Capítulo 8

Barriers and Pathways to Clinical Translation of AI in Automated Neonatal Pain Assessment

Leonardo Antunes Ferreira, Lucas Pereira Carlini, Gabriel de Almeida Sá Coutrin, Lucas Fontes Buzuti, Roberto Gonçalves de Magalhães Júnior, Rafael Nobre Orsi, Tatiany Marcondes Heiderich, Gabriela Cianciarullo, Salvador Pinillos Gimenez, Craig Pirie, Carlos Francisco Moreno-García, Marina Carvalho de Moraes Barros, Ruth Guinsburg, Carlos Eduardo Thomaz

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.

8.1 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

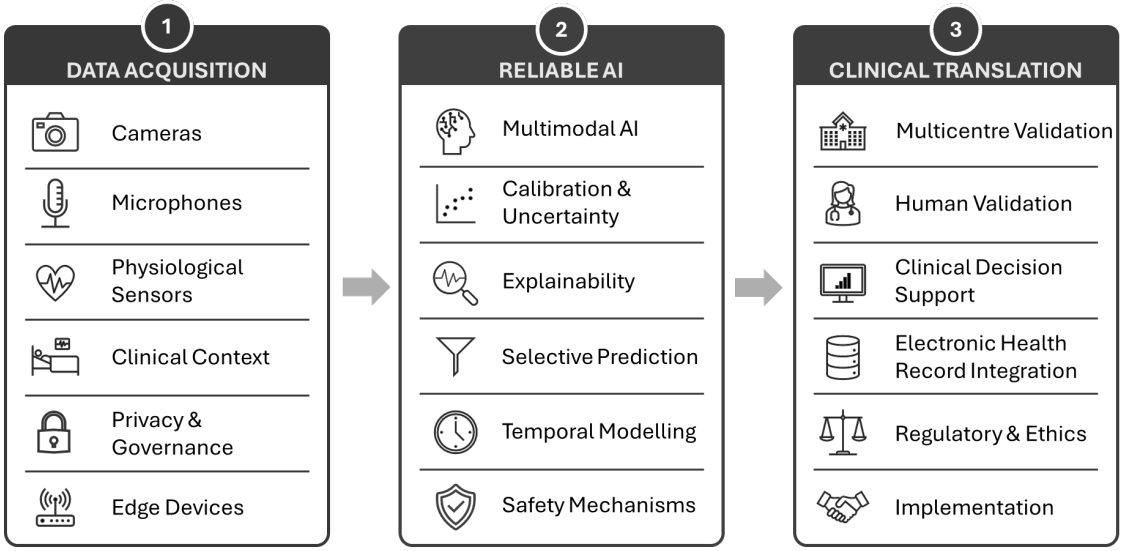

Figure 1.1. The three pillars of trustworthy AI-based neonatal assessment.

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, as illustrated in Figure 1.1, 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.

8.2 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.

8.3 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

8.4 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.

8.5 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/

8.5.1 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. 2 019]. 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].

8.5.2 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.

8.5.3 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].

8.5.4 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].

8.6 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.

