 0,28 \pm 0,19$).

No ViT-B/32, o MCDP resultou em correlações próximas de zero para $p_{\text{drop}} = 30\%$ e $p_{\text{drop}} = 50\%$, indicando baixa capacidade discriminativa da incerteza. Os *Deep Ensembles* apresentaram melhora progressiva com o aumento de k_{ens}, alcançando o melhor desempenho em $k_{\text{ens}} = 10$ ($\rho = r = 0,22 \pm 0,20$). Esse resultado evidencia que, nessa arquitetura e nas condições avaliadas, o *dropout* não capturou adequadamente a incerteza epistêmica, enquanto a diversidade entre modelos independentes se mostrou mais informativa.

As métricas de classificação de todas as variantes estão apresentadas no Apêndice A. Na comparação pareada de cada modelo com sua respectiva etapa anterior, realizada com o teste de Wilcoxon, a quantificação da incerteza não alterou de forma significativa a maior parte das métricas classificatórias. A única alteração estatisticamente significativa foi o aumento da AUC da VGGFace ($p = 0{,}014$). A Tabela 8 apresenta as métricas de classificação após a aplicação dos melhores métodos de quantificação da incerteza. Os valores-p da última coluna, obtidos pelo teste de Friedman, comparam as três arquiteturas após essa etapa e não indicam diferenças significativas. Observa-se, portanto, que o impacto desses métodos foi mínimo, com preservação do desempenho global dos modelos.

Tabela 8 – Métricas de classificação dos modelos após quantificação da incerteza (média ± DP, em %).

Métrica	**NCNN** (MCDP p_{drop} = 30%)	**VGGFace** (MCDP p_{drop} = 50%)	**ViT-B/32** (10 Ensembles)	**valor-p**
Acurácia	79,5 ± 8,7 ↓	**84,1 ± 5,0** ↓	82,8 ± 6,7 ↑	0,5836
F1-Score	74,8 ± 11,5 ↓	**81,5 ± 6,8** ↓	79,9 ± 7,4 ↑	0,3320
Sensibilidade	70,2 ± 17,2 ↓	**80,2 ± 10,3** ↓	75,9 ± 10,2 ↑	0,3474
Especificidade	87,7 ± 6,6 ↑	87,0 ± 8,5 =	**88,6 ± 6,8** ↑	0,9710
Precisão	83,0 ± 10,1 ↑	84,1 ± 8,8 ↓	**85,0 ± 6,8** ↑	0,7165
AUC	88,3 ± 6,4 ↓	**91,9 ± 3,4** ↑*	89,9 ± 5,0 ↓	0,1496

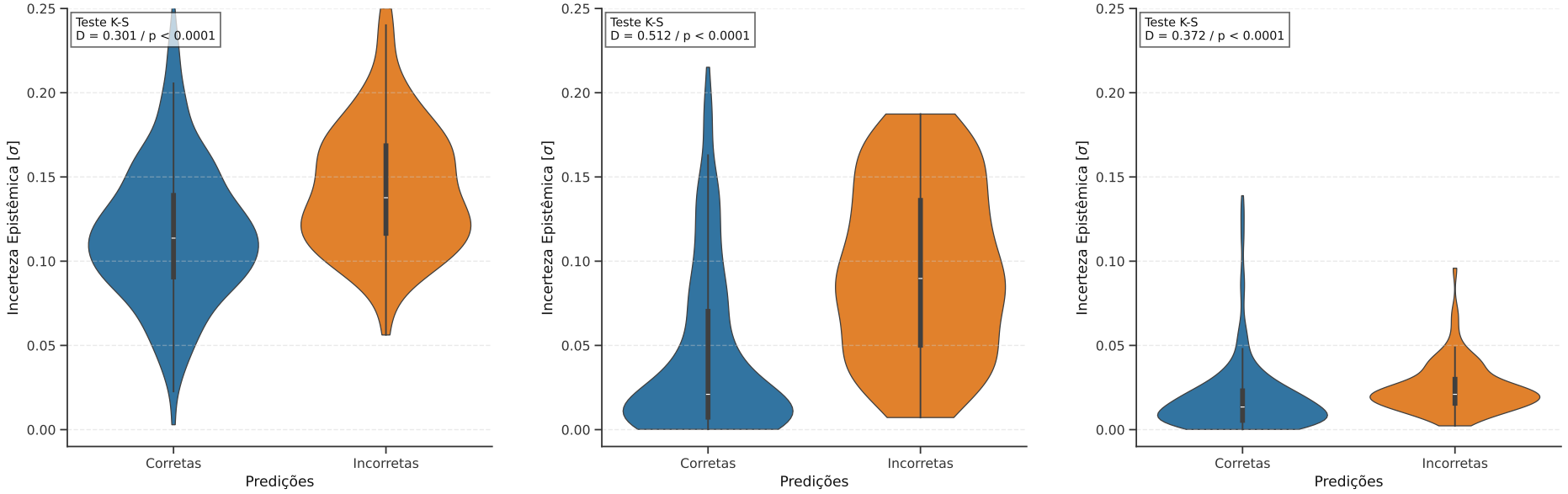
Fonte: Autor.

Legenda: Os símbolos indicam melhora (↑), piora (↓), melhora estatisticamente significativa (↑*), piora estatisticamente significativa (↓*) e resultado inalterado (=). O asterisco indica $p < 0{,}05$ no teste de Wilcoxon pareado com a etapa anterior. A última coluna apresenta o teste de Friedman entre os modelos.

4.4.2 Fatores associados à incerteza

As análises seguintes procuram identificar os fatores associados à incerteza dos modelos, considerando o melhor método de quantificação selecionado para cada arquitetura. A Figura 15 apresenta as distribuições da incerteza epistêmica para predições corretas (azul) e incorretas (laranja) nos três modelos. Em todos os casos, as distribuições das predições incorretas deslocaram-se para valores mais elevados de σ, indicando que exemplos classificados incorretamente tenderam a ser mais incertos. A separação entre as distribuições foi estatisticamente significativa nos três modelos ($p < 0{,}0001$), embora com magnitudes distintas. A VGGFace apresentou a maior distância K–S ($D = 0{,}512$), seguida pelo ViT-B/32 ($D = 0{,}372$) e pela NCNN ($D = 0{,}301$).

Figura 15 – Distribuição da incerteza epistêmica em função das respostas corretas e incorretas.

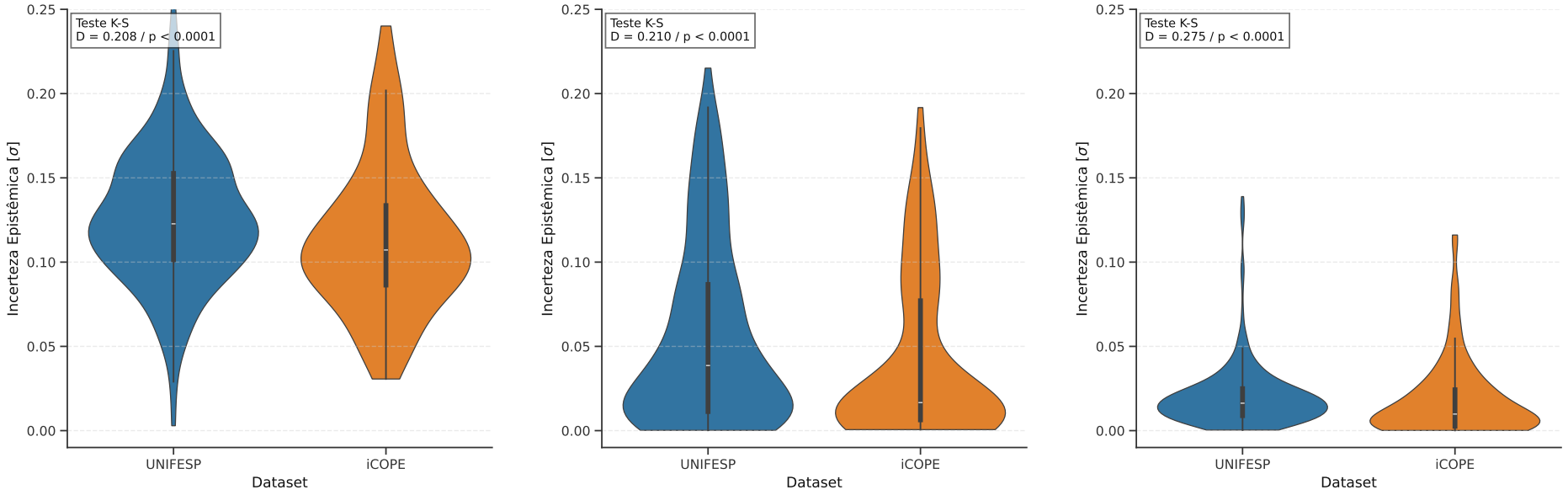


(a) NCNN. (b) VGGFace. (c) ViT-B/32.

Fonte: Autor.

A Figura 16 compara as distribuições da incerteza epistêmica entre os bancos UNIFESP (azul) e iCOPE (laranja). Em todos os casos, verificou-se separação estatisticamente significativa entre as distribuições ($p < 0{,}0001$), com magnitudes $D = 0{,}208$ na NCNN, $D = 0{,}210$ na VGGFace e $D = 0{,}275$ no ViT-B/32. Observa-se que o conjunto UNIFESP apresentou deslocamento para valores mais elevados de σ, enquanto o iCOPE concentrou maior densidade em níveis mais baixos de incerteza.

Figura 16 – Distribuição da incerteza epistêmica em função dos bancos de imagens.



(a) NCNN.

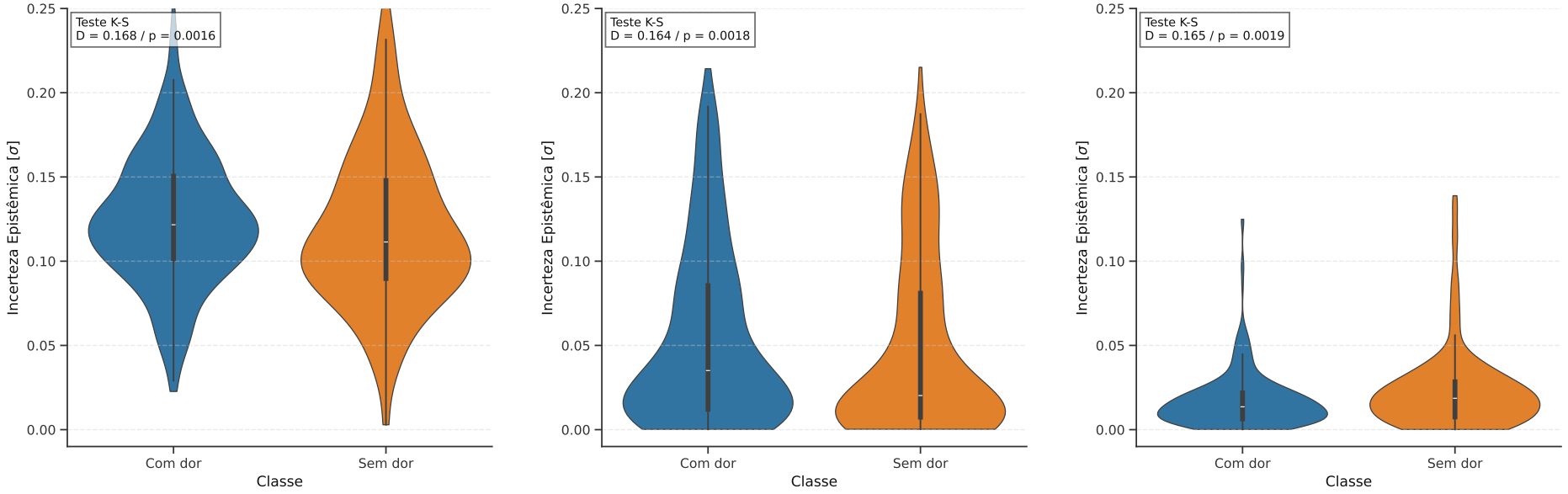
(b) VGGFace.

(c) ViT-B/32.

Fonte: Autor.

Por fim, a Figura 17 compara as distribuições da incerteza epistêmica para predições *com dor* (azul) e *sem dor* (laranja). Observou-se um leve deslocamento das predições *sem dor* para valores mais elevados de σ, enquanto as predições *com dor* se concentraram em níveis mais baixos de incerteza. As diferenças entre as distribuições foram estatisticamente significativas em todos os modelos, embora com magnitudes modestas (valores de D em torno de 0,16), o que indica separação parcial.

Figura 17 – Distribuição da incerteza epistêmica em função das predições dos modelos.



(a) NCNN.

(b) VGGFace.

(c) ViT-B/32.

Fonte: Autor.

4.4.3 Análise qualitativa da incerteza

Na análise qualitativa, a Figura 18 apresenta as faces médias dos RNs, alinhadas em um referencial comum e estratificadas pelos quartis da incerteza epistêmica (σ), para predições *com dor* e *sem dor* nos três modelos. O Apêndice A reúne exemplos individuais de predições certas e incertas de cada modelo.

Nas predições *com dor*, observa-se que, nos menores níveis de incerteza (Min–Q1), as faces médias exibiram de forma nítida marcadores clássicos da expressão dolorosa, como abertura ampla da boca, acentuação do sulco nasolabial e sinais de estreitamento palpebral. À medida que a incerteza aumentou (Q2–Q3–Max), essas características tornaram-se progressivamente mais difusas e menos pronunciadas, resultando em fisionomias mais ambíguas (boca apenas parcialmente aberta, sulco nasolabial pouco definido e menor contração das pálpebras). Esse gradiente foi mais evidente visualmente na VGGFace e na NCNN, enquanto, no ViT-B/32, não se observaram diferenças claras nas características faciais entre os intervalos de incerteza.

Para as predições *sem dor*, as imagens médias não apresentaram um padrão claro em relação às características faciais associadas à maior incerteza. Esse resultado reforça que a ausência de dor é um estado mais difícil de afirmar com alta certeza do que a presença de marcadores evidentes de dor. Entretanto, ao observar as imagens isoladas com baixa incerteza apresentadas no Apêndice A, foi possível identificar olhos abertos e boca fechada como os principais marcadores faciais associados a tais predições.

4.5 DISCUSSÃO

Os resultados mostram que a quantificação da incerteza acrescenta uma dimensão de confiabilidade que não está presente nas métricas discriminativas do capítulo anterior. A incerteza epistêmica mostrou-se informativa para distinguir, ao menos parcialmente, predições corretas de incorretas, o que é particularmente relevante em um cenário clínico no qual erros representam risco direto ao paciente.

Os testes estatísticos indicaram que o MCDP foi mais sensível às variações de seus parâmetros do que os *Deep Ensembles*, com aumento da incerteza à medida que a taxa de *dropout* se elevou. Esse comportamento é compatível com a natureza do método, já que taxas maiores de *dropout* ampliam a variabilidade entre inferências. Nos *Deep Ensembles*, em contraste, a diversidade decorre de trajetórias independentes de otimização, produzindo estimativas mais estáveis, embora menos responsivas ao aumento do número de modelos.

Na prática, a utilidade da incerteza variou conforme a arquitetura. A associação entre erro e incerteza foi mais forte na VGGFace, intermediária na NCNN e mais discreta no ViT-B/32. A incerteza não deve, portanto, ser tratada como uma propriedade universal, mas como uma grandeza dependente da arquitetura e do método de quantificação. Por essa razão, os valores absolutos de σ não são diretamente comparáveis entre modelos e devem ser interpretados com limiares específicos para cada arquitetura.

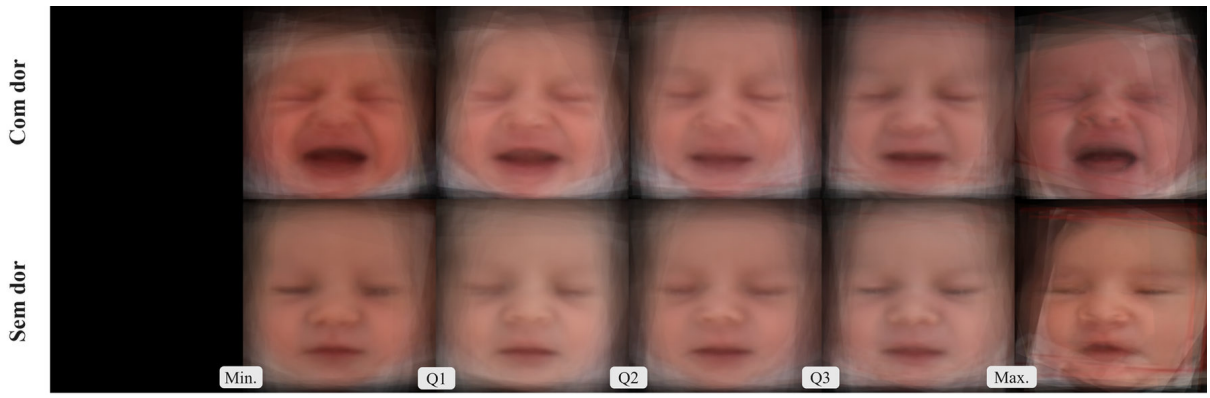
Figura 18 – Faces médias das predições estratificadas por quartis de incerteza epistêmica.



(a) NCNN.



(b) VGGFace.



(c) ViT-B/32.

Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

Legenda: Os quartis Q1, Q2 e Q3 representam, respectivamente, 25%, 50% e 75% dos valores da incerteza epistêmica. Quanto mais à direita, maior a incerteza.

A análise entre bancos de dados mostrou que a incerteza também capturou discrepâncias de domínio. O conjunto UNIFESP apresentou valores mais altos de σ do que o iCOPE. Esse achado é coerente com a interpretação da incerteza epistêmica como insuficiência de conhecimento do modelo diante de exemplos pouco representados no treinamento, uma vez que o banco UNIFESP é mais diverso quanto à etnia e às condições de aquisição, mas possui poucas amostras para representar toda essa variação.

Também se observou maior incerteza nas predições *sem dor*. Em conjunto com as faces médias da Figura 18, esse padrão evidencia, nos modelos avaliados, maior dificuldade para afirmar com alta certeza a ausência de dor do que para identificar marcadores faciais evidentes de dor. Essa dificuldade é compatível com a natureza do problema, pois a ausência de dor não se manifesta por um fenótipo específico, mas pela falta de marcadores claros. Predições *sem dor* com alta incerteza devem ser interpretadas com cautela, pois podem corresponder a casos ambíguos ou atípicos, pouco representados no conhecimento do modelo. Esse resultado também é coerente com a sensibilidade inferior à especificidade, condição em que os modelos mostraram maior propensão a falsos negativos (classificar como *sem dor* quando há dor) do que a falsos positivos (classificar como *com dor* quando não há dor).

Do ponto de vista computacional, tanto o MCDP quanto os *Deep Ensembles* adicionam complexidade ao processo de inferência devido às múltiplas passagens necessárias. Esse aspecto pode limitar a aplicabilidade prática de arquiteturas mais complexas, como a VGGFace e o ViT-B/32, em ambientes hospitalares com restrições de hardware, reforçando a necessidade de equilibrar desempenho preditivo, quantificação da incerteza e custo computacional.

4.5.1 Implicações para a prática clínica

Em UTINs, predições muito incertas devem receber tratamento específico e ser acompanhadas por profissionais de saúde. Limiares de incerteza podem filtrar predições pouco confiáveis, priorizar casos duvidosos para revisão humana e reduzir o risco de falsos negativos, diante da maior incerteza em casos *sem dor*.

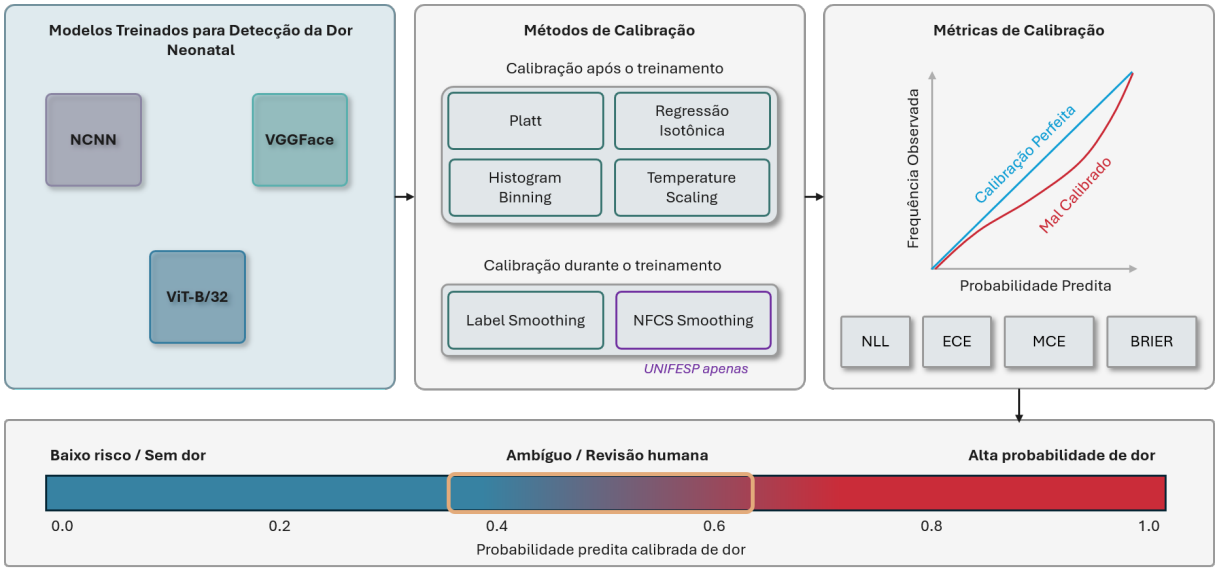
As diferenças observadas entre os bancos de dados iCOPE e UNIFESP também evidenciam a necessidade de considerar variabilidades demográficas e técnicas na implementação de sistemas de IA. O aumento da incerteza em dados provenientes de contextos distintos reforça a importância de validações multicêntricas para assegurar comportamento estável em diferentes populações neonatais. Em um cenário de implantação clínica, essa sensibilidade precisa ser monitorada para evitar que tais variações sejam confundidas com alterações reais do estado do RN.

No fluxo de trabalho, a quantificação da incerteza funciona como um mecanismo de triagem, ou seja, casos muito incertos podem ser priorizados para revisão humana, enquanto aqueles de baixa incerteza seguem para outras etapas de qualificação, como a calibração probabilística e a análise explicativa. Este capítulo acrescenta, assim, uma salvaguarda ao sistema e responde

parcialmente à questão de pesquisa 2 ao mostrar que a incerteza epistêmica ajuda a identificar predições potencialmente incorretas e a caracterizar a confiabilidade dos modelos.

5 CALIBRAÇÃO

Figura 19 – Visão geral da metodologia do Capítulo 5.



Fonte: Autor.

5.1 INTRODUÇÃO

O capítulo anterior mostrou que os modelos de classificação expressam incerteza preditiva e que essa informação ajuda a identificar situações em que o conhecimento modelado é limitado. A etapa seguinte verifica em que medida as probabilidades que formarão o sinal da dor constituem estimativas confiáveis de risco no âmbito da tarefa avaliada. Este capítulo examina, portanto, a calibração probabilística como um dos pilares de confiabilidade do sinal e como requisito para o uso responsável das saídas dos modelos.

O objetivo é avaliar e comparar métodos de calibração capazes de alinhar a probabilidade predita do modelo à frequência empírica dos eventos de dor (Figura 19). Essa etapa complementa a quantificação da incerteza e dá continuidade à resposta da segunda pergunta específica de pesquisa, ao investigar se as probabilidades têm significado empírico.

No contexto médico, probabilidades preditas por modelos computacionais podem ser interpretadas como estimativas de risco individual, contribuindo para a priorização de pacientes durante a triagem (Jiang *et al.*, 2012; Kompa; Snoek; Beam, 2021). Em cenários críticos, como a detecção da dor neonatal, essas estimativas permitem direcionar a atenção da equipe para casos com maior probabilidade de dor e reduzir intervenções desnecessárias em situações de menor risco. Contudo, essa interpretação só é válida quando o modelo está calibrado.

Um modelo é considerado perfeitamente calibrado quando a probabilidade predita coincide com a frequência observada. Por exemplo, se o modelo atribui probabilidade de 80% a

um evento, espera-se que esse evento ocorra em aproximadamente 80% dos casos em que tal predição é feita. A calibração é amplamente empregada em áreas como meteorologia (Murphy; Epstein, 1967; Bröcker, 2009) e tornou-se essencial em aplicações de alto risco, como diagnósticos médicos e condução autônoma (Jiang *et al.*, 2012; Guo *et al.*, 2017; Nixon *et al.*, 2019; Kompa; Snoek; Beam, 2021; Filho *et al.*, 2023).

Entretanto, é importante distinguir probabilidade predita de confiança, termos frequentemente utilizados como sinônimos. Em classificação binária, a probabilidade predita corresponde à saída do modelo após a ativação final e expressa a estimativa de que a amostra pertença à classe positiva. Probabilidades próximas às extremidades de 0% ou 100% indicam maior confiança na decisão, enquanto valores intermediários (próximos de 50%) refletem maior incerteza (Kuleshov; Liang, 2015; Kompa; Snoek; Beam, 2021).

A calibração perfeita de um modelo binário é expressa como:

$$\mathbb{P}(y_+ \mid \mathbb{F}(x) = p) = p, \tag{5.1}$$

em que a predição do modelo $\mathbb{F}(x) = p$ deve corresponder à verdadeira probabilidade de ocorrência do evento positivo y_+, neste trabalho definido como a presença de dor. Um modelo pode apresentar elevada capacidade discriminativa e, ainda assim, estar mal calibrado, e o inverso também é possível. Por exemplo, se um modelo atribui probabilidades muito elevadas e próximas entre si ($p_{y_+} = 99\%$ para amostras positivas e $p_{y_-} = 98\%$ para negativas), continuará mal calibrado, pois tais probabilidades não refletem a frequência real dos eventos, embora preservem perfeita discriminação entre classes (Niculescu-Mizil; Caruana, 2005; Jiang *et al.*, 2012; Lakshminarayanan; Pritzel; Blundell, 2017).

Em sistemas de IA aplicados à saúde, a calibração permite interpretar probabilidades como estimativas quantitativas de risco clínico e também fornece uma base mais consistente para definir limiares de decisão, mecanismos de abstenção e estratégias de priorização de casos.

5.2 MÉTODOS

De acordo com Karandikar *et al.* (2021), os métodos de calibração agrupam-se em duas categorias principais. A primeira compreende métodos *post-hoc*, nos quais as probabilidades preditas são ajustadas após o treinamento do modelo. Esses métodos são amplamente utilizados devido à sua simplicidade e ao fato de não exigirem modificações na arquitetura da rede. Exemplos incluem *Platt Scaling*, Regressão Isotônica e *Temperature Scaling* (Guo *et al.*, 2017).

A segunda categoria engloba métodos que incorporam mecanismos específicos para promover a calibração durante o treinamento, por meio de modificações na função de perda ou de técnicas de regularização. Entre eles, destacam-se *Label Smoothing* (Szegedy *et al.*, 2016) e *Mixup* (Zhang *et al.*, 2017), que visam reduzir a superconfiança característica de modelos de *deep learning*. As subseções seguintes descrevem os métodos de calibração adotados neste trabalho, tanto *post-hoc* quanto intrínsecos ao processo de aprendizado.

5.2.1 Funções de ativação de saída

Em modelos de *deep learning*, a conversão dos valores de saída da rede (*logits*) em probabilidades de pertencimento a uma determinada classe é realizada por funções de ativação específicas. Em tarefas de classificação multiclasse, utiliza-se tipicamente a função *Softmax*, enquanto, em problemas binários, a função *Sigmoid* é a mais empregada. Ambas transformam os *logits* em valores no intervalo $[0, 1]$, permitindo sua interpretação como probabilidades aparentes de pertencimento a uma classe.

As funções são definidas como:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \qquad \text{Sigmoid}(z_1) = \frac{1}{1 + e^{-z_1}}, \tag{5.2}$$

em que C representa o número total de classes e z_i o *logit* correspondente à i-ésima classe. No caso da *Sigmoid*, há apenas um *logit* z_1, associado à classe positiva.

Embora *Softmax* e *Sigmoid* produzam saídas com interpretação probabilística, isso não garante que o modelo esteja calibrado. Os valores resultantes refletem a distribuição produzida pelo modelo, mas não necessariamente a frequência empírica de ocorrência dos eventos preditos. Essa discrepância motiva a aplicação de métodos adicionais de calibração, com o objetivo de alinhar probabilidade predita e frequência observada.

5.2.2 Calibração de Platt

A calibração de Platt, proposta por Platt (1999), foi um dos primeiros métodos desenvolvidos para corrigir a descalibração de classificadores. Inicialmente concebida para SVMs, mostrou-se eficaz também na calibração de outros modelos estatísticos e redes neurais (Niculescu-Mizil; Caruana, 2005). O método ajusta as saídas do modelo, representadas por $\mathbb{F}(x)$, por meio de uma transformação sigmoidal parametrizada, de modo que o resultado possa ser interpretado como probabilidade calibrada:

$$\mathbb{P}(y_+ \mid \mathbb{F}(x)) = p = \frac{1}{1 + e^{(A\,\mathbb{F}(x) + B)}}, \tag{5.3}$$

em que A e B são parâmetros ajustáveis aprendidos a partir de um conjunto de validação. Esses parâmetros são estimados pela minimização do *Negative Log-Likelihood (NLL)* entre as classes reais y e as probabilidades preditas p:

$$\underset{A,B}{\operatorname{argmin}} \left\{ - \left[\, y \log(p) + (1 - y) \log(1 - p) \, \right] \right\}. \tag{5.4}$$

O conjunto de dados utilizado para calibração deve ser distinto daquele empregado no treinamento, a fim de evitar sobreajuste dos parâmetros A e B. Recomenda-se o uso de um conjunto independente ou de *cross-validation* (Niculescu-Mizil; Caruana, 2005).

Para mitigar o sobreajuste dos valores de A e B, Platt (1999) propuseram o uso de *soft-labels* em substituição às classes binárias (*hard-labels*). Considerando N_+ amostras positivas e N_- negativas, as classes ajustadas são definidas como:

$$y_+ = \frac{N_+ + 1}{N_+ + 2}, \qquad y_- = \frac{1}{N_- + 2}. \tag{5.5}$$

A calibração de Platt apresenta bom desempenho quando uma função sigmoidal aproxima adequadamente a relação entre *logits* e probabilidades calibradas. Em cenários mais complexos, métodos não paramétricos, como a Regressão Isotônica, podem oferecer melhor desempenho.

5.2.3 Regressão Isotônica

A Regressão Isotônica foi introduzida como método de calibração por Zadrozny e Elkan (2002). Diferentemente da calibração de Platt, que assume uma forma sigmoidal fixa, trata-se de um método não paramétrico que ajusta uma função monotonicamente crescente sem impor formato específico. Essa restrição garante que maiores probabilidades preditas estejam associadas, em média, a maiores frequências empíricas do evento positivo.

Formalmente, o método busca uma função $\mathcal{I}$ pertencente ao conjunto de funções isotônicas que minimize o erro quadrático ponderado entre as probabilidades preditas p_i e as classes verdadeiras y_i:

$$\underset{\mathcal{I}}{\operatorname{argmin}} \left\{ \sum_i w_i \big(\mathcal{I}(p_i) - y_i\big)^2 \right\}, \tag{5.6}$$

em que w_i representa o peso associado a cada amostra, usualmente definido como $w_i = 1$.

Assim como na calibração de Platt, o ajuste de $\mathcal{I}$ deve ser realizado em conjunto de dados independente ou com *cross-validation*. A Regressão Isotônica tende a superar a calibração de Platt quando há grande volume de dados disponíveis, beneficiando-se de sua maior flexibilidade. Em conjuntos menores, essa flexibilidade pode induzir sobreajuste (Niculescu-Mizil; Caruana, 2005).

5.2.4 Histogram Binning

O *Histogram Binning*, proposto por Zadrozny e Elkan (2001), é um método não paramétrico que ajusta as probabilidades por meio da discretização do intervalo $[0, 1]$ em N_{bins} intervalos mutuamente exclusivos $B_1, B_2, \ldots, B_{N_{\text{bins}}}$. Para cada m-ésimo intervalo B_m, calcula-se a probabilidade calibrada $\hat{p}_m$ como a fração de amostras positivas no respectivo intervalo:

$$\hat{p}_m = \frac{\sum_{i \in B_m} y_i}{|B_m|}, \tag{5.7}$$

em que $|B_m|$ representa o número de amostras em B_m e $y_i \in \{0, 1\}$ denota a classe verdadeira da amostra i. A probabilidade calibrada associada a cada amostra é então:

$$\hat{p}_i = \hat{p}_m \quad \text{se} \quad p_i \in B_m. \tag{5.8}$$

A escolha do número de intervalos N_{bins} envolve compromisso entre viés e variância. Poucos intervalos produzem estimativas mais estáveis, porém potencialmente enviesadas, enquanto muitos intervalos aumentam a flexibilidade, mas elevam o risco de sobreajuste.

5.2.5 Temperature Scaling

O *Temperature Scaling* é um método simples e computacionalmente eficiente (Guo *et al.*, 2017). Ele introduz um parâmetro de temperatura $T > 0$ que reescala uniformemente os *logits* z_i antes da aplicação da *Softmax*:

$$\text{Softmax}(z_i, T) = \frac{e^{z_i/T}}{\sum_{j=1}^{C} e^{z_j/T}}, \tag{5.9}$$

em que C é o número de classes. Valores de $T > 1$ tornam a distribuição mais uniforme, reduzindo a superconfiança. Valores de $T < 1$, por sua vez, tornam-na mais concentrada.

No caso binário, a formulação equivalente é:

$$\text{Sigmoid}(z, T) = \frac{1}{1 + e^{-z/T}}. \tag{5.10}$$

O parâmetro T é estimado pela minimização do NLL em conjunto independente ou com *cross-validation*. Como a transformação é monotônica, o método preserva a ordem entre as classes e não altera métricas discriminativas como acurácia ou AUC (Jiang *et al.*, 2012).

5.2.6 Label Smoothing Regularizer

O LSR, proposto por Szegedy *et al.* (2016), é uma técnica de regularização aplicada durante o treinamento, com potencial efeito de calibração (Müller; Kornblith; Hinton, 2019). A técnica substitui classes binárias por distribuições suavizadas, reduzindo a superconfiança do modelo.

Para um problema com C classes, a classe suavizada é definida por:

$$y'_c = (1 - \epsilon)\, y_c + \frac{\epsilon}{C}, \tag{5.11}$$

em que $\epsilon \in [0, 1]$ é o parâmetro de suavização. No caso binário, $C = 2$.

Valores elevados de ϵ aumentam a entropia da distribuição-alvo, podendo comprometer a capacidade discriminativa quando excessivos.

5.2.7 Soft-labels por meio do NFCS

Uma limitação do LSR é o uso de um único valor de ϵ para todas as amostras (Vyas; Saxena; Voice, 2020). Como alternativa, utilizou-se neste trabalho a escala NFCS para definir *soft-labels* específicas por amostra.

A NFCS baseia-se na observação de características faciais do RN, e pontuações iguais ou superiores a três indicam presença de dor (Grunau; Craig, 1987). No banco UNIFESP, estão disponíveis anotações de cinco características faciais (Heiderich; Leslie; Guinsburg, 2015). Essas pontuações refletem o grau de certeza clínica e permitem incorporar a incerteza médica ao treinamento.

Foram propostos três métodos de mapeamento, todos normalizados em $[0, 1]$:

$$\text{SoftLinear}(NFCS) = 0{,}2 \times NFCS. \tag{5.12}$$

$$\text{SoftIntervalo}(NFCS) = \begin{cases} 0\%, & \text{se } NFCS \leq 1, \\ 30\%, & \text{se } NFCS = 2, \\ 70\%, & \text{se } NFCS = 3, \\ 100\%, & \text{se } NFCS \geq 4. \end{cases} \tag{5.13}$$

$$\text{SoftSigmoid}(NFCS) = \frac{1}{1 + e^{-NFCS+2{,}5}}. \tag{5.14}$$

Como apenas o banco UNIFESP contém anotações NFCS, o treinamento com *soft-labels* foi realizado exclusivamente com essas amostras. As divisões de *folds* da validação cruzada foram mantidas, removendo-se as imagens do banco iCOPE apenas na fase de treinamento, mas preservando-as na validação. Essa estratégia assegura comparabilidade entre experimentos e permite avaliar isoladamente o impacto das *soft-labels* na calibração e no desempenho classificatório.

5.2.8 Métricas de calibração

As métricas de calibração organizam-se em dois grupos. O primeiro avalia diretamente a correspondência entre as probabilidades preditas e a frequência empírica dos eventos, incluindo a curva de calibração, o *Expected Calibration Error* (ECE) e o *Maximum Calibration Error* (MCE). O segundo grupo, conhecido como *proper scoring rules* (Filho *et al.*, 2023), estima indiretamente o erro de calibração ao penalizar predições incorretas associadas à alta confiança, como o *Negative Log-Likelihood* (NLL) e o *Brier Score*.

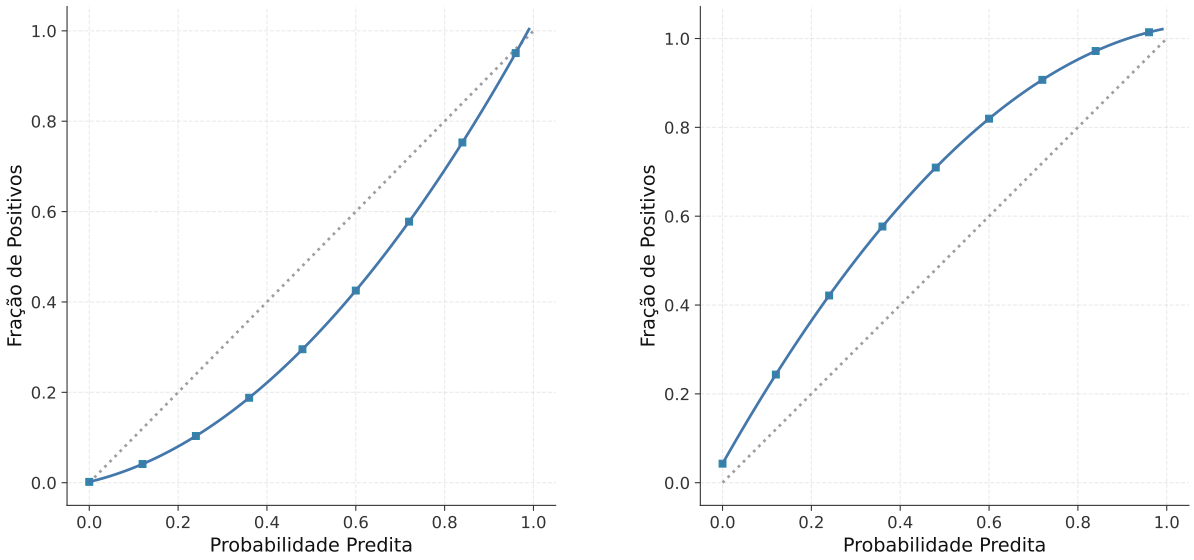
5.2.8.1 Curva de calibração

A curva de calibração é uma ferramenta visual utilizada para verificar a qualidade probabilística das predições de um modelo (DeGroot; Fienberg, 1983). Sua construção baseia-se na discretização das probabilidades preditas p em B intervalos. Para cada intervalo, calculam-se a média das probabilidades preditas (eixo x) e a fração de amostras positivas (eixo y), definida como a proporção de amostras da classe de interesse naquele intervalo.

A fração de amostras positivas difere da acurácia, pois enquanto a primeira expressa a frequência observada da classe positiva entre as amostras que recebem determinada probabilidade predita, a segunda descreve o desempenho global do modelo em todo o conjunto de dados.

Um modelo perfeitamente calibrado é representado por uma reta diagonal de 45°, na qual as probabilidades preditas coincidem com as frequências observadas. A Figura 20 ilustra os dois tipos de desvio: aqueles abaixo dessa linha (Figura 20a) indicam excesso de confiança (*overconfidence*), enquanto os desvios acima (Figura 20b) refletem subestimação (*underconfidence*).

Figura 20 – Curvas de calibração com excesso e falta de confiança.



(a) Excesso de confiança.

(b) Falta de confiança.

Fonte: Autor.

5.2.8.2 Expected Calibration Error

O ECE (Naeini; Cooper; Hauskrecht, 2015) é uma das métricas mais empregadas para quantificar a calibração em modelos de *deep learning* (Guo *et al.*, 2017; Nixon *et al.*, 2019). Ele mensura a discrepância média entre as probabilidades preditas e a frequência observada de eventos positivos, funcionando como uma estimativa numérica da curva de calibração.

O cálculo baseia-se na discretização das probabilidades preditas em B intervalos. Para cada intervalo i, obtêm-se a média das probabilidades preditas (p_i) e a fração de amostras positivas (o_i). O ECE é definido como a soma ponderada das diferenças absolutas entre esses valores, proporcional ao número de amostras n_i em cada intervalo:

$$ECE = \sum_{i=1}^{B} \frac{n_i}{N} \left| p_i - o_i \right|, \tag{5.15}$$

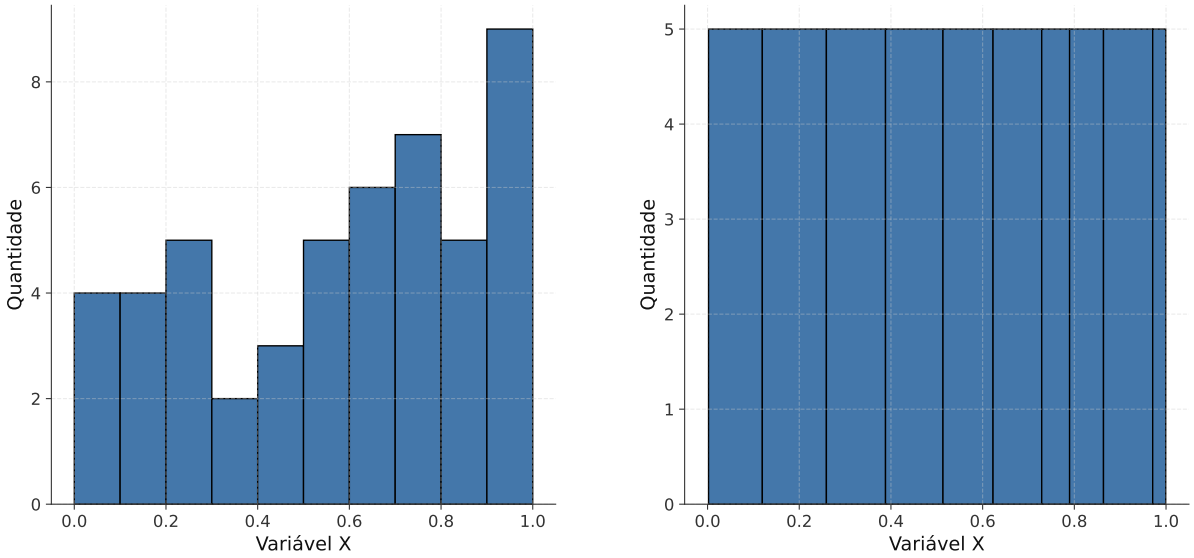
em que N representa o total de amostras. Valores menores indicam melhor calibração.

O *Maximum Calibration Error* (MCE), por outro lado, avalia o maior erro individual entre todos os intervalos:

$$MCE = \max_{i=1}^{B} \left(|p_i - o_i| \right). \tag{5.16}$$

Apesar de amplamente adotado, o ECE apresenta limitações. Como destacado por Nixon *et al.* (2019), a escolha do número de intervalos B é arbitrária e costuma variar entre 10 e 20. Um maior número de intervalos reduz o viés da estimativa, porém aumenta sua variância devido ao menor número de amostras por intervalo. A Figura 21 compara as duas abordagens. Para atenuar esse problema, Nixon *et al.* (2019) propuseram a discretização com número igual de amostras por intervalo (Figura 21b), em contraste com a abordagem tradicional de intervalos igualmente espaçados (Figura 21a). Neste trabalho, adotou-se a discretização com número igual de amostras.

Figura 21 – Métodos de discretização em intervalos.



(a) Intervalos igualmente espaçados.

(b) Intervalos com número igual de amostras.

Fonte: Autor.

5.2.8.3 Brier Score

O *Brier Score* (Brier, 1950) avalia simultaneamente a calibração e a acurácia, sendo definido como o erro quadrático médio entre a probabilidade predita e a classe real:

$$BrierScore = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2, \tag{5.17}$$

em que p_i representa a probabilidade predita para a amostra i e y_i a respectiva classe real. Valores menores indicam predições mais confiáveis probabilisticamente.

5.2.8.4 Negative log-likelihood

O NLL, também conhecido como erro de *Cross-Entropy*, é amplamente empregado como função de perda em modelos de *deep learning*. Ele penaliza fortemente predições incorretas associadas à alta confiança, oferecendo uma medida sensível da calibração probabilística (Hastie; Tibshirani; Friedman, 2009).

No caso binário, o NLL é definido como:

$$NLL = -\big[\, y \log(p) + (1-y) \log(1-p) \,\big], \tag{5.18}$$

em que p corresponde à probabilidade predita e y à classe real. Valores menores refletem melhor calibração.

Embora o NLL e o *Brier Score* mensurem a discrepância entre probabilidades e as classes reais, o NLL penaliza com maior intensidade as predições incorretas e excessivamente confiantes. Por exemplo, atribuir $p = 0{,}99$ a uma amostra negativa ($y = 0$) produz $NLL \approx 5$, enquanto o *Brier Score* é de aproximadamente $0{,}98$, exemplificando a maior sensibilidade do NLL a erros de alta confiança.

Apesar de o NLL ser empregado como função de otimização durante o treinamento, Guo *et al.* (2017) demonstraram que sua minimização não garante um modelo calibrado. Em redes neurais profundas, a otimização tende a favorecer a acurácia por meio de predições progressivamente mais confiantes, o que frequentemente resulta em sobreajuste e descalibração. Para mitigar esse efeito, técnicas como o LSR e o uso de *soft-labels* suavizam a função de perda, reduzindo o excesso de confiança sem comprometer o desempenho discriminativo.

5.3 EXPERIMENTO

O experimento deste capítulo foi conduzido a partir dos modelos com quantificação da incerteza selecionados no capítulo anterior. O objetivo foi avaliar o impacto dos métodos de calibração sobre a confiabilidade probabilística das predições e verificar se probabilidades mais altas correspondem, de fato, a expressões faciais mais intensas de dor neonatal.

Inicialmente, analisou-se a influência do número de intervalos B na métrica *Expected Calibration Error* (ECE), com o objetivo de definir um valor fundamentado, em vez de adotar uma escolha arbitrária, como frequentemente ocorre na literatura. Foram calculados valores de ECE sob diferentes configurações de B, examinando-se a estabilidade da métrica em função da granularidade dos intervalos. O valor final foi definido no ponto em que não se observaram variações relevantes nos desvios-padrão entre intervalos consecutivos de probabilidade.

A escolha do método considerou simultaneamente a calibração e a capacidade discriminativa. Selecionou-se o calibrador que apresentou diferença estatisticamente significativa no ECE em relação ao modelo não calibrado sem reduzir de forma relevante as métricas de desempenho. Esse critério procurou evitar que ganhos de calibração comprometessem a capacidade de detectar a dor.

Todos os parâmetros dos métodos de calibração foram ajustados por meio do mesmo procedimento de *cross-validation* em 10 *folds* empregado nos capítulos anteriores, assegurando comparabilidade entre modelos e consistência estatística das análises.

As métricas das três arquiteturas foram comparadas com o teste de Friedman (Friedman, 1937). Para avaliar o efeito da calibração dentro de cada arquitetura, comparou-se de forma pareada o modelo antes e depois da aplicação do calibrador por meio do teste de Wilcoxon (Wilcoxon, 1945). Em ambas as análises, adotou-se um nível de significância de 5% ($p < 0{,}05$).

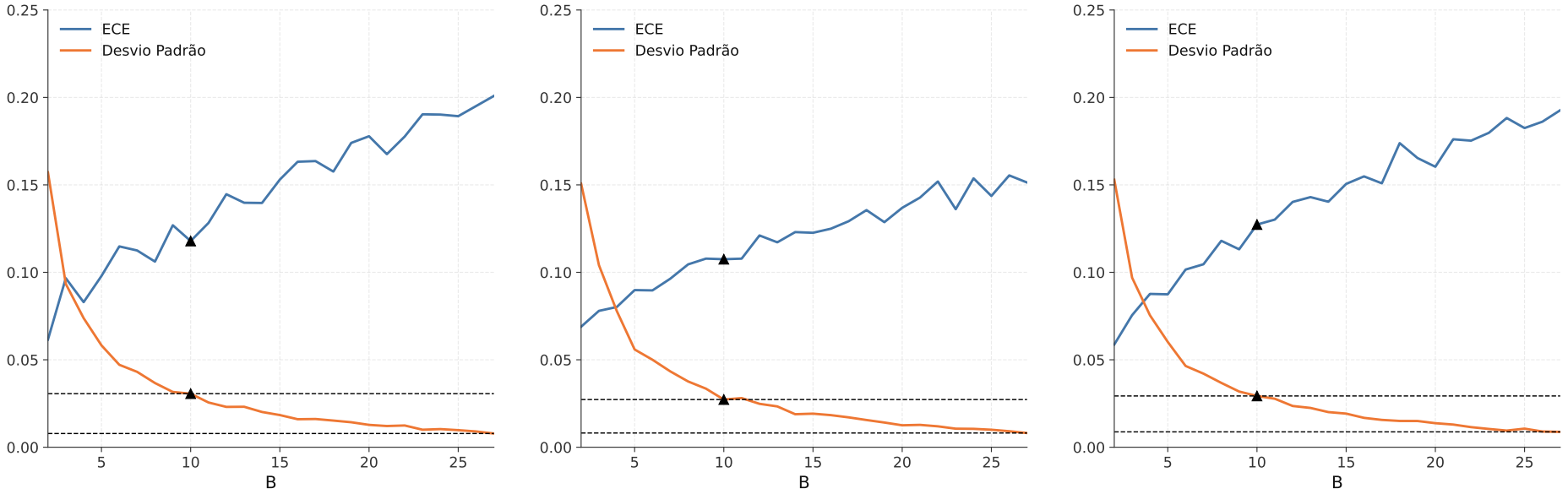
Por fim, investigou-se a correspondência entre probabilidades calibradas e expressões faciais dos recém-nascidos. Para esse fim, foram geradas faces médias agrupadas em intervalos de 10% da probabilidade predita, separando-se amostras corretamente e incorretamente classificadas. Essa análise qualitativa permitiu verificar se o aumento da probabilidade calibrada esteve associado a padrões faciais mais intensos de dor, contribuindo para a interpretação fisiológica e clínica dos resultados.

5.4 RESULTADOS

5.4.1 Avaliação dos calibradores

A Figura 22 apresenta a influência do número de intervalos B na métrica ECE para os modelos NCNN, VGGFace e ViT-B/32. Observa-se que, a partir de $B = 10$, o desvio-padrão entre os intervalos de probabilidade torna-se estável (diferenças < 0,025). A inclusão de mais intervalos aumentaria artificialmente o ECE em razão da redução do número de amostras por intervalo, sem ganho efetivo de precisão da métrica. Assim, adotou-se $B = 10$ nos experimentos subsequentes.

Figura 22 – Influência do número de intervalos no ECE.



(a) NCNN.

(b) VGGFace.

(c) ViT-B/32.

Fonte: Autor.

As métricas completas de classificação e calibração encontram-se no Apêndice B. A Tabela 9 sintetiza os resultados. As métricas da NCNN e da VGGFace permaneceram inalteradas, pois nenhum calibrador atingiu os critérios estabelecidos. No ViT-B/32, a única alteração estatisticamente significativa foi a melhora do ECE ($p = 0{,}037$), com a Regressão Isotônica. A VGGFace manteve as melhores médias de desempenho discriminativo e de calibração, embora o teste de Friedman não tenha indicado diferença estatística em relação aos outros modelos. O ViT-B/32 apresentou resultados intermediários, enquanto a NCNN obteve os menores valores de classificação e calibração.

Tabela 9 – Métricas de classificação e calibração dos modelos após quantificação da incerteza e calibração (média ± DP, em %).

Métrica	**NCNN** (MCDP p_{drop} = 30%, –)	**VGGFace** (MCDP p_{drop} = 50%, –)	**ViT-B/32** (10 Ens., Reg. Iso.)	**valor-p**
Acurácia	79,5 ± 8,7 =	**84,1 ± 5,0** =	82,7 ± 6,3 ↓	0,5754
F1-Score	74,8 ± 11,5 =	**81,5 ± 6,8** =	79,7 ± 6,8 ↓	0,5754
Sensibilidade	70,2 ± 17,2 =	**80,2 ± 10,3** =	75,0 ± 11,1 ↓	0,3474
Especificidade	87,7 ± 6,6 =	87,0 ± 8,5 =	**88,6 ± 8,2** =	0,8858
Precisão	83,0 ± 10,1 =	84,1 ± 8,8 =	**86,2 ± 5,9** ↑	0,7165
AUC	88,3 ± 6,4 =	**91,9 ± 3,4** =	89,8 ± 5,3 ↓	0,2725
ECE	12,9 ± 4,7 =	**10,3 ± 4,2** =	10,6 ± 3,7 ↑*	0,1495
MCE	**29,5 ± 12,6** =	33,5 ± 17,0 =	32,5 ± 14,7 ↑	0,4965
NLL	42,4 ± 10,1 =	**37,7 ± 9,3** =	41,7 ± 15,2 ↓	0,4965
Brier	13,7 ± 4,1 =	**11,7 ± 3,2** =	12,7 ± 4,1 ↓	0,3011

Fonte: Autor.

Legenda: Os símbolos indicam melhora (↑), piora (↓), melhora estatisticamente significativa (↑*), piora estatisticamente significativa (↓*) e resultado inalterado (=). Para ECE, MCE, NLL e Brier, reduções representam melhora. O asterisco indica $p < 0{,}05$ no teste de Wilcoxon pareado com a etapa anterior. A última coluna apresenta o teste de Friedman entre os modelos.

Os calibradores baseados na escala NFCS para geração de *soft-labels* apresentaram os piores resultados de classificação, conforme tabelas do Apêndice B. Esses métodos foram treinados exclusivamente com o banco UNIFESP, reduzindo substancialmente o número de imagens disponíveis e comprometendo a generalização. Em razão desse desempenho inferior, foram descartados das análises comparativas finais. Em relação à calibração, entretanto, apresentaram desempenho similar ao dos demais métodos e distribuições de probabilidades preditas mais dispersas e uniformes, como visualizado nas curvas de calibração apresentadas nas Figuras 23, 24 e 25.

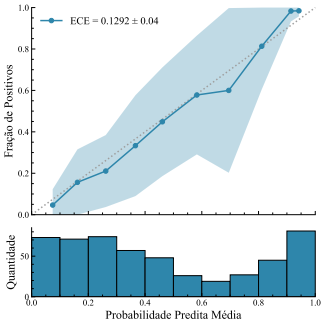
As Figuras 23, 24 e 25 complementam a análise quantitativa ao mostrar, para cada arquitetura, o alinhamento entre a probabilidade predita e a frequência observada de dor nos diferentes calibradores. Nas Figuras 23a, 24a e 25a, as curvas médias dos modelos originais permanecem próximas à calibração perfeita, embora apresentem comportamentos probabilísticos distintos. A NCNN e o ViT-B/32 produziram distribuições mais uniformemente dispersas,

enquanto a VGGFace concentrou as probabilidades preditas em torno de 0 e 1. A VGGFace, portanto, emitiu predições mais confiantes, mas que não necessariamente estão corretas.

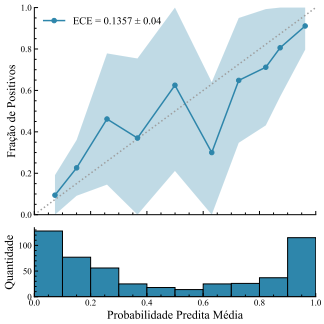
Nos diferentes níveis de LSR, verificou-se que o aumento da suavização das classes concentrou as probabilidades preditas em regiões próximas de 0,5, caracterizadas por menor confiabilidade. Esse efeito é evidenciado nas curvas de calibração e nos histogramas apresentados nas Figuras 23h, 24h e 25h. Como consequência, a calibração deteriorou-se à medida que o grau de suavização aumentou.

Entre os calibradores *post-hoc*, apenas no ViT-B/32 a Regressão Isotônica promoveu melhora nas métricas de calibração sem prejuízo relevante do desempenho discriminativo. Na NCNN e na VGGFace, nenhum calibrador atendeu simultaneamente aos critérios de calibração e discriminação estabelecidos, sugerindo que esses modelos já apresentavam boa calibração para os dados analisados.

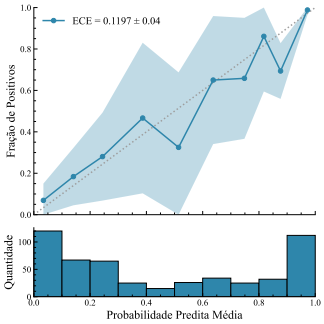
Figura 23 – Curvas de calibração da NCNN.



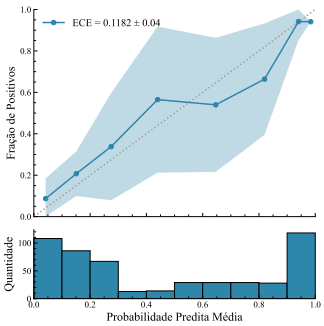
(a) Original.



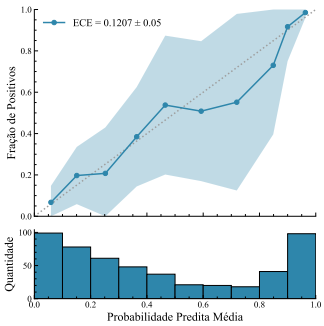
(b) Platt Scaling.



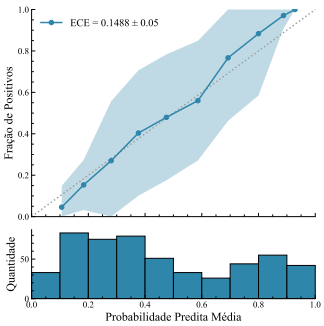
(c) Regressão Isotônica.



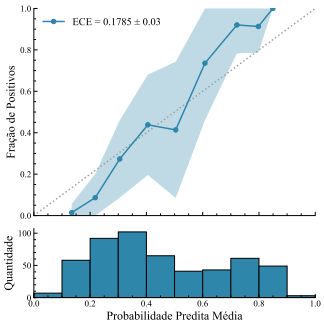
(d) *Histogram Binning*.



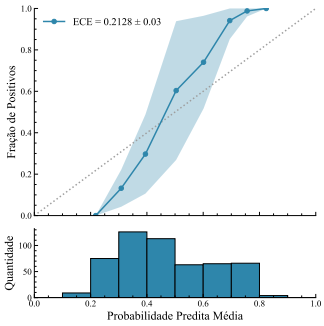
(e) *Temperature Scaling*.



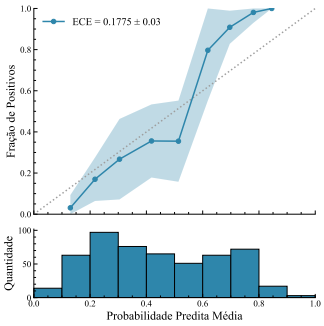
(f) LSR = 0,1.



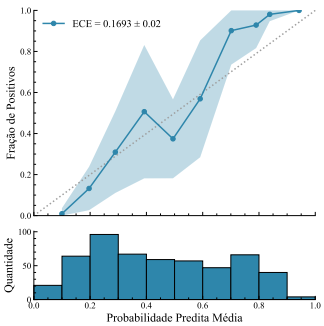
(g) LSR = 0,3.



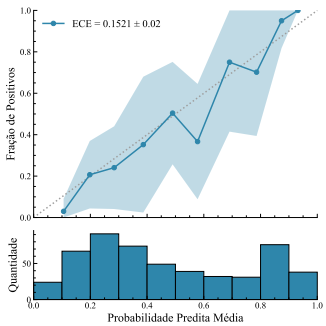
(h) LSR = 0,5.



(i) *Soft*-Sigmoid.



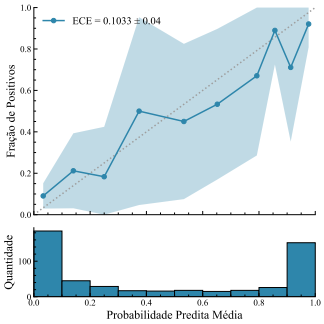
(j) *Soft*-Linear.



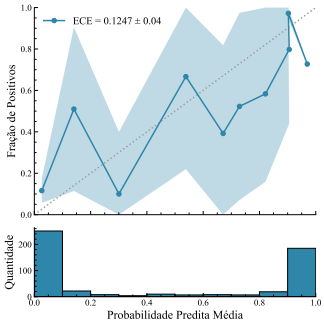
(k) *Soft*-Intervalos.

Fonte: Autor.

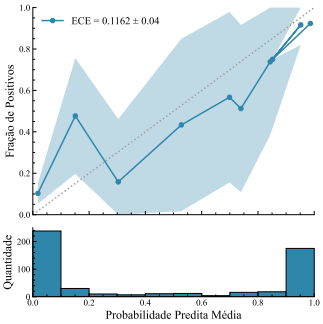
Figura 24 – Curvas de calibração da VGGFace.



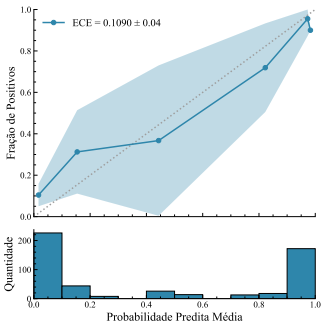
(a) Original.



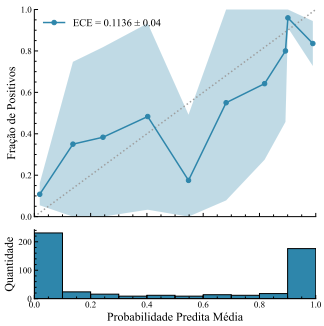
(b) Platt Scaling.



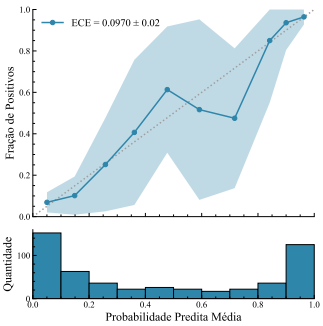
(c) Regressão Isotônica.



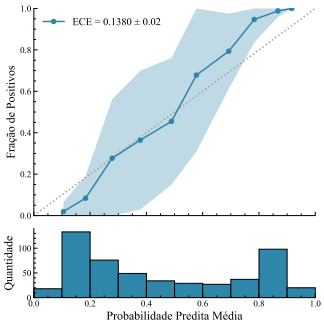
(d) *Histogram Binning*.



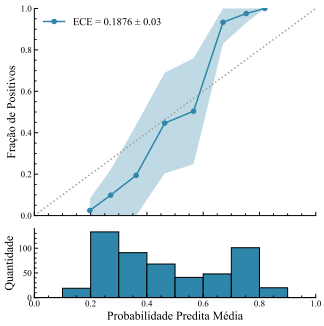
(e) *Temperature Scaling*.



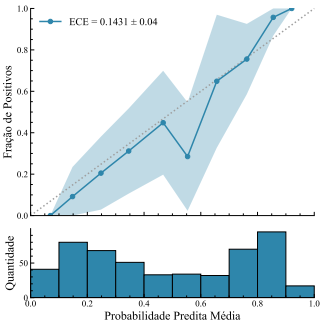
(f) LSR = 0,1.



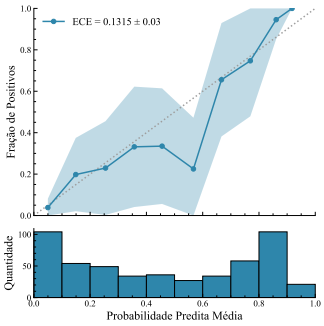
(g) LSR = 0,3.



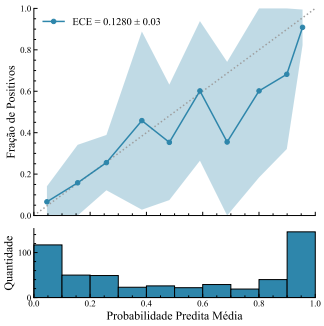
(h) LSR = 0,5.



(i) *Soft*-Sigmoid.



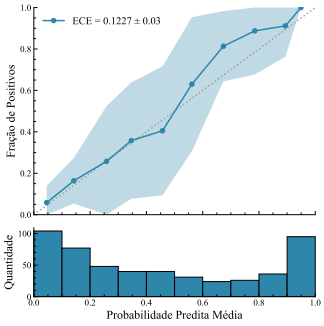
(j) *Soft*-Linear.



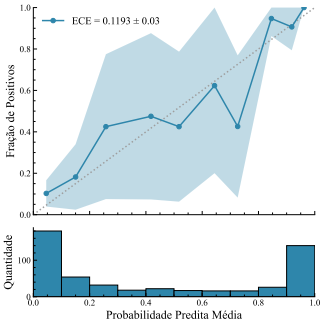
(k) *Soft*-Intervalos.

Fonte: Autor.

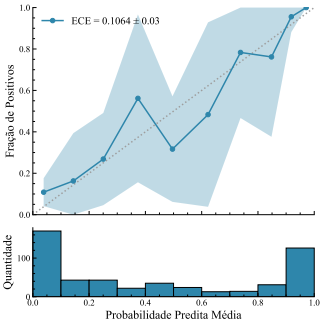
Figura 25 – Curvas de calibração do ViT-B/32.



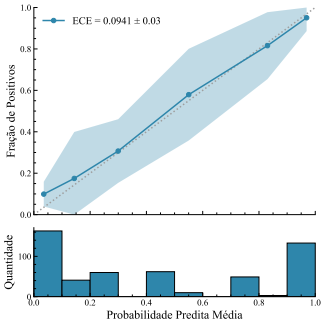
(a) Original.



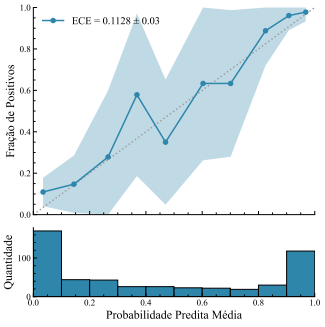
(b) Platt Scaling.



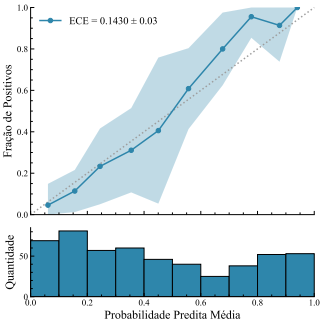
(c) Regressão Isotônica.



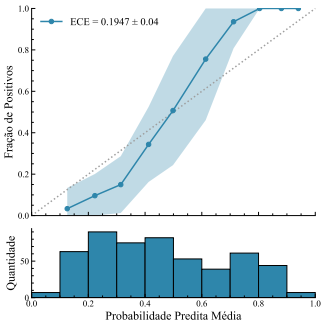
(d) *Histogram Binning*.



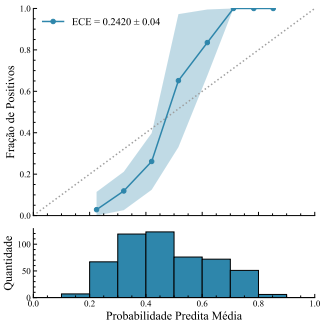
(e) *Temperature Scaling*.



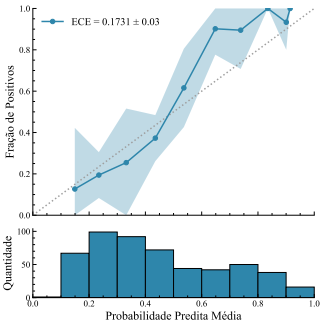
(f) LSR = 0,1.



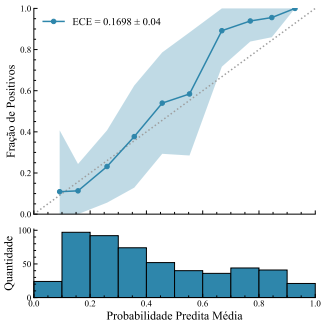
(g) LSR = 0,3.



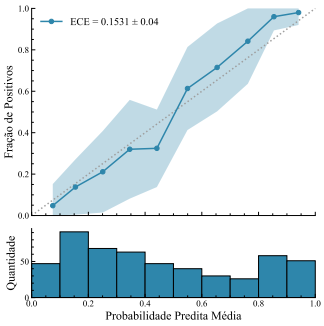
(h) LSR = 0,5.



(i) *Soft*-Sigmoid.



(j) *Soft*-Linear.



(k) *Soft*-Intervalos.

Fonte: Autor.

5.4.2 Análise qualitativa da calibração

Para investigar qualitativamente a calibração, foram geradas faces médias alinhadas em um referencial comum e agrupadas em intervalos de 10% da probabilidade predita. A Figura 26 separa os acertos e os erros de cada modelo. Nas predições corretas, a intensidade das expressões faciais de dor aumentou progressivamente com a probabilidade predita, principalmente em decorrência da abertura da boca, do cerramento dos olhos e do aprofundamento dos sulcos nasolabiais. Esse padrão é compatível com as escalas clínicas de dor neonatal: quanto mais presentes estiverem as características faciais associadas à dor, maior será a probabilidade estimada de o RN realmente estar com dor.

Nas predições incorretas, esse padrão foi menos acentuado. Em todos os intervalos, predominaram imagens com características faciais semelhantes. Esses achados evidenciam padrões faciais ambíguos entre as classes *com dor* e *sem dor*, associados às classificações incorretas. Essa ambiguidade evidencia a dificuldade da tarefa quando analisada de forma estática, uma vez que uma imagem isolada não fornece informações sobre a evolução temporal da condição observada. Assim, uma imagem isolada não permite determinar se a expressão facial registrada corresponde a um momento transitório de desconforto, a uma transição entre estados ou a um episódio de dor sustentado e contínuo.

5.5 DISCUSSÃO

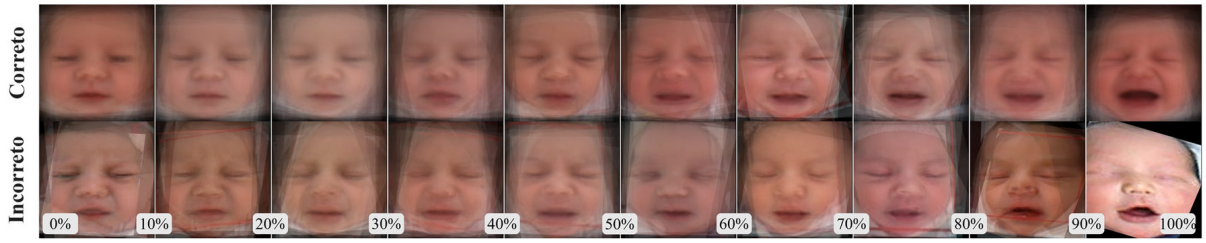
Os resultados evidenciam que desempenho discriminativo e qualidade probabilística permanecem propriedades distintas. Ainda é necessário avaliar em que medida as probabilidades preditas refletem adequadamente a realidade clínica. Nesse contexto, a calibração constitui um recurso para melhorar a consistência probabilística e a interpretação das predições produzidas pelos modelos.

A NCNN e a VGGFace já apresentaram comportamento probabilístico relativamente estável dentro dos critérios adotados, sem ganhos relevantes após calibração adicional. O ViT-B/32, por outro lado, beneficiou-se da Regressão Isotônica, indicando que arquiteturas distintas demandam estratégias específicas de ajuste probabilístico. Esse resultado reforça que não há um método universal de calibração aplicável de forma homogênea a diferentes modelos de *deep learning*, para a tarefa em questão.

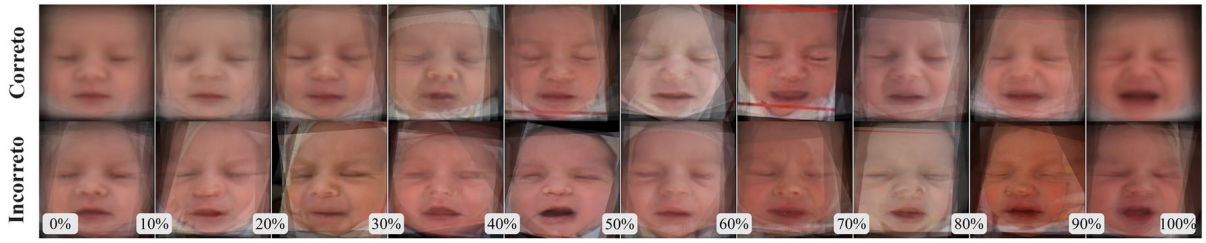
Entre as estratégias avaliadas, os métodos baseados na escala NFCS para a geração de *soft-labels* apresentaram desempenho inferior. Embora a incorporação explícita de conhecimento clínico ao treinamento seja uma abordagem promissora, sua aplicação ficou restrita ao banco UNIFESP, o que reduziu o volume efetivo de dados e, consequentemente, limitou a capacidade de generalização, conforme esperado. Apesar dessa limitação, esse método de calibração apresenta potencial para aumentar a interpretabilidade do modelo e incorporar explicitamente a incerteza clínica às predições, aspectos que podem ser explorados em trabalhos futuros.

Curvas de calibração e métricas como ECE e MCE mostraram-se importantes para

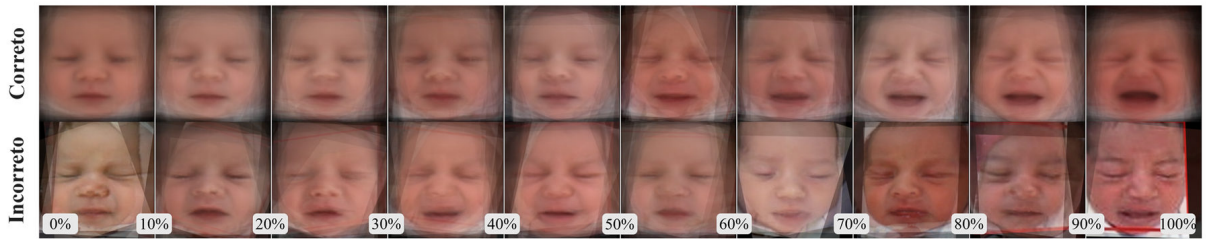
Figura 26 – Faces médias das predições em função da probabilidade predita.



(a) NCNN.



(b) VGGFace.



(c) ViT-B/32.

Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

monitorar a qualidade probabilística das predições. A VGGFace, por exemplo, apresentou ECE de $10{,}3\%$, indicando que as probabilidades preditas diferem, em média, cerca de 10 pontos percentuais da frequência empírica observada. No caso do MCE, discrepâncias locais podem atingir até 30 pontos percentuais, embora restritas a um intervalo específico de probabilidade, principalmente em regiões de baixa confiança. De forma geral, a curva de calibração fornece informações mais detalhadas sobre cada faixa de probabilidade predita e indica se ela está em uma região de excesso ou de falta de confiança, enquanto o ECE e o MCE sintetizam essas informações em um único valor.

A análise qualitativa das faces médias fornece evidências adicionais sobre a consistência probabilística dos modelos. Nas predições corretas, observou-se aumento progressivo da intensidade dos sinais faciais de dor à medida que a probabilidade predita aumentava. Esses achados sugerem que a escala probabilística produzida pelos modelos apresenta correspondência com fenótipos faciais clinicamente reconhecíveis e ressaltam a importância de explorar o fenômeno da dor em sua dimensão temporal, uma vez que imagens isoladas não fornecem informações sobre

a evolução da condição observada. Nas predições incorretas, esse gradiente não se manifestou de forma clara, e as imagens permaneceram visualmente ambíguas, independentemente do intervalo probabilístico analisado.

5.5.1 Implicações para a prática clínica

O uso de modelos de IA na prática clínica deve assumir caráter assistencial e complementar à avaliação humana, e não substituí-la. Para isso, os sistemas precisam produzir informações interpretáveis, consistentes e clinicamente confiáveis, permitindo sua integração ao processo de tomada de decisão.

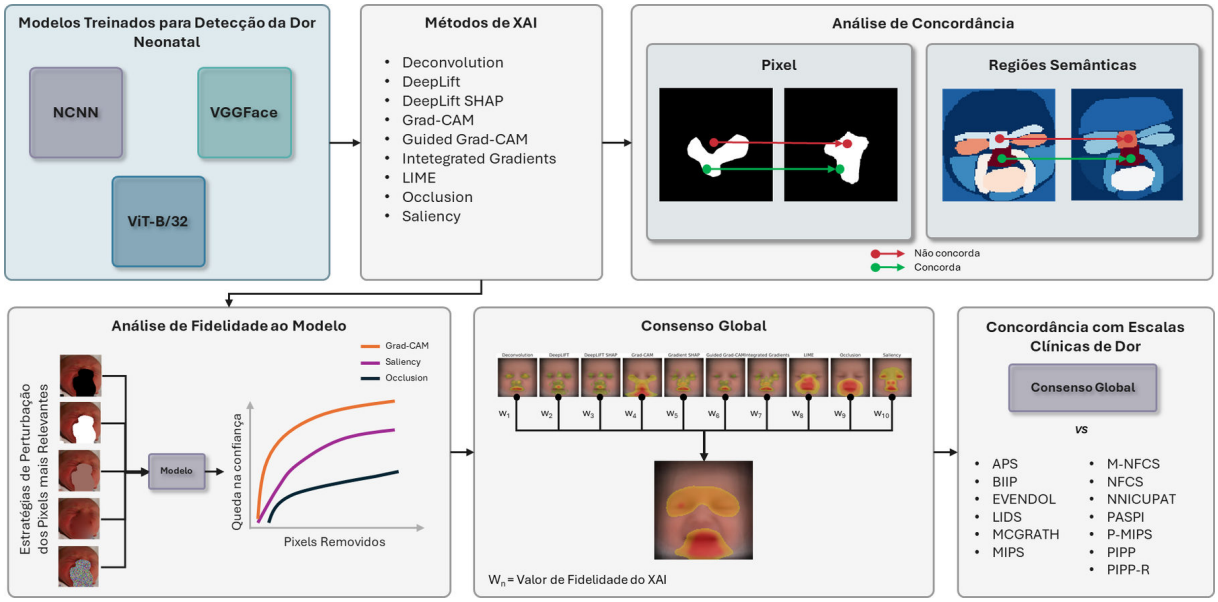
Nesse contexto, a calibração permite interpretar as probabilidades produzidas pelo modelo como indicadores mais confiáveis de risco clínico. Essa propriedade pode auxiliar a priorização assistencial, especialmente em ambientes como a UTIN, nos quais decisões rápidas e monitoramento contínuo devem ser priorizados. Em outras palavras, uma predição de 90% de probabilidade de dor indica que, em média, 9 em cada 10 RNs com características faciais semelhantes apresentarão dor, esses casos devem ser avaliados com prioridade. Por outro lado, uma predição de 10% pode indicar que apenas 1 em cada 10 RNs com características faciais semelhantes apresentará dor, o que admite monitoramento com menor urgência.

A correspondência entre probabilidades elevadas e expressões faciais mais intensas de dor sustenta a plausibilidade clínica das predições, embora os modelos não tenham sido treinados explicitamente para regressão. Essa associação também apoia a construção de um sinal contínuo da dor neonatal. Os resultados mostram que, além de distinguir presença e ausência de dor, os modelos organizam suas predições de forma probabilisticamente consistente com a intensidade das expressões faciais observadas. No monitoramento contínuo, a evolução temporal da dor pode ser mais informativa do que decisões binárias isoladas.

Do ponto de vista assistencial, modelos cujas probabilidades refletem padrões faciais clinicamente familiares são mais interpretáveis e potencialmente mais aceitos pelos profissionais de saúde. Essa transparência aumenta a confiança no sistema e favorece sua utilização como ferramenta de apoio à triagem, monitoramento e tomada de decisão clínica.

6 INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

Figura 27 – Visão geral da metodologia do Capítulo 6.



Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

6.1 INTRODUÇÃO

A etapa seguinte desenvolve o pilar de explicabilidade da metodologia e examina se as predições, além de probabilisticamente confiáveis, são clinicamente plausíveis. Este capítulo apresenta a Inteligência Artificial Explicável, do inglês *eXplainable Artificial Intelligence* (XAI), para analisar quais regiões faciais sustentam as predições dos modelos e se elas são compatíveis com o conhecimento clínico sobre a dor neonatal (Figura 27). A análise responde à terceira pergunta específica de pesquisa: as explicações produzidas pelos modelos são coerentes com regiões faciais reconhecidas na literatura e úteis para avaliar a plausibilidade clínica das predições?

Embora interpretação e explicação sejam frequentemente tratadas como sinônimos na literatura de XAI, os termos expressam dimensões distintas e complementares. Os métodos de explicação procuram evidenciar por que um modelo produziu determinada predição, enquanto a interpretação ocorre quando essas evidências adquirem significado para quem as examina, processo que depende de fatores cognitivos, do conhecimento prévio e do público ao qual a explicação se destina (Lipton, 2018; Miller, 2019).

Na visão computacional aplicada à medicina, por exemplo, mapas de saliência, também denominados mapas de atribuição, destacam os pixels que mais influenciam a saída do modelo (Arrieta *et al.*, 2020). Sua simples geração, contudo, não torna o resultado necessariamente compreensível nem garante sua utilidade clínica, pois, sem organização semântica e contextualização

adequada, esses mapas podem permanecer pouco interpretáveis para o usuário (Lipton, 2018). Essa limitação ganha relevância diante de marcos regulatórios recentes, como o *European Union AI Act* e a Resolução CFM n. 2.454/2026, que reforçam, em contextos sensíveis, requisitos como o direito a explicação do por quê uma decisão foi tomada (Goodman; Flaxman, 2017; European Union, 2024; Conselho Federal de Medicina, 2026).

A utilização de modelos de aprendizado profundo em contextos clínicos de alto risco impõe exigências adicionais de transparência, confiabilidade e auditabilidade (Arrieta *et al.*, 2020). Embora esses modelos possam atingir desempenho comparável ou superior ao de especialistas humanos, seus processos internos de decisão tendem a permanecer opacos (Chen *et al.*, 2022; Hossain *et al.*, 2025). Para esta tese, essa limitação é particularmente crítica, pois é necessário verificar se a predição de dor decorre de sinais faciais clinicamente plausíveis ou de padrões espúrios associados ao fundo da imagem, às condições de aquisição ou a artefatos presentes na UTIN.

A proliferação de métodos de explicação decorrente dos avanços regulatórios trouxe dúvidas sobre quais deles produzem resultados confiáveis. Um dos problemas identificados na literatura é o *disagreement problem*, em que diferentes explicadores podem destacar características distintas para o mesmo modelo e a mesma entrada, o que dificulta a interpretação e pode reduzir a confiança do usuário (Krishna *et al.*, 2022). Outra perspectiva relevante na avaliação de explicações é a fidelidade (*faithfulness*), por meio da qual se investiga se uma explicação reflete adequadamente o comportamento real do modelo (Yeh *et al.*, 2019).

Este capítulo tem como objetivo examinar em que medida as explicações produzidas para a detecção da dor neonatal são consistentes entre si, fiéis ao comportamento dos modelos e alinhadas às escalas clínicas validadas de dor neonatal. A análise foi organizada em quatro eixos: (I) discordância entre explicadores nos níveis de pixel e de região facial; (II) fidelidade por perturbação de pixels; (III) consenso global entre explicadores; e (IV) alinhamento com escalas clínicas de dor neonatal.

6.2 MÉTODOS

6.2.1 Explainable artificial intelligence

Para esta pesquisa, foram selecionados 10 explicadores comumente utilizados em XAI para visão computacional (Hossain *et al.*, 2025), conforme descrito a seguir:

a) **Deconvolution:** descrito inicialmente por Zeiler e Fergus (2014), realiza a retropropagação apenas de sinais positivos e reverte operações convolucionais para revelar estruturas da entrada que ativam fortemente filtros específicos;

b) **DeepLIFT (*Deep Learning Important FeaTures*):** proposto por Shrikumar, Greenside e Kundaje (2017), compara ativações neuronais com uma entrada de referência e retropropaga as respostas do modelo, permitindo atribuições mesmo em situações de gradientes nulos;

c) **DeepLIFT SHAP:** combina o DeepLIFT com o arcabouço de *Shapley Additive Explanations* (SHAP) (Lundberg; Lee, 2017), por meio da média de múltiplas avaliações sobre diferentes amostras de base, para aproximar valores de Shapley e reduzir a variabilidade das atribuições;

d) **Grad-CAM (*Gradient-weighted Class Activation Mapping*):** introduzido por Selvaraju *et al.* (2017), utiliza gradientes com respeito aos mapas de ativação da última camada convolucional para produzir mapas de atribuição que localizam regiões espaciais mais discriminativas;

e) **Gradient SHAP:** proposto por Lundberg e Lee (2017), estende o Integrated Gradients ao calcular a média de atribuições baseadas em gradiente ao longo de múltiplas interpolações ruidosas entre a imagem de referência e a imagem de entrada, gerando explicações mais suaves e estáveis;

f) **Guided Grad-CAM:** combina Guided Backpropagation com Grad-CAM (Selvaraju *et al.*, 2017), unindo a localização em alto nível com gradientes em nível de pixel para produzir mapas de atribuição mais detalhados;

g) **Integrated Gradients:** introduzido por Sundararajan, Taly e Yan (2017), integra os gradientes ao longo de um caminho linear entre uma imagem de referência e a entrada, gerando atribuições que atendem aos axiomas de sensibilidade e invariância de implementação;

h) **LIME (*Local Interpretable Model-Agnostic Explanations*):** o método introduzido por Ribeiro, Singh e Guestrin (2016) aproxima localmente o comportamento do modelo por meio de um classificador linear esparso, identificando as características que mais influenciam a predição na vizinhança imediata da entrada;

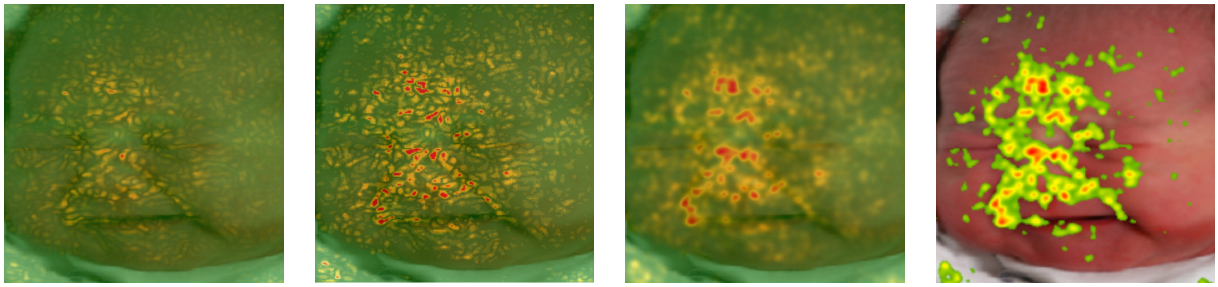
i) **Occlusion:** também descrito por Zeiler e Fergus (2014), mascara pequenos blocos da imagem e mede a variação na resposta do modelo, fornecendo uma estimativa direta da importância regional;

j) **Saliency:** proposto por Simonyan, Vedaldi e Zisserman (2013), calcula o gradiente da classe-alvo em relação a cada pixel da imagem de entrada, destacando regiões nas quais pequenas perturbações influenciam mais fortemente a predição.

6.2.2 Pós-processamento dos mapas de atribuição

Apesar de sua utilidade, os mapas de atribuição podem ser difíceis de visualizar e compreender. Carlini *et al.* (2024) propuseram uma técnica que aproxima sua apresentação dos mapas de saliência visual humana e facilita sua visualização e comparação. Essa técnica foi aplicada neste trabalho para padronizar os mapas de atribuição, como exemplificado na Figura 28.

Figura 28 – Pós-processamento aplicado ao *Integrated Gradients*.



(a) Mapeamento de cores. (b) Agrupamento pelo *k-means*. (c) Suavização com filtro gaussiano. (d) Adição do canal de transparência α.

Fonte: Adaptado de Carlini (2023).

Primeiro, definiu-se um esquema de cores para representar os níveis de ativação do modelo: verde para baixa, amarelo para média e vermelho para alta ativação (Figura 28a). Em seguida, o algoritmo *k-means* agrupou os pixels do mapa em cinco grupos, correspondentes aos três níveis de ativação e a dois grupos de transição entre cores (Figura 28b). Aplicou-se então um filtro Gaussiano de tamanho 11×11 e desvio-padrão 2 para suavizar as bordas e as transições (Figura 28c). Por fim, adicionou-se ao mapa um canal de transparência (α), que permite ocultar áreas com baixo valor de atribuição (Figura 28d). Para determinar quais pixels seriam ocultados, utilizou-se o seguinte critério no mapa pós-processado:

$$\alpha_{\text{channel}} = \begin{cases} 1 & \text{se } L_i \geq \mu_L + c \times \sigma_L \\ 0 & \text{se } L_i < \mu_L + c \times \sigma_L \end{cases} \tag{6.1}$$

em que L_i é o valor do *pixel* na posição i do mapa de atribuição, e μ_L e σ_L são, respectivamente, a média e o desvio-padrão do mapa. O valor escalar c foi definido experimentalmente como 0,5 (Carlini, 2023; Carlini *et al.*, 2024). Um valor igual a 1 no canal de transparência torna o *pixel* totalmente visível e um valor igual a 0 o torna totalmente oculto.

6.2.3 Discordância entre explicadores

A discordância foi avaliada em dois níveis complementares: pixel e regiões faciais semanticamente definidas (Pirie *et al.*, 2026). A análise em nível de pixel quantifica se dois explicadores destacam evidências espacialmente coincidentes. Seja $\mathcal{P} = \{1, \ldots, H\} \times \{1, \ldots, W\}$ o conjunto de coordenadas de pixels de um mapa de saliência S, e seja S_{ij} o valor de saliência no pixel (i, j). Para cada mapa, apenas os $k\%$ pixels mais relevantes são mantidos com base no percentil τ_k:

$$\tau_k = \text{Percentile}\big(\{S_{ij} \mid (i,j) \in \mathcal{P}\},\, 100 - k\big), \tag{6.2}$$

produzindo o mapa binário

$$M = \begin{cases} 1, & \text{se } S_{ij} \geq \tau_k, \\ 0, & \text{caso contrário.} \end{cases} \tag{6.3}$$

A concordância em nível de pixel entre dois explicadores $E^{(a)}$ e $E^{(b)}$ é então definida por:

$$\text{PixelAgreement}(E^{(a)}, E^{(b)}) = \frac{M^{(a)} \cap M^{(b)}}{M^{(a)} \cup M^{(b)}}. \tag{6.4}$$

Como a avaliação clínica de dor neonatal feita por humanos não é realizada em nível de pixel, a discordância também foi analisada em nível de região facial. Cada pixel foi associado a uma região facial semanticamente definida por meio de um mapeamento $\phi : \{1, \ldots, H\} \times \{1, \ldots, W\} \to \mathcal{R}$, em que $\mathcal{R} = \{r_1, \ldots, r_R\}$ representa o conjunto de regiões faciais. Para uma região $r \in \mathcal{R}$, define-se

$$\Omega_r = \big\{(i, j) \mid \phi(i, j) = r\big\}. \tag{6.5}$$

A importância regional é calculada pela média da saliência na região:

$$A_r = \frac{1}{|\Omega_r|} \sum_{(i,j)\in\Omega_r} S_{ij}. \tag{6.6}$$

Sejam $Top_k^{(a)} = [r_1^{(a)}, \ldots, r_k^{(a)}]$ e $Top_k^{(b)} = [r_1^{(b)}, \ldots, r_k^{(b)}]$ as listas das k regiões mais relevantes para dois explicadores. A concordância regional é dada por:

$$\text{RegionAgreement}(E^{(a)}, E^{(b)}) = \frac{\left|Top_k^{(a)} \cap Top_k^{(b)}\right|}{k}. \tag{6.7}$$

O escore de concordância em nível de região representa a fração de regiões que aparecem simultaneamente nos conjuntos top-k de ambas as explicações. Um valor igual a 1 indica concordância perfeita, enquanto um valor igual a 0 indica discordância completa. A avaliação em nível de região reduz a instabilidade em nível de pixel e está alinhada ao raciocínio clínico, que se baseia em regiões faciais, e não em pixels individuais.

6.2.4 Análise de fidelidade por perturbação de pixels

A fidelidade foi avaliada com base na premissa de que uma explicação é mais fiel quando os pixels por ela considerados mais importantes, ao serem perturbados, produzem maior degradação da predição (Ferreira *et al.*, 2027). Neste trabalho, os pixels foram perturbados na ordem *Most Relevant First* (MoRF), conforme os valores absolutos de atribuição de cada explicador.

Como não existe um protocolo de referência consolidado para esse problema em dor neonatal, foram comparadas cinco estratégias de substituição de pixels: preto, branco, média,

ruído Gaussiano e ROAD (*RemOve And Debias*) (Rong *et al.*, 2022). Exemplos são apresentados na Figura 29.

Figura 29 – Estratégias de perturbação avaliadas.



Fonte: Autor, utilizando banco de imagens iCOPE (Brahnam *et al.*, 2006).

A fidelidade foi quantificada pela métrica *Area Over the Perturbation Curve* (AOPC) (Samek *et al.*, 2017):

$$\mathrm{AOPC} = \frac{1}{L}\Big\langle \sum_{k=1}^{L} \big[\mathbb{F}(x) - \mathbb{F}(x^{(k)})\big] \Big\rangle, \tag{6.8}$$

em que $\mathbb{F}(x)$ representa a probabilidade de dor predita para a imagem original, $\mathbb{F}(x^{(k)})$ representa a probabilidade após a perturbação dos k pixels mais relevantes, $\langle\cdot\rangle$ denota a média sobre o conjunto de dados, e L representa o número total de etapas de perturbação. Valores mais altos indicam explicações mais fiéis.

Para evitar interpretações baseadas apenas no efeito bruto da perturbação, o AOPC foi centralizado em relação a um *baseline* aleatório, no qual o mesmo número de pixels é perturbado em ordem aleatória, em vez da ordem MoRF. Valores positivos indicam que os pixels priorizados pelo explicador afetam mais a predição do que pixels escolhidos aleatoriamente; valores próximos de zero sugerem que não há diferença relevante; e valores negativos indicam que a perturbação em si tem maior impacto do que a ordem definida pelo explicador. A ordem aleatória foi previamente fixada para todos os modelos e explicadores, o que evita que variações na sequência de perturbação influenciem a comparação.

6.2.5 Consenso global entre explicadores

Diante da discordância entre explicadores descrita em Krishna *et al.* (2022), foi calculado um mapa de consenso global por média ponderada, no qual cada explicador contribui de acordo com seu desempenho em fidelidade. Dado um conjunto de mapas de saliência $(S_1, S_2, \ldots, S_N)$ e seus respectivos valores de métrica $(m_1, m_2, \ldots, m_N)$, o consenso é calculado conforme a Equação (6.9):

$$S_{\text{global}} = \frac{\sum_{i=1}^{N} m_i \, S_i}{\sum_{i=1}^{N} m_i}, \tag{6.9}$$

em que S_{global} representa o mapa de saliência de consenso que enfatiza os pixels consistentemente destacados pelos explicadores de melhor desempenho. Como cada explicador produz valores de saliência em escalas diferentes, todos os mapas foram normalizados para o intervalo $[-1, 1]$ antes da média ponderada.

6.2.6 Concordância com escalas clínicas de dor

Para avaliar a consistência entre os modelos e o conhecimento clínico, as importâncias ranqueadas das regiões obtidas pelos métodos de explicação foram comparadas com as regiões faciais presentes em escalas neonatais de dor utilizadas para detecção de dor. Seja o conjunto de regiões faciais $\mathcal{R} = \{r_1, r_2, \ldots, r_R\}$, com as correspondentes importâncias regionais fornecidas pelos explicadores $A = (A_1, A_2, \ldots, A_R)$. As regiões foram ordenadas em ordem decrescente de importância, atribuindo-se a posição 1 à região mais relevante.

Foi definido um escore C que recompensa regiões enfatizadas tanto pelo explicador quanto pela escala clínica, ao mesmo tempo em que penaliza regiões altamente ranqueadas pelos explicadores, mas ausentes na escala. Seja $g_r \in \{0, 1\}$ um indicador de se a região r está incluída na escala de dor. O escore de concordância é definido pela Equação (6.10):

$$C = \sum_{r=1}^{R} (R + 1 - \text{rank}(r))(2g_r - 1), \tag{6.10}$$

em que $(R + 1 - \text{rank}(r))$ inverte o ranqueamento, de modo que valores maiores correspondam a maior importância, e $(2g_r - 1)$ atribui contribuições positivas ou negativas dependendo de a região facial estar ou não presente na escala clínica de dor.

Para garantir comparabilidade entre escalas com diferentes números de regiões, o escore foi normalizado pelo seu valor máximo teórico $C_{\max}$, calculado pela Equação (6.11) e correspondente ao caso em que todas as regiões clinicamente relevantes ocupam as P primeiras posições do ranking, com $P = \sum_r g_r$:

$$C_{\max} = 2 \left[P(R + 1) - \frac{P(P + 1)}{2} \right] - \frac{R(R + 1)}{2}. \tag{6.11}$$

O escore de concordância normalizado é dado pela Equação (6.12):

$$C_{\text{norm}} = \frac{C + C_{\max}}{2C_{\max}}, \tag{6.12}$$

em que $C_{\text{norm}} \in [0, 1]$. O valor 1 indica concordância perfeita com a escala clínica, 0 indica discordância completa, e $0{,}5$ representa uma correspondência neutra ou aleatória. Essa métrica

foi projetada para garantir a comparabilidade entre escalas com diferentes números de regiões, mantendo a mesma escala de comparação. Para ser conservadora, a métrica penaliza regiões ausentes das escalas estabelecidas e prioriza o conhecimento clínico validado, necessário à construção inicial de confiança em ambientes críticos de UTIN.

6.3 EXPERIMENTO

O experimento deste capítulo teve como objetivo comparar explicadores sob critérios simultaneamente relevantes do ponto de vista computacional e clínico. Os dez explicadores foram implementados com a biblioteca Python Captum[1], utilizando seus hiperparâmetros padrão. Para cada imagem de validação dos 10 *folds* e para cada modelo, foi gerado um mapa de atribuição escalado para o intervalo $[-1, 1]$, visto que alguns métodos produzem atribuições negativas, isto é, características que reduzem a predição do modelo. Após a geração, foi aplicado o pós-processamento descrito na Seção 6.2.2. Nas métricas de concordância e fidelidade, utilizou-se a magnitude das atribuições, de modo a avaliar a relevância atribuída pelo modelo independentemente do sinal da contribuição.

Na análise de discordância, a concordância em nível de pixel foi calculada com top-$k = 10\%$, e a concordância em nível de região, com top-$k = 3$ (Pirie *et al.*, 2026). Também foram realizados testes de sanidade (Adebayo *et al.*, 2018), nos quais cada explicador foi comparado com: (I) a mesma arquitetura inicializada com pesos aleatórios; (II) imagens de ruído com distribuição normal; e (III) mapas obtidos por um detector de bordas Canny. O objetivo foi verificar se as explicações respondiam aos parâmetros aprendidos, e não apenas a padrões estruturais triviais.

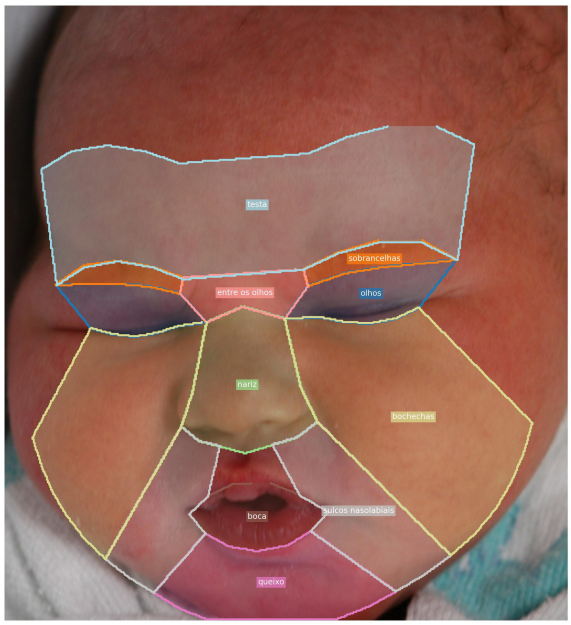
Para a análise em nível de região, as regiões faciais foram derivadas automaticamente dos 106 pontos fiduciais obtidos na detecção facial descrita no Capítulo 3, conforme a Figura 30. As características simétricas dos lados esquerdo e direito, como os olhos, foram combinadas em uma única região, pois a literatura clínica sobre dor neonatal não diferencia os lados do rosto. Foram consideradas nove regiões faciais: olhos, região entre os olhos, testa, sulcos nasolabiais, boca, nariz, queixo, bochechas e sobrancelhas. O fundo da imagem, que não corresponde a uma região facial, foi incluído como décima região para avaliar se os explicadores destacam evidências fora do rosto.

Na avaliação de fidelidade, os pixels top-k foram perturbados em ordem MoRF segundo a magnitude das atribuições. A saída avaliada correspondeu sempre à probabilidade predita pelo modelo para a classe *com dor*. Foram comparadas as cinco estratégias de perturbação apresentadas na Figura 29, e a fidelidade foi medida por AOPC centralizado em relação à perturbação aleatória.

Em seguida, os valores de AOPC foram utilizados para calcular o consenso global por média ponderada, conforme definido na Seção 6.2.5. Esse consenso foi comparado a um consenso

[1] https://captum.ai/

Figura 30 – Regiões faciais derivadas automaticamente dos pontos fiduciais.



Fonte: Autor, utilizando banco de imagens iCOPE (Brahnam *et al.*, 2006).

obtido por média simples, para verificar se a incorporação da fidelidade melhora a qualidade da explicação agregada.

Por fim, quantificou-se a concordância entre o consenso global dos explicadores e 13 escalas neonatais de dor clinicamente validadas. A Tabela 10 apresenta as escalas consideradas e as respectivas regiões faciais utilizadas. As escalas são descritas no Capítulo 2.

Tabela 10 – Presença de regiões faciais em escalas neonatais de dor.

Região	**APS**	**BIIP**	**EVENDOL**	**LIDS**	**MCGRATH**	**MIPS**	**M-NFCS**	**NFCS**	**NNICUPAT**	**PASPI**	**P-MIPS**	**PIPP**	**PIPP-R**	**Total**
Olhos	1	1	1	1	1	1	1	1	1	1	1	1	1	13
Entre os olhos	0	1	1	1	1	1	1	1	1	1	1	1	1	12
Testa	0	1	1	1	1	1	1	1	1	1	1	1	1	12
Sulcos nasolabiais	1	1	0	0	1	1	1	1	1	1	1	1	1	11
Boca	0	1	1	1	1	1	1	1	1	1	1	0	0	10
Queixo	0	0	0	1	1	1	0	1	0	0	1	0	0	5
Sobrancelhas	1	0	0	1	0	0	0	0	0	0	0	0	0	2
Total	3	5	4	6	6	6	5	6	5	5	6	4	4	65

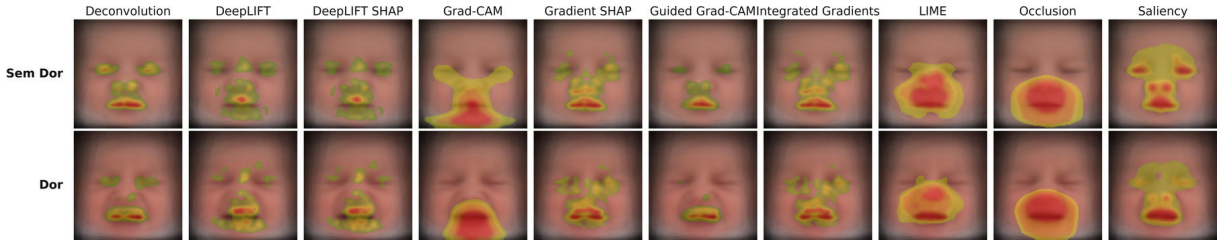
Fonte: Autor.

6.4 RESULTADOS

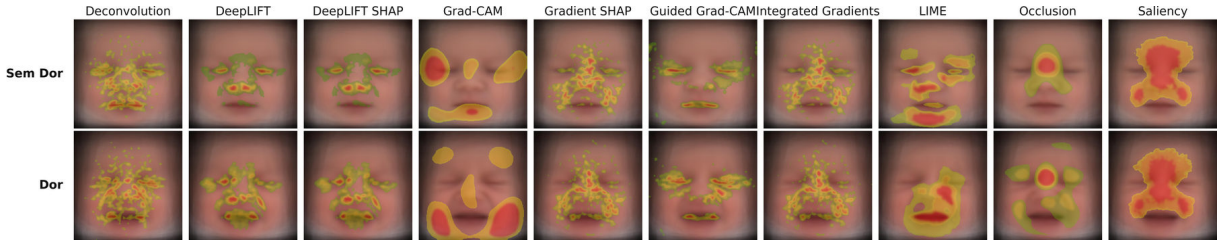
6.4.1 Discordância entre explicadores

A Figura 31 apresenta os mapas de saliência médios de todos os explicadores nos três modelos para as classes *Dor* e *Sem dor*. Todas as atribuições dos explicadores e as imagens faciais foram alinhadas em uma referência comum com base na extração de pontos-chave da detecção facial e, posteriormente, submetidas à média. Observa-se o problema de discordância entre explicadores, uma vez que diferentes métodos atribuíram níveis distintos de importância a

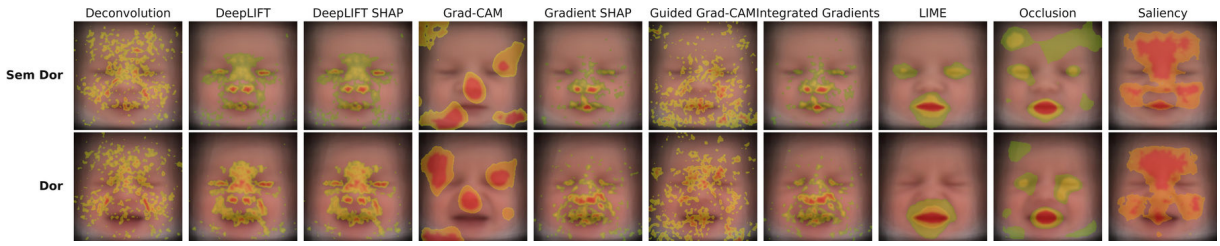
Figura 31 – Explicações médias e faces médias para as classes *Dor* e *Sem dor*.



(a) NCNN.



(b) VGGFace.



(c) ViT-B/32.

Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

diversas regiões da imagem. Métodos como Grad-CAM, LIME e Occlusion produziram padrões mais estruturados de *superpixels* (grupos de pixels contíguos), enquanto outros explicadores geraram atribuições mais dispersas.

Tabela 11 – Explicadores com concordância média em nível de pixel superior a 0,50.

NCNN	**VGGFace**	**ViT-B/32**
DL & DLS = 1,00	**DL & DLS = 1,00**	**S & D = 0,69**
DL & Não treinado DL = 1,00	IG & GS = 0,99	**IG & GS = 0,69**
DLS & Não treinado DLS = 1,00	Não treinado D & D = 0,79	-
IG & GS = 0,84	-	-
DL & ED = 0,77	-	-
DLS & ED = 0,77	-	-

Fonte: Autor.

Legenda: DL = DeepLIFT, DLS = DeepLIFT SHAP, ED = Edge Detector, IG = Integrated Gradients, GS = Gradient SHAP, D = Deconvolution, S = Saliency.

A Tabela 11 apresenta, para os três modelos, as concordâncias em nível de pixel superiores

Tabela 12 – Explicadores com pelo menos duas regiões concordantes na avaliação regional.

NCNN	VGGFace	ViT-B/32
DL & DLS = 1,00	**DL & DLS = 1,00**	**DLS & DL = 1,00**
IG & GS = 0,95	**IG & GS = 1,00**	S & D = 0,87
S & IG = 0,85	Não treinado D & D = 0,83	IG & GS = 0,86
S & GS = 0,84	Não treinado DL & DL = 0,80	Não treinado DL & DL = 0,71
GS & D = 0,78	Não treinado DLS & DLS = 0,80	Não treinado DLS & DLS = 0,71
IG & D = 0,78	DL & D = 0,76	-
S & D = 0,76	DLS & D = 0,76	-
DL & D = 0,67	Não treinado GS & GS = 0,73	-
DLS & D = 0,67	S & GS = 0,71	-
-	S & IG = 0,71	-
-	Não treinado IG & IG = 0,71	-

Fonte: Autor.

Legenda: DL = DeepLIFT, DLS = DeepLIFT SHAP, IG = Integrated Gradients, GS = Gradient SHAP, D = Deconvolution, S = Saliency.

a 0,50, o que indica que mais da metade dos pixels destacados ocupa as mesmas posições. A Tabela 12 complementa essa análise com concordâncias regionais superiores a 0,66, isto é, pelo menos duas das três regiões faciais foram identificadas por ambos os explicadores comparados. Os resultados completos de concordância são apresentados no Apêndice C. De modo geral, a concordância foi maior em nível de região do que em nível de pixel. Em outras palavras, embora os explicadores difiram quanto à posição exata dos pixels mais relevantes, eles convergem para as mesmas regiões faciais.

Ao comparar os explicadores com os testes de sanidade, observou-se que DeepLIFT e DeepLIFT SHAP apresentaram, de forma consistente, maior concordância em nível de pixel e de região em todos os modelos quando comparados às atribuições provenientes de redes não treinadas e, no caso da NCNN, também em relação à detecção de bordas. Para a VGGFace, também foi observada concordância em nível de região entre as versões treinadas e não treinadas de Deconvolution, Gradient SHAP e Integrated Gradients. Em geral, a baixa concordância entre a maioria dos explicadores e os testes de sanidade indicou que as explicações são guiadas por parâmetros aprendidos, e não por artefatos ou padrões arquiteturais dos modelos. Os testes de sanidade baseados em ruído produziram concordância negligenciável em nível de pixel ou de região (tipicamente <0,10 e <0,30, respectivamente), confirmando que os mapas de saliência não consistem simplesmente em ruído.

Na análise entre explicadores, métodos baseados em gradiente, incluindo Integrated Gradients, Gradient SHAP, DeepLIFT, DeepLIFT SHAP e Saliency, apresentaram maior concordância entre si do que métodos baseados em *superpixels*, como Grad-CAM, Occlusion e LIME, os quais exibiram pouca ou nenhuma concordância entre si ou com outros explicadores. Os pares mais consistentes foram DeepLIFT & DeepLIFT SHAP e Integrated Gradients & Gradient

SHAP, refletindo semelhanças em seus mecanismos de operação internos.

Entre os modelos, a NCNN apresentou maior número de pares com concordância em nível de pixel, enquanto a VGGFace e, especialmente, o ViT-B/32 apresentaram menos pares concordantes, indicando maior variabilidade em seus padrões espaciais de atribuição.

6.4.2 Fidelidade por perturbação de pixels

Para todos os modelos, a estratégia ROAD produziu consistentemente valores positivos de fidelidade AOPC (Tabela 13). Esse resultado era esperado, pois, na estratégia ROAD, os pixels são perturbados por meio de uma interpolação sem viés que preserva estatísticas locais da imagem (Rong *et al.*, 2022). Em contraste, outras estratégias de perturbação introduzem formas ou bordas artificiais que podem revelar informações de classe (Rong *et al.*, 2022). Por exemplo, substituições por valores preto e branco, particularmente em CNNs, resultaram em valores negativos de AOPC, indicando que o desempenho do modelo se degradou mais rapidamente em razão dos próprios artefatos da perturbação do que da relevância das atribuições fornecidas pelo método explicativo. Para testar a hipótese nula de que todas as estratégias de perturbação apresentam desempenho equivalente, foi aplicado o teste de Friedman (Friedman, 1937), obtendo-se um valor-p < 0,001 para todos os modelos.

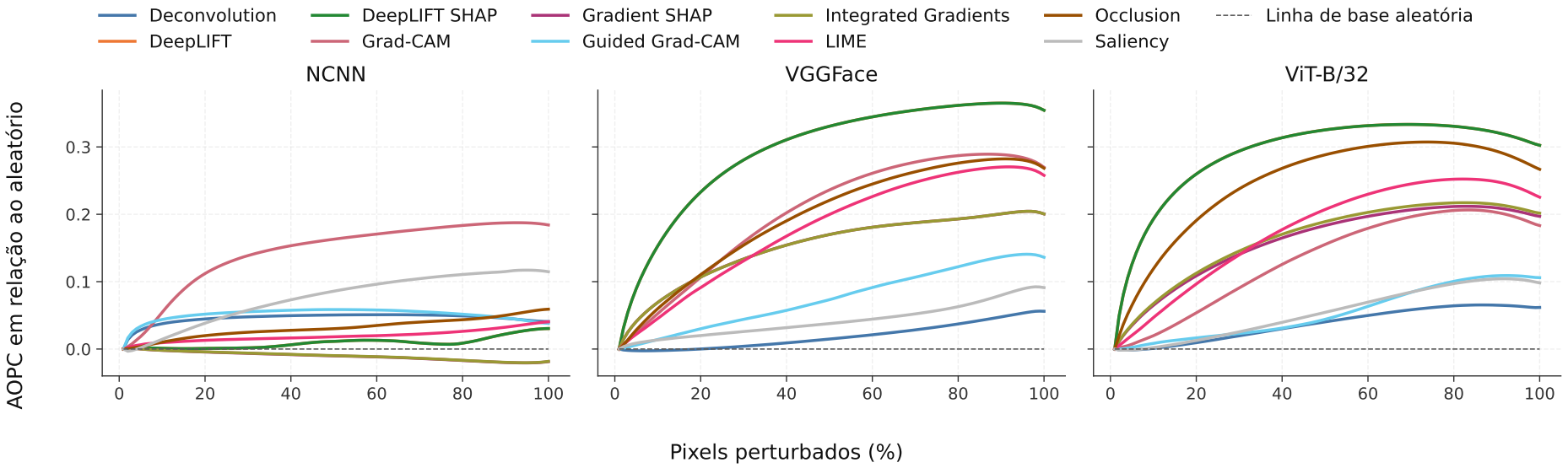
Tabela 13 – Ranking das estratégias de perturbação por modelo (AOPC entre parênteses).

Posição	**NCNN**	**VGGFace**	**ViT-B/32**
1º	ROAD (0,05)	ROAD (0,22)	Branco (0,20)
2º	Gaussiano (0,04)	Média (-0,11)	ROAD (0,19)
3º	Média (-0,12)	Gaussiano (-0,16)	Preto (0,14)
4º	Preto (-0,20)	Preto (-0,38)	Média (0,05)
5º	Branco (-0,24)	Branco (-0,42)	Gaussiano (-0,17)
valor-p	< 0,001	< 0,001	< 0,001

Fonte: Autor.

Os resultados da avaliação de fidelidade são apresentados na Figura 32 e na Tabela 14. Um valor alto de AOPC indica que a perturbação dos pixels mais relevantes altera significativamente a predição, sugerindo que os pixels removidos correspondem a características utilizadas pelo modelo. Assim, uma curva que cresce rapidamente indica maior influência dos pixels priorizados pelo explicador sobre a predição.

Figura 32 – AOPC da estratégia ROAD em relação à perturbação aleatória.



Fonte: Autor.

Tabela 14 – Ranking dos explicadores por modelo para a estratégia ROAD (AOPC entre parênteses).

Posição	**NCNN**	**VGGFace**	**ViT-B/32**
1º	Grad-CAM (0,18)	DeepLIFT (0,35)	DeepLIFT (0,30)
2º	Saliency (0,11)	DeepLIFT SHAP (0,35)	DeepLIFT SHAP (0,30)
3º	Occlusion (0,06)	Grad-CAM (0,27)	Occlusion (0,27)
4º	Deconvolution (0,04)	Occlusion (0,27)	LIME (0,23)
5º	Guided Grad-CAM (0,04)	LIME (0,26)	Gradient SHAP (0,20)
6º	LIME (0,04)	Gradient SHAP (0,20)	Integrated Gradients (0,20)
7º	DeepLIFT (0,03)	Integrated Gradients (0,20)	Grad-CAM (0,18)
8º	DeepLIFT SHAP (0,03)	Guided Grad-CAM (0,14)	Guided Grad-CAM (0,11)
9º	Gradient SHAP (–0,02)	Saliency (0,09)	Saliency (0,10)
10º	Integrated Gradients (–0,02)	Deconvolution (0,06)	Deconvolution (0,06)
valor-p	$< 0{,}001$	$< 0{,}001$	$< 0{,}001$

Fonte: Autor.

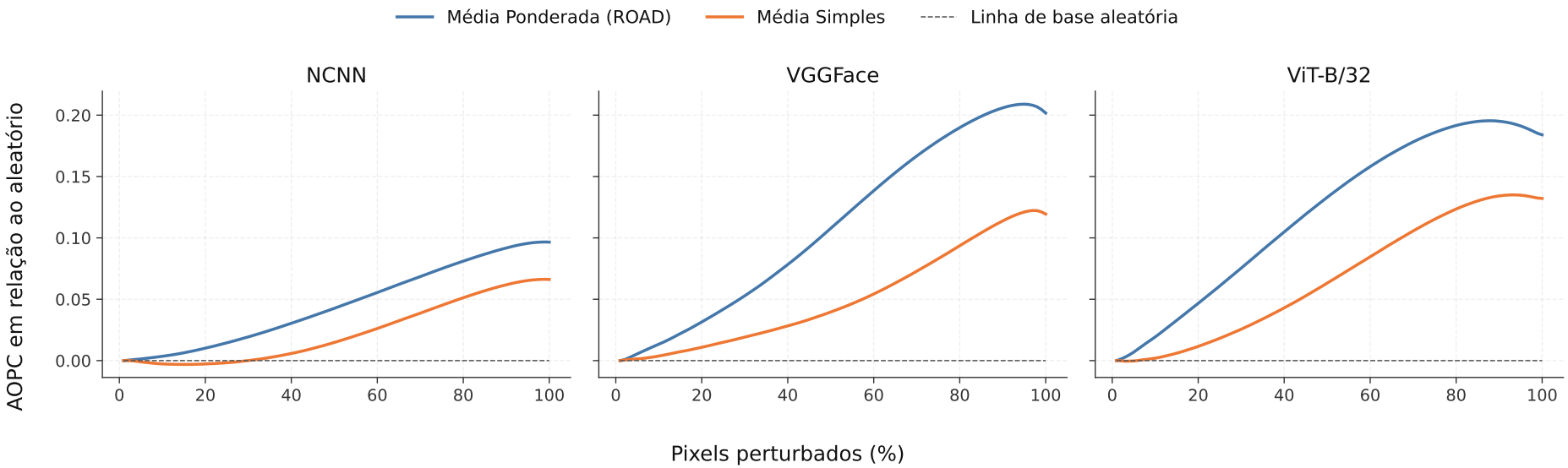
Ao utilizar ROAD como estratégia padrão de perturbação, nenhum explicador se destacou como o mais fiel em todos os modelos, embora diferenças estatisticamente significativas tenham sido observadas segundo o teste de Friedman. A forte concordância entre os pares DeepLIFT & DeepLIFT SHAP e Integrated Gradients & Gradient SHAP refletiu-se em valores idênticos de fidelidade. O modelo NCNN apresentou os explicadores menos fiéis, com valores de AOPC próximos de zero, indicando pouca diferença em relação à perturbação de pixels aleatórios. Em contraste, VGGFace e ViT-B/32 exibiram valores positivos mais elevados, demonstrando que seus explicadores foram mais fiéis ao comportamento do modelo, particularmente aqueles classificados nas primeiras posições.

Adicionalmente, os explicadores DeepLIFT & DeepLIFT SHAP aplicados aos modelos VGGFace e ViT-B/32 apresentaram os maiores valores de AOPC, apesar da elevada concordância entre os mapas gerados pelos modelos treinados e por seus correspondentes não treinados. Em conjunto, esses resultados sustentam a interpretação de que tais métodos capturam, ao menos em parte, padrões determinados pela própria arquitetura dos modelos, e não apenas pelas representações aprendidas durante o treinamento. A perturbação das regiões destacadas afeta negativamente a predição, possivelmente por interferir em elementos estruturalmente relevantes para o processamento interno dos modelos.

6.4.3 Consenso global entre explicadores

As explicações de consenso global obtidas pela média ponderada dos explicadores com base em seus valores de AOPC foram mais fiéis do que o consenso por média simples, conforme a Figura 33. Embora tenha superado a média simples, o consenso ponderado não foi consistentemente superior aos explicadores individuais mais bem classificados na Tabela 14.

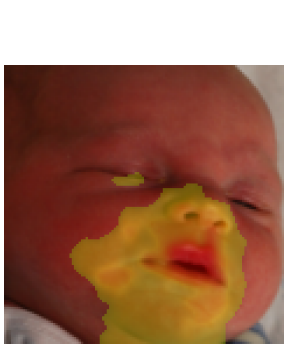
Figura 33 – AOPC do consenso global com média simples e ponderada.



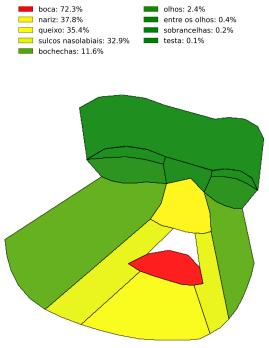
Fonte: Autor.

A Figura 34 apresenta exemplos do consenso global ponderado pelo AOPC e do consenso por média simples, além de indicar as regiões faciais consideradas mais importantes, ilustrando as diferenças entre as duas abordagens de consenso. As figuras mostram que a média ponderada concentra as explicações em pontos de maior relevância, como a boca, reduzindo as atribuições dispersas em regiões menos relevantes da face.

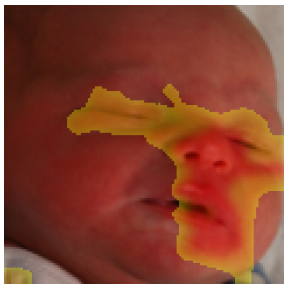
Figura 34 – Consenso global por média simples e ponderação por AOPC.



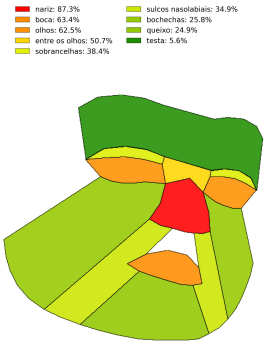
(a) NCNN – média simples.



(b) NCNN – média simples (regiões).



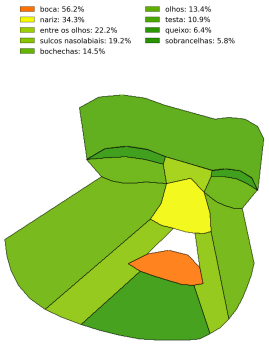
(c) NCNN – ponderação por AOPC.



(d) NCNN – AOPC (regiões).



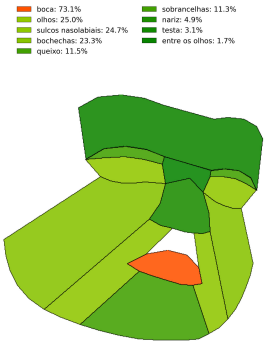
(e) VGGFace – média simples.



(f) VGGFace – média simples (regiões).



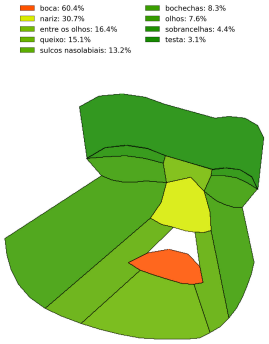
(g) VGGFace – ponderação por AOPC.



(h) VGGFace – AOPC (regiões).



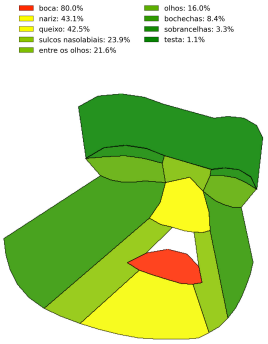
(i) ViT-B/32 – média simples.



(j) ViT-B/32 – média simples (regiões).



(k) ViT-B/32 – ponderação por AOPC.



(l) ViT-B/32 – AOPC (regiões).

Fonte: Autor, utilizando banco de imagens iCOPE (Brahnam *et al.*, 2006).

Nota: As porcentagens indicam a importância individual de cada região facial. Por isso, a soma difere de 100%.

6.4.4 Concordância global com escalas clínicas de dor

A Tabela 15 apresenta o ranking de importância das regiões faciais derivado do consenso ponderado para cada arquitetura. Para a NCNN, a boca é a região mais enfatizada, seguida pelo nariz e pelo queixo. Na VGGFace, o nariz e a região entre os olhos foram priorizados, enquanto a boca aparece apenas na terceira posição. O modelo ViT-B/32 também destacou a boca e o nariz, porém atribuiu importância substancial a áreas fora da face, sugerindo maior sensibilidade a artefatos contextuais ou ao fundo da imagem.

Tabela 15 – Ranking de importância das regiões faciais (em %).

Posição	**NCNN**	**VGGFace**	**ViT-B/32**
1º	Boca (32,2)	Nariz (29,9)	Boca (30,0)
2º	Nariz (23,2)	Entre os olhos (24,7)	Queixo (13,9)
3º	Queixo (17,7)	Boca (13,1)	Nariz (13,6)
4º	Sulcos nasolabiais (17,0)	Sulcos nasolabiais (11,6)	Fora da face (10,0)
5º	Olhos (6,2)	Olhos (9,9)	Sulcos nasolabiais (9,1)
6º	Entre os olhos (2,0)	Bochechas (3,6)	Olhos (8,8)
7º	Sobrancelhas (1,0)	Queixo (3,4)	Entre os olhos (8,4)
8º	Bochechas (0,8)	Sobrancelhas (2,3)	Bochechas (3,8)
9º	Fora da face (0,1)	Testa (1,4)	Sobrancelhas (1,5)
10º	Testa (0,0)	Fora da face (0,1)	Testa (0,9)

Fonte: Autor.

Tabela 16 – Concordância global entre o consenso dos explicadores e escalas clínicas de dor.

Escala de dor	**NCNN**	**VGGFace**	**ViT-B/32**
APS	0,59	0,49	0,37
BIIP	0,73	0,67	0,60
EVENDOL	0,65	0,62	0,55
LIDS	0,73	0,67	0,63
MCGRATH	0,77	0,68	0,66
MIPS	0,77	0,68	0,66
M-NFCS	0,73	0,67	0,60
NFCS	0,77	0,68	0,66
NNICUPAT	0,73	0,67	0,60
PASPI	0,73	0,67	0,60
PIPP	0,62	0,59	0,50
PIPP-R	0,62	0,59	0,50
P-MIPS	0,77	0,68	0,66
Média	**0,71**	0,64	0,59

Fonte: Autor.

A maior concordância média foi alcançada pela NCNN (0,71) (Tabela 16). Suas explicações concentraram-se principalmente na boca (32,2%), no nariz (23,2%), no queixo (17,7%) e

nos sulcos nasolabiais (17,0%). Entre essas regiões, a boca e os sulcos nasolabiais são frequentemente representados nas escalas neonatais de dor, enquanto o queixo também é incluído em diversas escalas. A NCNN também atribuiu importância mínima a regiões não representadas nas escalas, como as bochechas (0,8%), áreas fora da face (0,1%) e a testa (0,0%), o que contribuiu para seu maior nível de alinhamento clínico.

Uma concordância intermediária foi observada para a VGGFace (0,64), cujas explicações enfatizaram o nariz (29,9%), a região entre os olhos (24,7%), a boca (13,1%) e os sulcos nasolabiais (11,6%). Embora o nariz não esteja explicitamente incluído na maioria das escalas neonatais de dor, a forte ênfase na região entre os olhos, presente em diversas escalas, e na boca contribui para a manutenção de alinhamento moderado com regiões faciais clinicamente relevantes.

A menor concordância foi observada para o ViT-B/32 (0,59). Suas explicações destacaram a boca (30,0%), o queixo (13,9%) e o nariz (13,6%), mas também atribuíram importância considerável a áreas fora da face (10,0%), que não fazem parte das escalas neonatais de dor. A atribuição de relevância a regiões não clínicas reduz sua concordância global com as escalas.

De forma geral, conforme apresentado na Tabela 16, os valores de concordância mostraram-se relativamente homogêneos entre explicadores e entre escalas. A escala APS apresentou consistentemente os menores valores de concordância em todos os modelos, comportamento esperado, pois inclui apenas três regiões faciais (olhos, sobrancelhas e sulcos nasolabiais), o que limita o grau máximo de concordância possível com as regiões mais frequentemente priorizadas pelas explicações dos modelos.

6.5 DISCUSSÃO

Os resultados confirmam a presença do *disagreement problem* (Krishna *et al.*, 2022) no contexto da detecção da dor neonatal. Em nível de pixel, a concordância entre explicadores foi geralmente baixa, com poucos agrupamentos estáveis. Os pares Integrated Gradients & Gradient SHAP e DeepLIFT & DeepLIFT SHAP destacaram-se pela maior convergência, o que é coerente com a proximidade metodológica entre esses métodos. Esse achado sustenta a interpretação de que explicadores baseados em princípios semelhantes apresentam redundância funcional.

Quando as explicações foram agregadas em regiões faciais semanticamente definidas, a concordância aumentou substancialmente em todos os modelos. Esse resultado mostra que, embora os explicadores tenham discordado quanto à localização exata dos pixels mais relevantes, convergiram para regiões faciais semelhantes. Consequentemente, a análise em nível de região mostrou-se mais estável e mais compatível com o raciocínio clínico, que se baseia em regiões faciais e não em evidência pixel a pixel.

Em relação à comparação entre modelos, a NCNN apresentou maior número de pares concordantes, por operar com representações menos complexas devido à sua baixa complexidade e profundidade de camadas. VGGFace e ViT-B/32 exibiram maior variabilidade explicativa,

sugerindo que arquiteturas mais profundas ou baseadas em atenção sustentam múltiplos padrões de atribuição.

A avaliação de fidelidade indicou que a ROAD (Rong *et al.*, 2022) foi a estratégia de perturbação mais consistente. Em contraste, substituições por valores constantes, média ou ruído podem introduzir bordas artificiais e alterar padrões locais de forma não realista, afetando desproporcionalmente modelos sensíveis a texturas e estruturas espaciais. O comportamento do ViT-B/32 diferiu do observado nos modelos convolucionais. Nesse caso, os valores positivos de AOPC produzidos por diferentes estratégias evidenciaram maior robustez a perturbações locais nas condições avaliadas. Esse resultado é compatível com estudos prévios sobre ViTs (Chefer; Gur; Wolf, 2021; Paul; Chen, 2022), nos quais a modelagem global da imagem reduz a dependência de padrões locais específicos, tornando o ViT-B/32 mais robusto a perturbações de pixels individuais.

Entre os modelos avaliados, a NCNN apresentou os menores valores de fidelidade, frequentemente próximos de zero, mostrando que a remoção dos pixels classificados como mais relevantes pouco se diferenciou de uma perturbação aleatória. Em contraste, VGGFace e ViT-B/32 exibiram valores mais elevados e ordenações relativamente consistentes entre explicadores. Ainda assim, os resultados de fidelidade devem ser interpretados em conjunto com os testes de sanidade. DeepLIFT e DeepLIFT SHAP, por exemplo, apresentaram bom desempenho em AOPC para VGGFace e ViT-B/32, mas também elevada concordância com modelos não treinados, indicando que parte das atribuições pode refletir vieses arquiteturais que influenciam a predição do modelo, em vez de características aprendidas relacionadas à dor neonatal.

De modo geral, Grad-CAM, LIME e Occlusion figuraram entre os explicadores mais consistentes em termos de fidelidade. Uma possível explicação é que esses métodos produzem mapas de relevância mais agregados espacialmente, resultando em perturbações estruturalmente mais coerentes e, consequentemente, maior degradação da predição do modelo quando regiões relevantes são removidas. Contudo, como esses métodos dependem fortemente de escolhas de parametrização, esses resultados devem ser considerados à luz do uso de hiperparâmetros padrão e da ausência de uma análise sistemática de sensibilidade.

O consenso global ponderado por fidelidade mostrou que a agregação melhora a qualidade geral da explicação, embora não supere consistentemente os melhores explicadores individuais. Ainda assim, a apresentação de uma única explicação de consenso, em vez de múltiplas explicações individuais, reduz a carga cognitiva do usuário e facilita a interpretação clínica. Essa abordagem também evita que a equipe médica seja responsável por selecionar um único explicador com base apenas em critérios qualitativos ou em preferência subjetiva pela explicação apresentada, mesmo quando esta não representa fielmente o comportamento do modelo.

Por fim, as pontuações de concordância com escalas clínicas de dor neonatal situaram-se majoritariamente entre 0,60 e 0,70. Esse resultado indica que os modelos priorizaram regiões faciais clinicamente relevantes, apesar de essas escalas não terem sido utilizadas durante o treinamento. As regiões que mais contribuíram para essa concordância foram principalmente

boca, olhos, testa e a região entre os olhos, todas compatíveis com unidades de ação facial descritas em escalas clínicas de dor neonatal (Tamanaka *et al.*, 2023). Entretanto, regiões como nariz, bochechas e áreas fora da face também foram frequentemente destacadas, embora não estejam explicitamente representadas nas escalas clínicas. Esse resultado admite duas interpretações não mutuamente exclusivas: (I) tais regiões podem conter sinais faciais adicionais ainda pouco formalizados na literatura clínica, ou (II) podem refletir a influência de fatores contextuais presentes em imagens de unidades de terapia intensiva neonatal, como cobertores, dispositivos médicos e variações de iluminação.

6.5.1 Implicações para a prática clínica

Os achados deste capítulo indicam que a contribuição clínica da explicabilidade não reside na apresentação isolada de mapas de atribuição, mas na tradução do comportamento do modelo para uma forma compatível com o raciocínio empregado pelos profissionais na avaliação da dor neonatal. A XAI atua como uma camada de apoio à decisão: além da probabilidade de dor, o sistema informa quais regiões faciais sustentaram a predição e se a distribuição de relevância é coerente com sinais reconhecidos na prática clínica. Dessa forma, a explicação passa a apoiar a validação e a auditoria do modelo, além de favorecer a comunicação entre o sistema e o usuário clínico e contribuir para a interpretabilidade de suas predições.

Para esse uso, o consenso global organizado em regiões faciais semânticas mostrou-se o formato mais adequado entre os avaliados. Em vez de interpretar múltiplos mapas em nível de pixel, potencialmente instáveis e dependentes do explicador, o médico pode receber uma síntese por regiões, como boca, olhos, testa, nariz, queixo, bochechas e áreas externas à face. Essa representação aproxima a saída do modelo das escalas clínicas de dor neonatal, que se baseiam em sinais anatômicos observáveis, e não em pixels individuais. O consenso ponderado por fidelidade também reduz a variabilidade entre explicadores e privilegia os métodos mais aderentes ao comportamento do modelo, oferecendo uma explicação única e mais simples para o uso à beira do leito.

Na prática, essa organização ajuda a explicitar as limitações do sistema. Como a VGGFace tende a priorizar a região central da face, por exemplo, sondas, mãos, tampões oculares ou outros elementos que ocultem essa área podem reduzir a confiabilidade da predição. Regras operacionais, como "não utilizar o sistema se mais de X% da boca estiver ocluída", podem formalizar essas limitações e alinhar-se a iniciativas que ponderam regiões faciais conforme sua importância diagnóstica (Heiderich *et al.*, 2023).

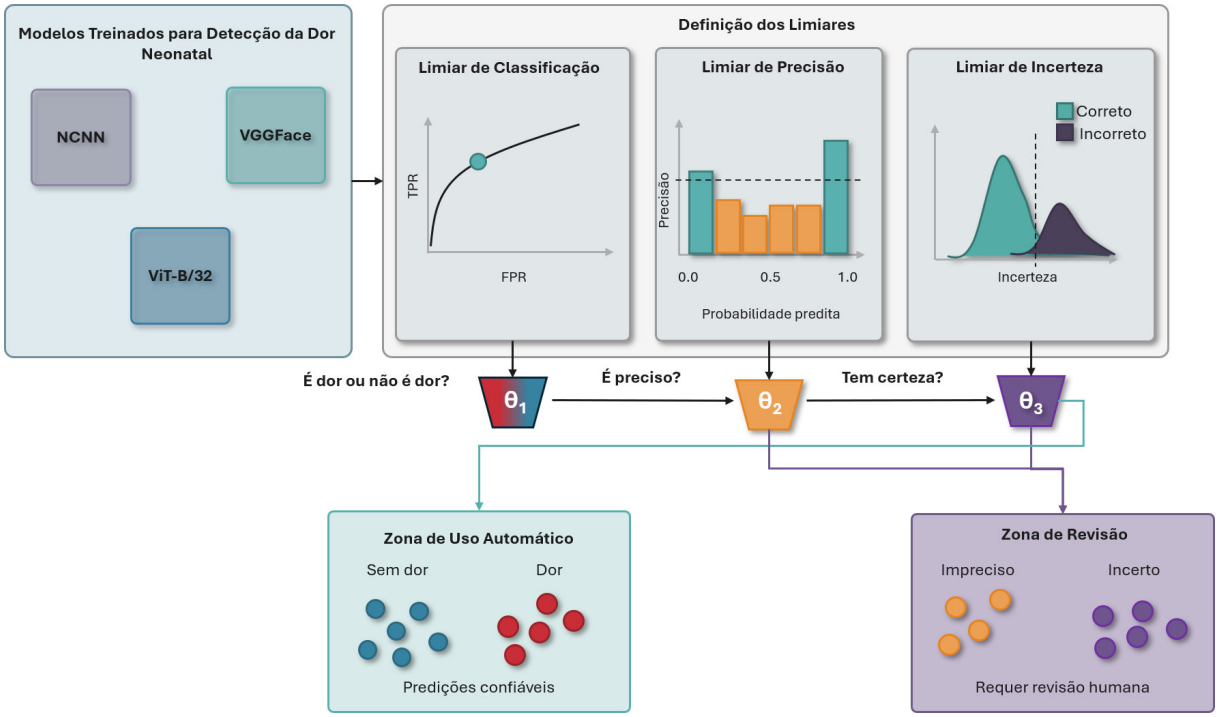
Os resultados obtidos com o ViT-B/32 reforçam a importância dessa forma de apresentação. Embora alguns explicadores tenham apresentado elevada fidelidade, regiões fora da face foram frequentemente destacadas. Em ambientes de UTIN, nos quais equipamentos de monitoramento, tubos, cobertores e variações de iluminação podem ocupar grande parte da imagem, esse comportamento deve ser interpretado como alerta clínico-operacional. O explicador pode estar correto ao mostrar que o modelo utilizou a parede, um tubo ou outro elemento do fundo,

entretanto, isso indica que a predição pode estar apoiada em confundidores, e não em sinais faciais de dor.

Desse modo, o consenso global por regiões faciais pode apoiar uma colaboração mais segura entre profissionais de saúde e modelos de IA. Quando a relevância se concentra em regiões clinicamente esperadas, a explicação pode aumentar a confiança na predição e auxiliar na priorização da avaliação clínica. Em contrapartida, quando há destaque persistente de áreas não faciais ou desalinhamento em relação às escalas de dor, o sistema deve sinalizar a necessidade de revisão humana. Essa abordagem acrescenta uma camada adicional de segurança à prática clínica, ao permitir que os modelos sejam auditados por meio de suas explicações e ao favorecer sua aderência aos requisitos regulatórios aplicáveis a sistemas de IA. A explicação, entretanto, pode induzir confiança excessiva quando apresentada de forma persuasiva, contribuindo para *automation bias*. Em trabalhos futuros, será necessário avaliar se profissionais compreendem essas limitações e se conseguem utilizar as explicações para questionar, e não apenas confirmar, a saída do modelo.

7 LIMIARES OPERACIONAIS

Figura 35 – Visão geral da metodologia do Capítulo 7.



Fonte: Autor.

7.1 INTRODUÇÃO

Os capítulos anteriores estabeleceram quatro componentes necessários à construção de um sistema de monitoramento contínuo e confiável: desempenho discriminativo, quantificação da incerteza, calibração probabilística e explicabilidade. Os resultados mostram que um bom classificador não é suficiente, por si só, para o uso clínico, mas que métodos adicionais fornecem informações convertidas neste capítulo em regras explícitas de decisão. Este capítulo desenvolve o pilar de segurança operacional ao transformar medidas de desempenho, incerteza e calibração em limiares formais, respondendo à quarta pergunta específica de pesquisa mediante a definição das condições de aceitação, rejeição e revisão das saídas.

No suporte à decisão em saúde, a ausência de limiares operacionais bem definidos aumenta o risco de uso inadequado do sistema, sobretudo diante de incerteza elevada, probabilidades descalibradas ou desbalanceamento entre classes (Guo *et al.*, 2017). Este capítulo define e analisa limiares baseados em três dimensões (Figura 35): (I) um limiar de classificação θ_1, que ajusta o ponto de corte entre as classes *com dor* e *sem dor*; (II) um limiar de precisão θ_2, que restringe as predições a faixas probabilísticas com maior confiabilidade empírica; e (III)

um limiar de incerteza θ_3, que rejeita predições cuja variabilidade estimada comprometa sua confiabilidade.

A definição desses limiares é particularmente relevante no monitoramento da dor neonatal, em que decisões inadequadas podem resultar tanto em subtratamento quanto em tratamento excessivo, com potenciais consequências imediatas e de longo prazo para o RN. Os limiares propostos não visam substituir o julgamento clínico, mas oferecem um mecanismo sistemático para distinguir predições que atendem aos critérios de aceitação daquelas que devem ser obrigatoriamente revisadas por profissionais de saúde.

7.2 MÉTODOS

7.2.1 Limiar de classificação θ_1

Na validação de modelos binários, o limiar de 50% de probabilidade predita é frequentemente adotado para classificar as respostas como positivas (*com dor*) ou negativas (*sem dor*). Esse valor, contudo, é arbitrário, e a otimização do equilíbrio entre sensibilidade e especificidade requer seu ajuste. Uma estratégia consolidada para determinar o limiar ótimo consiste na análise da curva *Receiver Operating Characteristic* (ROC), que descreve a relação entre a Taxa de Verdadeiros Positivos (TVP) e a Taxa de Falsos Positivos (TFP) em função de diferentes limiares de decisão (c).

O índice de Youden (Youden, 1950) permite identificar o ponto da curva ROC que maximiza a diferença entre TVP e TFP, fornecendo um critério objetivo para seleção do limiar. Assim, o limiar θ_1 é definido como:

$$\theta_1 = \underset{c}{\operatorname{argmax}}(TVP[c] - TFP[c]). \tag{7.1}$$

A classificação final do modelo é, então, estabelecida por:

$$\text{Classificação} = \begin{cases} Com\ dor & \text{se } p \geq \theta_1 \\ Sem\ dor & \text{se } p < \theta_1 \end{cases} \tag{7.2}$$

em que p representa a probabilidade predita pelo modelo.

7.2.2 Limiar de precisão θ_2

Para cada intervalo de probabilidade predita, a proporção de predições corretas é estimada por meio do Valor Preditivo Positivo (VPP) e do Valor Preditivo Negativo (VPN). O VPP é calculado como $\frac{VP}{VP+FP}$, e o VPN, como $\frac{VN}{VN+FN}$. Por exemplo, um VPP de 30% no intervalo [50% a 60%[indica que, nessa faixa, apenas 30% das predições *com dor* estão corretas. Essas métricas refletem a precisão empírica do modelo em cada intervalo de probabilidade e estão diretamente relacionadas à calibração.

Seja $\mathcal{B} = \{B_1, B_2, \ldots, B_K\}$ o conjunto de intervalos que particionam o domínio das probabilidades preditas. Para cada intervalo B_k, considera-se o valor preditivo correspondente à classe atribuída pelo modelo nessa faixa, definido por:

$$\text{Valor Preditivo}(B_k) = \begin{cases} \text{VPN}(B_k) & \text{se } B_k \text{ estiver associado à classe } \textit{sem dor}, \\ \text{VPP}(B_k) & \text{se } B_k \text{ estiver associado à classe } \textit{com dor}. \end{cases} \tag{7.3}$$

A definição de um nível mínimo de precisão permite restringir o uso clínico a regiões da distribuição probabilística nas quais o modelo apresenta desempenho consistente. Dada a criticidade das decisões em dor neonatal, adotou-se neste estudo $\tau = 0{,}90$. Diferentemente de um limiar escalar, θ_2 é definido como o conjunto dos intervalos cujo valor preditivo é igual ou superior a esse nível:

$$\theta_2 = \{B_k \in \mathcal{B} \mid \text{Valor Preditivo}(B_k) \geq \tau\}. \tag{7.4}$$

Assim, a classificação quanto à precisão é considerada precisa quando p pertence a pelo menos um dos conjuntos B_k que compõem θ_2. Caso contrário, ela é considerada imprecisa:

$$\text{Classificação quanto à precisão}(p) = \begin{cases} \text{Precisa} & \text{se } p \in \bigcup_{B_k \in \theta_2} B_k, \\ \text{Imprecisa} & \text{caso contrário.} \end{cases} \tag{7.5}$$

7.2.3 Limiar de incerteza θ_3

Além da precisão baseada em calibração, a incerteza associada às predições, denotada por σ, constitui um critério complementar para a avaliação da confiabilidade. Seguindo o método proposto por Dolezal *et al.* (2022), empregou-se o índice de Youden para determinar um limiar θ_3 que maximize a separação entre predições corretas e incorretas com base nos valores de incerteza estimados.

Dessa forma, estabelece-se um limiar de incerteza θ_3 definido por:

$$\text{Classificação quanto à incerteza} = \begin{cases} Certa & \text{se } \sigma \leq \theta_3 \\ Incerta & \text{se } \sigma > \theta_3 \end{cases} \tag{7.6}$$

7.3 EXPERIMENTO

O experimento definiu e avaliou os limiares operacionais dos três modelos estudados (NCNN, VGGFace e ViT-B/32) nas dimensões de classificação, precisão e incerteza. O procedimento experimental considerou os mesmos 10 *folds* dos capítulos anteriores. Em cada *fold*, os limiares foram estimados apenas a partir do subconjunto de treinamento.

Após a definição dos limiares em cada *fold*, estes foram aplicados ao respectivo conjunto de validação, de forma isolada e cumulativa, conforme o cenário analisado: apenas θ_1, apenas

θ_2, apenas θ_3 e a combinação completa θ_{1+2+3}. Para cada cenário, registraram-se métricas de desempenho discriminativo e métricas de calibração, além de métricas de cobertura, incluindo a porcentagem de dados removidos, a proporção de erros eliminados e a proporção de acertos eliminados.

Os valores finais dos limiares foram apresentados como média ± desvio-padrão, refletindo a variabilidade entre os *folds*. No caso do limiar de precisão θ_2, por envolver intervalos de probabilidade, os resultados foram reportados como as faixas de probabilidade utilizadas por cada modelo para atingir o nível mínimo de precisão estabelecido.

Também foi realizada uma análise de sensibilidade para θ_2 e θ_3, com a variação dos limiares entre os valores mínimo e máximo observados nos *folds*. Em cada ponto, comparou-se o percentual de dados retidos com um subconjunto de mesmo tamanho selecionado aleatoriamente, e calculou-se o F1-Score para ambos. Essa análise avaliou se os limiares identificavam, de forma sistemática, predições mais confiáveis do que a seleção aleatória e, portanto, se discriminavam estruturas de erro do modelo.

Em cada cenário, as comparações entre NCNN, VGGFace e ViT-B/32 foram realizadas com o teste de Friedman (Friedman, 1937). O efeito dos limiares dentro de uma mesma arquitetura foi avaliado de forma pareada com o teste de Wilcoxon (Wilcoxon, 1945), comparando-se as métricas antes e depois de sua aplicação. Adotou-se um nível de significância de 5% ($p < 0{,}05$).

7.4 RESULTADOS

A Tabela 17 apresenta os valores dos limiares θ_1, θ_2 e θ_3 obtidos para os modelos NCNN, VGGFace e ViT-B/32.

Tabela 17 – Valores dos limiares θ_1, θ_2 e θ_3 para os modelos (intervalo para θ_2 e média ± DP para θ_1 e θ_3).

Limiar	**NCNN**	**VGGFace**	**ViT-B/32**
θ_1	0,4551 ± 0,1175	0,5013 ± 0,0906	0,4743 ± 0,0210
θ_2	[0%, 20%] e [80%, 100%]	[0%, 20%] e [80%, 100%]	[0%, 20%] e [80%, 100%]
θ_3	0,1186 ± 0,0200	0,0621 ± 0,0065	0,0130 ± 0,0033

Fonte: Autor.

Com o limiar de classificação θ_1 (Tabela 18), o teste pareado de Wilcoxon não indicou diferenças estatisticamente significativas nas métricas de classificação nem nas métricas de calibração para NCNN, VGGFace ou ViT-B/32. Em termos práticos, o ajuste do ponto de corte preservou o comportamento de cada modelo, sem produzir ganhos relevantes.

Segundo o teste pareado de Wilcoxon, o limiar de precisão θ_2 (Tabela 19) produziu ganhos expressivos e estatisticamente significativos nas métricas de classificação e reduções significativas em NLL e Brier, enquanto ECE e MCE não apresentaram alterações significativas. O aspecto central desses resultados foi a quantidade de dados removidos por esse limiar. Em todos

Tabela 18 – Métricas de classificação e calibração dos modelos após aplicação do limiar θ_1 (média ± DP, em %).

Métrica	**NCNN**	**VGGFace**	**ViT-B/32**	**valor-p**
Acurácia	79,0 ± 5,6 ↓	**84,3 ± 4,9** ↑	82,0 ± 6,4 ↓	0,0347
F1-Score	75,8 ± 7,6 ↑	**81,9 ± 6,3** ↑	79,2 ± 7,1 ↓	0,1225
Sensibilidade	75,8 ± 16,7 ↑	**80,7 ± 10,5** ↑	75,6 ± 11,3 ↑	0,4914
Especificidade	81,5 ± 13,1 ↓	**86,9 ± 8,2** ↓	86,7 ± 8,7 ↓	0,7103
Precisão	80,1 ± 12,0 ↓	**84,4 ± 8,5** ↑	84,2 ± 6,7 ↓	0,4066
AUC	88,3 ± 6,4 =	**91,9 ± 3,4** =	89,8 ± 5,3 =	0,2725
ECE	12,9 ± 4,7 =	**10,3 ± 4,2** =	10,6 ± 3,7 =	0,1496
MCE	**29,5 ± 12,6** =	33,5 ± 17,0 =	32,6 ± 14,7 =	0,4966
NLL	42,4 ± 10,1 =	**37,8 ± 9,3** =	41,7 ± 15,2 =	0,4966
Brier	13,7 ± 4,1 =	**11,7 ± 3,2** =	12,7 ± 4,1 =	0,3012

Fonte: Autor.

Legenda: Os símbolos indicam melhora (↑), piora (↓), melhora estatisticamente significativa (↑*), piora estatisticamente significativa (↓*) e resultado inalterado (=). O asterisco indica $p < 0{,}05$ no teste de Wilcoxon pareado com a etapa anterior. A última coluna apresenta o teste de Friedman entre os modelos.

os modelos, pelo menos metade dos dados incorretos foi eliminada, com destaque para a NCNN, que removeu $81{,}2\%$ dos erros, ao custo de $47{,}9\%$ dos acertos. A VGGFace apresentou um perfil mais equilibrado, com $49{,}4\%$ dos erros eliminados e apenas $21{,}6\%$ dos acertos removidos. Já o ViT-B/32 eliminou $59{,}4\%$ dos erros, com remoção de $28{,}5\%$ dos acertos.

O limiar de incerteza θ_3 (Tabela 20) também elevou significativamente as métricas discriminativas, com exceção da sensibilidade do ViT-B/32, e reduziu significativamente NLL e Brier nos três modelos, enquanto ECE e MCE não apresentaram alterações significativas. No entanto, a remoção de dados foi mais acentuada do que com θ_2, sobretudo no ViT-B/32, que eliminou $74{,}8\%$ dos erros, mas também $54{,}2\%$ dos acertos. De modo geral, o limiar de incerteza mostrou-se mais rigoroso que o limiar de precisão, removendo uma parcela maior dos dados. O único modelo que reteve, em média, mais de 50% dos dados foi a VGGFace, que eliminou $27{,}9\%$ dos dados, com $68{,}6\%$ dos erros eliminados e $34{,}2\%$ dos acertos removidos.

A aplicação conjunta θ_{1+2+3} (Tabela 21) produziu as maiores médias de desempenho discriminativo, com aumentos estatisticamente significativos em todas as métricas, exceto na sensibilidade do ViT-B/32. Também foram observadas reduções significativas de NLL e Brier nos três modelos e de ECE no ViT-B/32, as demais alterações em ECE e MCE não foram significativas. Como esperado, a combinação dos limiares removeu uma parcela substancial dos dados. A NCNN eliminou $89{,}7\%$ dos erros, ao custo de $59{,}0\%$ dos acertos. O ViT-B/32 eliminou $82{,}8\%$ dos erros e $56{,}4\%$ dos acertos. A VGGFace apresentou um perfil intermediário, com $67{,}6\%$ dos erros eliminados e $34{,}3\%$ dos acertos removidos.

Para facilitar a compreensão dos dados removidos, o Apêndice D apresenta as imagens rejeitadas por cada limiar nos três modelos. A análise qualitativa mostrou que os casos rejeitados

Tabela 19 – Métricas de classificação e calibração dos modelos após aplicação do limiar θ_2 (média ± DP, em %).

Métrica	**NCNN**	**VGGFace**	**ViT-B/32**	**valor-p**
Acurácia	**93,5 ± 6,2** $\uparrow^*$	89,8 ± 4,7 $\uparrow^*$	90,0 ± 6,1 $\uparrow^*$	0,4966
F1-Score	**92,9 ± 7,1** $\uparrow^*$	88,3 ± 5,9 $\uparrow^*$	88,4 ± 6,5 $\uparrow^*$	0,2725
Sensibilidade	**91,0 ± 11,0** $\uparrow^*$	85,1 ± 9,3 $\uparrow^*$	83,4 ± 10,8 $\uparrow^*$	0,5669
Especificidade	**95,9 ± 4,6** $\uparrow^*$	92,9 ± 8,9 $\uparrow^*$	94,7 ± 7,5 $\uparrow^*$	0,8752
Precisão	**95,6 ± 5,6** $\uparrow^*$	92,8 ± 8,2 $\uparrow^*$	95,2 ± 5,4 $\uparrow^*$	0,8752
AUC	**96,3 ± 4,6** $\uparrow^*$	94,6 ± 3,0 $\uparrow^*$	93,2 ± 4,8 $\uparrow^*$	0,3012
ECE	11,8 ± 1,9 $\uparrow$	**9,7 ± 3,9** $\uparrow$	9,8 ± 4,6 $\uparrow$	0,1496
MCE	**30,0 ± 14,1** $\downarrow$	33,6 ± 14,3 $\downarrow$	30,8 ± 11,8 $\uparrow$	0,4966
NLL	**23,1 ± 11,6** $\uparrow^*$	30,7 ± 12,2 $\uparrow^*$	32,4 ± 19,2 $\uparrow^*$	0,3012
Brier	**6,0 ± 4,4** $\uparrow^*$	8,9 ± 3,9 $\uparrow^*$	8,8 ± 5,0 $\uparrow^*$	0,3012
% Dados corretos removidos	47,9 ± 15,3	**21,6 ± 5,2**	28,5 ± 7,9	
% Dados incorretos removidos	**81,2 ± 16,5**	49,4 ± 17,3	59,4 ± 17,4	
% Total removido	39,7 ± 15,3	**16,3 ± 4,3**	22,0 ± 9,4	

Fonte: Autor.

Legenda: Os símbolos indicam melhora ($\uparrow$), piora ($\downarrow$), melhora estatisticamente significativa ($\uparrow^*$), piora estatisticamente significativa ($\downarrow^*$) e resultado inalterado ($=$). Para ECE, MCE, NLL e Brier, reduções representam melhora. O asterisco indica $p < 0{,}05$ no teste de Wilcoxon pareado com a etapa anterior. A última coluna apresenta o teste de Friedman entre os modelos.

Tabela 20 – Métricas de classificação e calibração dos modelos após aplicação do limiar θ_3 (média ± DP, em %).

Métrica	**NCNN**	**VGGFace**	**ViT-B/32**	**valor-p**
Acurácia	91,8 ± 4,6 $\uparrow^*$	92,1 ± 4,5 $\uparrow^*$	**92,2 ± 4,2** $\uparrow^*$	0,8948
F1-Score	**90,9 ± 4,5** $\uparrow^*$	**90,9 ± 6,0** $\uparrow^*$	88,2 ± 6,9 $\uparrow^*$	0,8187
Sensibilidade	87,1 ± 9,7 $\uparrow^*$	**88,3 ± 9,5** $\uparrow^*$	80,3 ± 11,0 $\uparrow$	0,4018
Especificidade	97,2 ± 4,1 $\uparrow^*$	93,9 ± 10,3 $\uparrow^*$	**99,5 ± 1,4** $\uparrow^*$	0,7165
Precisão	96,2 ± 5,8 $\uparrow^*$	94,7 ± 7,8 $\uparrow^*$	**98,8 ± 3,5** $\uparrow^*$	0,6376
AUC	96,1 ± 3,8 $\uparrow^*$	96,0 ± 2,5 $\uparrow^*$	**96,4 ± 2,1** $\uparrow^*$	0,8948
ECE	13,0 ± 1,7 $\downarrow$	8,3 ± 4,0 $\uparrow$	**8,2 ± 2,8** $\uparrow$	0,0164
MCE	**35,4 ± 7,3** $\downarrow$	35,9 ± 22,3 $\downarrow$	43,6 ± 23,3 $\downarrow$	0,7165
NLL	26,9 ± 10,0 $\uparrow^*$	26,1 ± 12,7 $\uparrow^*$	**22,3 ± 8,5** $\uparrow^*$	1,0000
Brier	7,5 ± 3,5 $\uparrow^*$	7,2 ± 3,9 $\uparrow^*$	**6,2 ± 2,6** $\uparrow^*$	0,8948
% Dados corretos removidos	52,1 ± 20,5	**34,2 ± 6,2**	54,2 ± 21,1	
% Dados incorretos removidos	**79,5 ± 13,7**	68,6 ± 11,2	74,8 ± 25,7	
% Total removido	45,4 ± 23,1	**27,9 ± 6,5**	49,3 ± 22,7	

Fonte: Autor.

Legenda: Os símbolos indicam melhora ($\uparrow$), piora ($\downarrow$), melhora estatisticamente significativa ($\uparrow^*$), piora estatisticamente significativa ($\downarrow^*$) e resultado inalterado ($=$). Para ECE, MCE, NLL e Brier, reduções representam melhora. O asterisco indica $p < 0{,}05$ no teste de Wilcoxon pareado com a etapa anterior. A última coluna apresenta o teste de Friedman entre os modelos.

Tabela 21 – Métricas de classificação e calibração dos modelos após aplicação do limiar θ_{1+2+3} (média ± DP, em %).

Métrica	**NCNN**	**VGGFace**	**ViT-B/32**	**valor-p**
Acurácia	**94,7 ± 5,7** $\uparrow^*$	92,1 ± 4,5 $\uparrow^*$	94,5 ± 2,7 $\uparrow^*$	0,6412
F1-Score	**94,3 ± 5,9** $\uparrow^*$	90,9 ± 6,0 $\uparrow^*$	90,9 ± 6,8 $\uparrow^*$	0,4124
Sensibilidade	**90,5 ± 10,6** $\uparrow^*$	88,3 ± 9,5 $\uparrow^*$	84,7 ± 10,1 $\uparrow$	0,4281
Especificidade	99,2 ± 2,4 $\uparrow^*$	93,9 ± 10,3 $\uparrow^*$	**99,5 ± 1,4** $\uparrow^*$	0,2907
Precisão	**99,3 ± 2,0** $\uparrow^*$	94,7 ± 7,8 $\uparrow^*$	98,8 ± 3,5 $\uparrow^*$	0,2907
AUC	**96,8 ± 3,9** $\uparrow^*$	96,0 ± 2,5 $\uparrow^*$	96,4 ± 1,9 $\uparrow^*$	0,6412
ECE	10,3 ± 3,1 $\uparrow$	8,2 ± 4,1 $\uparrow$	**6,5 ± 2,3** $\uparrow^*$	0,2636
MCE	**30,4 ± 10,7** $\downarrow$	35,9 ± 22,3 $\downarrow$	35,6 ± 22,8 $\downarrow$	0,8948
NLL	20,9 ± 12,8 $\uparrow^*$	26,0 ± 12,7 $\uparrow^*$	**19,5 ± 8,3** $\uparrow^*$	0,7165
Brier	5,2 ± 4,5 $\uparrow^*$	7,2 ± 3,9 $\uparrow^*$	**5,1 ± 2,4** $\uparrow^*$	0,4594
% Dados corretos removidos	59,0 ± 17,6	**34,3 ± 6,1**	56,4 ± 20,0	
% Dados incorretos removidos	**89,7 ± 12,5**	67,6 ± 14,5	82,8 ± 20,4	
% Total removido	51,4 ± 20,1	**28,3 ± 6,3**	49,9 ± 22,6	

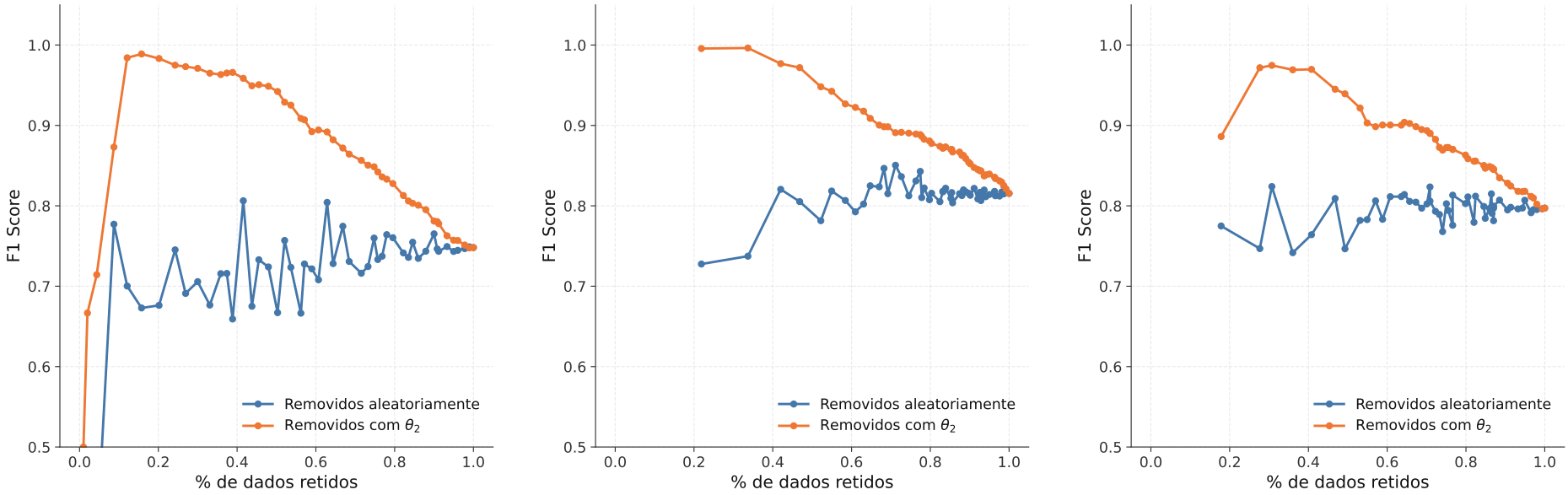
Fonte: Autor.

Legenda: Os símbolos indicam melhora ($\uparrow$), piora ($\downarrow$), melhora estatisticamente significativa ($\uparrow^*$), piora estatisticamente significativa ($\downarrow^*$) e resultado inalterado ($=$). Para ECE, MCE, NLL e Brier, reduções representam melhora. O asterisco indica $p < 0{,}05$ no teste de Wilcoxon pareado com a etapa anterior. A última coluna apresenta o teste de Friedman entre os modelos.

por θ_2 e θ_3 costumam apresentar características associadas a uma classificação mais difícil, como baixa qualidade da imagem e traços faciais atípicos em relação à classe original. Os limiares, portanto, capturaram aspectos relevantes da confiabilidade das predições, em vez de apenas descartarem amostras arbitrariamente.

Como observado nas Tabelas 19 e 20, os limiares θ_2 e θ_3 reduziram a cobertura dos modelos em proporções relevantes, mas essa redução só é útil se os casos preservados forem, de fato, mais confiáveis do que uma amostragem arbitrária do mesmo tamanho. A Figura 36 mostra que o limiar θ_2 ordenou os exemplos de forma consistente nos três modelos. Para a NCNN, os subconjuntos mais restritivos atingiram F1-Score próximo de 1,0 e a métrica diminuiu gradualmente à medida que a cobertura aumentou. O mesmo padrão foi observado na VGGFace e no ViT-B/32. A curva aleatória, por sua vez, oscilou em torno do desempenho médio do modelo, sem ganho sistemático associado à remoção de amostras.

Figura 36 – F1-Score em função da porcentagem de dados retidos por θ_2 e por seleção aleatória.



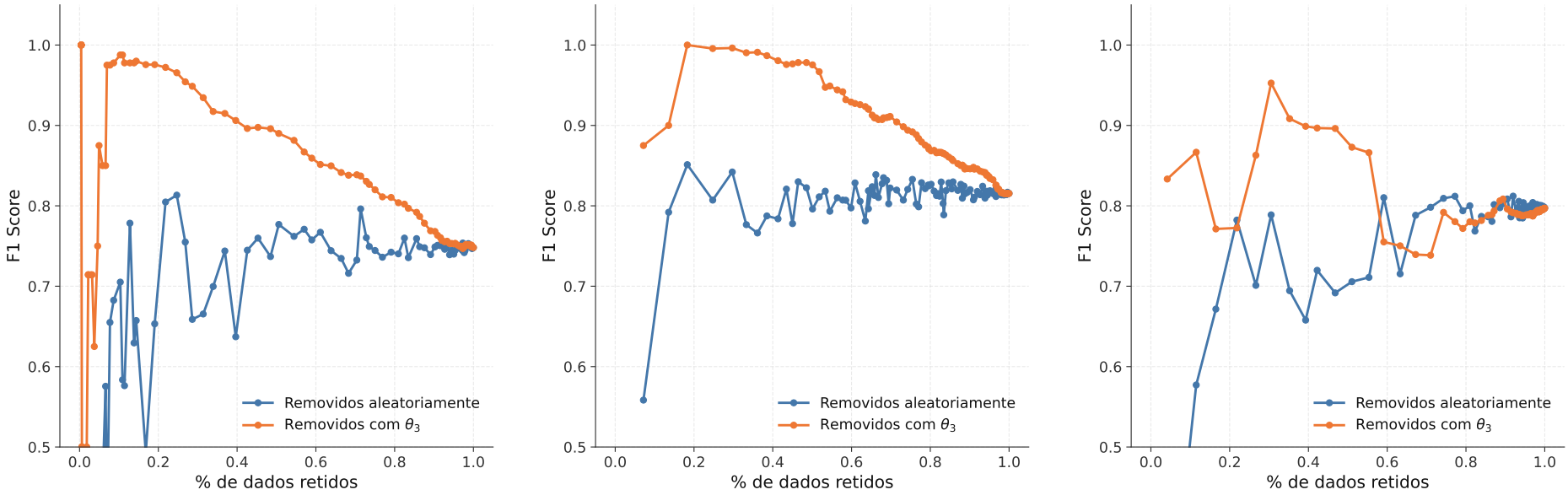
(a) NCNN. (b) VGGFace. (c) ViT-B/32.

Fonte: Autor.

Nota: O eixo y foi limitado a um F1-Score mínimo de 50% para melhor visualização. Valores abaixo de 50% indicam desempenho inferior ao esperado por uma escolha aleatória.

A análise de θ_3, apresentada na Figura 37, confirma que a incerteza contém informação sobre a qualidade das predições, mas seu comportamento depende mais da arquitetura. Na NCNN e na VGGFace, a remoção guiada por θ_3 produziu F1-Score superior ao controle aleatório em quase toda a faixa de retenção. No ViT-B/32, a separação foi menos estável. À medida que a porcentagem de dados retidos aumentou, a curva se aproximou da seleção aleatória e, em alguns trechos, sobrepôs-se a ela. Para esse modelo, a incerteza isolada foi menos eficiente do que θ_2 na ordenação dos erros.

Figura 37 – F1-Score em função da porcentagem de dados retidos por θ_3 e por seleção aleatória.



(a) NCNN. (b) VGGFace. (c) ViT-B/32.

Fonte: Autor.

Nota: O eixo y foi limitado a um F1-Score mínimo de 50% para melhor visualização. Valores abaixo de 50% indicam desempenho inferior ao esperado por uma escolha aleatória.

7.5 DISCUSSÃO

Os limiares operacionais alteraram de forma substancial o perfil de uso dos modelos. Em vez de responder a toda entrada, os modelos passaram a operar de forma seletiva, aceitando algumas predições e rejeitando outras com base em critérios explícitos. Com isso, as propriedades estudadas nos capítulos anteriores foram convertidas em regras de decisão.

Os experimentos indicaram que ajustes ao limiar de classificação (θ_1) foram insuficientes para produzir ganhos relevantes nas métricas discriminativas ou de calibração, o que corrobora a literatura segundo a qual a simples escolha do limiar pela curva ROC raramente resolve problemas de confiabilidade em sistemas probabilísticos (Guo *et al.*, 2017). Em contrapartida, os limiares de precisão (θ_2) e de incerteza (θ_3) produziram melhorias consistentes em acurácia, F1-Score e AUC, além de reduções em NLL e Brier, evidenciando que a qualidade das predições aceitas aumenta quando o sistema se torna seletivo.

O limiar de precisão (θ_2) mostrou-se mais compatível com cenários assistenciais de alta demanda, nos quais se deseja reduzir erros sem sacrificar excessivamente a cobertura. O limiar de incerteza (θ_3) correspondeu a um regime mais conservador, adequado a situações em que a prioridade é aceitar apenas predições de alta certeza, mesmo que mais casos sejam rejeitados. A VGGFace apresentou o perfil mais equilibrado, com boa retenção de acertos mesmo sob critérios rigorosos, e pode ser a candidata mais adequada ao monitoramento contínuo entre os modelos avaliados. A NCNN e o ViT-B/32, embora tenham apresentado maiores ganhos nas métricas após a seleção, exigiriam supervisão mais frequente devido à maior proporção de casos rejeitados.

A combinação dos três limiares (θ_{1+2+3}) produziu o cenário mais conservador e as melhores métricas entre as predições retidas. O custo foi uma redução substancial da cobertura, sobretudo na NCNN e no ViT-B/32. Aumentar a confiabilidade implica, portanto, reduzir a autonomia decisória do sistema, pelo menos com as bases de dados disponíveis para este estudo.

As análises de sensibilidade sustentaram a validade dos limiares, ou seja, para níveis equivalentes de remoção, os subconjuntos retidos por θ_2 e θ_3 apresentaram F1-Score superior ao dos subconjuntos selecionados aleatoriamente. Os limiares não descartaram amostras de modo arbitrário, mas exploraram a estrutura dos dados, o que justifica sua incorporação como mecanismo formal de segurança operacional.

7.5.1 Implicações para a prática clínica

A definição de limiares operacionais tem implicações diretas para a incorporação responsável de sistemas de IA na avaliação da dor neonatal. Em primeiro lugar, os resultados indicam que não é apropriado tratar todas as predições como equivalentes, pois uma fração relevante das saídas apresenta incerteza elevada e/ou baixa precisão empírica e, portanto, não deve ser utilizada sem revisão. A aplicação dos limiares permite identificar automaticamente essas situações, sinalizando ao profissional a necessidade de supervisão humana.

Entretanto, a redução da cobertura introduz consequências operacionais relevantes para

a prática clínica. Em ambientes como a UTIN, nos quais decisões precisam ser contínuas e oportunas, um sistema que rejeita uma parcela substancial das amostras pode gerar lacunas de monitoramento justamente em momentos críticos. Nesses casos, a ausência de uma predição utilizável pode ser tão problemática quanto uma predição incorreta, especialmente se a indisponibilidade ocorrer de forma recorrente em pacientes mais instáveis, durante procedimentos invasivos ou em condições de aquisição adversas.

Do ponto de vista do profissional de saúde, a baixa cobertura também afeta a usabilidade e a confiança no sistema. Um modelo que se abstém com frequência, como a NCNN e o ViT-B/32, transfere continuamente a decisão para a equipe clínica, reduz o potencial de ganho operacional da ferramenta e aumenta a carga cognitiva da supervisão. A VGGFace, que apresentou o melhor equilíbrio entre retenção de acertos e remoção de erros, pode ser mais adequada a cenários de monitoramento contínuo.

É indispensável, porém, definir protocolos para os casos rejeitados, incluindo o encaminhamento para avaliação clínica, o registro das causas prováveis da rejeição e uma eventual reavaliação após nova aquisição de imagem. O perfil das amostras rejeitadas também deve ser monitorado de forma sistemática. Perfis clínicos, faixas etárias, condições de iluminação ou padrões faciais excluídos com maior frequência podem introduzir vieses operacionais e limitar a generalização clínica do sistema. Após a análise, esses casos podem ser utilizados no retreinamento do modelo para reduzir a taxa de rejeição e aumentar a cobertura sem sacrificar a confiabilidade.

8 MODELOS FINAIS

8.1 INTRODUÇÃO

Este capítulo consolida, para cada arquitetura, as estratégias selecionadas nas etapas de treinamento, quantificação da incerteza, calibração, XAI e definição dos limiares operacionais. A integração produz os modelos finais utilizados para a construção do sinal da dor e sintetiza seus resultados antes da análise temporal apresentada no Capítulo 9.

8.2 EXPERIMENTO

Conforme descrito no Capítulo 3, os dados foram divididos em 80% para treinamento e 20% para teste. Na etapa anterior, o conjunto de treinamento foi subdividido em *folds* para a validação cruzada e a definição dos hiperparâmetros. Nesta etapa, os modelos foram treinados com todo o conjunto de treinamento, utilizando os hiperparâmetros apresentados na Tabela 5, e avaliados no conjunto de teste para a obtenção dos resultados finais.

A Tabela 22 sintetiza as estratégias selecionadas para cada modelo final.

Tabela 22 – Estratégias metodológicas aplicadas a cada modelo ao longo da metodologia proposta.

Estratégia	**NCNN**	**VGGFace**	**ViT-B/32**
QI	MCDP (N_{MC} = 50 e p_{drop} = 30%)	MCDP (N_{MC} = 50 e p_{drop} = 50%)	10 Ensembles
Calibração	—	—	Regressão Isotônica
θ_1	0,4551	0,5013	0,4743
θ_2	[0%, 20%] e [80%, 100%]	[0%, 20%] e [80%, 100%]	[0%, 20%] e [80%, 100%]
θ_3	0,1186	0,0621	0,0130

Fonte: Autor.

As etapas da metodologia foram aplicadas nesta ordem: quantificação da incerteza (QI), calibração, XAI e aplicação dos limiares operacionais. Os resultados foram registrados para cada modelo. As métricas de desempenho discriminativo e de qualidade probabilística foram calculadas em cada etapa, permitindo analisar o impacto de cada intervenção metodológica. Além disso, foi medida, em milissegundos, a latência entre o início do processamento, com a detecção facial, e a inferência final do modelo após todas as etapas da metodologia. Para tanto, foi utilizado o ambiente computacional do computador pessoal descrito no Capítulo 3, na Tabela 3.

8.3 RESULTADOS

As Tabelas 23, 24 e 25 sintetizam a evolução dos modelos ao longo das etapas metodológicas.

A quantificação da incerteza e a calibração produziram poucas alterações no desempenho discriminativo, e os ganhos ocorreram principalmente após a aplicação dos limiares operacio-

Tabela 23 – Evolução das métricas da NCNN ao longo das etapas metodológicas (valores em %).

Métrica	**Base**	**QI**	**Calib.**	θ_1	θ_2	θ_3	θ_{1+2+3}
Acurácia	84,9	84,9	84,9	84,9	**89,3**	85,7	88,5
F1-Score	81,8	81,0	81,0	81,0	**88,0**	81,5	85,7
Sensibilidade	75,0	70,8	70,8	70,8	**78,6**	73,3	75,0
Especificidade	93,1	96,6	96,6	96,6	**100,0**	95,0	**100,0**
Precisão	90,0	94,4	94,4	94,4	**100,0**	91,7	**100,0**
AUC	84,8	85,3	85,3	85,3	**89,3**	86,0	87,5
ECE	**7,1**	7,9	7,9	7,9	13,3	15,6	12,4
MCE	19,2	**14,4**	**14,4**	**14,4**	25,8	58,2	39,9
NLL	43,0	43,1	43,1	43,1	**33,4**	39,3	34,7
Brier Score	13,5	13,4	13,4	13,4	**9,7**	12,0	10,2
% Dados corretos removidos	–	–	–	–	44,4	**33,3**	48,9
% Dados incorretos removidos	–	–	–	–	**62,5**	37,5	**62,5**
% Total removido	–	–	–	–	47,2	**34,0**	50,9

Fonte: Autor.

Tabela 24 – Evolução das métricas da VGGFace ao longo das etapas metodológicas (valores em %).

Métrica	**Base**	**QI**	**Calib.**	θ_1	θ_2	θ_3	θ_{1+2+3}
Acurácia	81,1	81,1	81,1	81,1	88,1	**93,8**	**93,8**
F1-Score	77,3	77,3	77,3	77,3	85,7	**93,3**	**93,3**
Sensibilidade	70,8	70,8	70,8	70,8	79,0	**87,5**	**87,5**
Especificidade	89,7	89,7	89,7	89,7	95,7	**100,0**	**100,0**
Precisão	85,0	85,0	85,0	85,0	93,8	**100,0**	**100,0**
AUC	87,8	88,4	88,4	88,4	91,8	**94,5**	**94,5**
ECE	10,7	10,0	10,0	10,0	8,6	**7,5**	**7,5**
MCE	32,4	30,5	30,5	30,5	**18,5**	30,5	30,5
NLL	44,3	42,1	42,1	42,1	33,5	**22,6**	**22,6**
Brier Score	14,0	13,6	13,6	13,6	9,9	**5,8**	**5,8**
% Dados corretos removidos	–	–	–	–	**14,0**	30,2	30,2
% Dados incorretos removidos	–	–	–	–	50,0	**80,0**	**80,0**
% Total removido	–	–	–	–	**20,8**	39,6	39,6

Fonte: Autor.

Tabela 25 – Evolução das métricas do ViT-B/32 ao longo das etapas metodológicas (valores em %).

Métrica	**Base**	**QI**	**Calib.**	θ_1	θ_2	θ_3	θ_{1+2+3}
Acurácia	86,8	84,9	84,9	84,9	92,5	**100,0**	**100,0**
F1-Score	84,4	82,6	82,6	82,6	91,4	**100,0**	**100,0**
Sensibilidade	79,2	79,2	79,2	79,2	84,2	**100,0**	**100,0**
Especificidade	93,1	89,7	89,7	89,7	**100,0**	**100,0**	**100,0**
Precisão	90,5	86,4	86,4	86,4	**100,0**	**100,0**	**100,0**
AUC	95,8	94,7	94,8	94,8	99,3	**100,0**	**100,0**
ECE	10,2	5,6	7,2	7,2	7,0	**2,4**	**2,4**
MCE	23,3	14,2	18,9	18,9	80,0	**6,9**	**6,9**
NLL	29,6	29,6	27,5	27,5	16,9	**2,4**	**2,4**
Brier Score	9,0	9,1	9,1	9,1	5,3	**0,1**	**0,1**
% Dados corretos removidos	–	–	–	–	**17,8**	44,4	44,4
% Dados incorretos removidos	–	–	–	–	62,5	**100,0**	**100,0**
% Total removido	–	–	–	–	**24,5**	52,8	52,8

Fonte: Autor.

nais. Entre os modelos finais, a VGGFace apresentou o melhor equilíbrio entre desempenho e cobertura, com acurácia de 93,8%, AUC de 94,5% e remoção de 80,0% dos erros, ao custo de rejeitar 39,6% das amostras.

O ViT-B/32 alcançou 100,0% nas métricas discriminativas e removeu todos os erros no subconjunto aceito, mas apresentou a maior taxa de rejeição (52,8%). A NCNN também operou de forma conservadora, com especificidade e precisão de 100,0%, porém rejeitou 50,9% das amostras e removeu 62,5% dos erros. Assim, os resultados evidenciam o compromisso entre desempenho e cobertura: o ViT-B/32 foi o mais seletivo, enquanto a VGGFace preservou a maior cobertura entre os modelos de melhor desempenho.

Em relação à latência, a Tabela 26 apresenta os valores médios e os respectivos desvios-padrão para cada etapa da metodologia. A etapa de XAI foi a mais custosa em termos de processamento, correspondendo a 94,5%, 97,8% e 99,4% da latência total dos modelos NCNN, VGGFace e ViT-B/32, respectivamente. Esse custo elevado decorre da necessidade de gerar dez mapas de explicabilidade para uma única entrada e, posteriormente, combiná-los.

Nos experimentos sem XAI, a inferência com quantificação da incerteza representou 68,1% do tempo total para a NCNN com MCDP, 85,9% para a VGGFace com MCDP e 69,3% para o ViT-B/32 com *Deep Ensembles*. Quando a XAI foi incluída, essas proporções foram reduzidas para 3,8%, 1,9% e 0,5%, respectivamente, devido ao predomínio do custo computacional da geração das explicações.

Considerando todas as etapas da metodologia, a NCNN apresentou a menor latência (845,22 $\pm$ 48,61 ms), seguida pela VGGFace (4766,53 $\pm$ 66,36 ms) e pelo ViT-B/32 (8024,50 $\pm$

346,37 ms). Portanto, a adição dos métodos de quantificação da incerteza e a inclusão da XAI resultaram em um aumento substancial do tempo total de processamento.

Tabela 26 – Latência média ± desvio-padrão para cada etapa da metodologia. Valores em milissegundos.

Modelo	Detecção Facial	Pré-proc.	CPU→GPU	Inferência	XAI	Limiares	Latência Final
NCNN	10,20 ± 1,25	4,18 ± 2,35	**0,11 ± 0,03**	**1,03 ± 0,27**	–	**0,01 ± 0,00**	**15,53 ± 3,01**
NCNN + MCDP	10,27 ± 1,33	4,45 ± 2,40	**0,11 ± 0,07**	32,00 ± 4,82	–	0,02 ± 0,00	46,85 ± 6,24
NCNN + MCDP + XAI	10,22 ± 1,38	4,09 ± 2,73	**0,11 ± 0,03**	32,35 ± 3,24	**798,44 ± 45,62**	**0,01 ± 0,00**	845,22 ± 48,61
VGGFace	**10,01 ± 1,07**	4,92 ± 2,51	0,16 ± 0,03	2,15 ± 0,17	–	**0,01 ± 0,00**	17,25 ± 2,95
VGGFace + MCDP	10,18 ± 1,07	5,16 ± 2,54	0,17 ± 0,03	95,23 ± 3,21	–	0,02 ± 0,01	110,76 ± 5,45
VGGFace + MCDP + XAI	10,62 ± 0,93	3,63 ± 2,73	0,15 ± 0,03	92,69 ± 2,69	4659,43 ± 62,41	**0,01 ± 0,00**	4766,53 ± 66,36
ViT-B/32	10,11 ± 0,99	5,15 ± 2,56	0,15 ± 0,02	4,16 ± 1,00	–	**0,01 ± 0,00**	19,58 ± 2,92
ViT-B/32 + Ensembles	10,71 ± 1,60	6,02 ± 2,63	0,18 ± 0,12	38,58 ± 5,86	–	0,02 ± 0,00	55,51 ± 7,00
ViT-B/32 + Ensembles + XAI	10,26 ± 1,35	**2,67 ± 2,39**	0,14 ± 0,02	38,75 ± 4,55	7972,67 ± 342,11	**0,01 ± 0,00**	8024,50 ± 346,37

Fonte: Autor.

8.4 DISCUSSÃO

A quantificação da incerteza e a calibração pouco modificaram as métricas discriminativas, mas permitiram caracterizar melhor a confiança nas predições e estabelecer critérios de aceitação e rejeição. Os limiares θ_2 e θ_3, em particular, restringiram a decisão aos casos mais confiáveis e eliminaram parte relevante dos erros.

Por essa razão, desempenho e cobertura devem ser analisados conjuntamente. No subconjunto aceito, as métricas elevadas vieram acompanhadas de uma taxa expressiva de rejeição, sobretudo no ViT-B/32 e na NCNN. Na prática clínica, uma amostra rejeitada deve ser encaminhada para avaliação humana ou, quando possível, para uma nova aquisição, em vez de ser tratada como um caso sem decisão. A cobertura também permite estimar a demanda de supervisão, ou seja, quanto mais conservador o critério de aceitação, menor a exposição a decisões semi-automáticas pouco confiáveis, mas maior o número de casos transferidos para a equipe assistencial.

No ambiente computacional utilizado, após a inclusão da quantificação da incerteza e sem XAI, a latência final foi de 46,85 ms para a NCNN, 110,76 ms para a VGGFace e 55,51 ms para o ViT-B/32. Com a geração das explicações, esses valores aumentaram para 845,22 ms, 4766,53 ms e 8024,50 ms, respectivamente, o que corresponde a fatores de aproximadamente 18, 43 e 145 vezes. Diretrizes de interação humano-computador revisadas por Attig *et al.* (2017) associam respostas de até aproximadamente 100 ms à aparência de instantaneidade e situam respostas entre 500 ms e 1 s na faixa denominada imediata, embora já distinguível de uma resposta instantânea. Os autores ressaltam que esses valores não constituem limites universais, logo a percepção e a aceitação dependem do tipo e da complexidade da tarefa, da experiência prévia e das expectativas do usuário.

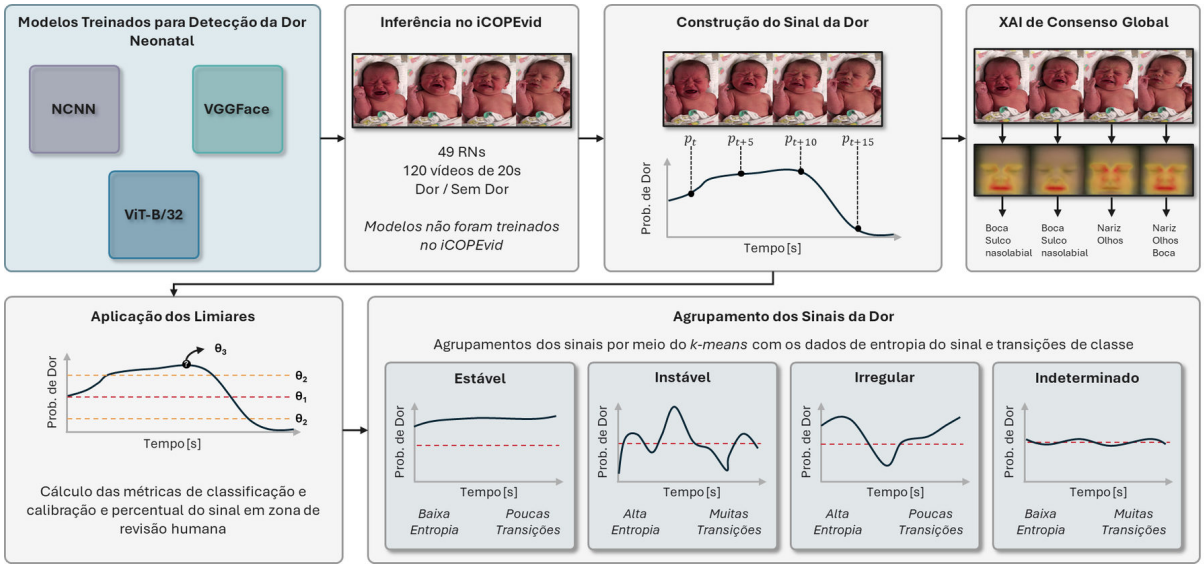
Assim, os 845,22 ms da NCNN com XAI permanecem próximos de uma resposta imediata, enquanto os 4766,53 ms da VGGFace e os 8024,50 ms do ViT-B/32 configuram espera

explícita. Estudos experimentais mostram que atrasos da ordem de segundos podem prejudicar a interação e que a apresentação de feedback aumenta a tolerância à espera (Nah, 2004; Szameitat *et al.*, 2009). Uma alternativa seria, portanto, retornar de forma síncrona a classificação e a incerteza, cuja latência máxima foi de 110,76 ms, e gerar o XAI em segundo plano ou apenas em casos específicos de auditoria.

A VGGFace apresentou o melhor equilíbrio entre desempenho e cobertura e, por isso, preservou maior proporção de saídas potencialmente utilizáveis. O ViT-B/32 obteve o melhor desempenho no subconjunto aceito, mas rejeitou mais amostras e operou de forma mais seletiva. Já a NCNN manteve a menor latência em todas as configurações, o que a favorece quando o tempo de resposta é prioritário. A escolha depende, assim, do equilíbrio pretendido entre desempenho, cobertura, confiabilidade e custo computacional, além da capacidade de supervisão e dos requisitos temporais do ambiente clínico.

9 DESENVOLVIMENTO E AVALIAÇÃO DO SINAL DA DOR

Figura 38 – Visão geral da metodologia do Capítulo 9.



Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

9.1 INTRODUÇÃO

O Capítulo 8 consolidou os modelos finais após integrar treinamento, quantificação da incerteza, calibração, XAI e limiares operacionais. Este capítulo materializa a contribuição central da tese ao transformar essas saídas, ainda organizadas no nível da imagem, em uma representação temporal contínua e interpretável da dor neonatal (Figura 38). Dessa forma, responde à quinta pergunta específica de pesquisa: como a integração das probabilidades, da incerteza, dos limiares operacionais e das explicações em um sinal contínuo permite caracterizar a dinâmica e a confiabilidade temporal das expressões faciais associadas à dor neonatal?

Grande parte da literatura em visão computacional para dor neonatal concentra-se em classificações estáticas, expressas como classes binárias ou estimativas pontuais em instantes isolados, como visto no Capítulo 2. Embora úteis para avaliar desempenho discriminativo, essas saídas não representam adequadamente a natureza dinâmica da experiência dolorosa nem oferecem, por si só, um instrumento compatível com a rotina assistencial, na qual é necessário acompanhar tendências, persistência, transições e episódios agudos ao longo do tempo (Anand, 2017; Llerena *et al.*, 2023).

Abordagens temporais baseadas apenas na agregação de predições quadro a quadro tendem a ignorar componentes da segurança clínica, como incerteza preditiva, calibração probabilística e limiares operacionais. Depois de desenvolvidos nos capítulos anteriores, esses elementos são integrados em uma representação temporal única, denominada sinal da dor. O

sinal descreve a evolução da probabilidade de dor ao longo do tempo e permite caracterizar sua estabilidade e confiabilidade.

O sinal da dor é definido como uma série temporal derivada das probabilidades de dor estimadas pelos modelos, acompanhada por medidas de incerteza e interpretada à luz dos limiares operacionais previamente estabelecidos. Diferentemente de uma classificação pontual, essa representação permite identificar padrões como estabilidade, transições abruptas, flutuações irregulares e estados de indeterminação. Conceitualmente, o sinal constitui uma camada de informação compatível com uma futura arquitetura HITL, pois organiza o que o modelo estima, quão frágil é essa estimativa e como ela evolui antes de sua interpretação pelo profissional.

Este capítulo integra os pilares desenvolvidos ao longo da tese e avalia o sinal da dor como uma representação destinada ao monitoramento contínuo. O sinal organiza as predições em categorias temporais interpretáveis e marca a passagem conceitual de classificadores estáticos para um monitor dinâmico. Ao integrar desempenho preditivo, incerteza, calibração, limiares operacionais, explicabilidade e modelagem temporal, a abordagem estabelece uma base formal para a avaliação automática e contínua da expressão facial associada à dor neonatal.

9.2 MATERIAIS

9.2.1 Banco de vídeos iCOPEVid

O banco iCOPEVid amplia o iCOPE ao incorporar sequências temporais de vídeo de outros RNs, o que permite capturar variações dinâmicas das expressões faciais neonatais (Brahnam *et al.*, 2020).

Figura 39 – Exemplos de *frames* do banco iCOPEVid.



Fonte: Banco de vídeos iCOPEVid (Brahnam *et al.*, 2020).

Foram incluídos 49 recém-nascidos (RNs), com idades entre 34 e 70 horas, no Mercy Hospital de Springfield (EUA), após consentimento informado dos pais. Os vídeos foram gravados com a câmera Sony HDR-XR520V HDD Camcorder, em resolução 1920×1080, sob iluminação ambiente, a 29,97 quadros por segundo (Figura 39). Ao todo, foram obtidos 246

vídeos de 1 minuto durante procedimentos rotineiros de enfermagem. Conforme Brahnam *et al.* (2020), 175 segmentos de 20 segundos foram anotados manualmente por profissionais de saúde, resultando em 49 segmentos classificados como dor, 71 como repouso, 34 durante fricção e 21 em movimentação. Neste estudo, utilizaram-se exclusivamente os segmentos rotulados como "dor" e "repouso". Destaca-se que os trechos de repouso foram selecionados manualmente para incluir expressões de difícil classificação, tornando a tarefa de detecção mais desafiadora.

A Tabela 27 apresenta as principais características de cada banco de dados empregado neste estudo.

Tabela 27 – Características dos bancos de dados utilizados.

Banco de dados	**iCOPE**	**UNIFESP**	**iCOPEVid**
Tipo de dados	Imagem	Imagem	Vídeo
Nº de RNs	26	30	49
Masculino	13	15	26
Feminino	13	15	23
Idade pós-natal (h)	18–72	24–144	34–70
Resolução da câmera	3008×2000	320×233	1920×1080
Total de amostras	123	360	120
Dor	60	162	49
Sem dor	63	198	71

Fonte: Autor.

9.3 MÉTODOS

9.3.1 Sinal da dor

Neste trabalho, o sinal da dor é definido como uma representação temporal composta pela trajetória suavizada da probabilidade de dor e por informações complementares de incerteza, limiares operacionais e explicabilidade. A curva de probabilidade constitui seu núcleo quantitativo, enquanto os demais componentes qualificam sua confiabilidade, suas condições de interpretação e a necessidade de supervisão humana ao longo do tempo.

A construção do núcleo quantitativo do sinal baseou-se nas probabilidades estimadas pelos modelos de IA para cada *frame* dos vídeos, organizadas como uma série temporal contínua. Para cada instante discreto t, o modelo fornece uma probabilidade de dor $\hat{p}_t \in [0, 1]$, de modo que o sinal bruto é definido como

$$s_{\text{bruto}}(t) = \hat{p}_t, \quad t = 1, 2, \ldots, T,$$

em que T corresponde ao número total de *frames* do vídeo.

Como a detecção facial pode falhar em determinados instantes, gerando valores ausentes, aplicou-se interpolação linear para reconstruir a continuidade temporal. Considerando dois

instantes válidos consecutivos t_a e t_b, com $t_a < t < t_b$, a probabilidade interpolada foi calculada da seguinte forma:

$$\hat{p}_t = \hat{p}_{t_a} + \frac{t - t_a}{t_b - t_a} \left(\hat{p}_{t_b} - \hat{p}_{t_a} \right).$$

Após a reconstrução da continuidade temporal, aplicou-se suavização por média móvel com janela $w = 30$ amostras (aproximadamente 1 segundo de vídeo), definida por

$$s(t) = \frac{1}{w} \sum_{i=0}^{w-1} \hat{p}_{t-i},$$

para $t \geq w$.

O sinal final $s(t)$ representa a evolução temporal suavizada da probabilidade de dor ao longo do vídeo. A suavização reduz flutuações de alta frequência e facilita a observação de padrões dinâmicos.

9.3.2 Métricas dinâmicas do sinal da dor

9.3.2.1 Média temporal da probabilidade

Para a classificação global do sinal da dor em cada clipe, calculou-se a média temporal das probabilidades ao longo do tempo. Seja $s(t)$ o sinal, com $t = 1, \ldots, T$. A média temporal foi definida como

$$\hat{p} = \frac{1}{T} \sum_{t=1}^{T} s(t),$$

sendo $\hat{p}$ utilizada como estimativa agregada da classificação final do modelo, tal que:

$$\text{Classificação de } s(t) = \begin{cases} Com\ dor & \text{se } \hat{p} \geq \theta_1 \\ Sem\ dor & \text{se } \hat{p} < \theta_1 \end{cases} \tag{9.1}$$

9.3.2.2 Incerteza temporal do modelo

Além da probabilidade de dor, considerou-se a incerteza associada às predições do modelo. Seja $\hat{\sigma}(t)$ a incerteza estimada no instante t, definida como o desvio-padrão da distribuição de probabilidades do modelo. A partir do limiar de incerteza θ_3, definiu-se que

$$\text{Classificação quanto à incerteza de } s(t) = \begin{cases} \text{Incerto}, & \text{se } \hat{\sigma}(t) > \theta_3, \\ \text{Confiável}, & \text{se } \hat{\sigma}(t) \leq \theta_3. \end{cases}$$

9.3.2.3 Entropia de Shannon

A entropia de Shannon (Shannon, 1948) foi empregada para mensurar a imprevisibilidade do sinal, sendo definida como

$$\mathbb{H} = -\sum_{i=1}^{B} P(\xi_i) \log_2 P(\xi_i),$$

em que ξ representa a variável aleatória associada aos valores discretizados do sinal e $P(\xi_i)$ corresponde à probabilidade de ocorrência do i-ésimo intervalo.

Neste trabalho, o sinal $s(t)$ foi discretizado em $B = 10$ intervalos igualmente espaçados no intervalo $[0, 1]$. A probabilidade empírica de cada intervalo foi estimada por

$$P(\xi_i) = \frac{n_i}{T},$$

em que n_i representa o número de amostras pertencentes ao intervalo i. Valores de entropia próximos de zero indicam concentração do sinal em poucos intervalos, caracterizando comportamento estável e previsível. À medida que a distribuição se aproxima da uniformidade, a entropia aumenta, refletindo maior imprevisibilidade. O valor máximo ocorre quando $P(\xi_i) = 0{,}1$ para todos os intervalos, isto é, quando todos os $B = 10$ intervalos têm a mesma probabilidade de ocorrer, resultando em

$$\mathbb{H}_{\text{max}} = -\sum_{i=1}^{10} 0{,}1 \log_2 0{,}1 \approx 3{,}32.$$

Dessa forma, a entropia de Shannon fornece uma métrica quantitativa para caracterizar a dinâmica temporal do sinal da dor, permitindo distinguir padrões estáveis de padrões instáveis ao longo do tempo.

9.3.2.4 Transições de classe

Para quantificar a alternância entre estados de dor e ausência de dor ao longo do tempo, contabilizou-se o número de transições de classe. Considerando o limiar de decisão θ_1, definiu-se a classe instantânea como

$$c(t) = \begin{cases} \textit{com dor}, & \text{se } s(t) \geq \theta_1, \\ \textit{sem dor}, & \text{se } s(t) < \theta_1. \end{cases}$$

O número total de transições N_{trans} foi obtido por

$$N_{\text{trans}} = \sum_{t=2}^{T} \mathbb{I}\big(c(t) \neq c(t-1)\big),$$

em que $\mathbb{I}(\cdot)$ é a função indicadora. Valores elevados indicam alternâncias frequentes entre estados e, consequentemente, maior instabilidade temporal.

9.4 EXPERIMENTO

O experimento avaliou se os modelos finais mantiveram sua utilidade quando as saídas foram projetadas no domínio temporal. Cada vídeo foi decomposto em *frames*, nos quais a detecção facial foi realizada com o mesmo SCRFD-10GF do Capítulo 3. Apenas os *frames* com detecção válida da face foram considerados. Os demais foram descartados da análise dos modelos. Em seguida, os modelos NCNN, VGGFace e ViT-B/32, com os métodos de quantificação da incerteza e calibração apresentados no Capítulo 8, produziram uma probabilidade de dor $\hat{p}_t \in [0, 1]$ para cada *frame*. Obteve-se, assim, uma sequência temporal de predições acompanhada do valor de incerteza $\hat{\sigma}$.

A partir dessas probabilidades, construiu-se o sinal temporal da dor. Lacunas decorrentes de falhas na detecção facial foram preenchidas por interpolação linear, tanto para as probabilidades quanto para os valores de incerteza associados, seguida de suavização por média móvel com janela de 30 amostras (aproximadamente 1 segundo), obtendo-se os sinais suavizados.

O método de unificação de XAI proposto no Capítulo 6 também foi aplicado para gerar explicações locais. Com isso, foi possível visualizar as regiões faciais mais relevantes para as predições em cada instante e relacioná-las aos padrões temporais identificados. Para compreender a correlação dos sinais de cada modelo com as expressões faciais dos RNs, foram calculadas as correlações de Pearson e de Spearman entre as probabilidades preditas e os valores de importância de cada região facial ao longo do tempo. Além da correlação temporal direta, calculou-se a correlação entre as derivadas temporais, a fim de verificar se mudanças nas probabilidades preditas estavam associadas a mudanças nas regiões faciais relevantes.

Com o sinal construído, procedeu-se à sua caracterização dinâmica por meio do cálculo da entropia de Shannon e da contagem de transições de classe com base no limiar θ_1. Após a extração das métricas dinâmicas, utilizou-se o algoritmo *k*-means para identificar padrões temporais recorrentes entre os diferentes clipes. O número ótimo de clusters foi determinado experimentalmente, variando-se a contagem de grupos de dois a dez e selecionando-se a configuração com o maior índice de silhueta (Rousseeuw, 1987), que avalia a coesão e separação dos grupos formados.

Cada sinal foi representado por um vetor de características definido como

$$\mathbf{x}_j = \begin{bmatrix} \mathbb{H}_j \\ N_{\text{trans},j} \end{bmatrix},$$

em que $\mathbb{H}_j$ corresponde à entropia do j-ésimo sinal e $N_{\text{trans},j}$ ao número total de transições no respectivo clipe.

O algoritmo busca minimizar a soma das distâncias quadráticas entre cada ponto e o centróide do grupo ao qual pertence, formalmente definida como

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2,$$

em que C_i representa o conjunto de sinais do grupo i e $\boldsymbol{\mu}_i$ seu centróide.

Antes da aplicação do algoritmo, as variáveis foram normalizadas para evitar que diferenças de escala entre entropia e número de transições influenciassem o agrupamento.

Após a categorização, os sinais foram avaliados segundo as métricas de classificação e calibração, considerando-se a média temporal da probabilidade $\hat{p}$ como critério classificatório. Adicionalmente, com a utilização dos limiares de precisão e incerteza, calculou-se o percentual de dados rejeitados para revisão humana.

9.5 RESULTADOS

9.5.1 Avaliação do sinal da dor e seus tipos

A Tabela 28 apresenta as métricas de classificação e calibração quando os clipes de 20 s são condensados em uma única predição por meio da média temporal $\hat{p}$. Observou-se redução consistente nas métricas de classificação em comparação com os resultados obtidos na validação cruzada dos capítulos anteriores. Ainda assim, os modelos apresentaram desempenho próximo ao relatado por Brahnam *et al.* (2020), que reportaram AUC de 79,8%, em comparação com 72,8% da NCNN, 74,8% da VGGFace e 68,3% da ViT-B/32 na mesma base de dados.

Tabela 28 – Métricas de classificação e calibração dos modelos no sinal da dor sem aplicação da metodologia proposta (valores em %).

Métrica	**NCNN**	**VGGFace**	**ViT-B/32**
Acurácia	**65,8**	65,0	60,0
F1-Score	**67,7**	64,4	61,9
Sensibilidade	**87,8**	77,6	79,6
Especificidade	50,7	**56,3**	46,5
Precisão	**55,1**	**55,1**	50,7
AUC	72,8	**74,8**	68,3
ECE	19,3	**15,4**	17,9
MCE	**30,2**	45,0	33,6
NLL	68,9	**66,7**	76,4
Brier	23,5	**23,0**	26,4

Fonte: Autor.

À primeira vista, esses resultados são compatíveis com *overfitting* no treinamento. A análise temporal, porém, mostrou que muitos falsos positivos estavam associados a quadros nos quais os RNs apresentavam expressões faciais compatíveis com dor, embora o vídeo tivesse sido rotulado como "sem dor". As expressões também variaram durante os 20 segundos de gravação, em contraste com a classe única atribuída a todo o vídeo.

A análise dos sinais da dor indicou que, para todos os modelos, o número ótimo de agrupamentos no *k*-means foi três, com base na maximização do índice de silhueta. Esses agrupamentos foram posteriormente denominados estável, irregular e instável, de acordo com

suas características distintivas. Os resultados, incluindo métricas de classificação e calibração para cada tipo de sinal, estão sumarizados na Tabela 29.

Exemplos de sinais da dor são apresentados nas Figuras 41 a 44. Nelas, as linhas coloridas representam a evolução da probabilidade predita de dor ao longo dos 20 segundos do vídeo para cada modelo. Os marcadores indicam os instantes em que o sinal da dor cruza o limiar de classificação específico de cada modelo. Abaixo de cada sinal, são exibidos quadros do vídeo em intervalos de 1 segundo, acompanhados de suas respectivas máscaras de explicação de consenso global. A legenda na parte superior de cada figura apresenta, na seguinte ordem: a classe real do vídeo, sua predição final com probabilidade média ($\hat{p}$), a classificação temporal do sinal e a incerteza média ($\hat{\sigma}$) do sinal, com base no limiar θ_3.

Esses achados foram sistematizados nas matrizes de decisão apresentadas na Figura 40, que orientam a necessidade de intervenção clínica de acordo com o tipo de sinal da dor. Para assegurar consistência entre os modelos, empregaram-se termos qualitativos, como entropia alta ou baixa e número elevado ou reduzido de cruzamentos de limiar, enquanto as definições quantitativas específicas de cada agrupamento foram detalhadas na Tabela 30. Por exemplo, na VGGFace, entropia alta corresponde aos intervalos observados para sinais irregulares ($1{,}8 \pm 0{,}6$) e instáveis ($2{,}6 \pm 0{,}5$).

Tabela 29 – Métricas de classificação e calibração por tipo de sinal da dor e modelo após a rejeição pelos limiares θ_2 e θ_3 (valores em %). "–" indica que não foi possível calcular a métrica.

Modelo	Tipo	Acurácia	F1-Score	Sensibilidade	Especificidade	Precisão	AUC	ECE	MCE	NLL	Brier
NCNN	Estável	**73,1**	**75,9**	**91,7**	**57,1**	**64,7**	**74,4**	**19,9**	**34,5**	**76,4**	**23,6**
	Irregular	58,1	52,6	76,9	50,0	40,0	63,1	31,4	62,8	91,1	30,9
	Instável	52,4	61,5	80,0	27,3	50,0	45,5	38,3	84,1	106,7	37,3
	Indeterminado	–	–	–	–	–	–	–	–	–	–
VGGFace	Estável	**80,0**	**72,2**	81,3	**79,4**	**65,0**	**88,6**	**17,4**	**30,8**	**62,4**	**18,2**
	Irregular	63,0	70,2	**87,0**	39,1	58,8	59,6	31,1	43,4	101,0	31,6
	Instável	41,7	46,2	60,0	28,6	37,5	50,0	42,9	76,5	116,9	40,8
	Indeterminado	–	–	–	–	–	–	–	–	–	–
ViT-B/32	Estável	**72,0**	**72,0**	**90,0**	**60,0**	**60,0**	**76,5**	**25,5**	**36,1**	**94,0**	**26,0**
	Irregular	53,3	51,2	68,8	44,8	40,7	54,1	39,9	75,3	119,9	38,0
	Instável	47,4	58,3	58,3	28,6	58,3	46,4	39,6	78,7	104,9	36,7
	Indeterminado	–	–	–	–	–	–	–	–	–	–

Fonte: Autor.

9.5.1.1 Sinais Estáveis

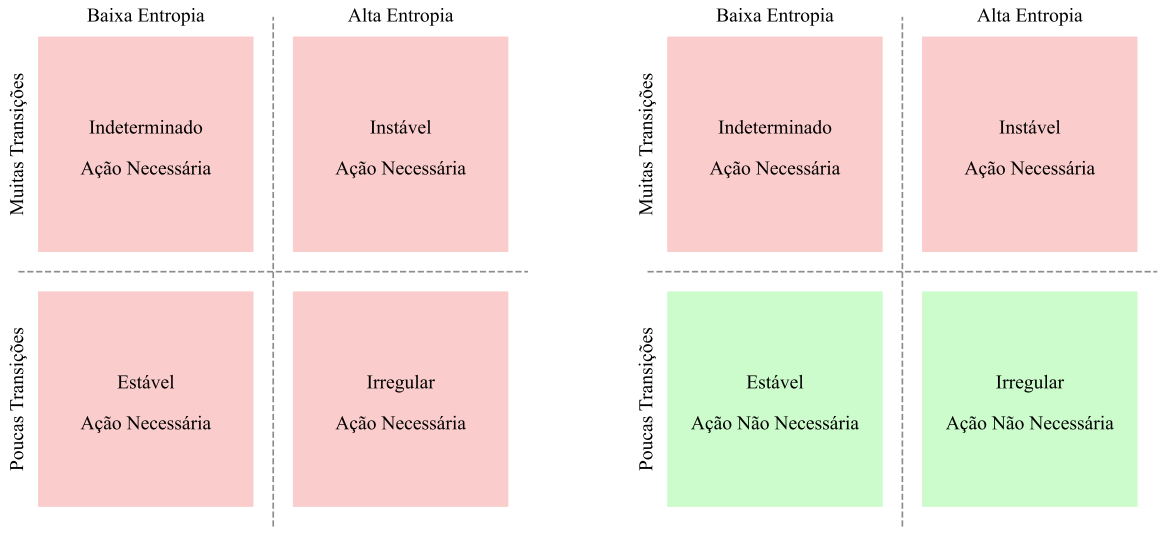
Os sinais classificados como estáveis (Figura 41) apresentaram baixa entropia e poucos cruzamentos do limiar de decisão θ_1, mantendo comportamento consistente ao longo do vídeo. Eles se concentraram nas regiões de alta confiança definidas por θ_2, tanto para "dor" quanto para "sem dor", e apresentaram baixa rejeição pelos limiares. Esse grupo apresentou as melhores métricas discriminativas e correspondeu ao regime de maior confiabilidade do sistema, como mostram os menores percentuais médios de quadros rejeitados por cada limiar na Tabela 30. Todos os modelos obtiveram melhor desempenho nos sinais estáveis, refletindo a consistência

Tabela 30 – Centroides dos agrupamentos e percentual médio de *frames* rejeitados pelos limiares θ_2 e θ_3, por tipo de sinal da dor.

Modelo	Tipo	Dor	Sem Dor	Entropia	Transições	Rejeitados por θ_2 (%)	Rejeitados por θ_3 (%)	Rejeitados por θ_{2+3} (%)
NCNN	Estável	24	29	0,8 ± 0,7	0,1 ± 0,3	20,5 ± 26,2	20,5 ± 26,0	26,60 ± 27,87
	Irregular	14	30	2,3 ± 0,5	2,4 ± 0,9	50,8 ± 26,3	53,1 ± 29,8	59,23 ± 27,30
	Instável	11	11	2,6 ± 0,5	6,8 ± 1,4	76,4 ± 20,0	72,9 ± 17,0	80,15 ± 17,23
	Indeterminado	0	1	1,5	2	100,0	19,0	100,00
VGGFace	Estável	16	34	0,4 ± 0,5	0,2 ± 0,4	10,6 ± 17,1	14,5 ± 20,7	15,97 ± 20,53
	Irregular	23	23	1,8 ± 0,6	2,5 ± 0,9	33,6 ± 19,3	42,2 ± 20,5	44,56 ± 20,34
	Instável	10	14	2,6 ± 0,5	6,8 ± 2,3	60,2 ± 17,5	73,0 ± 13,1	75,18 ± 13,18
	Indeterminado	0	0	—	—	—	—	—
ViT-B/32	Estável	20	31	1,2 ± 0,8	0,2 ± 0,4	16,0 ± 19,1	32,4 ± 28,6	33,94 ± 28,61
	Irregular	16	30	2,6 ± 0,5	3,0 ± 1,1	54,5 ± 22,0	68,2 ± 22,5	70,90 ± 22,13
	Instável	13	10	2,6 ± 0,4	8,0 ± 2,3	76,4 ± 17,4	83,1 ± 15,0	84,84 ± 14,08
	Indeterminado	0	0	—	—	—	—	—

Fonte: Autor.

Figura 40 – Matrizes de decisão para sinais da dor nos estados “com dor” e “sem dor”.



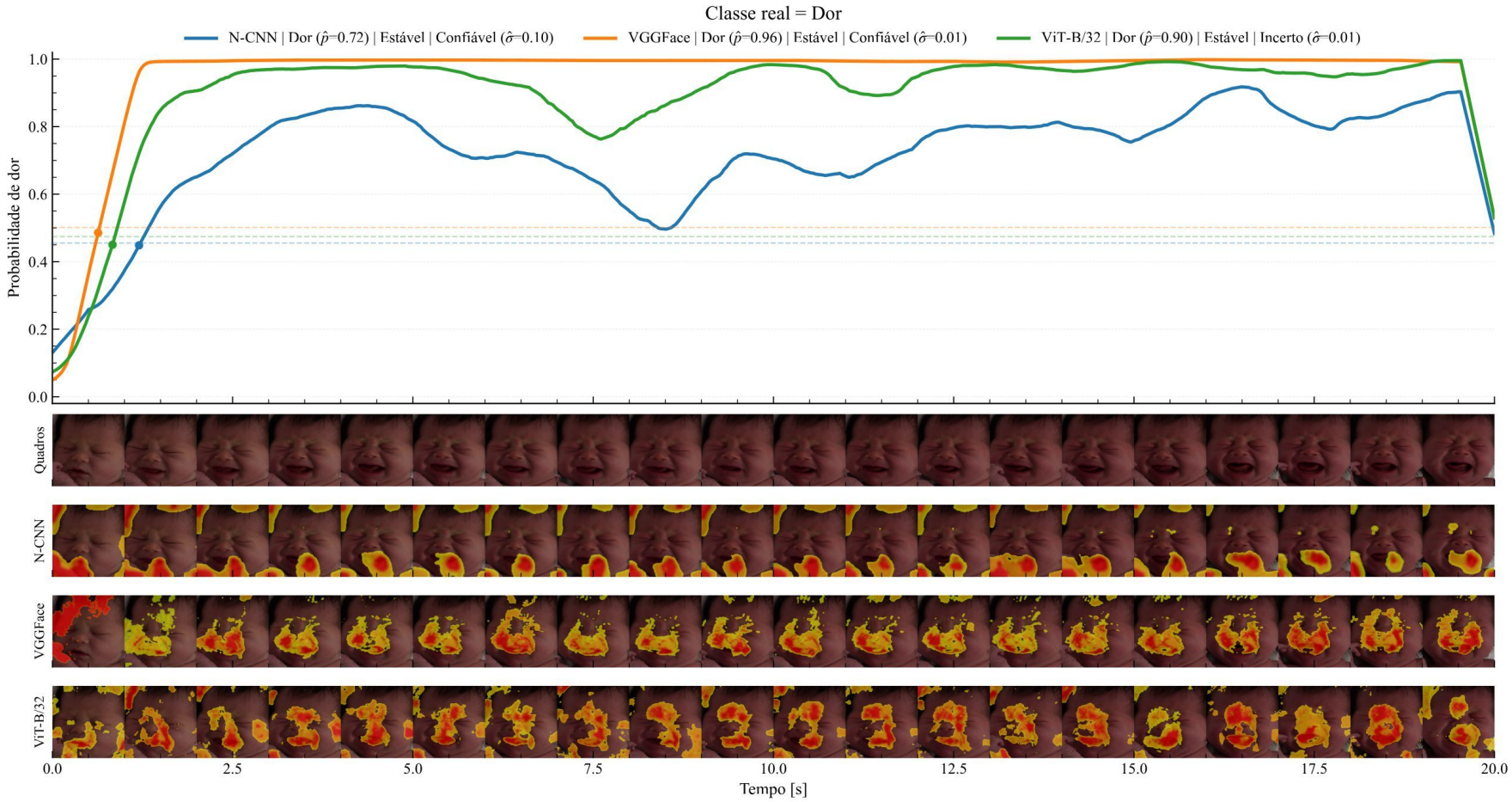
(a) Matriz de decisão — com dor.

(b) Matriz de decisão — sem dor.

Fonte: Autor.

das expressões faciais ao longo do vídeo. Portanto, quando a predição indica “dor”, a equipe de saúde deve priorizar a avaliação do RN.

Figura 41 – Sinal da dor do tipo estável.

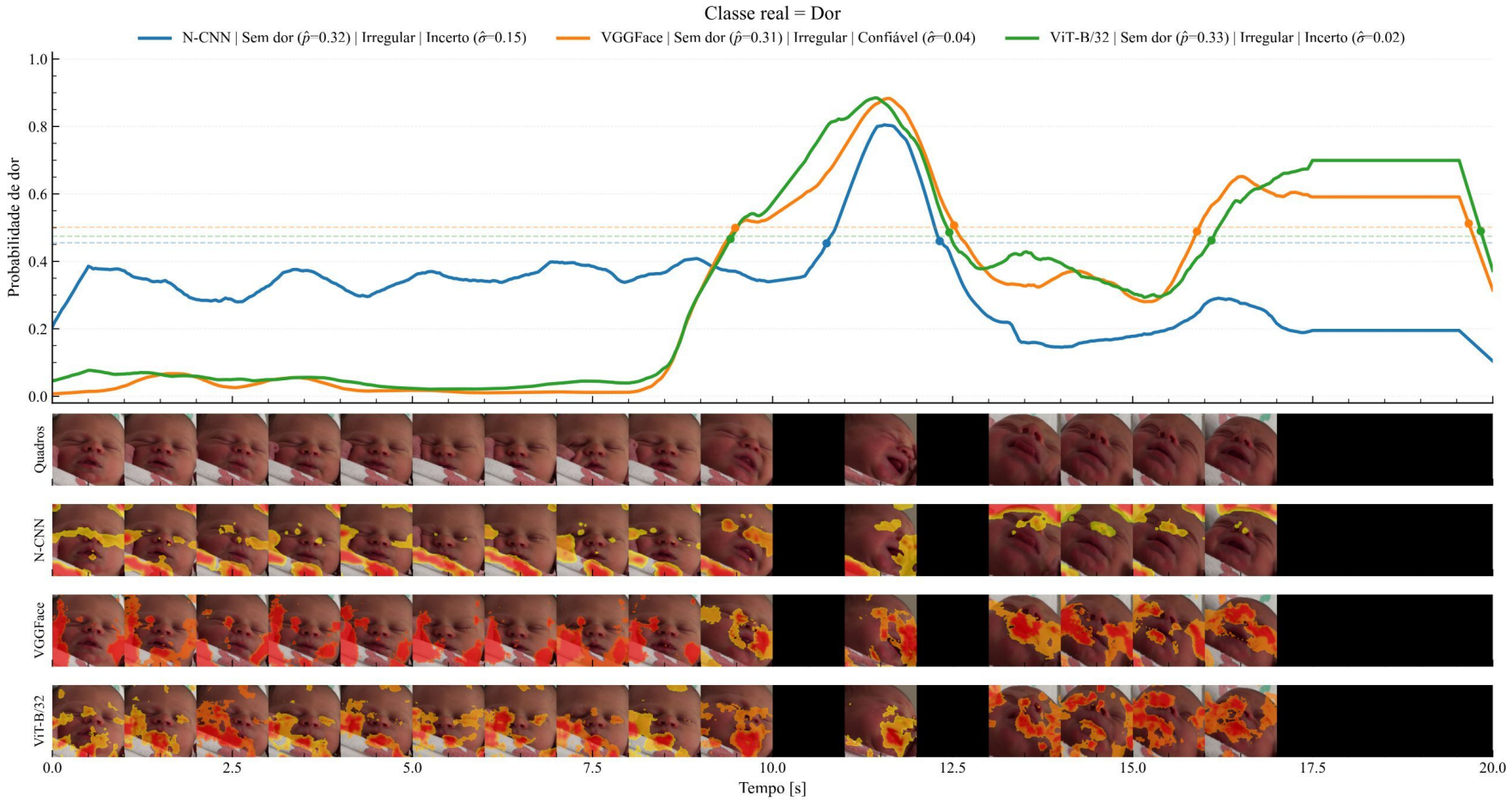


Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

9.5.1.2 Sinais Irregulares

Os sinais irregulares (Figura 42) exibiram entropia intermediária e poucos cruzamentos de θ_1. As probabilidades oscilaram predominantemente na região intermediária delimitada por θ_2, sem atingir de forma consistente níveis de confiança que caracterizassem mudanças sustentadas entre dor e ausência de dor. Essas irregularidades estiveram frequentemente associadas a oclusões faciais ou a variações abruptas da expressão. Nesses casos, a probabilidade apresentava picos bem definidos, mas retornava rapidamente ao estado anterior, sem confiança suficiente para alterar o estado predito do RN. Em comparação com os sinais estáveis, houve redução das métricas discriminativas, aumento dos erros probabilísticos e piora moderada da calibração. Esses sinais também demandam atenção clínica, especialmente quando a predição é “dor”.

Figura 42 – Sinal da dor do tipo irregular.

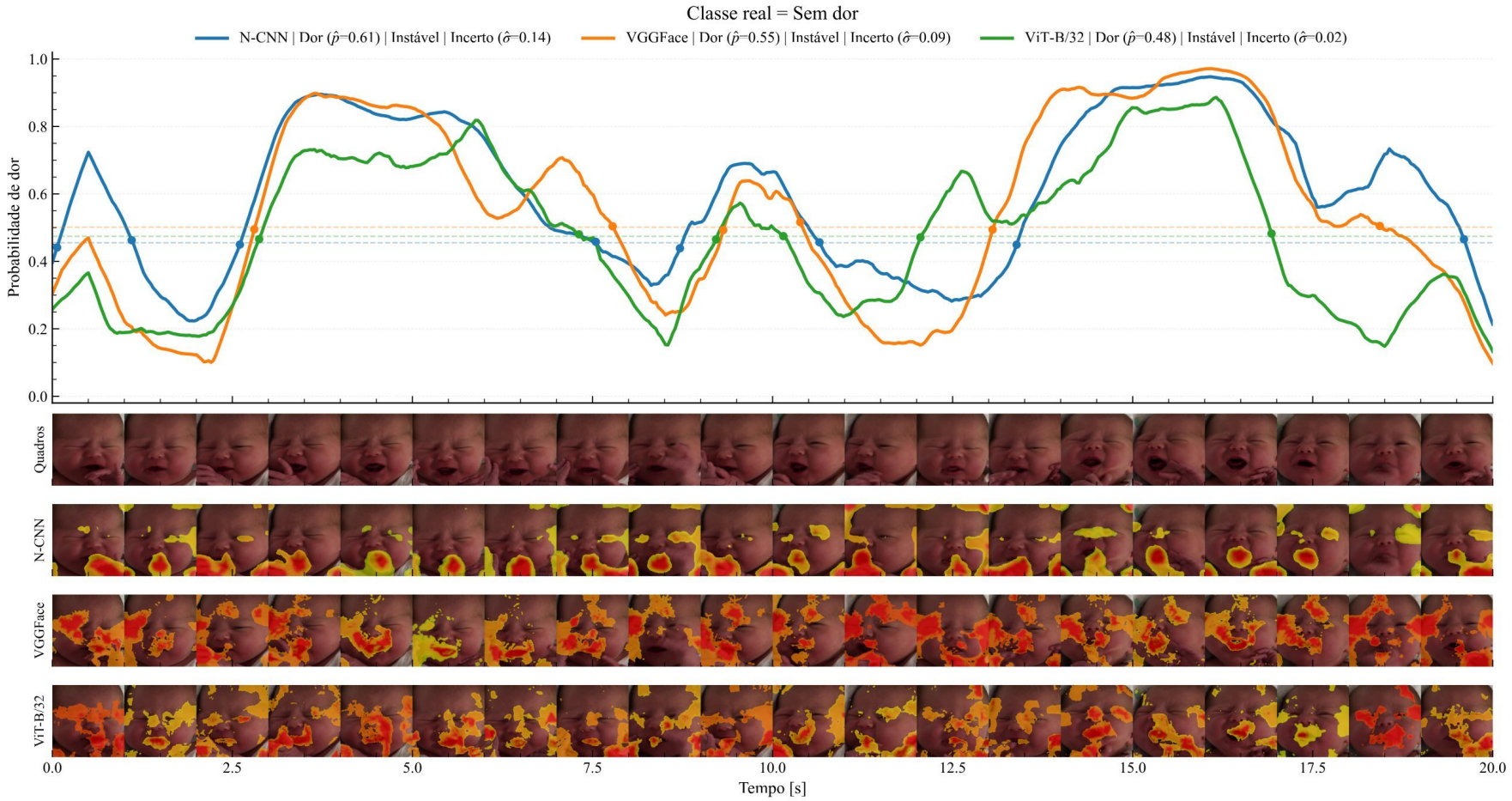


Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

9.5.1.3 Sinais Instáveis

Os sinais instáveis (Figura 43) foram definidos por alta entropia e elevado número de cruzamentos de θ_1, caracterizando alternâncias frequentes entre estados preditivos. Nesses casos, grande parte das amostras situou-se fora das regiões de alta confiança (θ_2) e apresentou valores acima do limiar de incerteza θ_3, o que resultou nos maiores percentuais de rejeição, tanto para os limiares isolados quanto para sua aplicação conjunta (Tabela 30). Esse padrão resultou nas piores métricas discriminativas, além de degradação substancial das métricas de calibração, evidenciada pelo aumento expressivo de ECE, MCE e NLL. Tal comportamento pode refletir tanto um evento clinicamente dinâmico quanto instabilidade preditiva estrutural do modelo, configurando o regime de menor confiabilidade operacional. Em sinais instáveis, uma única predição ou classe não representa adequadamente a variação contínua das expressões faciais, o que justifica a recomendação de supervisão humana tanto para predições de “dor” quanto de “sem dor”.

Figura 43 – Sinal da dor do tipo instável.

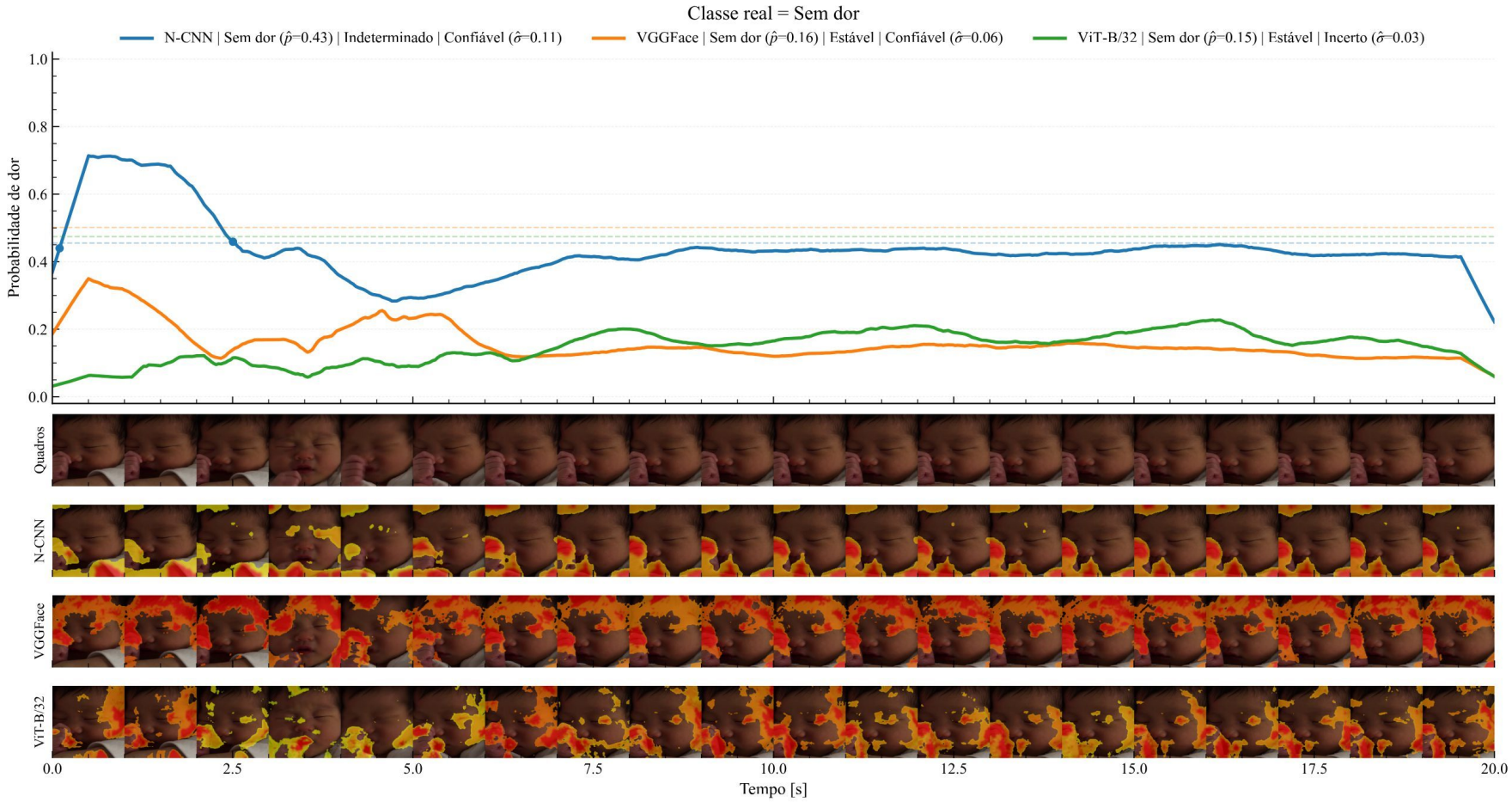


Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

9.5.1.4 Sinais Indeterminados

Identificou-se ainda um quarto padrão preliminar, denominado indeterminado e ilustrado no quadrante superior esquerdo das matrizes de decisão (Figura 40). Nesse padrão, o sinal flutua em torno do limiar de classificação θ_1, com baixa entropia. Não há predominância clara de "dor" ou "sem dor", e o sinal permanece, em sua maior parte, na região intermediária definida por θ_2. Esse padrão ocorreu em apenas um exemplo da NCNN no conjunto avaliado (Figura 44), caracterizando uma ocorrência empírica rara de ambiguidade.

Figura 44 – Sinal da dor do tipo indeterminado do modelo NCNN.



Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

9.5.1.5 Correlação entre os sinais da dor

Embora os modelos tenham apresentado desempenho classificatório semelhante, a correlação temporal entre seus sinais de dor foi apenas moderada. Nos 120 vídeos avaliados, o par de modelos NCNN e VGGFace apresentou as maiores correlações, enquanto as comparações envolvendo o ViT-B/32 mostraram correlações inferiores, conforme mostrado na Tabela 31. Ao analisar as derivadas temporais dos sinais, com o objetivo de avaliar a correlação nas variações instantâneas e nas transições entre estados, observou-se padrão semelhante.

Na concordância categórica dos tipos de sinal, estimada pelo coeficiente κ de Cohen (Cohen, 1960), adotaram-se as faixas interpretativas propostas por Landis e Koch (1977): concordância pobre para $\kappa < 0{,}00$, leve para $0{,}00 \leq \kappa \leq 0{,}20$, fraca para $0{,}21 \leq \kappa \leq 0{,}40$, moderada para $0{,}41 \leq \kappa \leq 0{,}60$, substancial (forte) para $0{,}61 \leq \kappa \leq 0{,}80$ e quase perfeita para $0{,}81 \leq \kappa \leq 1{,}00$. Segundo esses intervalos, observaram-se concordâncias fracas entre NCNN e VGGFace ($\kappa = 0{,}31$) e entre NCNN e ViT-B/32 ($\kappa = 0{,}31$), enquanto a concordância entre VGGFace e ViT-B/32 foi moderada ($\kappa = 0{,}44$).

Em conjunto, os achados mostram que, apesar de métricas globais comparáveis, os modelos não produziram dinâmicas temporais plenamente alinhadas nem tipos de sinais totalmente coincidentes. O desvio-padrão elevado das correlações nas curvas e nas derivadas reforça que arquiteturas distintas capturaram aspectos também distintos tanto do nível quanto da dinâmica das expressões faciais associadas à dor.

Tabela 31 – Correlação entre sinais da dor, suas derivadas e concordância categórica κ.

Comparação entre modelos	**Curva**		**Derivada**		$\boldsymbol{\kappa}$
	Pearson	**Spearman**	**Pearson**	**Spearman**	
NCNN vs. VGGFace	**0,69 ± 0,29**	**0,60 ± 0,32**	0,67 ± 0,22	**0,50 ± 0,22**	0,31
NCNN vs. ViT-B/32	0,63 ± 0,32	0,54 ± 0,32	0,64 ± 0,24	0,42 ± 0,20	0,31
VGGFace vs. ViT-B/32	0,67 ± 0,29	0,57 ± 0,32	**0,68 ± 0,22**	0,47 ± 0,19	**0,44**

Fonte: Autor.

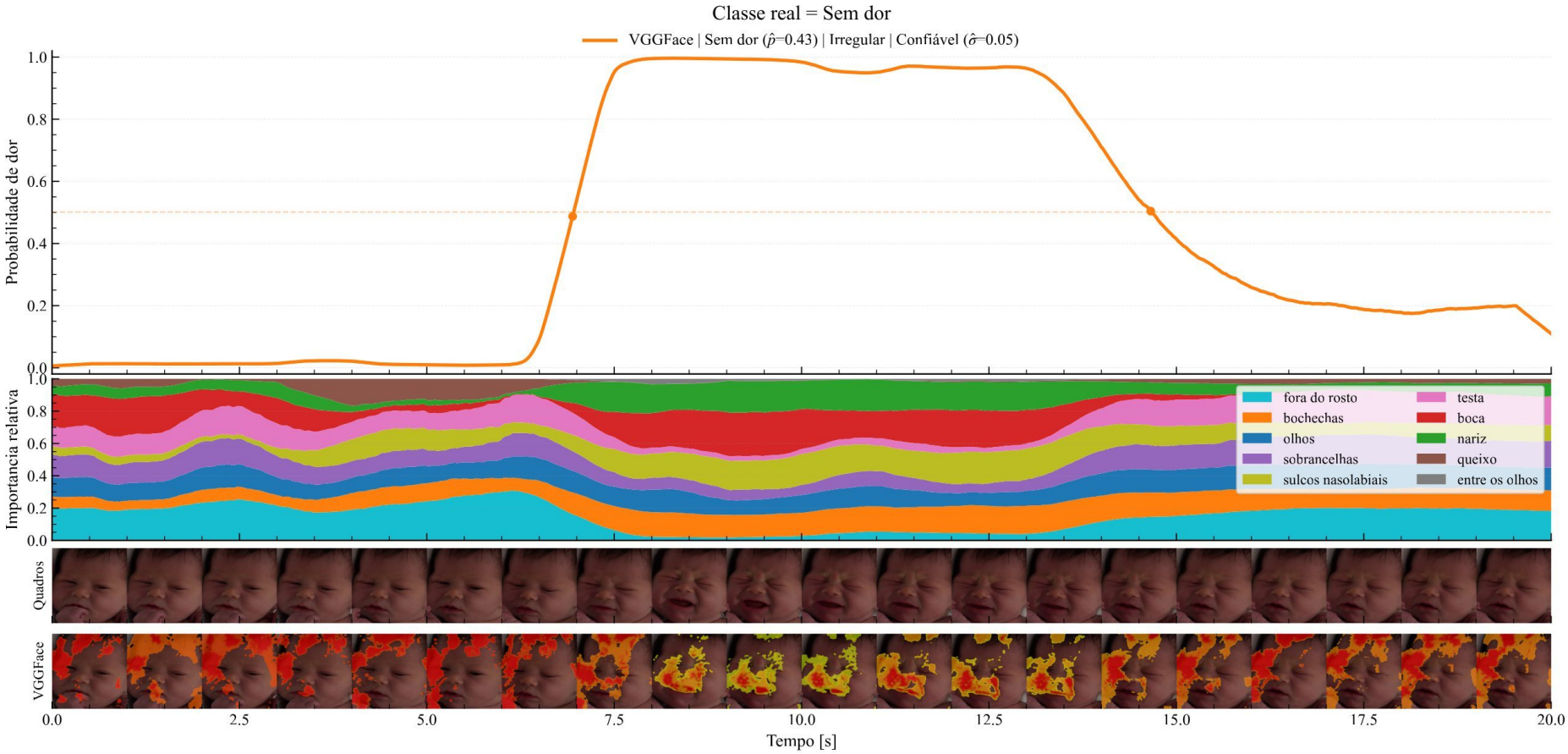
9.5.2 Associação entre o sinal da dor e as expressões faciais

Conforme observado nas Figuras 41 a 44, os mapas de explicação também apresentaram variação temporal. Nos sinais estáveis, as regiões de maior relevância mantiveram-se relativamente consistentes ao longo do vídeo. Em sinais irregulares e instáveis, por sua vez, as regiões relevantes oscilaram de forma mais acentuada, refletindo a variabilidade das expressões faciais e acompanhando as flutuações do sinal da dor.

A Figura 45 ilustra um exemplo dessa dinâmica, no qual as linhas empilhadas coloridas indicam a importância relativa das regiões faciais ao longo dos 20 segundos do vídeo. Quanto maior a altura da linha, maior a relevância daquela região para a predição naquele instante. No

início do vídeo, quando o RN não apresenta expressões faciais associadas à dor, o sinal mantém-se próximo de 0% de probabilidade de dor, e as regiões de maior relevância concentram-se predominantemente fora do rosto, na boca e na testa. A partir do segundo 6, o sinal da dor passa a aumentar, acompanhando o surgimento de expressões faciais compatíveis com dor, e as regiões de maior relevância tornam-se boca, nariz e sulco nasolabial. A partir do segundo 15, o sinal retorna à região “sem dor”, e a relevância volta a concentrar-se nas mesmas áreas observadas nos instantes iniciais do vídeo. Esse comportamento evidencia como a importância das características faciais varia ao longo do tempo e como essa variação se correlaciona com a dinâmica do sinal da dor.

Figura 45 – Evolução temporal do sinal da dor e da importância relativa das regiões faciais do modelo VGGFace.



Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

Para aprofundar a análise dessa associação, calculou-se a correlação entre as derivadas do sinal da dor e as derivadas da relevância das regiões faciais, com o objetivo de verificar se mudanças no sinal eram acompanhadas por alterações significativas nos mapas de atribuição.

Os resultados apresentados na Tabela 32 corroboram os achados anteriormente descritos no Capítulo 6, especialmente na Tabela 15. Observou-se que o modelo NCNN utiliza predominantemente regiões inferiores da face, como boca, queixo e sulco nasolabial. No contexto do sinal da dor, essas mesmas regiões apresentaram correlação moderada com as variações do sinal, indicando coerência entre o padrão explicativo e a dinâmica temporal da predição.

Para a VGGFace, verificou-se maior relevância atribuída a regiões centrais da face, especialmente o nariz, que também se mostraram associadas às variações do sinal da dor. Notavelmente, as correlações negativas da região externa à face evidenciaram relação inversa entre o sinal da dor e a relevância fora da face. Esse resultado é compatível com um deslocamento espacial da relevância atribuída pelo modelo, ou seja, à medida que a probabilidade de dor aumenta, as explicações concentram-se nas regiões faciais, reduzindo a importância atribuída a áreas externas, e o inverso ocorre quando a probabilidade diminui.

Por fim, o modelo ViT-B/32 apresentou correlação moderada entre a região da boca e as variações do sinal da dor, enquanto as demais regiões exibiram correlações fracas. Esse comportamento foi observado tanto em casos com dor quanto em casos sem dor.

9.6 DISCUSSÃO

Transformar predições *frame a frame* em um sinal contínuo revela aspectos temporais da dor neonatal que classificações pontuais não capturam. A decomposição do comportamento em padrões estáveis, irregulares, instáveis e, em casos raros, indeterminados desloca o foco de "acertar uma classe" para interpretar a dinâmica do fenômeno e avaliar a confiabilidade das predições ao longo do tempo.

A estratificação por tipo de sinal evidenciou uma relação consistente entre complexidade temporal e degradação de desempenho e confiabilidade (Tabela 30). Sinais estáveis apresentaram baixa entropia e poucas transições, concentrando-se predominantemente em regiões de maior confiança e resultando em melhor discriminação e menor erro probabilístico, enquanto sinais instáveis exibiram alta entropia e numerosas alternâncias, com piora acentuada das métricas. Essa tendência é compatível com a interpretação de que o regime instável agrega, potencialmente, dois fenômenos: (I) episódios clinicamente dinâmicos, nos quais a expressão facial oscila em decorrência de mudanças reais no estado do RN; e (II) instabilidade preditiva do modelo induzida por ruído visual, oclusões, pose ou iluminação, isto é, condições que elevam a própria incerteza do modelo. O sinal, portanto, não apenas estima a probabilidade de dor, mas também captura condições de observabilidade que afetam a confiabilidade do sistema.

A análise com filtragem por θ_2 e θ_3 reforça o papel complementar dos limiares operacionais na dimensão temporal. Em sinais estáveis, os baixos percentuais de rejeição por θ_2 e θ_3

Tabela 32 – Correlações regionais entre a derivada do sinal da dor e as derivadas das séries temporais de XAI por modelo e classe (média ± desvio-padrão).

Modelo	**Região**	**Dor**		**Sem dor**	
		Spearman	**Pearson**	**Spearman**	**Pearson**
NCNN	Olhos	-0,021 ± 0,229	0,110 ± 0,215	-0,025 ± 0,231	0,059 ± 0,280
	Sobrancelhas	0,013 ± 0,203	0,101 ± 0,176	0,054 ± 0,204	0,082 ± 0,203
	Bochechas	-0,026 ± 0,233	0,110 ± 0,249	-0,083 ± 0,237	0,040 ± 0,328
	Nariz	-0,023 ± 0,229	0,091 ± 0,242	0,038 ± 0,208	0,112 ± 0,287
	Boca	**0,417 ± 0,193**	**0,576 ± 0,177**	**0,403 ± 0,208**	**0,541 ± 0,227**
	Queixo	0,183 ± 0,237	0,353 ± 0,190	0,187 ± 0,216	0,311 ± 0,222
	Testa	-0,049 ± 0,181	0,042 ± 0,188	-0,024 ± 0,228	0,049 ± 0,238
	Entre os olhos	0,055 ± 0,208	0,113 ± 0,182	0,079 ± 0,189	0,133 ± 0,216
	Sulcos nasolabiais	0,260 ± 0,219	0,425 ± 0,183	0,282 ± 0,191	0,391 ± 0,206
	Fora da face	-0,170 ± 0,215	0,109 ± 0,277	-0,171 ± 0,230	0,034 ± 0,378
VGGFace	Olhos	0,113 ± 0,229	0,293 ± 0,222	0,173 ± 0,204	0,298 ± 0,253
	Sobrancelhas	-0,004 ± 0,213	0,165 ± 0,227	0,056 ± 0,212	0,156 ± 0,270
	Bochechas	0,181 ± 0,192	**0,409 ± 0,202**	0,194 ± 0,183	**0,340 ± 0,219**
	Nariz	**0,275 ± 0,197**	0,369 ± 0,203	**0,279 ± 0,220**	**0,340 ± 0,237**
	Boca	0,181 ± 0,214	0,361 ± 0,263	0,188 ± 0,231	0,325 ± 0,256
	Queixo	-0,181 ± 0,220	-0,112 ± 0,223	-0,076 ± 0,237	-0,031 ± 0,252
	Testa	-0,121 ± 0,192	0,004 ± 0,230	-0,080 ± 0,212	0,001 ± 0,293
	Entre os olhos	0,186 ± 0,219	0,258 ± 0,202	0,233 ± 0,199	0,263 ± 0,239
	Sulcos nasolabiais	0,119 ± 0,227	0,349 ± 0,258	0,122 ± 0,209	0,272 ± 0,252
	Fora da face	-0,302 ± 0,183	-0,139 ± 0,257	-0,218 ± 0,234	-0,098 ± 0,351
ViT-B/32	Olhos	-0,035 ± 0,182	0,096 ± 0,203	-0,013 ± 0,181	0,106 ± 0,221
	Sobrancelhas	-0,019 ± 0,152	0,098 ± 0,167	-0,009 ± 0,179	0,095 ± 0,205
	Bochechas	0,002 ± 0,181	0,167 ± 0,175	0,021 ± 0,187	0,166 ± 0,200
	Nariz	0,050 ± 0,176	0,148 ± 0,172	0,017 ± 0,162	0,134 ± 0,196
	Boca	**0,135 ± 0,141**	**0,365 ± 0,238**	**0,122 ± 0,175**	**0,314 ± 0,261**
	Queixo	0,038 ± 0,138	0,100 ± 0,161	0,067 ± 0,208	0,144 ± 0,248
	Testa	0,024 ± 0,134	0,123 ± 0,151	0,019 ± 0,163	0,112 ± 0,199
	Entre os olhos	0,082 ± 0,144	0,178 ± 0,148	0,033 ± 0,171	0,104 ± 0,199
	Sulcos nasolabiais	0,077 ± 0,150	0,244 ± 0,178	0,079 ± 0,185	0,226 ± 0,215
	Fora da face	0,009 ± 0,167	0,172 ± 0,192	0,023 ± 0,154	0,203 ± 0,240

Fonte: Autor.

indicam um regime em que decisões semi-automáticas são mais justificáveis. Em sinais instáveis, os elevados percentuais de quadros rejeitados, associados ao aumento da incerteza, mostram que a rejeição deve ser interpretada como mecanismo de segurança. Em outras palavras, os limiares operacionais continuam relevantes quando o sistema deixa o domínio estático e passa a operar como monitor temporal.

A comparação entre modelos mostrou que, nas análises realizadas, métricas globais semelhantes coexistiram com dinâmicas temporais apenas moderadamente concordantes e com características de sinal não totalmente coincidentes. No monitoramento contínuo, importa não apenas a classificação final, mas também quando o sistema sinaliza transições, com que estabilidade o faz e em que momentos entra em regime de ambiguidade. Isso reforça a necessidade de incluir medidas de concordância temporal na avaliação de modelos destinados ao monitoramento clínico, em vez de selecionar um único modelo com base apenas em métricas estáticas de classificação. Assim como ocorre entre avaliadores humanos, os modelos apresentaram padrões distintos de análise, retomando o problema de *disagreement* estudado no Capítulo 6.

A correlação regional entre as variações do sinal e as séries temporais de XAI (Tabela 32) evidencia que, além de indicar a probabilidade de dor, o sinal fornece informações sobre as regiões faciais associadas à predição e sobre como essa associação se altera ao longo do tempo. Isso permite, por exemplo, informar que o sinal aumentou em 40% enquanto a relevância atribuída à região da boca também aumentou.

Este capítulo integra os componentes desenvolvidos anteriormente em uma representação temporal contínua, interpretável e operacionalmente útil. Esses resultados materializam a contribuição central da tese: um sinal que reúne a dinâmica da expressão facial e as informações necessárias para interpretar a confiabilidade e os limites das predições ao longo do tempo.

9.6.1 Implicações para a prática clínica

Do ponto de vista clínico, o sinal da dor oferece uma representação contínua e rastreável da probabilidade de dor do recém-nascido, mais compatível com a rotina de monitoramento em UTIN do que classes binárias isoladas. A visualização do sinal permite acompanhar o valor atual, a tendência e o histórico recente, informações que ajudam a distinguir episódios agudos, persistentes ou recorrentes e a contextualizar as intervenções da equipe. Essas características são comumente avaliadas por enfermeiros e médicos, mas não são capturadas por classes pontuais nem pelos modelos tradicionalmente utilizados para a classificação da dor neonatal.

As características dos sinais podem orientar uma política de uso baseada em confiabilidade. Sinais estáveis têm maior potencial de automação, pois se concentram em regiões de maior confiança e apresentam melhor alinhamento probabilístico. Sinais irregulares e instáveis, em contraste, devem ser tratados como indicadores da necessidade de supervisão, e não apenas como falhas do sistema. Nesses casos, o sistema deve sinalizar a fragilidade da evidência e indicar a necessidade de reavaliação clínica dirigida. A presença do padrão indeterminado reforça a necessidade de um quarto desfecho operacional, o inconclusivo. Em vez de impor uma decisão

binária quando o sinal permanece próximo a θ_1, a conduta mais segura é registrar explicitamente a inconclusão e solicitar avaliação humana.

Adicionalmente, a dimensão temporal de XAI acrescenta informação útil para a interpretação e a auditoria do sistema. Ao associar as oscilações do sinal às mudanças nas regiões faciais relevantes, o sistema oferece não apenas um valor numérico, mas também uma justificativa visual para sua variação. A contribuição prática deste capítulo consiste em demonstrar a viabilidade técnica do monitoramento contínuo com níveis graduais de confiabilidade. A segurança e a usabilidade dessa operação em UTIN ainda requerem avaliação prospectiva.

Por fim, da mesma forma que um sinal de frequência cardíaca ou de saturação de oxigênio, o sinal da dor tem potencial para ser integrado a sistemas de monitoramento contínuo, fornecendo alertas e acionando protocolos de intervenção quando necessário. Ao longo da tese, foram desenvolvidos diversos componentes que, juntos, permitiram construir um sinal contínuo, interpretável e acompanhado por informações de confiabilidade e limiares operacionais, com potencial para apoiar a detecção e o manejo da dor neonatal em ambientes clínicos.

10 CONCLUSÃO

A ampliação do uso de sistemas de IA na medicina vem produzindo resultados promissores em diversas áreas. No entanto, a adoção desses sistemas não depende somente de desempenho discriminativo. Um sistema com 99,99% de acurácia não é necessariamente o mais adequado ao uso clínico, tendo em vista o Regulamento de Inteligência Artificial da União Europeia (*EU AI Act*) e a Resolução CFM n. 2.454/2026, que estabelecem requisitos de supervisão humana, confiabilidade, transparência e governança para sistemas de IA em contextos sensíveis (European Union, 2024; Conselho Federal de Medicina, 2026). Um sistema destinado a esse contexto deve ser confiável, explicável e auditável e deve ser capaz de se integrar ao fluxo assistencial, respeitando a autonomia do paciente e a responsabilidade da equipe, princípios centrais da IA responsável.

A avaliação automática da dor neonatal, tema desta tese, materializa esse desafio. A dor é um fenômeno subjetivo, e sua avaliação depende de múltiplos fatores, incluindo a experiência do avaliador, o contexto clínico e as características individuais do paciente. Além disso, seu manejo inadequado pode produzir consequências adversas de curto e longo prazo, incluindo alterações no desenvolvimento neurológico e comportamental do neonato. Nesse contexto, a avaliação manual contínua de todos os RNs internados em UTIN é inviável, o que motiva o desenvolvimento de instrumentos computacionais de apoio ao monitoramento.

O objetivo desta tese foi desenvolver e avaliar um sinal contínuo da dor neonatal baseado na expressão facial, sustentado pelos princípios de uma IA responsável. Para isso, foi construída uma metodologia formal, reprodutível e orientada à segurança operacional para a prática clínica, que integrou modelos discriminativos, quantificação da incerteza, calibração probabilística, explicabilidade e limiares operacionais. Esses componentes culminaram em uma representação temporal que evidencia não apenas a probabilidade estimada de dor, mas também sua estabilidade, sua confiabilidade e as evidências faciais associadas às variações do sinal. A contribuição central da tese é, portanto, o sinal contínuo da dor acompanhado do arcabouço metodológico necessário para sua construção e avaliação. Esse arcabouço é compatível com uma futura arquitetura HITL de apoio à avaliação neonatal, mas sua integração ao fluxo assistencial e a própria interação entre profissionais e sistema permanecem aplicações futuras que requerem validação prospectiva.

Em suma, a tese contribui com mecanismos de confiabilidade, explicabilidade e segurança operacional para sistemas de IA no contexto neonatal. O sinal da dor aproxima as saídas dos modelos das necessidades do monitoramento contínuo, ao mesmo tempo que preserva a supervisão humana e explicita situações de incerteza, ambiguidade ou baixa confiabilidade. No caso dos RNs, que não verbalizam a dor, detectá-la é apenas parte do problema. Também é necessário estabelecer quando as evidências disponíveis são suficientes para sustentar essa avaliação.

A seguir, são discutidos os principais resultados, limitações e perspectivas futuras desta pesquisa.

10.1 REALIZAÇÃO DO OBJETIVO GERAL E DOS OBJETIVOS ESPECÍFICOS

O objetivo geral foi alcançado pela integração progressiva de cinco objetivos específicos. Cada etapa forneceu um componente necessário à construção e à avaliação do sinal contínuo da dor, desde a geração das probabilidades até sua qualificação em termos de confiabilidade, explicabilidade e segurança operacional.

10.1.1 Objetivo específico 1: desenvolver e avaliar modelos de reconhecimento automático da dor neonatal

O primeiro objetivo específico foi atendido no Capítulo 3, com apoio da revisão crítica apresentada no Capítulo 2. Nessa etapa, os modelos NCNN, VGGFace e ViT-B/32 foram treinados e comparados em um protocolo controlado, estabelecendo o ponto de partida discriminativo da tese. Nas condições avaliadas, as três arquiteturas alcançaram desempenho global sem diferença estatisticamente significativa para a discriminação entre dor e não dor, com destaque numérico para a VGGFace na validação cruzada. Ao mesmo tempo, essa etapa evidenciou que acurácia e AUC, isoladamente, são insuficientes para sustentar o uso clínico, pois não informam a confiabilidade probabilística nem a estabilidade operacional das predições.

10.1.2 Objetivo específico 2: investigar incerteza e calibração probabilística

O segundo objetivo foi desenvolvido em dois capítulos complementares. No Capítulo 4, foram avaliados métodos de quantificação da incerteza, e verificou-se que a incerteza epistêmica se associa, em diferentes graus, às predições incorretas e às discrepâncias entre bancos de dados, oferecendo uma camada adicional de segurança para sinalizar predições potencialmente frágeis. No Capítulo 5, verificou-se que a qualidade probabilística das saídas depende da arquitetura e do método empregado, sendo que os modelos NCNN e VGGFace já apresentavam calibração satisfatória dentro dos critérios adotados, enquanto o ViT-B/32 se beneficiou de calibração adicional com a Regressão Isotônica. Em conjunto, essas etapas qualificaram as probabilidades que compõem o sinal e forneceram mecanismos para interpretar sua confiabilidade.

10.1.3 Objetivo específico 3: avaliar a explicabilidade das predições

O terceiro objetivo foi desenvolvido no Capítulo 6, no qual a explicabilidade foi examinada de forma sistemática. Os resultados mostraram que mapas em nível de pixel são instáveis, mas que explicações agregadas em regiões faciais semanticamente definidas oferecem interpretação mais robusta e mais próxima do raciocínio clínico. Essa organização semântica permitiu relacionar as predições a regiões reconhecidas em escalas neonatais e forneceu o pilar de explicabilidade posteriormente incorporado à análise temporal do sinal.

10.1.4 Objetivo específico 4: definir formalmente os limiares operacionais dos modelos

O quarto objetivo foi alcançado no Capítulo 7. Nessa etapa, foram definidos limiares explícitos para decisão (θ_1), precisão empírica (θ_2) e incerteza (θ_3), que distinguem as predições aceitas com maior confiança daquelas rejeitadas ou encaminhadas à revisão. Os resultados mostraram que os ganhos de desempenho obtidos após a aplicação desses limiares decorrem da rejeição seletiva de casos ambíguos, e não de uma melhoria intrínseca do classificador. Esses limiares constituíram o pilar de segurança operacional aplicado posteriormente ao sinal.

10.1.5 Objetivo específico 5: integrar e avaliar o sinal contínuo da dor

O quinto objetivo foi atingido no Capítulo 9, no qual as probabilidades produzidas pelos modelos finais foram transformadas em uma representação temporal contínua e analisadas em conjunto com a incerteza, os limiares operacionais e as explicações. Essa integração substituiu a lógica de classificações pontuais por uma lógica de monitoramento, com padrões estáveis, irregulares, instáveis e, em casos raros, indeterminados. O sinal mostrou que a confiabilidade do sistema varia ao longo do tempo e depende tanto da dinâmica da expressão facial quanto das condições de observação. Também foram examinadas a concordância temporal entre arquiteturas e a associação entre as variações do sinal e a relevância das regiões faciais. A metodologia chega, assim, a uma proposta mais próxima da necessidade de acompanhamento contínuo em UTIN.

10.2 LIMITAÇÕES DA TESE

Apesar das contribuições apresentadas, os resultados devem ser interpretados à luz de limitações importantes.

Em primeiro lugar, os experimentos foram conduzidos em um conjunto relativamente pequeno, composto por 483 imagens de 56 sujeitos, além de vídeos de duração limitada para a etapa do sinal da dor. Ainda que o protocolo experimental tenha buscado rigor e separação por sujeito, a dimensão amostral reduzida limita a generalização estatística dos achados e aumenta a sensibilidade a variações específicas do conjunto analisado.

Em segundo lugar, os dados utilizados não são plenamente representativos da variabilidade encontrada em ambientes reais de UTIN. Diferenças de iluminação, pose, etnia, idade gestacional, dispositivos médicos, oclusões faciais e protocolos locais podem alterar de forma substancial o comportamento dos modelos. A própria análise de incerteza indicou sensibilidade a diferenças entre bases, o que reforça a necessidade de validação multicêntrica.

Em terceiro lugar, a tese não definiu um limiar formal de XAI equivalente aos limiares de precisão e de incerteza. Esse limiar permitiria operacionalizar a explicabilidade como critério de rejeição. Além disso, a análise do desempenho dos modelos em tempo real não explorou técnicas de otimização para execução em dispositivos de baixo custo, como a quantização de pesos.

Por fim, a tese não realizou validação prospectiva em fluxo clínico real nem avaliação de impacto assistencial. Não foi investigado, por exemplo, se o uso do sistema melhora o tempo de

resposta à dor, reduz a variabilidade entre avaliadores ou modifica condutas analgésicas.

10.3 AGENDA DE PESQUISA

Os trabalhos futuros decorrentes desta tese organizam-se em uma agenda voltada à translação clínica:

a) **construção de bases multicêntricas e temporalmente anotadas:** uma das prioridades é ampliar a base empírica com dados de múltiplas UTINs, maior diversidade demográfica e protocolos padronizados de aquisição. Além da expansão amostral, é necessário substituir classes binárias por anotações temporais mais refinadas, capazes de representar intensidade, transições e ambiguidade clínica ao longo do tempo;

b) **modelagem temporal e multimodal da dor neonatal:** a etapa do sinal da dor mostrou o valor da dimensão temporal, mas ainda parte de probabilidades derivadas de *frames* independentes. Estudos futuros devem investigar modelos nativamente temporais, baseados em vídeo, e a fusão multimodal com choro, parâmetros fisiológicos e contexto do procedimento. A integração dessas modalidades pode reduzir ambiguidades faciais e melhorar o desempenho em cenários de oclusão;

c) **explicabilidade orientada por conhecimento clínico:** os resultados de XAI indicam a necessidade de avançar para explicações mais estáveis, semânticas e auditáveis. Isso inclui estudos com regularização por regiões faciais, avaliação sistemática de oclusões clinicamente realistas, explicações baseadas em conceitos e protocolos formais de comparação entre atribuições do modelo e o julgamento de especialistas. Nessa agenda, a explicabilidade deve ser tratada como critério de projeto do modelo, e não como etapa posterior de inspeção;

d) **validação prospectiva e integração ao fluxo assistencial:** a translação exige que o sistema seja avaliado em ambiente clínico real. Isso envolve desenvolver ferramentas para uso à beira do leito, incluindo interfaces de visualização do sinal da dor, e investigar como esses elementos afetam a tomada de decisão da equipe. Estudos prospectivos devem medir o desempenho algorítmico e os efeitos sobre o tempo de resposta, a carga de alarmes, a confiança dos profissionais, a segurança do paciente e o atendimento aos requisitos regulatórios;

e) **governança clínica, documentação e conformidade regulatória:** em consonância com o Estatuto dos Direitos do Paciente e com a Resolução CFM n. 2.454/2026, estudos futuros devem investigar como comunicar o uso da IA ao paciente ou a seus representantes, como registrar a participação do sistema no prontuário, como auditar falhas e padrões explicativos, e como definir responsabilidades entre equipe assistencial, instituição e desenvolvedores. Essa agenda é essencial para transformar o arcabouço técnico da tese em uma solução clinicamente governável.

O arcabouço de IA investigado nesta tese não deve ser entendido como substituto da avaliação clínica, mas como uma camada adicional de apoio à decisão, integrada ao monitoramento contínuo. A Figura 46 representa, de forma conceitual, uma interface futura que reúne sinais fisiológicos convencionais e o sinal de dor em um mesmo ambiente de visualização à beira do leito. Esta tese contribui, portanto, para uma etapa inicial relevante na concretização do sistema ilustrado na figura, que um dia poderá apoiar não só os profissionais de saúde, mas também os pacientes, promovendo maior segurança, transparência e confiança no uso da IA na avaliação da dor neonatal.

Figura 46 – Representação conceitual de um monitor futuro para suporte à avaliação da dor neonatal.



Fonte: Autor, utilizando o modelo de geração de imagens do ChatGPT.

10.3.1 Publicações

Os Apêndices F a J apresentam os artigos publicados durante o desenvolvimento desta tese. Em ordem cronológica, os artigos são:

a) apêndice F: "*Revisiting N-CNN for Clinical Practice*", publicado no workshop *Predictive Intelligence in Medicine* (PRIME) em 2023, em conjunto com a conferência *Medical Image Computing and Computer Assisted Intervention* (MICCAI);

b) apêndice G: "*Disclosing neonatal pain in real-time: AI-derived pain sign from continuous assessment of facial expressions*", publicado na revista *Computers in Biology and Medicine* em 2025;

c) apêndice H: "*Understanding Disagreement Between Humans and Machines in XAI: Robustness, Fidelity, and Region-Based Explanations in Automatic Neonatal Pain*

Assessment", publicado na conferência *World Conference on Explainable Artificial Intelligence* em 2025;

d) apêndice I: "*From Pixels to Pain Scales: Robust Evaluation of Explainable AI in Neonatal Pain Detection*", publicado na conferência *World Conference on Explainable Artificial Intelligence* em 2026;

e) apêndice J: "*Barriers and Pathways to Clinical Translation of AI in Neonatal Pain Assessment*", publicado como capítulo de livro para a coleção Desafios de IA na Saúde, durante o Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS) 2026.

Os artigos publicados e aceitos na *World Conference on Explainable Artificial Intelligence* resultaram de uma parceria internacional entre o Centro Universitário FEI e a Robert Gordon University, em Aberdeen, Escócia, Reino Unido. A colaboração também resultou em uma bolsa de Doutorado Sanduíche no Exterior (PDSE) da CAPES, que possibilitou a realização de parte do trabalho na Robert Gordon University, sob a supervisão do Prof. Dr. Carlos Francisco Moreno-García.

Além dos artigos publicados diretamente dos resultados desta tese, o autor também contribuiu com outros trabalhos relacionados à avaliação da dor neonatal, incluindo artigos publicados em periódicos e conferências internacionais, como Carlini *et al.* (2021), Coutrin *et al.* (2023), Carlini *et al.* (2024) e Carlini *et al.* (2025).

REFERÊNCIAS

ABD-ELSAYED, A. **Pain: A Review Guide**. 1. ed. Cham: Springer, 2019. 1334 p. ISBN 978-3-319-99124-5.

ABDAR, M. *et al.* A review of uncertainty quantification in deep learning: Techniques, applications and challenges. **Information fusion**, Elsevier, v. 76, p. 243–297, 2021. ISSN 1566-2535.

ADEBAYO, J. *et al.* Sanity checks for saliency maps. **Advances in neural information processing systems**, v. 31, 2018.

AMERSHI, S. *et al.* Guidelines for human–ai interaction. *In*: **Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems**. [*S.l.: s.n.*]: Association for Computing Machinery, 2019. p. 1–13.

ANAND, K. J. Defining pain in newborns: need for a uniform taxonomy? **Acta Paediatrica**, Wiley, v. 106, n. 9, p. 1438–1444, 2017. ISSN 0803-5253.

ANAND, K. J. *et al.* Assessment of continuous pain in newborns admitted to nicu's in 18 european countries. **Acta paediatrica**, Wiley Online Library, v. 106, n. 8, p. 1248–1259, 2017.

ANAND, K. J.; HICKEY, P. R. *et al.* Pain and its effects in the human neonate and fetus. **N Engl j Med**, v. 317, n. 21, p. 1321–1329, 1987.

APKARIAN, A. V.; BALIKI, M. N.; GEHA, P. Y. Towards a theory of chronic pain. **Progress in neurobiology**, Elsevier, v. 87, n. 2, p. 81–97, 2009. ISSN 0301-0082.

ARRIETA, A. B. *et al.* Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. **Information fusion**, Elsevier, v. 58, p. 82–115, 2020.

ATTIG, C. *et al.* System latency guidelines then and now: Is zero latency really considered necessary? *In*: **Engineering Psychology and Cognitive Ergonomics: Cognition and Design**. [*S.l.: s.n.*]: Springer International Publishing, 2017. (Lecture Notes in Computer Science, v. 10276), p. 3–14.

BALDA, R. d. C. X. *et al.* The recognition of facial expression of pain in full-term newborns by parents and health professionals. **Archives of pediatrics & adolescent medicine**, American Medical Association, v. 154, n. 10, p. 1009–1016, 2000. ISSN 1072-4710.

BARUA, P. D. *et al.* Automated detection of pain levels using deep feature extraction from shutter blinds-based dynamic-sized horizontal patches with facial images. **Scientific Reports**, Nature Publishing Group UK London, v. 12, n. 1, p. 17297, 2022. ISSN 2045-2322.

BASBAUM, A. I. *et al.* Cellular and molecular mechanisms of pain. **Cell**, Elsevier, v. 139, n. 2, p. 267–284, 2009. ISSN 0092-8674.

BRAHNAM, S. *et al.* Machine assessment of neonatal facial expressions of acute pain. **Decision Support Systems**, Elsevier, v. 43, n. 4, p. 1242–1254, 2007. ISSN 0167-9236.

BRAHNAM, S. *et al.* Machine recognition and representation of neonatal facial displays of acute pain. **Artificial intelligence in medicine**, Elsevier, v. 36, n. 3, p. 211–222, 2006. ISSN 0933-3657.

BRAHNAM, S. *et al.* Neonatal pain detection in videos using the icopevid dataset and an ensemble of descriptors extracted from gaussian of local descriptors. **Applied Computing and Informatics**, Emerald Publishing Limited, v. 19, n. 1/2, p. 122–143, 2020. ISSN 2634-1964.

Brasil. **Lei n. 15.378, de 6 de abril de 2026: Institui o Estatuto dos Direitos do Paciente**. 2026. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2026/lei/l15378.htm. Acesso em: 21 ago. 2026.

BRIER, G. W. Verification of forecasts expressed in terms of probability. **Monthly weather review**, American Meteorological Society, v. 78, n. 1, p. 1–3, 1950. ISSN 0027-0644.

BRÖCKER, J. Reliability, sufficiency, and the decomposition of proper scores. **Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography**, Wiley Online Library, v. 135, n. 643, p. 1512–1519, 2009.

BUCHHOLZ, M. *et al.* Pain scores in infants: a modified infant pain scale versus visual analogue. **Journal of pain and symptom management**, Elsevier, v. 15, n. 2, p. 117–124, 1998. ISSN 0885-3924.

BUCSEA, O. *et al.* Clinical thresholds in pain-related facial activity linked to differences in cortical network activation in neonates. **Pain**, LWW, v. 164, n. 5, p. 1039–1050, 2023. ISSN 0304-3959.

CARBAJAL, R. *et al.* Dan: une échelle comportementale d'évaluation de la douleur aiguë du nouveau-né. **Archives de pédiatrie**, Elsevier, v. 4, n. 7, p. 623–628, 1997. ISSN 0929-693X.

CARLINI, L. P. **Human vs machine towards neonatal pain assessment**: a comparison of the facial features extracted by adults and convolutional neural networks. 2023. 96 p p. Dissertação ((Mestrado Engenharia Elétrica)) — Centro Universitário FEI, São Bernardo do Campo, 2023. Disponível em: https://doi.org/10.31414/EE.2023.D.131608.

CARLINI, L. P. *et al.* Human vs machine towards neonatal pain assessment: A comprehensive analysis of the facial features extracted by health professionals, parents, and convolutional neural networks. **Artificial Intelligence in Medicine**, Elsevier, v. 147, p. 102724, 2024. ISSN 0933-3657.

CARLINI, L. P. *et al.* A convolutional neural network-based mobile application to bedside neonatal pain assessment. *In*: **2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [*S.l.: s.n.*]: IEEE, 2021. p. 394–401.

CARLINI, L. P. *et al.* Is this neonate feeling pain? leveraging clinical knowledge towards high-precision large language model-based neonatal pain assessment. **Pediatric Research**, Nature Publishing Group US New York, p. 1–8, 2025. ISSN 0031-3998.

CARR, D. B.; GOUDAS, L. C. Acute pain. **The Lancet**, Elsevier, v. 353, n. 9169, p. 2051–2058, 1999. ISSN 0140-6736.

CHEFER, H.; GUR, S.; WOLF, L. Transformer interpretability beyond attention visualization. *In*: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [*S.l.: s.n.*]: IEEE, 2021. p. 782–791.

CHEN, H. *et al.* Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. **NPJ digital medicine**, Nature Publishing Group UK London, v. 5, n. 1, p. 156, 2022. ISSN 2398-6352.

CHEN, X. *et al.* Video-based versus on-site neonatal pain assessment in neonatal intensive care units: the impact of video-based neonatal pain assessment in real-world scenario on pain diagnosis and its artificial intelligence application. **Diagnostics**, MDPI, v. 13, n. 16, p. 2661, 2023. ISSN 2075-4418.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 1960.

COHEN, J. **Statistical Power Analysis for the Behavioral Sciences**. 2. ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988. 567 p.

COLLINS, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. **BMJ**, v. 385, p. e078378, 2024.

Conselho Federal de Medicina. **Resolução CFM n. 2.454, de 11 de fevereiro de 2026: Normatiza o uso da inteligência artificial na medicina**. 2026. Disponível em: https://www.in.gov.br/web/dou/-/resolucao-cfm-n-2.454-de-11-de-fevereiro-de-2026-689247948. Acesso em: 17 jun. 2026.

COUTRIN, G. A. *et al.* Convolutional neural networks for newborn pain assessment using face images: A quantitative and qualitative comparison. *In*: SU, R. *et al.* (ed.). **International Conference on Medical Imaging and Computer-Aided Diagnosis**. Singapore, 2023. p. 503–513.

CRUZ, M. D.; FERNANDES, A.; OLIVEIRA, C. Epidemiology of painful procedures performed in neonates: a systematic review of observational studies. **European Journal of Pain**, Wiley Online Library, v. 20, n. 4, p. 489–498, 2016. ISSN 1090-3801.

DAVIS, J. *et al.* **Quantifying uncertainty in deep learning systems**. Amazon Web Services, 2020. AWS Prescriptive Guidance. Acesso em: 23 fev. 2026. Disponível em: https://d1.awsstatic.com/APG/quantifying-uncertainty-in-deep-learning-systems.pdf.

DEGROOT, M. H.; FIENBERG, S. E. The comparison and evaluation of forecasters. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 32, n. 1-2, p. 12–22, 1983.

DOLEZAL, J. M. *et al.* Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. **Nature communications**, Nature Publishing Group UK London, v. 13, n. 1, p. 6572, 2022. ISSN 2041-1723.

DOSOVITSKIY, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020.

EGEDE, J. *et al.* Automatic neonatal pain estimation: An acute pain in neonates database. *In*: IEEE. **2019 8th international conference on affective computing and intelligent interaction (ACII)**. [*S.l.: s.n.*]: IEEE, 2019. p. 1–7.

European Union. **Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence**. 2024. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Acesso em: 21 ago. 2026.

EVANS, J. D. **Straightforward Statistics for the Behavioral Sciences**. [*S.l.: s.n.*]: Brooks/Cole Publishing Company, 1996. 600 p. ISBN 9780534231002.

FERREIRA, L. A. *et al.* Disclosing neonatal pain in real-time: Ai-derived pain sign from continuous assessment of facial expressions. **Computers in Biology and Medicine**, Elsevier, v. 189, p. 109908, 2025. ISSN 0010-4825.

FERREIRA, L. A. *et al.* Revisiting ncnn for clinical practice. *In*: REKIK, I. *et al.* (ed.). **International Workshop on PRedictive Intelligence In MEdicine**. Cham, 2023. p. 231–240.

FERREIRA, L. A. *et al.* From pixels to pain scales: Robust evaluation of explainable ai in neonatal pain detection. *In*: JUNIOR, S. B. *et al.* (ed.). **Explainable Artificial Intelligence**. Cham: Springer Nature Switzerland, 2027. p. 521–544.

FIELDS, H. State-dependent opioid control of pain. **Nature Reviews Neuroscience**, Nature Publishing Group UK London, v. 5, n. 7, p. 565–575, 2004. ISSN 1471-003X.

FILHO, N. M.; SILVA, G. A. Invalidez por dor nas costas entre segurados da previdência social do brasil. **Revista de Saúde Pública**, SciELO Public Health, v. 45, n. 3, p. 494–502, 2011. ISSN 0034-8910.

FILHO, T. S. *et al.* Classifier calibration: a survey on how to assess and improve predicted class probabilities. **Machine Learning**, Springer, v. 112, n. 9, p. 3211–3260, 2023.

FITZGERALD, M.; WALKER, S. M. Infant pain management: a developmental neurobiological approach. **Nature Clinical Practice Neurology**, Nature Publishing Group UK London, v. 5, n. 1, p. 35–50, 2009. ISSN 1745-834X.

FOURNIER-CHARRIèRE, E. *et al.* Evendol, a new behavioral pain scale for children ages 0 to 7 years in the emergency department: design and validation. **Pain**, Elsevier, v. 153, n. 8, p. 1573–1582, 2012. ISSN 0304-3959.

FRANCK, L. S.; MIASKOWSKI, C. Measurement of neonatal responses to painful stimuli: a research review. **Journal of Pain and Symptom Management**, Elsevier, v. 14, n. 6, p. 343–378, 1997. ISSN 0885-3924.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the american statistical association**, Taylor & Francis, v. 32, n. 200, p. 675–701, 1937. ISSN 0162-1459.

GAL, Y.; GHAHRAMANI, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *In*: BALCAN, M. F.; WEINBERGER, K. Q. (ed.). **Proceedings of The 33rd International Conference on Machine Learning**. New York, New York, USA: PMLR, 2016. (Proceedings of Machine Learning Research, v. 48), p. 1050–1059.

GAWLIKOWSKI, J. *et al.* A survey of uncertainty in deep neural networks. **Artificial Intelligence Review**, Springer, v. 56, n. S1, p. 1513–1589, 2023. ISSN 0269-2821.

GKIKAS, S.; TSIKNAKIS, M. Automatic assessment of pain based on deep learning methods: A systematic review. **Computer Methods and Programs in Biomedicine**, Elsevier, v. 231, p. 107365, 2023. ISSN 0169-2607.

GODDARD, K.; ROUDSARI, A.; WYATT, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. **Journal of the American Medical Informatics Association**, v. 19, n. 1, p. 121–127, 2012.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [*S.l.: s.n.*]: MIT Press, 2016. 800 p. http://www.deeplearningbook.org.

GOODMAN, B.; FLAXMAN, S. European union regulations on algorithmic decision making and a "right to explanation". **AI Magazine**, Wiley, v. 38, n. 3, p. 50–57, 2017. ISSN 0738-4602.

GRUNAU, R. E. Neonatal pain in very preterm infants: long-term effects on brain, neurodevelopment and pain reactivity. **Rambam Maimonides medical journal**, Rambam Health Care Campus, v. 4, n. 4, 2013.

GRUNAU, R. E.; HOLSTI, L.; PETERS, J. W. Long-term consequences of pain in human neonates. *In*: ELSEVIER. **Seminars in Fetal and Neonatal Medicine**. [*S.l.: s.n.*]: Elsevier BV, 2006. v. 11, n. 4, p. 268–275. ISSN 1744-165X.

GRUNAU, R. V.; CRAIG, K. D. Pain expression in neonates: facial action and cry. **Pain**, Elsevier, v. 28, n. 3, p. 395–410, 1987. ISSN 0304-3959.

GUINSBURG, R. Arte e ciência na UTI neonatal: avaliação e tratamento da dor. **Revista Brasileira de Psicanálise**, v. 45, n. 4, p. 109–116, 2011. ISSN 0486-641X.

GUINSBURG, R. *et al.* Differences in pain expression between male and female newborn infants. **Pain**, Elsevier, v. 85, n. 1-2, p. 127–133, 2000. ISSN 0304-3959.

GUO, C. *et al.* On calibration of modern neural networks. *In*: PRECUP, D.; TEH, Y. W. (ed.). **Proceedings of the 34th International Conference on Machine Learning**. [*S.l.: s.n.*]: PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 1321–1330.

GUO, J. *et al.* Sample and computation redistribution for efficient face detection. **arXiv preprint arXiv:2105.04714**, 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**: Data mining, inference, and prediction. 2. ed. New York, NY: Springer, 2009. 745 p. (Springer Series in Statistics). ISBN 978-0-387-84858-7.

HATFIELD, L. A. Neonatal pain: What's age got to do with it? **Surgical neurology international**, Scientific Scholar, v. 5, n. Suppl 13, p. S479–S489, 2014. ISSN 2152-7806.

HEIDERICH, T. M. *et al.* Face-based automatic pain assessment: challenges and perspectives in neonatal intensive care units. **Jornal de Pediatria**, Elsevier BV, v. 99, n. 6, p. 546–560, 2023. ISSN 0021-7557.

HEIDERICH, T. M.; LESLIE, A. T. F. S.; GUINSBURG, R. Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. **Acta Paediatrica**, Wiley Online Library, v. 104, n. 2, p. e63–e69, 2015. ISSN 0803-5253.

HOLSTI, L.; GRUNAU, R. E. Initial validation of the behavioral indicators of infant pain (biip). **Pain**, Elsevier, v. 132, n. 3, p. 264–272, 2007. ISSN 0304-3959.

HOLZINGER, A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? **Brain Informatics**, v. 3, n. 2, p. 119–131, 2016.

HORGAN, M.; CHOONARA, I. Measuring pain in neonates: an objective score. **Paediatric nursing**, v. 8, n. 10, p. 24–27, 1996.

HOSSAIN, M. I. *et al.* Explainable ai for medical data: current methods, limitations, and future directions. **ACM Computing Surveys**, ACM New York, NY, v. 57, n. 6, p. 1–46, 2025. ISSN 0360-0300.

HÜLLERMEIER, E.; WAEGEMAN, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. **Machine Learning**, Springer, v. 110, n. 3, p. 457–506, 2021. ISSN 0885-6125.

HUXLEY, T. H. On the reception of the 'origin of species'. *In*: DARWIN, F. (ed.). **The Life and Letters of Charles Darwin**. New York: D. Appleton and Company, 1887. v. 1, p. 533–558. Disponível em: https://aleph0.clarku.edu/huxley/Book/Recep.html. Acesso em: 9 set. 2026.

IASP, I. Subcommittee on taxonomy. pain terms. a list with definitions and notes on usage. **Pain**, v. 6, n. 3, p. 249–52, 1979.

JIANG, X. *et al.* Calibrating predictive model estimates to support personalized medicine. **Journal of the American Medical Informatics Association**, BMJ Group, v. 19, n. 2, p. 263–274, 2012. ISSN 1067-5027.

KARANDIKAR, A. *et al.* Soft calibration objectives for neural networks. **Advances in Neural Information Processing Systems**, v. 34, p. 29768–29779, 2021.

KHAN, S. *et al.* Transformers in vision: A survey. **ACM computing surveys (CSUR)**, ACM New York, NY, v. 54, n. 10s, p. 1–41, 2022. ISSN 0360-0300.

KIM, Y.-T. *et al.* Defining operational safety in clinical artificial intelligence systems. **npj Digital Medicine**, v. 9, n. 1, p. 281, 2026.

KOLMOGOROV, A. N. Sulla determinazione empirica di una legge di distribuzione. **Giornale dell'Istituto Italiano degli Attuari**, v. 4, p. 83–91, 1933.

KOMPA, B.; SNOEK, J.; BEAM, A. L. Second opinion needed: communicating uncertainty in medical machine learning. **NPJ Digital Medicine**, Nature Publishing Group UK London, v. 4, n. 1, p. 4, 2021. ISSN 2398-6352.

KRISHNA, S. *et al.* The disagreement problem in explainable machine learning: A practitioner's perspective. **arXiv preprint arXiv:2202.01602**, 2022.

KULESHOV, V.; LIANG, P. S. Calibrated structured prediction. **Advances in Neural Information Processing Systems**, v. 28, 2015.

LAKSHMINARAYANAN, B.; PRITZEL, A.; BLUNDELL, C. Simple and scalable predictive uncertainty estimation using deep ensembles. **Advances in neural information processing systems**, v. 30, 2017.

LAMBERT, B. *et al.* Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. **Artificial Intelligence in Medicine**, Elsevier, v. 150, p. 102830, 2024. ISSN 0933-3657.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159–174, 1977.

LECUN, Y. *et al.* Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Ieee, v. 86, n. 11, p. 2278–2324, 1998. ISSN 0018-9219.

LEKADIR, K. *et al.* FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. **BMJ**, v. 388, p. e081554, 2025.

LIAW, J.-J. *et al.* Psychometric analysis of a taiwan-version pain assessment scale for preterm infants. **Journal of clinical nursing**, Wiley Online Library, v. 21, n. 1-2, p. 89–100, 2012. ISSN 0962-1067.

LIPTON, Z. C. The mythos of model interpretability. **Communications of the ACM**, Association for Computing Machinery, v. 61, n. 10, p. 36–43, 2018. ISSN 0001-0782.

LITJENS, G. *et al.* A survey on deep learning in medical image analysis. **Medical image analysis**, Elsevier, v. 42, p. 60–88, 2017. ISSN 1361-8415.

LLERENA, A. *et al.* Neonatal pain assessment: Do we have the right tools? **Frontiers in Pediatrics**, Frontiers Media SA, v. 10, p. 1022751, 2023. ISSN 2296-2360.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

MACKAY, D. J. A practical bayesian framework for backpropagation networks. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals, v. 4, n. 3, p. 448–472, 1992. ISSN 0899-7667.

MANWORREN, R. C. *et al.* Performance evaluation of a supervised machine learning pain classification model developed by neonatal nurses. **Advances in Neonatal Care**, LWW, v. 24, n. 3, p. 301–310, 2024. ISSN 1536-0903.

MARCEAU, J. Pilot study of a pain assessment tool in the neonatal intensive care unit. **Journal of paediatrics and child health**, Wiley Online Library, v. 39, n. 8, p. 598–601, 2003. ISSN 1034-4810.

MCGRATH, P. A. Multidimensional pain assessment in children. **Advances in pain research and therapy**, Raven Press, p. 387–393, 1985.

MCGRATH, P. A. **Pain in children : nature, assessment, and treatment**. Guilford Press, 1990. Disponível em: https://cir.nii.ac.jp/crid/1130000797702541696.

MERSKEY, H. Pain terms: a list with definitions and notes on usage. recommended by the iasp subcommittee on taxonomy. **Pain**, v. 6, p. 249–252, 1979.

MILLAN, M. J. The induction of pain: an integrative review. **Progress in neurobiology**, Elsevier, v. 57, n. 1, p. 1–164, 1999. ISSN 0301-0082.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. **Artificial Intelligence**, Elsevier BV, v. 267, p. 1–38, 2019. ISSN 0004-3702.

MÜLLER, R.; KORNBLITH, S.; HINTON, G. E. When does label smoothing help? **Advances in neural information processing systems**, v. 32, 2019.

MURPHY, A. H.; EPSTEIN, E. S. Verification of probabilistic predictions: A brief review. **Journal of Applied Meteorology and Climatology**, American Meteorological Society, v. 6, n. 5, p. 748–755, 1967. ISSN 0021-8952.

NAEINI, M. P.; COOPER, G.; HAUSKRECHT, M. Obtaining well calibrated probabilities using bayesian binning. *In*: **Proceedings of the AAAI conference on artificial intelligence**. [*S.l.: s.n.*], 2015. v. 29, n. 1.

NAH, F. F.-H. A study on tolerable waiting time: How long are web users willing to wait? **Behaviour & Information Technology**, v. 23, n. 3, p. 153–163, 2004.

NICULESCU-MIZIL, A.; CARUANA, R. Predicting good probabilities with supervised learning. *In*: **Proceedings of the 22nd International Conference on Machine Learning**. New York, NY, USA: Association for Computing Machinery, 2005. (ICML '05), p. 625–632.

NIXON, J. *et al.* Measuring calibration in deep learning. *In*: **CVPR workshops**. [*S.l.: s.n.*], 2019. v. 2, n. 7.

PARKHI, O.; VEDALDI, A.; ZISSERMAN, A. Deep face recognition. *In*: BRITISH MACHINE VISION ASSOCIATION. **BMVC 2015-Proceedings of the British Machine Vision Conference 2015**. [*S.l.: s.n.*]: British Machine Vision Association, 2015. p. 41.1–41.12.

PASZKE, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. **Advances in neural information processing systems**, v. 32, 2019.

PAUL, S.; CHEN, P.-Y. Vision transformers are robust learners. *In*: **Proceedings of the AAAI conference on Artificial Intelligence**. [*S.l.: s.n.*]: Association for the Advancement of Artificial Intelligence (AAAI), 2022. v. 36, n. 2, p. 2071–2081. ISSN 2374-3468.

PEREIRA, A. L. d. S. T. *et al.* Validity of behavioral and physiologic parameters for acute pain assessment of term newborn infants. **Sao Paulo Medical Journal**, v. 117, n. 2, p. 72–80, 1999. ISSN 1516-3180.

PERRY, M. *et al.* Neonatal pain: perceptions and current practice. **Critical Care Nursing Clinics**, Elsevier, v. 30, n. 4, p. 549–561, 2018.

PIRIE, C. *et al.* Understanding disagreement between humans and machines in xai: Robustness, fidelity, and region-based explanations in automatic neonatal pain assessment. *In*: GUIDOTTI, R.; SCHMID, U.; LONGO, L. (ed.). **Explainable Artificial Intelligence**. Cham: Springer Nature Switzerland, 2026. p. 274–298.

PLATT, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. **Advances in Large Margin Classifiers**, p. 61–74, 1999.

PRESTES, A. C. Y. *et al.* Painful procedures and analgesia in the NICU: what has changed in the medical perception and practice in a ten-year period? **Jornal de Pediatria**, Elsevier, v. 92, n. 1, p. 88–95, 2016. ISSN 0021-7557.

PRKACHIN, K. M. *et al.* Does experience influence judgements of pain behaviour? evidence from relatives of pain patients and therapists. **Pain Research and Management**, Hindawi, v. 6, n. 2, p. 105–112, 2001. ISSN 1203-6765.

RAJA, S. N. *et al.* The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. **PAIN**, v. 161, n. 9, p. 1976–1982, 2020.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. *In*: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**. [*S.l.: s.n.*], 2016. p. 1135–1144.

RICHTER, L. M. *et al.* Investing in the foundation of sustainable development: pathways to scale up for early childhood development. **The Lancet**, Elsevier BV, v. 389, n. 10064, p. 103–118, 2017. ISSN 0140-6736.

RODKEY, E. N.; RIDDELL, R. P. The infancy of infant pain research: the experimental origins of infant pain denial. **The Journal of Pain**, Elsevier, v. 14, n. 4, p. 338–350, 2013. ISSN 1526-5900.

RONG, Y. *et al.* A consistent and efficient evaluation strategy for attribution methods. **arXiv preprint arXiv:2202.00449**, 2022.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987. ISSN 0377-0427.

RUSHFORTH, J. A.; LEVENE, M. I. Behavioural response to pain in healthy neonates. **Archives of Disease in Childhood-Fetal and Neonatal Edition**, BMJ Publishing Group, v. 70, n. 3, p. F174–F176, 1994. ISSN 1359-2998.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3. ed. [*S.l.: s.n.*]: Prentice Hall, 2009. ISBN 978-0136042594.

SALEKIN, M. S. *et al.* First investigation into the use of deep learning for continuous assessment of neonatal postoperative pain. *In*: IEEE. **2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)**. [*S.l.: s.n.*]: IEEE, 2020. p. 415–419.

SALEKIN, M. S. *et al.* Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. **Computers in biology and medicine**, Elsevier, v. 129, p. 104150, 2021. ISSN 0010-4825.

SAMEK, W. *et al.* Evaluating the visualization of what a deep neural network has learned. **IEEE transactions on neural networks and learning systems**, IEEE, v. 28, n. 11, p. 2660–2673, 2017. ISSN 2162-237X.

SARIO, G. D. D. *et al.* Using ai to detect pain through facial expressions: A review. **Bioengineering**, MDPI, v. 10, n. 5, p. 548, 2023. ISSN 2306-5354.

SCHIAVENATO, M. *et al.* Neonatal pain facial expression: Evaluating the primal face of pain. **Pain**, Elsevier, v. 138, n. 2, p. 460–471, 2008. ISSN 0304-3959.

SELVARAJU, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. *In*: **Proceedings of the IEEE international conference on computer vision**. [*S.l.: s.n.*]: IEEE, 2017. p. 618–626.

SEONI, S. *et al.* Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023). **Computers in Biology and Medicine**, v. 165, p. 107441, 2023.

SHANNON, C. E. A mathematical theory of communication. **The Bell system technical journal**, Nokia Bell Labs, v. 27, n. 3, p. 379–423, 1948.

SHNEIDERMAN, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. **International Journal of Human–Computer Interaction**, v. 36, n. 6, p. 495–504, 2020.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. *In*: PMLR. **International conference on machine learning**. [*S.l.: s.n.*], 2017. p. 3145–3153.

SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. **arXiv preprint arXiv:1312.6034**, 2013.

SMIRNOV, N. Table for estimating the goodness of fit of empirical distributions. **The annals of mathematical statistics**, Institute of Mathematical Statistics, v. 19, n. 2, p. 279–281, 1948. ISSN 0003-4851.

SOUZA, J. B. d. *et al.* Prevalence of chronic pain, treatments, perception, and interference on life activities: Brazilian population-based survey. **Pain Research and Management**, Wiley Online Library, v. 2017, n. 1, p. 4643830, 2017. ISSN 1203-6765.

STEVENS, B. *et al.* Premature infant pain profile: development and initial validation. **The Clinical journal of pain**, LWW, v. 12, n. 1, p. 13–22, 1996. ISSN 0749-8047.

STEVENS, B. J. *et al.* The premature infant pain profile-revised (pipp-r): initial validation and feasibility. **The Clinical journal of pain**, LWW, v. 30, n. 3, p. 238–243, 2014.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. *In*: PMLR. **International conference on machine learning**. [*S.l.: s.n.*], 2017. p. 3319–3328.

SWAFFORD, L. I.; ALLAN, D. Pain relief in the pediatric patient. **Medical Clinics of North America**, Elsevier, v. 52, n. 1, p. 131–136, 1968. ISSN 0025-7125.

SZAMEITAT, A. J. *et al.* Behavioral and emotional consequences of brief delays in human–computer interaction. **International Journal of Human-Computer Studies**, v. 67, n. 7, p. 561–570, 2009.

SZEGEDY, C. *et al.* Rethinking the inception architecture for computer vision. *In*: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S.l.: s.n.*]: IEEE, 2016. p. 2818–2826.

TAMANAKA, F. G. *et al.* Neonatal pain assessment: A kendall analysis between clinical and visually perceived facial features. **Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization**, Taylor & Francis, v. 11, n. 3, p. 331–340, 2023. ISSN 2168-1163.

TOUVRON, H. *et al.* Training data-efficient image transformers & distillation through attention. *In*: PMLR. **International conference on machine learning**. [*S.l.: s.n.*], 2021. p. 10347–10357.

TRACEY, I.; MANTYH, P. W. The cerebral signature for pain perception and its modulation. **Neuron**, Elsevier, v. 55, n. 3, p. 377–391, 2007. ISSN 0896-6273.

TREEDE, R.-D. *et al.* A classification of chronic pain for icd-11. **Pain**, LWW, v. 156, n. 6, p. 1003–1007, 2015.

VASWANI, A. *et al.* Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

VYAS, N.; SAXENA, S.; VOICE, T. Learning soft labels via meta learning. **arXiv preprint arXiv:2009.09496**, 2020.

WALKER, S. M. Pain in children: recent advances and ongoing challenges. **British journal of anaesthesia**, Oxford University Press, v. 101, n. 1, p. 101–110, 2008. ISSN 0007-0912.

WILCOXON, F. Individual comparisons by ranking methods. **Biometrics Bulletin**, International Biometric Society, v. 1, n. 6, p. 80–83, 1945. ISSN 0099-4987.

WILLIAMS, A. C. d. C.; CRAIG, K. D. Updating the definition of pain. **Pain**, LWW, v. 157, n. 11, p. 2420–2423, 2016. ISSN 0304-3959.

YANG, N. *et al.* Feasibility and effectiveness of a low-code ai platform for developing a neonatal multimodal pain classification model. **Journal of Multidisciplinary Healthcare**, Taylor & Francis, v. 18, p. 5771–5780, 2025. ISSN 1178-2390.

YEH, C.-K. *et al.* On the (in) fidelity and sensitivity of explanations. **Advances in neural information processing systems**, v. 32, 2019.

YOON, J. S. *et al.* Domain generalization for medical image analysis: a review. **Proceedings of the IEEE**, IEEE, v. 112, n. 10, p. 1583–1609, 2024. ISSN 0018-9219.

YOUDEN, W. J. Index for rating diagnostic tests. **Cancer**, Wiley Online Library, v. 3, n. 1, p. 32–35, 1950. ISSN 0008-543X.

ZADROZNY, B.; ELKAN, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *In*: **Icml**. [*S.l.: s.n.*], 2001. v. 1, p. 609–616.

ZADROZNY, B.; ELKAN, C. Transforming classifier scores into accurate multiclass probability estimates. *In*: **Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining**. [*S.l.: s.n.*]: ACM, 2002. p. 694–699.

ZAMZMI, G. *et al.* Neonatal pain expression recognition using transfer learning. **arXiv preprint arXiv:1807.01631**, 2018.

ZAMZMI, G. *et al.* A review of automated pain assessment in infants: features, classification tasks, and databases. **IEEE reviews in biomedical engineering**, IEEE, v. 11, p. 77–96, 2018. ISSN 1937-3333.

ZAMZMI, G. *et al.* Pain assessment from facial expression: Neonatal convolutional neural network (n-cnn). *In*: IEEE. **2019 International Joint Conference on Neural Networks (IJCNN)**. [*S.l.: s.n.*]: IEEE, 2019. p. 1–7.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. *In*: SPRINGER. **European conference on computer vision**. [*S.l.: s.n.*], 2014. p. 818–833.

ZHANG, H. *et al.* mixup: Beyond empirical risk minimization. **arXiv preprint arXiv:1710.09412**, 2017.

ZHAO, Y. *et al.* Pose-invariant and occlusion-robust neonatal facial pain assessment. **Computers in Biology and Medicine**, Elsevier BV, v. 165, p. 107462, 2023. ISSN 0010-4825.

APÊNDICES

APÊNDICE A – RESULTADOS AUXILIARES DA QUANTIFICAÇÃO DA INCERTEZA

Este apêndice reúne resultados complementares da quantificação da incerteza descrita no Capítulo 4. São apresentadas, de forma consolidada, as métricas de classificação obtidas com diferentes configurações de Monte Carlo Dropout (MCDP) e Deep Ensembles, além de exemplos qualitativos do comportamento dos modelos em amostras com maior e menor incerteza epistêmica.

A.1 MÉTRICAS DE CLASSIFICAÇÃO SOB DIFERENTES CONFIGURAÇÕES DE INCERTEZA

As Tabelas 33, 34 e 35 apresentam as métricas de classificação da NCNN, da VGGFace e do ViT-B/32 sob diferentes configurações de MCDP e Deep Ensembles. Os símbolos indicam melhora (↑), piora (↓), melhora estatisticamente significativa (↑*), piora estatisticamente significativa (↓*) e resultado inalterado (=). Como cada configuração aprimorada é comparada de forma pareada com o *baseline* da mesma arquitetura, o asterisco indica $p < 0{,}05$ no teste de Wilcoxon. Comparações simultâneas entre as arquiteturas são realizadas com o teste de Friedman, conforme descrito no corpo da tese.

Tabela 33 – Métricas de classificação da NCNN sob diferentes configurações de MCDP e Deep Ensembles (média ± DP, em %).

	MCDP (prob. de dropout p_{drop})			**Deep Ensemble** (nº de redes k_{ens})		
Métrica	p_{drop} = 10%	p_{drop} = 30%	p_{drop} = 50%	k_{ens} = 3	k_{ens} = 5	k_{ens} = 10
Acurácia	79,6 ± 8,6 ↓	79,5 ± 8,7 ↓	79,3 ± 8,8 ↓	**82,1 ± 7,6** ↑	81,0 ± 7,9 ↑	81,2 ± 7,9 ↑
F1-Score	75,4 ± 11,2 ↓	74,8 ± 11,5 ↓	73,9 ± 11,2 ↓	**79,0 ± 7,8** ↑	77,6 ± 9,3 ↑	77,1 ± 9,0 ↑
Sensibilidade	71,7 ± 16,5 ↓	70,2 ± 17,2 ↓	67,2 ± 15,2 ↓	**73,7 ± 8,3** ↑	73,4 ± 10,3 ↑	70,6 ± 10,8 ↓
Especificidade	86,6 ± 7,6 ↓	87,7 ± 6,6 ↑	89,1 ± 7,3 ↑	88,8 ± 11,6 ↑	87,4 ± 10,4 ↑	**89,3 ± 10,7** ↑
Precisão	82,2 ± 10,6 =	83,0 ± 10,1 ↑	84,3 ± 11,2 ↑*	86,1 ± 10,9 ↑	83,9 ± 12,9 ↑	**86,3 ± 11,3** ↑
AUC	88,4 ± 5,8 ↓*	88,3 ± 6,4 ↓	88,1 ± 6,4 ↓	88,1 ± 7,8 ↓	88,3 ± 6,9 ↓	**88,5 ± 6,7** ↓

Fonte: Autor.

Tabela 34 – Métricas de classificação da VGGFace sob diferentes configurações de MCDP e Deep Ensembles (média ± DP, em %).

	MCDP (prob. de dropout p_{drop})			**Deep Ensemble** (nº de redes k_{ens})		
Métrica	p_{drop} = 10%	p_{drop} = 30%	p_{drop} = 50%	k_{ens} = 3	k_{ens} = 5	k_{ens} = 10
Acurácia	84,5 ± 4,4 ↑	84,3 ± 4,5 =	84,1 ± 5,0 ↓	83,7 ± 4,4 ↓	**84,8 ± 3,1** ↑	84,5 ± 2,5 ↑
F1-Score	82,1 ± 5,7 ↑	81,9 ± 5,7 =	81,5 ± 6,8 ↓	81,2 ± 5,7 ↓	**82,2 ± 4,8** ↑	81,8 ± 4,2 ↓
Sensibilidade	**80,8 ± 8,8** =	80,4 ± 8,8 ↓	80,2 ± 10,3 ↓	78,8 ± 9,2 ↓	79,2 ± 9,6 ↓	78,3 ± 8,3 ↓
Especificidade	87,4 ± 8,6 ↑	87,4 ± 8,6 ↑	87,0 ± 8,5 =	87,4 ± 7,3 ↑	89,1 ± 7,2 ↑	**89,4 ± 7,3** ↑
Precisão	84,7 ± 8,9 ↑	84,6 ± 8,9 ↑	84,1 ± 8,8 ↓	84,5 ± 7,7 ↑	86,6 ± 7,2 ↑	**86,8 ± 7,6** ↑
AUC	91,6 ± 3,3 =	91,6 ± 3,2 =	91,9 ± 3,4 ↑*	91,6 ± 3,1 ↑	91,9 ± 2,9 ↑	**92,0 ± 3,3** ↑

Fonte: Autor.

Tabela 35 – Métricas de classificação do ViT-B/32 sob diferentes configurações de MCDP e Deep Ensembles (média ± DP, em %).

	MCDP (prob. de dropout p_{drop})			**Deep Ensemble** (nº de redes k_{ens})		
Métrica	$p_{\text{drop}} = 10\%$	$p_{\text{drop}} = 30\%$	$p_{\text{drop}} = 50\%$	$k_{\text{ens}} = 3$	$k_{\text{ens}} = 5$	$k_{\text{ens}} = 10$
Acurácia	76,9 ± 8,0 $\downarrow^*$	46,0 ± 7,1 $\downarrow^*$	46,1 ± 5,2 $\downarrow^*$	82,5 ± 6,8 $\uparrow$	**82,8 ± 6,6** $\uparrow$	**82,8 ± 6,7** $\uparrow$
F1-Score	68,8 ± 11,7 $\downarrow^*$	62,6 ± 6,7 $\downarrow^*$	52,8 ± 10,8 $\downarrow^*$	79,4 ± 7,5 $\uparrow$	79,8 ± 7,2 $\uparrow$	**79,9 ± 7,4** $\uparrow$
Sensibilidade	58,7 ± 17,2 $\downarrow^*$	**100,0 ± 0,0** $\uparrow^*$	70,7 ± 21,9 $\downarrow$	75,2 ± 10,7 $\uparrow$	75,5 ± 10,4 $\uparrow$	75,9 ± 10,2 $\uparrow$
Especificidade	**91,8 ± 13,0** $\uparrow$	0,7 ± 2,1 $\downarrow^*$	26,5 ± 16,5 $\downarrow^*$	88,6 ± 6,8 $\uparrow$	88,9 ± 6,7 $\uparrow$	88,6 ± 6,8 $\uparrow$
Precisão	**90,2 ± 11,5** $\uparrow$	45,9 ± 7,0 $\downarrow^*$	43,8 ± 5,7 $\downarrow^*$	84,9 ± 6,7 $\uparrow$	85,3 ± 6,4 $\uparrow$	85,0 ± 6,8 $\uparrow$
AUC	86,3 ± 6,1 $\downarrow^*$	53,6 ± 10,3 $\downarrow^*$	48,8 ± 4,1 $\downarrow^*$	89,5 ± 5,6 $\downarrow$	89,6 ± 5,3 $\downarrow$	**89,9 ± 5,0** $\downarrow$

Fonte: Autor.

A.2 EXEMPLOS QUALITATIVOS DE PREDIÇÕES COM ALTA E BAIXA INCERTEZA

As Figuras 47 a 52 mostram exemplos de predições com alta e baixa incerteza para cada modelo, utilizando os métodos selecionados no Capítulo 4. O canto superior esquerdo apresenta a probabilidade predita e a incerteza epistêmica. No topo, indicam-se o banco de origem e a classe verdadeira. A borda verde corresponde a um acerto do modelo, e a vermelha, a um erro.

Figura 47 – Exemplos de predições com alta incerteza da NCNN.



(a) Com dor.



(b) Sem dor.

Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

Figura 48 – Exemplos de predições com baixa incerteza da NCNN.



(a) Com dor.



(b) Sem dor.

Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

Figura 49 – Exemplos de predições com alta incerteza da VGGFace.



(a) Com dor.



(b) Sem dor.

Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

Figura 50 – Exemplos de predições com baixa incerteza da VGGFace.



(a) Com dor.



(b) Sem dor.

Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

Figura 51 – Exemplos de predições com alta incerteza do ViT-B/32.



(a) Com dor.



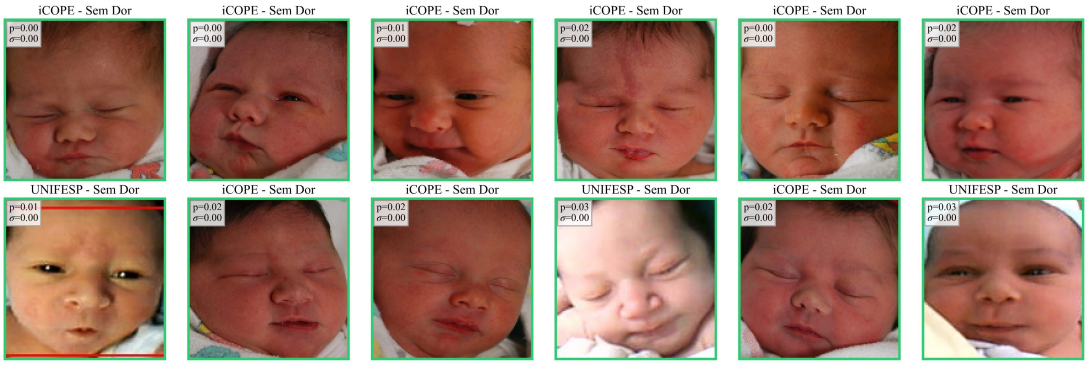
(b) Sem dor.

Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

Figura 52 – Exemplos de predições com baixa incerteza do ViT-B/32.



(a) Com dor.



(b) Sem dor.

Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

APÊNDICE B – RESULTADOS AUXILIARES DE CALIBRAÇÃO

Este apêndice apresenta resultados complementares da calibração probabilística discutida no Capítulo 5. As métricas discriminativas e de calibração permitem comparar os modelos sem calibrador com suas versões calibradas.

Em todas as tabelas deste apêndice, os símbolos indicam melhora (↑), piora (↓), melhora estatisticamente significativa (↑*), piora estatisticamente significativa (↓*) e resultado inalterado (=). Como cada configuração aprimorada é comparada de forma pareada com o *baseline* da mesma arquitetura, o asterisco indica $p < 0{,}05$ no teste de Wilcoxon. Comparações simultâneas entre as arquiteturas são realizadas com o teste de Friedman, conforme descrito no corpo da tese. Para as métricas de calibração ECE, MCE, NLL e Brier, valores menores representam melhor desempenho. Portanto, reduções são indicadas por ↑ e aumentos por ↓.

As Tabelas 36 a 41 apresentam, respectivamente, as métricas de classificação e de calibração da NCNN, da VGGFace e do ViT-B/32.

Tabela 36 – Métricas de classificação da NCNN (média ± DP, em %).

Modelo	**Acurácia**	**F1-Score**	**Sensibilidade**	**Especificidade**	**Precisão**	**AUC**
NCNN	79,5 ± 8,7	74,8 ± 11,5	70,2 ± 17,2	87,7 ± 6,6	83,0 ± 10,1	88,3 ± 6,4
NCNN + Platt	79,8 ± 6,2 ↑	75,9 ± 9,1 ↑	73,7 ± 16,2 ↑	84,7 ± 9,9 ↓	81,8 ± 10,9 ↓	88,3 ± 6,4 =
NCNN + Reg. Iso.	79,0 ± 5,5 ↓	75,6 ± 8,0 ↑	75,4 ± 17,2 ↑	81,8 ± 12,3 ↓	80,2 ± 11,7 ↓	88,2 ± 6,2 ↓
NCNN + Hist.	78,7 ± 6,1 ↓	75,7 ± 7,9 ↑	**75,8 ± 16,7** ↑	81,1 ± 14,1 ↓	80,2 ± 14,0 ↓	87,7 ± 6,2 ↓*
NCNN + Temp.	79,5 ± 8,7 =	74,8 ± 11,5 =	70,2 ± 17,2 =	87,7 ± 6,6 =	83,0 ± 10,1 =	88,3 ± 6,4 =
NCNN + LSR 0,1	78,1 ± 8,2 ↓	73,1 ± 11,2 ↓	68,6 ± 16,0 ↓	85,9 ± 8,3 ↓	81,2 ± 11,4 ↓*	87,2 ± 7,5 ↓
NCNN + LSR 0,3	79,7 ± 9,0 ↑	74,9 ± 11,9 ↑	69,6 ± 15,1 ↓	**88,4 ± 8,7** ↑	83,9 ± 13,2 ↑	88,3 ± 6,0 =
NCNN + LSR 0,5	**80,3 ± 7,0** ↑	**76,1 ± 8,8** ↑	70,4 ± 12,2 ↑	88,3 ± 7,9 ↑	**84,9 ± 11,1** ↑	**89,3 ± 5,4** ↑
NCNN + Linear	74,7 ± 8,1 ↓	70,3 ± 9,7 ↓	67,5 ± 14,5 ↓	80,3 ± 12,4 ↓	75,1 ± 9,5 ↓	83,5 ± 8,2 ↓
NCNN + Intervalo	75,9 ± 9,5 ↓	71,3 ± 13,0 ↓	68,8 ± 15,3 ↓	81,5 ± 5,9 ↓*	74,5 ± 11,4 ↓	84,6 ± 8,4 ↓
NCNN + Sigmoid	78,0 ± 8,7 ↓	72,9 ± 12,8 ↓	69,4 ± 17,3 ↓	84,9 ± 5,9 ↓	78,6 ± 9,3 ↓	85,4 ± 7,2 ↓

Fonte: Autor.

Tabela 37 – Métricas de calibração da NCNN (média ± DP, em %).

Modelo	**ECE**	**MCE**	**NLL**	**Brier**
NCNN	12,9 ± 4,7	**29,5 ± 12,6**	42,4 ± 10,1	**13,7 ± 4,1**
NCNN + Platt	13,6 ± 4,3 ↓	42,0 ± 16,8 ↓*	45,9 ± 13,1 ↓*	14,6 ± 4,3 ↓*
NCNN + Reg. Iso.	12,0 ± 4,4 ↑	42,9 ± 16,0 ↓*	45,8 ± 12,3 ↓	14,7 ± 4,2 ↓*
NCNN + Hist.	**11,8 ± 4,1** ↑	38,2 ± 15,0 ↓	45,6 ± 12,3 ↓*	14,8 ± 4,3 ↓*
NCNN + Temp.	12,1 ± 4,8 ↑*	31,8 ± 13,2 ↓	**41,9 ± 11,0** ↑	**13,7 ± 4,3** =
NCNN + LSR 0,1	14,9 ± 5,0 ↓	35,6 ± 14,3 ↓	44,9 ± 9,3 ↓*	14,5 ± 3,9 ↓
NCNN + LSR 0,3	17,9 ± 3,6 ↓*	36,7 ± 11,9 ↓*	46,7 ± 7,7 ↓*	14,8 ± 3,4 ↓
NCNN + LSR 0,5	21,3 ± 3,0 ↓*	38,4 ± 5,1 ↓	50,6 ± 5,0 ↓*	16,2 ± 2,3 ↓*
NCNN + Linear	16,9 ± 2,4 ↓*	36,3 ± 8,2 ↓	50,1 ± 7,8 ↓*	16,5 ± 3,4 ↓*
NCNN + Intervalo	15,2 ± 2,4 ↓	36,2 ± 6,9 ↓	47,5 ± 9,5 ↓	15,6 ± 4,0 ↓
NCNN + Sigmoid	17,8 ± 3,5 ↓*	35,6 ± 6,3 ↓	49,6 ± 7,0 ↓*	16,0 ± 3,1 ↓

Fonte: Autor.

Tabela 38 – Métricas de classificação da VGGFace (média ± DP, em %).

Modelo	Acurácia	F1-Score	Sensibilidade	Especificidade	Precisão	AUC
VGGFace	84,1 ± 5,0	81,5 ± 6,8	80,2 ± 10,3	87,0 ± 8,5	84,1 ± 8,8	91,9 ± 3,4
VGGFace + Platt	83,6 ± 4,6 ↓	80,9 ± 6,5 ↓	79,0 ± 10,8 ↓	86,9 ± 9,0 ↓	84,3 ± 9,3 ↑	91,9 ± 3,4 =
VGGFace + Reg. Iso.	84,1 ± 4,2 =	81,6 ± 5,6 ↑	79,5 ± 9,8 ↓	87,5 ± 8,7 ↑	85,2 ± 9,1 ↑	91,6 ± 3,3 ↓
VGGFace + Hist.	**84,9 ± 3,4** ↑	**82,1 ± 5,5** ↑	78,7 ± 8,8 ↓	**90,0 ± 5,9** ↑	**86,7 ± 7,8** ↑	91,4 ± 3,1 ↓*
VGGFace + Temp.	84,1 ± 5,0 =	81,5 ± 6,8 =	80,2 ± 10,3 =	87,0 ± 8,5 =	84,1 ± 8,8 =	91,9 ± 3,4 =
VGGFace + LSR 0,1	84,1 ± 4,6 ↓	81,0 ± 7,6 ↓	78,1 ± 10,3 ↓	88,1 ± 6,8 ↑	84,9 ± 8,2 ↑	92,1 ± 3,2 ↑
VGGFace + LSR 0,3	83,3 ± 5,3 ↓	79,7 ± 8,5 ↓	75,4 ± 10,7 ↓	89,5 ± 7,2 ↑	85,8 ± 10,5 ↑	**92,6 ± 3,3** ↑
VGGFace + LSR 0,5	83,9 ± 5,3 ↓	80,3 ± 8,2 ↓	75,8 ± 11,7 ↓	**90,0 ± 6,6** ↑	86,6 ± 7,9 ↑	92,3 ± 3,7 ↑
VGGFace + Linear	80,4 ± 4,5 ↓*	78,4 ± 6,0 ↓	80,0 ± 9,6 ↓	80,7 ± 6,9 ↓*	77,9 ± 7,2 ↓*	89,7 ± 2,3 ↓*
VGGFace + Intervalo	80,0 ± 4,4 ↓*	78,3 ± 6,9 ↓	**81,8 ± 9,9** ↑	78,4 ± 7,6 ↓*	76,0 ± 8,9 ↓*	89,5 ± 2,7 ↓*
VGGFace + Sigmoid	79,1 ± 5,1 ↓*	77,1 ± 7,3 ↓*	79,7 ± 10,5 ↓	78,9 ± 8,7 ↓*	76,1 ± 10,5 ↓*	89,7 ± 3,1 ↓*

Fonte: Autor.

Tabela 39 – Métricas de calibração da VGGFace (média ± DP, em %).

Modelo	ECE	MCE	NLL	Brier
VGGFace	10,3 ± 4,2	33,5 ± 17,0	37,7 ± 9,3	11,7 ± 3,2
VGGFace + Platt	12,5 ± 4,0 ↓*	42,0 ± 19,2 ↓*	45,6 ± 10,7 ↓*	12,9 ± 3,1 ↓*
VGGFace + Reg. Iso.	11,6 ± 3,8 ↓	54,7 ± 25,8 ↓*	58,2 ± 32,2 ↓*	12,7 ± 3,0 ↓*
VGGFace + Hist.	10,9 ± 3,9 ↓	36,0 ± 9,4 ↓	57,9 ± 31,5 ↓*	12,6 ± 2,9 ↓*
VGGFace + Temp.	11,4 ± 4,6 ↓*	38,8 ± 17,9 ↓*	46,2 ± 13,8 ↓*	12,5 ± 3,5 ↓*
VGGFace + LSR 0,1	**9,7 ± 2,0** ↑	**27,2 ± 7,5** ↑	**34,9 ± 6,8** ↑	**11,0 ± 2,5** ↑
VGGFace + LSR 0,3	13,8 ± 2,3 ↓	31,7 ± 5,6 ↑	38,4 ± 4,5 ↓	11,6 ± 1,9 ↑
VGGFace + LSR 0,5	18,8 ± 2,8 ↓*	33,1 ± 3,7 ↑	45,2 ± 2,9 ↓*	13,8 ± 1,3 ↓*
VGGFace + Linear	13,2 ± 3,5 ↓	31,8 ± 12,7 ↑	42,8 ± 5,2 ↓*	13,8 ± 2,2 ↓*
VGGFace + Intervalo	12,8 ± 3,7 ↓*	38,8 ± 13,6 ↓	44,0 ± 8,5 ↓*	14,3 ± 2,9 ↓*
VGGFace + Sigmoid	14,3 ± 3,7 ↓	32,6 ± 11,9 ↑	43,8 ± 5,2 ↓*	14,0 ± 2,2 ↓*

Fonte: Autor.

Tabela 40 – Métricas de classificação do ViT-B/32 (média ± DP, em %).

Modelo	Acurácia	F1-Score	Sensibilidade	Especificidade	Precisão	AUC
ViT-B/32	82,8 ± 6,7	79,9 ± 7,4	75,9 ± 10,2	88,6 ± 6,8	85,0 ± 6,8	89,9 ± 5,0
ViT-B/32 + Platt	82,2 ± 6,4 ↓	79,3 ± 7,2 ↓	75,9 ± 10,2 =	87,4 ± 6,0 ↓	83,6 ± 6,1 ↓	89,9 ± 5,0 =
ViT-B/32 + Reg. Iso.	82,7 ± 6,3 ↓	79,7 ± 6,8 ↓	75,0 ± 11,1 ↓	88,6 ± 8,2 =	86,2 ± 5,9 ↑	89,8 ± 5,3 ↓
ViT-B/32 + Hist.	82,1 ± 6,0 ↓	77,9 ± 7,7 ↓	71,5 ± 13,4 ↓*	91,1 ± 6,9 ↑	87,6 ± 6,9 ↑	89,4 ± 5,1 ↓
ViT-B/32 + Temp.	82,8 ± 6,7 =	79,9 ± 7,4 =	75,9 ± 10,2 =	88,6 ± 6,8 =	85,0 ± 6,8 =	89,9 ± 5,0 =
ViT-B/32 + LSR 0,1	83,2 ± 6,4 ↑	80,2 ± 6,8 ↑	75,4 ± 10,4 ↓	89,5 ± 6,9 ↑	86,3 ± 5,9 ↑	90,3 ± 4,4 ↑
ViT-B/32 + LSR 0,3	83,9 ± 5,7 ↑	80,8 ± 6,3 ↑	75,5 ± 10,4 ↓	90,9 ± 5,9 ↑	87,9 ± 5,8 ↑	90,7 ± 4,5 ↑
ViT-B/32 + LSR 0,5	**84,9 ± 4,7** ↑	**81,9 ± 5,4** ↑	76,7 ± 9,4 ↑	**91,6 ± 5,4** ↑	**88,9 ± 6,0** ↑*	**90,9 ± 4,4** ↑
ViT-B/32 + Linear	78,7 ± 6,7 ↓*	73,1 ± 8,1 ↓*	**85,6 ± 5,9** ↓	65,0 ± 12,7 ↑*	85,6 ± 5,9 ↑	89,8 ± 6,8 ↓
ViT-B/32 + Intervalo	80,2 ± 7,9 ↓*	76,4 ± 8,7 ↓*	71,3 ± 13,0 ↓*	86,9 ± 8,6 ↓	83,5 ± 7,2 ↓	87,4 ± 6,9 ↓*
ViT-B/32 + Sigmoid	80,6 ± 5,6 ↓	75,9 ± 6,7 ↓*	68,4 ± 11,6 ↓*	90,0 ± 7,1 ↑	86,9 ± 4,9 ↑	87,2 ± 7,0 ↓*

Fonte: Autor.

Tabela 41 – Métricas de calibração do ViT-B/32 (média ± DP, em %).

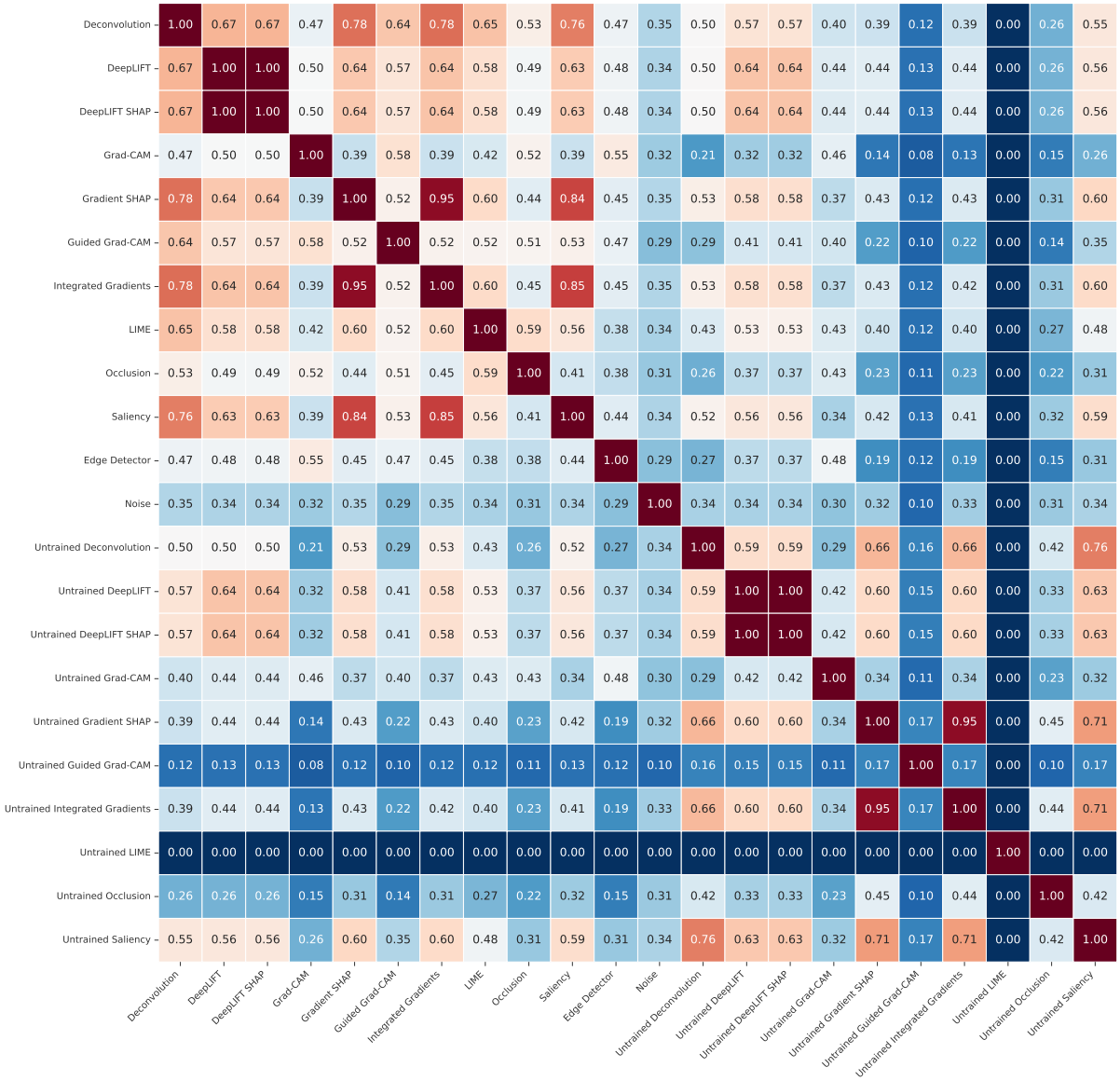
Modelo	**ECE**	**MCE**	**NLL**	**Brier**
ViT-B/32	12,3 ± 3,3	35,1 ± 10,9	39,5 ± 9,7	**12,5 ± 3,4**
ViT-B/32 + Platt	11,9 ± 3,4 ↑	40,2 ± 12,4 ↓*	41,5 ± 13,3 ↓	12,8 ± 4,2 ↓
ViT-B/32 + Reg. Iso.	10,6 ± 3,7 ↑*	32,5 ± 14,7 ↑	41,7 ± 15,2 ↓	12,7 ± 4,1 ↓
ViT-B/32 + Hist.	**9,4 ± 3,6** ↑*	27,9 ± 12,3 ↑*	42,0 ± 15,1 ↓	12,9 ± 4,4 ↓
ViT-B/32 + Temp.	11,3 ± 3,4 ↑	39,1 ± 11,7 ↓*	41,2 ± 13,6 ↓	12,6 ± 3,9 ↓
ViT-B/32 + LSR 0,1	14,3 ± 3,4 ↓*	36,4 ± 11,5 ↓	40,3 ± 7,5 ↓	12,6 ± 2,9 ↓
ViT-B/32 + LSR 0,3	19,5 ± 3,9 ↓*	36,5 ± 6,3 ↓	45,0 ± 5,3 ↓*	14,1 ± 2,3 ↓*
ViT-B/32 + LSR 0,5	24,2 ± 4,1 ↓*	40,8 ± 5,3 ↓	50,8 ± 3,8 ↓*	16,2 ± 1,8 ↓*
ViT-B/32 + Linear	86,4 ± 7,6 ↓*	**17,0 ± 4,5** ↑*	**34,6 ± 9,6** ↓	47,8 ± 8,5 ↓*
ViT-B/32 + Intervalo	15,3 ± 4,4 ↓*	37,1 ± 16,2 ↓	44,9 ± 8,9 ↓*	14,4 ± 3,5 ↓*
ViT-B/32 + Sigmoid	17,3 ± 2,7 ↓*	39,9 ± 11,8 ↓	48,1 ± 6,8 ↓*	15,4 ± 2,9 ↓*

Fonte: Autor.

APÊNDICE C – RESULTADOS AUXILIARES DE XAI

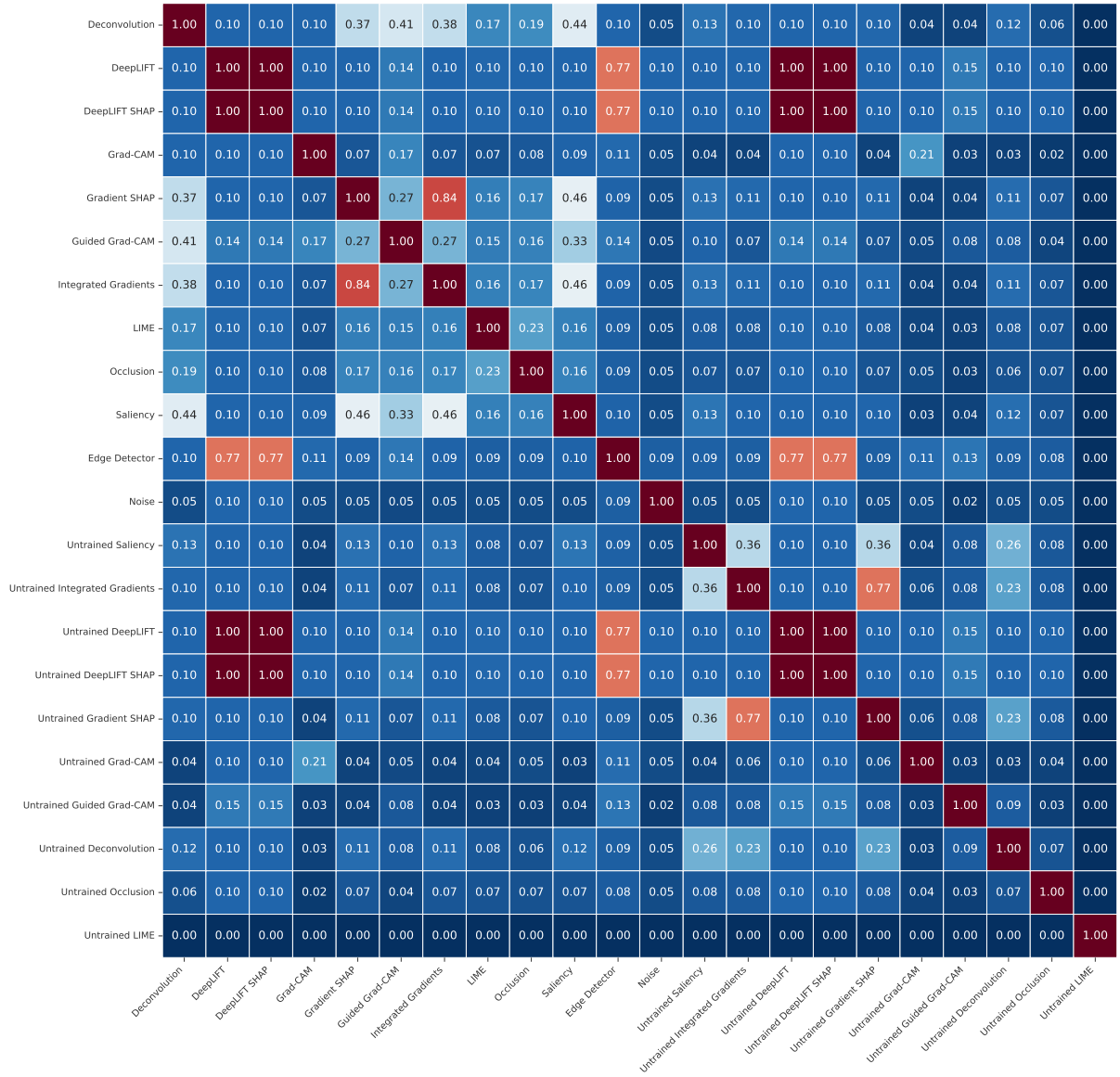
Este apêndice apresenta os resultados completos da concordância entre os explicadores em nível de região e de *pixel* para os modelos NCNN, VGGFace e ViT-B/32, complementando a análise apresentada no Capítulo 6. As Figuras 53 a 58 apresentam esses resultados para cada modelo e nível de análise.

Figura 53 – Concordância em nível de região entre os explicadores para o modelo NCNN.



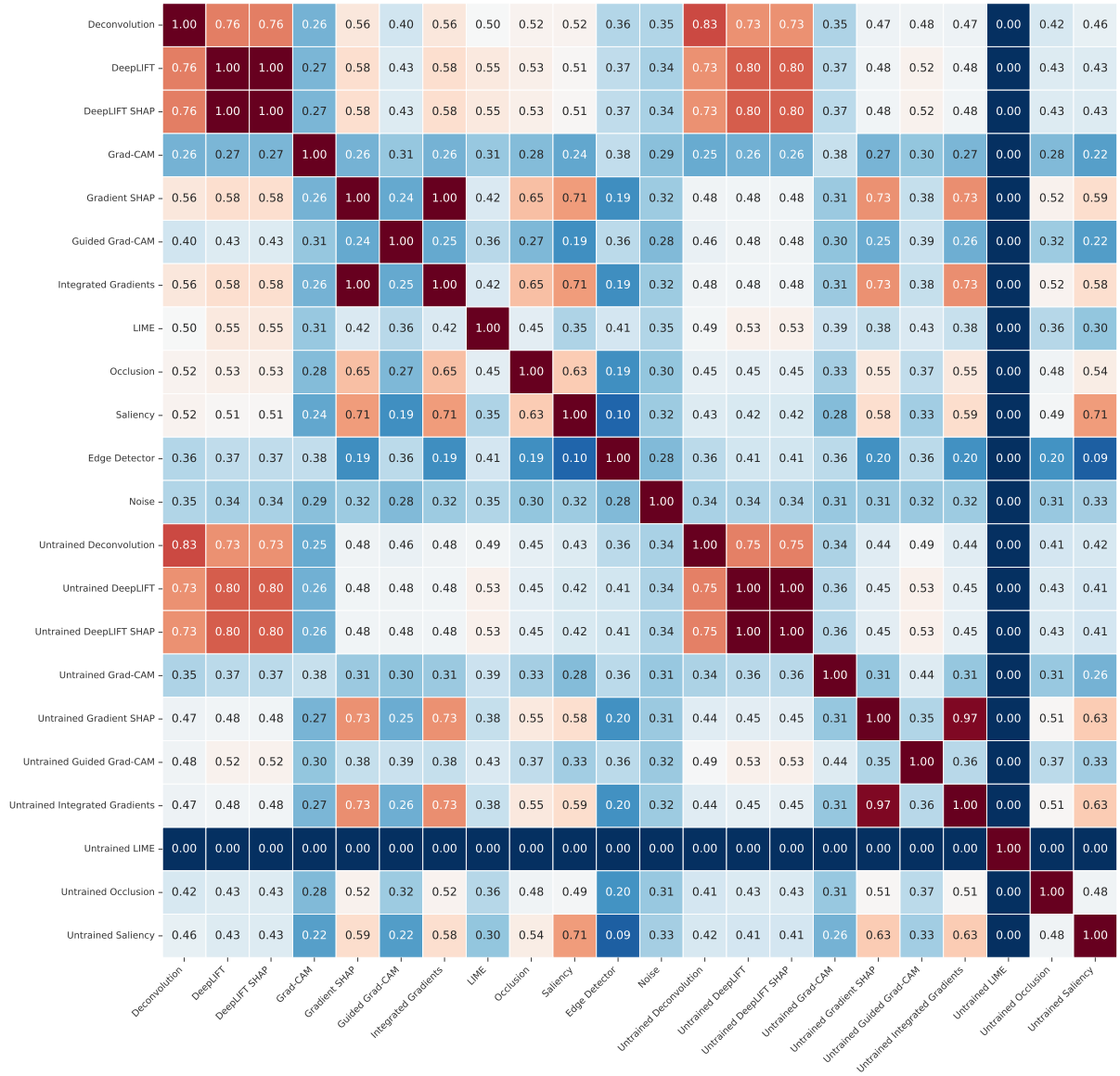
Fonte: Autor.

Figura 54 – Concordância em nível de *pixel* entre os explicadores para o modelo NCNN.



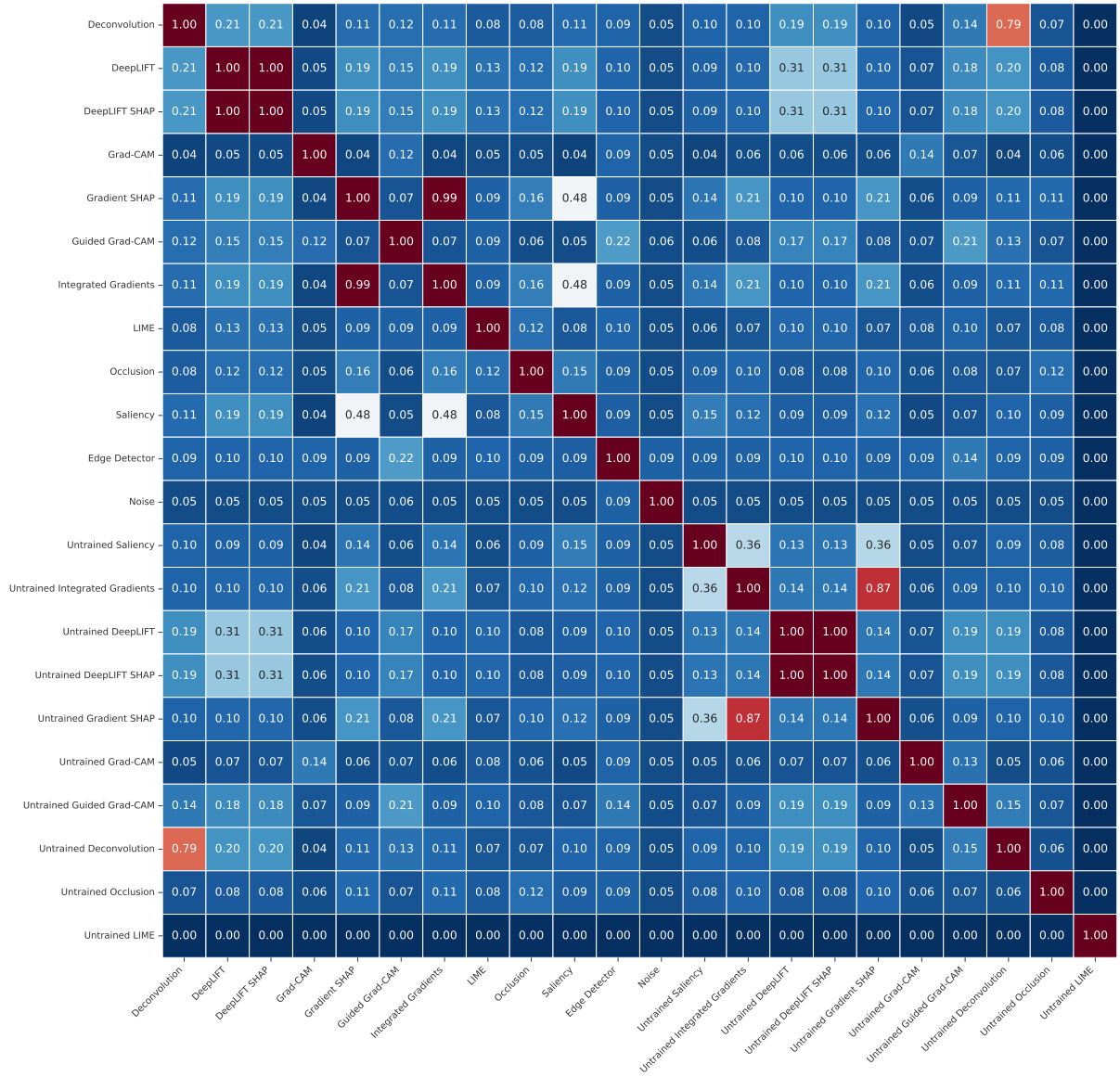
Fonte: Autor.

Figura 55 – Concordância em nível de região entre os explicadores para o modelo VGGFace.



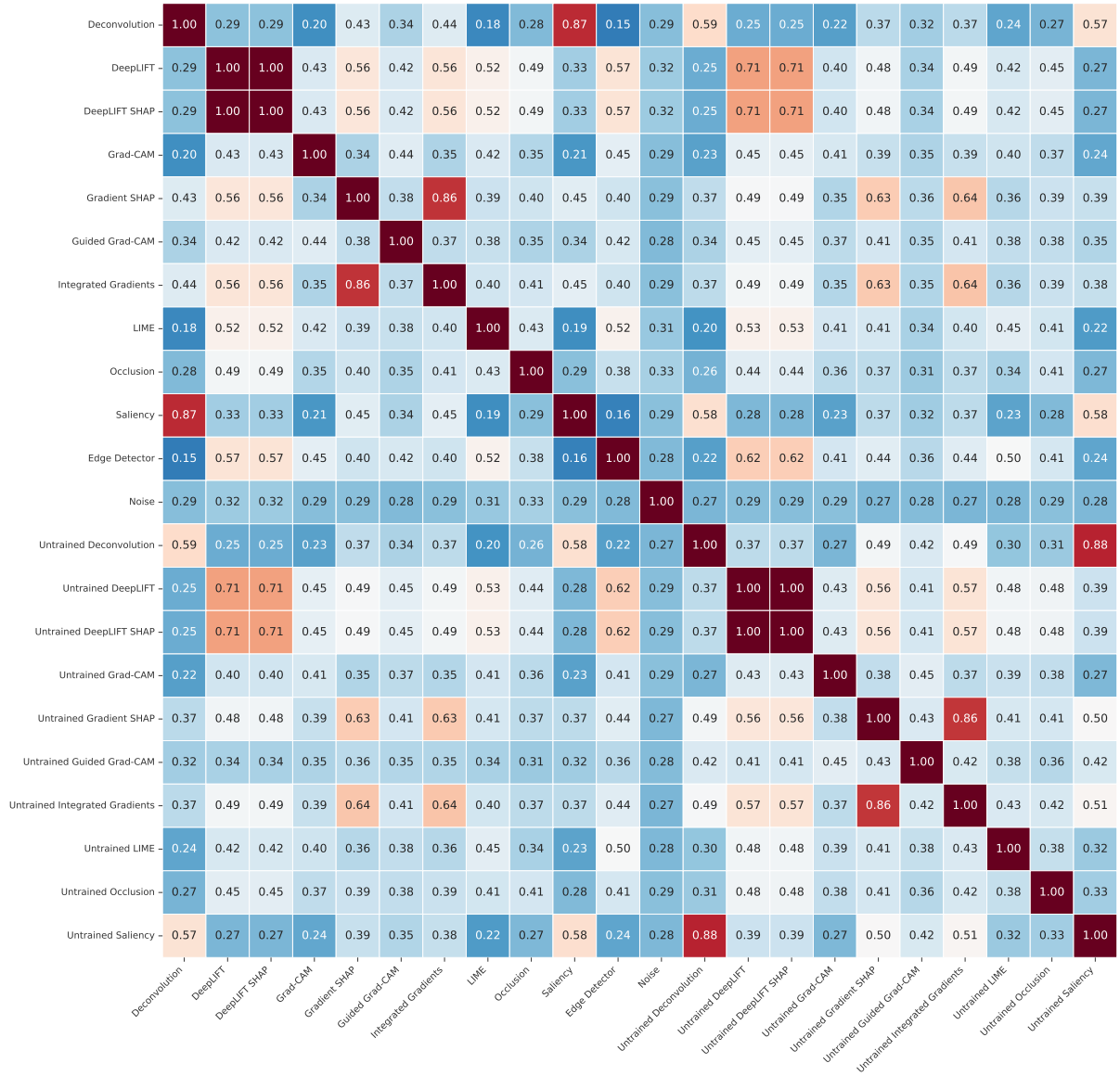
Fonte: Autor.

Figura 56 – Concordância em nível de *pixel* entre os explicadores para o modelo VGGFace.



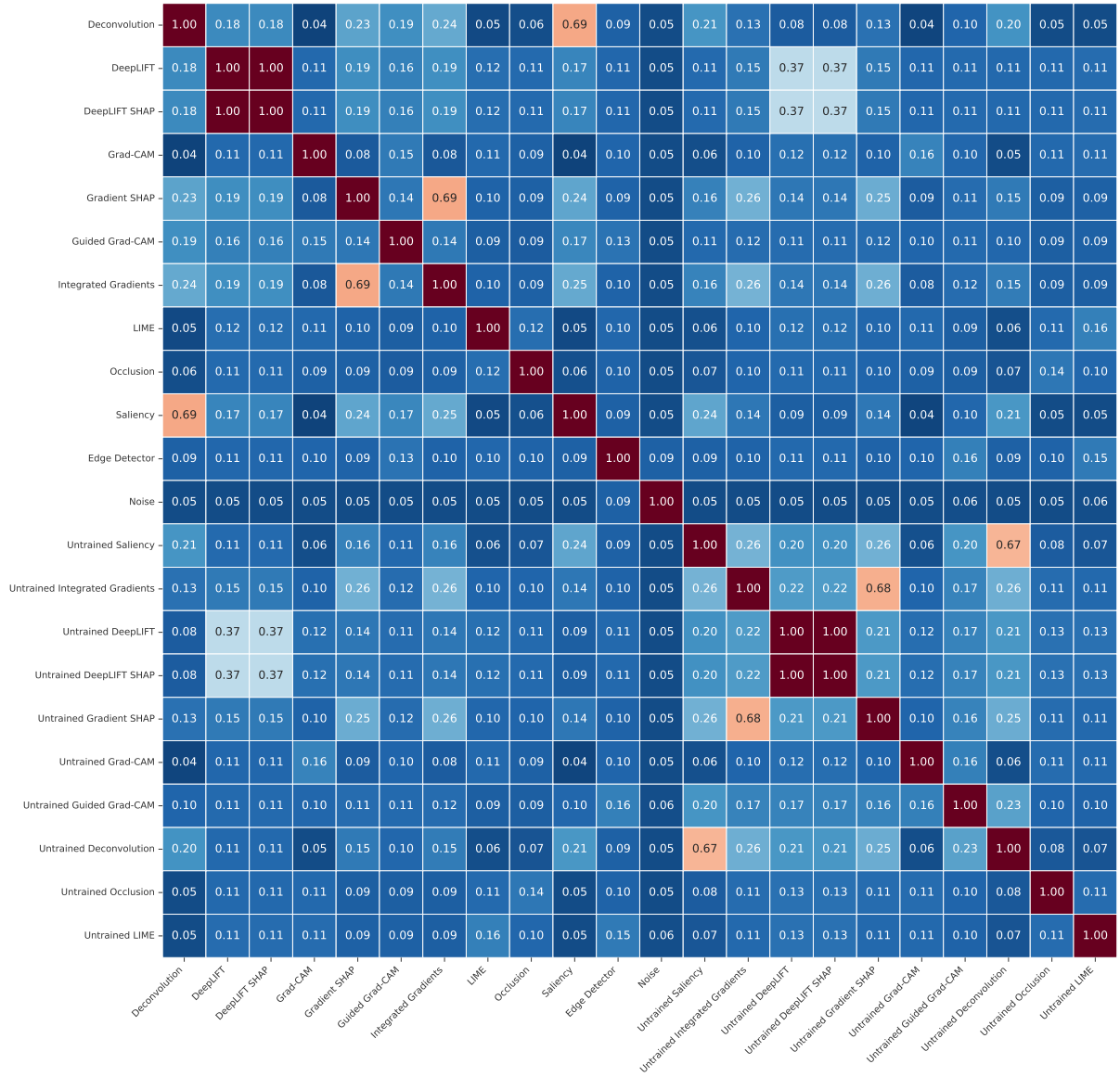
Fonte: Autor.

Figura 57 – Concordância em nível de região entre os explicadores para o modelo ViT-B/32.



Fonte: Autor.

Figura 58 – Concordância em nível de *pixel* entre os explicadores para o modelo ViT-B/32.



Fonte: Autor.

APÊNDICE D – RESULTADOS AUXILIARES DOS LIMIARES

Este apêndice documenta visualmente as amostras removidas pelos critérios de precisão e incerteza. As Figuras 59, 60 e 61 apresentam, respectivamente, as amostras removidas pelo limiar de precisão (θ_2), pelo limiar de incerteza (θ_3) e pela aplicação conjunta dos dois. As imagens exibidas foram rejeitadas simultaneamente pelos três modelos e, portanto, consideradas de baixa confiabilidade por todos eles. Sua análise ajuda a identificar padrões associados à baixa confiabilidade e a refinar os critérios de exclusão em futuras iterações dos modelos.

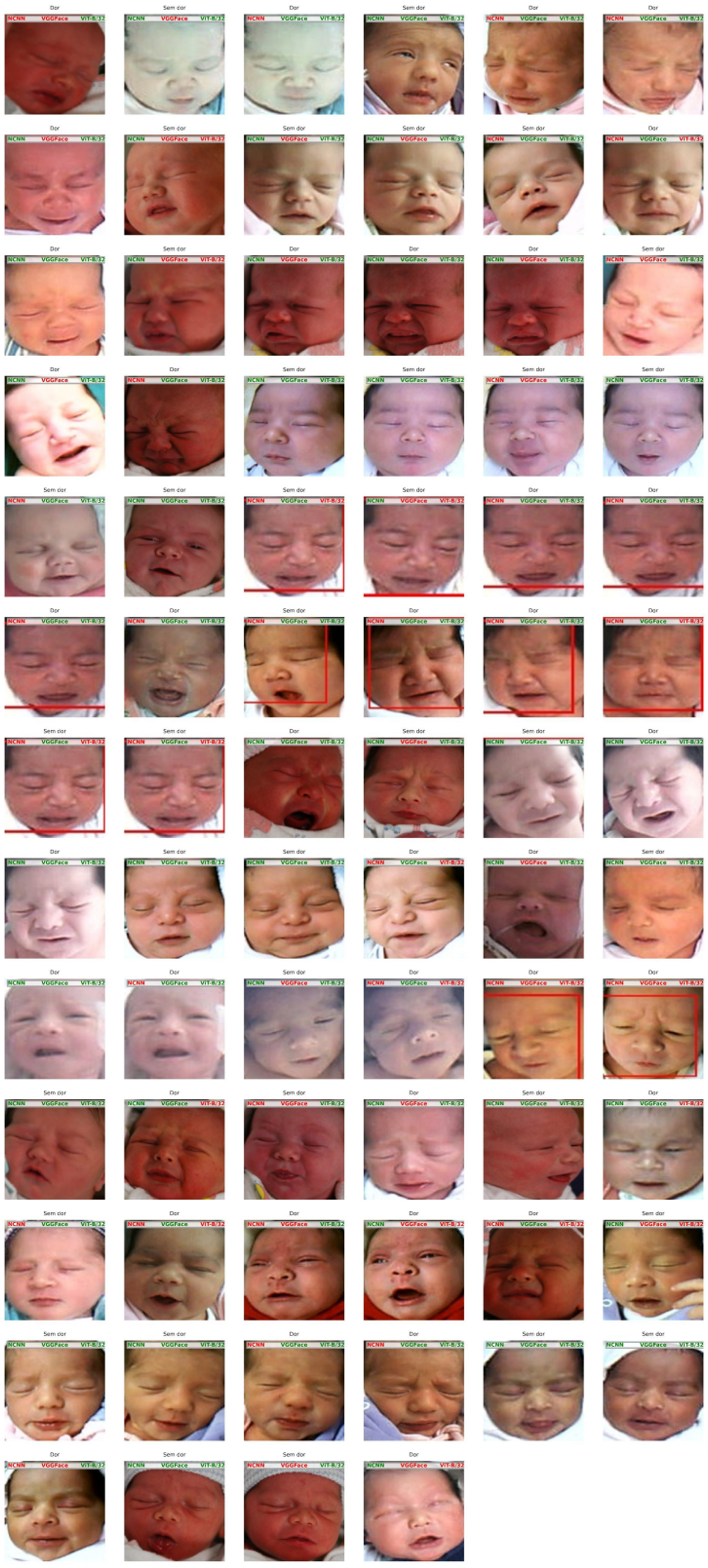
Na parte superior de cada amostra, apresentam-se a classe original da imagem (dor ou sem dor) e a indicação de acerto ou erro do modelo. A cor verde representa um acerto, e a vermelha, um erro.

Figura 59 – Amostras removidas pelo limiar de precisão θ_2.



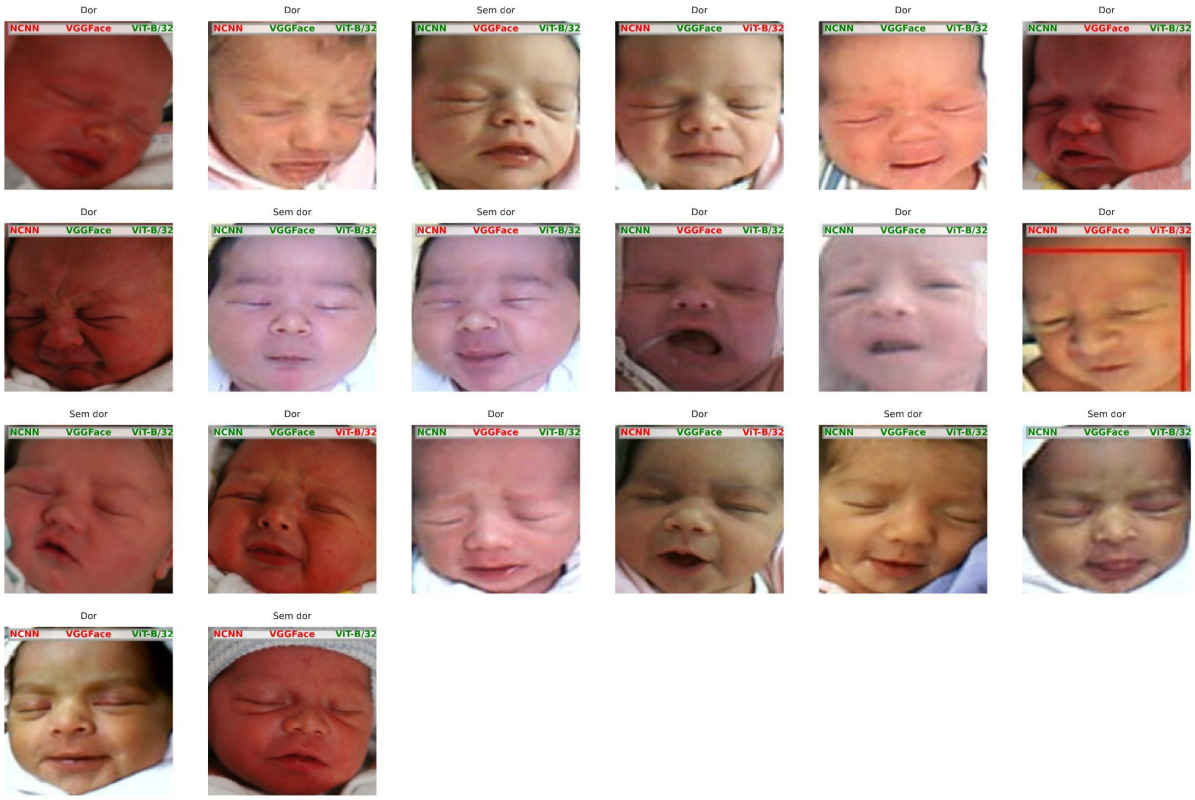
Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

Figura 60 – Amostras removidas pelo limiar de incerteza θ_3.



Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

Figura 61 – Amostras removidas pela aplicação conjunta dos limiares θ_2 e θ_3.

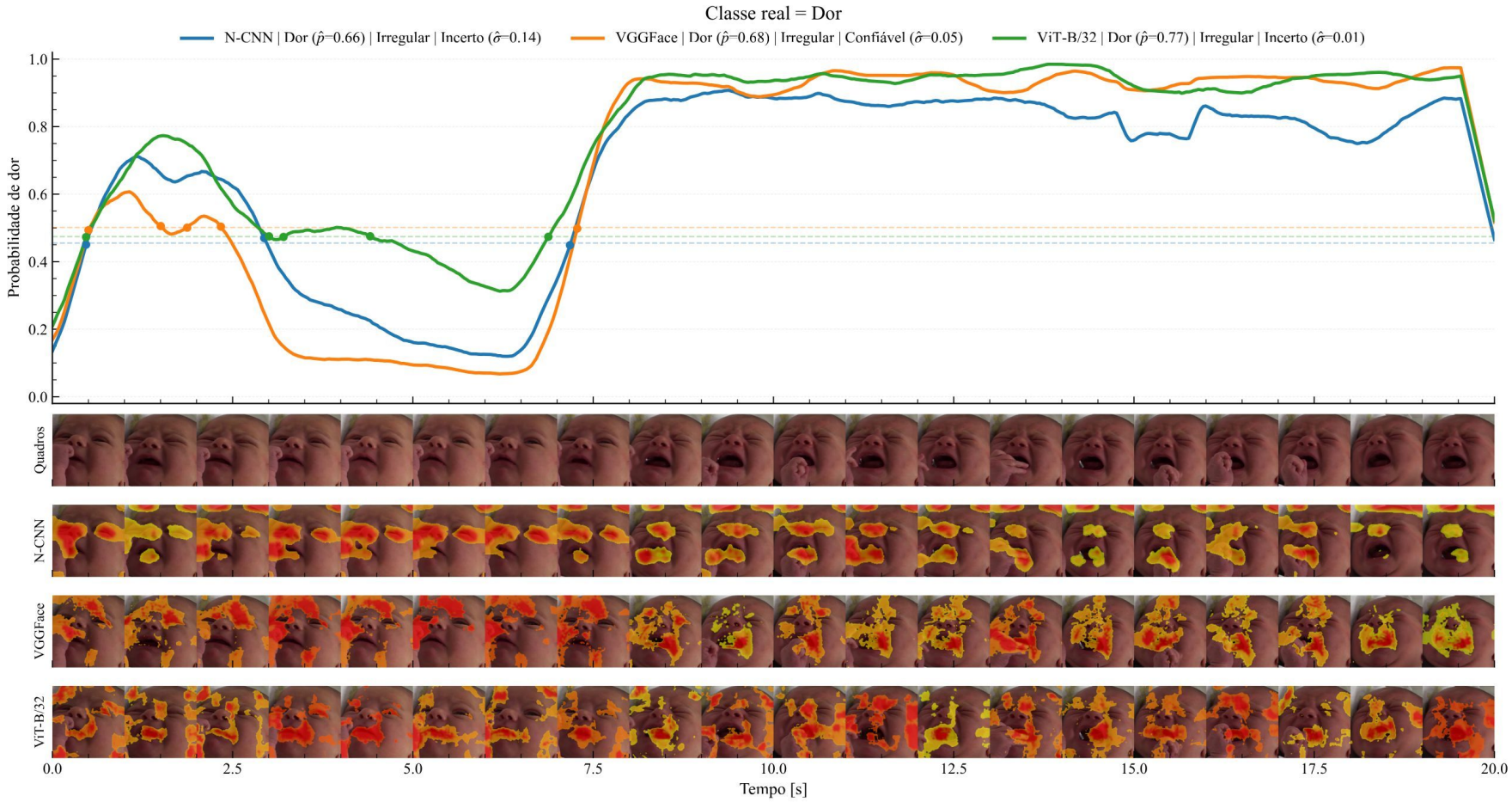


Fonte: Autor, utilizando banco de imagens da UNIFESP (Heiderich; Leslie; Guinsburg, 2015) e iCOPE (Brahnam *et al.*, 2006).

APÊNDICE E – RESULTADOS AUXILIARES DO SINAL DA DOR

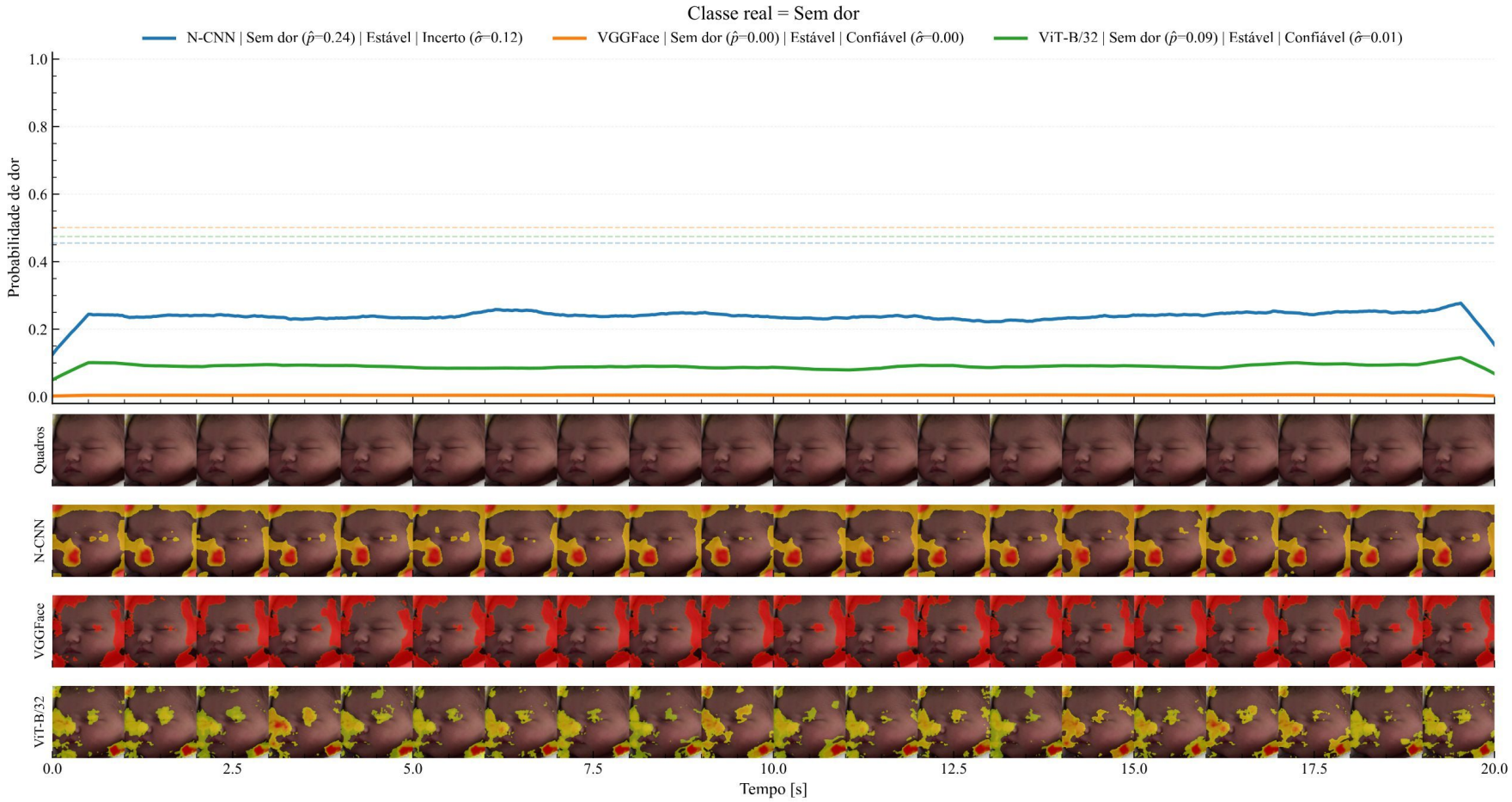
Este apêndice apresenta exemplos complementares dos sinais da dor descritos no Capítulo 9. Foram selecionados seis segmentos do banco iCOPEVid, três rotulados como “dor” e três como “sem dor”, para ampliar a inspeção qualitativa da evolução temporal das probabilidades estimadas pela NCNN, pela VGGFace e pelo ViT-B/32. Esses segmentos são apresentados nas Figuras 62 a 67.

Figura 62 – Sinal da dor do segmento de dor do recém-nascido S002.



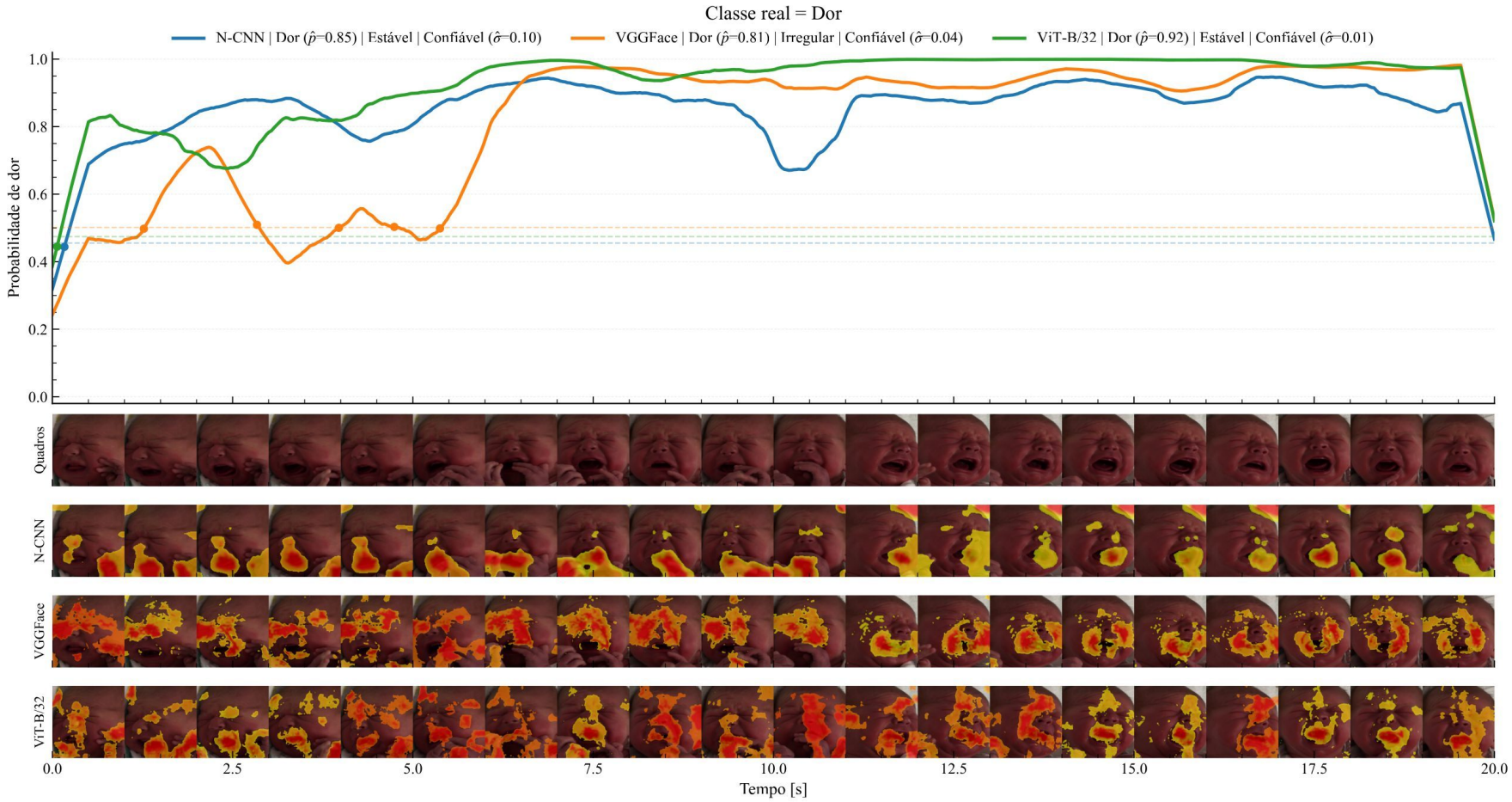
Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

Figura 63 – Sinal da dor do segmento sem dor do recém-nascido S003.



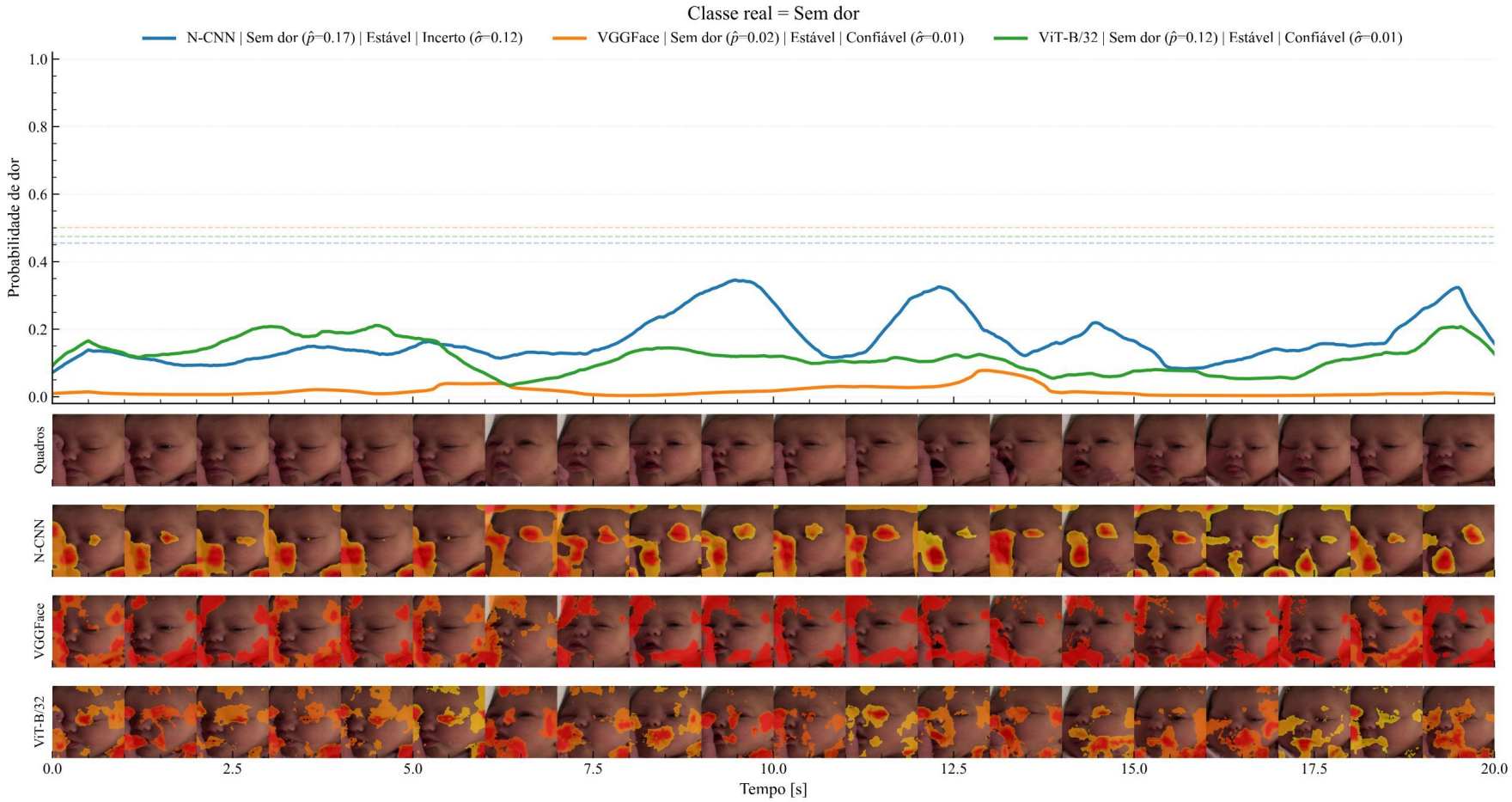
Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

Figura 64 – Sinal da dor do segmento de dor do recém-nascido S005.



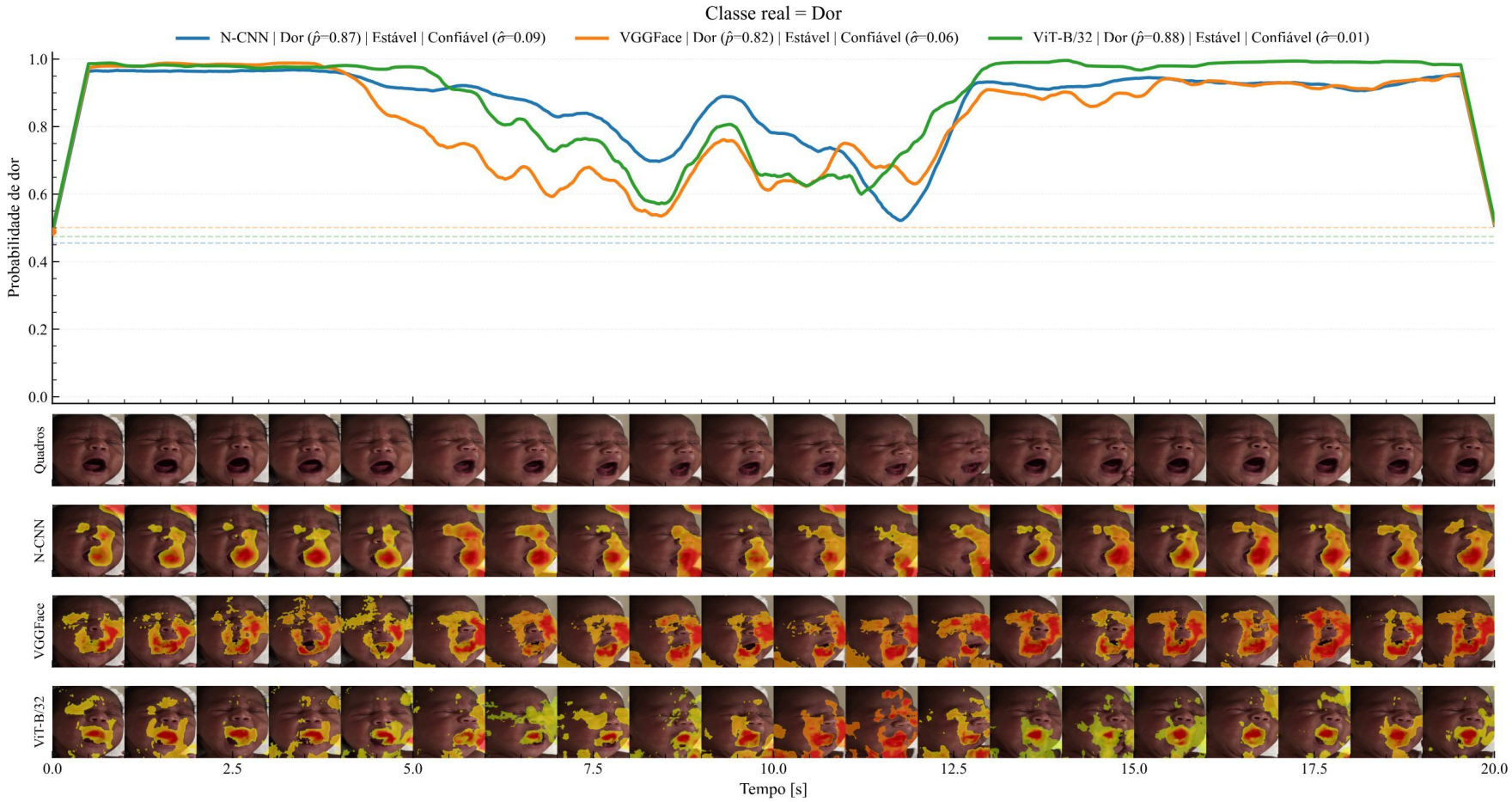
Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

Figura 65 – Sinal da dor do segmento sem dor do recém-nascido S006.



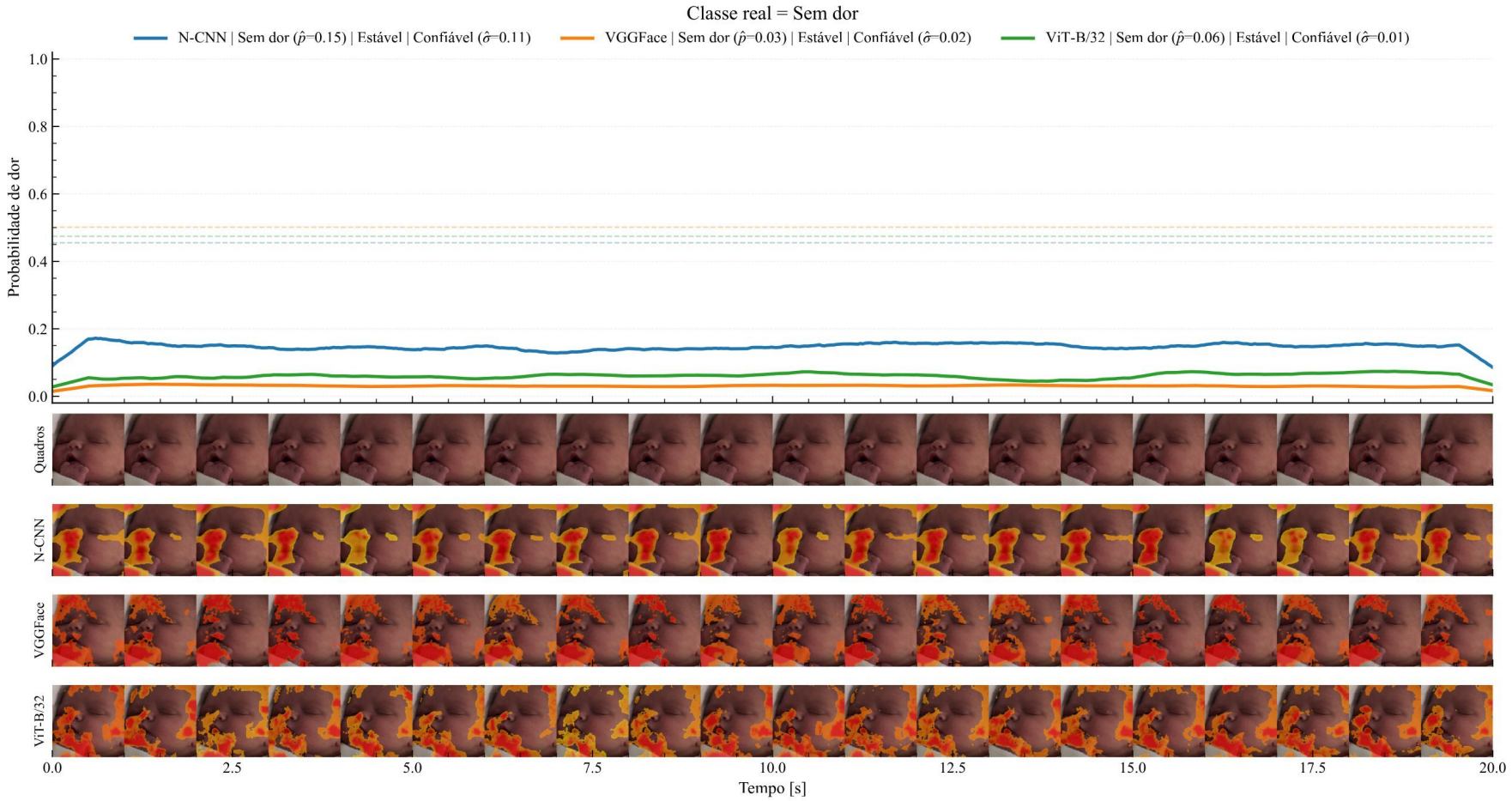
Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

Figura 66 – Sinal da dor do segmento de dor do recém-nascido S007.



Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

Figura 67 – Sinal da dor do segmento sem dor do recém-nascido S008.



Fonte: Autor, utilizando banco de vídeos iCOPEvid (Brahnam *et al.*, 2020).

APÊNDICE F – ARTIGO PUBLICADO NO WORKSHOP *PREDICTIVE INTELLIGENCE IN MEDICINE (PRIME 2023)*



Revisiting N-CNN for Clinical Practice

Leonardo Antunes Ferreira[1(✉)], Lucas Pereira Carlini[1], Gabriel de Almeida Sá Coutrin[1], Tatiany Marcondes Heideirich[1], Marina Carvalho de Moraes Barros[2], Ruth Guinsburg[2], and Carlos Eduardo Thomaz[1]

[1] FEI University Center, Sao Bernardo do Campo, SP 09850-901, Brazil
{leferr,cet}@fei.edu.br
[2] Federal University of Sao Paulo, Sao Paulo SP 04024–002, Brazil

Abstract. This paper revisits the Neonatal Convolutional Neural Network (N-CNN) by optimizing its hyperparameters and evaluating how they affect its classification metrics, explainability and reliability, discussing their potential impact in clinical practice. We have chosen hyperparameters that do not modify the original N-CNN architecture, but mainly modify its learning rate and training regularization. The optimization was done by evaluating the improvement in F1 Score for each hyperparameter individually, and the best hyperparameters were chosen to create a Tuned N-CNN. We also applied soft labels derived from the Neonatal Facial Coding System, proposing a novel approach for training facial expression classification models for neonatal pain assessment. Interestingly, while the Tuned N-CNN results point towards improvements in classification metrics and explainability, these improvements did not directly translate to calibration performance. We believe that such insights might have the potential to contribute to the development of more reliable pain evaluation tools for newborns, aiding healthcare professionals in delivering appropriate interventions and improving patient outcomes.

Keywords: Neonatal pain · AI · Explainability · Reliability

1 Introduction

In recent years, Artificial Intelligence (AI) models capable of automatically detecting pain through facial expression analysis have gained significant attention, addressing the issues of human subjectivity and untreated pain [4]. Among these models, the Neonatal Convolutional Neural Network (N-CNN) stands out as the first end-to-end Deep Learning model specifically designed and trained for neonatal pain detection based on facial expressions [14]. In fact, the N-CNN offers a lightweight architecture, that requires less memory and lower computational resources, making it suitable for embedding into mobile or clinical devices within neonatal intensive care units (NICU), ensuring faster training and recall capabilities.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
I. Rekik et al. (Eds.): PRIME 2023, LNCS 14277, pp. 231–240, 2023.
https://doi.org/10.1007/978-3-031-46005-0_20

Models intended for clinical practice must exhibit qualities such as accuracy, explainability and reliability [4,8,9]. While the N-CNN has already proven to be accurate [14], limited research has been conducted regarding its hyperparameters, explainability and reliability. Hyperparameter optimization, or tuning, is often used to improve model performance without altering its architecture. Although this performance improvement can be verified by traditional classification metrics and eXplainable Artificial Intelligence (XAI) methods, less attention has been given to model reliability. Being reliable, or calibrated, is a highly desirable property in automatic medical diagnosis, as it enables the estimation of associated risks during clinical practice [8,9]. For instance, high-confidence predictions made by an AI model should exactly match the likelihood of that event occurring, otherwise, overconfident predictions can potentially lead to harmful consequences. In this context, this paper investigates the effects of hyperparameter optimization on classification metrics, explainability and reliability discussing their potential impact in clinical practice.

2 Materials and Methods

For training and testing the N-CNN, we utilized two facial image datasets: iCOPE [1] and UNIFESP [7]. In this way, the model was trained from scratch. The iCOPE dataset comprised 60 images labeled as *"Pain"* and 63 images labeled as *"No Pain"* from 26 newborns (NBs). The UNIFESP dataset included 164 *"Pain"* images and 196 *"No Pain"* images from 30 NBs. To ensure robust evaluation, we employed a leave-sample-subjects-out cross-validation method [3] with 10 folds, using the same 10 folds in all evaluations. Data augmentation was applied to training images in each fold, generating 20 new images with varied parameters such as width and height shift (0.20), rotation (30°), shear (0.15), brightness (0.50–1.10), zoom (0.70–1.50), and horizontal flip.

2.1 N-CNN Performance Evaluation

We selected hyperparameters according to Table 1 that do not modify the original N-CNN architecture [14], primarily affecting learning rate and training regularization. However, a slight modification was required in the N-CNN output activation function to accommodate the Label Smoothing Regularizer (LSR) [13] applied here. Specifically, we replaced the single output neuron activated by a sigmoid function with two neurons activated by a softmax function, being each neuron responsible for providing normalized probabilities for the respective classes. We exclusively focused on the *"Pain"* neuron output, where a threshold of 50% was set to consider a prediction belonging to this class.

Since we separately evaluated each hyperparameter, only those that demonstrated an improvement in F1 Score were selected to make the Tuned N-CNN. To assess the performance of the final model, we employed standard evaluation metrics including Accuracy, F1 Score, Precision, and Sensibility. We also used Grad-CAM (*GC*) [11] and Integrated Gradients (*IG*) [12]. Both are considered

to be attribution methods and provide further insights into the model's decision-making process, but GC focuses on regions of the images deemed discriminant to classification, whereas IG assigns to each pixel an importance value to the final classification. Regarding the GC, we generated the attribution mask based on the last convolutional layer of the N-CNN, where the model should capture high-level features and semantic information [11].

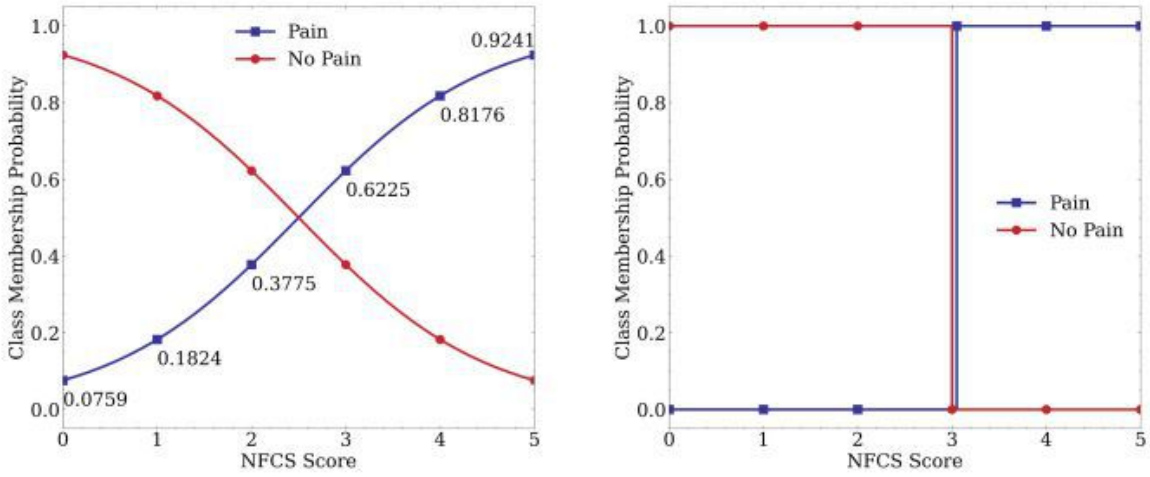
Additionally, we employed calibration curves to validate whether the confidence scores yielded by the N-CNN are aligned with the frequency of the *"Pain"* class occurrences. The calibration curve plots the frequency of the positive class against the model's confidence scores. A perfectly calibrated model is represented by a 45° diagonal line, that is, the confidence is exactly equal to the actual frequency of events [6]. Points below the diagonal indicate overconfidence, where the model assigns higher probabilities to the occurrence of events than what is observed in reality. Conversely, points above the diagonal indicate a lack of confidence, where the model's confidence is lower than expected despite correctly identifying most of the positive class samples. To quantify if a model is calibrated, we used the Expected Calibration Error (ECE) [10]. This metric divides predictions confidences into K equally-spaced bins, here $K = 10$, and calculates the weighted average of the differences between the frequency of the positive class and the confidence within each bin, providing a quantification of the model's miscalibration [10]. It is important to note that the calibration curve focuses solely in the positive class, considered here as the *"Pain"* class.

2.2 NFCS as Soft Label

One of the studied hyperparameters is the value ϵ of the LSR [13]. It involves adjusting the ground truth labels during training by introducing a small amount of smoothing. Instead of assigning a hard label (binary target) of 0 or 1 to a particular class, LSR assigns a slightly softened label (non-binary target), such as [0.1, 0.9]. This regularization technique helps prevent the model from becoming overconfident and overfitting [13]. Nonetheless, applying the same smoothing value for all labels can potentially result in information loss as we increase ϵ, encouraging the model to be less confident about the true class. Therefore, we propose a novel approach for establishing independent class membership probabilities for each image using the Neonatal Facial Coding System (NFCS) [5], as it is widely used and available in the UNIFESP dataset [7].

The NFCS aims to quantify the presence of 8 facial action units (FAUs) related to pain facial expressions [5]. However, the UNIFESP dataset considered only 5 FAUs due to acquisition limitations [7]. A $NFCS \geq 3$ indicates a facial expression of pain [5], which is easily mapped to a hard label as seen in Fig. 1b. For mapping to a soft label, the sigmoid in Eq. 1 was used, as it offers a degree of smoothness in the probability distribution and provides normalized probabilities between 0 and 1. To assure the sigmoid stays in the range (0–5) of the NFCS scores available in the UNIFESP dataset we added the value 2.5:

$$S(NFCS) = (1 + e^{-NFCS+2.5})^{-1}. \tag{1}$$



(a) NFCS soft labels. (b) NFCS hard labels.

Fig. 1. NFCS mapping to soft and hard labels.

Equation 1 maps the NFCS score to the *"Pain"* class membership probability, while the complement is used for the *"No Pain"* class. The final soft label is made of $[1 - S(NFCS), S(NFCS)]$ (Fig. 1a).

Using the NFCS as a soft label, we encourage the output neuron of the *"Pain"* class to be more confident in images with more FAUs of pain. In contrast, the output neuron of the *"No Pain"* class will be more confident when there are few or no FAUs of pain. It is important to acknowledge that the availability of NFCS scores was limited to the UNIFESP dataset. Therefore, to ensure consistency in our evaluations and fair comparisons with other hyperparameters, we excluded the iCOPE images from the original 10 training folds. However, the iCOPE images remained a part of the test folds.

3 Results

As we evaluated each hyperparameter individually, some were kept fixed when not being assessed, based on [14]: learning rate ($1e^{-4}$), total number of epochs (100), and batch size (16). Model checkpoints were created if the test loss value decreased, and we only selected the lowest loss value checkpoint for evaluation.

3.1 Metrics

Changing image size and optimizer did not improve F1 Score. The LSR improved F1 Score in +4.21% when using $\epsilon = 0.3$. Cosine Annealing presented the best results to the learning rate schedulers, with an improvement of +0.84%. Furthermore, increasing the number of epochs led to a notable improvement of +2.99%. Surprisingly, even with 30% fewer images for training, the proposed NFCS soft

label achieved the same F1 Score as the Original N-CNN 74.10%, demonstrating the positive impact of integrating probabilities during training.

Table 2 presents the classification metrics of the test folds for the Original and Tuned N-CNN, with standard deviations. All metrics are statistically significant by the *t-test* ($p < 0.05$), highlighting that hyperparameter optimization alone can lead to improvements in classification results without altering the N-CNN architecture.

Table 1. Hyperparameters search space and selected values.

Hyperparameter	Search Space	Selected	F1 Score	Δ
Image Size	64×64, 120×120, 224×224	120×120	74.10%	0%
Optimizer	Adam, Adagrad, RMSProp, SGD	RMSprop	74.10%	0%
Epochs	50 to 120	120	77.09%	+2.99%
Label Smoothing	0.1, 0.3, 0.5 and NFCS Soft Label	0.3	78.31%	+4.21%
Schedulers	Step, Exponential, Cosine Annealing	Cosine Annealing	74.94%	+0.84%

Table 2. Classification metrics for Original and Tuned N-CNN.

Model	Accuracy	F1 Score	Precision	Recall
Original	78.69% ± 5%	74.10% ± 6%	83.17% ± 8%	68.96% ± 9%
Tuned	82.97% ± 5%	79.74% ± 6%	85.63% ± 8%	75.43% ± 9%
Δ	+4.28%	+5.64%	+2.46%	+6.47%

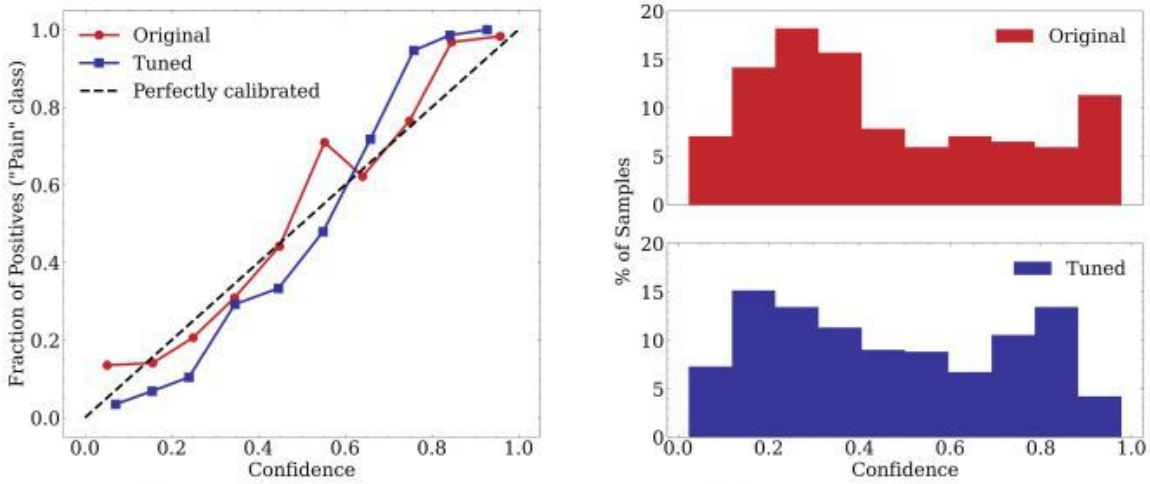
3.2 Reliability

Despite achieving higher classification metrics, the Tuned N-CNN did not exhibit the same level of calibration as the Original N-CNN. Specifically, the calibration curve of the Tuned N-CNN displayed an "S" shape, indicating overestimation of predictions with low confidences and underestimation of predictions with high confidences (Fig. 2a). This difference in calibration is confirmed by the ECE metric [10], with the Tuned N-CNN yielding an $ECE = 0.091$, approximately three times greater than the Original N-CNN's $ECE = 0.035$.

As shown in Fig. 2a, of the predicted samples for which the model exhibits confidence around 20%, less than 10% of them were actually in the *"Pain"* class.

On the other hand, almost all samples with 80% confidence were of the *"Pain"* class, indicating that the Tuned N-CNN has a lower confidence in this interval. Conversely, when considering the Original N-CNN, the model exhibits limited confidence in the 50% to 60% range, indicating challenges in confidently classifying samples within the threshold region of the *"Pain"* class.

The low confidence phenomenon in Tuned N-CNN, is further depicted in Fig. 2b, showing how the percentage of samples with confidences over 90%, from Original N-CNN, shifted to a less confident range due to the LSR technique which tends to reduce model overconfidence.



(a) Calibration curves. (b) Confidences histogram.

Fig. 2. Calibration results of the Original and Tuned N-CNN.

3.3 Explainability

Figures 3 and 4 depict images correctly classified by both models, providing insights into the explainability improvements when accessing the Tuned N-CNN. For example, the Tuned N-CNN focused on regions of the NBs face, ignoring artifacts such as clothing (Fig. 3d). In addition, it was found that the IG_{Tuned} acquired a more uniform distribution across the NBs face, unlike the $IG_{Original}$, which gave more attention to the eyes and mouth. As for the *"No Pain"* images, GC_{Tuned} concentrated on the forehead and cheeks (Fig. 3e), contrasting with $GC_{Original}$, which did not obtain a central area of focus.

Our results revealed a correlation between XAI and the predicted confidence in both Original and Tuned N-CNNs. Confidences ranging from 40% to 60% were observed to explain features not related to the NB face, as it can be seen in Fig. 3c and 3f, that is, an overconfidence in the *"Pain"* class. This interpretation was also observed for images that both models incorrectly classified. In contrast, confidences above 80% or below 20% in the *"Pain"* class that were correctly classified presented an attribution mask deemed more relevant to clinical practice.

Moreover, both *IG* and *GC* results mainly gave more importance to the mouth region, which agrees with one of the observed regions by neonatologists [2].

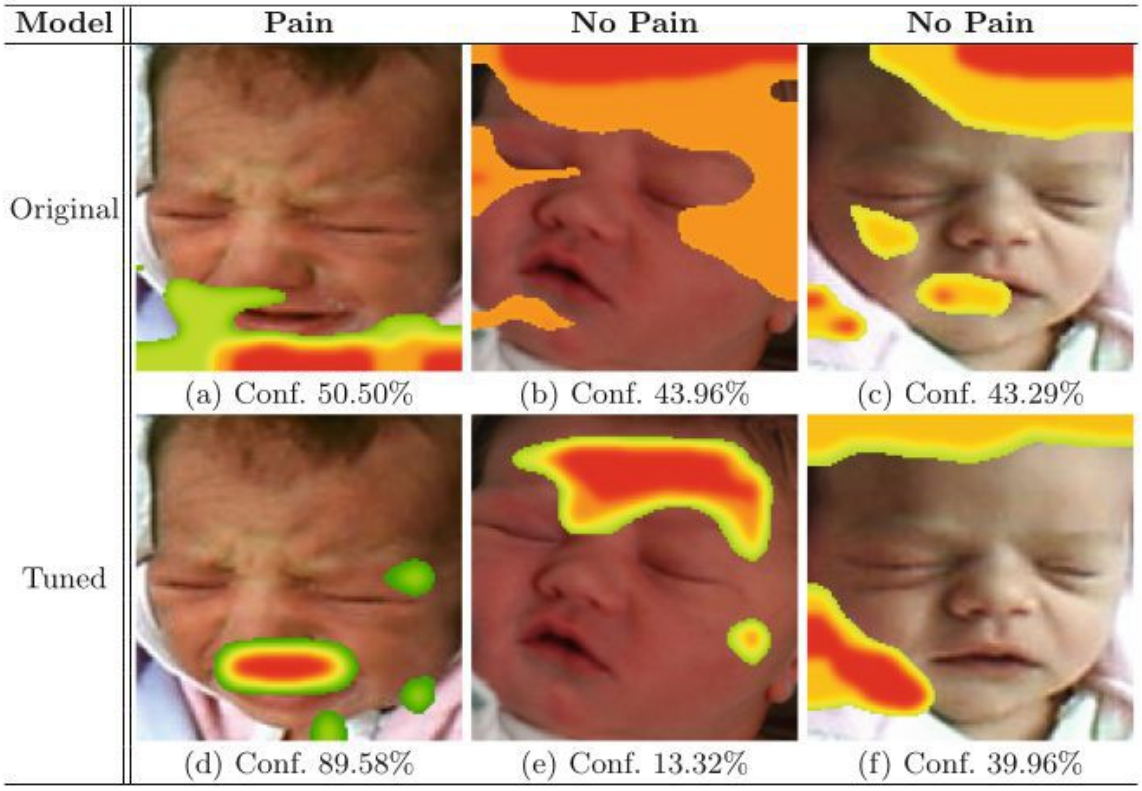


Fig. 3. *GC* attribution masks and predictions confidence in the *"Pain"* class.

4 Discussion

Based on our observations, training with soft labels instead of hard labels proved beneficial to the N-CNN classification metrics, which has not been extensively explored in the existing literature. We believe that the NFCS soft label proposed here can surpass the best LSR result found if iCOPE NFCS scores were available, emphasizing the necessity for standardization across neonatal pain scales in datasets, promoting consistency and comparability in future research. This technique can also be applied to various clinical pain scales that quantify the presence or absence of pain, not only to NFCS.

Regarding reliability, calibrated predictions are essential to accurately estimate the associated risks in critical settings like in NICU. For instance, predictions with 90% confidence in the *"Pain"* class should prompt immediate action as the NB has a real 90% chance of being in pain, while predictions with 60% confidence require more careful analysis by healthcare professionals, as the AI model exhibits uncertainty in its prediction for the specific situation of the NB. The fact that the calibration of the Tuned N-CNN has worsened raises an important

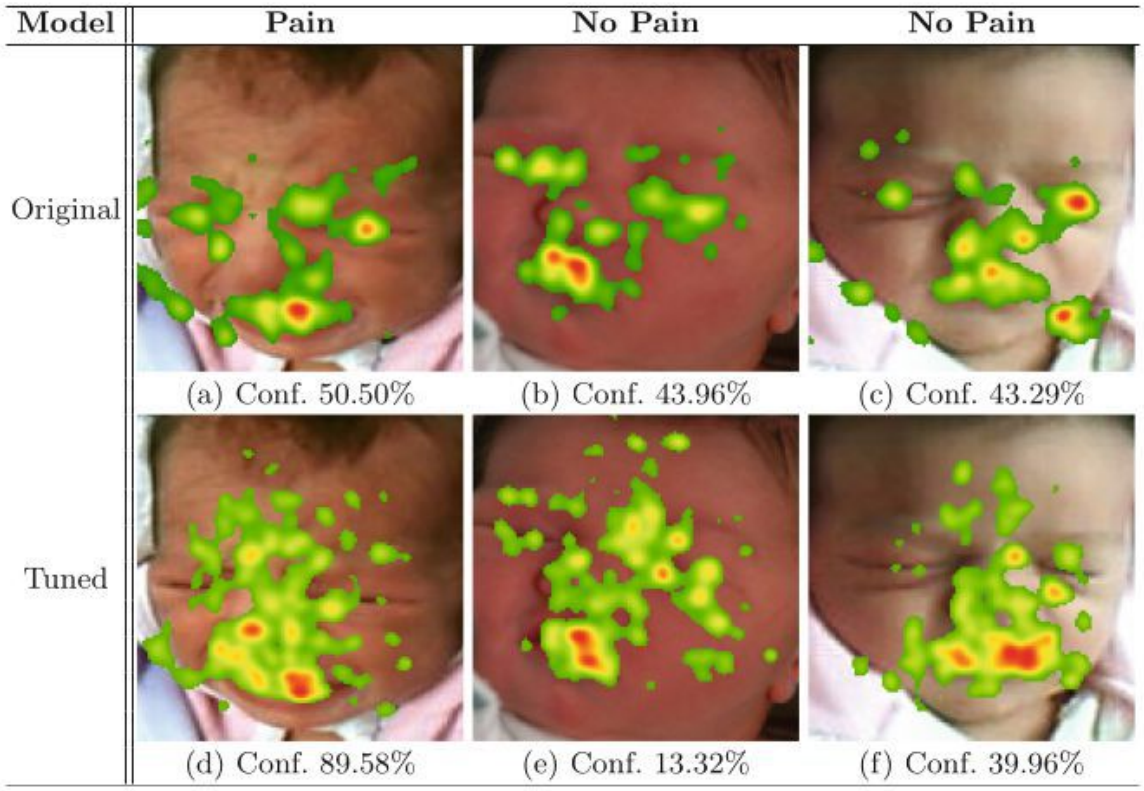


Fig. 4. *IG* attribution masks and predictions confidence in the *"Pain"* class.

question: to what extent does improving classification metrics benefit clinical practice? While a +5% increase in the Tuned N-CNN metrics may lead to more accurate automatic predictions, healthcare professionals may only rely on these predictions if the model's decision-making process is explainable and aligned with the actual probability of event occurrence.

The use of *IG* on the Tuned N-CNN revealed that more pixels contributed to the final prediction. We believe that this increase in relevant pixels were a contributing factor to the Tuned N-CNN achieving higher metrics. However, the N-CNN still has some explainability limitations when compared to other Deep-Learning models [3], due to the shallow architecture which may limit its capacity to capture complex patterns and nuances in neonatal pain assessment. Nevertheless, by visualizing these attribution masks, healthcare professionals can better understand the model's reasoning and gain trust in its predictions.

5 Conclusion

In this work, we revisited the N-CNN model by optimizing its hyperparameters and evaluating how tuning affects its classification metrics, explainability and reliability. Although we observed improvements in classification metrics and explainability, these improvements did not simply translate to calibration performance. An uncalibrated model applied in clinical practice can lead to a lack of

trust by the health professional in the model's predictions, resulting in the rejection of the technology. While enhancements in traditional metrics are notable, achieving reliability and explainability in any AI model intended for the health sector is imperative to ensure reliable and actionable predictions.

We believe that the new approach proposed here of using the NFCS as a soft label might disclose a new paradigm in training classification models for neonatal pain analysis based on facial expressions. As future work, we intend to explore soft labels based on clinical pain scales and calibration methods, ensuring that confidences align with the actual frequency of both pain and no-pain events. Overall, we understand that it is crucial to address the issue of uncalibrated confidences as risk estimators in the medical field, not only to prevent erroneous automatic decision-making and foster trust among healthcare professionals, but also to facilitate the responsible and effective use of AI in clinical practice.

Acknowledgments. This study was financed in part by CAPES (Finance Code 001) and FAPESP (2018/13076-9).

References

1. Brahnam, S., Chuang, C.F., Shih, F.Y., Slack, M.R.: Machine recognition and representation of neonatal facial displays of acute pain. Artif. Intell. Med. **36**(3), 211–222 (2006)
2. Carlini, L.P., et al.: A visual perception framework to analyse neonatal pain in face images. In: Campilho, A., Karray, F., Wang, Z. (eds.) ICIAR 2020. LNCS, vol. 12131, pp. 233–243. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50347-5_21
3. Coutrin, G.A., et al.: 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. LNEE. Springer, Cham (2024). ISSN: 1876-1100
4. Gkikas, S., Tsiknakis, M.: Automatic assessment of pain based on deep learning methods: a systematic review. Comput. Methods Programs Biomed. **231**, 107365 (2023)
5. Grunau, R.V., Craig, K.D.: Pain expression in neonates: facial action and cry. Pain **28**(3), 395–410 (1987)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
7. Heiderich, T.M., Leslie, A.T.F.S., Guinsburg, R.: Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. Acta Paediatr. **104**(2), e63–e69 (2015)
8. Jiang, X., Osl, M., Kim, J., Ohno-Machado, L.: Calibrating predictive model estimates to support personalized medicine. J. Am. Med. Inform. Assoc. **19**(2), 263–274 (2012)
9. Kompa, B., Snoek, J., Beam, A.L.: Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit. Med. **4**(1), 4 (2021)
10. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29 (2015)

11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
12. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328. PMLR (2017)
13. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
14. Zamzmi, G., Paul, R., Goldgof, D., Kasturi, R., Sun, Y.: Pain assessment from facial expression: neonatal convolutional neural network (n-CNN). In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2019)

APÊNDICE G – ARTIGO PUBLICADO NA REVISTA *COMPUTERS IN BIOLOGY AND MEDICINE*

Computers in Biology and Medicine 189 (2025) 109908



Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed





Disclosing neonatal pain in real-time: AI-derived pain sign from continuous assessment of facial expressions

Leonardo Antunes Ferreira [a,*], Lucas Pereira Carlini [a], Gabriel de Almeida Sá Coutrin [a], Tatiany Marcondes Heiderich [a,b], Rita de Cássia Xavier Balda [b], Marina Carvalho de Moraes Barros [b], Ruth Guinsburg [b], Carlos Eduardo Thomaz [a]

[a] Department of Electrical Engineering, Centro Universitário FEI, Av. Humberto de Alencar Castelo Branco, 3972-B, São Bernardo do Campo, 09850-901, São Paulo, Brazil

[b] Department of Paediatrics, Federal University of São Paulo, R. Botucatu, 740, São Paulo, 04024-002, São Paulo, Brazil

ARTICLE INFO

Keywords:
Pain assessment
Deep-learning
Continuous monitoring
Pain sign

ABSTRACT

This study introduces an AI-derived pain sign for continuous neonatal pain assessment, addressing the limitations of existing pain scales and computational approaches. Traditional pain scales, though widely used, are hindered by inter-rater variability, discontinuity, and subjectivity. While AI, particularly Deep-Learning, has shown promise, prior research has largely prioritized model performance over clinical applicability, often delivering static, binary predictions that lack interpretability in clinical practice. To bridge this gap, we developed a real-time pain sign tracking tool using facial expression analysis, a primary and non-invasive pain indicator in neonates. Leveraging benchmark datasets (iCOPE, iCOPEvid, and UNIFESP) and Deep-Learning frameworks (VGG-Face, N-CNN, and ViT-B/16), the models analyze video frames to generate a continuous visual representation of pain probability. Our results reveal the limitations of single-label predictions for time intervals, emphasizing the utility of a continuous monitoring visualization tool. The proposed pain sign effectively tracks dynamic changes in neonatal facial expressions, providing actionable and interpretable insights for healthcare professionals. We categorized these insights into a novel classification scheme, such as stable, irregular, unstable, and indeterminate pain signs. By integrating this pain sign into clinical workflows as a potential vital sign, this approach enables personalized pain management and continuous monitoring of both current and historical pain states in neonates, enhancing neonatal care and improving outcomes for these vulnerable patients.

1. Introduction

In the past, neonatal pain was disregarded under the common assumption that the neural pathways responsible for pain in newborns were not fully developed, resulting in invasive procedures being performed without the use of analgesic or anesthetic agents [1,2]. Medical advancements have since debunked the notion that newborns do not experience pain [3]. However, even with this understanding, neonates admitted to NICUs (Neonatal Intensive Care Units) may still undergo more than 300 painful procedures during their stay [4] and two-thirds of neonates admitted into NICUs do not receive continuous pain assessment [5]. In fact, it is known that prolonged or repeated exposure to painful stimuli can lead to detrimental lifelong consequences [1,4,6].

Over 40 neonatal pain scales have since been proposed [4], like the Neonatal Facial Coding System (NFCS) [2], Neonatal Pain, Agitation and Sedation Scale (N-PASS) [7], Neonatal Infant Pain Scale (NIPS) [8], and Face, Legs, Activity, Cry, Consolability (FLACC) scale [9]. The existence of numerous scales underscores the importance of identifying and addressing pain, as well as the difficulty in establishing a universal method for neonatal pain assessment. Most of these scales share the analysis of facial expressions as a common feature, capturing highly sensitive and specific indicators of pain intensity and nature, such as a furrowed brow, narrowed eyes, nasolabial furrow, tense tongue, and a widened mouth angle [2]. Facial expression is, in fact, a primary and non-invasive form of manifestation of pain, being a common and universal expression present from birth regardless of gender and ethnicity [10], facilitating effective communication between non-verbal neonates and healthcare professionals.

Although pain scales are widely employed, their precision is compromised by the lack of inter-rater agreement among healthcare professionals [11–13]. Moreover, these scales are not continuous due to

* Corresponding author.
E-mail address: leferr@fei.edu.br (L.A. Ferreira).

https://doi.org/10.1016/j.compbiomed.2025.109908
Received 2 October 2024; Received in revised form 4 February 2025; Accepted 19 February 2025
Available online 1 March 2025
0010-4825/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

the labor-intensive nature of the process, requiring the participation of multiple skilled professionals, resulting in intermittent monitoring [5]. Additionally, their subjectivity is rooted in the innate human tendency to perceive pain subjectively in others [14].

To address these challenges, researchers have explored the use of Artificial Intelligence (AI), particularly Deep-Learning, for detecting neonatal pain through facial expressions [12,15,16]. These studies have demonstrated accurate and objective results and highlight the potential applicability of AI as a continuous tool for predictive early pain management [17]. However, a significant gap remains between computational and medical literature, specifically the need for a clinically comprehensible and user-friendly tool for neonatal pain assessment that delivers actionable insights with minimal training [13]. While previous computational studies have primarily focused on improving performance metrics by developing Deep-Learning architectures or integrating multidimensional data, they often fall short in producing practical clinical tools. In other words, if current research methods were implemented in clinical settings, healthcare professionals would still rely on static, discontinuous predictions, such as binary labels ("pain" or "no pain") or single pain intensity estimates for a given time interval. These approaches lack the interpretability and visual clarity required in NICUs, where clinicians must frequently assess both the historical and current pain states of neonates to distinguish between acute, prolonged, persistent, or chronic pain [18].

To bridge the gap between computational methodologies and clinical practice in neonatal pain assessment, our study brings the following main contributions:

- *A real-time, continuous pain sign*: Proposed and implemented to track dynamic changes in neonatal facial expressions, addressing the limitations of static, single-label pain assessments;
- *A novel set of temporal classifications for the pain phenomenon*: Developed specifically for healthcare professionals in clinical practice, supporting actionable decision-making in NICU settings;
- *Correlation between model probabilities and pain facial expression intensity*: Our experimental results show a strong correlation between the probabilities generated by AI models and the intensity of pain-related facial expressions, even though the models were not explicitly trained for regression tasks.

2. Material and methods

2.1. Data sources and pre-processing

We utilized three benchmark datasets of neonatal pain focused on capturing neonatal facial expressions: iCOPE [19], UNIFESP [20], and iCOPEvid [21]. Both UNIFESP and iCOPE consist of images, whereas iCOPEvid is a video dataset. The characteristics of each dataset are detailed in Table 1.

Brahnam et al. [19] conducted one of the pioneering studies investigating neonatal pain through facial expressions, and subsequently created one of the earliest datasets for this purpose, named iCOPE dataset. Adhering to protocols and ethics directives for research involving human subjects at St. John's Health System Inc., the authors obtained images of healthy neonates during predefined procedures, including transport from one crib to another, air stimulus, friction, and pain caused by a heel puncture. The captured images were categorized as follows: 63 images classified as resting, 18 depicting crying, 23 during air stimulus, 36 during friction, and 60 exhibiting pain. For our study, only images labeled as "rest" and "pain" were utilized.

The UNIFESP dataset was established with the objective of developing a software capable of automatically detecting facial expressions of pain in newborns by Heiderich et al. [20]. All newborns involved in this data collection were healthy, free from congenital malformations, and were hospitalized at Hospital São Paulo of Escola Paulista de Medicina at the Federal University of São Paulo. Image acquisition was granted approval by the Research Ethics Committee of the university under protocol number 1299/09. For each second, three images were extracted from 10-min videos recorded before, during, and after painful procedures such as injections, heel puncture, and venipuncture. Out of all the images acquired, 360 images were clinically evaluated by neonatal healthcare professionals using the NFCS [2] resulting in "no pain" and "pain" labels. Consequently, only these labeled images were included in this study.

Table 1
Datasets characteristics.

Dataset	iCOPE	UNIFESP	iCOPEvid
Data type	Image	Image	Video
Newborns	26	30	49
Males	13	15	26
Females	13	15	23
Postnatal age (hours)	24 h–144 h	18 h–72 h	34 h–70 h
Camera resolution	3008 × 2000	320 × 233	1920 × 1080
Total samples	123	360	120
Pain	60	164	71
No Pain	63	196	49

Later, Brahnam et al. [21] also introduced the iCOPEvid dataset adheres to the same procedures and ethical considerations as iCOPE for data acquisition. However, in this instance, videos of neonates experiencing both painful (heel puncture) and non-painful (movement, friction, and rest) stimuli were recorded for at least 1 min during each procedure. From the raw footage, a total of 234 video segments, each lasting 20-s, were extracted, capturing the precise moment when the noxious stimulus was administered. According to the authors of this dataset, when feasible, the non-painful video segments were hand-picked by two investigators to present challenging facial expressions for classification. This implies that non-painful facial expressions exhibit characteristics similar to those of painful expressions. To maintain consistency against datasets, we solely utilized the "rest" and "pain" categories, same and analogous as for the iCOPE and UNIFESP datasets.

All images and videos underwent a pre-processing step of face detection, wherein we employed the RetinaFace [22] face detector. This process involved cropping out only the faces of the neonates as the regions of interest.

2.2. Deep-learning framework

We utilized two benchmark models for automatic neonatal pain assessment [21,23–30], VGG-Face [31] and N-CNN (Neonatal Convolutional Neural Network) [32]. Additionally, we introduced a third model, a ViT (Vision Transformer), specifically the ViT-B/16 [33], to assess whether newer Deep-Learning architectures can perform comparably to established CNNs in this task.

UNIFESP and iCOPE datasets were merged for training and validating the Deep-Learning models, whereas iCOPEvid was reserved solely for testing, providing a verification task that is closer to clinical reality. To ensure robust evaluation and prevent data leakage during training, we employed the cross-validation method known as leave-some-subjects-out [27]. In this approach, the neonates were evenly divided into ten folds. During each iteration of cross-validation, one fold was set aside for validating the model, while the remaining folds were used for training. For each training fold, we applied standard data augmentation techniques, presented in Appendix A. All models were trained in a standard binary format, employing a single output neuron activated by a Sigmoid function. This neuron predicts the probability, ranging from zero to one, of the prediction belonging to the positive class, namely "pain". For additional details on the training protocol, refer to Appendix B.

Table 2
Classification performance on validation and test datasets.

Datasets	Metrics	VGG-Face	N-CNN	ViT-B/16	*p*-value
Validation results on iCOPE + UNIFESP	Accuracy	88.6% ± 6.0%	83.6% ± 7.2%	82.9% ± 4.2%	0.2405
	Sensitivity	88.4% ± 9.0%	82.5% ± 10.8%	87.6% ± 6.9%	0.5916
	Specificity	89.3% ± 9.7%	85.3% ± 13.7%	79.5% ± 9.5%	0.0550
	AUC	93.0% ± 4.3%	89.0% ± 5.4%	89.1% ± 2.9%	0.2016
Test results on iCOPEvid	Accuracy	63.3%	66.7%	56.7%	–
	Sensitivity	85.7%	81.6%	83.7%	–
	Specificity	47.9%	56.3%	38.0%	–
	AUC	71.2%	74.0%	68.8%	–

Footnote: The *p*-values are regarding the statistical difference between models during validation. There is no standard deviation in the iCOPEvid results because only the best model across all available folds was used.

2.3. Explainability

To better understand the model's decision-making process towards its predictions, we employed two explainability methods: Grad-CAM (GC) [34] and Integrated Gradients (IG) [35]. Both are classified as attribution methods, generating heatmap masks highlighting the most crucial regions of the input image influencing the model's prediction. Nonetheless, GC and IG operate differently. GC identifies relevant regions of the image, and IG assesses the individual contribution of each pixel to the final classification.

To improve the visualization of these masks, we applied a post-processing technique proposed by Carlini et al. [29], where the color red indicates the most important image features, yellow represents moderately important features, and green highlights the least important regions of the image to the model's predictions.

2.4. Pain sign

With the goal of providing real-time and continuous insight into the occurrence probability of painful events in neonates, we introduced the pain sign real-time tracking concept. Through the analysis of facial expressions using Deep-Learning, the pain sign effectively communicates all changes in the neonate's facial expressions in a visual, time-sensitive, and continuous manner to the health professionals.

The pain sign is two-dimensional, with the x-axis representing time and the y-axis indicating the predicted probability outputted by the Deep-Learning model for the "pain" class. For each frame captured in a video recording, a prediction probability is associated with that frame. If no face is detected during the video, the model generates no prediction. In such cases, we employed linear interpolation to fill in the missing values. As a final step, we applied a moving average filter of 30 samples, equaling to 1-s of video, to reduce potential noise.

2.5. Evaluation metrics and statistical analysis

For each validation fold, we assessed the models' performance using the following metrics with standard deviation: accuracy, sensitivity, specificity, and AUC (Area Under the receiver operating characteristic Curve). To determine the optimal threshold between "pain" and "no pain" predictions, we utilized Youden's Index [36], calculated during the training folds to prevent information leakage. The statistical differences between models were evaluated using the Friedman test with a significance level set at $p < 0.05$.

For the temporal evaluation of the pain sign, we computed the average prediction probability, which served as the final classification of the videos. Additionally, we calculated entropy as a measure of sign unpredictability [37], and determined the total number of classification threshold crossings indicating transitions between facial expressions from "pain" to "no pain", and vice versa. For comparability between pain signs, we used the Pearson correlation coefficient (r). To further analyze the pain patterns represented in the pain signs, we applied the well known k-means unsupervised learning clustering algorithm using as input the proposed entropy and threshold crossing metrics for each video. The optimal number of clusters was determined experimentally by varying the cluster count from two to ten and selecting the configuration with the highest silhouette score [38].

3. Results

Table 2 presents classification results on iCOPE and UNIFESP validation folds. VGG-Face achieved higher classification metrics than N-CNN and ViT-B/16, although when looking at the overall performance, there was no relevant statistical significance in any metric. Nevertheless, the models exhibited higher levels of standard deviation, meaning that these models are highly susceptible to the neonates inside each fold.

After validation, only the top-performing VGG-Face, N-CNN and ViT-B/16 models were chosen to generate the pain sign, specifically the models with the highest AUC across all folds. When assessing the performance of these models on 20-s video recordings condensed into a single label, and consequently one prediction, the metrics in Table 2 indicate a decrease in performance compared to the validation data. For instance, the best AUC obtained by Brahnam et al. [21] was 79.8%, against ours VGG-Face 71.2%, N-CNN 74.0%, and ViT-B/16 68.8%.

Despite the high sensitivity at 80% level, the specificity suggests a high number of false positives. Initially, these results might appear indicative of overfitting during training. However, upon incorporating the temporal dimension into the predictions, we observed that the occurrence of false positives was associated with frames where newborns exhibited facial expressions of pain, despite the video being labeled as "no pain". Moreover, the facial expressions of newborns varied throughout the 20-s of video, contradicting the single label assigned to the data.

When addressing the pain signs, for all models, the optimal number of clusters was three. These clusters were then labeled as stable, irregular, and unstable based on their distinct characteristics. The results, including classification and temporal metrics for each sign type, are summarized in Table 3. Examples of the pain signs are shown in Figs. 1, 2, 3, and 4. In these figures, the colored lines depict the evolution of pain predicted probability over the 20-s duration of the video for each model. The markers indicate the moments when the pain sign crosses the classification threshold specific to each model. Below each sign, video frames are displayed at 1-s intervals, accompanied by their corresponding GC and IG heatmap masks. The legend on top of each figure includes, in order: the ground-truth label of the video, each model's average probability and final prediction, the calculated entropy, and the number of threshold crossings.

Stable signs, illustrated in Fig. 1, exhibited low entropy and few threshold crossings, remaining relatively constant throughout the video. These signs were characterized by high-confidence predictions close to zero, indicating a period of rest, or close to one, suggesting a prolonged painful event. The irregular pain sign, portrayed in Fig. 2, showed higher entropy but few threshold crossings, indicating irregularities in the models' predictions but without sufficient confidence to definitively switch between "pain" or "no pain". These irregularities

Table 3
Classification and temporal metrics for each type of pain sign and model.

Model	Sign type	Accuracy	Sensitivity	Specificity	AUC	Entropy[a]	Threshold crossings[a]	Samples Pain[b]	Samples No Pain[b]
VGG-Face	Stable	70.6%	93.8%	50.0%	70.9%	0.35 ± 0.46	0.16 ± 0.37	32	36
	Irregular	60.6%	69.2%	55.0%	50.8%	1.75 ± 0.64	2.52 ± 0.91	13	20
	Unstable	42.1%	75.0%	33.3%	68.3%	2.56 ± 0.46	6.79 ± 2.30	4	15
N-CNN	Stable	72.2%	89.5%	62.9%	83.6%	0.75 ± 0.66	0.11 ± 0.32	19	35
	Irregular	71.4%	80.0%	63.6%	75.2%	2.31 ± 0.47	2.43 ± 0.89	20	22
	Unstable	45.8%	70.0%	28.6%	50.7%	2.61 ± 0.52	6.75 ± 1.42	10	14
ViT-B/16	Stable	71.9%	89.7%	57.1%	74.9%	1.23 ± 0.79	0.19 ± 0.39	29	35
	Irregular	32.4%	69.2%	9.5%	33.0%	2.64 ± 0.46	3.00 ± 1.13	13	21
	Unstable	50.0%	85.7%	33.3%	61.0%	2.62 ± 0.39	8.00 ± 2.26	7	15

Footnote:
[a] Data are in mean ± standard deviation.
[b] The samples for "Pain" and "No Pain" are based on the original labels of the videos.

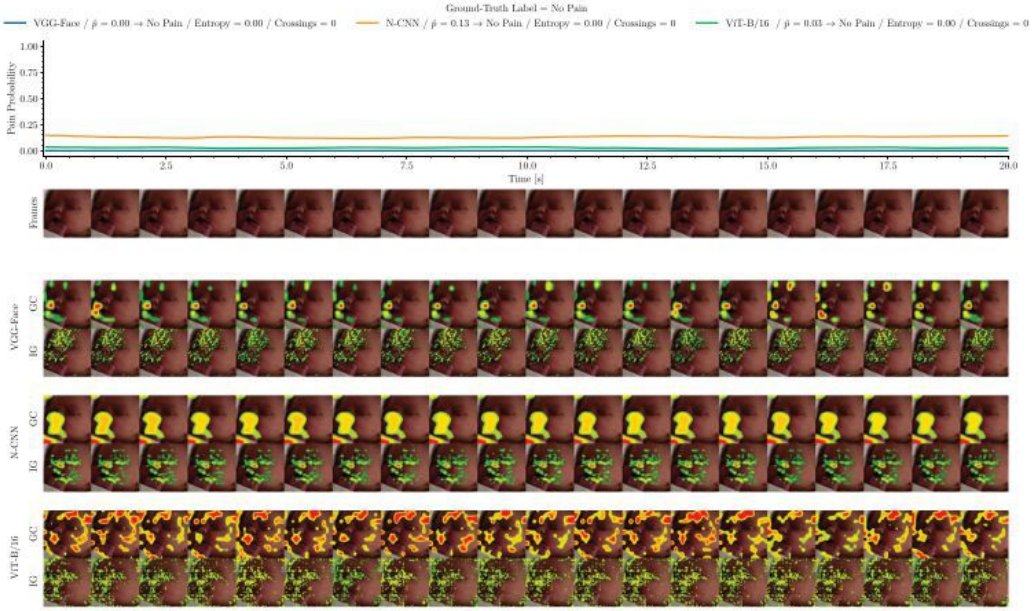


Fig. 1. Examples of stable pain signs.

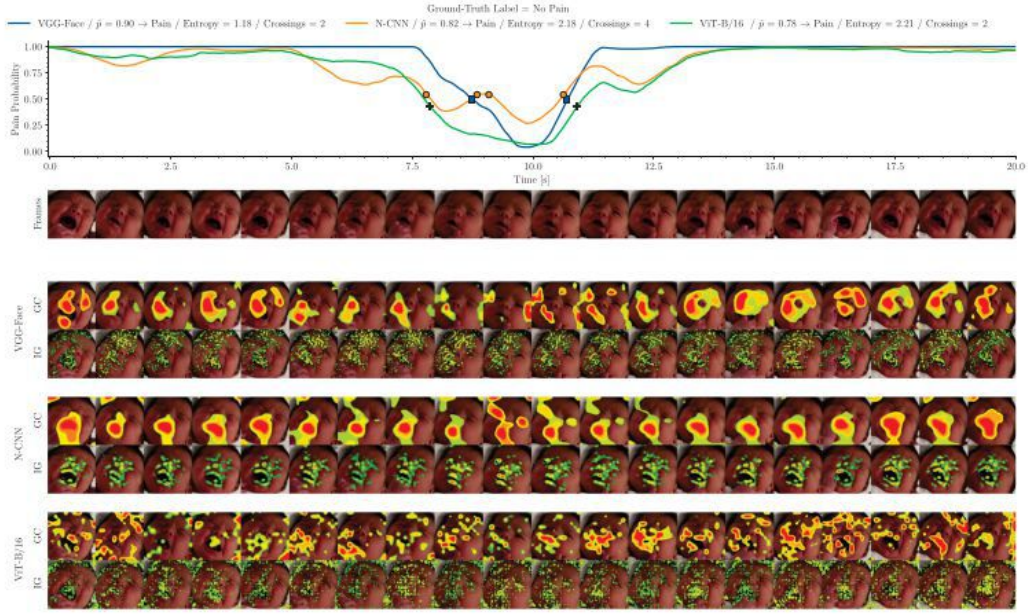


Fig. 2. Examples of irregular pain signs.

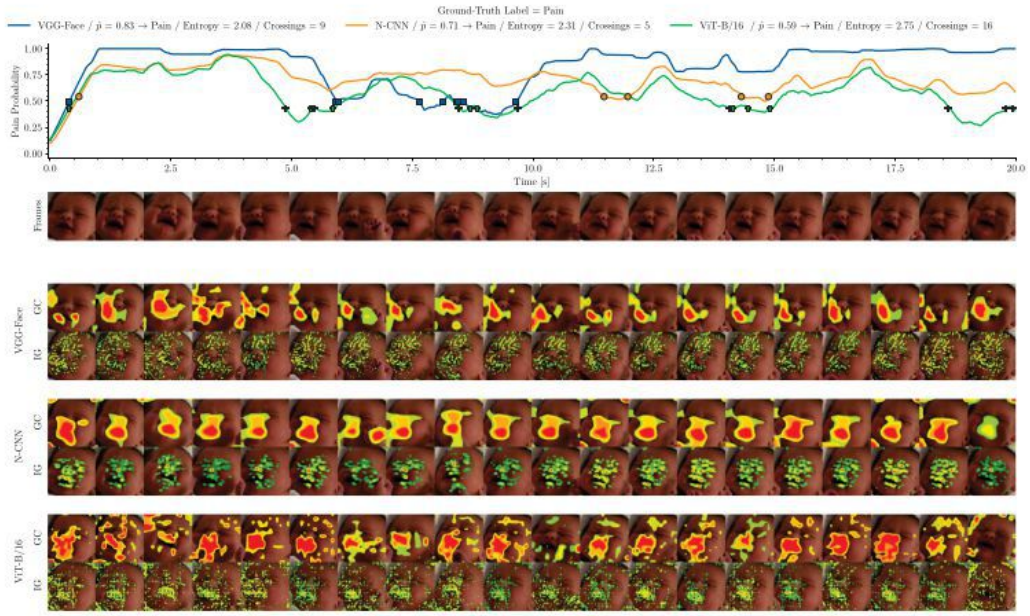


Fig. 3. Examples of unstable pain signs.

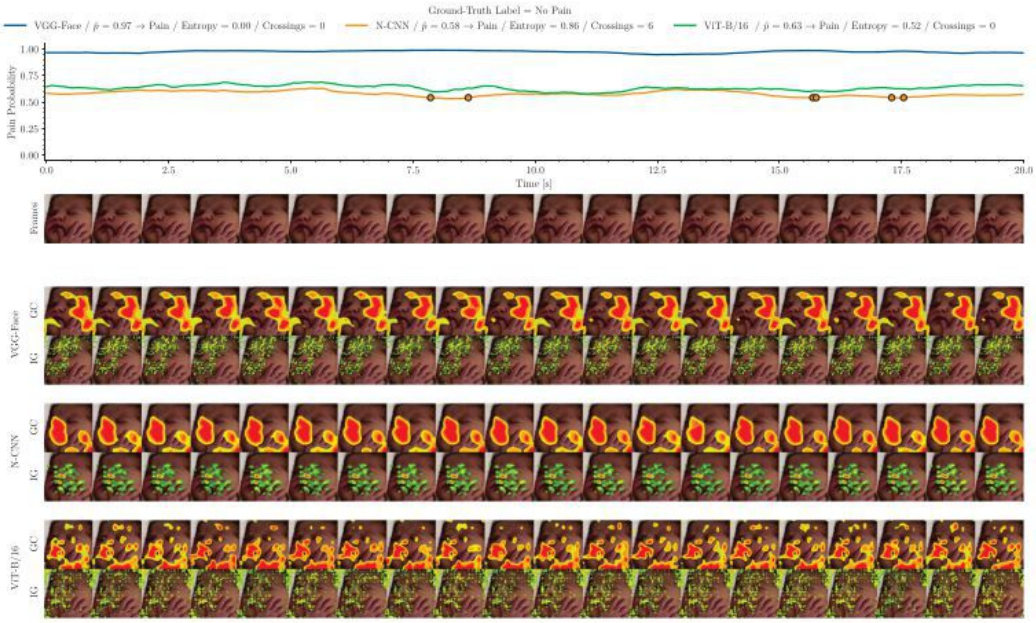
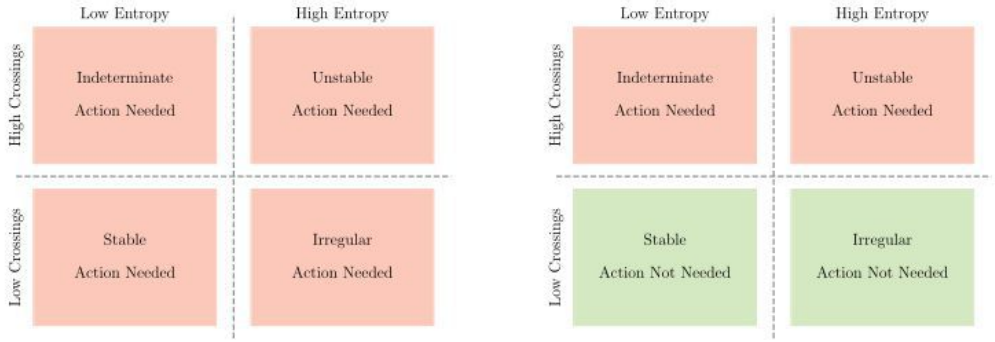


Fig. 4. Example of N-CNN indeterminate pain sign.

were often caused by hands occluding the face or sudden variations in facial expression. The unstable signs in Fig. 3 were defined by high entropy and numerous threshold crossings, where the models detected repeated painful and non-painful events, which could be interpreted as highly dynamic event or prediction instability.

As shown in Table 3, all models performed better when evaluating stable pain signs, reflecting the consistency of facial expressions throughout the video. If the prediction indicates "pain", immediate action by healthcare professionals is required. In contrast, irregular pain signs, characterized by fluctuations that impact the final prediction, reduced the models' predictive power. These signs also demand healthcare intervention, particularly if "pain" is predicted. For more complex pain signs, such as unstable ones, classification metrics declined significantly, as a single prediction or label could not adequately represent the constantly varying facial expressions, necessitating human supervision for both "pain" and "no pain" predictions. We translated these findings into decision matrices showed in Figs. 5(a) and 5(b), which guide healthcare professionals on whether to take action for specific pain sign types. For consistency across models, we used abstract terms such as high/low entropy and high/low threshold crossings in Figs. 5(a) and 5(b), with each sign type's cluster characteristics defined in Table 3. For instance, high entropy for VGG-Face corresponds to values within the intervals for irregular (1.75 ± 0.64) and unstable (2.56 ± 0.46) signs.

In fact, a fourth type of sign is also possible, as observed in the upper left quadrant of the decision matrices in Figs. 5(a) and 5(b). The fourth pain sign, referred to here as "indeterminate", is characterized by a high number of threshold crossings but low entropy, indicating a sign that fluctuates around the classification threshold without a clear designation as "pain" or "no pain". Only one sign from N-CNN met this criterion, as shown in Fig. 4. Due to the limited number of



(a) Pain predictions. (b) No pain predictions.

Fig. 5. Decision matrices for actionable responses for each type of pain sign.

samples for this pain sign type, it was not included in Table 3. Even though the ViT-B/16 pain sign closely resembles that of N-CNN, its lower classification threshold of 43.1%, compared to N-CNN's 54.2%, resulted in no detected threshold crossings, leading to its categorization as stable. Additionally, the mean predicted probability of pain for N-CNN is 58%, which is closer to the classification threshold, making it more indicative of an indeterminate prediction.

In Figs. 1, 2, 3, and 4, we also present the results of the GC and IG explainability methods. For VGG-Face, both GC and IG highlighted the entire face of the neonate as important. However, when the "pain" probability nears zero, GC emphasized peripheral regions of the image, as no face regions were deemed significant for classification. Conversely, for N-CNN, the mouth of the neonate consistently emerged as a region of interest, although GC also identified some regions outside the face as relevant. For ViT-B/16, the GC resulting masks appear more dispersed, due to its transformer-based architecture, which relies on a global attention mechanism rather than the localized operations typical of CNNs. Additionally, the ViT-B/16 used in this study was pre-trained on ImageNet, which lacks domain-specific features like neonatal facial expressions reflecting general patterns from pre-training rather than task-specific focus. However, we observed a tendency for the ViT-B/16 model to highlight the mouth region, especially when predicting pain.

Interestingly, IG remained mostly static for all models, highlighting the same face regions regardless of the predicted probability. Additionally, it is worth noting that GC is highly sensitive to the model's predicted pain probability, unlike IG. This granularity can offer real-time insights for healthcare professionals into why the pain probability fluctuates across different time intervals.

Even with comparable classification performance, the average Pearson correlation between signs was moderate: $r = 0.52 \pm 0.36$ for VGG-Face and N-CNN, $r = 0.49 \pm 0.33$ for VGG-Face and ViT-B/16, and $r = 0.57 \pm 0.37$ for N-CNN and ViT-B/16. Despite this, 60% of the videos were classified under the same sign type across models, as reflected in the number of "pain" and "no pain" samples in Table 3. Furthermore, applying the Friedman test to the entropy and threshold metrics for each model and its corresponding sign type revealed no significant differences, suggesting that the characterization of pain signs is not exclusive of the underlying model architecture.

The models effectively tracked subtle changes in facial expressions, predicting "pain" probabilities in a smooth and consistent manner. These probabilities demonstrated a clear correlation with the intensity of the pain-related facial expressions. Thus, expressions with more pronounced pain characteristics are more likely to be predicted as "pain". Fig. 6 provides a frame-by-frame analysis of all videos in the iCOPEvid database, comparing the ground-truth video labels with the models' frame-level predictions. Frames where the predicted label matched the video label were classified as correct (Fig. 6(a)), whereas mismatched predictions were classified as incorrect (Fig. 6(b)). To enhance interpretability, the facial images were centered to a common reference point and averaged within probability intervals of 10%. From 0% to 50%, there is no evident facial expression of pain, but for probabilities over 50%, an increase in facial expression intensity is clear as the predicted probability rises. This finding aligns with the classification thresholds found by the Youden Index, 54.2% for N-CNN, 49.4% for VGG-Face, and 43.1% for ViT-B/16. Additionally, these figures demonstrate that a single label cannot accurately capture the nuanced facial expressions of neonates within a given time interval. This is evident in the incorrectly classified frames shown in Fig. 6(b), where consistent pain-related facial characteristics appear in "no pain" videos, and vice versa.

4. Discussion

Overall, our study highlights that while previous research on computational neonatal pain assessment has prioritized enhancing model performance using multidimensional data or developing new AI methodologies [21,23,24,26,28,30,39], practical usability in clinical settings by clinicians remains largely overlooked. By aiming to bridge the gap between computational and clinical approaches, our experimental results show the potential of an AI-derived pain sign that is both clinically comprehensible and user-friendly, providing actionable and interpretable information for healthcare professionals.

This work employed two well-established Deep-Learning models in this field, VGG-Face and N-CNN [21,23–30], along with the ViT-B/16 architecture [33], to showcase the versatility and potential of our proposed methodology across different model frameworks. We utilized three distinct benchmark datasets containing images and videos of newborns exposed to both painful and non-painful stimuli for training, validation, and testing. The validation results revealed no statistically significant differences in the classification metrics of the three models. In short, while VGG-Face, N-CNN, and ViT-B/16 achieved high AUC values during validation (93.0% ± 4.3%, 89.0% ± 5.4%, and 89.1% ± 2.9%, respectively), their performance dropped significantly when tested on 20-s video segments (71.2%, 74.0%, and 68.8%, respectively). This decline points to the limitations of assigning a single label to an entire time interval and subsequently generating only one prediction.

Similarly, Manworren et al. [40] observed that using frame-level labels obtained from nurses, significantly improved AUC values to 97%. However, 12% of frames labeled as "pain" were classified as "no pain" by nurses, while 40% of "no pain" frames were labeled as "pain". These findings highlight the inherent variability and subjectivity in pain assessment and support our results, as our models also mirror this disagreement, with lower classification performance. Thus, according to the authors, frame-level ground-truth labels are recommended for training AI models, as not all frames correspond to the single assigned



(a) Correct predictions.



(b) Incorrect predictions.

Fig. 6. Average facial expressions by predicted pain probability and model.

label to the video. Even so, obtaining clinically accurate labels for over 72,000 frames in the iCOPEvid dataset is both labor-intensive and time-consuming for clinical staff.

The results in Fig. 6 highlight the potential of the employed models to automatically generate frame-level labels, thereby facilitating future model training. This is evident from the positive correlation between the intensity of pain-related facial expressions and the predicted pain probability, similarly to the NFCS score [2], where the presence of more facial characteristics indicative of pain increases the likelihood of the neonate being in pain. The models' ability to capture this relationship, despite not being trained for regression tasks, explains visually the pain sign contribution. Consequently, the pain sign continuously monitors frame-by-frame subtle changes in neonatal facial expressions in real-time, offering a visual, time-sensitive explanation of pain probability that assists healthcare professionals in managing NICU workloads.

While similar signs have been identified in adult pain estimation research [41–43], to the best of our knowledge, this study is the first to apply such an analysis to the neonatal domain. Furthermore, we improved the interpretability of these signs by incorporating two metrics – entropy and threshold crossings – that quantitatively defined three distinct and novel sign types: stable, irregular, and unstable. We also introduced a potential fourth type, the indeterminate pain sign. This temporal classification scheme, though separately defined for each model, exhibited consistent characteristics across all models, suggesting that sign types are not exclusive of the underlying AI model architecture. However, each model demonstrated varying interpretive capabilities when analyzing videos from the same time intervals, often leading to different predictions and sign types, resembling the inter-rater variability observed among human evaluators in the neonatal pain assessment task [11–13].

For clinical practice, we translated these computational and mathematical findings into decision matrices, illustrated in Fig. 5, to guide healthcare professionals. The matrices suggest prioritizing actions based on the following observations: (I) Stable signs predicting painful events, where the neonate's facial expression remains consistent, indicating a prolonged painful state; (II) Irregular signs predicting painful events, similar to stable signs but with facial occlusions or sudden variations in expression; (III) Unstable signs, regardless of the final prediction, representing highly dynamic events that the model cannot reliably predict; and (IV) Indeterminate signs, also regardless of the final prediction, due to significant uncertainty in the model's outputs. During clinical practice, the pain sign would automatically be assigned to one of these categories using the clustering algorithm proposed here. Additionally, triggers can be implemented to detect sudden surges or drops in the pain sign, promptly alerting healthcare professionals to changes in the neonate's condition.

However, it is important to highlight that the 20-s duration of iCOPEvid videos, which were recorded in a controlled environment, are not fully representative of real-world NICU scenarios, where neonates often have medical apparatus obscuring their faces and need to be monitored full-time. Our approach can be adapted to longer videos by segmenting them into smaller intervals and analyzing sign behavior over extended periods, such as the past minute, hour, or even longer. Therefore, it might be possible to align our temporal classification scheme with clinical terms of acute, prolonged, persistent, or chronic pain [18].

5. Conclusion

To the best of your knowledge, this retrospective multi-dataset study is the first to implement an AI-derived pain sign for continuous tracking of neonatal facial expressions with a novel set of temporal classifications designed specifically for healthcare professionals usage during clinical practice.

Unlike previous methods that offer single predictions and lack NICU-specific interpretability, our approach provides continuous, real-time insights into neonates' current and historical pain states. Bridging computational and clinical needs, it delivers actionable, user-friendly information with minimal training, supporting timely decisions and improving care in NICU settings. Furthermore, our approach can be extended to other AI models in the pain assessment domain that generate a prediction probability.

We believe that the proposed pain sign has the potential to serve as a vital sign, seamlessly integrating into vital sign monitors for ongoing visualization and monitoring of newborns. By categorizing pain signs into distinct types, healthcare professionals can effectively screen neonates and gain deeper insights into individual pain patterns experienced by newborns, enabling personalized treatment. With the implementation of the pain sign, we can advance towards ensuring optimal pain management and promoting the well-being of these vulnerable patients.

CRediT authorship contribution statement

Leonardo Antunes Ferreira: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lucas Pereira Carlini:** Writing – review & editing, Validation, Methodology, Formal analysis. **Gabriel de Almeida Sá Coutrin:** Writing – review & editing, Validation, Methodology, Formal analysis. **Tatiany Marcondes Heiderich:** Data curation. **Rita de Cássia Xavier Balda:** Project administration, Data curation, Conceptualization. **Marina Carvalho de Moraes Barros:** Project administration, Data curation, Conceptualization. **Ruth Guinsburg:** Writing – review & editing, Project administration, Funding acquisition, Data curation, Conceptualization. **Carlos Eduardo Thomaz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Ethics statement

All data used are benchmark databases already approved by the appropriate institutional committees.

Code availability

All codes are open source and available at https://github.com/leferr-code/article. Trained models can be requested to the corresponding author.

Funding

This work was supported by FAPESP (2018/13076-9) and CAPES (Finance Code 001).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Leonardo Antunes Ferreira reports financial support was provided by Coordination of Higher Education Personnel Improvement. Carlos Eduardo Thomaz reports financial support was provided by State of Sao Paulo Research Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Centro Universitário FEI, and the Brazilian funding agencies FAPESP and CAPES.

Appendix A. Data augmentation

Data augmentation is the process of generating artificially new data from existing data and is commonly used in Deep-Learning to prevent overfitting and regularizing the training procedure [44]. For our purposes, the available face images went through a series of transformations altering its characteristics, including rotation (30°), shear (0.15), width and height shift (0.20), brightness adjustment (0.50–1.1), zoom (0.70–1.5), and horizontal flip. These augmentations resulted in 20 new images for each original image. We ensured the neonate's face was fully detectable by applying the RetinaFace detector [22].

Appendix B. Training protocol

During training, the value of the Cross-Entropy error function was used to determine the convergence or divergence between training and validation. In each epoch of the 100 epochs scheduled for training, if the error value decreased in the validation data, a model checkpoint was saved. Otherwise, if there was no improvement in this error for five consecutive epochs, training was stopped. Only the best model, that is, the one that obtained the lowest error value in the validation data during training was used in the following steps. In the case of the VGG-Face, if the error did not improve for five consecutive epochs, the fine-tuning process started, allowing the training of the last two convolution layers groups. For ViT-B/16, only the classifier head was trained, and N-CNN was trained from scratch.

We also used the DropOut technique during training, for the VGG-Face the value was set at 50%, for N-CNN we followed what Zamzmi et al. [32] proposed, and for ViT-B/16 we used a rate of 10%. For all models, the batch size was fixed on 16, the optimizer used was RMSProp with a learning rate of $\eta = 1 \times 10^{-4}$.

References

[1] K.J. Anand, P.R. Hickey, Pain and its effects in the human neonate and fetus, N. Engl. J. Med. 317 (21) (1987) 1321–1329.
[2] R.E. Grunau, K.D. Craig, Pain expression in neonates: facial action and cry, Pain 28 (3) (1987) 395–410.
[3] K.J. Anand, D.B. Carr, The neuroanatomy, neurophysiology, and neurochemistry of pain, stress, and analgesia in newborns and children, Pediatr. Clin. North Am. 36 (4) (1989) 795–822.
[4] M. Perry, Z. Tan, J. Chen, T. Weidig, W. Xu, X.S. Cong, Neonatal pain: perceptions and current practice, Crit. Care Nurs. Clin. 30 (4) (2018) 549–561.
[5] K.J. Anand, M. Eriksson, E.M. Boyle, A. Avila-Alvarez, R.D. Andersen, K. Sarafidis, T. Polkki, C. Matos, P. Lago, T. Papadouri, et al., Assessment of continuous pain in newborns admitted to NICU's in 18 European countries, Acta Paediatr. 106 (8) (2017) 1248–1259.
[6] R.E. Grunau, Neonatal pain in very preterm infants: long-term effects on brain, neurodevelopment and pain reactivity, Rambam Maimonides Med. J. 4 (4) (2013) e0025.
[7] P. Hummel, M. Puchalski, S. Creech, M. Weiss, Clinical reliability and validity of the N-PASS: neonatal pain, agitation and sedation scale with prolonged pain, J. Perinatol. 28 (1) (2008) 55–60.
[8] J. Lawrence, D. Alcock, P. McGrath, J. Kay, S.B. MacMurray, C. Dulberg, The development of a tool to assess neonatal pain., Neonatal Netw.: NN 12 (6) (1993) 59–66.
[9] D.J. Crellin, D. Harrison, N. Santamaria, F.E. Babl, Systematic review of the face, legs, activity, cry and consolability scale for assessing pain in infants and children: is it reliable, valid, and feasible for use? Pain 156 (11) (2015) 2132–2151.
[10] M. Schiavenato, J.F. Byers, P. Scovanner, J.M. McMahon, Y. Xia, N. Lu, H. He, Neonatal pain facial expression: Evaluating the primal face of pain, Pain 138 (2) (2008) 460–471.
[11] R.X. Balda, R. Guinsburg, M.F.B. de Almeida, C. de Araújo Peres, M.H. Miyoshi, B.I. Kopelman, The recognition of facial expression of pain in full-term newborns by parents and health professionals, Arch. Pediatr. Adolesc. Med. 154 (10) (2000) 1009–1016.
[12] G.D. De Sario, C.R. Haider, K.C. Maita, R.A. Torres-Guzman, O.S. Emam, F.R. Avila, J.P. Garcia, S. Borna, C.J. McLeod, C.J. Bruce, et al., Using AI to detect pain through facial expressions: A review, Bioeng. 10 (5) (2023) 548.
[13] A. Llerena, K. Tran, D. Choudhary, J. Hausmann, D. Goldgof, Y. Sun, S.M. Prescott, Neonatal pain assessment: Do we have the right tools? Front. Pediatr. 10 (2023) 1022751.

[14] M.D. Cruz, A. Fernandes, C. Oliveira, Epidemiology of painful procedures performed in neonates: a systematic review of observational studies, Eur. J. Pain 20 (4) (2016) 489–498.
[15] G. Zamzmi, R. Kasturi, D. Goldgof, R. Zhi, T. Ashmeade, Y. Sun, A review of automated pain assessment in infants: features, classification tasks, and databases, IEEE Rev. Biomed. Eng. 11 (2017) 77–96.
[16] T.M. Heiderich, L.P. Carlini, L.F. Buzuti, R.C. Balda, M.C. Barros, R. Guinsburg, C.E. Thomaz, Face-based automatic pain assessment: challenges and perspectives in neonatal intensive care units, J. Pediatr. (RioJ) 99 (6) (2023) 546–560.
[17] M.S. Salekin, P.R. Mouton, G. Zamzmi, R. Patel, D. Goldgof, M. Kneusel, S.L. Elkins, E. Murray, M.E. Coughlin, D. Maguire, et al., Future roles of artificial intelligence in early pain management of newborns, Paediatr. Neonatal Pain 3 (3) (2021) 134–145.
[18] K.J. Anand, Defining pain in newborns: need for a uniform taxonomy? Acta Paediatr. 106 (9) (2017) 1438–1444.
[19] S. Brahnam, C.-F. Chuang, F.Y. Shih, M.R. Slack, Machine recognition and representation of neonatal facial displays of acute pain, Artif. Intell. Med. 36 (3) (2006) 211–222.
[20] T.M. Heiderich, A.T.F.S. Leslie, R. Guinsburg, Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements, Acta Paediatr. 104 (2) (2015) e63–e69.
[21] S. Brahnam, L. Nanni, S. McMurtrey, A. Lumini, R. Brattin, M. Slack, T. Barrier, Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from Gaussian of local descriptors, Appl. Comput. Inform. 19 (1/2) (2020) 122–143.
[22] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 5202–5211.
[23] C. Li, A. Pourtaherian, L. van Onzenoort, W.E.T.a. Ten, P.H.N. de With, Infant facial expression analysis: Towards a real-time video monitoring system using R-CNN and HMM, IEEE J. Biomed. Heal. Inform. 25 (5) (2021) 1429–1440.
[24] M.S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, Y. Sun, First investigation into the use of deep learning for continuous assessment of neonatal postoperative pain, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 415–419.
[25] L.P. Carlini, L.A. Ferreira, G.A. Coutrin, V.V. Varoto, T.M. Heiderich, R.C. Balda, M.C. Barros, R. Guinsburg, C.E. Thomaz, A convolutional neural network-based mobile application to bedside neonatal pain assessment, in: 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI, IEEE, 2021, pp. 394–401.
[26] M.S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, Y. Sun, Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment, Comput. Biol. Med. 129 (2021) 104150.
[27] G.A. Coutrin, L.P. Carlini, L.A. Ferreira, T.M. Heiderich, R.C. Balda, M.C. Barros, R. Guinsburg, C.E. Thomaz, Convolutional neural networks for newborn pain assessment using face images: A quantitative and qualitative comparison, in: R. Su, Y. Zhang, H. Liu, A. F. Frangi (Eds.), International Conference on Medical Imaging and Computer-Aided Diagnosis, Springer Nature Singapore, Singapore, 2023, pp. 503–513.
[28] Y. Zhao, H. Zhu, X. Chen, F. Luo, M. Li, J. Zhou, S. Chen, Y. Pan, Pose-invariant and occlusion-robust neonatal facial pain assessment, Comput. Biol. Med. 165 (2023) 107462.
[29] L.P. Carlini, G.A. Coutrin, L.A. Ferreira, J. do Carmo Azevedo Soares, G.V.T. Silva, T.M. Heiderich, R.C. Balda, M.C. Barros, R. Guinsburg, C.E. Thomaz, Human vs machine towards neonatal pain assessment: A comprehensive analysis of the facial features extracted by health professionals, parents, and convolutional neural networks, Artif. Intell. Med. 147 (2024) 102724.
[30] J. Hausmann, M.S. Salekin, G. Zamzmi, P.R. Mouton, S. Prescott, T. Ho, Y. Sun, D. Goldgof, Accurate neonatal face detection for improved pain classification in the challenging NICU setting, IEEE Access 12 (2024) 49122–49133.
[31] O. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: BMVC 2015-Proceedings of the British Machine Vision Conference 2015, British Machine Vision Association, 2015.
[32] G. Zamzmi, R. Paul, D. Goldgof, R. Kasturi, Y. Sun, Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN), in: 2019 International Joint Conference on Neural Networks, IJCNN, IEEE, 2019, pp. 1–7.
[33] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[34] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 618–626.
[35] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.
[36] W.J. Youden, Index for rating diagnostic tests, Cancer 3 (1) (1950) 32–35.
[37] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[38] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[39] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, Y. Sun, A comprehensive and context-sensitive neonatal pain assessment using computer vision, IEEE Trans. Affect. Comput. 13 (1) (2022) 28–45.
[40] R.C. Manworren, S. Horner, R. Joseph, P. Dadar, N. Kaduwela, Performance evaluation of a supervised machine learning pain classification model developed by neonatal nurses, Adv. Neonatal Care 24 (3) (2024) 301–310.
[41] J. Zhou, X. Hong, F. Su, G. Zhao, Recurrent convolutional neural network regression for continuous pain intensity estimation in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 84–92.
[42] S. Thuseethan, S. Rajasegarar, J. Yearwood, Deep hybrid spatiotemporal networks for continuous pain intensity estimation, in: Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26, Springer, 2019, pp. 449–461.
[43] D. Erekat, Z. Hammal, M. Siddiqui, H. Dibeklioğlu, Enforcing multilabel consistency for automatic spatio-temporal assessment of shoulder pain intensity, in: Companion Publication of the 2020 International Conference on Multimodal Interaction, in: ICMI '20 Companion, Association for Computing Machinery, New York, NY, USA, 2021, pp. 156–164.
[44] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

APÊNDICE H – ARTIGO PUBLICADO NA CONFERÊNCIA *WORLD CONFERENCE ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI 2025)*



Understanding Disagreement Between Humans and Machines in XAI: Robustness, Fidelity, and Region-Based Explanations in Automatic Neonatal Pain Assessment

Craig Pirie[1(✉)], Leonardo Antunes Ferreira[4], Gabriel de Almeida Sá Coutrin[4], Lucas Pereira Carlini[4], Carlos Francisco Moreno-García[1], Marina Carvalho de Moraes Barros[2], Ruth Guinsburg[2], Carlos Eduardo Thomaz[4], Rafael Nobre[3], and Nirmalie Wiratunga[1]

[1] Robert Gordon University, Aberdeen, U.K.
{c.pirie11,c.moreno-garcia,n.wiratunga}@rgu.ac.uk
[2] Federal University of São Paulo, São Paulo, Brazil
[3] Centro Paula Souza - UPEP, São Paulo, Brazil
[4] Centro Universitário FEI, São Bernardo do Campo, Brazil
{leferr,cet}@fei.edu.br

Abstract. Artificial Intelligence (AI) offers a promising approach to automating neonatal pain assessment, improving consistency and objectivity in clinical decision-making. However, differences between how humans and AI models perceive and explain pain-related features present challenges for adoption. In this study, we introduce a region-based explanation framework that improves interpretability and agreement between XAI methods and human assessments. Alongside this, we present a multi-metric evaluation protocol that jointly considers robustness, faithfulness, and agreement to support informed explainer selection. Applied to neonatal pain classification, our approach reveals several key insights: region-based explanations are more intuitive and stable than pixel-based methods—leading to higher consensus amongst explainer ensembles; both humans and machines focus on central facial features, such as the nose, mouth, and eyes; agreement is higher in "pain" cases than "no-pain" cases likely due to clearer visual cues; and robustness positively correlates with agreement, while higher faithfulness can reduce pixel-level consensus. Our findings highlight the value of region-based evaluation and multi-perspective analysis for improving the transparency and reliability of AI systems in clinical settings. We hope that this framework can support clinicians in better understanding model decisions, enabling more informed trust and integration of AI support in neonatal care.

Keywords: Neonatal Pain Assessment · Human Facial Perception · Explainer Disagreement

© The Author(s) 2026
R. Guidotti et al. (Eds.): xAI 2025, CCIS 2580, pp. 274–298, 2026.
https://doi.org/10.1007/978-3-032-08333-3_13

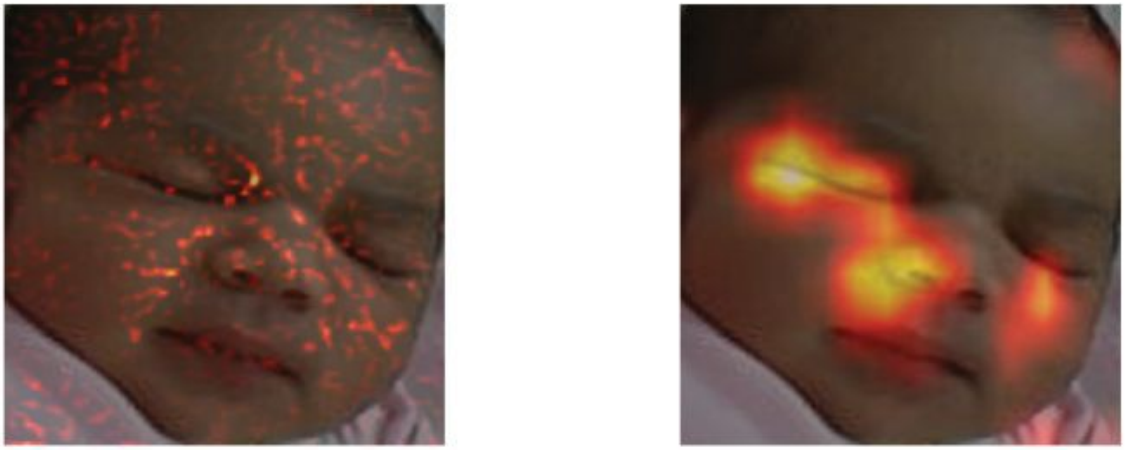
1 Introduction

Neonatal pain detection is critical in paediatric healthcare, particularly within Neonatal Intensive Care Units (NICUs) [3], where infants often undergo an average of 14 painful procedures per day [37]. Accurate pain assessment in neonates is essential not only for immediate comfort but also for safeguarding long-term neuro-development. Evidence suggests that inadequately managed pain in NICUs can lead to altered brain development, heightened stress responses, and persistent neuro-behavioural complications, which may result in increased mortality [42].

Assessing neonatal pain is a complex task, as infants cannot verbally communicate their discomfort, requiring reliance on non-verbal indicators such as facial expressions, body movements, and physiological changes. To support this, over 40 neonatal pain assessment scales have been developed, most of which use facial expression analysis as a primary criterion [41]. Among the most widely used are the Neonatal Facial Coding System (NFCS) [17] and the Neonatal Infant Pain Scale (NIPS) [23]. However, these scales are inherently subjective and prone to inter-rater variability among healthcare professionals [11,24]. They are also time-consuming to apply, both in training and in clinical practice, and offer only discontinuous assessments—pain is evaluated only when a human assessor is present. These limitations underscore the pressing need for objective, consistent, and continuous assessment tools, highlighting the potential of AI-based methods in this domain.

Advances in computational methodologies offer transformative potential in this field. On the one hand, eye-tracking technologies have been employed to explore how healthcare and non-healthcare professionals perceive neonatal facial regions during rest and painful procedures, identifying the most distinctive facial indicators of pain [2,10,16,27,36,38,41]. On the other hand, Artificial Intelligence (AI) approaches, particularly those leveraging facial expression analysis, have demonstrated their ability to overcome the limitations of traditional subjective scales, paving the way for objective and automated pain detection [5,9,13,18,19,32,33].

Although AI has demonstrated significant promise in medical imaging, its "black-box" nature remains a barrier to trust and widespread adoption among healthcare professionals [12]. Post-hoc Explainable AI (XAI) methods seek to address these challenges by elucidating how AI models arrive at their decisions. This transparency is especially crucial in clinical environments where decision-making carries substantial medico-legal and humanistic implications. By disclosing the reasoning process of AI systems, XAI enables clinicians to understand and trust model predictions, fostering a collaborative framework in which AI complements, rather than replaces, human expertise. Nonetheless, the adoption of XAI is undermined by the disagreement problem [21]—where two or more explainers produce different (and often conflicting) insights. For example, in Fig. 1, Grad-CAM highlights the nose and left eye as key regions, whereas Integrated Gradients (IG) distribute pixel importance more diffusely across the face. Such inconsistencies can create uncertainty for clinicians, making it difficult to



(a) Integrated Gradients. (b) Grad-CAM.

Fig. 1. An example case of disagreement between Integrated Gradients and Grad-CAM regarding the most important regions for a prediction within neonatal pain classification.

determine which features are truly relevant and potentially reducing trust in the model's explanations. This presents a need to systematically understand, mitigate and overcome disagreement, to enable the acceptance of ensemble XAI methods in practice.

In this paper, we make two closely linked contributions: (1) a region-based explanation framework that improves interpretability and agreement between XAI methods and human assessments in neonatal pain classification; and (2) a supporting multi-metric evaluation protocol that assesses explanation robustness, faithfulness, and agreement to guide explainer selection in neonatal settings. Our study reveals several important insights: (i) both humans and machines tend to focus on the nose, mouth, and area between the eyebrows in neonatal pain classification; (ii) region-based explanations significantly increase agreement across explainers and between humans and machines, while also yielding more robust attributions; (iii) "pain" cases yield higher agreement than "no-pain" cases, likely due to clearer facial cues; and (iv) higher faithfulness is associated with lower pixel-level agreement, though this effect is less pronounced at the region level.

The remainder of this paper is structured as follows: Sect. 2 reviews related work in neonatal pain analysis and the XAI disagreement problem. Section 3 outlines our methodology, including explanation aggregation and agreement metrics. Section 4 details the experimental setup. Section 5 presents our results which are then discussed in Sect. 6 alongside the limitations of our approach. Finally, Sect. 7 concludes the study, highlighting avenues for future work.

2 Related Works

2.1 Neonatal Pain Analysis

Assessing pain, including in newborns, is inherently subjective for two main reasons: individual variations in pain responses and thresholds, and potential misinterpretation by the assessor [1,14]. Healthcare professionals, for example, may misinterpret pain because of their habituation to frequent exposure to painful situations or the influence of sociocultural and environmental factors [11,24]. This challenge is particularly evident in busy neonatal intensive care units (NICUs), where recognising subtle subjective signs can be difficult.

Research using eye-tracking technology to study how individuals interpret facial expressions as indicators of pain has shown that different observer groups, such as parents and healthcare professionals, tend to focus on similar facial regions, including the nose, mouth, and forehead [38,41]. However, healthcare professionals generally demonstrate greater accuracy in pain detection [2,36]. Additionally, a study by Orsi et al. [27] found that pupil dilation patterns suggest that just two seconds of exposure to a newborn's facial expression are sufficient for an accurate assessment of pain, regardless of medical expertise. These findings point to an intrinsic, shared mechanism in humans for processing facial features related to pain perception.

Numerous AI frameworks have been proposed, achieving notable classification performance [19]. These approaches leverage facial images [9,13,44], videos [5,18,32], and multimodal data [33] to classify neonatal pain. However, most of these frameworks treat AI as black boxes, focusing solely on the final output without examining how the AI reaches its conclusions. A pioneering study by Carlini et al. [9] introduced XAI methods for neonatal pain assessment. This study identified that the model VGG-Face [28] highlighted key facial regions, including the nasolabial furrow, nose contour, and eyes, which align with clinical pain assessment scales. Subsequently, Coutrin et al. [13] applied XAI methods across multiple deep-learning architectures, finding that VGG-Face, pre-trained on facial image datasets, provided the most coherent explanations compared to other models trained on the standard ImageNet dataset. More recently, Carlini et al. [8] observed that, while the identified regions by AI are clinically relevant, they do not entirely agree with the visual attention patterns of healthcare professionals or parents.

2.2 XAI and the Disagreement Problem

Several methods have been developed that generate explanations by attributing importance to pixels. In computer vision, these are usually referred to as "saliency maps". Two prominent methods, particularly in healthcare, include Grad-CAM (GC) [35] and Integrated Gradients (IG) [40].

GC is a widely used explainability method for Convolutional Neural Networks (CNNs) that provides visual explanations for model predictions by highlighting important regions in the input image. GC computes the importance of a feature

map by leveraging the gradients of the model's prediction score with respect to the final convolutional layer's activations. This process highlights regions that positively influence the prediction for a given class, enabling interpretable visualisations of the model's decision-making process.

IG is an attribution method designed to explain the predictions of Neural Networks (NNs) by assigning importance scores to input features based on their contribution to the model's output. IG computes the path integral of the model's gradients along a straight-line path from a baseline input to the input of interest. The method adheres to several axioms including *Sensitivity*, ensuring that non-influential features receive zero attribution, and *Implementation Invariance*, ensuring that the attributions remain unchanged for functionally equivalent models. By integrating gradients over the input space, IG provides a theoretically grounded and robust explanation of feature importance.

Evaluating XAI methods is crucial to ensure that the explanations they provide are meaningful and reliable. Without robust evaluation, there is a risk that explanations might misrepresent the model's decision-making process, leading to misplaced trust or incorrect insights. The Co-12 framework [26] consolidates various desirable explanation properties such as Correctness (faithful representation of model behaviour), Consistency (determinism), and Completeness (coverage of model logic). Sensitivity and Infidelity (proposed in [43] and described by Eqs. 1 & 2 respectively) are two metrics that can be used to evaluate these properties and measure how robust the explanation is under slight variations and how well it represents the inner-workings of the model. We refer to Sensitivity as *robustness* since it reflects stability to input changes, and Infidelity as *faithfulness* as it captures how well explanations align with model outputs.

$$\text{Sensitivity} = \sup_{\|\delta\| \leq \epsilon} \|\mathcal{E}(f, x) - \mathcal{E}(f, x + \delta)\| \tag{1}$$

$$\text{Infidelity}(\mathcal{E}, f, x, \delta) = \mathbb{E}_\delta \left[(\delta^\top \mathcal{E}(f, x) - (f(x) - f(x - \delta)))^2 \right] \tag{2}$$

where: $\mathcal{E}(f, x)$ is the explanation for model f at input x; δ is a perturbation vector drawn from a normal distribution; $f(x) - f(x - \delta)$ is the model's change in prediction due to the perturbation δ; $\mathbb{E}_\delta$ denotes the expectation over the perturbation distribution; and $\delta^\top \mathcal{E}(f, x)$ represents the dot product of the perturbation and the explanation

Another important Co-12 property is Coherence, which can refer to the level of agreement between different explanations. In practice, explainers like IG and GC often provide conflicting outputs for the same input, a challenge referred to as the "disagreement problem". This inconsistency can be confusing not only for clinicians but also for patients (or parents in the neonatal case), especially when AI is used to support high-stakes medical decisions. It raises concerns about "Appropriate Reliance"—the extent to which users trust or distrust AI recommendations in a manner aligned with their accuracy. Prior studies show that detailed but misleading explanations can cause users to over-rely on incorrect predictions, while incoherent or poorly presented explanations may lead them to reject accurate ones [7]. Addressing disagreement and improving coherence

is therefore critical not only for interpretability but also for fostering calibrated trust and safe decision-making in clinical contexts.

Krishna et al. [21] were the first to formally investigate this issue. Their user study revealed that practitioners struggled to reconcile disagreements, often defaulting to subjective heuristics such as favouring the most recent or popular method. To address this, they introduced several metrics to quantify disagreement by measuring the overlap of the top-k features identified by each explainer. These metrics provide insight into how diverging explanations may impact user interpretation and decision-making. Several follow-up works have incorporated these metrics. Roy et al. [31] apply them to filter explanations, retaining only the top-k features with shared attribution signs. Pirie et al. [29] use weighted rankings to aggregate different explanations. In contrast, Schwarzschild et al. [34] propose a "consensus loss" during model training to encourage more aligned explainer outputs downstream.

Laberge et al. [22] further explore the roots of disagreement, suggesting that it arises from different assumptions made by XAI methods about feature interactions. They introduce the concept of regional explanations, using an FD-tree to segment the feature space into regions where the model behaves approximately additively. They report higher levels of agreement and interpretability by applying their approach.

3 Methodology

A key limitation of prior work is its reliance on pixel-level disagreement analysis, which lacks the semantic structure needed for human-interpretable explanations, and the assumption that all explainers are equally reliable. We propose a region-based framework inspired by Laberge et al. [22], which aggregates explanations over semantically meaningful facial regions while presenting multiple explainers alongside objective quality metrics. This supports more interpretable and structured comparisons between explainers, helping end-users identify which explanations to trust while preserving the diversity of insights. In this section, we outline our method consisting of five steps: (1) generating visual explanations for both machine and human decisions; (2) defining and aggregating explanations over semantically meaningful facial regions; and (3–5) applying our multi-metric evaluation protocol, which involves (3) computing feature agreement at pixel and region levels; (4) evaluating robustness and faithfulness using sensitivity and infidelity metrics, and (5) analysing these metrics across human–human, human–machine, and machine–machine perspectives.

3.1 Explanation Generation

Our framework begins by generating visual explanations for both machine and human decision processes.

Machine-Generated Explanations. Saliency maps are produced using post-hoc attribution methods, such as gradient-based or integrated-path techniques. These methods assign importance scores to input features (e.g., pixels in an image) that contribute most to the model's output. Heatmaps are standardised by normalising values and spatially cropping to the relevant region of interest to ensure comparability.

Human-Derived Visual Attention. To capture human evaluators' visual attention patterns toward newborns in painful and non-painful situations, we follow the methodology previously proposed and validated in prior research within this domain [2,10,27,36,38,41]. Eye movements are recorded using the Tobii TX300, a table-mounted eye tracker with infrared illumination and dual cameras integrated into a 23-inch display. The device performs binocular tracking 300 Hz, with 0.6° angular accuracy and 0.04° precision. Calibration and data acquisition are managed using Tobii Studio software.

The stimuli comprise 20 facial images of 10 full-term newborns from the UNIFESP neonatal face dataset [20] (For more information about the dataset, see Sect. 4.3). Each neonate contributes a pair of images: one following a clinically indicated painful procedure ("pain" sample, shown in Fig. 4a), such as vaccination or blood collection, and another during a painless resting state ("no-pain" sample, shown in Fig. 4b). Images are in colour, with upright frontal positioning, and free of obstructions. Only 20 images are selected to avoid excessively long eye-tracking sessions for participants. These images are handpicked by neonatalogists from UNIFESP to represent realistic and clinically relevant scenarios.

Volunteers are positioned in front of the eye-tracker at a distance of 60 cm and briefed on the task in a controlled environment with fixed lighting. The procedure begins with an introductory screen and two practice trials. Each trial starts with a 2-second fixation cross, followed by a randomly positioned neonatal face image displayed for 7 s. Participants then have 3 s to verbally rate the pain on a scale from 0 (no pain) to 10 (severe pain). A fixation cross reappears before the next trial. The numerical ratings provided verbally by the volunteers for both positive ("pain") and negative ("no-pain") facial stimuli are categorised using the scoring system from prior research [2]: scores of 0–2 indicated "no-pain", scores of 3–5 signified "pain" at mild or low-level pain, and scores of 6–10 also represent "pain" but at a moderate or severe level. For our purposes, we only consider the binary labels "no-pain" and "pain".

Eye movement data is processed using Tobii Studio, with fixation defined as two or more samples within a 50-pixel radius. Only participants with $\geq 70\%$ gaze sample captured are included. Heatmaps of fixation durations, generated by Tobii Studio, described accumulated gaze on facial regions. These heatmaps are then averaged across participants within the same group for each facial stimulus.

Both machine and human heatmaps are aligned in scale and cropped consistently to ensure fair comparison in downstream agreement analysis.

3.2 Region Annotation and Explanation Aggregation

To enable semantically meaningful comparison of explanations, we define a set of annotated facial regions that serve as higher-level features. The region masks (Fig. 2) are defined based on 14 neonatal scales that utilise facial action units as indicators of pain [10]. All masks are hand-drawn by a single researcher, under the supervision of a team of medical experts, using Tobii Studio Software and guided by a predefined template to ensure annotation consistency. Each mask segments the facial area into distinct regions (e.g., eyes, mouth, nose, etc.), enabling structured aggregation of saliency information in later steps.

These regions can then be used to aggregate the explanations as described in Eq. 3. The explanation function $\mathcal{E}$ applies an attribution method to a model f and an input image x, generating a saliency map: $\mathcal{E}(f, x) \rightarrow E$. Each pixel, $p \in P$ in the image maps to one of these predefined regions through a mapping function $f : P \rightarrow R, \quad \forall p \in P, \exists r \in R$ such that $f(p) = r$. A heatmap's relevance to a region is determined by aggregating the saliency values of all pixels within that region:

$$E^R = \frac{1}{|R|} \sum_{i \in R} E_i \tag{3}$$

where R is the set of pixels belonging to the region, E_i is the saliency value of pixel i within a region, $|R|$ is the total number of pixels in the region R, and E_R is the region-based explanation.

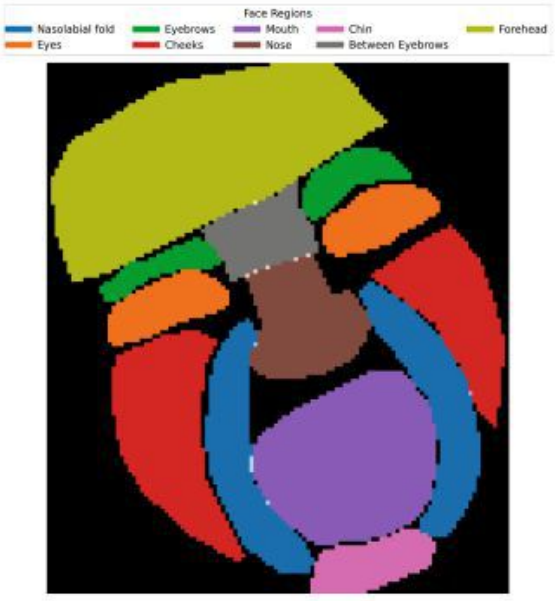


Fig. 2. Example of facial regions derived from human annotations.

3.3 Agreement Measurement

We adopt the feature agreement metric as defined by Krishna et al. [21] to quantitatively compare the heatmaps produced by different explainers and human annotations. This metric focuses on measuring agreement based on the most important features (or pixels in the context of images). The fraction of common elements in the sets of top-k features across two explanations defines their agreement.

Building upon this, we extend the application of the feature agreement metric from individual pixels to annotated regions. Here, regions are treated as macro-level features, enabling us to analyse agreements at a coarser but more interpretable level. These regions are pre-defined through human annotations, such as bounding boxes or segmentation masks, and allow for a domain-specific understanding of importance.

For heatmaps, the top-k most important pixels are extracted based on their saliency scores. Agreement (Eq. 4) is then computed as the overlap of these sets between two heatmaps. This quantifies how similarly two explanations highlight fine-grained areas of the image:

$$\text{FeatureAgreement}(E^a, E^b, k) = \frac{|TF(E^a, k) \cap TF(E^b, k)|}{k} \tag{4}$$

where $TF(E, k)$ returns the set of top-k features of explanation E based on the magnitude of feature importance values. The feature agreement metric remains the same to compare region-level importance. However, in this case, k refers to the number of 'most important' regions to compare, and E_a and E_b are the two region-based explanations where salience values have been aggregated, such that:

$$E^a, E^b \subset \begin{cases} \text{P} & \text{for pixel-based explanations} \\ \text{R} & \text{for region-based explanations} \end{cases}$$

3.4 Robustness and Faithfulness Evaluation

To evaluate the reliability of the generated explanations, we compute two widely used metrics: sensitivity and infidelity, as defined by Yeh et al. [45]. **Sensitivity** (used here as a measure of robustness) quantifies how much an explanation changes in response to small perturbations in the input. Lower sensitivity values indicate more stable and robust explanations. **Infidelity** (used here as a measure of faithfulness) measures how well the explanation captures the actual change in model output due to input perturbations. A lower infidelity score indicates higher alignment with the model's internal reasoning. To generate the perturbations needed for the metrics, we induce a small amount of Gaussian noise with a mean of 0 and standard deviation of 0.1 into every image, to produce 30 test samples each. As the two trained models vary in dimensions, we then scale the metrics by the number of pixels in the image. These metrics provide a principled

way to assess explanation quality beyond visual interpretation, allowing for the identification of explainers that are not only interpretable but also stable and faithful to the underlying model.

3.5 Multi-perspective Comparison

The final step in our framework involves systematically comparing explanations across different sources to assess alignment and identify areas of disagreement. Comparisons are conducted across the following perspectives:

- **Machine–Machine:** Agreement between different explanation methods or different models applied to the same input.
- **Machine–Human:** Comparison of machine-generated saliency maps with human-derived visual attention heatmaps.
- **Human–Human:** Evaluation of agreement between human groups with different backgrounds or expertise.

These comparisons are conducted at both the pixel-level and region-level to capture fine-grained and conceptually-structured differences, respectively. This dual approach enables the analysis of whether disagreements arise from local variation or from broader conceptual differences in how features are interpreted.

By combining region-based aggregation, agreement metrics, and quality evaluations across multiple sources of explanation, this framework provides a principled approach for understanding and interpreting disagreement in AI decision-making. The following section applies this framework to a real-world neonatal pain classification task involving both human and machine explanations.

4 Experiments

This section details the experimental setup used to apply and evaluate our proposed methodology. We describe the participant groups, deep learning models, and datasets used, followed by implementation specifics for training and explanation generation. All experimental procedures involving human data were approved by the Ethics Committee for Research at the Federal University of São Paulo (UNIFESP), under protocols 1299/09 and 3116/14.

4.1 Subjects

A total of 73 participants took part in the study, divided into two distinct groups. The first group named "experts" comprised 44 healthcare professionals, including 4 paediatricians and 40 neonatologists (mean age: 33.48 $\pm$ 7.01 years). The second group consisted of 29 parents (mean age: 30.48 $\pm$ 6.95 years) of neonates hospitalised in the Neonatal Unit of Hospital of São Paulo—Escola Paulista de Medicina—UNIFESP.



(a) Pain.



(b) No pain.

Fig. 3. iCOPE facial images.

4.2 Deep-Learning Models

We utilised benchmark CNN architectures that are commonly used in neonatal pain assessment research [5,8,9,13,18,32,33]:

- VGG-Face (VGG): Originally introduced by Parkhi et al. [28], this CNN was adapted for neonatal pain assessment by incorporating two additional fully connected layers, each with 512 neurons, while retaining the pre-trained convolutional layers for transfer learning [9];
- Neonatal Convolutional Neural Network (NCNN): First proposed by Zamzmi et al. [44], this was specifically designed for neonatal pain analysis. No pre-trained NCNN was available, thus it needs to be trained from scratch.

4.3 Datasets

We utilised two benchmark face image datasets, where each neonate's face was extracted using the RetinaFace detector [15]:

- infant Classification Of Pain Expression (iCOPE) (Fig. 3): Brahnam et al. [4] developed this dataset at the St. John Hospital (now Mercy Hospital) at the Neonatology Department in Missouri, USA. It comprises 200 high-resolution images (3008 × 2000 pixels) captured from 26 Caucasian neonates aged between 18 h and 3 days. The neonates were photographed during sessions involving four distinct stimuli: (1) transfer between cribs, (2) exposure to an air stimulus, (3) heel friction, and (4) heel puncture for blood collection. The dataset includes 63 images of neonates at rest ("no-pain"), 18 crying, 23 during air stimulation, 36 during friction, and 60 during painful procedures ("pain"). For this study, we focused exclusively on images labelled as "no-pain" and "pain";
- UNIFESP (Fig. 4): Developed by Heiderich et al. [20] at the Federal University of São Paulo, Brazil, this dataset comprises 30 healthy neonates aged 24 to 168 h. For each neonate, 12 images (resolution: 320 × 233 pixels)

were collected before, during, and after clinically indicated painful procedures performed by neonatologists. All photographs were captured in 2014 under maternal consent, and no subsequent photos were taken after the experiment. Experienced neonatologists from NICUs independently evaluated the images, resulting in 164 being classified as "pain" and 196 as "no-pain".



(a) Pain. (b) No pain.

Fig. 4. UNIFESP facial images.

4.4 Training Protocol

To mitigate the risk of data leakage and rigorously evaluate the models' generalisation to unseen neonates, we employed a leave-some-subjects-out cross-validation strategy [13], also known as leave-P-groups-out. In this approach, the iCOPE and UNIFESP datasets were merged and then partitioned across ten folds, ensuring that images from the same neonate were exclusively assigned to either the training or testing set, thereby preventing overlap and preserving the independence of evaluation.

We applied data augmentation only to the training data after assigning subjects to each cross-validation fold. A total of 20 augmented training images were generated through a series of standard transformations, including random rotations (up to 30°), shear (0.15), width and height shifts (0.20), brightness adjustments (0.50–1.1), zoom (0.70–1.5), and horizontal flips. To ensure the integrity of the augmented dataset, we utilised the RetinaFace [15] model to validate that the neonate's entire face was present in each image, replacing any that failed verification with a new augmented image.

4.5 Producing and Assessing Explanations

We utilised two XAI methods: (1) GC [35] and (2) IG [40] since our literature review found that these were commonly used in medical applications. These

were implemented in Python following guidance from their respective papers[1]. For GC, we selected the feature maps from the last convolutional layer of both VGG and NCNN, as these layers are expected to capture the most relevant and discriminative spatial information for classification [35]. For IG, we used the standard black image baseline (an all-zero embedding) with 50 interpolation steps. Explanations for the predicted class were generated using both explainers. These were then aggregated to produce region-level explanations and were evaluated as per the methodology (Fig. 5). For pixel-level agreement, the top 10% of pixels were selected ($k = 10\%$), while region-level agreement focused on the top three facial regions ($k = 3$). This thresholding ensured a comparable number of features across both levels, balancing the inclusion of meaningful explanatory content with the exclusion of low-salience noise.

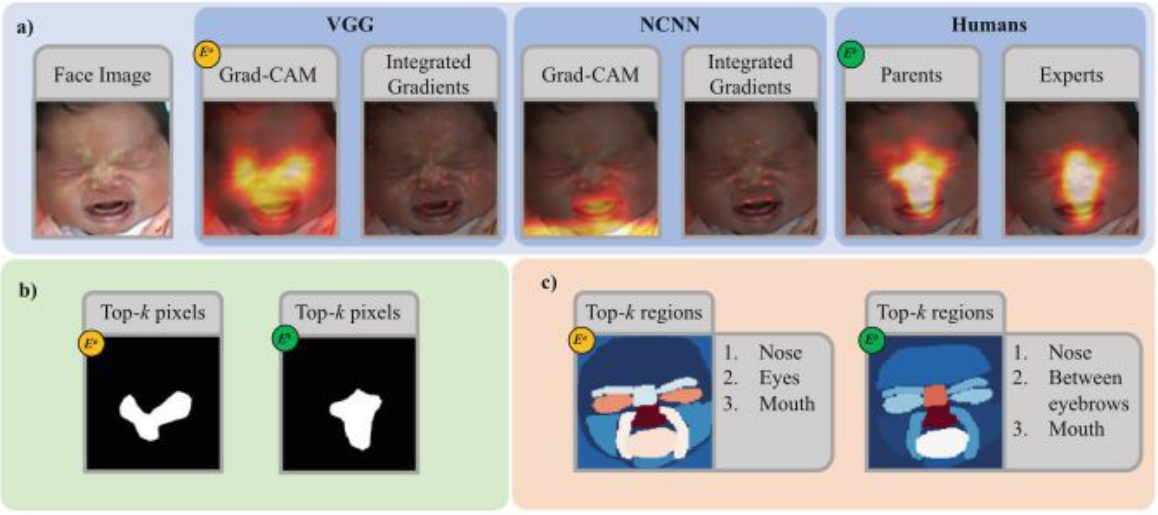


Fig. 5. Agreement experiment. (a) Pairwise comparisons were made based on features extracted by machines and humans; (b) pixel-level agreement of the top-k pixels; (c) region-based considering the top-k agreed regions. The yellow dot represents one explanation (E^a) and the green dot represents the other (E^b). (Color figure online)

5 Results

5.1 Human and Machine Classification Performance

The binary classification performance ("no-pain" or "pain") of both machine and human models was evaluated on the standardised set of 20 images from UNIFESP using multiple metrics, including accuracy, precision, recall, F1-score, and AUC (for machine models). As shown in Table 1, human experts achieved the highest classification accuracy (92.3% ± 7%) with superior recall (96.0% ±

[1] The code for the experiments can be found at https://github.com/leferr-code/XAI-NeonatalPain.

8%) and F1-score (92.8% ± 6%), highlighting their ability to correctly identify neonatal pain with minimal false negatives. Parents also demonstrated strong performance, with an accuracy of 86.5% ± 9% and a balanced precision-recall tradeoff (87.5% ± 14% precision and 89.6% ± 13% recall), indicating their capacity to recognise pain cues in neonates effectively, though with slightly higher variability compared to experts.

Among the machine models, VGG outperformed NCNN in all metrics, due to its pre-training in facial images [13]. VGG achieved an accuracy of 86.2% ± 7%, closely matching that of parents, and exhibited strong recall (90.3% ± 9%), suggesting that it effectively captures neonatal facial expressions associated with pain. In contrast, NCNN had lower accuracy (77.1% ± 7%) and a lower AUC (76.0% ± 7%), indicating reduced reliability in distinguishing pain-related features. However, both machine models showed higher recall than precision, suggesting a tendency to classify more instances as pain, which may be beneficial for minimising missed pain cases but could also increase false positives.

Table 1. Average classification performance of each machine and human model.

Model	Accuracy	Precision	Recall	F1 score	AUC
VGG	86.2% ± 7%	85.9% ± 7%	90.3% ± 9%	87.7% ± 6%	85.4% ± 8%
NCNN	77.1% ± 7%	74.6% ± 8%	89.0% ± 9%	80.8% ± 6%	76.0% ± 7%
Experts	92.3% ± 7%	90.9% ± 10%	96.0% ± 8%	92.8% ± 6%	–
Parents	86.5% ± 9%	87.5% ± 14%	89.6% ± 13%	87.0% ± 9%	–

5.2 Explainer Insights

Across all 20 images investigated, both experts and parents consistently identified the nose as a primary feature, followed by the region between the eyebrows and the mouth (see Table 2). This suggests a shared human visual perception focused on central facial regions, reflecting a holistic visual strategy that persists regardless of the infant's pain state and medical expertise of the evaluators. Similarly, XAI methods highlighted the nose and mouth, particularly in "pain" classifications (Table 2), reinforcing their importance in pain assessment. However, for "no-pain" cases (Table 2), explanations often emphasised peripheral or unexpected regions reducing the number of agreed regions. This could be due to the binary nature of model training, where a single Sigmoid-activated neuron generates a probability for the "pain" class. Consequently, for low pain probabilities (<50%), explanations focus on features that differentiate "pain" rather than distinct indicators of the absence of pain.

Table 3 illustrates the agreement levels across machine-machine, machine-human, and human-human comparisons, further highlighting key areas of consensus and divergence. The nose exhibited the highest agreement across all three

Table 2. Total sum of top-k most important regions for both **Pain** and **No-Pain** classes across 20 images. Each cell indicates the combined count (*Pain*/*No-Pain*). **V** = VGG, **N** = NCNN, **I** = Integrated Gradients, **G** = Grad-CAM, **E** = Experts, **P** = Parents. Values in **bold** are the most important regions for both classes, as identified by a pair of explainers.

Region	VG-NG	VG-NI	VG-VI	VI-NI	VI-NG	NG-NI	VG-E	VG-P	VI-E	VI-P	NG-E	NG-P	NI-E	NI-P	E-P	Total
Background	2(0/2)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2(0/2)
Between Eyebrows	–	1(1/0)	2(2/0)	5(1/4)	–	–	4(4/0)	4(4/0)	14(6/8)	14(5/9)	–	–	5(2/3)	5(2/3)	16(8/8)	70(35/35)
Cheeks	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0(0/0)
Chin	2(0/2)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2(0/2)
Eyebrows	–	3(1/2)	2(0/2)	3(1/2)	–	–	–	1(1/0)	–	–	–	–	–	1(1/0)	–	10(4/6)
Eyes	1(0/1)	**9(6/3)**	–	–	–	1(1/0)	2(1/1)	4(1/3)	–	–	1(0/1)	2(1/1)	2(1/1)	5(2/3)	3(1/2)	30(14/16)
Forehead	–	–	2(1/1)	–	–	–	–	–	–	–	–	–	–	–	–	2(1/1)
Mouth	**7(4/3)**	7(4/3)	2(2/0)	5(5/0)	**5(5/0)**	**15(9/6)**	5(3/2)	6(4/2)	4(4/0)	5(5/0)	**15(8/7)**	**13(8/5)**	**17(8/9)**	14(8/6)	15(8/7)	135(85/50)
Nasolabial fold	–	–	1(0/1)	–	3(1/2)	–	–	–	–	–	–	–	–	–	–	4(1/3)
Nose	3(2/1)	**9(7/2)**	**8(7/1)**	**14(7/7)**	4(2/2)	5(2/3)	**10(8/2)**	**10(8/2)**	**16(9/7)**	**16(9/7)**	5(2/3)	5(2/3)	**17(9/8)**	**17(9/8)**	**20(10/10)**	**159(93/66)**

groups, reaching 36% for machine-machine, 60% for machine-human, and 100% for human-human agreement, confirming its central role in neonatal pain perception. The mouth also showed strong agreement, particularly between humans (75%), and to a lesser extent between machines (34%) and machine-human comparisons (49%). In contrast, the region between the eyebrows demonstrated substantial agreement between humans (80%) but lower agreement with machines (7–29%), suggesting that while humans perceive this as important, machine explainers often overlook it. Peripheral regions, such as the chin, forehead, and cheeks, had minimal to no agreement, indicating their limited relevance in pain classification. These results emphasise the systematic patterns of alignment between human perception and AI explanations, particularly in distinguishing pain-related facial features.

Table 3. Agreement percentages between machines and humans for different facial regions. Colour intensity represents the percentage values.

	Machine-Machine	Machine-Humans	Humans-Humans
Background	2%	0%	0%
Between Eyebrows	7%	29%	80%
Cheeks	0%	0%	0%
Chin	2%	0%	0%
Eyebrows	7%	1%	0%
Eyes	9%	10%	15%
Forehead	2%	0%	0%
Mouth	34%	49%	75%
Nasolabial Fold	3%	0%	0%
Nose	36%	60%	100%

5.3 Agreement

Overall, our results demonstrate that region-based explanations reach higher levels of agreement than pixel-based variants, as seen in Fig. 6. This can be due to a simplification of the explanation problem. Aggregation of saliency values over regions reduces the dimensionality of the problem, therefore increasing the likelihood of an explanation pair reaching a consensus on the most important features. Further, as noted in [22] different explainers handle implicit feature interdependencies differently, which can lead to disagreement. By explaining regions, we make these relationships explicit, reducing semantic discrepancies.

In general, we observe higher agreement levels for the "pain" class compared to the "no-pain" class across all explainer pairs. This can be attributed to the well-defined nature of the "pain" class, which makes it easier for explainers to identify and highlight the key features contributing to the model's predictions. The "pain" class is characterised by distinct and prototypical features, such as specific facial expressions (e.g., widened mouth, squeezed eyes, tense tongue) or physiological cues (e.g., increased heart rate or variations in oxygen saturation). These features are consistent and directly related to the concept of pain, providing a clear and interpretable signal for explainers to focus on.

In contrast, the "no-pain" class is more ambiguous, often defined by the absence of such features rather than the presence of distinct patterns. This lack of clear and consistent attributes creates a challenge for explainers, as the model's decision for "no-pain" may be influenced by a more diffuse set of features or the relative absence of features indicative of pain. The ambiguity of the "no-pain" class can lead to greater variability in the highlighted regions, resulting in lower agreement across explainer pairs.

Consequently, the clear prototypical features of the "pain" class not only make it easier for explainers to identify the most important features but also ensure greater consistency in their outputs, whereas the lack of such clarity in the "no-pain" class introduces variability and disagreement.

Table 4 illustrates that humans, regardless of medical training, predominantly focus on the central facial region, including the nose, the area between the eyebrows, and the mouth. Notably, this visual pattern remains consistent across both "pain" and "no pain" conditions. This consistency led to high levels of agreement, with average values across classes reaching 70% ± 6% for pixel-level analysis and 90% ± 16% for region-based analysis (Fig. 6), underscoring the robust alignment in human visual attention.

VGG-GC and NCNN-IG demonstrated the highest agreement levels among machine-machine pairs for both pixel-based and region-based analyses, with averages of 8% ± 5% and 48% ± 30%, respectively. Table 4 highlights the distinct operational mechanisms of these methods: GC emphasises salient "regions", while IG focuses on fine-grained pixel-level details. However, this disparity is substantially mitigated by our region-based approach, which achieved an average 40%-point increase in agreement values, underscoring the importance of facial features such as the eyes, mouth, and nose to the final classification. These results are particularly notable given that VGG achieved the highest classifica-

Table 4. Visualisation of the agreement for a sample neonate across different explainers, classes (pain/no-pain) and explanation type (pixel-/region-level). Warmer colours indicate the most important facial regions, while cooler colours represent less important features.

Type	Class	VGG-GC	VGG-IG	NCNN-GC	NCNN-IG	Experts	Parents
Pixel	Pain						
	No-Pain						
Region	Pain						
	No-Pain						

tion accuracy among the machine models. Interestingly, this suggests the model may have relied on less intuitive features, such as facial texture or contour, rather than central regions like the nose and mouth that humans typically focus on, highlighting that high predictive performance does not necessarily equate to human-aligned reasoning. GC likely paired best with VGG because it depends on well-structured spatial activations from the final convolutional layers to generate class-relevant heatmaps. VGG, with its deep and pretrained architecture, provides rich and consistent spatial feature maps that support this process. In contrast, IG aligned more closely with human attention when used with NCNN, possibly because its simpler, task-specific architecture allowed for more direct and localised attributions. These findings suggest that the quality and interpretability of explanations may depend not only on the model or explainer individually, but also on the compatibility between the two.

NCNN-IG achieved the highest agreement with human observers at pixel-level (14% ± 2%) and region-level (70% ± 17%), including both parents and experts. While IG heatmaps are visually challenging to compare with other types of heatmaps, our quantitative analyses revealed that the highlighted regions, such as the mouth and nose, align closely with areas of human focus. Moreover, IG consistently emphasised the same facial regions—regardless of whether the classification was "pain" or "no-pain"—further demonstrating its tendency to identify stable and key features across conditions.

Among the least agreeable pairs, VGG-IG and NCNN-GC exhibited the lowest agreement levels, with 4% ± 3% at the pixel-level and 20% ± 20% at the region-level. Interestingly, this result is the inverse of the highest-agreement pair, VGG-GC and NCNN-IG, highlighting a notable contrast in the explainers' ability to capture shared features. This discrepancy is further evident when comparing the same model with different explainers. For instance, VGG-GC and VGG-IG achieved only 6% ± 2% pixel-level and 30% ± 26% region-level agreement, underscoring the disagreement problem in these XAI methods.

NCNN-GC exhibited the lowest agreement metrics with human observers, including both parents and experts, among machine-human pairs. This discrepancy underscores, again, the disagreement between GC and IG, as the regions highlighted by NCNN-GC diverge significantly from those identified by NCNN-IG when compared against the same human observers. NCNN-GC predominantly emphasised the lower regions of the images, including the background, chin, and mouth, being the mouth region predominantly focused in "pain" images.

5.4 Robustness and Infidelity

Table 5 outlines the stability of the explainers with respect to changes in: 1) the data; and 2) the model. The average robustness (Eq. 1) and infidelity (Eq. 2) over all explanations in the sample set is reported per explainer and model. The results suggest that IG is less sensitive to small changes in the data and is more faithful to the black-box model than GC; regardless of the type of underlying model used. NCNN exhibits a higher level of faithfulness on average than VGG, yet is more sensitive.

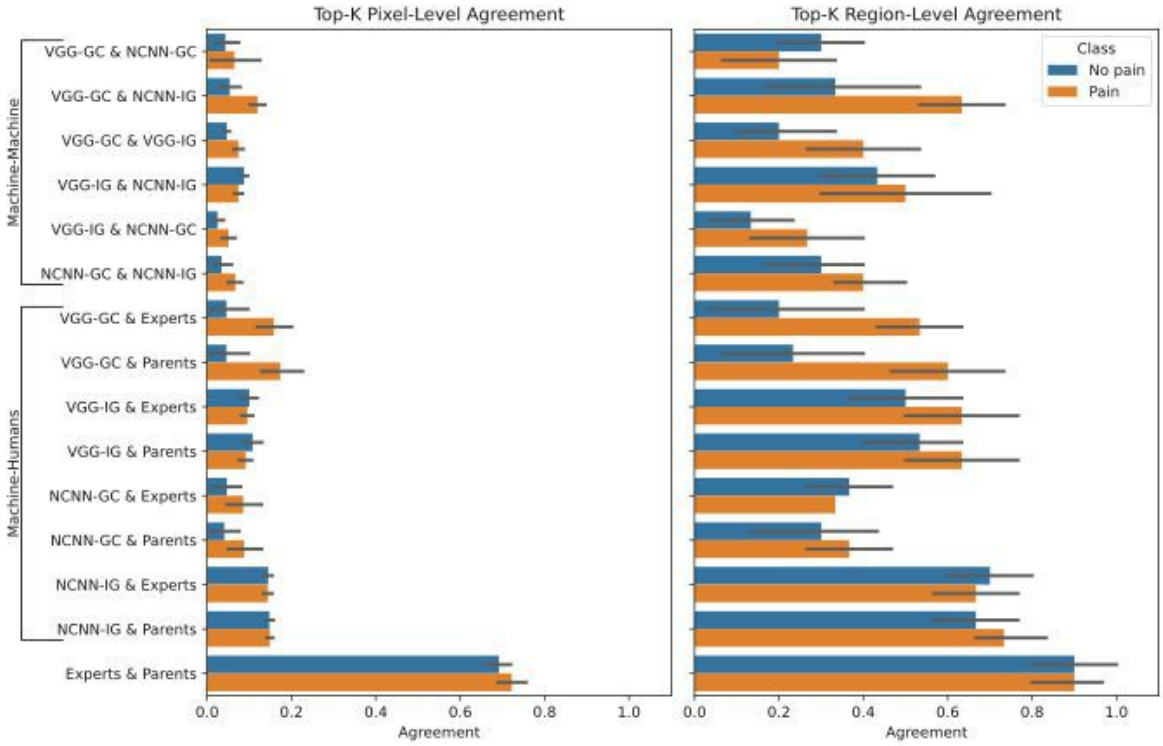


Fig. 6. Average top-k agreement at pixel-level and region-level for the pain (orange) and no-pain (blue) classes. (Color figure online)

Our results in Table 6 indicate that robustness exhibits a weak positive correlation with both pixel-level (0.2060) and region-level (0.0892) agreement, suggesting that as explanations become more stable under perturbations, different explainers tend to agree slightly more. The stronger correlation at the pixel-level implies that explainers producing highly localised attributions are more sensitive to robustness, whereas region-based explanations demonstrate greater inherent stability. This reinforces the notion that regional explanations provide a more intuitive and interpretable representation of model behavior, as they aggregate importance over broader facial features rather than being influenced by small perturbations in individual pixels.

Conversely, infidelity demonstrates a moderate negative correlation with pixel-level agreement (-0.3891) and a weaker negative correlation with region-level agreement (-0.0776). This suggests that as explanations become more faithful to the model's true reasoning, pixel-level agreement decreases, highlighting the potential for misleading consensus in pixel-based attributions. High agreement at the pixel-level may not necessarily indicate a more reliable explanation but could instead reflect a tendency for explainers to converge on superficial but non-causal features. Region-based explanations, however, appear to be more resistant to this effect, as their agreement is less impacted by changes in fidelity, further supporting their suitability for clinical applications where interpretability and stability are critical.

Table 5. Averaged Robustness and Infidelity Results. Models are most faithful and robust as values $\rightarrow 0$. Values in **bold** indicate the best score.

Model	Explainer	Robustness ↓	Infidelity ↓
NCNN	GC	1.56×10^{-6}	4.67×10^{-4}
NCNN	IG	8.69×10^{-8}	$\mathbf{2.00 \times 10^{-5}}$
VGG	GC	2.67×10^{-7}	9.08×10^{-4}
VGG	IG	$\mathbf{6.79 \times 10^{-8}}$	3.40×10^{-5}

Table 6. Pearson Correlation (ρ) between Robustness and Infidelity with Agreement metrics. $\rho > 0$ indicates a positive correlation with agreement, whereas $\rho < 0$ indicates an inverse relationship. $\rho \approx 0$ suggests no correlation.

Metric	Pixel Agreement	Region Agreement
Robustness	0.2060	0.0892
Infidelity	−0.3891	−0.0776

6 Discussion

Our findings demonstrate that the proposed methodology is applicable to both machine-generated explanations and human visual pattern data. We implemented two complementary approaches to measure agreement: a pixel-based analysis, which quantifies numerical similarity without semantic context, and a region-based analysis, which segments the face into anatomically relevant regions associated with neonatal pain. The latter approach provides a more intuitive representation, where analysing facial regions rather than individual pixels better aligns with human perception, easing comparisons of facial feature importance.

Yet, a clear distinction emerges in the features extracted by machine models and their respective explainers. While model architecture and pre-training contribute to differences in the features being explained, our findings also revealed notably low agreement levels between explainers applied to the same model. This underscores that machine-extracted features are not universally consistent, owing to methodological variations in XAI methods, thus the disagreement problem. GC typically emphasised broader, salient regions, whereas IG focused on fine-grained, pixel-level importance. This demonstrates how the use of a single XAI approach can be deceiving to the end user. Using multiple methods and measuring their agreement, as proposed here, helps identify the most important and consistent features.

Our region-based approach addresses a key limitation in XAI methods: the interpretability of their results. Presenting both GC and IG heatmaps to healthcare professionals could lead to confusion and uncertainty about the differences between them and which to trust. By segmenting and annotating facial regions with their relative importance, our region-based approach enables healthcare professionals to better understand the specific facial areas analysed by the AI

model. This facilitates more confident and informed decision-making, as the focus shifts from the visualisation differences of XAI methods to the regions that are most relevant to the model's assessment. Ultimately, this approach ensures that medical staff can rely on the AI's outputs without being hindered by inconsistencies in the visualisation procedures of XAI methods.

These findings underscore the importance of evaluating multiple XAI properties when assessing explanation quality, as high agreement does not inherently imply better fidelity. In the context of neonatal pain assessment, where clinical decision-making relies on trustworthy and stable explanations, region-based approaches should be prioritised over pixel-level attributions. The reduced sensitivity of region-based agreement to robustness and fidelity changes allows for a higher level of consensus to be reached, suggesting that such methods can offer a more reliable and clinically meaningful assessment of pain-related facial features.

Beyond technical advancements, our findings hold direct implications for clinical practice. The identified key facial features (e.g., nose, eyes, mouth) can help guide improved neonatal pain assessments. For instance, trainee clinicians can be instructed to focus on central facial regions during evaluations and ensure these areas are unobstructed in both manual and automated assessments. Our region-based method also holds potential to enhance the safety of automated decisions by indicating whether a prediction is based on clinically relevant characteristics. Furthermore, automated pain detection alleviates the burden on medical personnel, who cannot continuously monitor every neonate in the NICU, supporting more accurate and timely pain recognition and improving neonatal care.

6.1 Limitations

This study has some limitations that should be acknowledged. First, although the proposed region-based agreement approach improves interpretability, it has not been tested in real clinical settings. How healthcare professionals perceive, trust, and apply these explanations in high-stakes environments like NICUs remains to be explored. Second, the availability of high-quality, annotated neonatal datasets is a significant constraint. Due to the ethical and logistical complexities of data collection in NICUs, datasets are scarce, and access is often restricted. For instance, the UNIFESP dataset is not publicly available due to parental consent requirements and ethical regulations; to comply with these restrictions, we shared only heatmap masks generated from eye-tracking and XAI methods between institutions. Third, while region-based explanations improved agreement between humans and models, they do not eliminate inconsistencies across different XAI methods. The persistent variation in attribution highlights the need for further refinement of explanation techniques to better align AI outputs with human understanding. Lastly, in our experiments, we relied on human-engineered regions for aggregating explanations. In practice, these would be infeasible to obtain, and further work should explore incorporating machine-generated concepts (such as [39]) into the pipeline to expedite the process.

7 Conclusion

In this study, we investigated both low-level (pixel) and higher-order (facial-region) features extracted by deep learning models and human evaluators when assessing pain in neonatal images. By comparing human visual analysis patterns with XAI methods—and examining differences among XAI techniques—we highlighted the disagreement problem. However, by introducing a region-based agreement framework, we provided an interpretable approach to understanding and increasing the alignment between human expertise and machine-extracted features. Our findings show that human evaluators primarily focus on central facial features, such as the nose and mouth, while XAI methods demonstrate varying degrees of agreement depending on the technique and model architecture. Nonetheless, both humans and machines consistently identify the mouth and nose as the most important features for neonatal pain assessment.

We demonstrated that region-level analyses significantly enhance interpretability and mitigate the visualisation and comprehension challenges inherent in pixel-level XAI approaches. Our findings show that explanations generated through this approach are also more faithful to the black-box model and less sensitive to noise. This advancement may support more transparent and trustworthy AI applications in neonatal care. Specifically, the region-based agreement framework could be integrated into real-time clinical decision support systems, providing clinicians with immediate feedback on agreed facial regions between XAI methods and thus helping determine the reliability of AI predictions.

7.1 Future Work

In future work, we aim to expand our dataset by collecting additional eye-tracking data from individuals of diverse nationalities to investigate the potential influence of cultural and social factors on assessors' visual attention patterns. Furthermore, capturing real-time eye-tracking data from healthcare professionals in NICUs could offer valuable insights into clinical decision-making processes. To address data limitations, we plan to explore generative AI for creating synthetic neonatal facial image datasets [6]. Additionally, we will explore alternative deep learning architectures like Vision Transformers, and other explainers such as LIME [30] and SHAP [25]. We also plan to investigate alternative solutions to the disagreement problem, including aggregation approaches that leverage metrics such as robustness and infidelity as weighting factors. Finally, automated concept-generation techniques will be explored, to increase the practicability of our method in clinical settings.

References

1. Balda, R.D.C.X., Guinsburg, R., de Almeida, M.F.B., de Araújo Peres, C., Miyoshi, M.H., Kopelman, B.I.: The recognition of facial expression of pain in full-term newborns by parents and health professionals. Arch. Pediatrics Adolescent Med. **154**(10), 1009–1016 (2000)

2. Barros, M.C.D.M., et al.: Identification of pain in neonates: the adults' visual perception of neonatal facial features. J. Perinatol. **41**(9), 2304–2308 (2021)
3. Beam, K., Sharma, P., Levy, P., Beam, A.L.: Artificial intelligence in the neonatal intensive care unit: the time is now. J. Perinatol. **44**(1), 131–135 (2024)
4. Brahnam, S., Chuang, C.F., Shih, F.Y., Slack, M.R.: Machine recognition and representation of neonatal facial displays of acute pain. Artif. Intell. Med. **36**(3), 211–222 (2006)
5. Brahnam, S., et al.: Neonatal pain detection in videos using the Icopevid dataset and an ensemble of descriptors extracted from gaussian of local descriptors. Appl. Comput. Inf. **19**(1/2), 122–143 (2020)
6. Buzuti, L., Giraldi, G., Heiderich, T., Barros, M., Guinsburg, R., Thomaz, C.: Generative AI for neonatal pain assessment: a sound approach to improve data-driven automatic recognition in intensive care unit. SSRN Electron. J. (2024). SSRN: https://ssrn.com/abstract=4994076
7. Cabitza, F., Fregosi, C., Campagner, A., Natali, C.: Explanations considered harmful: the impact of misleading explanations on accuracy in hybrid human-AI decision making. In: World Conference on Explainable Artificial Intelligence, pp. 255–269. Springer (2024)
8. Carlini, L.P., et al.: Human vs machine towards neonatal pain assessment: a comprehensive analysis of the facial features extracted by health professionals, parents, and convolutional neural networks. Artif. Intell. Med. **147**, 102724 (2024)
9. Carlini, L.P., et al.: A convolutional neural network-based mobile application to bedside neonatal pain assessment. In: 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 394–401. IEEE (2021)
10. Carlini, L.P., et al.: Neonatal pain scales and human visual perception: an exploratory analysis based on facial expression recognition and eye-tracking. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12662, pp. 62–76. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68790-8_6
11. de Cassia Xavier Balda, R., Guinsburg, R.: Perceptions of neonatal pain. Neoreviews **8**(12), e533–e542 (2007)
12. Chen, H., Gomez, C., Huang, C.M., Unberath, M.: Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Dig. Med. **5**(1), 156 (2022)
13. Coutrin, G.A., et al.: Convolutional neural networks for newborn pain assessment using face images: a quantitative and qualitative comparison. In: Su, R., Zhang, Y., Liu, H., F Frangi, A. (eds.) International Conference on Medical Imaging and Computer-Aided Diagnosis. pp. 503–513. Springer, Singapore (2023)
14. Cruz, M.D., Fernandes, A., Oliveira, C.: Epidemiology of painful procedures performed in neonates: a systematic review of observational studies. Eur. J. Pain **20**(4), 489–498 (2016)
15. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: single-shot multi-level face localisation in the wild. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5202–5211 (2020)
16. Giordano, V., et al.: Comparative analysis of artificial intelligence and expert assessments in detecting neonatal procedural pain. Sci. Rep. **14**(1), 20374 (2024)
17. Grunau, R.V., Craig, K.D.: Pain expression in neonates: facial action and cry. Pain **28**(3), 395–410 (1987)
18. Hausmann, J.: Accurate neonatal face detection for improved pain classification in the challenging NICU setting. IEEE Access **12**, 49122–49133 (2024)
19. Heiderich, T.M., et al.: Face-based automatic pain assessment: challenges and perspectives in neonatal intensive care units. Jornal de Pediatria **99**(6), 546–560 (2023)

20. Heiderich, T.M., Leslie, A.T.F.S., Guinsburg, R.: Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. Acta Paediatr. **104**(2), e63–e69 (2015)
21. Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: a practitioner's perspective. Trans. Mach. Learn. Res. (2024). https://openreview.net/forum?id=jESY2WTZCe
22. Laberge, G., Pequignot, Y.B., Marchand, M., Khomh, F.: Tackling the XAI disagreement problem with regional explanations. In: International Conference on Artificial Intelligence and Statistics. pp. 2017–2025. PMLR (2024)
23. Lawrence, J., Alcock, D., McGrath, P., Kay, J., MacMurray, S.B., Dulberg, C.: The development of a tool to assess neonatal pain. Neonatal Network: NN **12**(6), 59–66 (1993)
24. Llerena, A., et al.: Neonatal pain assessment: do we have the right tools? Front. Pediatr. **10**, 1022751 (2023)
25. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
26. Nauta, M., et al.: From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. ACM Comput. Surv. **55**(13s), 1–42 (2023)
27. Orsi, R.N., et al.: Visual attention during neonatal pain assessment: a 2-s exposure to a facial expression is sufficient. Electron. Lett. **59**(6), e12756 (2023)
28. Parkhi, O., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association (2015)
29. Pirie, C., Wiratunga, N., Wijekoon, A., Moreno-Garcia, C.F.: Agree: a feature attribution aggregation framework to address explainer disagreements with alignment metrics. In: CEUR Workshop Proceedings (2023)
30. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
31. Roy, S., Laberge, G., Roy, B., Khomh, F., Nikanjam, A., Mondal, S.: Why don't XAI techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In: 2022 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 444–448. IEEE (2022)
32. Salekin, M.S., Zamzmi, G., Goldgof, D., Kasturi, R., Ho, T., Sun, Y.: First investigation into the use of deep learning for continuous assessment of neonatal postoperative pain. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 415–419. IEEE (2020)
33. Salekin, M.S., Zamzmi, G., Goldgof, D., Kasturi, R., Ho, T., Sun, Y.: Multimodal Spatio-temporal deep learning approach for neonatal postoperative pain assessment. Comput. Biol. Med. **129**, 104150 (2021)
34. Schwarzschild, A., Cembalest, M., Rao, K., Hines, K., Dickerson, J.: Reckoning with the disagreement problem: Explanation consensus as a training objective. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pp. 662–678 (2023)
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
36. da Silva, G.V.T., et al.: What facial features does the pediatrician look to decide that a newborn is feeling pain? Am. J. Perinatol. **40**(08), 851–857 (2023)

37. Simons, S.H., Dijk, M., Anand, K.S., Roofthooft, D., Lingen, R.A., Tibboel, D.: Do we still hurt newborn babies?: a prospective study of procedural pain and analgesia in neonates. Arch. Pediatrics Adolescent Med. **157**(11), 1058–1064 (2003)
38. Soares, J.D.C.A., et al.: Looking at neonatal facial features of pain: do health and non-health professionals differ? Jornal de Pediatria **98**(4), 406–412 (2022)
39. Sun, A., Ma, P., Yuan, Y., Wang, S.: Explain any concept: Segment anything meets concept-based explanation. Adv. Neural. Inf. Process. Syst. **36**, 21826–21840 (2023)
40. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328. PMLR (2017)
41. Tamanaka, F.G., et al.: Neonatal pain assessment: a Kendall analysis between clinical and visually perceived facial features. Comput. Meth. Biomech. Biomed. Eng. Imaging Vis. **11**(3), 331–340 (2023)
42. Williams, M.D., Lascelles, B.D.X.: Early neonatal pain—a review of clinical and experimental implications on painful conditions later in life. Front. Pediatr. **8**, 30 (2020)
43. Yeh, C.K., Hsieh, C.Y., Suggala, A., Inouye, D.I., Ravikumar, P.K.: On the (in) fidelity and sensitivity of explanations. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
44. Zamzmi, G., Paul, R., Goldgof, D., Kasturi, R., Sun, Y.: Pain assessment from facial expression: neonatal convolutional neural network (N-CNN). In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



APÊNDICE I – ARTIGO PUBLICADO NA CONFERÊNCIA *WORLD CONFERENCE ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI 2026)*

From Pixels to Pain Scales: Robust Evaluation of Explainable AI in Neonatal Pain Detection

Leonardo Antunes Ferreira[1,2][0000−0002−9856−2460], Craig Pirie[2][0000−0002−6799−0497], Gabriel de Almeida Sá Coutrin[1][0000−0002−9211−6101], Lucas Pereira Carlini[1][0000−0003−0999−186X] Carlos Francisco Moreno-García[2][0000−0001−7218−9023], Marina Carvalho de Moraes Barros[3][0000−0001−6989−3474], Ruth Guinsburg[3][0000−0003−1967−9861], and Carlos Eduardo Thomaz[1][0000−0001−5566−1963]

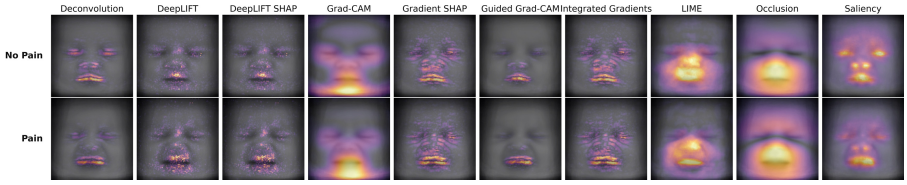
[1] Centro Universitário FEI, São Bernardo do Campo, Brazil
[2] Robert Gordon University, Aberdeen, U.K.
[3] Universidade Federal de São Paulo, São Paulo, Brazil
{leferr,cet}@fei.edu.br

Abstract. Accurate and trustworthy neonatal pain assessment is essential in intensive care, yet deep learning systems remain opaque. eXplainable Artificial Intelligence aims to clarify model decision-making, but no gold-standard exists due to persistent disagreement between explainers. In neonatal pain assessment, only Grad-CAM and Integrated Gradients have been examined, and few studies have aligned explanations with clinical knowledge. We present here four contributions. First, we evaluated ten post-hoc explainers across three architectures (NCNN, VGGFace, ViT-B/32) at both pixel and semantic-region levels, showing that even with pixel-level variability, explainers consistently converged on key facial regions such as the mouth and nose. Second, we assessed explainers faithfulness using multiple pixel-replacement strategies and found that black and white perturbations produced misleadingly scores for NCNN and VGGFace due to unrealistic image alterations, while ViT-B/32 was less affected because of its patch-based architecture. Third, we compared explainer-derived regions with 13 neonatal pain scales and concluded that NCNN achieved the highest alignment despite being less predictive, yet all models focused on clinically relevant regions (mouth, eyes, forehead) alongside some attention to areas outside established scales (nose and cheeks). Fourth, we presented new standards for collaboration between clinicians and artificial intelligence researchers to develop a reliable and trustworthy neonatal pain assessment tool.

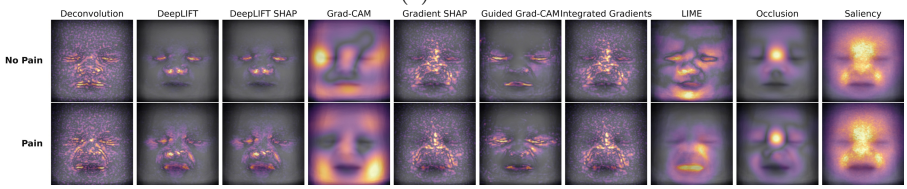
Keywords: Neonatal Pain · Healthcare XAI · Disagreement · Faithfulness · Clinical Interpretability.

1 Introduction

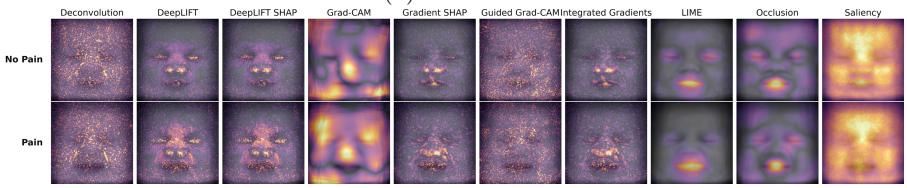
Neonatal pain remains one of the most difficult clinical assessments performed in intensive care [4]. Newborns cannot verbalize their pain, facial cues change



(a) NCNN



(b) VGGFace



(c) ViT-B/32

Fig. 1: Average explanations and average faces for *Pain* and *No pain* classes.

rapidly, and widely used neonatal behavioural and multidimensional pain scales require trained personnel and have intermittent scoring [19]. These constraints lead to delayed recognition, high inter-rater variability, and inconsistent pain management. Advances in computer vision and deep learning have enabled automatic pain detection from facial images [15]. By offering continuous and real-time monitoring, such systems could support Neonatal Intensive Care Unit (NICU) practice and improve analgesic management [13].

However, deploying opaque models in high-stakes clinical settings raises substantial concerns. While deep learning models can perform at or above expert level, their internal decision-making processes are typically inaccessible to users [9,17]. This opacity limits clinical adoption, as it becomes unclear whether predictions rely on valid information or on inaccurate patterns introduced by data imbalance, imaging conditions, or medical devices. The result is a paradox, in one hand Artificial Intelligence (AI) is introduced to mitigate human subjectivity, on the other hand its lack of transparency introduces a new barrier to trust.

Explainable AI (XAI) seeks to clarify how models arrive at their decisions. In medical imaging, saliency maps are widely used to highlight image pixels most influential to the model's prediction [2]. These explanations can strengthen clinician trust, support model debugging, and promote fairness and transparency

in clinical systems. A growing range of *post-hoc* XAI techniques has been used for deep learning models in healthcare [17], but without a universal or golden-standard method. Recent regulatory developments, including the European Union's AI Act and its associated "right to explanation" [14], further reinforce the need for clinically interpretable AI.

Despite progress in automatic neonatal pain detection [15], few studies have examined the role of XAI [7,10,13] or proposed frameworks for aligning explanations with neonatal clinical expertise [6,23]. Existing work has relied primarily on Grad-CAM and Integrated Gradients, without broader evaluation of alternative explainers. A rigorous, quantitative assessment of multiple XAI methods is therefore needed to better characterize the disagreement problem [18] in neonatal pain assessment and to evaluate the reliability of explainers as clinical tools.

This study extends the framework proposed in Pirie et al. [23] by expanding the analysis from two to ten commonly used XAI explainers [17] and introducing a Vision Transformer (ViT) model [12], whose architecture differs fundamentally from the Convolutional Neural Networks (CNNs) previously evaluated, and is still not well explored in this domain. In addition, we reformulate explainer faithfulness by incorporating a perturbation-based metric that better reflects disruptions likely to affect model predictions, and we introduce a new agreement measure comparing explainer-derived regions with clinically validated neonatal pain scales, further bridging the gap between human and AI knowledge. Our contributions are fourfold:

1. **Multi-model and multi-explainer agreement:** We quantify agreement across top-k pixels and top-k facial regions for three architectures (NCNN, VGGFace, ViT-B/32) and ten explainers, providing a comprehensive characterization of the disagreement problem.
2. **Faithfulness evaluation:** We evaluate how accurately each explainer reflects the model's reasoning using pixel-replacement perturbation, assessing the differences in black, white, mean, Gaussian noise, and RemOve And Debias (ROAD) strategies, based on the Area Over the Perturbation Curve (AOPC) metric.
3. **Agreement with neonatal pain scales:** We introduce an agreement metric linking explainer-derived regions to 13 clinically validated neonatal pain scales, enabling direct comparison between model attributions and expert-defined facial indicators of pain.
4. **Implications for neonatal pain monitoring**: We outline how our findings can guide the development of reliable and clinically interpretable systems for real-time neonatal pain assessment.

The remainder of the paper is organized as follows. Section 2 reviews prior work in XAI evaluation and neonatal pain assessment. Section 3 details the methods. Section 4 outlines the experimental setup. Section 5 reports the results. Section 6 discusses the findings with attention to clinical implications, and Section 7 concludes with key insights and future directions.

2 Related Works

2.1 eXplainable Artificial Intelligence

With the growing popularity of deep learning, AI systems have achieved human-level or superior performance across many tasks [17]. However, transparency is frequently sacrificed for these benefits, leading to "black box" models with internal decision-making processes that users cannot access. XAI methods were developed to mitigate this opacity by providing interpretable insights into how complex models generate predictions [9,17].

XAI techniques can be categorized along several dimensions:

- **Local vs. Global:** *Local* methods explain a single prediction (e.g., why an image was classified as pain), whereas *Global* methods describe overall model behaviour across a dataset.
- **Intrinsic vs. Post-hoc:** *Intrinsic* methods rely on models that are interpretable by design, while *Post-hoc* methods are applied after training to interpret an existing model without altering its structure.
- **Model-specific vs. Model-agnostic:** *Model-specific* techniques exploit a model's internal mechanisms (such as gradients), whereas *Model-agnostic* methods can be applied to any classifier or regressor.
- **Attribution-based vs. Non-attribution-based:** *Attribution-based* explainers assign relevance scores to features or pixels, while *Non-attribution-based* methods provide explanations through rules, examples, or surrogate models.

Popular XAI methods for medical applications include mostly local and *post-hoc* methods [17], such as:

- **Deconvolution [36]:** Backpropagates only positive signals and inverts convolutional operations to reveal input structures that strongly activate specific convolutional filters.
- **DeepLIFT [30]:** Compares neuron activations against a reference input and backpropagates contribution scores, enabling attribution even in cases where gradients vanish.
- **DeepLIFT SHAP [20]:** Combines DeepLIFT with SHAP's game-theoretic framework, averaging multiple DeepLIFT evaluations over different background samples to approximate Shapley values and reduce attribution variance.
- **Grad-CAM [29]:** Gradient-weighted Class Activation Mapping uses gradients with respect to the final convolutional feature maps to produce coarse heatmaps that localize the most discriminative spatial regions.
- **Gradient SHAP [20]:** Extends Integrated Gradients by averaging gradient-based attributions across multiple noisy interpolations between the baseline and the input, yielding smoother and more robust explanations.
- **Guided Grad-CAM [29]:** Combines Guided Backpropagation with Grad-CAM, merging high-level localization with fine-grained pixel-level gradients to generate sharper attribution maps.

- **Integrated Gradients [32]:** Integrates gradients along a straight-line path from a baseline image to the input, producing attributions that satisfy the axioms of sensitivity and implementation invariance.
- **LIME [24]:** Local Interpretable Model-Agnostic Explanations approximates the model locally using a sparse linear surrogate, identifying the features that most influence the prediction within the immediate neighbourhood of the input.
- **Occlusion [36]:** Perturbs the input by masking small patches and measuring changes in the output score, providing a direct estimate of regional importance.
- **Saliency [31]:** Computes the gradient of the target class score with respect to each input pixel, highlighting regions where small input changes most strongly influence the prediction.

2.2 Disagreement and XAI Metrics

The proliferation of XAI techniques has created uncertainty about which methods provide reliable explanations. A central issue is the disagreement problem, where different explainers often highlight distinct features for the same model and input, complicating interpretation and undermining user trust [18].

To quantify this variability, Krishna et al. [18] introduced a suite of metrics that compare the top-k most relevant features across explainers. *Feature Agreement* measures the overlap in selected features. *Rank Agreement* assesses whether shared features appear in the same rank positions. *Sign Agreement* checks whether the direction of contribution (positive or negative) is consistent. *Signed Rank Agreement* combines all three, identifying features that are shared, similarly ordered, and directionally consistent across explanations.

Faithfulness offers another perspective, assessing whether an explanation reflects the true behaviour of the model. A faithful explanation should identify features that, when perturbed, meaningfully change the model's prediction. One of the earliest fidelity tests, Pixel Flipping [3], removes pixels in order of attributed importance and measures how rapidly the model's output degrades. Samek et al. [28] extended this to Region Perturbation, masking larger image segments and summarizing the resulting prediction decline using the AOPC.

Rieger et al. [25] introduced the Iterative Removal of Features (IROF), which quantifies the smallest fraction of features that must be removed for the prediction to fall below a specified threshold, indicating how concentrated the salient evidence is. ROAD [26] evaluates prediction drop as the top-k pixels are removed. Unlike earlier perturbation strategies that replace removed pixels with constant values (e.g., black, white, or noise), ROAD replaces them using a debiased interpolation procedure that reconstructs missing pixels from their surrounding neighbourhood. This interpolation estimates the expected value of the removed pixels conditioned on nearby pixels, producing a locally plausible reconstruction that preserves natural image statistics.

A further dimension in XAI evaluation is the use of sanity checks [1], such as randomizing model weights, to verify that attribution maps change accordingly.

Failure to respond to these checks suggests that the explainer may be insensitive to the model's learned parameters and therefore unreliable.

2.3 XAI in Neonatal Pain Assessment

Automatic neonatal pain assessment has progressed substantially over the past decade [15]. Early work used handcrafted features combined with traditional classifiers to detect pain-related facial expressions [5]. Deep learning later enabled end-to-end models capable of analysing static images, video, and multimodal data [13,27,35].

Despite these advances, most models still function as black boxes, prioritizing accuracy over interpretability. The first application of XAI to neonatal pain detection was in Carlini et al. [7], using Grad-CAM and Integrated Gradients to visualize regions driving predictions in a VGGFace model. The qualitative findings suggested a focus on facial areas aligned with clinical expertise, like the mouth and eyes.

Coutrin et al. [10] evaluated multiple deep learning architectures (VGGFace, ResNet, InceptionV3, SENet) using Grad-CAM and reported, qualitatively, that VGGFace produced more coherent explanations to facial features, likely due to its pre-training on facial recognition. Carlini et al. [6] extended this analysis by comparing Grad-CAM and Integrated Gradients attributions overlap with human gaze patterns, highlighting substantial disagreement across explainers.

More recently, Pirie et al. [23] examined quantitatively how Grad-CAM and Integrated Gradients align with human gaze, and found that both humans and models often focus on central facial regions such as the eyes, nose, and mouth. The study introduced a region agreement metric, evaluating attribution patterns across semantic facial areas defined by neonatal pain scales. This region-level perspective enhanced clinical interpretability, enabling statements such as: "Grad-CAM identifies the mouth as the most influential region for this prediction".

3 Methods

We build on prior work, using the agreement framework of Pirie et al. [23] to analyse explainer alignment and the AOPC formulation of Samek et al. [28] to assess faithfulness. We further introduce a method to quantify agreement between explainer-derived attributions and predefined facial regions from clinical pain scales.

3.1 Explainers Agreement

Pixel agreement quantifies whether two explainers highlights spatially overlapping evidence for a given input. Let $\mathcal{P} = \{1, \ldots, H\} \times \{1, \ldots, W\}$ denote the set of pixel coordinates inside a saliency map S from explainer E, with height H and width W, and let S_{ij} represent the saliency value at pixel (i, j).

For each saliency map, we retained only the top $k\%$ most relevant pixels obtained by applying a percentile threshold (τ_k) such that:

$$\tau_k = \text{Percentile}\big(\{S_{ij} \mid (i,j) \in \mathcal{P}\},\, 100 - k\big), \tag{1}$$

pixels exceeding this threshold form a binary importance mask M:

$$M = \begin{cases} 1, & \text{if } S_{ij} \geq \tau_k, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Pixel-level agreement between explainers $E^{(a)}$ and $E^{(b)}$ is:

$$\text{PixelAgreement}(E^{(a)}, E^{(b)}) = \frac{M^{(a)} \cap M^{(b)}}{M^{(a)} \cup M^{(b)}}. \tag{3}$$

A value of 1 indicates perfect overlap between the two explanations, whereas 0 indicates that the top ($k\%$) salient pixels do not intersect.

Region agreement extends pixel agreement analysis to semantically defined facial regions (mouth, eyes, nose, etc.). For each explanation, the k most important regions are extracted based on their region-level saliency.

Each pixel is associated with a semantically defined facial region through a mapping $\phi : \{1, \ldots, H\} \times \{1, \ldots, W\} \to \mathcal{R}$, where $\mathcal{R} = \{r_1, \ldots, r_R\}$ represents facial regions. For a given region $r \in \mathcal{R}$, the corresponding pixel set is:

$$\Omega_r = \big\{(i,j) \mid \phi(i,j) = r\big\}. \tag{4}$$

Region-level importance is obtained by averaging the saliency values within each region, resulting in the vector $A = (A_1, \ldots, A_R)$, where

$$A_r = \frac{1}{|\Omega_r|} \sum_{(i,j) \in \Omega_r} S_{ij}. \tag{5}$$

Let $Top_k^{(a)} = [r_1^{(a)}, \ldots, r_k^{(a)}]$ and $Top_k^{(b)} = [r_1^{(b)}, \ldots, r_k^{(b)}]$ be the lists of the k most relevant regions from two distinct explainers, $E^{(a)}$ and $E^{(b)}$. The agreement score is then formulated as:

$$\text{RegionAgreement}(E^{(a)}, E^{(b)}) = \frac{\big|Top_k^{(a)} \cap Top_k^{(b)}\big|}{k}. \tag{6}$$

The region agreement score represents the fraction of regions that appear in the top-k sets of both explanations. A value of 1 indicates perfect agreement, whereas a value of 0 indicates complete disagreement. Region-level evaluation mitigates pixel-level instability and aligns with clinical reasoning, which is based on facial action units rather than individual pixels.

3.2 Faithfulness Evaluation by Pixel Perturbation

We reformulated the faithfulness analysis proposed in Pirie et al. [23] to better capture perturbations that meaningfully influence model predictions. Instead of relying on infidelity [34], which quantifies the mean squared error between (i) a dot product of explainer attributions and random input perturbations, and (ii) the resulting change in model output, we adopted an ordered pixel-replacement strategy [3,26,28].

Our approach perturbs pixels in Most Relevant First (MoRF) order according to each explainer attribution values, enabling a direct measurement of their causal impact on the prediction, something infidelity cannot guarantee due to its reliance on stochastic noise injection. Furthermore, the pixel-replacement method allows extrapolation from individual pixels to broader facial regions, highlighting which facial areas most strongly drive model behaviour and should therefore remain visible and emphasized during clinical assessment.

Because no established baseline exists for perturbation strategies in neonatal pain detection, we evaluated five pixel-replacement approaches (black, white, mean, Gaussian noise, and ROAD), illustrated in Figure 2.

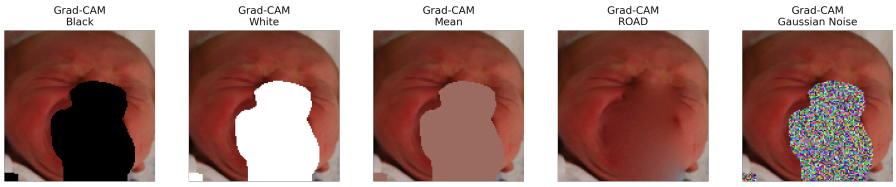


Fig. 2: Example of the perturbation strategies.

To quantify faithfulness, we computed the AOPC [28], which captures the average drop in model confidence as important pixels are iteratively perturbed:

$$\mathrm{AOPC} = \frac{1}{L+1}\Big\langle \sum_{k=0}^{L} \big[f(x) - f(x^{(k)})\big] \Big\rangle, \tag{7}$$

where $f(x)$ is the original model predicted pain probability, $f(x^{(k)})$ is the perturbed predicted pain probability after removing the top-k most relevant pixels, $\langle\cdot\rangle$ denotes dataset averaging, and L is the total number of perturbation steps. Higher values reflect more faithful explanations.

To confirm that perturbations to top-k pixels meaningfully affect the model, we centred AOPC against a random baseline, which perturbs the same amount of pixels but in random order instead of MoRF order. Values near zero suggest no meaningful difference from random perturbation, implying that the explainer does not identify truly faithful features, negative values indicates that perturbing in random order is more relevant than the explainers top-k pixels, and positive values relates to the explainers top-k pixels being faithful to the model.

3.3 Agreement with Clinical Pain Scales

To evaluate consistency between models and clinical knowledge, we compared the ranked region importances from explainers with the facial regions present in neonatal pain scales that are used for pain detection. Let the facial regions be $\mathcal{R} = \{r_1, r_2, \ldots, r_R\}$ with corresponding explainers region-importances $A = (A_1, A_2, \ldots, A_R)$. Regions were ranked in descending importance, with rank 1 assigned to the most relevant region.

We defined a score C that rewards regions emphasized both by explainer and the clinical scale, while penalizing regions highly ranked by explainers but absent from the scale. Let $g_r \in \{0, 1\}$ indicate whether region r is included in the pain scale. The agreement score is:

$$C = \sum_{r=1}^{R} (R + 1 - \text{rank}(r))(2g_r - 1), \tag{8}$$

where $(R + 1 - \text{rank}(r))$ inverts the ranking so that higher values correspond to higher importance, and $(2g_r - 1)$ assigns positive or negative contributions depending on if the facial regions are present or not in the clinical pain scale.

To ensure comparability across scales with different numbers of regions, the score was normalized by its theoretical maximum $C_{\max}$, the case in which all clinically relevant regions occupy the top-P ranks, with $P = \sum_r g_r$:

$$C_{\max} = 2\left[P(R+1) - \frac{P(P+1)}{2}\right] - \frac{R(R+1)}{2}. \tag{9}$$

The normalized agreement score is:

$$C_{\text{norm}} = \frac{C + C_{\max}}{2C_{\max}}, \tag{10}$$

where $C_{\text{norm}} \in [0, 1]$. A value of 1 indicates perfect agreement with the clinical scale, 0 indicates complete disagreement, and 0.5 represents a neutral or random correspondence. It is important to note that this agreement metric is designed to be conservative. By penalizing regions outside established scales, prioritizing validated clinical knowledge, a necessary step for building initial trust in safety-critical NICU environments.

4 Experimental Setup

4.1 Models

We extended the set of evaluated models in Pirie et al. [23] by introducing the ViT-B/32 architecture. ViTs are inherently different architectures than CNNs and process images as sequences of patch embeddings, enabling fundamentally different feature integration and attention patterns. Including ViT-B/32 allows us to evaluate whether XAI methods generalize beyond CNNs, to test if pixel-based perturbation faithfully captures Transformer behaviour, and to determine

whether agreement with clinical pain scales holds for models that rely on patch-level reasoning rather than localized pixel features.

- **NCNN (Neonatal CNN) [35]:** This architecture was specifically designed for neonatal pain analysis. As no pre-trained NCNN weights are available, the model was trained from scratch.
- **VGGFace [21]:** This CNN was adapted for neonatal pain assessment by adding two fully connected layers with 512 neurons each, while retaining the pre-trained convolutional backbone for transfer learning [7].
- **ViT-B/32 [12]:** The ViT applies transformer mechanisms to image classification. In this study, the ViT-B/32 model pre-trained on ImageNet was fine-tuned for neonatal pain detection, as no public ViT pre-trained on faces was available.

All models were modified to output a single sigmoid-activated neuron in the final classification layer, with outputs being interpreted as the *Pain* probability.

4.2 Datasets

We recognize that acquiring new datasets of newborns experiencing pain is ethically sensitive, typically requiring approval from research ethics committees and maternal consent. Because most such datasets remain private, our work relies on well-established neonatal datasets available in the literature and released under formal licence agreements, ethical committees approval and maternal consent.

- **infant Classification Of Pain Expression (iCOPE):** Developed by Brahnam et al. [5] at St. John Hospital (now Mercy Hospital) in Missouri, USA, this dataset contains 200 high-resolution images (3008 × 2000 pixels) from 26 Caucasian neonates aged 18 hours to 3 days. Images were captured under four stimuli: (1) crib transfer, (2) air exposure, (3) heel friction, and (4) heel puncture for blood collection. The dataset includes 63 images of neonates at rest (*No Pain*), 18 crying, 23 during air stimulation, 36 during friction, and 60 during painful procedures (*Pain*). For this study, we used only the *No Pain* and *Pain* categories.
- **UNIFESP:** Developed by Heiderich et al. [16] at the Federal University of São Paulo (Brazil), this dataset includes 30 healthy neonates aged 24 to 168 hours. Each neonate was photographed 12 times (resolution: 320 × 233 pixels) before, during, and after clinically indicated painful procedures performed by neonatologists. Expert neonatologists independently evaluated the images, producing 164 *Pain* and 196 *No Pain* labels.

All images from both datasets underwent standardized pre-processing to ensure consistency across models. Facial regions were detected and cropped using the RetinaFace detector [11], which also provided 106 keypoints later used to define the facial regions (between eyes, cheeks, chin, eyebrows, eyes, forehead, mouth, nasolabial folds, nose, and outside the face) automatically. This improves the work of Pirie et al. [23] where regions were manually annotated. Images were then resized according to each model's input specification.

Table 1: Hyperparameters used in the training of the models.

Hyperparameter	**NCNN**	**VGGFace**	**ViT-B/32**
Input size	120×120 (RGB)	224×224 (BGR)	224×224 (RGB)
pre-training	None	VGGFace dataset	ImageNet
Output activation	Sigmoid	Sigmoid	Sigmoid
Loss function	Cross-Entropy	Cross-Entropy	Cross-Entropy
Batch size	16	16	16
Initial learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Fine-tuning learning rate	–	1×10^{-6} (last 2 conv.)	1×10^{-6} (last 2 blocks)
Regularization	$L_2 = 0.01$	$L_1 = 5 \times 10^{-4}$	–
Dropout	10%	50%	0%
Early stopping	5 epochs	5 epochs	5 epochs

4.3 Training Recipe

To mitigate data leakage and evaluate model generalization to unseen neonates, we employed a leave-some-subjects-out cross-validation strategy [10]. The iCOPE and UNIFESP datasets were merged and partitioned into ten folds, ensuring that all images from a given neonate were assigned exclusively to either the training or testing set. This prevented subject-level overlap and preserved the independence of model evaluation.

Data augmentation consisted of mild transformations designed to improve model robustness given the limited dataset size. Augmentations included random rotations (up to 30°), shear (0.15), width and height shifts (0.20), brightness adjustments (0.50–1.1), zoom (0.70–1.5), and horizontal flips. These transformations were applied only to the training set, generating 20 augmented samples per image. No augmentations were applied to the testing set, and all explainability analyses were computed on the original, unmodified test images. The RetinaFace [11] detector ensured that all augmented samples retained a complete face.

All three models were trained and evaluated on identical folds. Within each fold, NCNN was trained from scratch, while VGGFace and ViT-B/32 were fine-tuned on the same training set and evaluated on the same test set. Table 1 summarizes the hyperparameters used for each model.

4.4 XAI Methods

We applied ten commonly used post-hoc explainable AI techniques in medical domain [17]. All methods were implemented using the CAPTUM[4] Python library with its default hyperparameters. For each test image and model, one attribution map per explainer was obtained using each explainer default CAPTUM parameters, and with values scaled to $[-1, 1]$ to indicate the degree to which individual pixels contributed positively or negatively to the model's output. The explainers used were: Deconvolution [36], DeepLIFT [30], DeepLIFT

[4] `https://captum.ai/`

Table 2: Presence of facial regions across neonatal pain scales.

Region	APS	BIIP	EVENDOL	LIDS	MCGRATH	MIPS	M-NFCS	NFCS	NNICU	PAT	PASPI	P-MIPS	PIPP	PIPP-R	Total
Eyes	1	1	1	1	1	1	1	1	1	1	1	1	1	1	13
Between Eyes	0	1	1	1	1	1	1	1	1	1	1	1	1	1	12
Forehead	0	1	1	1	1	1	1	1	1	1	1	1	1	1	12
Nasolabial Folds	1	1	0	0	1	1	1	1	1	1	1	1	1	1	11
Mouth	0	1	1	1	1	1	1	1	1	1	1	1	0	0	10
Chin	0	0	0	1	1	1	0	1	0	0	0	1	0	0	5
Eyebrows	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2
Total	3	5	4	6	6	6	5	6	5	5	6	4	4		65

SHAP [20], Grad-CAM [29], Gradient SHAP [20], Guided Grad-CAM [29], Integrated Gradients [32], LIME [24], Occlusion [36], and Saliency [31].

4.5 Assessing Explanations

For agreement metrics, we used the magnitude of attribution values to assess model attention independently of sign, focusing on whether the model attended to a region rather than on the direction of its contribution.

For each explainer, we computed pixel agreement using top-$k = 10\%$ and region agreement using top-$k = 3$ [23]. To ensure the robustness and validity of the generated explanations, we performed sanity checks [1] by comparing each explainer against: (i) a randomly initialized version of the same model architecture with untrained weights, (ii) random noise images, and (iii) Canny edge-detector maps. This enabled identification of explainers that exhibited distinct and meaningful patterns rather than random artifacts.

In a second stage, we compared explainers directly to one another to examine inter-method similarity. For region agreement, the semantic facial regions were automatically computed over the 106 keypoints extracted during face detection.

We assessed explainer faithfulness using pixel-replacement perturbations applied to the top-k pixels ranked by absolute attribution values. Although the raw attributions are signed, the sign does not affect the AOPC computation in our setup because pixel ranking and replacement were based solely on attribution magnitude, and the perturbed output is always the model's predicted probability of *Pain*. Because replacement strategy can strongly influence conclusions, we assessed multiple strategies shown in Figure 2, ranging from aggressive (black-/white/Gaussian noise) to more naturalistic (mean and ROAD), and evaluated using the AOPC metric centralized against a random perturbation.

Lastly, we quantified the agreement between each explainer and clinically validated neonatal pain scales to determine whether the AI models emphasized the same facial regions that clinicians rely on for pain assessment. Table 2, shows the scales and its used facial regions (for more information about the scales, see Tamanaka et al. [33]).

5 Results

The average classification performance of the models on the pain detection task was around 80% accurate (Table 3). Confirming that they learned pain patterns

Table 3: Classification metrics (average ± standard deviation, values in %).

Metric	NCNN	VGGFace	ViT-B/32
Accuracy	80.0 ± 9.2	**84.2 ± 4.8**	81.8 ± 6.9
F1-Score	76.0 ± 11.8	**82.0 ± 5.9**	78.5 ± 7.7
Sensitivity	72.5 ± 16.2	**80.8 ± 8.8**	74.3 ± 11.2
Specificity	86.7 ± 7.7	87.0 ± 9.6	**88.1 ± 7.2**
AUC	88.7 ± 5.6	**91.5 ± 3.3**	90.1 ± 4.7

Table 4: Explainers with average pixel agreement values above 0.50 (DL = DeepLIFT, DLS = DeepLIFT SHAP, ED = Edge Detector, IG = Integrated Gradients, GS = Gradient SHAP, D = Deconvolution, S = Saliency).

NCNN	VGGFace	ViT-B/32
DL & DLS = 1.00	DL & DLS = 1.00	S & D = 0.69
DL & Untrained DL = 1.00	IG & GS = 0.99	IG & GS = 0.69
DLS & Untrained DLS = 1.00	Untrained D & D = 0.79	-
IG & GS = 0.84	-	-
DL & ED = 0.77	-	-
DLS & ED = 0.77	-	-

from the face images. This is important because explanations are only meaningful if the model has learned something.

5.1 Multi-model and Multi-explainer Agreement

In Figure 1, we present the average heatmaps of all explainers across the three models for both *Pain* and *No Pain* classes, all explainers attributions and facial images were aligned on a common reference based on the face detection keypoints extraction and then averaged. The figure illustrates the disagreement problem, as different explainers assign different importance to various regions of the image. Notably, methods such as Grad-CAM, LIME, and Occlusion produce more structured "superpixel" patterns (groups of contiguous pixels) whereas other explainers yield more scattered attributions.

Table 4 reports the pixel-level agreements for the three models that achieved average values above 0.50, indicating that more than half of the highlighted pixels occupy the same spatial locations and therefore reflect similar feature attribution beyond random chance. Table 5 complements this analysis with region-level agreement values exceeding 0.66, meaning that at least two out of three facial regions were consistently identified across explainers, effectively functioning as a majority-vote consensus at the semantic level.

First, we reiterate the findings of Pirie et al. [23] that region-level agreement is substantially higher than pixel-level agreement, as reflected in the number of agreeing explainers and their corresponding values when comparing Tables 4

Table 5: Explainers with average region agreement with at least two regions agreed upon (DL = DeepLIFT, DLS = DeepLIFT SHAP, IG = Integrated Gradients, GS = Gradient SHAP, D = Deconvolution, S = Saliency).

NCNN	**VGGFace**	**ViT-B/32**
DL & DLS = 1.00	DL & DLS = 1.00	DLS & DL = 1.00
IG & GS = 0.95	IG & GS = 1.00	S & D = 0.87
S & IG = 0.85	Untrained D & D = 0.83	IG & GS = 0.86
S & GS = 0.84	Untrained DL & DL = 0.80	Untrained DL & DL = 0.71
GS & D = 0.78	Untrained DLS & DLS = 0.80	Untrained DLS & DLS = 0.71
IG & D = 0.78	DL & D = 0.76	-
S & D = 0.76	DLS & D = 0.76	-
DL & D = 0.67	Untrained GS & GS = 0.73	-
DLS & D = 0.67	S & GS = 0.71	-
-	S & IG = 0.71	-
-	Untrained IG & IG = 0.71	-

Table 6: Ranking of importance (in %) of facial features averaged by explainer.

Rank	**NCNN**	**VGGFace**	**ViT-B/32**
1st	Mouth (29.69)	Nose (26.28)	Mouth (19.51)
2nd	Nose (21.59)	Between Eyes (18.20)	Nose (14.99)
3rd	Nasolabial Folds (16.16)	Mouth (14.66)	Outside Face (12.74)
4th	Eyes (10.70)	Eyes (13.16)	Chin (10.08)
5th	Chin (9.41)	Nasolabial Folds (6.41)	Between Eyes (9.75)
6th	Between Eyes (5.99)	Cheeks (5.37)	Eyes (8.98)
7th	Eyebrows (3.48)	Eyebrows (5.13)	Nasolabial Folds (7.46)
8th	Outside Face (1.27)	Forehead (4.66)	Forehead (6.95)
9th	Cheeks (1.17)	Chin (3.43)	Cheeks (5.83)
10th	Forehead (0.51)	Outside Face (2.67)	Eyebrows (3.73)

and 5. Although explainers differ in their precise pixel attributions, they tend to converge on the same broader facial regions.

When evaluating explainers against their baselines, we observed that DeepLIFT and DeepLIFT SHAP consistently showed higher pixel- and region-level agreement across all models with attributions from untrained networks and, for NCNN, with edge detection. For VGGFace, we also noted higher region agreement between untrained and trained versions of Deconvolution, Gradient SHAP, and Integrated Gradients. Overall, the generally low agreement between most explainers and untrained or edge detector baselines indicates that explanations are driven by learned parameters rather than artefacts. Noise baselines produced negligible pixel- or region-level agreement (typically <0.10 and <0.30, respectively), confirming that saliency maps are not simply noise.

In the explainer-versus-explainer analysis, gradient-based methods, including Integrated Gradients, Gradient SHAP, DeepLIFT, DeepLIFT SHAP, and Saliency, showed higher agreement than superpixel-based methods such as Grad-CAM, Occlusion, and LIME, which exhibited little to no agreement with each other or with other explainers. The tightest pairs were DeepLIFT & DeepLIFT SHAP and Integrated Gradients & Gradient SHAP, reflecting similarities in their underlying mechanisms.

Across models, NCNN showed more pixel-level agreement pairs, whereas VGGFace and especially ViT-B/32 displayed fewer agreed pairs, indicating greater variability in their spatial attribution patterns.

Table 6 presents the ranked importance of facial regions averaged across all explainers for each architecture. For NCNN, the mouth is the most emphasized region, followed by the nose and the nasolabial folds. VGGFace, by contrast, prioritizes the nose and the region between the eyes, with the mouth appearing only in third place. ViT-B/32 also highlights the mouth and nose as its top regions, but unlike the CNN models, it assigns substantial importance to areas outside the face, indicating that the ViT-B/32 pre-trained on ImageNet may be more sensitive to contextual or background cues.

5.2 Faithfulness Evaluation

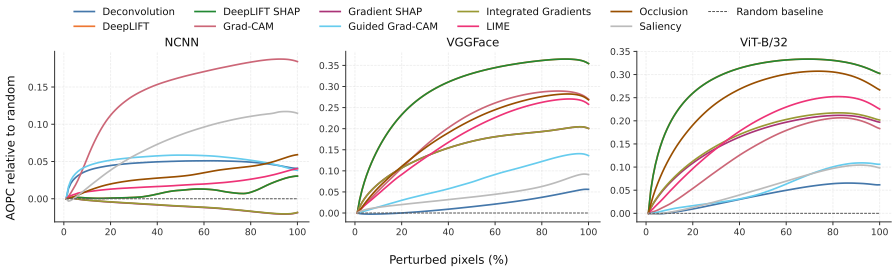


Fig. 3: AOPC relative to random perturbation for ROAD strategy.

The faithfulness evaluation results, are illustrated in Figure 3 and summarized in Tables 7 and 8. A high AOPC, indicates that perturbing the most relevant pixels leads to a significant change in model prediction, in other words, the curve should rise steeply if the perturbed pixels are important to the model.

For all models, ROAD consistently produced positive AOPC faithfulness values (Table 7). This is expected, as ROAD perturbs pixels through a debiased interpolation that preserves local image statistics, thus random perturbation have less of an impact in model prediction [26]. Whereas other perturbation strategies introduce artificial shapes or edges that can leak class information [26]. For example, black and white replacements, particularly for CNNs, often

yielded negative AOPC values. Indicating that model performance degraded faster due to the perturbation artefacts themselves rather than the relevance of the explainer's attributions. To test the null hypothesis that all perturbation strategies perform equally, we applied the Friedman test and obtained a p-value < 0.001 for all models, rejecting the null hypothesis.

As for ViT-B/32 the pattern was the opposite, black and white replacements provided positive faithfulness values whereas only Gaussian noise exhibited negative results. Reinforcing how ViTs operate differently at architecture and prediction level, thus being more likely robust to perturbation than CNNs.

Using ROAD as the default perturbation strategy (Table 8), no explainer emerged as the most faithful across all models, although their results were significantly different according to the Friedman test of the null hypothesis that all explainers perform equally. Nonetheless, several patterns were evident. First, the strong agreement between the pairs DeepLIFT & DeepLIFT SHAP and Integrated Gradients & Gradient SHAP was reflected in their identical faithfulness values. Second, despite differences in pre-training, architecture, and region importance, VGGFace and ViT-B/32 showed remarkably similar explainer faithfulness, both in absolute values and ranking. Third, Grad-CAM, LIME, and Occlusion were consistently top-ranked, as for the least faithful it was dependent on model. For VGGFace and ViT-B/32 were Saliency and Deconvolution, whereas for NCNN it was the Gradient SHAP and Integrated Gradients.

Interestingly, the NCNN model had the least faithful explainers, with AOPC values close to zero, indicating little difference from perturbing random pixels. In contrast, VGGFace and ViT-B/32 showed higher positive values, demonstrating that their explainers were more faithful to model behaviour, particularly those ranked highest.

5.3 Agreement with Neonatal Pain Scales

When we considered explainers and pain scales agreement (full tables in Appendix A) alongside explainers most emphasized facial regions (Table 6), several patterns emerged. As summarized in Table 2, the most common regions across pain scales are the eyes (13/13), the region between the eyes and the forehead (12/13 each), the nasolabial folds (11/13), and the mouth (10/13).

Table 7: Ranking of perturbation strategies by model, AOPC in parentheses.

Rank	**NCNN**	**VGGFace**	**ViT-B/32**
1st	ROAD (0.05)	ROAD (0.22)	White (0.20)
2nd	Gaussian (0.04)	Mean (-0.11)	ROAD (0.19)
3rd	Mean (-0.12)	Gaussian (-0.16)	Black (0.14)
4th	Black (-0.20)	Black (-0.38)	Mean (0.05)
5th	White (-0.24)	White (-0.42)	Gaussian (-0.17)
p-value	< 0.001	< 0.001	< 0.001

Table 8: Ranking of explainers by model for ROAD strategy, AOPC in parentheses.

Rank	**NCNN**	**VGGFace**	**ViT-B/32**
1st	Grad-CAM (0.18)	DeepLIFT (0.35)	DeepLIFT (0.30)
2nd	Saliency (0.11)	DeepLIFT SHAP (0.35)	DeepLIFT SHAP (0.30)
3rd	Occlusion (0.06)	Grad-CAM (0.27)	Occlusion (0.27)
4th	Deconvolution (0.04)	Occlusion (0.27)	LIME (0.23)
5th	Guided Grad-CAM (0.04)	LIME (0.26)	Gradient SHAP (0.20)
6th	LIME (0.04)	Gradient SHAP (0.20)	Integrated Gradients (0.20)
7th	DeepLIFT (0.03)	Integrated Gradients (0.20)	Grad-CAM (0.18)
8th	DeepLIFT SHAP (0.03)	Guided Grad-CAM (0.14)	Guided Grad-CAM (0.11)
9th	Gradient SHAP (–0.02)	Saliency (0.09)	Saliency (0.10)
10th	Integrated Gradients (–0.02)	Deconvolution (0.06)	Deconvolution (0.06)
p-value	< 0.001	< 0.001	< 0.001

NCNN achieved the highest mean agreement (0.71), it prioritized the mouth (29.69%), nose (21.59%), and nasolabial folds (16.16%). The mouth and nasolabial folds are highly prevalent in clinical scales, and NCNN assigned relatively little importance to regions absent from the scales, such as the cheeks and outside the face. VGGFace showed intermediate agreement (0.64), focusing on the nose (26.28%), region between the eyes (18.20%), and mouth (14.66%). Although the nose is not explicitly represented in any scale, its strong focus on the region between the eyes (present in 12 scales) and the mouth supports moderate clinical alignment. ViT-B/32 exhibited the lowest agreement (0.59), it highlighted the mouth (19.51%) and nose (14.99%), it also assigned notable importance to areas outside the face (12.74%), which are not included in neonatal pain scales, penalizing its agreement scores.

Across models and explainers, agreement values were relatively homogeneous, with no explainer consistently outperforming the others. Similarly, most pain scales yielded comparable scores, except for APS, which showed the lowest agreement (NCNN = 0.59, VGGFace = 0.49, ViT-B/32 = 0.37). APS includes only three facial regions (eyes, eyebrows, nasolabial folds), that are rarely prioritized by any model, naturally limiting the achievable agreement.

6 Discussion

6.1 Multi-model and Multi-explainer Agreement

Our results confirmed the disagreement problem described by Krishna et al. [18] in neonatal pain domain, and reinforced the results of Pirie et al. [23]. At the pixel level, most explainers highlighted markedly different areas of the image for the same prediction. Pixel agreement scores were generally low across models, and only a few stable clusters emerged, most notably the very high agreement between Integrated Gradients & Gradient SHAP, and the near-perfect agreement

between DeepLIFT & DeepLIFT SHAP. This is expected, as each pair is derived from similar methodological principles, suggesting that only one explainer from each pair may be needed in practice.

However, DeepLIFT and DeepLIFT SHAP also showed high agreement with their untrained-model attributions across all architectures, indicating that they do not reliably capture information learned from pain detection training and may therefore be unsuitable for clinical interpretation. We observed similar results on VGGFace region agreement with Integrated Gradients, Gradient SHAP, and Deconvolution, where model architecture biases were more important than the learned pain features during training. These findings underscore the challenge of treating any single pixel-level explainer as a reliable representation of model reasoning, specially DeepLIFT and DeepLIFT SHAP.

Once we aggregated explanations into semantic facial regions, agreement scores were at least twice as high as pixel-wise scores across all models, and different explainers tended to converge on similar regions even when they disagreed at the pixel level. This supports the idea, proposed by Pirie et al. [23], that region-based evaluation provides a more stable and clinically interpretable lens for also other explainers, beyond Grad-CAM and Integrated Gradients, closer to how neonatologists reason using pain scales, which are defined over facial action movements rather than individual pixels.

Architectural differences also manifested in the disagreement patterns. NCNN, a comparatively shallow and task-specific network, exhibited higher number of pixel agreement pairs than VGGFace and ViT-B/32. We interpret this as a consequence of lower representational flexibility: with a simpler latent space due to its shallow architecture, different explainers were more likely to highlight similar structures. In contrast, VGGFace and ViT-B/32 support more diverse attribution patterns due to their pre-training and deeper architectures, resulting in lower agreement and underscoring that model choice is as important as explainer choice when designing explainable systems, especially models pre-trained on face datasets [10].

6.2 Faithfulness Evaluation

Our results suggest ROAD [26] is the most reliable perturbation strategy in this setting in automatic neonatal pain assessment, and we believe, due to its nature, that it should also provide reasonable results in any task that involves faces. Since any other perturbation strategy like black, white, Gaussian noise, or mean replacements introduce sharp, artificial edges that CNNs are particularly highly sensitive to.

In contrast, ViT-B/32 relies on patch-level and attention-based mechanisms, making it more robust to perturbations, as shown in previous research [8,22] and confirmed here for neonatal pain assessment. Most perturbation strategies yielded positive AOPC values, indicating that ViT-B/32 remained stable even under random pixel disruptions.

At the model level, NCNN explainers were the least faithful to model behaviour, as indicated by AOPC values near zero in Table 8. As discussed pre-

viously, this likely reflects the model's shallow architecture, which is more sensitive to low-level artefacts introduced by perturbations, as well as the absence of large-scale pre-training. In contrast, VGGFace and ViT-B/32 showed very similar AOPC rankings, an interesting result given their markedly different architectures and pre-training. Although each model emphasized different facial features, the consistency of their highest-ranked explainers suggests that these explainers captured meaningful model behaviours when predicting pain.

Yet, as shown in the sanity checks, DeepLIFT and DeepLIFT SHAP despite appearing highly faithful to VGGFace and ViT-B/32, showed strong region agreement with their untrained counterparts, suggesting that part of their explanation may reflect architectural biases that affect model prediction rather than learned neonatal pain features, thus needing caution interpretation of their explanations.

For individual explainers, Grad-CAM, LIME, and Occlusion were consistently top-ranked in faithfulness, because their coarse, superpixel-based attributions perturb larger contiguous regions rather than isolated pixels, producing greater degradation during perturbation. However, these methods are highly sensitive to their chosen hyperparameters, and a dedicated evaluation of this sensitivity should be pursued in future work.

6.3 Agreement with Neonatal Pain Scales

Across models, explainers, and scales, agreement scores ranged from 0.60 to 0.70 on a normalized scale where 0.50 represents random agreement. This shows that the models prioritized clinically pain-relevant regions, despite these scales not being part of the training data. Although, they still attributed meaningful importance to areas outside traditional scales.

The facial regions that contributed positively to agreement were those that neonatal pain experts routinely monitor: mouth, eyes, forehead, and the region between the eyes. These areas encode action units such as brow bulge, eye squeeze, and open mouth, which are behind used scales [33]. At the same time, the nose, cheeks, and areas outside the face were often ranked as relevant by the models, even though they are not explicitly present in any of the scales.

We see two main interpretations. On one hand, these regions might carry genuine but unmodelled pain information, especially the nose and cheeks areas. On the other hand, they were also locations where blankets, monitoring equipment, and lighting artefacts frequently appear in NICU images. Without additional evidence, we argue that attributions outside scale-defined facial areas should be treated with caution and regarded as hypotheses rather than validated pain indicators.

6.4 Implications for Neonatal Pain Monitoring

Our findings carry several implications for the development and safe deployment of explainable neonatal pain assessment in the NICU. Models whose explainers demonstrate high faithfulness are preferable for clinical auditing because they

more accurately reveal the features that drive predictions. However, faithfulness alone is insufficient. Agreement across explainers and sanity checks remain essential to detect duplicated behaviour, uninformative saliency patterns, or failures of the explainer to meaningfully reflect internal model reasoning.

The ViT-B/32 results illustrated why explainability must be considered early in system design. Although its explainers were highly faithful, showing which pixels influenced the prediction, they frequently highlighted regions outside the face. In the NICU, where monitoring equipment occupy much of the frame, such behaviour poses a safety risk. The explainer may be "correct" in revealing that the model is focusing on the wall or a tube, yet the model itself is relying on contextual confounders rather than facial indicators of pain. We believe the use of pre-trained ViT on facial datasets may improve alignment with facial features.

Explanations must also be presented in clinician-centred formats. Region-level summaries, grounded in semantic facial areas rather than raw saliency pixels, offer a clearer bridge between continuous AI monitoring and intermittent clinical scoring. When aggregated over time, these regional attributions can form a "pain sign" [13], simultaneously reporting pain probability and the facial features supporting that estimate. Such outputs may support real-time cross-checking of AI predictions against bedside observations and help guide decisions around analgesia and procedure timing.

Our analysis also revealed operational constraints for reliable model use. Faithfulness linked to specific facial regions suggests that the mouth and nose (identified as the most influential areas) should remain unobstructed during model assessment. Practical rules such as "do not use the system if more than X% of the mouth is occluded" could clarify the system's limitations and align with efforts to weight facial regions according to their diagnostic importance [15]. These constraints are crucial in NICU environments where feeding tubes, eye patches, and hands frequently obscure critical facial cues.

Finally, the contrasting behaviour of NCNN underscored why both alignment and faithfulness must be jointly evaluated. Although NCNN matched pain-scales, highlighting clinically relevant regions, its faithfulness scores were near zero, indicating that perturbing those "important" pixels were no different from random perturbation. This paradox, where the model is spatially correct but mechanistically unstable, suggests sensitivity to noise or low-level artefacts rather than robust facial features. Identifying such fragile characteristics is essential for safe deployment. Persistent relevance assigned to non-facial areas or inconsistent alignment with pain scales should trigger model recalibration or human review, consistent with regulatory expectations such as the EU AI Act's requirements for human oversight and monitoring of explanation patterns.

6.5 Limitations

We acknowledge limitations in our work, particularly regarding generalizability to diverse NICU populations. Our findings were based on a controlled set of 483 images that are not publicly available, which limits reproducibility. Moreover, the integration of XAI into routine clinical practice and its real-world performance

remain largely unexplored due to limited access to real NICU environments. Accordingly, our study focused on establishing the theoretical foundations needed for trustworthy and responsible use of AI in neonatal pain detection.

7 Conclusion

Across three model architectures and ten explainers, pixel-level attributions showed substantial variability, whereas region-based explanations consistently converged on clinically meaningful facial areas. Explainer faithfulness was highly dependent on the perturbation strategy, with ViT-B/32 being more robust to perturbations than NCNN or VGGFace, however the ROAD method proved to be a robust default for our domain. All models attended to key facial regions defined by neonatal pain scales (mouth, eyes, forehead, and the region between the eyes), while attention to areas outside these scales (nose, cheeks, and outside the face regions) requires cautious interpretation. This raises an important question for safety-critical applications: whether to prioritize models that align most closely with expert-defined constructs (NCNN) or those that are more predictive and exhibit stronger explainer faithfulness to model behaviour (VGGFace and ViT-B/32), even when they rely on features beyond existing pain scales.

7.1 Future Directions

Future work should expand this research along several directions. Larger, more diverse, and publicly accessible datasets that reflect real NICU variability would enable more robust evaluation of model behaviour across clinical contexts. Extending faithfulness analyses to include clinically realistic occlusions (such as tubes or masks) may further clarify how models respond to common NICU scenarios. Extending this framework to video models, multimodal inputs, and region-regularised training may further improve stability and clinical usefulness. While concept-based explainers are promising, semantically defined facial regions remain more interpretable and better aligned with established clinical knowledge, though future concept discovery may still reveal additional pain-relevant features.

Acknowledgments. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The authors acknowledge the National Laboratory for Scientific Computing (LNCC/MTCI, Brazil) for providing HPC resources of the SDumont supercomputer, which have contributed to the research results reported within this paper.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Agreement with Neonatal Pain Scales

Table 9: Average agreement between pain scales and explainers for each model.

(a) NCNN

Pain Scale	Deconv.	DeepLIFT	DeepLIFT SHAP	Grad-CAM	Grad. SHAP	Guided Grad-CAM	Int. Grad.	LIME	Occlusion	Saliency	Average
APS	0.62	0.57	0.57	0.52	0.65	0.56	0.63	0.50	0.63	0.65	0.59
BIIP	0.74	0.73	0.73	0.69	0.78	0.65	0.75	0.70	0.75	0.78	0.73
EVENDOL	0.66	0.65	0.65	0.60	0.72	0.54	0.69	0.62	0.66	0.72	0.65
LIDS	0.72	0.73	0.73	0.71	0.77	0.66	0.74	0.69	0.77	0.77	0.73
MCGRATH	0.76	0.78	0.78	0.78	0.79	0.71	0.77	0.74	0.81	0.79	0.77
MIPS	0.76	0.78	0.78	0.78	0.79	0.71	0.77	0.74	0.81	0.79	0.77
M-NFCS	0.74	0.73	0.73	0.69	0.78	0.65	0.75	0.70	0.75	0.78	0.73
NFCS	0.76	0.78	0.78	0.78	0.79	0.71	0.77	0.74	0.81	0.79	0.77
NNICUPAT	0.74	0.73	0.73	0.69	0.78	0.65	0.75	0.70	0.75	0.78	0.73
PASPI	0.74	0.73	0.73	0.69	0.78	0.65	0.75	0.70	0.75	0.78	0.73
PIPP	0.63	0.62	0.62	0.59	0.68	0.53	0.65	0.59	0.65	0.68	0.62
PIPP-R	0.63	0.62	0.62	0.59	0.68	0.53	0.65	0.59	0.65	0.68	0.62
P-MIPS	0.76	0.78	0.78	0.78	0.79	0.71	0.77	0.74	0.81	0.79	0.77
Average	0.71	0.71	0.71	0.68	0.75	0.64	0.73	0.67	0.74	0.75	0.71

(b) VGGFace

Pain Scale	Deconv.	DeepLIFT	DeepLIFT SHAP	Grad-CAM	Grad. SHAP	Guided Grad-CAM	Int. Grad.	LIME	Occlusion	Saliency	Average
APS	0.56	0.56	0.56	0.34	0.43	0.46	0.43	0.56	0.48	0.55	0.49
BIIP	0.71	0.70	0.70	0.54	0.71	0.56	0.71	0.68	0.68	0.68	0.67
EVENDOL	0.67	0.65	0.65	0.43	0.69	0.51	0.69	0.59	0.62	0.65	0.62
LIDS	0.71	0.69	0.69	0.57	0.69	0.60	0.69	0.69	0.64	0.72	0.67
MCGRATH	0.70	0.70	0.70	0.64	0.69	0.59	0.69	0.73	0.68	0.68	0.68
MIPS	0.70	0.70	0.70	0.64	0.69	0.59	0.69	0.73	0.68	0.68	0.68
M-NFCS	0.71	0.70	0.70	0.54	0.71	0.56	0.71	0.68	0.68	0.68	0.67
NFCS	0.70	0.70	0.70	0.64	0.69	0.59	0.69	0.73	0.68	0.68	0.68
NNICUPAT	0.71	0.70	0.70	0.54	0.71	0.56	0.71	0.68	0.68	0.68	0.67
PASPI	0.71	0.70	0.70	0.54	0.71	0.56	0.71	0.68	0.68	0.68	0.67
PIPP	0.61	0.60	0.60	0.45	0.64	0.44	0.64	0.57	0.65	0.67	0.59
PIPP-R	0.61	0.60	0.60	0.45	0.64	0.44	0.64	0.57	0.65	0.67	0.59
P-MIPS	0.70	0.70	0.70	0.64	0.69	0.59	0.69	0.73	0.68	0.68	0.68
Average	0.68	0.67	0.67	0.54	0.67	0.54	0.67	0.66	0.65	0.67	0.64

(c) ViT-B/32

Pain Scale	Deconv.	DeepLIFT	DeepLIFT SHAP	Grad-CAM	Grad. SHAP	Guided Grad-CAM	Int. Grad.	LIME	Occlusion	Saliency	Average
APS	0.27	0.43	0.43	0.43	0.30	0.35	0.31	0.52	0.44	0.26	0.37
BIIP	0.57	0.61	0.61	0.55	0.59	0.59	0.59	0.68	0.65	0.60	0.60
EVENDOL	0.57	0.56	0.56	0.46	0.54	0.50	0.54	0.59	0.62	0.60	0.55
LIDS	0.63	0.63	0.63	0.61	0.58	0.61	0.58	0.69	0.71	0.66	0.63
MCGRATH	0.59	0.67	0.67	0.66	0.61	0.68	0.61	0.76	0.73	0.62	0.66
MIPS	0.59	0.67	0.67	0.66	0.61	0.68	0.61	0.76	0.73	0.62	0.66
M-NFCS	0.57	0.61	0.61	0.55	0.59	0.59	0.59	0.68	0.65	0.60	0.60
NFCS	0.59	0.67	0.67	0.66	0.61	0.68	0.61	0.76	0.73	0.62	0.66
NNICUPAT	0.57	0.61	0.61	0.55	0.59	0.59	0.59	0.68	0.65	0.60	0.60
PASPI	0.57	0.61	0.61	0.55	0.59	0.59	0.59	0.68	0.65	0.60	0.60
PIPP	0.52	0.49	0.49	0.44	0.46	0.49	0.46	0.56	0.53	0.53	0.50
PIPP-R	0.52	0.49	0.49	0.44	0.46	0.49	0.46	0.56	0.53	0.53	0.50
P-MIPS	0.59	0.67	0.67	0.66	0.61	0.68	0.61	0.76	0.73	0.62	0.66
Average	0.55	0.59	0.59	0.56	0.55	0.58	0.55	0.67	0.64	0.57	0.59

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
4. Beam, K., Sharma, P., Levy, P., Beam, A.L.: Artificial intelligence in the neonatal intensive care unit: the time is now. Journal of Perinatology **44**(1), 131–135 (2024)
5. Brahnam, S., Chuang, C.F., Shih, F.Y., Slack, M.R.: Machine recognition and representation of neonatal facial displays of acute pain. Artificial intelligence in medicine **36**(3), 211–222 (2006)
6. Carlini, L.P., Coutrin, G.d.A.S., Ferreira, L.A., Soares, J.d.C.A., Silva, G.V.T., Heiderich, T.M., Balda, R.d.C.X., de Moraes Barros, M.C., Guinsburg, R., Thomaz,

C.E.: Human vs machine towards neonatal pain assessment: A comprehensive analysis of the facial features extracted by health professionals, parents, and convolutional neural networks. Artificial Intelligence in Medicine **147**, 102724 (2024)
7. Carlini, L.P., Ferreira, L.A., Coutrin, G.A., Varoto, V.V., Heiderich, T.M., Balda, R.C., Barros, M.C., Guinsburg, R., Thomaz, C.E.: A convolutional neural network-based mobile application to bedside neonatal pain assessment. In: 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). pp. 394–401. IEEE (2021)
8. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 782–791 (2021)
9. Chen, H., Gomez, C., Huang, C.M., Unberath, M.: Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ digital medicine **5**(1), 156 (2022)
10. Coutrin, G.A., Carlini, L.P., Ferreira, L.A., Heiderich, T.M., Balda, R.C., Barros, M.C., Guinsburg, R., Thomaz, C.E.: Convolutional neural networks for newborn pain assessment using face images: A quantitative and qualitative comparison. In: Su, R., Zhang, Y., Liu, H., F Frangi, A. (eds.) International Conference on Medical Imaging and Computer-Aided Diagnosis. pp. 503–513. Springer Nature Singapore, Singapore (2023)
11. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5202–5211 (2020)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Ferreira, L.A., Carlini, L.P., Coutrin, G.d.A.S., Heiderich, T.M., Balda, R.d.C.X., de Moraes Barros, M.C., Guinsburg, R., Thomaz, C.E.: Disclosing neonatal pain in real-time: Ai-derived pain sign from continuous assessment of facial expressions. Computers in Biology and Medicine **189**, 109908 (2025)
14. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision making and a "right to explanation". AI Magazine **38**(3), 50–57 (2017)
15. Heiderich, T.M., Carlini, L.P., Buzuti, L.F., Balda, R.C., Barros, M.C., Guinsburg, R., Thomaz, C.E.: Face-based automatic pain assessment: challenges and perspectives in neonatal intensive care units. Jornal de Pediatria **99**(6), 546–560 (2023)
16. Heiderich, T.M., Leslie, A.T.F.S., Guinsburg, R.: 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 (2015)
17. Hossain, M.I., Zamzmi, G., Mouton, P.R., Salekin, M.S., Sun, Y., Goldgof, D.: Explainable ai for medical data: current methods, limitations, and future directions. ACM Computing Surveys **57**(6), 1–46 (2025)
18. Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602 (2022)
19. Llerena, A., Tran, K., Choudhary, D., Hausmann, J., Goldgof, D., Sun, Y., Prescott, S.M.: Neonatal pain assessment: Do we have the right tools? Frontiers in Pediatrics **10**, 1022751 (2023)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

21. Parkhi, O., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC 2015-Proceedings of the British Machine Vision Conference 2015. pp. 1–12. British Machine Vision Association (2015)
22. Paul, S., Chen, P.Y.: Vision transformers are robust learners. In: Proceedings of the AAAI conference on Artificial Intelligence. vol. 36, pp. 2071–2081 (2022)
23. Pirie, C., Ferreira, L.A., Coutrin, G.d.A.S., Carlini, L.P., Moreno-García, C.F., Barros, M.C.d.M., Guinsburg, R., Thomaz, C.E., Nobre, R., Wiratunga, N.: Understanding disagreement between humans and machines in xai: Robustness, fidelity, and region-based explanations in automatic neonatal pain assessment. In: World Conference on Explainable Artificial Intelligence. pp. 274–298. Springer (2026)
24. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
25. Rieger, L., Hansen, L.K.: Irof: a low resource evaluation metric for explanation methods. In: Workshop AI for Affordable Healthcare at ICLR 2020 (2020)
26. Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E.: A consistent and efficient evaluation strategy for attribution methods. In: International Conference on Machine Learning. pp. 18770–18795. PMLR (2022)
27. Salekin, M.S., Zamzmi, G., Goldgof, D., Kasturi, R., Ho, T., Sun, Y.: Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. Computers in biology and medicine **129**, 104150 (2021)
28. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems **28**(11), 2660–2673 (2016)
29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
30. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMlR (2017)
31. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
32. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
33. Tamanaka, F.G., Carlini, L.P., Heiderich, T.M., Balda, R.C., Barros, M.C., Guinsburg, R., Thomaz, C.E.: Neonatal pain assessment: A kendall analysis between clinical and visually perceived facial features. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(3), 331–340 (2023)
34. Yeh, C.K., Hsieh, C.Y., Suggala, A., Inouye, D.I., Ravikumar, P.K.: On the (in) fidelity and sensitivity of explanations. Advances in neural information processing systems **32** (2019)
35. Zamzmi, G., Paul, R., Goldgof, D., Kasturi, R., Sun, Y.: Pain assessment from facial expression: Neonatal convolutional neural network (n-cnn). In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2019)
36. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)

APÊNDICE J – CAPÍTULO PUBLICADO NO LIVRO *GRANDES DESAFIOS DA COMPUTAÇÃO APLICADA À SAÚDE*

Chapter 1

Barriers and Pathways to Clinical Translation of AI in Automated Neonatal Pain Assessment

Leonardo Antunes Ferreira (FEI, SBC, Brazil)
Lucas Pereira Carlini (FEI, SBC, Brazil)
Gabriel de Almeida Sá Coutrin (FEI, SBC, Brazil)
Lucas Fontes Buzuti (FEI, SBC, Brazil)
Roberto Gonçalves de Magalhães Júnior (FEI, SBC, Brazil)
Rafael Nobre Orsi (CEETEPS, SP, Brazil)
Tatiany Marcondes Heiderich (FEI, SBC, Brazil)
Gabriela Cianciarullo (FEI, SBC, Brazil)
Salvador Pinillos Gimenez (FEI, SBC, Brazil)
Craig Pirie (Robert Gordon University, Aberdeen, U.K.)
Carlos Francisco Moreno-García (Robert Gordon University, Aberdeen, U.K.)
Marina Carvalho de Moraes Barros (UNIFESP, SP, Brazil)
Ruth Guinsburg (UNIFESP, SP, Brazil)
Carlos Eduardo Thomaz (FEI, SP, Brazil)

Abstract

This work highlights the grand challenge of developing trustworthy Artificial Intelligence (AI) systems for neonatal pain assessment, structured around three pillars: (I) standardized, privacy-preserving data acquisition; (II) clinically safe and reliable AI; and (III) multicentre validation. We argue that predictive performance alone is insufficient for clinical translation. Safe clinical deployment requires explainability, calibration, uncertainty estimation, temporal modelling, equity-aware evaluation, and robust acquisition infrastructures. A multidimensional metric framework is proposed to track technical, clinical, systemic, and equity-related progress over the next decade. Only a coordinated and multidisciplinary effort can enable responsible and sustainable AI-driven improvements in automated neonatal pain management.

1.1. Introduction

Pain during the neonatal period represents a major clinical concern rather than a niche issue. Observational studies and systematic reviews indicate that newborns in intensive

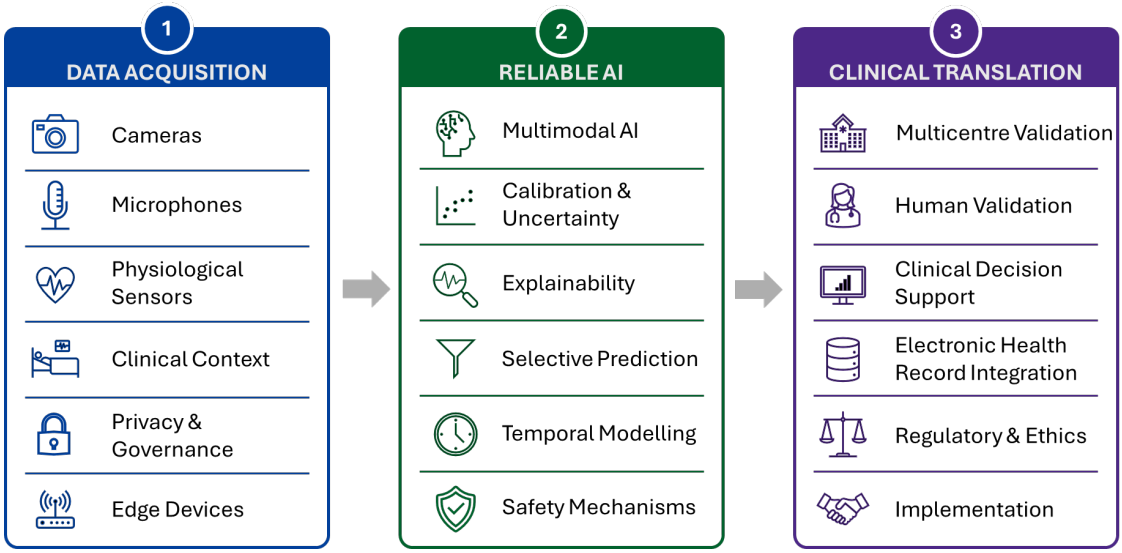


Figure 1.1. The three pillars of trustworthy AI-based neonatal assessment.

care are exposed to multiple painful procedures each day [Perry et al. 2018; Llerena et al. 2023], reinforcing the need for systematic pain recognition and consistent implementation of evidence-based analgesia and comfort management strategies in Neonatal Intensive Care Units (NICUs) [Sunwoo and El-Dib 2026]. However, substantial variability persists across institutions in the application of these practices. In clinical routine, neonatal pain assessment still relies predominantly on intermittent, observer-based scales that evaluate facial expressions and body movements during potentially painful events. These methods exhibit considerable inter-observer variability and inconsistent reporting practices [Llerena et al. 2023; Sunwoo and El-Dib 2026], limiting their suitability for continuous monitoring and timely intervention. Beyond the immediate clinical burden, repeated exposure to pain in preterm infants has been associated with alterations in brain development and later neurodevelopmental outcomes, including measurable differences in brain structure, connectivity, and subsequent cognitive, behavioural, and pain responses [R. E. Grunau 2013].

In Brazil, neonatal pain management is closely linked to the national burden of preterm birth, neonatal morbidity and mortality, and structural inequalities [Victor et al. 2025]. National analyses report persistent regional and sociodemographic disparities in preterm birth trends, and population-based studies demonstrate ethnic-racial inequalities in adverse birth and neonatal outcomes [Victor et al. 2025]. At the same time, the Brazilian Ministry of Health is consolidating a national digital health infrastructure with the development of national interoperability frameworks [Ministério da Saúde 2020], thereby creating a policy window for neonatal digital innovation as a coordinated, equity-oriented national program.

Automated pain assessment should therefore not be conceptualized solely as a computational modelling task, but as a multimodal sociotechnical challenge. It involves integrating facial, audio, physiological, and contextual data with governance frameworks, regulatory requirements, and clinical workflows. Although technical feasibility has al-

ready been demonstrated from a computational perspective [Salekin et al. 2021; Ferreira et al. 2025], safe clinical deployment requires a systems-level architecture rather than incremental gains in predictive performance. In neonatology, where errors may have long-term consequences for patients and families, AI systems must extend beyond accuracy to incorporate interpretability, calibration, uncertainty quantification, and rigorous validation across heterogeneous and extremely sensitive NICU populations.

The central challenge for the next decade is to design, validate, and build a trustworthy, continuously operating, and regulatorily compliant AI infrastructure (hardware and software) for neonatal pain assessment that is clinically actionable, probabilistically reliable, equitable across populations, and demonstrably robust under real-world conditions.

This paper is structured into five sections. The first three examine the challenges and potential pathways related to three interdependent pillars essential for the transition from research to clinical implementation: data (Section 2), reliability and safety (Section 3), and multicentre validation (Section 4). Section 5 proposes a multidimensional metric framework to track technical, clinical, systemic, and equity-related progress over the next decade. Section 6 concludes the paper.

1.2. Pillar I: Standardized and Privacy-Preserving Data Acquisition

AI systems require large volumes of high-quality and representative data to learn robust patterns and generate reliable predictions. In neonatal pain assessment, however, datasets are typically limited in size, heterogeneous, and frequently restricted to individual institutions. This structural constraint limits generalizability, weakens reproducibility across independent studies, and ultimately hinders clinical translation.

Data acquisition in neonatal populations is constrained by ethics, law, and clinical feasibility. Research involving newborns requires institutional ethics approval and explicit parental informed consent under Brazilian regulations [Conselho Nacional de Saúde 2012], while intentional pain induction for research is unacceptable. Consequently, data collection must be limited to clinically indicated painful procedures, reducing sample diversity and making data acquisition dependent on routine NICU care.

Beyond ethical and governance constraints, data acquisition in neonatal environments depends on complex sensing infrastructures. Continuous monitoring systems typically rely on multimodal instrumentation, including cameras for facial analysis, microphones for cry detection, and physiological sensors for heart rate and oxygen saturation [Sanga, Fosah, and Ejuh 2025]. In practice, signal stability may be affected by motion artifacts, calibration errors, partial occlusions caused by medical equipment, and routine caregiving activities. Acquisition conditions in NICUs are therefore inherently noisy and heterogeneous, with variations in lighting and patient positioning that complicate dataset standardization.

Within this constrained environment, bias frequently originates from data acquisition processes rather than solely from algorithmic design. Single-centre datasets reflect local clinical routines, demographic characteristics, equipment availability, and documentation practices. Models trained on such data risk systematic underperformance when

deployed in underrepresented populations or regions. Documented regional and sociodemographic disparities in neonatal outcomes [Victor et al. 2025] further emphasize the need to explicitly address distributional heterogeneity when designing multicentre studies and evaluating deployment strategies.

Privacy introduces an additional layer of complexity because neonatal pain assessment involves highly sensitive health information and may include biometric identifiers such as facial images. Under the *Lei Geral de Proteção de Dados Pessoais* (LGPD), health and biometric data are classified as sensitive personal data and therefore require enhanced safeguards and clearly defined legal bases for processing and storage [Brasil 2018]. Continuous monitoring systems amplify exposure risks due to prolonged data capture, high temporal resolution, and the possibility of secondary data use beyond the originally authorized clinical or research purpose. Governance frameworks must therefore reconcile data utility with strict requirements for purpose limitation, long-term protection, traceability, and accountability.

The field therefore needs harmonized, privacy-preserving acquisition standards. These should define modality-specific protocols, device configurations, contextual metadata, and labelling procedures grounded in validated clinical scales such as the Neonatal Facial Coding System (NFCS) [R. V. Grunau and Craig 1987]. Synthetic data generation may reduce scarcity [Buzuti et al. 2024], but it cannot replace multicentre acquisition because it remains constrained by the source data distribution. Secure infrastructures, including protected on-site storage [World Health Organization B 2023], edge-based privacy-preserving representations [Heydari and Mahmoud 2025], and federated learning [Pati et al. 2024], provide a bridge between acquisition and validation without requiring unrestricted circulation of raw neonatal data.

1.3. Pillar II: Clinically Safe and Reliable AI

In neonatal care, a statement such as "the model predicted pain" is not clinically actionable unless clinicians understand the basis of the prediction, its uncertainty, and its implications within established protocols. From a regulatory perspective, explainability is also aligned with data protection frameworks such as the LGPD, which establishes the right to review decisions based solely on automated processing [Brasil 2018]. Explainability, calibration, and rejection mechanisms should therefore be treated as safety requirements rather than optional enhancements [World Health Organization A 2021].

Current research often reports incremental classification gains on non-standardized and frequently private datasets [Heiderich et al. 2023]. Without shared benchmarks and clinically meaningful evaluation criteria, such gains do not establish clinical utility. The reliability pillar must therefore connect model development to the acquisition standards described above (Pillar I) and to the external validation protocols described below (Pillar III).

Explainability is commonly pursued through eXplainable Artificial Intelligence (XAI) techniques that attribute relevance to input features. In image-based systems, heatmaps are often presented as evidence of plausible focus regions. However, empirical studies have shown that some XAI methods fail to faithfully represent internal model reasoning [Adebayo et al. 2018]. Furthermore, the coexistence of multiple explainers,

including Grad-CAM, SHAP, and LIME, introduces the disagreement problem, whereby different methods highlight distinct regions for the same input and model [Pirie et al. 2025]. Visual plausibility alone does not guarantee clinical validity, and uncritical reliance on these heatmaps may lead to overinterpretation [Ghassemi, Oakden-Rayner, and Beam 2021]. Future research should therefore prioritize concept-based and clinically grounded approaches that align computational representations with established pain assessment constructs, such as facial action units or validated clinical pain scales.

Reliable deployment also requires calibrated probabilities, uncertainty-aware outputs, selective abstention, and temporal modelling. Calibration links predicted probabilities to observed outcome frequencies [Guo et al. 2017], uncertainty estimation identifies low-confidence cases [Kendall and Gal 2017], selective prediction defers unreliable cases to expert review [Geifman and El-Yaniv 2017], and temporal representations capture pain trajectories across procedures and recovery phases, capable of supporting early warning signals and clinically actionable interventions [Ferreira et al. 2025].

Recent advances in foundation models and large language models (LLMs) expand the capacity to process heterogeneous and unstructured data, including facial video, audio signals, physiological measurements, and clinical documentation [Bommasani et al. 2021]. Multimodal architectures may support integrated reasoning, while LLMs may transform validated outputs into structured summaries for electronic health records or clinical decision support systems [Buzuti et al. 2024; Pereira Carlini et al. 2025]. Nevertheless, AI-generated outputs must undergo explicit human validation before inclusion in clinical records or automated decision-making, as healthcare professionals retain ultimate responsibility for clinical decisions under Brazilian Resolution CFM No. 2.454[1].

A modular and hierarchical architecture offers a promising design strategy. Task-specific predictive models can generate calibrated predictions, uncertainty estimates, and clinically grounded explanations. Safety mechanisms enforce predefined reliability thresholds through selective prediction. A higher-level reasoning component, such as LLMs can synthesize these outputs into structured summaries adapted to clinical context and professional roles. Within such systems, any personalization mechanisms must remain transparent in order to prevent hidden biases and preserve accountability.

Even with advances in algorithmic safety, AI systems for NICUs must operate under strict computational and spatial constraints. Operational feasibility therefore becomes a central requirement. Edge inference, model compression, and hardware-aware optimization are essential for reliable deployment in resource-constrained clinical settings [Shi et al. 2016; Heydari and Mahmoud 2025]. Neonatal monitoring platforms often rely on distributed architectures in which embedded devices handle signal acquisition and lightweight preprocessing, while computationally intensive tasks, including deep learning inference, multimodal fusion, and temporal analysis, are executed on hospital servers or external workstations [Sanga, Fosah, and Ejuh 2025]. These systems must also ensure secure communication and management of sensitive neonatal data across devices and clinical infrastructures.

[1] https://www.in.gov.br/web/dou/-/resolucao-cfm-n-2.454-de-11-de-fevereiro-de-2026-689247948

1.4. Pillar III: Multicentre Validation and Clinical Translation

The absence of multicentre validation remains a central barrier to the clinical legitimacy of AI systems for neonatal pain assessment. Favourable results obtained in single-centre studies are insufficient to justify deployment in NICU environments characterized by variability in patient populations, clinical practices, equipment, staffing, and environmental conditions. In Brazil, this variability is further amplified by geographic, socioeconomic, and infrastructural heterogeneity [Victor et al. 2025].

Single-centre datasets encode local clinical practices, equipment configurations, documentation standards, and patient case-mixing. Scalable deployment therefore requires empirical evidence across regions and hospital types that differ in operational routines and demographic composition. Although pain in neonates is a universal physiological phenomenon [Schiavenato et al. 2008], dataset bias related to ethnicity, imaging conditions, clinical protocols, and caregiving practices can compromise external validity.

Addressing these challenges requires sustained coordination among hospitals, universities, and research networks operating within interoperable governance frameworks aligned with LGPD requirements. Standardized experimental designs should become routine practice, including leave-one-centre-out validation and cross-regional testing. Such protocols quantify performance degradation under distribution shift, estimate model transportability, and reveal centre-specific vulnerabilities that may remain hidden in single-centre evaluations.

Because unrestricted data sharing is often unfeasible, privacy-preserving benchmarking mechanisms are essential. Federated evaluation frameworks [Pati et al. 2024] and secure research environments can enable cross-site assessment while complying with legal and ethical constraints. Interoperability standards, harmonized metadata schemas, and consistent labelling procedures are critical for enabling reproducible multicentre analyses. In addition, heterogeneity in hardware configurations, staffing patterns, workflow organization, and environmental conditions should be systematically documented to support meaningful cross-site interpretation of results.

Existing national infrastructures provide an opportunity to operationalize such collaboration. Networks such as the Rede Brasileira de Pesquisas Neonatais[2] could support an AI-oriented extension defining shared reference datasets, standardized reporting templates, and minimum evaluation criteria. These criteria should include calibration metrics, uncertainty estimation procedures, selective rejection policies, and temporal evaluation protocols. Multicentre validation should then feed back into model refinement and acquisition standards, ensuring that deployment evidence remains iterative rather than a one-time endpoint.

1.5. Metrics for Tracking Progress Over the Next Decade

This section proposes metrics for monitoring the roadmap over the next decade. The goal is to move beyond isolated classification results and evaluate whether acquisition standards, reliability mechanisms, validation evidence, workflow integration, and equity safeguards are improving together.

[2]https://redeneonatal.com.br/

1.5.1. Technical Metrics

Calibration: Monitor Expected Calibration Error (ECE) and Brier score across institutions and over time, given the known susceptibility of neural networks to miscalibration under distribution shift [Guo et al. 2017]. Reliable probability estimates are essential for clinical decision-making and risk communication.

Uncertainty quantification: Evaluate uncertainty estimates under controlled perturbations and predefined distribution shifts [Ovadia et al. 2019]. Selective prediction should be quantified using risk–coverage curves and coverage levels at fixed risk thresholds, explicitly assessing abstention behaviour in safety-critical scenarios [Kendall and Gal 2017].

Explainability: Evaluate the clinical plausibility and stability of explanations produced by XAI methods. Quantitative assessment should include metrics to measure the causal contribution of predominant features to model predictions [Pirie et al. 2025].

1.5.2. Clinical Metrics

Concordance: Measure agreement between AI predictions and validated neonatal pain scales, enabling systematic comparison between computational outputs and established clinical knowledge. Such concordance metrics can support bidirectional learning between clinicians and intelligent systems.

Decision impact: Evaluate the effect of AI systems on clinical workflows and decision-making. Relevant indicators may include reductions in time to analgesia reassessment, decreases in unrecognized pain episodes, and reductions in unnecessary interventions.

1.5.3. Systemic Metrics

Adoption and coverage: Monitor the number and proportion of NICUs integrating AI-based monitoring systems, including geographic distribution, system uptime, and degree of integration with clinical workflows and information systems.

Cost-effectiveness: Estimate cost per NICU bed, maintenance costs, incremental staff workload, and computational requirements under edge-computing constraints [Heydari and Mahmoud 2025].

Governance compliance: Document the proportion of sites implementing data protection impact assessments, audit logging mechanisms, monitoring dashboards, and model update governance aligned with medical device lifecycle principles [World Health Organization A 2021].

1.5.4. Equity Metrics

Stratified performance: Report predictive performance, calibration quality, uncertainty estimates, and interpretability analyses stratified by relevant demographic and contextual factors, including region, facility type, gestational age, and sex. Where ethically justified, stratification by skin tone categories may also be considered to monitor known risks of algorithmic bias [Obermeyer et al. 2019].

Coverage equity under rejection: Ensure that selective prediction mechanisms do not disproportionately reduce coverage in under-resourced populations or regions. Risk–coverage

profiles should therefore be stratified to verify equitable behaviour across groups [Geifman and El-Yaniv 2017].

1.6. Conclusion

The grand challenge for the coming decade is to operationalize a clinically safe, regulatorily compliant, and continuously monitored AI infrastructure for neonatal pain assessment. The unresolved problem is not only technical accuracy, but the absence of an integrated translation pathway linking reliable ground-truth practices, sensitive data acquisition, privacy-preserving governance, calibrated and uncertainty-aware modelling, clinical interpretability, multicentre validation, and hospital workflow integration.

The roadmap proposed here treats standardized data acquisition, reliable AI, and multicentre validation as mutually reinforcing pillars. Acquisition standards make model evidence reproducible, calibrated and rejectable models make clinical review safer, external validation exposes weaknesses that should refine both datasets and algorithms, and post-deployment monitoring closes the lifecycle through recalibration, audit, and governance revision. Progress should therefore be assessed through reliability, transparency, equity, governance, clinical impact, and sustainability, not through predictive accuracy alone. Only sustained collaboration among clinicians, engineers, ethicists, regulators, hospital managers, and public digital health programs can transform automated neonatal pain assessment from promising research into responsible clinical infrastructure.

About the Proponents

Leonardo A. Ferreira is currently pursuing a PhD degree in Electrical Engineering, with an emphasis on image and signal processing, at FEI. Since 2020, he has conducted research on the application of AI to neonatal pain assessment, with a focus on explainability, uncertainty quantification, and probabilistic reliability in clinical AI systems.

Lucas P. Carlini is an Applied Scientist at Amazon in São Paulo and holds an MSc in Electrical Engineering at FEI. His work focuses on LLMs and AI agent systems for multilingual content generation, while he also conducts academic research on automatic neonatal pain assessment. He has authored more than 27 publications across venues including *Pediatric Research*, *Artificial Intelligence in Medicine*, CVPR, and MICCAI, and received the National Academy of Medicine's Fernandes Figueira Award.

Gabriel A. S. Coutrin holds a Master's degree in Electrical Engineering from FEI, with emphasis on image and signal processing. Since 2020, he has been engaged in research on AI methods for neonatal pain assessment.

Lucas F. Buzuti is an AI Researcher at the Samsung R&D Institute Brazil and a member of the Image Processing Lab at FEI. He received his BSc degree in Control and Automation Engineering from FEI, in 2018. He subsequently obtained his MSc in Electrical Engineering in 2020 and his PhD in Electrical Engineering in March 2025, both from the same institution. His main research interests include Pattern Recognition, Self-Supervised Learning, Deep Learning, Generative Models, and Large Vision Models.

Roberto G. M. Júnior received his BSc degree in Automation and Control Engineering, and his MSc and DSc degrees in Electrical Engineering from FEI, São Paulo, Brazil, in

2016, 2019 and 2026 respectively. His research interests include Cognitive Perception, Pattern Recognition and Discrete Mathematics.

Rafael N. Orsi is an Associate Professor and Researcher at the Paula Souza Center - CEETEPS, Government of the State of Sao Paulo. He holds the MSc and DSc degrees in Electrical Engineering from FEI, Sao Paulo, Brazil, and the BSc degree in Mathematics. His research interests include Artificial Intelligence, Machine Learning, and Statistical methods for Pattern Recognition.

Tatiany M. Heiderich is Physiotherapist (2000) from Universidade de Mogi das Cruzes (UMC), Sao Paulo, Brazil, Master's in Biomedical Engineering (UMC, 2006), PhD in Pediatric Sciences (2015) from Escola Paulista de Medicina of the Federal University of São Paulo (EPM-UNIFESP), and PhD in Electrical Engineering from FEI (2024). Her research interest is in automated neonatal pain assessment.

Gabriela Cianciarullo is a Biomedical Engineer and a PhD student in Electrical Engineering at FEI. She earned her BSc in Biomedical Engineering from Pontifícia Universidade Católica de São Paulo, in 2025. Her research focuses on nanoelectronics and integrated circuits, with a particular emphasis on biomedical signal processing, embedded systems, and AI applied to healthcare, especially for multimodal neonatal pain assessment. Her main research interests include medical devices, assistive technologies, and AI-driven healthcare systems.

Salvador P. Gimenez is an Electrical Engineer (UMC, 1984), MSc (1990), and PhD (2004) in Electrical Engineering from the Polytechnic School of the University of São Paulo (EPUSP). Full Professor at FEI and Assistant Professor at PUC-SP, working in Nano/Microelectronics, CMOS integrated circuits, and microprocessors/microcontrollers. CNPq Research Productivity Fellow (PQ-2) and member of the SBMicro Council (2025–2027). Author of several books on microcontrollers and power electronics, with industrial experience at DIMEP, TRACECOM, and Ford - Electronic Division/Visteon. Holder of patents on innovative SOI MOSFET structures and biomedical devices. Founder of the startup MTG2i Solutions (2019), supported by FAPESP PIPE, focused on analog and RF CMOS IC design and technological applications.

Craig Pirie is a Research Associate and PhD researcher in Artificial Intelligence at the School of Computing, Engineering and Technology at Robert Gordon University. His research focuses on eXplainable Artificial Intelligence (XAI), particularly the evaluation and reconciliation of disagreements between explanation methods such as feature attribution and counterfactual explanations. His work explores explanation robustness and coherence, and develops new methods for evaluating and improving explanation quality.

Carlos F. Moreno-García is an Associate Professor at the School of Computing, Engineering and Technology at Robert Gordon University, Aberdeen, Scotland, UK. His research is centred around pattern recognition, computer vision and visualisation to tackle problems in the energy, healthcare and food industries, amongst others. He has been the recipient of numerous awards and grants, including CONACyT Doctoral Programme, MINECO (Spain) national projects, The Data Lab, UKRI NERC and KTP, and the British Council Newton Fund, amongst others.

Marina C. M. Barros graduated in Medicine from the Escola Paulista de Medicina - Fed-

eral University of São Paulo, with specializations in Pediatrics (1990) and Neonatology (1992), and Master's (1996) and Doctorate (2005) from the same institution, including Master's in Business Administration from Ibmec - São Paulo (2009). She is Professor of the Postgraduate Program in Pediatrics and Sciences Applied to Pediatrics (since 2011) and Executive Editor of the Editorial Board of Revista Paulista de Pediatria since July 2015.

Ruth Guinsburg is Full Professor of Neonatal Medicine at Escola Paulista de Medicina of the Federal University of São Paulo (EPM-Unifesp), Head of the Neonatal Intensive Care Unit at São Paulo Hospital/EPM-Unifesp, Co-chair of the Neonatal Resuscitation Program of the Brazilian Society of Pediatrics, Scientific Coordinator of the Brazilian Network on Neonatal Research and Emeritus Editor of Revista Paulista de Pediatria.

Carlos E. Thomaz is Full Professor of Electrical Engineering at FEI, São Paulo, Brazil. He is Head of the Image Processing Lab at FEI. He received, in 1993, his BSc degree in Electronic Engineering from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil. After working for six years in industry, he obtained the MSc degree in Electrical Engineering from PUC-Rio in 1999. In October 2000, he joined the Department of Computing at Imperial College London, UK, where he obtained the PhD degree in Computing in 2004. His main research interests are in Pattern Recognition, Cognitive Perception and Machine Learning.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) - 2018/13076-9.

References

Adebayo, Julius et al. (2018). "Sanity checks for saliency maps". In: *Advances in neural information processing systems* 31.

Bommasani, Rishi et al. (2021). "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258*.

Brasil (2018). *Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais – LGPD)*. https://www.gov.br/anpd/pt-br/centrais-de-conteudo/outros-documentos-e-publicacoes-institucionais/lgpd-en-lei-no-13-709-capa.pdf. Access in: 23 fev. 2026.

Buzuti, Lucas et al. (2024). "Generative AI for neonatal pain assessment: a sound approach to improve data-driven automatic recognition in intensive care unit". In: *Available at SSRN 4994076*.

Conselho Nacional de Saúde (2012). *Resolução nº 466, de 12 de dezembro de 2012*. https://bvsms.saude.gov.br/bvs/saudelegis/cns/2013/res0466_12_12_2012.html. Access in: 23 fev. 2026.

Ferreira, Leonardo Antunes et al. (2025). "Disclosing neonatal pain in real-time: AI-derived pain sign from continuous assessment of facial expressions". In: *Computers in Biology and Medicine* 189, p. 109908.

Geifman, Yonatan and Ran El-Yaniv (2017). "Selective classification for deep neural networks". In: *Advances in neural information processing systems* 30.

Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L Beam (2021). "The false hope of current approaches to explainable artificial intelligence in health care". In: *The lancet digital health* 3.11, e745–e750.

Grunau, Ruth Eckstein (2013). "Neonatal pain in very preterm infants: long-term effects on brain, neurodevelopment and pain reactivity". In: *Rambam Maimonides medical journal* 4.4, e0025.

Grunau, Ruth VE and Kenneth D Craig (1987). "Pain expression in neonates: facial action and cry". In: *Pain* 28.3, pp. 395–410.

Guo, Chuan et al. (2017). "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR, pp. 1321–1330.

Heiderich, Tatiany M. et al. (2023). "Face-based automatic pain assessment: challenges and perspectives in neonatal intensive care units". In: *Jornal de Pediatria* 99.6, pp. 546–560. ISSN: 0021-7557.

Heydari, Soroush and Qusay H Mahmoud (2025). "Tiny machine learning and on-device inference: A survey of applications, challenges, and future directions". In: *Sensors* 25.10, p. 3191.

Kendall, Alex and Yarin Gal (2017). "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems* 30.

Llerena, Amelia et al. (2023). "Neonatal pain assessment: Do we have the right tools?" In: *Frontiers in pediatrics* 10, p. 1022751.

Ministério da Saúde (2020). *Estratégia de Saúde Digital para o Brasil 2020–2028*. `https://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf`. Access in: 23 fev. 2026.

Obermeyer, Ziad et al. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464, pp. 447–453.

Ovadia, Yaniv et al. (2019). "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift". In: *Advances in neural information processing systems* 32.

Pati, Sarthak et al. (2024). "Privacy preservation for federated learning in health care". In: *Patterns* 5.7.

Pereira Carlini, Lucas et al. (2025). "Is this neonate feeling pain? Leveraging clinical knowledge towards high-precision Large Language Model-based neonatal pain assessment". In: *Pediatric Research*, pp. 1–8.

Perry, Mallory et al. (2018). "Neonatal pain: perceptions and current practice". In: *Critical care nursing clinics of North America* 30.4, p. 549.

Pirie, Craig et al. (2025). "Understanding Disagreement Between Humans and Machines in XAI: Robustness, Fidelity, and Region-Based Explanations in Automatic Neonatal Pain Assessment". In: *World Conference on XAI*. Springer, pp. 274–298.

Salekin, Md Sirajus et al. (2021). "Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment". In: *Computers in biology and medicine* 129, p. 104150.

Sanga, Mbah Carlson, Ngum Precious Fosah, and Geh Wilson Ejuh (2025). "Design and realization of an autonomous multi-powered IOT-based infant incubator monitor-

ing system hardware module". In: *International Journal of Research in Advanced Electronics Engineering* 6.1, pp. 56–65.

Schiavenato, Martin et al. (2008). "Neonatal pain facial expression: Evaluating the primal face of pain". In: *Pain* 138.2, pp. 460–471.

Shi, Weisong et al. (2016). "Edge computing: Vision and challenges". In: *IEEE internet of things journal* 3.5, pp. 637–646.

Sunwoo, John and Mohamed El-Dib (2026). "Beyond the face: advancing multimodal AI for neonatal pain assessment". In: *Pediatric Research*, pp. 1–3.

Victor, Audêncio et al. (2025). "National and regional Temporal trends and forecasting of preterm birth in brazil: evidence from National birth data (2014–2023) with projections to 2030". In: *BMC pregnancy and childbirth* 25.1, p. 1339.

World Health Organization A (2021). *Ethics and Governance of Artificial Intelligence for Health*. https://iris.who.int/server/api/core/bitstreams/4f2c477c-4b72-4ca1-9a78-a1e73af64e50/content. Access in: 23 fev. 2026.

World Health Organization B (2023). *Regulatory Considerations on Artificial Intelligence for Health*. https://iasalut.cat/wp-content/uploads/2023/10/OMS-AI-Health-Regulatory-considerations.pdf. Access in: 23 fev. 2026.