Interdisciplinary Research in Medical Sciences Specialty
ISSN: 3062-4401
2026, Volume 6, Issue 1, Page No: 284-298

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Deep Learning–Based Video Analysis for Identifying Neurologic Changes in Critically Ill Infants: A Retrospective Single-Center Cohort Study

Elena Petrova[1]*, Ivan Georgiev[1], Nikolay Stoyanov[2], Petar Kolev[3]

[1]Department of Pediatric Neurocritical Care and Video Analysis, Faculty of Medicine, Sofia University, Sofia, Bulgaria.
[2]Department of Deep Learning and Neurologic Monitoring, Faculty of Medicine, University of Plovdiv, Plovdiv, Bulgaria.
[3]Department of Critical Care Neurology, Faculty of Medicine, University of Ruse, Ruse, Bulgaria.

***E-mail** ✉ elena.petrova@uni-sofia.bg

Received: 18 April 2026; **Revised:** 16 June 2026; **Accepted:** 19 June 2026

ABSTRACT

Changes in infant alertness and neurological status may indicate serious underlying conditions, yet these are typically evaluated through physical examinations that occur intermittently and involve subjective judgment. There is a clear need for dependable, ongoing monitoring techniques. We proposed that a computer vision technique using pose estimation artificial intelligence (AI) to monitor movement could help forecast neurological alterations in the neonatal intensive care unit (NICU). Video recordings synchronized with electroencephalography (video-EEG) were gathered from infants under 1 year corrected age at Mount Sinai Hospital, a level IV urban NICU in New York City, spanning February 1, 2021, to December 31, 2022. A deep learning pose estimation model was trained on these video recordings, with 14 key anatomical landmarks annotated across 25 frames per infant. Classifiers were subsequently developed using these landmark positions to forecast cerebral dysfunction (identified via EEG interpretation by a neurologist specializing in epilepsy) and sedation status (determined by the delivery of sedative drugs). The study assembled the most extensive video-EEG dataset currently available, comprising 282,301 video minutes from 115 infants drawn from a varied demographic. Infant pose estimation proved accurate in cross-validation, unseen frames, and unseen infants, yielding receiver operating characteristic area under the curve (ROC-AUC) values of 0.94, 0.83, and 0.89, respectively. Median movement levels rose with advancing age; after adjusting for age, movement decreased in association with sedative administration and among infants exhibiting cerebral dysfunction (all $P < 5 \times 10^{-3}$, based on 10,000 permutations). The model demonstrated strong performance in predicting sedation across cross-validation, held-out time segments, and held-out infants (ROC-AUCs 0.90, 0.91, 0.87). Similarly, it effectively predicted cerebral dysfunction (ROC-AUCs 0.91, 0.90, 0.76). These results establish that pose AI can be effectively deployed in an intensive care environment and that an EEG-based diagnosis of cerebral dysfunction can be inferred solely from video recordings. Deep learning combined with pose AI represents a promising, scalable, and minimally intrusive tool for neurological remote monitoring in the NICU.

Keywords: Neonatology, Sedation, Encephalopathy, Artificial intelligence, Deep learning

How to Cite This Article: Petrova E, Georgiev I, Stoyanov N, Kolev P. Deep Learning–Based Video Analysis for Identifying Neurologic Changes in Critically Ill Infants: A Retrospective Single-Center Cohort Study. Interdiscip Res Med Sci Spec. 2026;6(1):284-98. https://doi.org/10.51847/9jax59rqy6

Introduction

Infant alertness represents the most sensitive component of the neurological assessment, mirroring the functional status of the entire central nervous system [1, 2]. Alterations in mental status among infants may stem from encephalopathy, sedative effects, or other factors and tend to fluctuate rapidly, requiring ongoing evaluation. Encephalopathy arises from multiple causes demanding prompt recognition and intervention (such as hypoxic, metabolic, or infectious origins) [3]. Lethargy, often a marker of encephalopathy, serves as a key clinical clue for neonatal sepsis [4]. While encephalopathy exemplifies pathological effects on alertness, sedative medications are

frequently used to intentionally modify mental status. Adjusting sedation levels proves difficult across patient groups but is especially complex in infants owing to their inability to verbalize and marked variability in drug responses. In this population, sedatives exhibit prolonged half-lives, rapid tolerance development, active metabolites that antagonize opioids, and greater risk of respiratory suppression [5, 6]. Consequently, frequent mental status evaluation is essential for critical decision-making in the NICU.

Despite its central role, continuous and objective quantification of infant alertness remains constrained. Physical examinations offer only momentary insights, are prone to observer bias, and may be postponed [1, 2]. Standardized tools such as the N-PASS for sedation [7] or the modified Sarnat score for encephalopathy [8, 9] help reduce subjectivity and achieve good inter-rater agreement [10], yet they demand significant staff time and lack continuity. EEG provides a continuous electrophysiological record but necessitates specialized personnel and apparatus unavailable in all NICUs [11]. Extended EEG use also risks skin breakdown, prompting protocols aimed at minimizing monitoring duration [12, 13]. Moreover, EEG's relevance to alertness assessment is restricted; although some data suggest it may correlate with sedation depth [14], routine standalone EEG is not employed for this purpose in practice [15]. While continuous EEG can gauge encephalopathy severity through cerebral dysfunction patterns, its routine clinical use is mostly confined to therapeutic hypothermia candidacy [16], partly due to substantial inter-rater differences and interpretive complexity for frontline clinicians [17, 18]. Emerging EEG deep learning applications for seizure detection [19], enhanced interpretability [16, 20], and outcome prediction [21] hold promise but have not yet been incorporated into standard care or alertness tracking. Thus, reliable assessment of infant alertness continues to pose a significant challenge in NICU practice.

A practical solution to this gap involves enhancing the traditional exam through computer vision. We hypothesized that pose AI [22, 23]—a deep learning method for real-time anatomical landmark tracking—could enable continuous characterization of neurological features in the NICU. In the present investigation, pose AI was successfully implemented on an extensive dataset (282,301 video-EEG minutes captured at 30 frames per second from 115 infants) and accurately forecasted both sedation and cerebral dysfunction in these critically ill patients **(Figure 1)**.

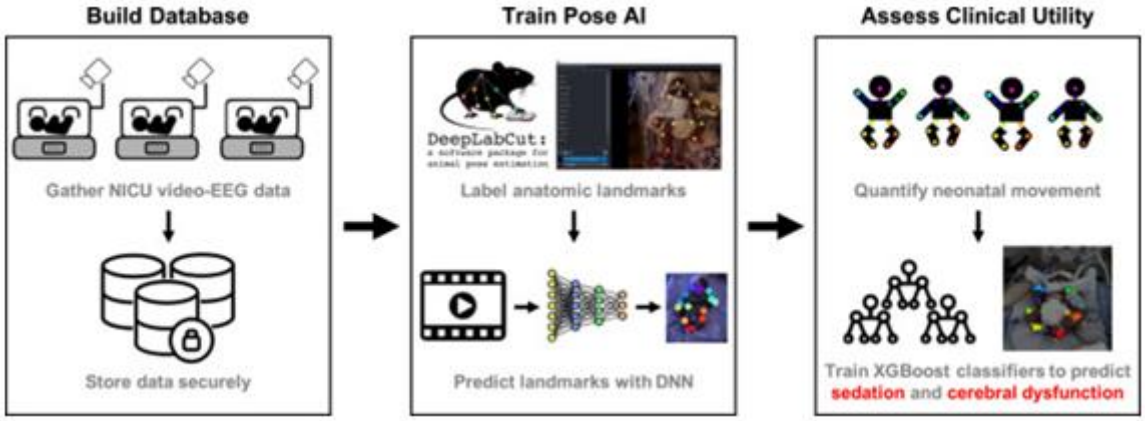


Figure 1. Deployment of Pose AI among critically ill newborns.

We assembled an extensive collection of synchronized video and EEG recordings (115 infants, 282,301 minutes captured at 30 frames per second, equaling 10.4 terabytes) housed securely on a HIPAA-compliant high-performance computing platform. A pose estimation AI system was developed and validated on 2,712 manually annotated frames using DeepLabCut to detect essential body landmarks in infants. In the final step, we highlight the clinical relevance of tracking motion patterns and leverage features derived from infant movements to anticipate both sedation levels and cerebral dysfunction. DNN stands for deep neural network.

Materials and Methods

Study population and data collection

Information was obtained retrospectively through an observational study design covering the period from February 1, 2021, to December 31, 2022. Broad enrollment standards were deliberately chosen to support the creation of robust algorithms that perform consistently regardless of race, sex, or gestational maturity, utilizing

data from a heterogeneous patient group. The number of participants was determined by the quantity of existing recordings. Eligibility encompassed every infant possessing video-EEG recordings whose chronologic age (time since birth) was no greater than one year when monitoring began. Postmenstrual age was computed by adding gestational age at birth to chronologic age, while corrected age was derived by subtracting 40 weeks from postmenstrual age [24].

The study site, Mount Sinai Hospital located in New York City, was particularly suitable due to its substantial annual birth volume of 6,000 to 8,000, along with a highly diverse patient demographic (6% Asian, 13% Black, 38% White, 27% Other, and 19% Unknown) and broad referral patterns (38% from Manhattan, 37% from other boroughs, and 25% from outside New York City), based on available maternal data for newborns entering the NICU or nursery. All video-EEG studies were performed as part of usual clinical management employing Natus XLTEK devices, Natus NeuroWorks Version 9 platform, and the established 10-20 international electrode placement system. The overall data handling process, which adheres to TRIPOD guidelines for prediction model reporting, appears in **Figure 2** [25, 26].

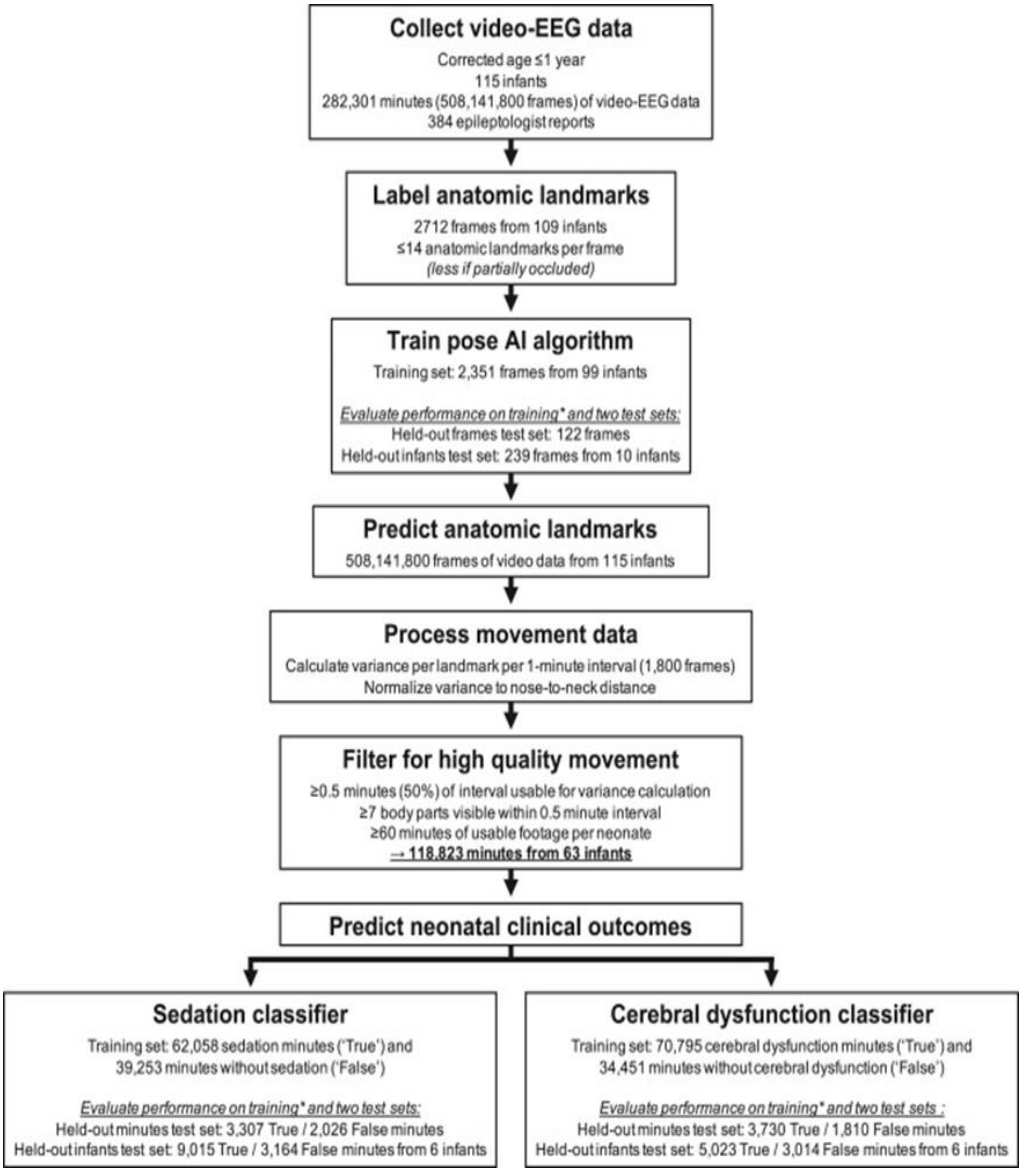


Figure 2. Overview of the data processing and filtering pipeline.

This diagram presents the entire data preparation sequence, including applied filters, participant counts at each stage, and the division into training and testing sets, in line with TRIPOD reporting recommendations [25, 26]. *Model performance was assessed on the training set via three repetitions of five-fold cross-validation. Additional

evaluations were conducted on two separate held-out validation sets: entirely new infants and held-out segments (either individual frames or time intervals) from infants included in the training process.

Ethics

The retrospective review of existing records received approval from the Institutional Review Board at the Icahn School of Medicine at Mount Sinai (IORG0000124), and consent procedures followed the approved study protocol. All video-EEG files along with corresponding electronic health information were securely moved from the Epilepsy Monitoring Unit via an encrypted channel to a dedicated HIPAA-compliant supercomputing environment at Mount Sinai for storage. Additional explicit written permission was secured from the families of infants whose images appear in the publication.

Clinical outcomes

The practical value and predictive power of pose AI were examined for two key clinical conditions: sedation and cerebral dysfunction. Phenobarbital along with continuous sedative infusions (midazolam, dexmedetomidine, and fentanyl) were categorized as sedating agents [27]. In contrast, levetiracetam and other infrequently administered antiseizure drugs such as phenytoin or oxcarbazepine were not classified as sedating in this analysis [27]. Sedation status was considered present on any day involving a sedative infusion or within four days following phenobarbital administration, accounting for the medication's extended half-life of 70 to 140 hours in infants [28]. The specific combinations of antiseizure medications observed are detailed in **Table 1**. Cerebral dysfunction (also referred to as encephalopathy) was determined daily by a board-certified epileptologist reviewing the EEG tracings. This diagnosis relied on the presence of focal or widespread disruptions in background EEG patterns, including issues with reactivity, synchrony, continuity, and rhythm slowing. A straightforward binary distinction between normal and abnormal background activity (indicating cerebral dysfunction) was selected as the outcome measure because it has demonstrated strong consistency, including in newborn populations [18]. Earlier investigations noted considerable disagreement among raters when evaluating specific individual EEG features [18] or when applying more intricate scoring systems for cerebral dysfunction [29]; therefore, those approaches were avoided as primary endpoints. While infants diagnosed with cerebral dysfunction tended to receive sedatives more often, 33% of the cohort (N = 38) belonged exclusively to one category or the other. This separation enabled subsequent statistical analyses that could help distinguish the effects of these frequently co-occurring conditions.

Table 1. Clinical characteristics of all patients included in this study.

Descriptive factor	**All (*N* = 115)**
Gestational Age at Birth, *median (range)*	38w1d[a] (23w0d–41w1d)
Age in months, *N (%)*	
<1	77 (67.0)
[1–3)	21 (18.3)
[3–6)	10 (8.7)
≥6	7 (6.1)
Sex, *N (%)*	
Female	53 (46.1)
Male	62 (53.9)
Race, *N (%)*	
American Indian or Alaskan	5 (4.3)
Asian	9 (7.8)
Black or African American	27 (23.5)
Native Hawaiian or Pacific Islander	1 (0.9)
White	32 (27.8)
Other[b]	33 (28.7)
Unknown	8 (7.0)
Ethnicity, *N (%)*	
Hispanic/Latino	33 (28.7)
Non-Hispanic	58 (50.4)
Unknown	24 (20.9)
Neurologic pathology, *N (%)*	
Hypoxic ischemic encephalopathy	24 (20.9)

Descriptive factor	All (*N* = 115)
Genetic or idiopathic epilepsy	14 (12.2)
Intraventricular hemorrhage	7 (6.1)
Stroke	7 (6.1)
Structural brain malformation	4 (3.5)
Other	15 (13.0)
None	44 (38.3)
CNS active medications[c], *N (%)*	
Levetiracetam	5 (4.3)
Levetiracetam, phenobarbital	9 (7.8)
Levetiracetam, phenobarbital, fos/phenytoin	4 (3.5)
Phenobarbital	16 (13.9)
Sedative infusion and any ASMs[d]	12 (10.4)
Other combination of ASMs[e]	5 (4.3)
None	64 (55.7)

a. w, weeks; d, days.
b. Other = option if the caregiver did not self-identify with a pre-specified racial category.
c. All medications acting on the central nervous system (CNS) that were used at any point while the patient was undergoing the video-EEG study.
d. ASMs, antiseizure medications.
e. Other ASMs were lacosamide, oxcarbazepine, topiramate, and vigabatrin.

Software development

Several pose estimation algorithms exist for analyzing human movement in videos. However, the majority are developed using adult data and perform suboptimally on infants, primarily due to differences in body proportions between infants and adults [30]. We selected DeepLabCut because of its proven adaptability across various species [22, 31], including humans [32] and robotic systems [33]. DeepLabCut [22] is a comprehensive toolkit offering a user-friendly interface for marking images, a Python-based setup that supports transfer learning, and flexible output formats suitable for further processing. Its training framework relies on transfer learning with ResNet [34], which represents the conventional method for computer vision deep learning tasks [35, 36]. Specifically, the transfer learning utilized the 50-layer ResNet architecture as the image feature extractor.

To develop the model for recognizing infant body positions, 25 frames were randomly selected from each infant, with as many as 14 anatomical landmarks annotated per frame. Achieving strong generalization in deep learning requires broad diversity within the training images [37]; therefore, selecting 25 frames per infant allowed inclusion of all available infants instead of concentrating on numerous frames from a smaller subset. Prior experience indicated limited gains beyond approximately 20 labeled images per recording [33], making 25 frames a cautious yet sufficient choice. Frames were excluded from labeling if the infant was not visible or if portions of the video were unsuitable for DeepLabCut, resulting in a final set of 2,712 frames from 109 infants available for model development and testing. All landmark annotations were completed by a single person. The precise center pixel of each landmark defined its position; enlarged representations appear in certain figures for clarity. To properly evaluate generalization and prevent information overlap [38, 39], two separate validation groups were created: 10% of infants (239 frames from 10 infants) were completely withheld from training, and an additional 5% of frames (122 frames) were held out from the remaining infants included in training. Random assignment used Python's built-in randomization tools. The model was trained on 2,351 frames from 99 infants. Hyperparameter tuning was performed with Bayesian optimization [40]. Once trained, the resulting model could transform any infant video into a set of vectorized pose data, providing X and Y coordinates along with confidence scores for each landmark in every frame.

Multiple assessment measures were applied to evaluate the infant-specific DeepLabCut model on both the training data and the randomized held-out sets of frames and infants. All frames were included in testing, even those with partial obstructions. These evaluations, along with landmark-specific analyses, proved essential for reliable deployment of computer vision models, as results can be influenced by factors such as video compression [41], labeling precision [42], and other pose-specific challenges [43]. L2 pixel error was computed as the straight-line distance between predicted and manually marked landmark positions. Effective robustness [44] was derived from differences in these L2 errors. Receiver operating characteristic area under the curve (ROC-AUC), sensitivity, specificity, and accuracy were calculated to evaluate the model's capacity to detect occluded landmarks using its

internal confidence values, with a focus on partially obscured frames. Additionally, the percentage of correct keypoints (PCK) [32]—a widely accepted pose evaluation standard—was used. PCK measures the proportion of predicted landmarks falling within 26.2 pixels of the true position (equivalent to roughly half the average infant head height in the study videos). A PCK exceeding 90% aligns with leading performance levels achieved by contemporary adult pose estimation systems [45, 46].

Statistics

An initial aggregate measure of infant neurological condition was created. Baseline movement was quantified by determining the variance of X and Y coordinates for each body part within every one-minute segment. This variance was normalized by dividing it by the median nose-to-neck distance in that interval to account for variations in infant size and camera distance. Variance was selected as it effectively captures both the frequency and magnitude of motion while remaining resistant to extreme values. Kinematic approaches employed in other infant pose studies [47] were not used here, as they proved difficult to apply reliably given the diverse camera angles and frequent partial occlusions in the video-EEG recordings. Variance was further justified by its established link to clinical alertness: a related metric based on standard deviation of motion artifacts in ECG signals had previously been associated with lethargy in late-onset sepsis [48]. Movement variance was computed per landmark over 1-minute windows and then summarized as the median value across all landmarks.

Comparisons of 1-minute movement segments across groups (for example, sedated versus non-sedated) employed a resampling method to properly handle repeated observations within subjects [49, 50]. The observed difference in median movement between groups was contrasted against a null distribution. This null distribution was generated by randomly resampling subject-level information (such as recording date, medications, and EEG findings) with replacement for each interval, then recalculating the median movement difference. The process was repeated 10,000 times. The resulting P value represented the fraction of resampled differences that were as extreme as or more extreme than the actual observed difference. Statistical significance was defined as $P < 0.05$.

Machine learning

One-minute movement summaries were subsequently used to build predictive models for sedation and cerebral dysfunction. Since classifier success depends heavily on input data quality, several preprocessing filters were applied **(Figure 2)**: postmenstrual age below 44 weeks for greater uniformity, a minimum of seven visible body parts per minute to guarantee adequate landmark data, variance calculations requiring at least 30 seconds of usable data per minute for reliable estimates, and a threshold of at least 60 minutes of valid recordings per infant to provide sufficient data volume. After filtering, the dataset for model training contained 118,823 minutes from 63 infants **(Figure 2)**.

Three algorithms were developed for sedation prediction: logistic regression, support vector machines, and XGBoost. XGBoost hyperparameters [51] were optimized through 5,000 iterations of random search, with the optimal configuration identified using three repetitions of five-fold cross-validation. In each random search trial, 12 hyperparameters were sampled from defined ranges, and performance was measured by the average macro-averaged F1 score across the cross-validation repeats to address potential class imbalance. The final model was chosen based on a combination of metrics including F1 score, accuracy, ROC-AUC, and precision-recall AUC [51, 52]. Performance of the selected model was then tested on three datasets **(Figure 2)**: the training data via repeated cross-validation and two independent test sets consisting of completely held-out infants and held-out time segments from training infants. Random splits again used Python's native randomization. To explore the contribution of individual landmarks, feature importance was assessed using XGBoost's 'gain' metric. As XGBoost captures interactions among features, 'gain' offers relative importance but not absolute standalone performance. Univariate metrics from XGBoost are often unreliable [53]. Therefore, a distribution of gain values was produced by repeatedly subsampling 80% of the data at random and recalculating gain 500 times.

Role of funding source

Funding for this project was provided by the Friedman Brain Institute Fascitelli Scholar Junior Faculty Grant and the Thrasher Research Fund Early Career Award. Additional support came from the computational infrastructure and expert staff at Scientific Computing and Data, Icahn School of Medicine at Mount Sinai, backed by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. The research also benefited from the Office of Research Infrastructure of the National Institutes of Health under awards S10OD026880 and S10OD030463. The authors bear sole responsibility for the

content, which does not necessarily reflect the official position of the National Institutes of Health. The funding organizations had no involvement in study design, data gathering, analysis, interpretation, or manuscript preparation.

Results and Discussion

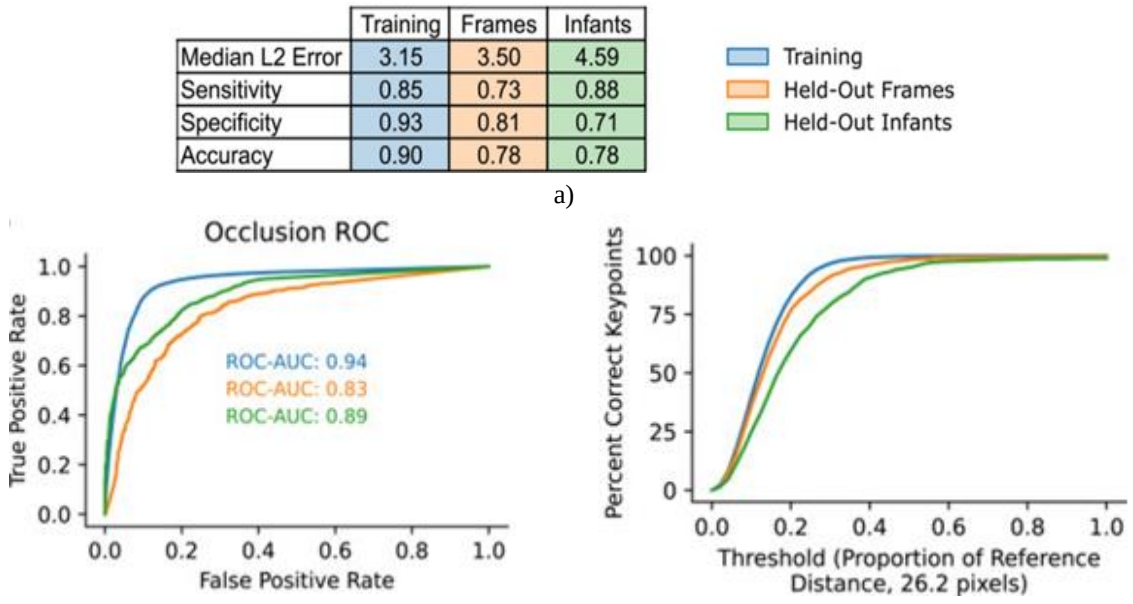
Cohort characteristics

Video-EEG recordings and associated clinical information were gathered between February 2021 and December 2022, yielding 115 eligible infants **(Table 1)**. Race and ethnicity were based on caregiver reports, while sex and age were extracted from the electronic health record. Reflecting standard video-EEG application in therapeutic hypothermia, the majority of infants were under one month of age (67%), and hypoxic ischemic encephalopathy (HIE) was the leading suspected or confirmed condition (21%). Levetiracetam and phenobarbital represented the most frequently used initial antiseizure medications (ASMs).Training and evaluating an infant pose recognition algorithm

A DeepLabCut computer vision model was trained to estimate infant body pose. Training utilized 2,351 manually annotated frames from 99 infants, with performance assessed on the training set as well as two validation groups: held-out frames from training infants and entirely new infants **(Figure 3)** [38, 39]. Low median L2 pixel errors were achieved (3.2 pixels on training data, 3.5 on held-out frames, and 4.6 on held-out infants); **(Figure 3a)** for videos with 320 × 240 resolution (76,800 total pixels). These errors were similar to or better than those reported in previous studies and fell below typical human annotation variability (6.25 pixels) [22]. Effective robustness [44] measured only 1.1 pixels, indicating comparable performance across validation sets and alignment within 10% of the camera's spatial resolution limit (1 pixel). Strong overall accuracy was further confirmed by ROC-AUC values above 0.83 for occlusion detection **(Figure 3b)** and over 95% of predictions falling within half the infant head size **(Figure 3c)**. Performance ranked highest on training data, followed by held-out frames and then held-out infants, except for sensitivity and ROC-AUC measures. The PCK metric, standard for pose tracking evaluation, reached levels equivalent to current best-performing adult human systems [45, 46].

	Training	Frames	Infants
Median L2 Error	3.15	3.50	4.59
Sensitivity	0.85	0.73	0.88
Specificity	0.93	0.81	0.71
Accuracy	0.90	0.78	0.78

a)



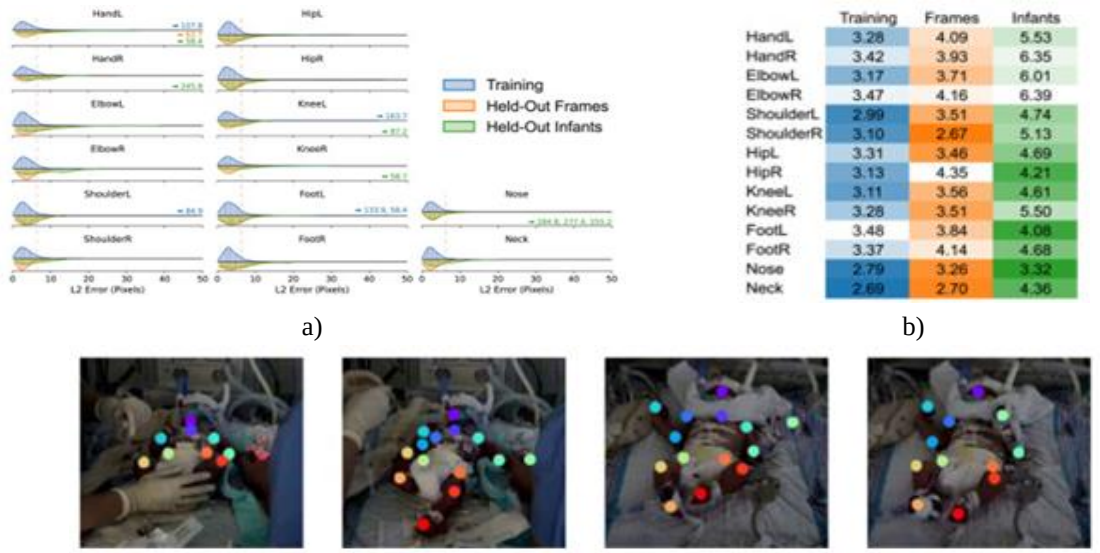
b) c)

Figure 3. Evaluation metrics assessing the effectiveness of pose AI for infants in the NICU, shown across three distinct datasets.

(a) L2 pixel error remained low throughout: on training data (blue, median 3.2 pixels, N = 18,399 landmarks), held-out frames from training infants (orange, median 3.5 pixels, N = 797 landmarks), and frames from completely held-out infants (green, median 4.6 pixels, N = 973 landmarks). These results were in line with published literature and below average human inter-rater variability when using the identical annotation software [22]. (b) The system achieved strong ROC-AUC values for identifying landmark occlusion. (c) Percentage of correct keypoints (PCK) exceeded 95% within the reference threshold of 26.2 pixels (half an infant head height) for both training and test

sets, matching the performance of leading adult pose estimation systems [45, 46]. All measures generally showed the anticipated pattern of strongest results on training data, followed by held-out frames and then held-out infants, except for sensitivity and ROC-AUC, which performed better than anticipated on frames from the held-out infants (green).

Consistent strong accuracy, as indicated by L2 pixel error, was maintained across every anatomical landmark **(Figure 4a)**. Violin plots [54] illustrated the distribution of these errors, which stayed well below typical human annotation variability (6.25 pixels) and contained very few outliers (0.06%), addressing a limitation noted in earlier methods [47]. The nose and neck landmarks displayed the smallest errors, whereas elbows and hands showed the largest **(Figure 4b)**, a pattern also observed in adult-focused pose estimation tools [55].



a) b)

c)

Figure 4. Assessment of AI-based pose estimation performance for NICU infants by individual anatomical landmarks.

(a) L2 pixel error distributions, displayed as violin plots with median and interquartile ranges, remained low and consistent with existing literature across all annotated landmarks. Results are shown for training data (blue, median 3.2 pixels, N = 18,399), held-out frames from training infants (orange, median 3.5 pixels, N = 797), and frames from held-out infants (green, median 4.6 pixels, N = 973). Thirteen landmarks (0.06%) exceeded 50 pixels in error and are indicated by arrows. The vertical red dashed line marks the standardized mean human inter-rater variability (6.25 pixels) for the camera resolution used [22]. (b) Median L2 pixel error was smallest for the nose and neck landmarks and largest for the elbows and hands. (c) Representative video frames from an infant receiving intravenous cannulation in the left arm, with AI-predicted anatomical landmarks overlaid on expected positions.

Example frames from this infant undergoing left arm intravenous cannulation demonstrated close alignment between predicted (colored) landmarks and actual locations **(Figure 4c)**. Importantly, the pose AI system avoided mistakenly tracking adult hands present in the scene, successfully identified the nose despite a continuous positive airway pressure device and two varying chin straps, accurately detected the left arm following cannulation (as seen in the right panels), and maintained reliable landmark detection across two different camera perspectives.

Once the infant pose recognition model was established, X and Y coordinates for all landmarks were generated across the complete video dataset from every participant (508,141,800 frames, equivalent to 282,301 minutes from 115 infants). This expansion incorporated videos from six additional infants who were not part of the initial training or validation process, increasing the total from 109 to 115 infants for all subsequent analyses.

Association between infant movement predicted by pose AI and neurologic changes

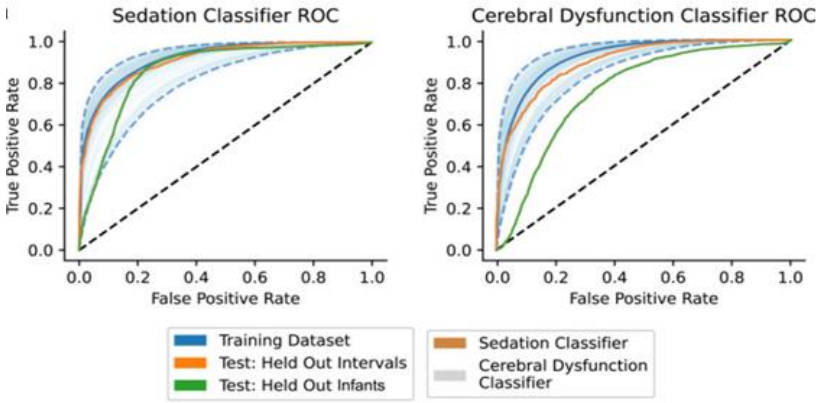
To examine the real-world usefulness of pose estimation in a standard intensive care setting, movement variance served as a straightforward indicator of infant activity expected to correspond with major clinical variables such as postmenstrual age (PMA), cerebral dysfunction, and sedative medication exposure. Forty-seven percent of the video recordings (range 0–100% per infant) contained adequate data to compute variance over 1-minute intervals for at least seven body parts, yielding 132,964 usable 1-minute segments out of the total 282,301 minutes.

Movement levels rose with increasing corrected age both in infants exhibiting EEG abnormalities or receiving sedatives (47-fold difference between the youngest and oldest age groups, permutation $P = 1 \times 10^{-4}$, 10,000 permutations) and in those without such factors (15-fold difference, permutation $P = 2.4 \times 10^{-3}$). When stratified by corrected age at the time of recording, term-born infants generally displayed greater movement than preterm infants, though this pattern was less pronounced among those with abnormal EEG findings or sedative use.

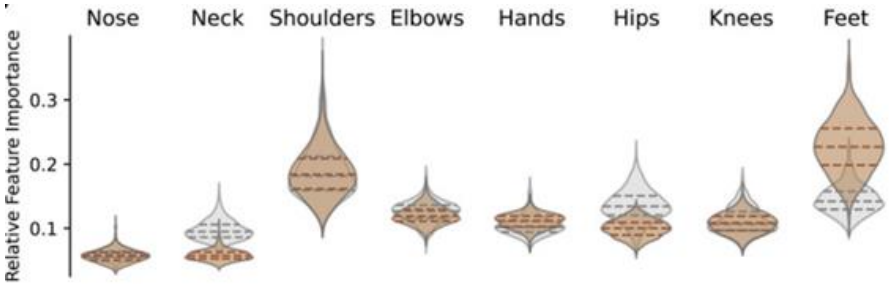
Subsequent analyses tested whether movement declined in infants on sedative medications or those with cerebral dysfunction. These comparisons were limited to neonates with PMA below 44 weeks, given the non-linear relationship between age and movement that appeared influenced by underlying brain conditions. Reduced movement was documented in infants with cerebral dysfunction ($P = 1 \times 10^{-4}$), those receiving phenobarbital ($P = 3.6 \times 10^{-3}$), cases involving both conditions ($P = 1 \times 10^{-4}$), and infants on sedative infusions ($P = 1 \times 10^{-4}$). Comparable patterns emerged when examining the hypoxic ischemic encephalopathy (HIE) subgroup separately from other neurological conditions. Overall, these findings confirm that movement increases with age and decreases in the presence of sedative medications or cerebral dysfunction, with the greatest reduction observed when both factors coincide.

Pose AI predicts sedation and cerebral dysfunction

The movement variance metric described earlier served to confirm the clinical relevance of pose AI but did not incorporate differential contributions or interactions among specific landmarks. To address this, dedicated predictive classifiers were constructed for sedation and cerebral dysfunction. Additional data filtering was applied to improve variance stability and ensure adequate per-patient data volume **(Figure 2)**, producing a dataset of 118,823 minutes from 63 infants (ranging from 65 to 14,409 minutes per infant). A large proportion of these intervals involved partial landmark occlusion. Among the tested algorithms, Extreme Gradient Boosting (XGBoost) outperformed logistic regression and support vector machines on all evaluation metrics for sedation prediction [51] and was therefore selected for further use. For sedation classification **(Figure 5a)**(left panel), the model demonstrated robust performance on training data (median ROC-AUC = 0.90), held-out time intervals (ROC-AUC = 0.91), and a completely independent set of infants (ROC-AUC = 0.87). Comparable results were obtained for the cerebral dysfunction classifier **(Figure 5a)** (right panel): median ROC-AUC of 0.91 on training data, 0.90 on held-out intervals, and 0.76 on held-out infants. Randomization procedures were reviewed to determine whether data distribution differences might account for performance variations. Most baseline characteristics were balanced across groups. In the held-out infant cohort used for cerebral dysfunction evaluation, no participants received centrally acting medications (Chi-squared P = 0.02), which may have contributed to the comparatively lower performance in that group. Feature importance analysis for both classifiers revealed that feet and shoulders consistently ranked among the most influential landmarks **(Figure 5b)**.



a)



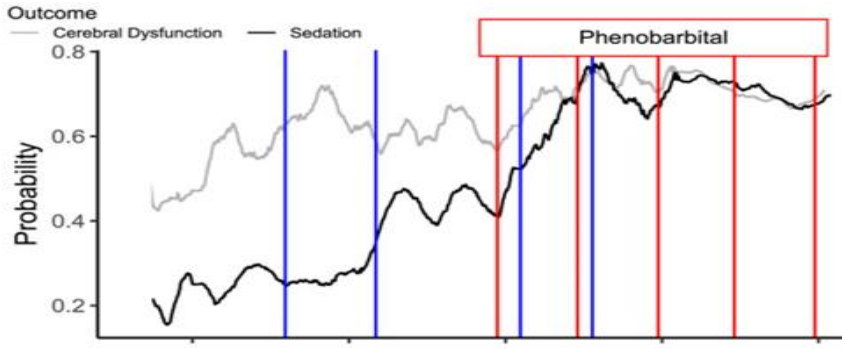
b)

Figure 5. Classifiers based on infant movement patterns successfully forecast sedation and cerebral dysfunction.

(a) Receiver operating characteristic curves generated by XGBoost models for sedation prediction (left panel) and cerebral dysfunction (right panel), developed using features from pose AI. ROC-AUC results for XGBoost are displayed from portions of the training set (blue, with median as solid line and 2.5th/97.5th percentiles as dashed lines) as well as on independent test sets consisting of held-out time intervals (orange) and held-out infants (green). The black dashed line represents chance-level performance (ROC-AUC = 0.5). (b) Relative feature importance was analyzed for each XGBoost model in panel (a). As anticipated, landmarks from the limbs contributed more substantially than those from the nose and neck in predicting both sedation and cerebral dysfunction. Distributions of relative importance are presented as violin plots showing median and 25th/75th quartiles, derived from 100 repetitions of five-fold cross-validation for each classifier.

Sedation prediction in an infant receiving therapeutic hypothermia

To demonstrate the practical application of sedation prediction at the individual level, sedation probability over time was visualized for a specific case: an infant born at 37 + 1 weeks gestation with Apgar scores of 1, 4, and 5 at 1, 5, and 10 minutes respectively, who underwent therapeutic hypothermia for suspected hypoxic ischemic encephalopathy (HIE) **(Figure 6)**. This infant was administered a 20 mg/kg loading dose of phenobarbital prior to rewarming in response to rising epileptiform activity on EEG. A therapeutic trough level was achieved following the load (C_trough = 20.9 mg/L), followed by maintenance doses of 2 mg/kg every 12 hours. During the video-EEG monitoring period, the infant also received antibiotics and electrolyte support but no additional medications. Using the two-step prediction method, sedation probability rose initially during the cooling phase and increased further after phenobarbital administration (median P(Sedation) = 0.23 across 1,476 minutes before phenobarbital versus 0.59 across 2,268 minutes afterward; Wilcoxon rank-sum test $P < 10^{-10}$). The infant carried a clinical diagnosis of mild cerebral dysfunction throughout monitoring, consistent with predicted probabilities ranging from 0.42 to 0.90 **(Figure 6a)**, (grey shading). Cerebral dysfunction probability also rose following phenobarbital, though to a lesser extent, underscoring the partial overlap between these two conditions. Sample video frames with overlaid heatmaps illustrating infant motion over the prior 5 minutes provide a visual representation of the observed reduction in activity **(Figure 6d)**.



a)

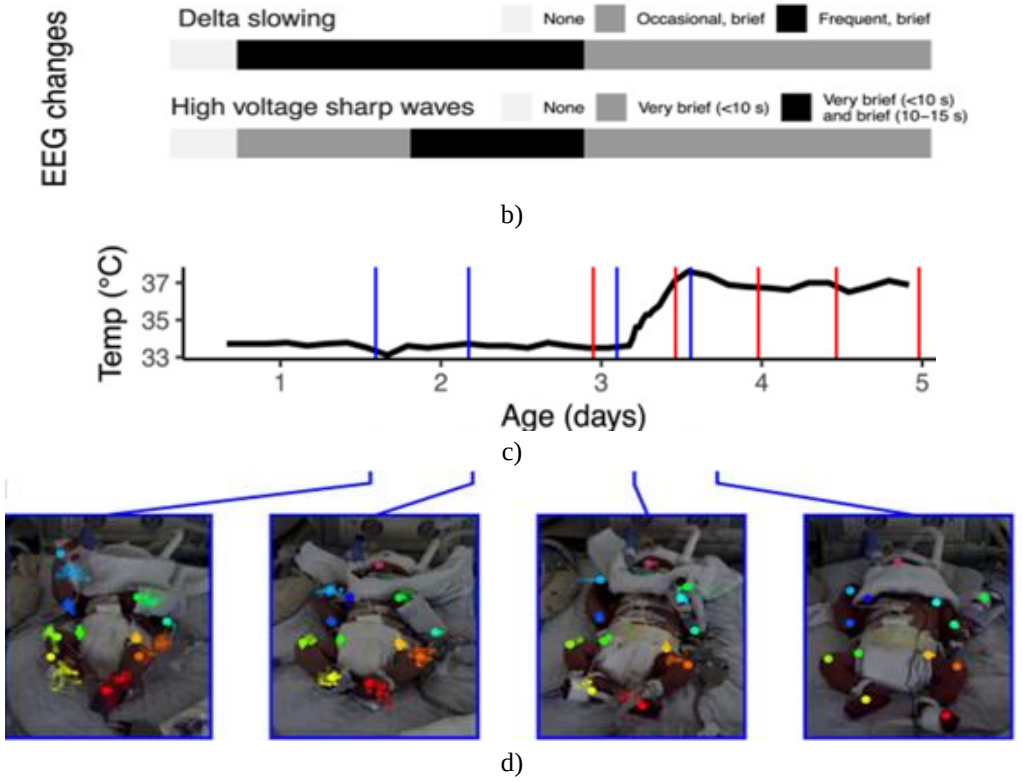


Figure 6. Pose AI forecasts sedation levels during therapeutic hypothermia in an infant with suspected hypoxic ischemic encephalopathy (HIE).

(a) Predicted probabilities (y-axis) of sedation (black line) and cerebral dysfunction (grey line) plotted against postnatal age in days (x-axis) for a term infant with probable HIE who was transferred to Mount Sinai Hospital for therapeutic hypothermia. Sedation probability rose following the start of cooling (panel c) and again after phenobarbital dosing (marked by red vertical lines). Cerebral dysfunction probability remained elevated for the entire video monitoring period, consistent with the infant's mild cerebral dysfunction. (b) EEG recordings revealed heightened epileptiform activity at two and three days of life, prompting phenobarbital administration to prevent seizures during the rewarming phase. (c) Core temperature trajectory throughout the cooling and rewarming process. (d) Representative video frames captured at rising sedation levels, each accompanied by a motion heatmap summarizing infant activity over the preceding 5 minutes. Blue vertical lines link these frames to the corresponding clinical events shown in panels (a)–(c).

This study demonstrates that pose AI effectively tracks infant movement and can forecast both sedation status and cerebral dysfunction among critically ill newborns in their routine clinical environment. Similar workflow approaches are already standard in neonatal cardiorespiratory monitoring systems used to detect arrhythmias and apnea-bradycardia episodes.

The technology holds wide potential because sedation is commonplace in neonatal care. Typical indications include mechanical ventilation, procedures at the bedside, diagnostic imaging, therapeutic hypothermia, extracorporeal membrane oxygenation support, and conditions such as pulmonary hypertension or unrepaired Tetralogy of Fallot [5]. Excessive sedation can prolong ventilator dependence, contribute to brain injury, and trigger withdrawal symptoms [56]. Insufficient sedation, on the other hand, is linked to pain, pulmonary hypertensive episodes, and complications including repeated imaging, catheter dislodgement, and unplanned extubation [57]. Pose-based tracking could help mitigate these issues by enabling more accurate sedation adjustment.

Infants also commonly undergo evaluation for encephalopathy. Cerebral dysfunction, identified on EEG, may signal encephalopathy arising from iatrogenic causes or unidentified clinical issues, such as medication effects, low blood sugar, infection, elevated ammonia levels, or lasting damage from stroke or seizures [3]. Our results show that pose tracking can reliably predict cerebral dysfunction, indicating its value for detecting and following encephalopathy.

Key strengths of the study include the use of interpretable inputs (movement of specific body landmarks) derived from a sizable and demographically diverse cohort (25.2% white non-Hispanic, 46% female), which helps address concerns regarding bias and transparency in AI applications [58]. A known obstacle is that adult-trained pose models typically underperform on infants because of dissimilar body proportions [59]. This was resolved by training exclusively on patients under one year of age. Both the pose estimation and downstream neurological predictions maintained high accuracy, even on independent held-out datasets. This contrasts with earlier research, which involved small samples ($N < 30$) and did not link derived infant pose data to neurological status in the NICU. One prior effort associated pose with neuromotor risk scores [47], yet it was limited (19 infants, 9.7 hours of video) and conducted outside the NICU in a controlled setup. Here, we connected movement variance to clinical neurological state, drawing a parallel to how heart rate variability reflects autonomic function [60]. Additionally, we provide assessments of both within-subject and between-subject reliability by testing on held-out frames/intervals and held-out infants, respectively. Strong within-subject consistency is especially relevant, as early clinical deployment would likely involve longitudinal monitoring of disease course and recovery in the same infant over weeks to months, potentially using edge computing devices with options for on-site calibration.

A constraint of pose AI is the requirement for adequate visualization of the infant to accurately locate landmarks. In our retrospective dataset, 47% of video recordings yielded usable data. Several factors contributed to this proportion, and multiple avenues exist for improvement. The original videos were recorded at low 240p resolution, which can hinder landmark detection amid the busy ICU environment; higher-resolution cameras could resolve this. Strict quality filters were applied to maintain reliable tracking. Expanded datasets encompassing more infants and the adoption of advanced transformer-based vision models [61] may enhance results in challenging, low-confidence segments. The retrospective video-EEG cameras were bulky, restricting optimal placement and causing obstructions. Dedicated compact edge devices could minimize such visual interference. Another difficulty in building clinically useful classifiers is the partial overlap between neurological outcomes. Future studies incorporating more detailed medication timing, pharmacokinetic modeling, and direct EEG signal correlation could refine labeling accuracy and better separate these interrelated effects. This would be especially useful for phenobarbital, given its well-characterized pharmacokinetics in newborns [28, 62, 63]. A study-specific limitation involves the modest size of held-out groups for clinical outcome validation. Only 10% of infants and 5% of frames/minutes from training infants were withheld **(Figure 2)**. While this approach maximized training data, it constrained thorough assessment of generalizability. Moreover, because randomization occurred independently for the pose model and subsequent classifiers, the reported metrics reflect the reliability of each pipeline stage rather than the integrated system. Collectively, these findings indicate that pose AI-based monitoring is viable in the NICU, although larger-scale training with refined labels and broader validation will be necessary prior to routine clinical implementation.

The current work introduces novel neuro-telemetry techniques along with their initial application, while acknowledging several constraints and outlining numerous directions for future research. Models were developed using data from a single center, so both the pose algorithm and neurological predictions require testing on recordings from other hospitals and varying camera systems. The analyses relied on retrospectively collected video-EEG data. Prospective studies in NICU populations with continuous video from birth will be needed to train more refined models. For instance, our sedation labels were based on medication administration; future prospective designs could incorporate validated sedation or analgesia scoring tools [57]. Emerging evidence suggests that combining multiple monitoring modalities improves sedation assessment [14], raising the possibility that integration with EEG could yield the most objective measure of neonatal sedation. Cerebral dysfunction represents an EEG indicator of encephalopathy, but subsequent investigations might incorporate additional markers such as alterations in feeding, breathing patterns, or autonomic activity. Prospective video studies could also examine movement in relation to clinical events not routinely captured by video-EEG, including sepsis, pre-treatment seizures, and withdrawal syndromes.

Conclusion

Although artificial intelligence holds transformative potential for pediatrics, its use for one of the most critical outcomes—neurological changes in the NICU—has been limited [58]. Neurological evaluation carries profound implications for long-term development, yet continuous neuro-monitoring remains unavailable in most NICUs despite years of EEG research and specialized neuro-intensive care programs. With ongoing NICU

deregionalization [64], pose AI could help meet this need by providing a low-burden, interpretable, and expandable solution for ongoing neurological surveillance. Validation in larger, prospective, and geographically diverse cohorts represents an essential step toward widespread adoption of pose AI in neonatal intensive care.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Khan OA, Garcia-Sosa R, Hageman JR. Core concepts: neonatal neurological examination. Neoreviews. 2014;15:e316-e24.
2. Volpe JJ, Inder TE, Darras BT. Volpe's neurology of the newborn. Elsevier; 2018:1-224.
3. Russ JB, Simmons R, Glass HC. Neonatal encephalopathy: beyond hypoxic-ischemic encephalopathy. Neoreviews. 2021;22:e148-e62.
4. Verstraete EH, Blot K, Mahieu L. Prediction models for neonatal health care-associated sepsis: a meta-analysis. Pediatrics. 2015;135:e1002-e14.
5. Mayock DE, Gleason CA. Pain and sedation in the NICU. Neoreviews. 2013;14:e22-e31.
6. Donato J, Rao K, Lewis T. Pharmacology of common analgesic and sedative drugs used in the neonatal intensive care unit. Clin Perinatol. 2019;46:673-92.
7. Morgan ME, Kukora S, Nemshak M. Neonatal Pain, Agitation, and Sedation Scale's use, reliability, and validity: a systematic review. J Perinatol. 2020;40(12):1753-63.
8. Sarnat HB, Sarnat MS. Neonatal encephalopathy following fetal distress. A clinical and electroencephalographic study. Arch Neurol. 1976;33:696-705.
9. Shankaran S, Laptook AR, Ehrenkranz RA. Whole-body hypothermia for neonates with hypoxic-ischemic encephalopathy. N Engl J Med. 2005;353:1574-84.
10. Pavageau L, Sánchez PJ, Steven Brown L. Inter-rater reliability of the modified Sarnat examination in preterm infants at 32-36 weeks' gestation. Pediatr Res. 2020;87:697-702.
11. Abend NS, Topjian AA, Williams S. How much does it cost to identify a critically ill child experiencing electrographic seizures? J Clin Neurophysiol. 2015;32:257-64.
12. Luton A, Hernandez J, Patterson CR. Preventing pressure injuries in neonates undergoing therapeutic hypothermia for hypoxic-ischemic encephalopathy: an interprofessional quality improvement project. Adv Neonatal Care. 2017;17:237-44.
13. Moyer LB, Lauderbaugh DL, Worten K. High-stage device-related pressure injury reduction in a neonatal intensive care unit: a quality improvement project. Pediatr Qual Saf. 2022;7:e554.
14. Giordano V, Deindl P, Goeral K. The power of N-PASS, aEEG, and BIS in detecting different levels of sedation in neonates: a preliminary study. Paediatr Anaesth. 2018;28:1096-104.
15. Schultz B, Schultz M, Boehne M. EEG monitoring during anesthesia in children aged 0 to 18 months: amplitude-integrated EEG and age effects. BMC Pediatr. 2022;22:156.
16. Moghadam SM, Airaksinen M, Nevalainen P. An automated bedside measure for monitoring neonatal cortical activity: a supervised deep learning-based electroencephalogram classifier with external cohort validation. Lancet Digit Health. 2022;4:e884-e92.
17. Shellhaas RA, Chang T, Tsuchida T. The American clinical neurophysiology society's guideline on continuous electroencephalography monitoring in neonates. J Clin Neurophysiol. 2011;28:611-7.
18. Massey SL, Shou H, Clancy R. Interrater and intrarater agreement in neonatal electroencephalogram background scoring. J Clin Neurophysiol. 2019;36:1-8.
19. Hogan R, Mathieson SR, Luca A. Scaling convolutional neural networks achieves expert-level seizure detection in neonatal EEG. 2024.
20. Barnett AJ, Guo Z, Jing J. Improving clinician performance in classifying EEG patterns on the ictal–interictal injury continuum using interpretable machine learning. NEJM AI. 2024;1.

21. Ansari A, Pillay K, Arasteh E. Resting state electroencephalographic brain activity in neonates can predict age and is indicative of neurodevelopmental outcome. Clin Neurophysiol. 2024;163:226-35.
22. Mathis A, Mamidanna P, Cury KM. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018;21:1281-9.
23. Cao Z, Hidalgo G, Simon T. OpenPose: realtime multi-person 2D pose estimation using Part Affinity fields. IEEE Trans Pattern Anal Mach Intell. 2018;43:172-86.
24. Blackmon LR. Age terminology during the perinatal period. Pediatrics. 2004;114:1362-4.
25. Moons KGM, Altman DG, Reitsma JB. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1-W73.
26. Collins GS, Reitsma JB, Altman DG. Transparent reporting of a multivariable prediction model for individual Prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55-63.
27. Bättig L, Dünner C, Cserpan D. Levetiracetam versus phenobarbital for neonatal seizures: a retrospective cohort study. Pediatr Neurol. 2023;138:62-70.
28. Pacifici GM. Clinical pharmacology of phenobarbital in neonates: effects, metabolism and pharmacokinetics. Curr Pediatr Rev. 2016;12:48-54.
29. Dhakar MB, Sheikh ZB, Desai M. Developing a standardized approach to grading the level of brain dysfunction on EEG. J Clin Neurophysiol. 2023;40:553-61.
30. Sciortino G, Farinella GM, Battiato S. On the estimation of children's poses. Lect Notes Comput Sci. 2017;10485:410-21.
31. Suryanto ME, Saputra F, Kurnia KA. Using DeepLabCut as a real-time and markerless tool for cardiac physiology assessment in zebrafish. Biology. 2022;11.
32. Insafutdinov E, Pishchulin L, Andres B. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. Lect Notes Comput Sci. 2016;9910:34-50.
33. Lu J, Jayakumari A, Richter F. SuPer deep: a surgical perception framework for robotic tissue manipulation using deep learning for feature extraction. Proc IEEE Int Conf Robot Autom. 2020;4783-9.
34. He K, Zhang X, Ren S. Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2015;770-8.
35. Zhuang F, Qi Z, Duan K. A comprehensive survey on transfer learning. Proc IEEE. 2021;109:43-76.
36. Ramanujan V, Nguyen T, Oh S. On the connection between pre-training data diversity and fine-tuning robustness. Adv Neural Inf Process Syst. 2023;36.
37. Mathis A, Schneider S, Lauer J. A primer on motion capture with deep learning: principles, pitfalls, and perspectives. Neuron. 2020;108:44-65.
38. Tomašev N, Harris N, Baur S. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. Nat Protoc. 2021;16(6):2765-87.
39. Whalen S, Schreiber J, Noble WS. Navigating the pitfalls of applying machine learning in genomics. Nat Rev Genet. 2021;23(3):169-81.
40. Nogueira F. Bayesian Optimization: open source constrained global optimization tool for Python. 2014.
41. Dodge S, Karam L. Understanding how image quality affects deep neural networks. 2016 8th Int Conf Qual Multimed Exp QoMEX. 2016.
42. Song H, Kim M, Park D. Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Networks Learn Syst. 2020;34:8135-53.
43. Huang J, Zhu Z, Guo F. The devil is in the details: delving into unbiased data processing for human pose estimation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2019;5699-708.
44. Taori R, Dave A, Shankar V. Measuring robustness to natural distribution shifts in image classification. Adv Neural Inf Process Syst. 2020;33:18583-99.
45. Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2013;1653-60.
46. Andriluka M, Pishchulin L, Gehler P. 2D human pose estimation: new benchmark and state of the art analysis. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2014;3686-93.
47. Chambers C, Seethapathi N, Saluja R. Computer vision to automatically assess infant neuromotor risk. IEEE Trans Neural Syst Rehabil Eng. 2020;28:2431-42.

48. Cabrera-Quiros L, Kommers D, Wolvers MK. Prediction of late-onset sepsis in preterm infants using monitoring signals and machine learning. Crit Care Explor. 2021;3:e0302.
49. Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman and Hall/CRC; 1994.
50. Ho J, Tumkaya T, Aryal S. Moving beyond P values: data analysis with estimation graphics. Nat Methods. 2019;16:565-6.
51. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016.
52. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-30.
53. Cava W, Bauer C, Moore JH. Interpretation of machine learning predictions for patient outcomes in electronic health records. AMIA Annu Symp Proc. 2019;2019:572.
54. Hintze JL, Nelson RD. Violin plots: a box plot-density trace synergism. Am Stat. 1998;52:181-4.
55. Bulat A, Kossaifi J, Tzimiropoulos G. Toward fast and accurate human pose estimation via soft-gated skip connections. Proc 15th IEEE Int Conf Autom Face Gesture Recognit FG. 2020;8-15.
56. Giordano V, Deindl P, Kuttner S. The Neonatal Pain, Agitation and Sedation Scale reliably detected oversedation but failed to differentiate between other sedation levels. Acta Paediatr. 2014;103:e515-e21.
57. McPherson C, Ortinau CM, Vesoulis Z. Practical approaches to sedation and analgesia in the newborn. J Perinatol. 2020;41(3):383-95.
58. Carroll AE, Christakis DA. Call for papers on artificial intelligence applied to pediatric care. JAMA Pediatr. 2023;177:884-5.
59. Antink CH, Ferreira JCM, Paul M. Fast body part segmentation and tracking of neonatal video data using deep learning. Med Biol Eng Comput. 2020;58:3049-61.
60. Berntson GG, Bigger JT, Eckberg DL. Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology. 1997;34:623-48.
61. Karaev N, Rocco I, Rocco I. CoTracker: it is better to track together. 2023.
62. Šíma M, Michaličková D, Slanař O. What is the best predictor of phenobarbital pharmacokinetics to use for initial dosing in neonates? Pharmaceutics. 2021;13:1-15.
63. Thibault C, Massey SL, Abend NS. Population pharmacokinetics of phenobarbital in neonates and infants on extracorporeal membrane oxygenation and the influence of concomitant renal replacement therapy. J Clin Pharmacol. 2021;61:378-87.
64. Boghossian NS, Geraci M, Phibbs CS. Trends in resources for neonatal intensive care at delivery hospitals for infants born younger than 30 Weeks' gestation, 2009-2020. JAMA Netw. 2023;6:e2312107.