KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Hermans T, Dereymaeker A, Lemmens K, Jansen K, Usman F, Robinson S, Naulaers G, De Vos M and Hartley C (2026) Toward automated neonatal EEG analysis: multi-center validation of a reliable deep learning pipeline. Front. Neurosci. 20:1750045. doi: 10.3389/fnins.2026.1750045
Archived version	Published manuscript: the content is identical to the content of the published paper.
Published version	https://doi.org/10.3389/fnins.2026.1750045
Journal homepage	https://www.frontiersin.org/journals/neuroscience
Author contact	Tim.hermans@esat.kuleuven.be Klik hier als u tekst wilt invoeren.
Abstract	ObjectivesTo evaluate the reliability and generalization of NeoNaid, a fully automated software tool for neonatal EEG analysis, based on functional brain age (FBA) estimation and sleep staging.MethodsNeoNaid combines a multi-task deep learning model with proposed quality control routines detecting artifacts, out-of-distribution inputs, and uncertain predictions. Based on a raw EEG input, it outputs one global FBA estimate and a continuous 2-state hypnogram. We validated performance on two independent hospital settings: an internal dataset (33 EEGs, 17 infants, median 900 min/recording) and an external dataset (38 EEGs, 24 infants, median 124 min/recording).ResultsQuality control rejected a comparable number of segments in the internal and external datasets, reducing extreme errors in FBA estimation, and modestly improving sleep staging accuracy. Across the internal and external data, NeoNaid achieved median absolute FBA errors of 0.50 and 0.55 weeks and Cohen's

	Kappa values of 0.89 and 0.87 for quiet sleep detection, respectively.DiscussionNeoNaid demonstrated improved reliability through integrated quality control and maintained performance across two independent datasets. By focusing on validation and trustworthiness, this work takes an essential step toward clinical adoption of automated neonatal EEG analysis and supports its utility for both NICU practice and large-scale research.
IR	https://lirias.kuleuven.be/4409396?&lang=en

(article begins on next page)

frontiers in Neuroscience

TYPE Original Research
PUBLISHED 27 February 2026
DOI 10.3389/fnins.2026.1750045

Check for updates

OPEN ACCESS

EDITED BY
Layla Banihashemi,
University of Pittsburgh, United States

REVIEWED BY
Rabindra Khadka,
Oslo Metropolitan University, Norway
Johannes Mader,
Medical University of Vienna, Austria

*CORRESPONDENCE
Tim Hermans
✉ tim.hermans@esat.kuleuven.be

RECEIVED 19 November 2025
REVISED 29 January 2026
ACCEPTED 05 February 2026
PUBLISHED 27 February 2026

CITATION
Hermans T, Dereymaeker A, Lemmens K, Jansen K, Usman F, Robinson S, Naulaers G, De Vos M and Hartley C (2026) Toward automated neonatal EEG analysis: multi-center validation of a reliable deep learning pipeline.
Front. Neurosci. 20:1750045.
doi: 10.3389/fnins.2026.1750045

COPYRIGHT
© 2026 Hermans, Dereymaeker, Lemmens, Jansen, Usman, Robinson, Naulaers, De Vos and Hartley. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Toward automated neonatal EEG analysis: multi-center validation of a reliable deep learning pipeline

Tim Hermans[1]*, Anneleen Dereymaeker[2,3], Katrien Lemmens[2,3], Katrien Jansen[2,4], Fatima Usman[5], Shellie Robinson[5], Gunnar Naulaers[2,3], Maarten De Vos[1,2] and Caroline Hartley[5]

[1]Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium, [2]Department of Development and Regeneration, KU Leuven, Leuven, Belgium, [3]Neonatal Intensive Care Unit, University Hospitals Leuven, Leuven, Belgium, [4]Child Neurology, University Hospitals Leuven, Leuven, Belgium, [5]Department of Paediatrics, University of Oxford, Oxford, United Kingdom

Objectives: To evaluate the reliability and generalization of NeoNaid, a fully automated software tool for neonatal EEG analysis, based on functional brain age (FBA) estimation and sleep staging.

Methods: NeoNaid combines a multi-task deep learning model with proposed quality control routines detecting artifacts, out-of-distribution inputs, and uncertain predictions. Based on a raw EEG input, it outputs one global FBA estimate and a continuous 2-state hypnogram. We validated performance on two independent hospital settings: an internal dataset (33 EEGs, 17 infants, median 900 min/recording) and an external dataset (38 EEGs, 24 infants, median 124 min/recording).

Results: Quality control rejected a comparable number of segments in the internal and external datasets, reducing extreme errors in FBA estimation, and modestly improving sleep staging accuracy. Across the internal and external data, NeoNaid achieved median absolute FBA errors of 0.50 and 0.55 weeks and Cohen's Kappa values of 0.89 and 0.87 for quiet sleep detection, respectively.

Discussion: NeoNaid demonstrated improved reliability through integrated quality control and maintained performance across two independent datasets. By focusing on validation and trustworthiness, this work takes an essential step toward clinical adoption of automated neonatal EEG analysis and supports its utility for both NICU practice and large-scale research.

KEYWORDS
automated analysis, clinical validation, functional brain age, neonatal EEG, quality control, sleep staging

1 Introduction

Neonatal electroencephalography (EEG) is the gold standard for assessing brain function in newborns and has proven to be a valuable monitoring tool in the neonatal intensive care unit (NICU) (McCoy and Hahn, 2013; Karamian and Wusthoff, 2021). With its high temporal resolution and multi-channel recordings, EEG provides rich information on brain maturation, sleep stages, and pathological activity such as seizures. However, neonatal EEG is challenging to interpret due to the complexity of the signal, and recordings often span many hours. Therefore, visual analysis requires expertise and is time-consuming. These factors limit the routine clinical use of EEG, despite its potential to provide valuable insights into neonatal brain health.

Over the past decade, a range of data-driven and artificial intelligence (AI) methods have been developed to assist with neonatal EEG interpretation. These include automated approaches for seizure detection (Temko et al., 2015; Ansari et al., 2019; Borovac et al., 2022; Raeisi et al., 2022), background grading (Dereymaeker et al., 2019; Moghadam et al., 2021; Raurale et al., 2021), sleep staging (Piryatinska et al., 2009; Koolen et al., 2017; Ansari et al., 2020), and functional brain age (FBA) estimation (Pillay et al., 2020; Stevenson et al., 2020; Ansari et al., 2024). Other advancements include the deployment of a cloud-based service, Babacloud, that implements an automated analysis pipeline computing neonatal EEG summary metrics (including measures of brain state and artifact detection) from uploaded EEG recordings (Montazeri et al., 2024). Furthermore, to improve robustness and generalization in EEG deep learning, recent work has also shown the power of self-supervised and multi-task training strategies (Hermans et al., 2023; Mohammadi Foumani et al., 2024; Hojjati et al., 2025). Such strategies allow models to be trained using multiple datasets from different cohorts, which may be unlabeled, or labeled for different tasks. By not restricting training to data collected for a single specific task, this approach enables the integration of heterogeneous datasets and learning from larger and more diverse data collections, thereby improving generalizability. Collectively, these developments highlight the potential of AI to scale EEG analysis beyond the limits of manual review and to provide decision-support tools for the clinical environment.

Of the various AI approaches to neonatal EEG, sleep staging and FBA estimation are especially informative for assessing neurodevelopment, and are the focus of this study. Sleep organization is an important marker of neurological development (Dereymaeker et al., 2017; Shellhaas et al., 2017). Automated sleep staging can provide continuous, objective measurements that would otherwise be impractical for clinicians to obtain. Similarly, FBA estimation offers a quantitative measure of brain maturation by comparing EEG-derived estimates of age with the infant's postmenstrual age (PMA). Deviations between the FBA and PMA can indicate atypical development and have prognostic value (Stevenson et al., 2020; Ansari et al., 2024). Together, these applications can support both clinical decision-making and long-term research on neonatal neurodevelopment.

Despite this potential, significant barriers remain to clinical adoption. Most published models are validated only on internal test data, raising concerns about their robustness to data from different hospitals, recording systems, or electrode montages. Long NICU recordings also unavoidably contain artifacts caused by movement, poor electrode contact, or physiological interference. Models trained primarily on clean data may fail when applied to such segments. For clinical usefulness, automated EEG tools must not only be accurate but also usable in practice and reliable across diverse datasets. Another important aspect for clinical deployment of automated EEG algorithms is their robustness to differences in recording setups. In neonatal care, EEG systems can vary in electrode montages, hardware characteristics, and acquisition protocols across centers. To be trusted in clinical workflows, algorithms must maintain performance despite these variations, including when applied to data from a hospital other than the one in which they were developed. Finally, to be adopted into clinical practice, tools must be easy to use with software that can aid interpretation.

To meet this need, we developed NeoNaid, a software tool that integrates a multi-task deep learning model for neonatal EEG analysis into a user-friendly graphical interface. Its design has been refined in discussion with clinicians to ensure interpretability and usability. This tool automatically processes long EEG recordings and provides robust, clinically relevant interpretations of the EEG, including sleep staging and FBA estimates. The underlying AI model builds on our previously published work and was trained on a large in-house dataset of neonatal EEG. Beyond this, NeoNaid implements quality control routines designed to improve the reliability and trustworthiness of the tool when used in clinical practice. These routines flag EEG segments likely to yield unreliable predictions by detecting artifacts, out-of-distribution inputs, or high model uncertainty.

In this paper, we focus on validating NeoNaid as a tool for neonatal EEG analysis. Unlike prior studies that primarily introduce new model architectures, our emphasis is on quality control and external validation. Specifically, we evaluate how NeoNaid performs on two independent datasets: an internal cohort from Leuven and an external cohort from Oxford. This cross-center validation is crucial for assessing generalizability and for building trust in real-world clinical use.

The aims of this study are 2-fold: (i) to assess the contribution of quality control to improving the reliability of automated FBA and sleep analysis, and (ii) to validate NeoNaid's performance for FBA estimation and sleep staging across both internal and external datasets. Together, these analyses address the key requirements for clinical usefulness: trustworthiness and generalizability.

2 Materials and methods

2.1 Automated EEG analysis using NeoNaid

NeoNaid is an in-house developed software tool for automated neonatal EEG analysis. It integrates preprocessing, deep learning–based predictions, and quality control routines within a graphical user interface designed for clinical use (Figure 1). A key feature of NeoNaid is its ability to handle variable input montages. The underlying model uses a channel-agnostic architecture with shared weights to process each channel independently, followed by an attention-based mechanism to aggregate per-channel predictions.



FIGURE 1
Screenshot of the NeoNaid GUI. The graphs show the predicted sleep hypnogram and the brain age as a function of time during a 16 h 45 min recording. Besides these visual outputs, the GUI produces result files which were used to analyze the results for this paper.

The input to NeoNaid is a raw multi-channel EEG recording. Predictions are generated for non-overlapping 30-s segments. For each channel in the input EEG, NeoNaid produces four main outputs: (i) an artifact mask indicating which samples in the input are likely to be noise; (ii) a prediction of the sleep stage: quiet sleep (QS), active sleep (AS), or wake; (iii) an estimate of functional brain age (FBA); (iv) an attention weight for downstream aggregation. After computation of these per-channel outputs, the attention weights are used to aggregate the per-channel sleep and FBA predictions into a single global output by means of a weighted average. This mechanism allows NeoNaid to prioritize informative signals and mitigate the impact of noisy signals.

The core of the software is a multi-task deep learning model, developed by converging our previously published models (Ansari et al., 2019, 2020, 2024; Dereymaeker et al., 2019; Hermans et al., 2023) into one common methodology. Its architecture builds on our previously published convolutional neural network (Hermans et al., 2023) and consists of a shared encoder connected to multiple output heads, one for each task. Unlike prior single-task models, NeoNaid was trained in a multi-task setting, enabling simultaneous supervised learning from different neonatal EEG datasets labeled for different tasks. The training data included approximately 1,326 h of EEG from 124 recordings with age labels for FBA estimation, 565 h from 132 recordings with sleep annotations, and 44 h from 73 recordings with artifact annotations.

2.2 Dataset

We evaluated NeoNaid using two independent neonatal EEG datasets: one internal dataset collected at University Hospitals Leuven, Belgium (Dataset A) and one external dataset recorded at the John Radcliffe Hospital in Oxford, UK (Dataset B). These datasets differ in acquisition hardware, electrode configuration,

and local recording protocols, allowing us to assess generalization of the algorithm to different recording conditions. The internal Dataset A originated from the same hospital as the data used for the development of the NeoNaid algorithms. However, all recordings Dataset A were independent of the development data and thus represent previously unseen cases.

2.2.1 Dataset A (internal)

Dataset A consists of 33 EEG recordings from 17 neonates monitored at the NICU of University Hospitals Leuven, Belgium. The cohort includes both term and preterm infants with postmenstrual ages (PMA) at the time of recording ranging from 27.3 to 47 weeks (median 37.7). EEGs were acquired using the BrainRT EEG system (Onafhankelijke Software Groep (OSG), Kontich, Belgium) with a sampling rate of 250 or 256 Hz. Electrodes were positioned according to a modified 10–20 system, including the following channels: Fp1, Fp2, C3, C4, T3, T4, O1, and O2, with Cz as the reference electrode.

The EEG recordings had a median duration of 900 (IQR: 90–980) minutes and were collected as part of routine clinical care. All recordings came from neonates without major developmental abnormalities, allowing PMA at recording to serve as a reference measure for FBA. Two-class sleep annotations (QS versus AS or wake) were available for a subset of 28 recordings, scored by trained experts. All data were anonymized prior to analysis. The study was approved by the local ethics committee of University Hospitals Leuven, in accordance with the Declaration of Helsinki.

2.2.2 Dataset B (external)

Dataset B comprises 38 EEG recordings from 24 neonates recorded at the Newborn Care Unit of the John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. The PMA at the time of recording ranged from 29.4 to 41.4 (median 34.5) weeks. Compared to Dataset A, Dataset B used a different EEG system and montage. EEGs were recorded using the SynAmps RT 64-channel headbox and amplifiers and CURRYscan7 neuroimaging suite (Compumedics Neuroscan) with a sampling rate of 2,000 Hz. The electrode configuration consisted of FCz, C3, C4, Cz, CPz, T3, T4, and Oz, referenced to Fz.

Recordings had a median duration of 124 (IQR: 91–143) minutes. Similar to dataset A, all recordings were from patients without any neurological abnormalities (infants were excluded from studies if they had a grade III or IV intraventricular hemorrhage, hypoxic ischemic encephalopathy, or major congenital malformations), making the PMA at time of recording a suitable reference for functional brain age. Of the 38 EEGs, 18 were randomly selected for sleep staging, ensuring good coverage across PMA and recording durations, and were subsequently labeled for sleep by a trained expert in Leuven.

The EEG recordings were collected for research purposes, as part of an independent study (Usman et al., 2025). Eligible families were given verbal and written information about the study, and written parental consent was signed before inclusion in the study. This dataset was fully anonymized and provided under a data sharing agreement between the University of Oxford and KU Leuven. Ethical approval was obtained through the relevant UK regulatory bodies (National Research Ethics Service reference: 12/SC/0447).

2.2.3 Pre-processing

To ensure consistency across datasets and to reduce the influence of differing reference electrodes, EEG recordings were transformed into a common bipolar montage using electrode pairs present in both datasets (C3–C4, C3–Cz, C3–T4, C4–Cz, C4–T3). This was achieved by re-referencing channel signals accordingly using electrode pairs present in both datasets. No other channel interpolation was applied. Following this, signals were bandpass-filtered (0.25–30 Hz) and downsampled to 64 Hz. The recordings were then segmented into 30-s non-overlapping epochs. Within each recording, channel amplitudes were normalized by the median standard deviation across all segments. These pre-processing steps ensured consistency across acquisition systems and were automatically executed by the NeoNaid software.

2.3 Quality control

Typically, the majority of training data is from a clean and labeled dataset, thus making existing trained models unreliable for artifact-containing data segments. Additionally, deep learning models perform well on data that is similar to the training data, but their performance becomes unreliable when applied to data that has significantly different characteristics compared to the training data (out-of-distribution data).

A central feature of NeoNaid is its quality control algorithm, which evaluates the reliability of each 30-s EEG segment before downstream interpretation. This process involves evaluating three independent reliability criteria: artifact content, novelty detection (to detect out-of-distribution inputs), and (un)certainty level.

2.3.1 Artifacts

Each segment is assigned an artifact score based on the percentage of samples identified as noise by the dedicated detection head of the model. Segments with more than 50% artifact content are flagged as unreliable, therefore preventing the model from producing predictions on segments where brain activity is largely obscured. This threshold was informed by an analysis of the relationship between performance and artifact content during NeoNaid's development using independent datasets. As most segments exhibited either very low or very high artifact content, the precise threshold was not critical and the 50% cutoff was adopted as a pragmatic criterion. Segments that do not pass the artifact check are flagged as unreliable and either excluded (for FBA) or explicitly marked (for sleep staging).

2.3.2 Novelties

Out-of-distribution inputs (novelties) are automatically identified by applying a novelty detection model to each channel in a segment. The novelty detection model in NeoNaid is an isolation forest that uses a set of nine spectral features as input (Piryatinska et al., 2009), predicting for each channel in every segment whether it is an inlier or a novelty (with respect to the NeoNaid training data). A separate novelty detection model was fitted per task, using the data that was used for training that part of the multi-task model. Channels that was labeled as a novelty by NeoNaid were

excluded from the model's channel aggregation, reducing their impact on the global predictions. Furthermore, if more than half of the channels in a segment were labeled as novelties, the entire segment was flagged.

2.3.3 Uncertainty

One of the outputs of NeoNaid are channel-wise attention weights. After normalization, these weights are used to compute a weighted average of channel-wise predictions. Prior to normalization, these channel-specific attention weights reflect the model's certainty regarding each channel's input. To identify segments where all channels are deemed uncertain, the maximum unnormalized attention weight across channels was computed for each segment. Segments where this maximum attention weight falls below a predefined threshold are flagged as unreliable. This threshold was defined and fixed during NeoNaid's development as the 1st percentile of maximum attention weights observed in an independent calibration dataset. In addition to attention-based flagging, for the sleep staging outputs, segments with QS probabilities near 0.5 were also marked as uncertain.

Together, these three criteria provide a conservative safeguard against unreliable predictions. NeoNaid then aggregates the segment-wise outputs and quality flags into clinically interpretable results. For FBA, a single robust estimate is obtained as the median across reliable segments, while for sleep staging, a continuous hypnogram is constructed by smoothing probabilities and interpolating over short, unreliable intervals using simple heuristic rules.

2.4 Performance metrics

2.4.1 Functional brain age

For each recording, the global FBA estimate was defined as the median of all segment-wise predictions that passed quality control. The performance was quantified in terms of the absolute error, which was defined as the absolute difference between the global FBA estimate and the infant's PMA at the time of recording. Lower errors indicate better performance. The interquartile range (IQR) of the retained segment-wise estimates was reported as a measure of prediction confidence, with wider IQRs indicating a lower prediction certainty.

2.4.2 Sleep staging

For sleep analysis, we evaluated NeoNaid's ability to detect QS. To this end, the model's AS and wake predictions were combined into a single category, representing the non-quiet sleep class. Predictions on segments flagged as unreliable were excluded. The performance was measured in terms of Cohen's kappa score, where a higher score indicates better agreement between predicted sleep stages and expert annotations.

2.5 Analysis

Our analyses were designed to evaluate both the impact of quality control and the generalizability of NeoNaid across datasets. We applied the full processing pipeline to both datasets: Dataset A (internal) and Dataset B (external). We compared two approaches: a naïve approach that included all segment-wise predictions, and a robust approach that excluded segments flagged by quality control.

2.5.1 Effect of quality control

To check for differences in data quality and characteristics between the two datasets, we first quantified the occurrence of quality control flags within each dataset. More specifically, we computed the proportion of EEG segments flagged for artifacts and novelty detection; these were independently assessed for the FBA and sleep tasks.

For FBA estimation, performance was evaluated as a function of EEG recording duration, as the effect of the robust method is most clear for EEGs of shorter durations. To simulate varying EEG durations, we extracted sub-epochs ranging from 30 s to 1 h from each recording. For each duration, 1,000 sub-epochs were randomly selected per recording. The median FBA and corresponding performance metrics were calculated for each duration, allowing assessment of how recording length and the inclusion of quality control routines affect prediction errors.

For sleep staging, we computed Cohen's kappa score for quiet sleep detection on the full recordings using both the naive and robust approaches and compared the performance.

2.5.2 Cross-center validation

Finally, we validated the robust methodology (i.e., including quality control) on the complete EEG recordings. Per-recording results were analyzed and visualized in two ways. Firstly, FBA and sleep performance metrics are reported for each channel separately, i.e., using the per-channel predictions prior to channel aggregation, as well as for the global result obtained after aggregation of the per-channel predictions. Secondly, we showed the performance as a function of PMA to investigate whether prediction accuracy is systematically affected by the age of the neonates.

3 Results

3.1 NeoNaid quality control reduces errors

We investigated how the quality control routine affects the automated analysis. For FBA, the median rejection rate was 21.5% in the internal dataset and 16.1% in the external dataset (Figure 2). These were lower for sleep staging, with medians of 2.0% (internal) and 2.4% (external), mainly due to the heuristic postprocessing, which interpolates short unreliable intervals, and retains segments with high-amplitude movement artifacts when they occur during predicted wake cycle.

Analyzing the rejection rates according to the three quality control criteria (which are not mutually exclusive), segments were most frequently flagged as artifacts. In the FBA data, 22.9%

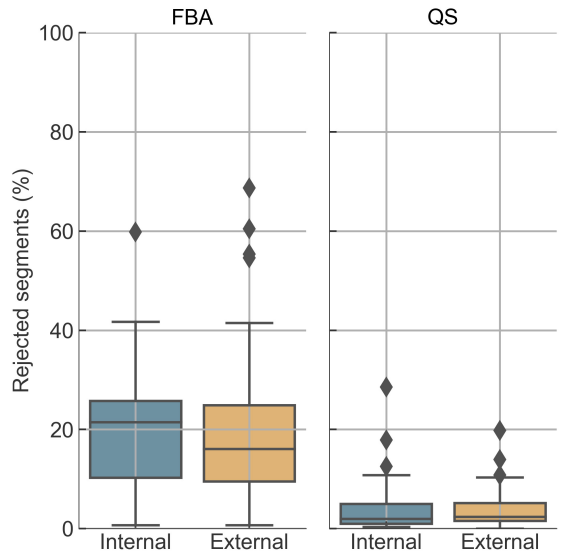

FIGURE 2
Rejection rates due to quality control in the two datasets. Rejected %: the total percentage of segments that were rejected in each recording.

of segments in the internal dataset and 19.6% in the external dataset were marked as artifactual; similar rates were observed for sleep data (22.3% and 19.3%, respectively). Novelty detection contributed less, flagging 1.1% and 2.4% of FBA segments, and 4.1% and 3.4% of sleep segments in the internal and external data, respectively. The contribution of segments flagged as uncertain for FBA was 3.3% (internal) and 7.2% (external), and for sleep 4.6% (internal) and 2.2% (external). For sleep staging, these values are observed prior to heuristic postprocessing, thus accounting for the generally lower rejection percentages than the proportion of initially flagged segments. Overall, data quality was comparable between centers, and the external data did not appear out-of-distribution despite differences in acquisition systems and protocols.

Next, we investigated the effect of the quality control on FBA performance in both datasets. Overall, the robust method (which applies segment rejection) and the naive method (which does not) yielded similar median FBA error values (Figure 3). This is expected, as the use of the median as an aggregation metric reduces the influence of outlier segments. Nevertheless, the robust method consistently showed a lower likelihood of producing extreme outliers, particularly in shorter recordings, demonstrating its value in minimizing risk. The median IQR of the FBA estimate initially increased with data length and then plateaued after approximately 20 min of usable data. This suggests that IQR values in recordings with less than 20 min of non-rejected EEG should be interpreted with care, as limited data availability may underestimate the IQR of the actual underlying distribution.

Quiet sleep detection performance using the naive and robust methods (Figure 4) showed modest performance improvement following quality control on the already relatively cleaned and labeled sleep data. Unusable segments were not annotated by the experts and, therefore, were not included in the performance evaluation. While the benefit of quality control is less pronounced in this evaluation setup, its primary value lies in preventing unreliable predictions when poor-quality data are encountered in practice.

3.2 Cross-center validation of NeoNaid

We investigated the results of the robust method (with quality control) of both datasets to validate the cross-center generalization of NeoNaid. Using the robust methodology on full recordings, the mean absolute FBA error across the datasets is 0.60 (median: 0.50, IQR: 0.21–0.78) weeks for Dataset A, and 0.69 (median: 0.55, IQR: 0.23–1.02) weeks for Dataset B. The percentage of recordings within 1 week error is 79% (100% within 2 weeks) and 74% (97% within 2 weeks) for Dataset A and Dataset B, respectively. Moreover, the true PMA for Dataset A fell within the IQR of the per-segment predictions in 70% of recordings, compared to 58% for Dataset B.

For QS detection, performance was high in both datasets. In Dataset A, per-recording kappa scores averaged 0.86 (median: 0.89, IQR: 0.80–0.96), while in Dataset B they averaged 0.79 (median: 0.87, IQR: 0.70–0.94). Comparable results were obtained when restricting the analysis to single channels, particularly for bipolar derivations around C3, Cz, and C4 (Figure 5). When pooling all recordings, the overall kappa was 0.874 (95% CI: 0.865–0.884) in Dataset A and 0.831 (95% CI: 0.811–0.851) in Dataset B.

The performance of FBA and quiet sleep detection was relatively consistent across channels (Figure 5), suggesting that the software can deliver reliable results even with limited or single-channel input. However, while combining all channels did not always outperform the best individual channel in terms of median performance, it generally helped reduce errors in outlier recordings, offering improved robustness.

Finally, we assessed the model performance compared with the age of the infants. For the FBA model, there is no significant linear correlation between the FBA error and PMA (Figure 6), indicating that the model performed equally at all ages. In contrast, for sleep staging, there was a trend that QS performance improves with increasing age (Figure 6). In one case in Dataset B, the model failed to detect any quiet sleep, whereas the expert annotation indicated a 15-min QS bout, resulting in a Kappa score of zero. The 15-min QS epoch included multiple high-amplitude artifacts, which resulted in the model misclassifying it as wake.

4 Discussion

The aim of this study was 2-fold: first, to assess the contribution of quality control routines in improving the reliability of automated neonatal EEG analysis, and second, to validate the performance of NeoNaid across independent datasets from two hospital settings.

We found that quality control reduced extreme errors and provided transparent confidence measures, particularly for functional brain age estimation in shorter or noisier recordings. NeoNaid's integrated quality control framework assesses the reliability of each EEG segment through artifact detection, novelty detection, and attention-based certainty scoring. Our results show that this quality control framework improves the

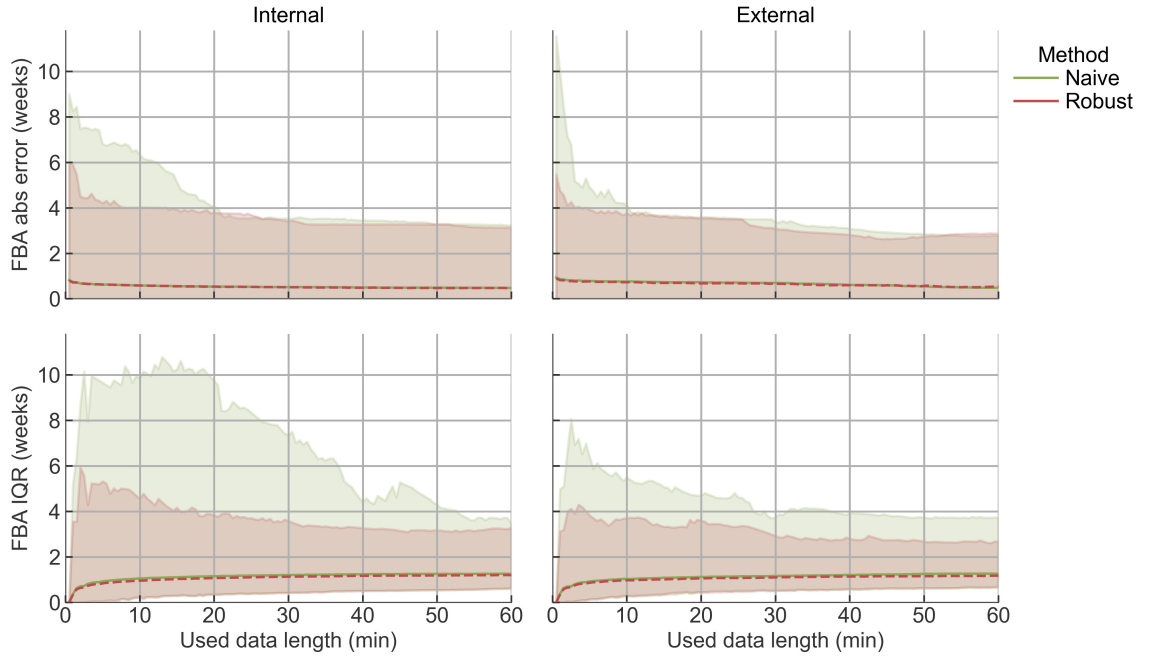

FIGURE 3
FBA performance for different simulated data lengths (1,000 repetitions per window, per recording). The x-axis represents the duration of data used to compute the estimates (excluding the rejected segments in the robust case), of which median + min–max is shown. Robust method includes quality control and subsequent segment rejection, naïve method does not. Left: internal Dataset A, right: external Dataset B.

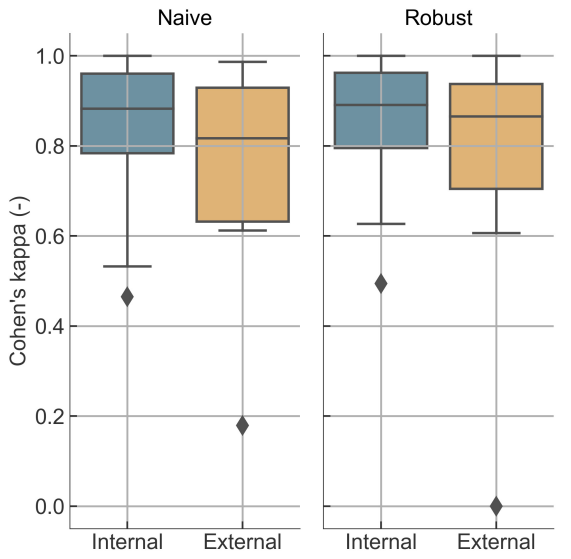

FIGURE 4
Sleep performance with (robust) and without (naive) artifact rejection and heuristic postprocessing.

reliability of automated neonatal EEG analysis by reducing the likelihood of extreme errors while preserving valid information. While average performance metrics changed only modestly, the safeguards proved valuable in minimizing risk, particularly in shorter or lower-quality recordings (Figures 3, 4). Therefore, these routines increase the clinical trust in the outputs generated by NeoNaid.

NeoNaid maintained performance when validated on an independent, external dataset, despite differences in acquisition hardware, montages, and protocols. Additionally, the external data were not identified as out-of-distribution by the quality control routines. This cross-center validation provides preliminary evidence that the approach can generalize across clinical environments, addressing an important barrier to the adoption of automated neonatal EEG analysis.

A key feature of NeoNaid is that it is not restricted to a specific montage or number of channels. As shown in Figure 5, the model produces comparable results when operating on a single bipolar derivation compared to the full five-channel setup. This is particularly relevant for clinical scenarios with limited channel availability, such as amplitude-integrated EEG (aEEG) monitoring, where only one or two channels may be accessible. This flexibility further enhances NeoNaid's clinical applicability across different monitoring environments.

Performance did not vary significantly with PMA for FBA, although there was a tendency toward slightly larger errors in recordings below 30 or above 40 weeks PMA. This pattern likely reflects the limited amount of training and validation data available in these extreme age ranges rather than a systematic bias. In contrast, sleep staging performance clearly improved with age. As shown in Figure 6, older neonates achieved higher kappa scores for quiet sleep detection, whereas younger infants showed slightly lower performance. This is likely due to the less distinct differentiation between active and quiet sleep at earlier developmental stages.

Several prior studies have developed deep learning models for brain age estimation (Pillay et al., 2020; Stevenson et al., 2020; Ansari et al., 2024) and sleep staging. While direct comparison of performance metrics across studies should be done cautiously

10.3389/fnins.2026.1750045

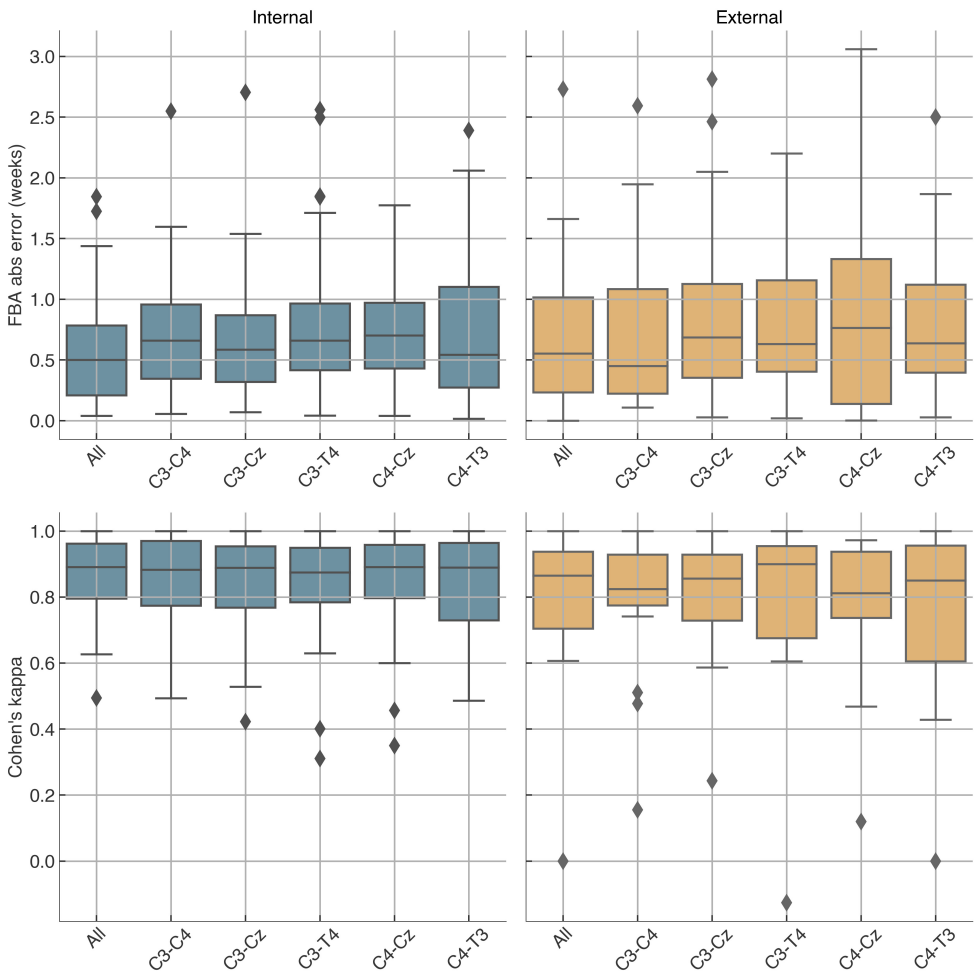

FIGURE 5
Per-recording performance per channel. "All" refers to attention-weighted average of single-channel predictions. Top: brain age error; bottom: quiet sleep detection score. Left: internal Dataset A, right: external Dataset B.

due to differences in test datasets, our findings show that the FBA performance of NeoNaid (mean absolute errors of 0.60 and 0.69 on internal and external datasets, respectively) is comparable, if not superior, to these earlier models, which typically report mean absolute errors between 0.7 and 1.0 weeks. Similarly, NeoNaid's performance in QS detection aligns with prior models that have reported kappa values up to 0.77.

NeoNaid offers value for both clinical and research applications. In clinical practice, automatic sleep staging and FBA estimation can aid in monitoring brain development, particularly in preterm infants. The IQR accompanying the FBA estimate provides a practical measure of confidence, helping clinicians interpret results more effectively. The built-in quality control indicators can alert users to unreliable segments, reducing the risk of misinterpretation due to artifacts or signal degradation. For researchers, NeoNaid provides a scalable solution for annotating large EEG datasets in a standardized way. It is especially useful in studies investigating neurodevelopmental trajectories, sleep–wake organization, and responses to therapeutic interventions.

This study has several limitations. The number of recordings in the external dataset was relatively small, limiting the statistical power of generalization claims. Sleep annotations were made by different raters at each site, without a formal assessment of inter-rater reliability, which could introduce bias. Moreover, the external recordings were shorter in duration and lacked accompanying physiological or video data, which made them more challenging to annotate. Finally, external validation was limited to a single center. Future work will focus on expanding collaborations to include additional external datasets and broader populations to further validate and refine the NeoNaid platform.

To conclude, we have shown that NeoNaid is a robust tool for automated neonatal EEG analysis, maintaining performance across two datasets with different recording setups. Its integrated quality control routines reduce extreme errors and improve trustworthiness, addressing a critical requirement for clinical adoption. This represents an important step toward the broader use of AI tools in neonatal EEG, where differences in acquisition setups are common and difficult to standardize. Ultimately, these result support the potential of NeoNaid for both NICU practice and large-scale research on neonatal brain monitoring.

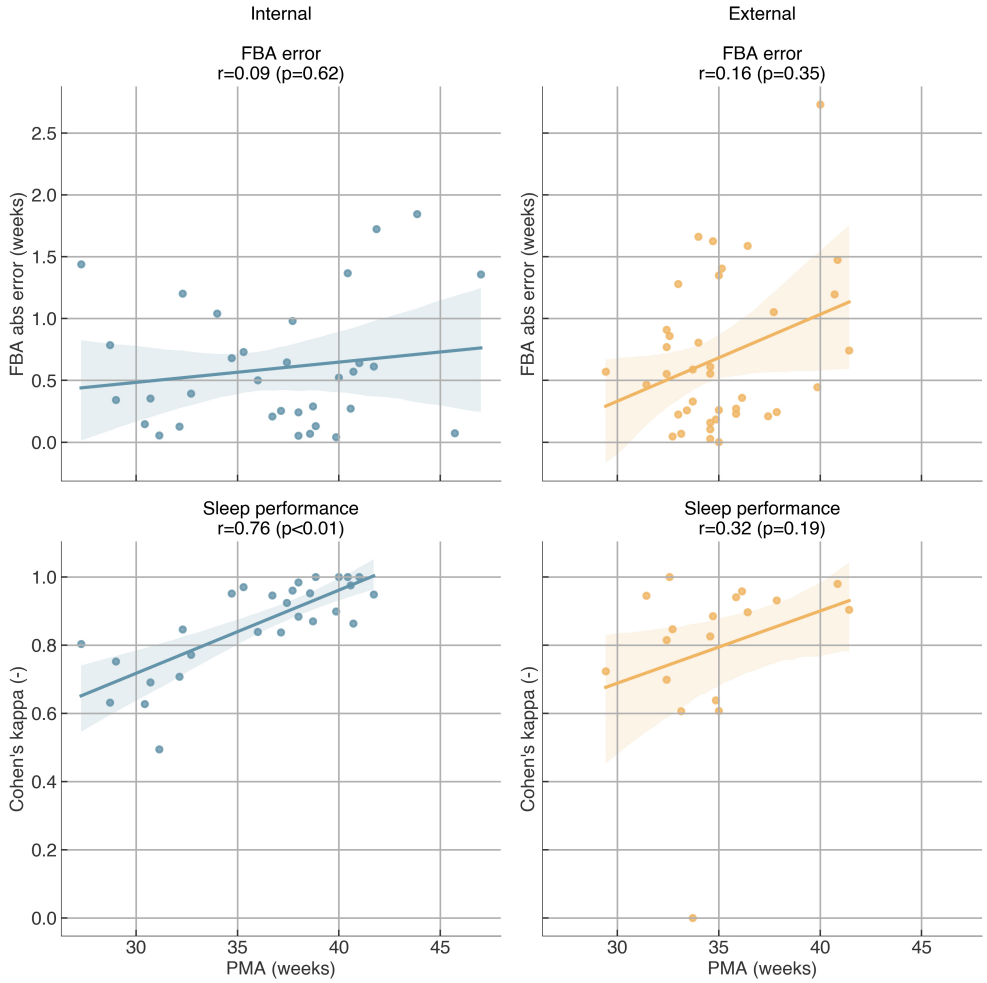

FIGURE 6
The relation between the performance and postmenstrual age (PMA). Left: internal Dataset A, right: external Dataset B.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the datasets presented in this article are not readily available because they contain sensitive medical data and cannot be publicly shared for privacy and ethical reasons. Requests to access the Leuven data should be directed to TH, tim.hermans@esat.kuleuven.be, and requests to access the Oxford data should be directed to CH, caroline.hartley@paediatrics.ox.ac.uk.

Ethics statement

The studies involving humans were approved by Medical Ethics Committee of University Hospitals Leuven (UZ Leuven), Belgium, and NHS Health Research Authority — National Research Ethics Service (NRES), United Kingdom. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

TH: Conceptualization, Methodology, Software, Validation, Writing – original draft. AD: Conceptualization, Investigation, Methodology, Writing – review & editing. KL: Writing – review & editing, Investigation. KJ: Conceptualization, Investigation, Writing – review & editing. FU: Investigation, Writing – review & editing. SR: Investigation, Writing – review & editing. GN: Conceptualization, Investigation, Writing – review & editing. MDV: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing. CH: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research received funding from the European Union as part of the Cost action "Maximizing impact of multidisciplinary research in early diagnosis of neonatal

brain injury" (AI-4-NICU) CA20124, Industrieel Onderzoeksfonds KU Leuven (IOF): C3 "NEONAID - Neonatal Online Neuro-Analysis for Interpretation and Decision-support - Bringing automated neonatal monitoring to the bedside," C3/24/023, the Horizon program: HORIZON-HLTH-2022-IND-13: "Privacy compliant health data as a service for AI development (PHASE IV AI)," funded by the European Union, under Grant Agreement #101095384, the Flemish Government (AI Research Program), MDV and TH are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium, and CH and SR were funded by the Wellcome Trust/Royal Society through a Sir Henry Dale Fellowship awarded to CH (grant reference: 213486/Z/18/Z). FU was funded by the Commonwealth Scholarship Commission.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author GN declared that they were an editorial board member of Frontiers at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. During the preparation of this work, the authors used ChatGPT in order to improve readability and language in portions of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ansari, A., Cherian, P. J., Caicedo, A., Naulaers, G., De Vos, M., and Van Huffel, S. (2019). Neonatal seizure detection using deep convolutional neural networks. *Int. J. Neural Syst.* 29:1850011. doi: 10.1142/s0129065718500119

Ansari, A., De Wel, O., Pillay, K., Dereymaeker, A., Jansen, K., Van Huffel, S., et al. (2020). A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants. *J. Neural Eng.* 17:16028. doi: 10.1088/1741-2552/ab5469

Ansari, A., Pillay, K., Arasteh, E., Dereymaeker, A., Mellado, G. S., Jansen, K., et al. (2024). Resting state electroencephalographic brain activity in neonates can predict age and is indicative of neurodevelopmental outcome. *Clin. Neurophysiol.* 163, 226–235. doi: 10.1016/j.clinph.2024.05.002

Borovac, A., Gudmundsson, S., Thorvardsson, G., Moghadam, S. M., Nevalainen, P., Stevenson, N., et al. (2022). Ensemble Learning Using Individual Neonatal Data for Seizure Detection. *IEEE J. Transl. Eng. Health Med.* 10:4901111. doi: 10.1109/jtehm.2022.3201167

Dereymaeker, A., Matic, V., Vervisch, J., Cherian, P. J., Ansari, A., De Wel, O., et al. (2019). Automated EEG background analysis to identify neonates with hypoxic-ischemic encephalopathy treated with hypothermia at risk for adverse outcome: A pilot study. *Pediatr. Neonatol.* 60, 50–58. doi: 10.1016/j.pedneo.2018.03.010

Dereymaeker, A., Pillay, K., Vervisch, J., De Vos, M., Van Huffel, S., Jansen, K., et al. (2017). Review of sleep-EEG in preterm and term neonates. *Early Hum. Dev.* 113, 87–103. doi: 10.1016/j.earlhumdev.2017.07.003

Hermans, T., Smets, L., Lemmens, K., Dereymaeker, A., Jansen, K., Naulaers, G., et al. (2023). A multi-task and multi-channel convolutional neural network for semi-supervised neonatal artefact detection. *J. Neural Eng.* 20:26013. doi: 10.1088/1741-2552/acbc4b

Hojjati, A., Li, L., Hameed, I., Yazidi, A., Lind, P. G., and Khadka, R. (2025). From video to EEG: Adapting joint embedding predictive architecture to uncover visual concepts in brain signal analysis. *arXiv [Preprint]* doi: 10.48550/arXiv.2507.03633

Karamian, A. G. S., and Wusthoff, C. J. (2021). Current and future uses of continuous EEG in the NICU. *Front. Pediatr.* 9:768670. doi: 10.3389/fped.2021.768670

Koolen, N., Oberdorfer, L., Rona, Z., Giordano, V., Werther, T., Klebermass-Schrehof, K., et al. (2017). Automated classification of neonatal sleep states using EEG. *Clin. Neurophysiol.* 128, 1100–1108. doi: 10.1016/j.clinph.2017.02.025

McCoy, B., and Hahn, C. D. (2013). Continuous EEG monitoring in the neonatal intensive care unit. *J. Clin. Neurophysiol.* 30, 106–114. doi: 10.1097/WNP.0b013e3182872919

Moghadam, S. M., Pinchefsky, E., Tse, I., Marchi, V., Kohonen, J., Kauppila, M., et al. (2021). Building an open source classifier for the neonatal EEG background: A systematic feature-based approach from expert scoring to clinical visualization. *Front. Hum. Neurosci.* 15:675154. doi: 10.3389/fnhum.2021.675154

Mohammadi Foumani, N., Mackellar, G., Ghane, S., Irtza, S., Nguyen, N., and Salehi, M. (2024). "EEG2Rep: Enhancing self-supervised EEG representation through informative masked inputs," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY: Association for Computing Machinery), 5544–5555. doi: 10.1145/3637528.3671600

Montazeri, S., Nevalainen, P., Metsäranta, M., Stevenson, N. J., and Vanhatalo, S. (2024). Clinical outcome prediction with an automated EEG trend, Brain State of the Newborn, after perinatal asphyxia. *Clin. Neurophysiol.* 162, 68–76. doi: 10.1016/J.CLINPH.2024.03.007

Pillay, K., Dereymaeker, A., Jansen, K., Naulaers, G., and De Vos, M. (2020). Applying a data-driven approach to quantify EEG maturational deviations in preterms with normal and abnormal neurodevelopmental outcomes. *Sci. Rep.* 10:7288. doi: 10.1038/s41598-020-64211-0

Piryatinska, A., Terdik, G., Woyczynski, W. A., Loparo, K. A., Scher, M. S., and Zlotnik, A. (2009). Automated detection of neonate EEG sleep stages. *Comput. Methods Programs Biomed.* 95, 31–46. doi: 10.1016/j.cmpb.2009.01.006

Raeisi, K., Khazaei, M., Croce, P., Tamburro, G., Comani, S., and Zappasodi, F. (2022). A graph convolutional neural network for the automated detection of seizures in the neonatal EEG. *Comput. Methods Programs Biomed.* 222:106950. doi: 10.1016/j.cmpb.2022.106950

Raurale, S. A., Boylan, G. B., Mathieson, S. R., Marnane, W. P., Lightbody, G., and O'Toole, J. M. (2021). Grading hypoxic-ischemic encephalopathy in neonatal EEG with convolutional neural networks and quadratic time–frequency distributions. *J. Neural Eng.* 18:046007. doi: 10.1088/1741-2552/abe8ae

Shellhaas, R. A., Burns, J. W., Hassan, F., Carlson, M. D., Barks, J. D. E., and Chervin, R. D. (2017). Neonatal sleep-wake analyses predict 18-month neurodevelopmental outcomes. *Sleep* 40:zsx144. doi: 10.1093/sleep/zsx144

Stevenson, N. J., Oberdorfer, L., Tataranno, M.-L., Breakspear, M., Colditz, P. B., Vries, L. S., et al. (2020). Automated cot-side tracking of functional brain age in preterm infants. *Ann. Clin. Transl. Neurol.* 7, 891–902. doi: 10.1002/acn3.51043

Temko, A., Marnane, W., Boylan, G., and Lightbody, G. (2015). Clinical implementation of a neonatal seizure detection algorithm. *Decis. Support Syst.* 70, 86–96. doi: 10.1016/j.dss.2014.12.006

Usman, F., Zandvoort, C. S., Robinson, S., Peck, M., Cobo, M. M., Adjei, T., et al. (2025). Brain activity as a biomarker for personalised caffeine treatment in premature neonates. *medRxiv [Preprint]* doi: 10.1101/2025.05.02.25326856