Donna Dowling, PhD, RN, Desi M. Newberry, DNP, NNP-BC, and Leslie Parker, PhD, APRN, FAAN ❍ Section Editors

Original Research

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

Performance Evaluation of a Supervised Machine Learning Pain Classification Model Developed by Neonatal Nurses

Renee C. B. Manworren, PhD, APRN, PCNS-BC, AP-PMN, FAAN; Susan Horner, PhD, APRN-CNS, PNC-NIC; Ralph Joseph, BS; Priyansh Dadar, MSBA; Naomi Kaduwela, MSIA

Background: Early-life pain is associated with adverse neurodevelopmental consequences; and current pain assessment practices are discontinuous, inconsistent, and highly dependent on nurses' availability. Furthermore, facial expressions in commonly used pain assessment tools are not associated with brain-based evidence of pain.
Purpose: To develop and validate a machine learning (ML) model to classify pain.
Methods: In this retrospective validation study, using a human-centered design for Embedded Machine Learning Solutions approach and the Neonatal Facial Coding System (NFCS), 6 experienced neonatal intensive care unit (NICU) nurses labeled data from randomly assigned iCOPEvid (infant Classification Of Pain Expression video) sequences of 49 neonates undergoing heel lance. NFCS is the only observational pain assessment tool associated with brain-based evidence of pain. A standard 70% training and 30% testing split of the data was used to train and test several ML models. NICU nurses' interrater reliability was evaluated, and NICU nurses' area under the receiver operating characteristic curve (AUC) was compared with the ML models' AUC.
Results: Nurses weighted mean interrater reliability was 68% (63%-79%) for NFCS tasks, 77.7% (74%-83%) for pain intensity, and 48.6% (15%-59%) for frame and 78.4% (64%-100%) for video pain classification, with AUC of 0.68. The best performing ML model had 97.7% precision, 98% accuracy, 98.5% recall, and AUC of 0.98.
Implications for Practice and Research: The pain classification ML model AUC far exceeded that of NICU nurses for identifying neonatal pain. These findings will inform the development of a continuous, unbiased, brain-based, nurse-in-the-loop Pain Recognition Automated Monitoring System (PRAMS) for neonates and infants.
Key Words: computer vision, interrater reliability, neonatal pain, pain classification, supervised machine learning

BACKGROUND AND SIGNIFICANCE

Current pain assessment practices are discontinuous, inconsistent, highly variable, and dependent on nurses' availability and methods used to alert nurses to the presence of pain.[1-4] Inability to self-report pain makes neonates vulnerable to underrecognition and both under- and overtreatment of pain.[1,5] In neonatal intensive care units (NICUs), vulnerable neonates have a median of 16 painful procedures per day; only 21% are treated in anticipation of pain.[6-9] Because of their underdeveloped pain inhibitory pathways, neonates are 30% to 50% more sensitive to pain than adults and have reduced pain tolerance as compared with children.[10,11] Early-life pain is associated with abnormal structural and functional brain development and results in adverse neurodevelopmental consequences, including cognitive and memory impairments, altered emotional functioning, psychopathologies, and global pain sensitivity.[6,12,13]

Observational tools are the primary technique nurses use for assessing neonatal pain; however, popular observational tools used to assess neonatal pain include facial expressions that are not associated with brain-based evidence of pain.[5,14-18] Advances in computer software solutions can assist in identifying facial indicators of pain, but lack of standardized reporting has limited comparison of these computer techniques, such as machine learning (ML) and deep learning (DL) artificial intelligence (AI) models, and their reproducibility.[19]

Author Affiliations: Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, Illinois (Drs Manworren and Horner); Feinberg School of Medicine, Northwestern University, Chicago, Illinois, and College of Nursing and Health Innovation, University of Texas at Arlington, Arlington, Texas (Dr Manworren); and Kavi Global, Barrington, Illinois (Messrs Joseph and Dadar and Ms Kaduwela).

This study was supported by a Perinatal Origins of Disease grant from Stanley Manne Children's Research Institute and a grant from the National Science Foundation grant number #2205472. Dr Manworren is supported by the Posey and Fred Love Endowment of Nursing Research at Ann & Robert H. Lurie Children's Hospital of Chicago.

Authors Manworren and Horner have no competing interests to declare. Author Joseph is a contractor with KaviLab at Kavi Global and has been financially compensated for performing services. Authors Dadar and Kaduwela are employees at KaviLab and Kavi Global and have not been financially compensated for their work on this study and manuscript.

Ethical approval was provided by the Institutional Review Board of the Ann & Robert H. Lurie Children's Hospital of Chicago (2021-4348). The iCOPEvid (infant Classification Of Pain Expression video dataset) was obtained with permission from Dr Sheryl Brahnam.

Renee C. B. Manworren: https://orcid.org/0000-0003-3626-0990

Correspondence: *Renee C. B. Manworren, PhD, APRN, PCNS-BC, AP-PMN, FAAN, Ann & Robert H. Lurie Children's Hospital of Chicago, 225 E Chicago Ave, Box 101, Chicago, IL 60611 (Renee.Manworren@uta.edu).*

Copyright © 2024 by The National Association of Neonatal Nurses.

DOI: 10.1097/ANC.0000000000001145

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.

The purpose of this study was to engage experienced NICU nurses in the development and validation of a ML model that would classify neonatal pain based on facial expressions associated with brain-based evidence of pain.

LITERATURE REVIEW

The only validated observational pain assessment tool that codes facial actions associated with brain-based evidence of pain is the Neonatal Facial Coding System (NFCS).[14,15,20-22] NFCS facial actions of neonatal acute pain include: (1) eyebrows lowered and drawn together to form a vertical furrow; (2) tightly closed eyes; (3) deepened nasolabial furrow; (4) open, (5) vertical, and (6) horizontal mouth stretch; (7) taut tongue; (8) chin quiver; (9) lip pursing; and (10) in preterm infants, tongue protrusion (in full-term infants, this is a no pain response). Coded as occurring or not occurring, NFCS is a 0 to 10 scale in preterm neonates and a 0 to 9 scale in term neonates.[5,19-21] Construct validity of NFCS had been established previously by the scale's ability to discriminate needle pain from touch,[21,22] pharmacologic treatments,[23-25] and postoperative pain assessments.[26] Despite the sensitivity and specificity of these facial actions as indicators of pain, NFCS has poor clinical utility with less than ideal (0.67) interrater reliability for real-time bedside NFCS assessments.[5,22]

Facial landmark detection is a foundational task for computer vision facial action unit detection. Objective facial action coding is time and labor-intensive.[3] In their neonatal pain detection work, Brahnam et al[27-29] used the Gaussian of Local Descriptors approach to extract facial features. This is a mathematically complicated and time-consuming 4-step process that involves Scale-Invariant Feature Transform descriptors and is computed based on the histogram of the gradient, making it computationally heavy.

Different ML models have been used for pain classification; for example, K-Nearest Neighbors was used to classify pain[30,31] and Support Vector Machines (SVM) was used to perform pain classification on Brahnam's iCOPEvid (infant Classification Of Pain Expression video) Database.[27,28] Salekin et al[31] used a DL method, Recurrent Neural Network (RNN), to model the temporal pattern of acute postoperative pain. RNNs can capture sequential information as they perform the same task for every element of a sequence, with the output being dependent on the previous outputs. Nonetheless, RNN suffers from the vanishing gradient and exploding gradient problem and cannot process considerably long sequences. Long Short-Term Memory, a variant of RNN, can address this issue by passing information through a mechanism known as cell states.

In a sequence of frames, not all frames corresponded to pain. Brahnam et al[29] validated their neonatal pain detection model based on assessments by 185 college students with no appreciable experience in neonatal pain assessment or management. We determined that having a frame-level, nurses-in-the-loop, ground truth to train a neonatal pain detection model would improve model performance. This level of expertise and granularity in data labeling is necessary to ensure clinician trust in automated clinical decision support solutions for pain detection.[32-34]

The goal of this study was to capture and transform NICU nurses' labeling of pain assessment data to train a supervised, accurate, unbiased, and precise pain classification ML model. Although NICU nurses' expertise in pain assessment was leveraged to develop the supervised ML model, we predict that accuracy and precision of the ideal ML model will exceed the abilities of these same nurses. The resulting ML model will inform future development of a continuous, automated AI-empowered, nurse-in-the-loop, computational Pain Recognition Automated Monitoring System (PRAMS) as an efficient clinical decision support software solution for neonatal pain assessment, pain treatment stratification, and pain treatment evaluation.

What This Study Adds

- A machine learning (ML) model based on frame-level, nurse-informed, ground truth data exceed that of neonatal intensive care unit (NICU) nurses for the classification of neonatal pain with 98% accuracy, 97.7% precision, and 98.5% recall.
- Nurses have a high interrater variability using current observational pain assessment tools; however, as subject matter experts, NICU nurses can label training data to improve pain classification with ML models.
- ML models using NICU nurse-labeled data can be used to develop a supervised, continuous, automated Pain Recognition Automated Monitoring System (PRAMS) to support nurse evaluation of neonatal pain in the NICU.
- PRAMS will alert healthcare professionals to neonatal pain with greater precision for more timely management.

METHODS

In this retrospective validation study, a human-centered design for Embedded Machine Learning Solution approach and NFCS were used. A key evaluation metric for a trustworthy AI model of pain assessment is the model's ability to both reflect nurses' expertise and exceed nurses' ability to consistently identify neonatal pain experiences. In our effort to develop an automated system to continuously monitor neonates' facial actions for pain, the human-centered design for Embedded Machine Learning Solution approach addressed 2

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

methodological challenges: the extensive time required for labeling of data with the NFCS[3,29] and the need to evaluate the model against nurses' expert assessment of neonatal pain.

Sample: iCOPEvid Neonatal Pain Video Database

The iCOPEvid database was obtained with permission and used for this study.[29] The iCOPEvid data set contains 20-second video sequences of neonatal facial expressions of 49 healthy term neonates (26 male, 23 female; 41 White, 1 African American, 1 Korean, 2 Hispanic, 4 interracial; 34-70 hours of age) undergoing a heel lance procedure. Videos were obtained in the following sequence: Movement I: transport from one crib to another, the state of the neonate was noted as either crying or resting; Resting I: the neonate was left undisturbed; Movement II: physically disturbing the neonate; Resting II: the neonate was again left undisturbed; Friction Stimulus: vigorous friction with an alcohol swab was applied to the external lateral surface of the heel; Resting III: the neonate was again left undisturbed; Pain Stimulus: the heel was punctured for a routine blood test and the heel was squeezed. This sequence resulted in 49 pain videos and 185 no pain videos; of these, we received a total of 175 videos, 49 pain and 126 no pain videos, for our study. Videos collected during rest periods included neonates who were crying, awake, or asleep. The iCOPEvid database included neonates who were swaddled (n = 16) and not swaddled (n = 33), as flailing hands and legs might occlude facial expressions.

Frames were extracted from every half second of the pain videos and from every second of the no pain videos to account for sample imbalance. Of the extracted image frames, 70% were used in the study, as the remaining 30% were blurry from rapid neonatal movement. The final sample included 1918 frames from 120 videos, with 762 frames from 34 pain videos and 1156 frames from 86 no pain videos.

To develop the ML classification model, the final sample of pain and no pain video sequences were split into training and testing (validation) data sets at a 7:3 frame ratio. This is a standard procedure for randomly selecting data from a common data set for training and testing. The training data set was used to train the ML models, while the testing data set was used to evaluate ML model performance. Testing the ML models on new data provides a more accurate assessment of the ML classification model performance.

Instrument

Good reliability (>0.80) had been demonstrated when NFCS was used to code slow motion and stop frame video in previous studies,[20,23-26] but coding took 60 minutes for every 20 seconds of video.[22-26] In this study, nurses coded each NFCS facial action as occurring or not occurring. Because chin quiver cannot be assessed using a frame-by-frame approach, this item was removed, bringing the number of NFCS items to 8 for term neonates in this study. All 8 facial actions were coded by nurses for all video frames in this study.

Participants

Six nurse volunteers, with a mean of 18.7 years of NICU nursing experience (ranging from 5 to 42 years) labeled data for this study. These nurses worked in a 64-bed level IV NICU, part of a 364-bed, free-standing, university-affiliated, not-for-profit urban children's hospital in Illinois that cares for neonates with complex medical needs. The neonatal pain assessment standard at this hospital had recently changed from the Neonatal Infant Pain Scale to the Neonatal Pain, Agitation, and Sedation Scale through the efforts of these same NICU nurses.[35] Thus, these nurses had demonstrated sustained interrater reliability and a commitment to advancing assessment and management strategies that are sensitive to neonates' pain care needs.

ML Pain Classification Model Development

Human to artificial intelligence (H2AI), an intuitive software solution developed by our research team to maximize neonatal nurses' experience, engagement, and productivity, was applied to PRAMS development. H2AI facilitates nurses' labeling of ML model training data and captures labels at the lowest level of granularity. A pretrained model, MobileFaceNet, was integrated into H2AI for facial detection and landmarking, foundational tasks for NFCS detection.[36]

To optimize the 6 experienced NICU nurses' interrater reliability in this study, the nurses were trained in the 4 H2AI workflows. First, the practice workflow allows nurses to become familiar with H2AI features and tasks. With a neonatal nurse scientist, NICU nurses scored 5 fixed frames to practice before starting the training workflow. The same neonatal nurse scientist then labeled an additional 5 random frames in parallel with each study nurse during the training workflow. The neonatal nurse scientist and study nurse then met to reconcile any labeling differences. If 88% agreement on NFCS (7 of 8 NFCS) and agreement on binary pain classification were achieved, and agreement on pain intensity scores were ±10 points before meeting, the nurse was "passed" to the next workflow; however, if these thresholds were not attained, parallel labeling continued until the interrater reliability threshold was reached. Labeling, the third workflow, was identical to the training workflow, except the data labels for NFCS generated by the study nurses were

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

stored and used to train the pain classification ML models. To ensure the study nurses met interrater reliability thresholds throughout labeling, nurse scientists randomly labeled up to 10% of videos labeled by each study nurse. The final review workflow gave nurse scientists a real-time dashboard of each study nurses' time labeling, labeling progress (Figure 1), and interrater reliability (Figure 2).

Each video was randomly assigned to the 6 study nurses, and each video frame was labeled by at least 2 study nurses. Nurses were blinded to whether the frames were from pain (heel lance) or no pain (rest, movement, or friction) videos; nurses were also blinded to data labeling by other nurses. The data labeling workflow included 6 tasks. First, nurses labeled each video frame to indicate if the 8 individual NFCS items occurred or did not occur. Second, nurses rated their perception of the neonate's pain intensity on a Visual Analog Scale of 0 to 100 (0 = no pain, 100 = worst possible pain). Third, nurses identified facial landmarks, and fourth, nurses identified when facial landmarks were occluded by neonates' hands or blankets. Fifth, nurses classified images as pain or no pain at the frame level, and sixth, they classified images as pain or no pain at the video level.

Data generated in H2AI for labeling NFCS, facial landmarks and occlusions, and pain classification at the frame and video levels were then used to train supervised computer vision models to classify pain. Features extracted from H2AI were fed into following 3 ML models: Logistic Regression, Random Forest, and SVM, to determine the best performing model. All data labeled by nurses, including pain intensity, were stored and interrater reliability was calculated (Figure 2).

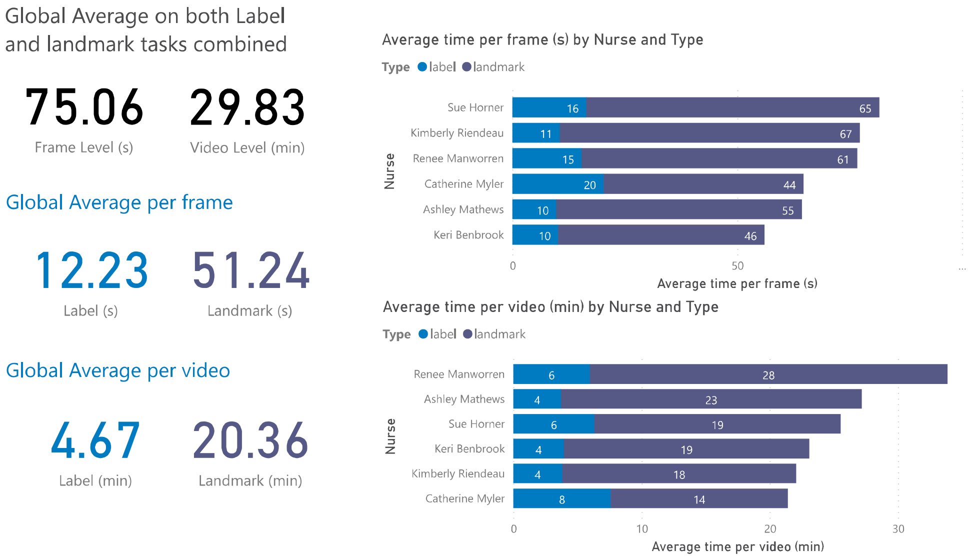
Statistical Analysis

Nurses' interrater reliability was calculated using 2 methods. During training, interrater reliability was calculated by dividing the number of instances of agreement by the total number of ratings, then multiplying by 100 to provide interrater reliability as a percentage. During data labeling, Cohen κ was calculated as the measure of agreement. Cohen κ is an estimate of the proportion of agreement between raters that is better than would occur by chance. Kappa values above 0.8 are considered very good agreement.[37]

NICU nurses' NFCS coding and pain correlations were plotted to identify key model features (Figure 3); these were compared with feature importance plots, which are provided by the model output, and features were ranked by their significance in model classification (Figure 4). Then, results of pain classification using the ML models were evaluated for accuracy (correct classifications/total classifications), precision (true positives/true and false positives), and recall (true positives/true positives and false negatives). Finally, the area under the receiver operating characteristic curve was calculated and compared for nurses and the ML models. The receiver operating characteristic curve is a graph of model recall versus the false positive rate at all

FIGURE 1

Timing Summary

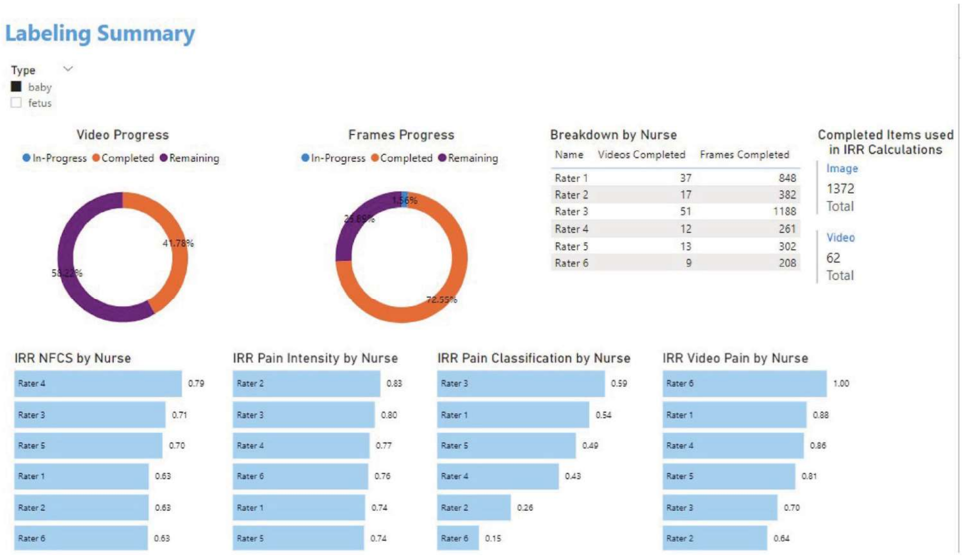


Dashboard for labeling progress and time to complete data labeling tasks.

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

FIGURE 2



Breakdown by Nurse

Name	Videos Completed	Frames Completed
Rater 1	37	848
Rater 2	17	382
Rater 3	51	1188
Rater 4	12	261
Rater 5	13	302
Rater 6	9	208

Completed Items used in IRR Calculations

Image
1372
Total

Video
62
Total

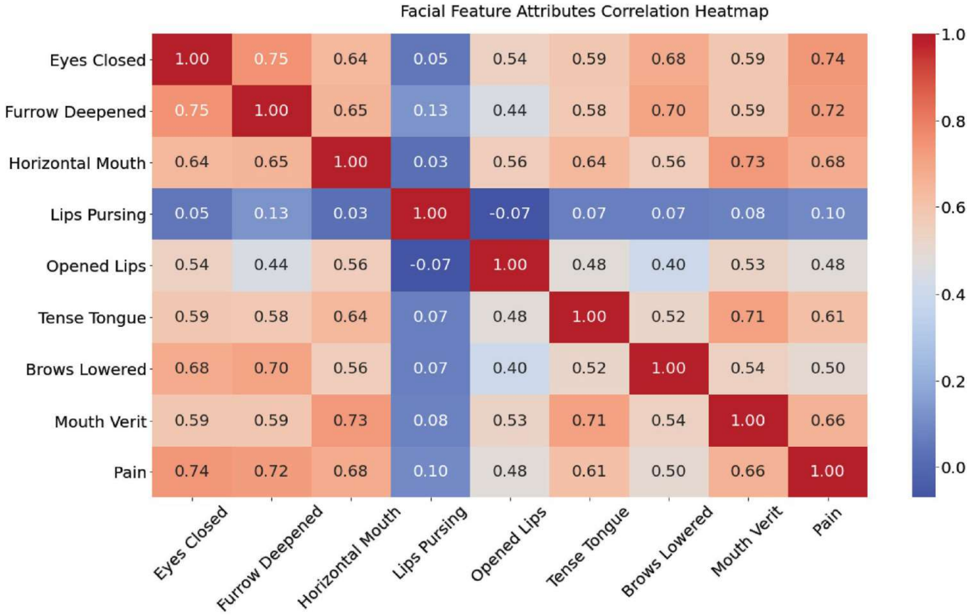
IRR dashboard for the labeling process. IRR indicates interrater reliability; NFCS, Neonatal Facial Coding System.

classification thresholds; the area under the receiver operating characteristic curve is therefore an aggregate measure of performance for a binary classification system. Area under the receiver operating characteristic curve values above 0.9 are considered outstanding, 0.8 to 0.89 is excellent, 0.7 to 0.79 is acceptable, and less than 0.69 is considered poor.

RESULTS

With 71.5% of the frames of 51.7% of the videos labeled, nurses weighted mean interrater reliability was 68% (63%-79%) for NFCS tasks, 77.7% (74%-83%) for video pain intensity, 48.6% (15%-59%) for frame pain classification, and 78.4%

FIGURE 3

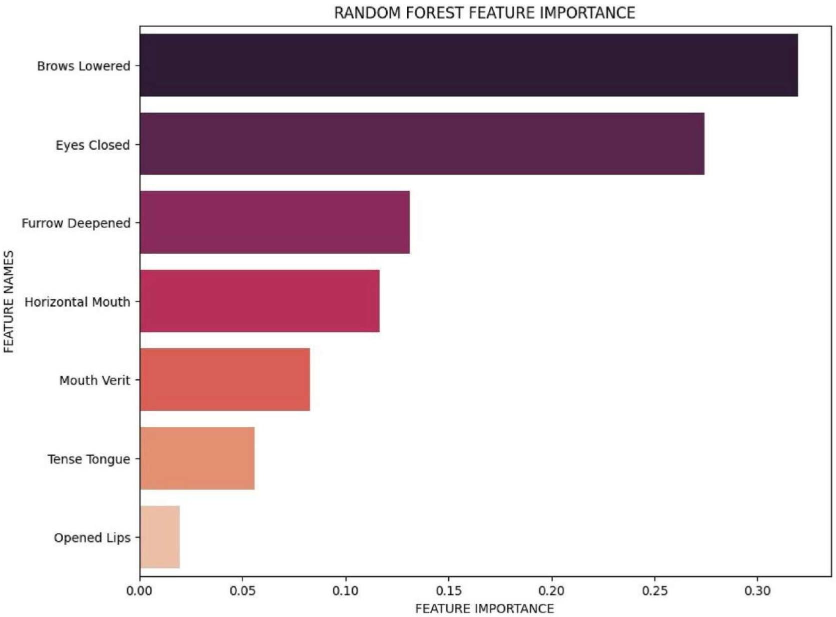


	Eyes Closed	Furrow Deepened	Horizontal Mouth	Lips Pursing	Opened Lips	Tense Tongue	Brows Lowered	Mouth Verit	Pain
Eyes Closed	1.00	0.75	0.64	0.05	0.54	0.59	0.68	0.59	0.74
Furrow Deepened	0.75	1.00	0.65	0.13	0.44	0.58	0.70	0.59	0.72
Horizontal Mouth	0.64	0.65	1.00	0.03	0.56	0.64	0.56	0.73	0.68
Lips Pursing	0.05	0.13	0.03	1.00	-0.07	0.07	0.07	0.08	0.10
Opened Lips	0.54	0.44	0.56	-0.07	1.00	0.48	0.40	0.53	0.48
Tense Tongue	0.59	0.58	0.64	0.07	0.48	1.00	0.52	0.71	0.61
Brows Lowered	0.68	0.70	0.56	0.07	0.40	0.52	1.00	0.54	0.50
Mouth Verit	0.59	0.59	0.73	0.08	0.53	0.71	0.54	1.00	0.66
Pain	0.74	0.72	0.68	0.10	0.48	0.61	0.50	0.66	1.00

Nurses labeling of Neonatal Facial Coding System feature correlation plot.

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.

FIGURE 4



Supervised machine learning champion model: random forest feature plot.

(64%-100%) for video pain classification (Figure 2), with an area under the receiver operating characteristic curve of 0.68. Of 563 frames from 16 pain videos, nurses agreed and classified 463 frames (82%) as pain and 68 frames (12%) as no pain; nurses disagreed in their classification of 32 pain frames (6%). All nurses classified all videos depicting the "pain" condition as pain except for one classification of one pain video.

Of the 46 no pain videos, 9 were collected during movement, 10 during friction, and 27 during rest periods. Of 809 frames from these 46 no pain videos, nurses agreed and classified 254 of 462 (55%) rest frames, 84 of 177 (47%) movement frames, and 65 of 170 (38%) friction frames correctly as no pain. Nurses agreed but incorrectly classified 161 (35%) rest frames, 72 (41%) movement frames, and 84 (50%) friction frames as pain. Nurses disagreed in their classification of 47 rest frames (10%), 21 (12%) movement frames, and 21 (12%) friction frames.

Consistent with previous research, all NFCS items except "lips pursing" were highly correlated with nurses' classification of pain (Figure 3).[21] Thus, using the 7 remaining NFCS features, 3 ML models (Logistic Regression, Random Forest, and SVM) were trained on the nurse-labeled data to classify pain. The best performing model was a tree-based Random Forest model; "brows lowered" was the most important feature for pain classification (Figure 4). This initial minimal viable product ML model proved to be reliable, with 95.5% accuracy, 95.4% precision, and 97.8% recall (Table 1). When we included all data labeled during the nurse training and practice workflows, the results of pain classification for all ML models improved. The Random Forest model again had the best classification results, with 98% accuracy, 97.7% precision, 98.5% recall, and area under the receiver operating characteristic curve of 0.98 (Table 1). The supervised ML model provided outstanding discrimination and the ML model area under the receiver operating characteristic curve far exceeded that of the NICU nurses.

DISCUSSION

The goal of developing a supervised ML model to classify neonatal acute pain with high accuracy, precision, and recall was achieved. The model was trained using NICU nurse labeling data, and nurses' interrater reliability was fair at the frame level but very good at the video level. Nurses weighted mean interrater reliability for labeling was lower than the interrater reliability required during training. This was expected since, unlike kappa, the percent agreement calculated during training did not take chance agreement into account.[37]

Despite efforts to maximize NICU nurses' interrater reliability, nurses assessed NFCS and pain differently, demonstrating the interrater variability in pain assessment of nurses in clinical practice.[22,35] Nurses agreed and classified 94% of videos of pain and more frames depicting pain (82%) than they did frames of no pain

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

TABLE 1. Pain Classification Results Using the Machine Learning Models

Model	Logistic Regression	SVM	Random Forest
Including only data from videos blindly labeled by at least 2 nurses			
Accuracy	0.930	0.875	0.955
Precision	0.912	0.865	0.954
Recall	0.982	0.956	0.978
AUC	0.904	0.842	0.946
Including data from practice and training frames and videos blindly labeled by at least 2 nurses			
Accuracy	0.972	0.976	0.98
Precision	0.970	0.970	0.977
Recall	0.977	0.985	0.985
AUC	0.972	0.976	0.98

Abbreviations: AUC, area under the receiver operating characteristic curve; SVM, Support Vector Machine.

activities (50%). Yet, the nurses' area under the receiver operating characteristic curve of 0.68 (0.59-0.74) suggests poor to acceptable discrimination and was only slightly better than a previous study using this same data set and college students' classifications area under the receiver operating characteristic curve (95% confidence interval, 0.665-0.677).[30]

Although other pain scales that include facial actions have been validated against nurses' diagnosis of pain, nurses' interrater reliability, and area under the receiver operating characteristic curve in this study provides additional evidence of the variability and inconsistency of diagnosing pain, even among experienced NICU nurses.[14-22,35] Notably, the facial expressions included in other popular neonatal pain assessment scales are not associated with brain-based evidence of pain.[15] In addition, several of these commonly used multidimensional pain scales include physiological measures.[5,15-19] Despite extensive research investigating changes in heart rate, heart rate variability, and other physiological measures, none have been found to be sensitive or specific for pain.[5,15]

Nurses' area under the receiver operating characteristic curves for each no pain stimulus was poor, at 0.62 for movement, 0.49 for rest, and 0.44 for friction videos. This may reflect nurses' heightened vigilance for identifying pain given this data labeling task. These classification data also support the data set developers' conclusion that this is a challenging data set to assess for pain.[29] Based on the neonates' age, the video was not of their first heel stick. Therefore, it is also possible that our experienced NICU nurses identified the neonates' anticipation of needle pain from facial expression changes during the friction stimulus.

In contrast to nurses, the best performing supervised ML model identified at the frame level more frames in pain videos as pain and more frames in no pain videos as not depicting pain. These data will be important for determining the time from when the model identifies pain until the PRAMS software alerts the nurse to the neonates' pain. This latency is critical to ensuring alerts are not discounted by nurses as false alarms.[2] The challenge, ultimately, is to continuously monitor hospitalized neonates for acute pain, such as surgical pain, and to evaluate the need to treat and the duration and effectiveness of pain treatments.[1,2,5,11,38,39] To date, only Salekin et al[31] have developed a novel multimodal spatiotemporal approach for assessing neonatal postoperative pain using visual and vocal signals to achieve 79% accuracy.

Because still frames were used, chin quivering was not labeled, and since neonates in the data set were full-term, tongue protrusion was not identified as a pain classifier. In addition, we found that lips pursing did not correlate with nurses' classification of pain. This finding is consistent with previous research that has discounted the value of lip pursing for pain classification.[26]

The association of eyes closed, horizontal mouth, vertical mouth, and open lips with pain was expected and was reflected in the feature correlation plot (Figure 3) and feature importance in the supervised ML model (Figure 4). However, brows lowered, eyes closed, and nasolabial furrow deepened had higher feature importance in the supervised ML model than horizontal mouth, vertical mouth, tense tongue, or opened lips. This was surprising because brow lowering and nasolabial furrow deepening is difficult to observe in real time. The feature importance of brows lowered and nasolabial furrow deepened is significant because it suggests that we can extend this supervised ML solution to earlier gestational ages, since microfacial expressions such as subtle brow changes require less energy expenditure. These feature importance findings also suggest that the supervised ML solution can be extended to the intubated neonate, whose mouth is occluded in

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.

continuous video surveillance by respiratory support equipment.

LIMITATIONS

The iCOPEvid neonatal pain database is small and lacks racial/ethnic, postconceptual age, and illness severity diversity.[18,28,29,40] There is tremendous variability in expression of pain across painful conditions and gestational ages.[2,5,18,20,21] Therefore, algorithms and models based on the homogenous iCOPEvid data set will be inaccurate when used in real-world samples of hospitalized neonates with pain.

The iCOPEvid database contained video; like other researchers, we converted these data to frames for data labeling granularity.[28-30] However, by using frame-level data, the resulting supervised ML model may fail to capture the dynamic patterns of facial expressions, which may provide important features for discriminating pain. Furthermore, pain in this video data set came from predictable needle procedures.[29]

Nurses' interrater reliability of 77.7% (74%-83%) for video pain intensity was very good; we gathered these data to replicate the work of others. Nevertheless, NFCS and other observational pain assessment methods are not measures of pain intensity.[5] Simply counting number of behaviors, responses, or robustness of responses does not imply pain severity or guide treatment. Nurses' interrater reliability for video pain intensity may instead reflect their knowledge that the videos were uniform in their inclusion of a heel lance procedure, a single source of pain with little variability in intensity, unlike surgical or disease-related pain. Nurses' pain intensity ratings did not help distinguish pain videos from videos depicting neonates experiencing friction, movement, or crying during periods of rest. An efficient clinical decision support software solution for neonatal pain assessment from a variety of painful conditions, pain treatment stratification, and pain treatment evaluation is still needed.

CLINICAL IMPLICATIONS

Trustworthy AI models of pain assessment must reflect nurses' expertise and exceed nurses' interrater reliability and area under the receiver operating characteristic curve for consistent identification of neonatal pain experiences. In this study, H2AI facilitated data labeling by NICU nurses as subject matter experts in neonatal care, and the 6 experienced NICU nurses demonstrated good to very good interrater reliability. Interrater reliability was calculated to ensure reliability and consistency of H2AI labeled data. These data were then used to train a supervised, accurate, and precise ML pain classification model. Our high-quality nurse-labeled data in the training data set are reflected in the superior performance of our model with the testing data set.

With 98% accuracy, 97.7% precision, 98.5% recall, and area under the receiver operating characteristic curve of 0.98, the supervised ML pain classification model far exceeded NICU nurses' area under the receiver operating characteristic curve. Of course, when providing direct patient care, NICU nurses' decisions to assess pain are based on context, including patients' conditions and the nurses' availability. When taken out of context, as demonstrated by nurses' variability with the friction stimulus video interrater reliability, nurses may under- or overdiagnose pain.

We are the only scientific group led by nurses that is working to develop a continuous, automated AI-empowered, nurse-in-the-loop, computational PRAMS as an efficient clinical decision support software solution for neonatal pain assessment, pain treatment stratification, and pain treatment evaluation. Despite our ML models' superior area under the receiver operating characteristic curve compared with that of our expert nurses and our use of brain-based facial actions of pain, nurses remain important for evaluating the context of neonates' facial actions, with the PRAMS as a tool to support nurses' decision making. We also believe the advantage of this ML model is yet to be realized.

FUTURE RESEARCH

This supervised ML model will inform future development of a continuous, automated AI-empowered, nurse-in-the-loop, computational PRAMS. This particular ML model provides the potential for continuous assessment of brain-based facial actions associated with pain. However, a gestational-, gender-, racial-, ethnic-, and condition-diverse data set is needed for further PRAMS development and testing. Recent federal data sharing requirements may facilitate access to robust and diverse video and clinical data sets that will accelerate further development of models like ours to promote equity in neonatal pain care.

We are moving forward with a clinical trial of continuous monitoring of perioperative neonates with our ML model. We will explore if there is added value to including physiologic features in the model, such as heart rate, heart rate variability, or pupillometry. An important specific aim of our clinical trial is to determine the latency of alert, specifically, how many frames of sustained brow lowering, eyes closed, and deepened nasal labial furrowing (the 3 most important features in the current ML model) should occur before the nurse is alerted to assess the neonate for pain. Direct care NICU nurses

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

Summary of Recommendations for Practice and Research

What we know:	• Inability to self-report pain makes neonates vulnerable to underrecognition and both undertreatment and overtreatment of pain. • Current pain assessment practices are discontinuous, inconsistent, and highly dependent on nurses' availability. • The Neonatal Facial Coding System (NFCS) is the only validated observational pain assessment tool that codes facial actions associated with brain-based evidence of pain, but is associated with high clinical variability.
What needs to be studied:	• Advances in computer software solutions can assist in identifying NFCS indicators of pain. • Frame-level, nurse-informed, ground truth data are needed to train a neonatal pain detection machine learning (ML) model and ensure nurses' trust in automated decision support solutions for pain detection.
What we can do today:	• As subject matter experts, neonatal intensive care unit (NICU) nurses' expertise in pain assessment can be leveraged to develop supervised ML models that perform with high accuracy and precision. • Nurses' involvement in clinical decision software development is critical for ensuring optimal and unbiased patient-centered care from artificial intelligence.

will be critical for evaluating the feasibility of PRAMS and innovating neonatal pain care.

Acknowledgments

The authors thank Keri Benbrook, BSN, RN, and Ashley Entler, BSN, RN, NICU nurses at Ann & Robert H. Lurie Children's Hospital of Chicago for their extensive coding of data to train the ML model. The authors thank Rajesh Inbasekaran and Balakumaran Manoharan for their guidance in overall software solution architecture and application design and development, and Rahul Dhanasiri for his help in building the MobileFaceNet model. Finally, the authors acknowledge Jun Hua Wong, Kosha Soni, Trishla Mishra, and Stacey Tobin for their assistance in literature review, manuscript preparation, and editing.

References

1. Anand KJS, Eriksson M, Boyle EM, et al. Assessment of continuous pain in newborns admitted to NICUs in 18 European countries. *Acta Paediatr.* 2017;106(8):1248-1259.
2. Manworren RCB, Atabek A. Time from pain assessment to pain intervention. *J Nurs Adm.* 2021;51(7/8):389-394.
3. Zamzmi G, Kasturi R, Goldgof D, Zhi R, Ashmeade T, Sun Y. A review of automated pain assessment in infants: features, classification tasks, and databases. *IEEE Rev Biomed Eng.* 2018;11:77-96.
4. Zhi R, Zamzmi GZD, Goldgof D, Ashmeade T, Sun Y. Automatic infants' pain assessment by dynamic facial representation: effects of profile view, gestational age, gender, and race. *J Clin Med.* 2018;7(7):173.
5. Herr K, Coyne PJ, Ely E, Gelinas C, Manworren RCB. Pain assessment in the patient unable to self-report: clinical practice recommendations in support of the ASPMN 2019 position statement. *Pain Manag Nurs.* 2019;20(5):404-417.
6. Carbajal R, Rousset A, Danan C, et al. Epidemiology and treatment of painful procedures in neonates in intensive care units. *JAMA.* 2008;300(1):60-70.
7. Johnston C, Barrington KJ, Taddio A, Carbajal R, Fillion F. Pain in Canadian NICUs: have we improved over the last 12 years? *Clin J Pain.* 2011;27(3):225-232.
8. Simons SH, van Dijk M, Anand KS, Roofthooft D, van Lingen RA, Tibboel D. Do we still hurt newborn babies? A prospective study of procedural pain and analgesia in neonates. *Arch Pediatr Adolesc Med.* 2003;157(11):1058-1064.
9. Manworren RCB. We do still hurt babies. *J Perinat Neonatal Nurs.* 2017;31(2):89-90.
10. Eriksson M, Campbell-Yeo M. Assessment of pain in newborn infants. *Semin Fetal Neonatal Med.* 2019;24(4):101003.
11. Ilhan E, Galea C, Pacey V, et al. Trajectories of post-surgical pain in infants admitted to neonatal intensive care. *Eur J Pain.* 2020;24(9):1822-1830.
12. Walker SM. Biological and neurodevelopmental implications of neonatal pain. *Clin Perinatol.* 2013;40(3):471-491.
13. Walker SM. Long-term effects of neonatal pain. *Semin Fetal Neonatal Med.* 2019;24(4):101005.
14. Chang J, Versloot J, Fashler SR, McCrystal KN, Craig KD. Pain assessment in children: validity of facial expression items in observational pain scales. *Clin J Pain.* 2015;31(3):189-197.
15. Relland LM, Gehred A, Maitre NL. Behavioral and physiological signs for pain assessment in preterm and term neonates during a nociception-specific response: a systematic review. *Pediatr Neurol.* 2019;90:13-23.
16. Andersen RD, Langius-Eklof A, Nakstad B, Bernklev T, Jylli L. The measurement properties of pediatric observational pain scales: a systematic review of reviews. *Int J Nurs Stud.* 2017;73:93-101.
17. Desai A, Aucott S, Frank K, Silbert-Flagg J. Comparing N-PASS and NIPS: improving pain measurement in the neonate. *Adv Neonatal Care.* 2018;18(4):260-266.
18. Stevens B, Johnston C, Taddio A, Gibbins S, Yamada J. The premature infant pain profile: evaluation 13 years after development. *Clin J Pain.* 2010;26(9):813-830.
19. Magesti BN, Christoffel MM, Fernandes AM, da Silva Dias C, da Silva Melo A, Possi JCS. Facial expression as an indicator of neonatal pain in randomized clinical trials: an integrative review. *J Neonatal Nurs.* 2023;29(3):260-266.
20. Grunau RVE, Craig KD. Pain expression in neonates: facial action and cry. *Pain.* 1987;28(3):395-410.
21. Grunau R, Craig K. Facial activity as a measure of neonatal pain expression. In: Tyler DC, Krane EJ, eds. *Advances in Pain Research and Therapy.* Vol. 15. New York, NY: Raven Press; 1990:147-155.
22. Grunau RE, Oberlander T, Holsti L, Whitfield MF. Bedside application of the Neonatal Facial Coding System in pain assessment of premature neonates. *Pain.* 1998;76(3):277-286.
23. Benini F, Johnston CC, Faucher D, Aranda JV. Topical anesthesia during circumcision in newborn infants. *JAMA.* 1993;270(7):850-853.
24. Scott CS, Riggs KW, Ling E, Grunau RVE, Craig K, Solimano A. Morphine pharmacokinetics and pain assessment in premature neonates. *Pediatr Res.* 1994;35:254(A).
25. Taddio A, Stevens B, Craig K, et al. Efficacy and safety of lidocaine-prilocaine cream for pain during circumcision. *N Eng J Med.* 1997;336(17):1197-1201.
26. Peters JWB, Koot HM, Grunau RE, et al. Neonatal Facial Coding System for assessing postoperative pain in infants: item reduction is valid and feasible. *Clin J Pain.* 2003;19(6):353-363.
27. Brahnam S, Chuang CF, Shih FY, Slack MR. SVM classification of neonatal facial images of pain. In: Block I, Petrosino A, Tettamanzi AGB, eds; *Fuzzy Logic and Applications (revised selected papers from the 6th International Workshop, WIKF 2005, Crema, Italy, September 15-17, 2005).* 2006:121-128.
28. Brahnam S, Chuang CF, Sexton RS, Shih FY. Machine assessment of neonatal facial expressions of acute pain. *Decis Support Syst.* 2007;43(4):1242-1254.
29. Brahnam S, Nanni L, McMurtrey S, et al. Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from Gaussian of Local Descriptors. *Appl Comput Informat.* 2020;19:122-143.
30. Zamzmi G, Goldgof D, Kasturi R, Sun Y. Toward ubiquitous assessment of neonates' health condition. UbiComp '18: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (pp. 952-955); 2018. doi:org/10.1145/3267305.3267688.
31. Salekin MS, Zamzmi G, Goldgof D, Kasturi R, Ho T, Sun Y. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Comput Biol Med.* 2021;129:104150.
32. Kaduwela NA, Dadar P, Horner S, Manworren RCB. Human centered design for embedded machine learning data labeling software solution: H2AI. *Int J Med Inform.* 2023.
33. Garcia MB, Pilueta NU, Jardiniano MF. VITAL APP: Development and user acceptability of an IoT based patient monitoring device for synchronous measurements of vital signs. IEEE, 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM) 2019;1-6.
34. Hogail AA. Improving IoT technology adoption through improving consumer trust. *Technologies.* 2018;6(3):64-81.

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.

35. Benbrook K, Manworren RCB, Zuravel R, et al. Agreement of Neonatal Pain, Agitation, and Sedation Scale (N-PASS) with NICU nurses' assessments. *Adv Neonatal Care*. 2023;23(2):173-181. doi:10.1097/ANC.0000000000000968.
36. Deng J, Guo J, Xue N, Zafeiriou S, et al. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans Pattern Anal Mach Intell*. 2022;44(10):1-1.
37. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. Hoboken, NJ: John Wiley; 2003.
38. Sposito NPB, Rossato LM, Bueno M, Kimura AF, Costa T, Guedes DMB. Assessment and management of pain in newborns hospitalized in a neonatal intensive care unit: a cross-sectional study. *Rev Lat Am Enfermagem*. 2017;25:e2931.
39. Riddell RP, Flora DB, Stevens SA, et al. Variability in infant acute pain responding meaningfully obscured by averaging pain responses. *Pain*. 2013;154(5):714-721.
40. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. *Proc Machine Learn Res*. 2018;81:1-15.

Downloaded from http://journals.lww.com/advancesinneonatalcare by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC4/OAVpDDa8K2+Ya6H515kE= on 05/22/2024

Copyright © 2024 National Association of Neonatal Nurses. Unauthorized reproduction of this article is prohibited.