IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 January 2024, accepted 18 March 2024, date of publication 1 April 2024, date of current version 10 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3383789

APPLIED RESEARCH

Accurate Neonatal Face Detection for Improved Pain Classification in the Challenging NICU Setting

JACQUELINE HAUSMANN[1], (Student Member, IEEE), MD SIRAJUS SALEKIN[1], (Member, IEEE), GHADA ZAMZMI[1], (Member, IEEE), PETER R. MOUTON[2], (Member, IEEE), STEPHANIE PRESCOTT[3], THAO HO[4], YU SUN[1], (Senior Member, IEEE), AND DMITRY GOLDGOF[1], (Fellow, IEEE)

[1]Department of Computer Science and Engineering, College of Engineering, University of South Florida, Tampa, FL 33620, USA
[2]SRC Biosciences, Tampa, FL 33606, USA
[3]College of Nursing, USF Health, University of South Florida, Tampa, FL 33620, USA
[4]Department of Pediatrics, College of Medicine, University of South Florida, Tampa, FL 33606, USA

Corresponding author: Jacqueline Hausmann (hausmannj@usf.edu)

This work was supported in part by the National Institutes of Health under Grant NIH R21NR018756 and Grant R41HD109086, and in part by the University of South Florida Nexus Initiative (UNI) Grant.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of South Florida Ethics Internal Review Board under IRB NO. #Pro00014318.

ABSTRACT There is a tendency for object detection systems using off-the-shelf algorithms to fail when deployed in complex scenes. The present work describes a case for detecting facial expression in post-surgical neonates (newborns) as a modality for predicting and classifying severe pain in the Neonatal Intensive Care Unit (NICU). Our initial testing showed that both an off-the-shelf face detector and a machine learning algorithm trained on adult faces failed to detect facial expression of neonates in the NICU. We improved accuracy in this complex scene by training a state-of-the-art "You-Only-Look-Once" (YOLO) face detection model using the USF-MNPAD-I dataset of neonate faces. At run-time our trained YOLO model showed a difference of 8.6% mean Average Precision (mAP) and 21.2% Area under the ROC Curve (AUC) for automatic classification of neonatal pain compared with manual pain scoring by NICU nurses. Given the challenges, time and effort associated with collecting ground truth from the faces of post-surgical neonates, here we share the weights from training our YOLO model with these facial expression data. These weights can facilitate the further development of accurate strategies for detecting facial expression, which can be used to predict the time to pain onset in combination with other sensory modalities (body movements, crying frequency, vital signs). Reliable predictions of time to pain onset in turn create a therapeutic window of time wherein NICU nurses and providers can implement safe and effective strategies to mitigate severe pain in this vulnerable patient population.

INDEX TERMS Convolutional neural network, face detection, neonate, neonatal intensive care unit, pain classification, recurrent neural network.

I. INTRODUCTION

Each year a large and increasing number of newborns are admitted to Neonatal Intensive Care Units (NICUs) worldwide following life-saving or corrective surgery during the immediate post-natal period. We propose that an automatic early pain detection (EPD) system would allow nurses and providers to implement "stay ahead of the pain" strategies for assessing and reducing post-surgical pain in this highly vulnerable population [1], [2]. Currently there

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li.

is a gap in the tools used in clinical care to asses Neonatal pain [3]. One such tool which we will reference in this work is the Neonatal Pain, Agitation, and Sedation scale(NPASS) [5], a pain measurement system currently used in practice by bedside caregivers for neonatal pain assessment. Effective based on theory, NPASS has varied success in practice, due to factors such as perceived bias by bedside nurses, delayed detection, etc. Further, there is substantial support for *a priori* pain avoidance strategies which could effectively minimize short- and long-term damage from severe pain, chronic pain sensitization and powerful analgesics exposure on the developing newborn's brain and their spinal cord [4], [6], [7].

Automatic approaches for early pain detection focus on the earliest evidence of pain expression in the faces of neonates emerging from sedation in the NICU, starting with face detection within the frame of an image or video [8], [9], [10], [11]. Face detection is a long-standing field of research in computer science with state-of-the-art algorithms for automatic detection of acute and chronic [12], [13], [14] pain in adult faces. Like adults, newborns perceive the full range of pain associated with knives, (scalpels), needles and other sharp objects passing through their tissues [15], [16], [17], [18], [19]. Furthermore, neonatal pain is strongly expressed through facial expressions [20], [21], [22], [23], providing an important focal point for early pain detection. As detailed in the following section, among the challenges in neonatal pain detection is that state-of-the-art face detectors for adults perform poorly or not at all as detectors of neonate faces in the NICU. It has been shown in the research that often facial expressions are of high importance to pain classification [24]. Thus, a new class of facial detection tools is needed for improving performance accuracy as part of an overall pain/no-pain classification system for post-surgical neonates.

There are notable differences between the composition of neonate and adult human faces. In addition to smaller size, neonate faces are truncated, resulting in pinched appearing eyes and mouths. Furthermore, the neonate face is generally lacking in contextual information, such as hair on the head or face to outline features [25], [26]. Additional challenges are associated with face detection within the clinical NICU setting, a complex and uncontrolled environment with variable ambient noise and lighting conditions. Neonates often sleep with lights on and off for different intervals or in isolettes covered by thick blankets for noise protection. Other obstacles include occlusion of faces by wires, instruments and devices required for medical care and complicated background settings with soft and textured blankets leading to unclear boundary distinctions. Finally, neonates are unable to participate in their health care; unable to follow directions (e.g., stay still or don't move), which further complicates face detection with extreme pose variations; and lack the verbal fluency to rate their pain, complicating ground truth determinations.

Some current works have looked at methods to overcome the difficulties associated with facial feature extraction of the neonate. In [27] and [28], Dosso et al. utilize extracted facial features via different intensity based approaches after the neonate face region of interest has already been determined by manual annotation. Dosso et al. extend this work by [29] focusing on the neonate face detection while emphasizing how the difficulty of the setting will largely impact final detection results.

iCOPEvid [30] is one of the limited number available of publicly accessible neonatal face datasets and is similar to USF-MNPAD-I in terms of difficulties. Brahnam et al. collect the iCOPEvid in [30], using it for procedural pain classification with a Discriminative Response Map Fitting [31] approach for face detection with results of this algorithm on iCOPEvid not reported. Olmi et al [32] attempt to detect neurological dysfunctions, such as seizure episodes in the neonate in the NICU utilizing an Aggregate Channel Feature algorithm achieving an average precision recall of 0.61 ± 0.05 for neonate face detection on the iCOPEvid dataset. Grooby et al. [33] further reinforce the problem of current off-the-shelf solutions being insufficient for neonatal face detection by fine-tuning YOLOv7 on 3 publicly available datasets including the iCOPEvid [30], but achieved only 86% accuracy which our work shows will still result in decreased performance on an automated system which relies on face detection as the first pre-processing step.

In this work, we propose a neonatal face detector that attempts to overcome the aforementioned difficulties by training on model with data containing real-world conditions. With initial results showing high accuracy, we compare the performance upon multiple datasets to demonstrate the need for a generalized neonatal face detector. To this end, we emphasis that off-the-shelf solutions used for adult face detection are insufficient for use in a neonatal intensive care setting. We enforce this hypothesis by also investigating the overall impact of differing face detectors on neonatal pain/no-pain classification using a state-of-the-art spatial temporal deep learning system. Pain classification of the neonate is only one application that would benefit from the use of an accurate neonate face detector. Other applications that could benefit from a robust neonatal face detector include facial recognition and surveillance [34], [35], [36], with further extensions being any automated computing system utilizing a dataset which exemplifies large data distribution shift, i.e. many differences between examples in the same class.

Contrary to current neonatal face detectors [37], [38], the proposed detector requires little pre-processing operations, allowing real-time monitoring of pain in a clinical setting. An additional strength of our algorithm is that it can achieve real-time detection performance by detecting neonatal faces with the same or similar rate (∼15 fps) as current real-time methods [39], [40]. In practice, a detection rate of 15 fps is sufficient speed for the neonatal face detection application as well as neonatal pain assessment. This work demonstrates a

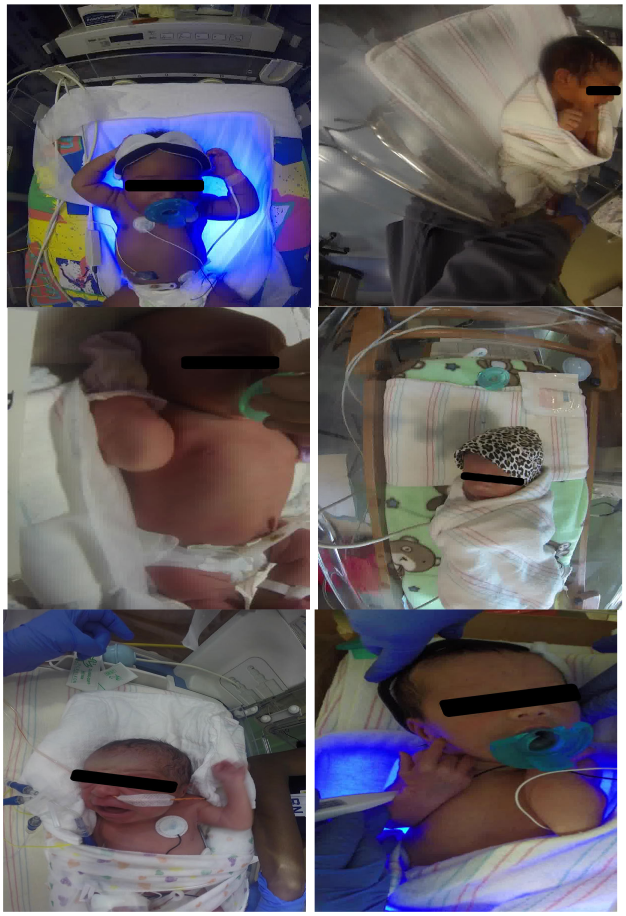

FIGURE 1. Examples of images in the USF-MNPAD-I dataset. From these images, we can notice the challenging conditions (e.g., low-light and occlusion) of this dataset.

framework which could be followed for additional automated systems.

The rest of this paper is organized as follows. Section II. presents the datasets used within this work, as well as describes the scientific methodology used upon those datasets. This includes any data preparation, evaluation protocols, and deep learning models. Section III. elaborates on the various experiments performed, including results of face detection, how face detection accuracy effects the results of an overall pain classification system, and further analysis of those experiments. Section IV. provides a discussion on the importance of the work presented here and Section V. concludes this work.

II. MATERIALS AND METHODS

A. DATASETS

This work utilizes three different datasets discussed in further detail below. These datasets are the WIDER FACE dataset [41], USF-MNPAD-I dataset [42] and the USF-MNPAD-II dataset for which data collection is still currently ongoing. All of the clinical data associated with the USF-MNPAD-I and II datasets were properly collected and monitored by the University of South Florida Ethics Internal Review Board (IRB #Pro00014318). We have chosen an adult face's in-the-wild dataset to compare with a neonate face in-the-wild as there are lacking available neonate datasets publicly available. Additionally, by both datasets being in-the-wild, they are able to capture similar difficulties such as pose variations, lighting concerns, etc which are traditional roadblocks for accurate computer vision algorithms.

1) WIDER FACE

WIDER FACE is a large scale dataset with 32,203 total images containing one or more adult faces which have been identified and labeled [41]. This dataset has a variety of depths, face postures, resolutions, backgrounds, and subjects making it ideal for training a generalized adult face detector. However, this dataset is insufficient for training a neonatal face detector as will be demonstrated in our experiments. In previous works, Salekin et al. [10], [11] used YOLOv3 face detector trained on WIDER FACE [41] for pain classification which is why we will consider this to be our baseline.

2) USF-MNPAD-I

USF-MNPAD-I [42] is a partially available multi-modal dataset, containing videos of neonates' responses to different procedural pain (36 neonates) and postoperative pain (9 neonates) stimuli. With a total of 45 subjects, this dataset will have two subsets used during our experimentation. Procedural pain can be defined as the initial and often intense pain felt directly after an unwelcome stimuli, where postoperative pain is characterized by lingering pain felt as the body heals from surgery, a more drastic negative stimuli. Figure 1 shows examples of images extracted from videos comprising USF-MNPAD-I.The ground truth pain/no-pain labels are derived from the Neonatal Pain, Agitation, and Sedation scale(NPASS) [5], which is a standard of care measurement.

USF-MNPAD-I dataset was collected in the Neonatal Intensive Care Unit (NICU) of Tampa General Hospital as a part of a collaborative study between USF's Computer Science and Engineering Department, USF Health, and Tampa General Hospital (TGH). Please refer to the project[1] for more information about accessing this dataset.

3) USF-MNPAD-II

USF-MNPAD-II is the second iteration of data collection at TGH after USF-MNPAD-I. As research progressed, postoperative pain of the neonate and its management became more important, and thus USF-MNPAD-II is comprised solely of postoperative pain subjects. We modified the data collection set-up for USF-MNPAD-II after observing initial difficulties faced when collecting videos for USF-MNPAD-I. For example, one such modification was utilizing more compact hardware to remove interference with hospital staff. In addition to physical difficulties, there are many privacy concerns associated with neonate healthcare data that

[1] https://rpal.cse.usf.edu/project_neonatal_pain/dataset.html

contributes towards the challenges of data collection of this vulnerable population. We recognize the lacking availability of public datasets and will be releasing the weights associated with the model's described in this work for benefit of the research community. The NPASS scale was used for ground truth labeling. Instead of short video segments as was done for USF-MNPAD-I, all of the data collected for USF-MNPAD-II consists of long continuous videos ranging from one to twenty hours. We then extracted smaller segments from the longer videos for experimentation. As data collection is still ongoing (60 subjects collected to date), we processed the initial 15 subjects for preliminary analysis.

B. METHOD

This section presents the proposed face detectors followed by the method used for neonatal pain expression classification. Finally, we describe our evaluation protocol as well as training and testing splits.

1) YOU-ONLY-LOOK-ONCE (YOLO) FOR FACE DETECTION

You-Only-Look-Once (YOLO) [43] is a popular detection algorithm developed by Joseph Redmon in 2012. In this study, we utilize recent versions of YOLOv3 [44], YOLOv5 [40], and YOLOv6 [46] for our neonate face detectors, all of which are open-sourced. There are many competing face detection algorithms, some of which we mentioned in the introduction. However, YOLO was chosen for this work due to its competing accuracy and superior ability in computing speed during test time which is key for a real world clinical application. While there are many additional versions of the YOLO algorithm being released, of which we will keep in consideration for future work, this work limits its use to the mentioned versions at this time. Due to only limited modifications between versions after YOLOv6, most likely minuscule improvements in speed and accuracy would have resulted for using newer algorithms and balanced this notion with time constraints required to retrain & test additional models.

YOLO is a one-stage anchor-based object detector that divides the entire image into a grid of size $S*S$, where each cell in the grid will produce B bounding boxes, each with confidence of C, *class, x, y, w,* and h. YOLOv3 [44] and later versions utilized 53 convolutional neural net layers, garnering the backbone of the architecture the name Darknet-53 (pre-trained on ImageNet [45]) to produce these B bounding boxes. YOLOv4, v5, & v6 contain an added neck and head component to the architecture [39], [40], [46]. After passing through the convolutional layers, the final reduced tensor is passed through two fully connected dense layers resulting in the final parameterized space of $S*S*(B * 5 + C)$, which is then thresholded by a confidence level via non-maximal suppression to help eliminate incorrect detections and capture correct detections.

In this work, we train a YOLOv5 and YOLOv6 face detector using the images of 36 neonates collected during procedural painful stimuli (see Section II-A.2) and a YOLOv3 face detector using the images of WIDER FACE (Section II-A.1). All face detectors are then evaluated using the images of 9 neonates collected during postoperative pain (see Section II-A.2). Additionally, we compare the highest performing face detector for USF-MNPAD-I postoperative (YOLOv6 trained on USF-MNPAD-I procedural vs YOLOv3 on WIDER FACE) by further testing on the initial subjects from USF-MNPAD-II. Finally, we then integrate each trained face detector into a neonatal pain expression classification system and evaluate the impact of each detector in said system.

2) Bi-LINEAR VGG16 & LSTM FOR PAIN CLASSIFICATION

To evaluate the impact of a robust neonate face detector on a useful automated system which can be used in the NICU, we integrate our face detector with the following system for pain expression classification [9].

Our pain expression classification network has two stages: (1) a bi-linear convolution neural network and (2) a long-short term memory (LSTM) network. The bi-linear network of this system uses two branches or streams of VGGFace [37], a derivative of the VGG16 model [47]. This model has 13 convolution blocks, composed of convolution layers with 3 × 3 kernel filter size and subsequent pooling layer. After the convolutional layers, the model has 2 dense layers each followed by a dropout layer with a relu activation function. The final dense layer has a linear activation with a size equal the number of classes (pain and no-pain). Each of these VGGFace streams (or branches) are applied to various locations of the input image in order to extract differing sets of spatial features. These separate feature vectors are then combined through the use of sum pooling resulting in a final feature vector u. This process is described mathematically in the following equations, where I is input image, L is location in the image, and F_x, F_y are the feature vectors.

$$b = (I, L, F_x, F_y) \to F_x(I, L)^\intercal F_y(I, L) \tag{1}$$

$$u = \sum b(I, L) \tag{2}$$

We train the bi-linear network for 100 epochs with early stopping and a mean square error loss function. To enlarge the training data, we use augmentation on the fly and perform random rotations of 30°, horizontal flipping, and brightness intensity modification in the range of [0.75,1.25]. After the bi-linear network stage, a Long-Short Term Memory (LSTM) [48] network is used to model the temporal representation of pain. LSTM is a common type of Recurrent Neural Network (RNN) [49], which has seen success maintaining long-term dependencies while resolving the vanishing gradient problem. This is done through the use of 3 gates: input, output, and forget gates. The input gate maintains saved information over time, the forget gate ignores unimportant information, and the output gate controls what is passed to the next node in the stream. This network outputs a value of 1 (pain expression) or 0 (no pain expression). The

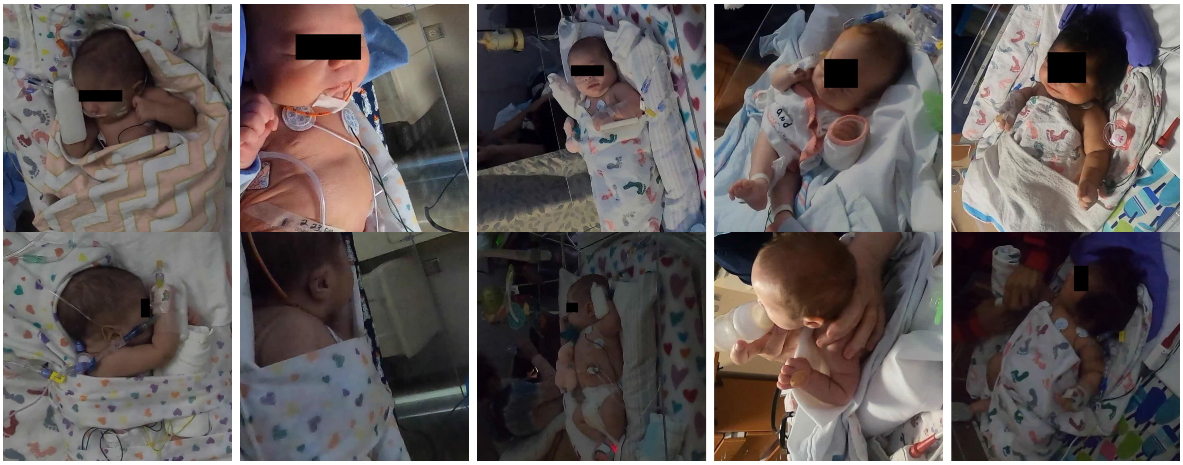

FIGURE 2. Figure giving examples from USF-MNPAD-II of cropped face images. This is passed as input to the face modality. From left to right, each of the columns is an individual subject. Top to bottom shows different complexity of images where top is a easier example ("Level 1"), and bottom is more complex ("Level 2", e.g. darker, more extreme pose variations and obstructions).

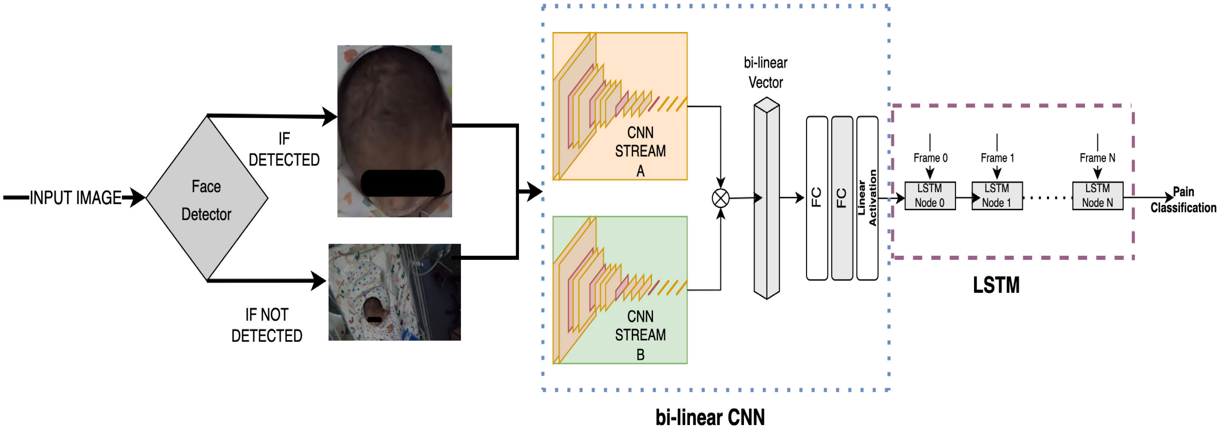

FIGURE 3. Figure of full system pipeline. An input set of images are provided to the face detection layer, which then determines the corresponding images passed to the bi-linear CNN & LSTM layers. Different subsets resulting from the use of differing face detection models are outlined in Section II-B.3.

LSTM model used in this work utilized 2 LSTM layers of unit 16, one layer returning all hidden states with the other not. The LSTM layers are followed by two dense layers with dropout of 30%, and a final single node dense layer with sigmoid activation. Training was done for 100 epochs (early stopping based on validation data), with an Adam optimizer (initial learning rate of $1e^{-4}$ with learning rate reduced on plateau) and a binary cross entropy loss function.

Fig. 3 shows a visual representation of the pain classification system with its two networks: bi-linear CNN and LSTM and the added neonatal face detector performing face detection at run-time. It is worth noting that due to the need for a set segment length in the LSTM stage any missed detections will still provide the original image to the bi-linear CNN. How to handle missing detections is something that can be explored further in future works.

3) DATA PREPARATION & EVALUATION PROTOCOL

For the training of our face detector the procedural pain subset of USF-MNPAD-I subjects were used. As these subjects were filmed during short procedures, video segments ranged from 1 min to 6 minutes in length. The key frames of these video segments were extracted to obtain a total of 8,826 images in the full training set. During experimentation, we utilized a subject-wise 10-fold cross-validation approach to test face detection accuracy on the procedural subset, with a full neonate face detector used for subsequent testing on the postoperative subset of USF-MNPAD-I being trained on all 36 procedural subjects. We further experiment as to the generalization of this trained face detector by testing it on the initial 15 (face data missing from one subject, 14 subjects total) subjects of the USF-MNPAD-II dataset.

For initial neonate face detection testing purposes, the postoperative portion (9 subjects) of USF-MNPAD-I was utilized. This subset consists of 218 video segments of a minimum of 9 seconds in length. Initially, all key frames are extracted from each of the 218 video segments resulting in a full testing set of 12,416 images (slightly reduced to the largest multiple of 32 images per video segment). Being a much larger and more complex dataset, for the USF-MNPAD-II testing we used 14 subjects consisting of

TABLE 1. **Distribution of training and testing across datasets. This table displays the training and testing split of subsequent results. ∗ Results using this test set are in Table 2 & Table 4. ⊙ results using this test set are in Table 3.**

Testset	Training	Validation
USF-MNPAD-I ∗	80% USF-MNPAD-I procedural	20% USF-MNPAD-I procedural
USF-MNPAD-II⊙	80% USF-MNPAD-I procedural & postoperative	20% USF-MNPAD-I procedural & postoperative

2,486 segments of exactly 30 seconds in length. We extracted frames at a rate of 15 fps.

For pain classification experimentation purposes, the USF-MNPAD-I postoperative test set was split into further subsets. Neonate face detection was applied to the 12,416 images using differing versions of a face detector. Then for each video segment, the correctly detected key frame images were further limited by a set length, taking images in order until the set length requirement was met. For the YOLOv5 Max Multiple subset, we took the largest grouping of found detections in a multiple of 32, allowing for multiple bi-linear vectors per video to be passed to the LSTM network. The set segment lengths are required for LSTM training, which is used in the pain/no-pain classification algorithm [9]. To note, not all of the video segments had enough correctly detected neonate faces to fulfill the set length requirement and thus this video segment is excluded for the pain/no-pain classification testing subset.

1) YOLOv3-SEG16: Set Length 16; 3,072 images from 192 video segments
2) YOLOv3-SEG32: Set Length 32; 5,920 images from 185 video segments
3) YOLOv5-SEG32: Set Length 32; 6,880 images from 215 video segments
4) YOLOv5-MM: Max Multiple of 32; 9,856 images from 215 video segments

After initial experimentation, two final subsets were created for equal comparison purposes. Instead of taking the first sequentially correctly detected images which met the defined set length, we equally distributed the key frame images which were correctly detected across the entire video segment. Initially, frames per segment were kept in an increase order over time, however for a more even comparison if a segment was able to detect more than the LSTM segment length, frames were utilized in an equally distributed temporal order. While still maintaining temporal order, this equal distribution saw an increase in accuracy for pain/no-pain classification most likely due to a more complete expression of pain being represented across the entire segment as compared to the initial images. Segments which did not have enough correctly detected faces to satisfy the 32 minimum were discarded.

5) YOLOv3-Equal: Equal Distribution of Set Length 32, 5920 images from 185 video segments
6) YOLOv5-Equal: Equal Distribution of Set Length 32, 6880 images from 215 video segments

All of these different subsets are then used as input to the bi-Linear CNN-LSTM to demonstrate how the performance of different face detection algorithms will impact pain/no pain classification accuracy, as shown in Table 4.

III. EXPERIMENTS

A. EVALUATION METRICS

To quantify the similarities between the predicted and ground truth samples, we used the Sørensen-Dice Coefficient (SDC) [50] which quantifies the amount of correct overlap between the predicted and found bounding boxes. This allows for us to have a metric determine not only if a neonate's face was detected within the frame but how well the detector was able to determine where in the image the face was located. Mathematically, SDC can be represented in the context of the bounding box problem by four metrics: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

$$SDC = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3}$$

To accept the output of the detector as accurate, the SDC value must be higher than 0.5; otherwise, the detection was deemed inaccurate. We only considered the bounding boxes that have a class confidence score higher than 40%. At runtime, any detection which had a confidence of 40% or lower was automatically discarded. These thresholds were empirically derived metrics. While in practice our system is able to remove the step of manual annotation, for this work we used human-defined bounding boxes for our ground truth. These are the same bounding boxes provided to the model during training.

B. FACE DETECTION RESULTS

We have included Table 1 in order to illustrate the different distribution of training and testing splits that were utilized to calculate the results in Table 2 and Table 3. Table 2 shows the performance of differing face detectors (v3, v5, v6) on the USF-MNPAD-I postoperative dataset. Recall that these detectors were trained on the procedural subset of USF-MNPAD-I, so no training data was seen at test time. We used the bounding boxes manually annotated by a human to provide our ground truth labels, and all of the labels were the same for each of the individual models.

When evaluated on the training set, YOLOv5 completely off-the-shelf achieved an average of 24.0% across USF-MNPAD-I procedural folds, with YOLOv3 trained on WIDER-FACE achieving an average of 68.96%, YOLOv5 achieves an average of 98.12% and YOLOv6 achieved an average of 99.92% for the same folds. This indicates that YOLOv5 or YOLOv6 trained on the USF-MNPAD-I

TABLE 2. Performance of face detection on USF-MNPAD-I postoperative subjects, subject-wise.

Subject	Images	V5 No Training	V3 Detection	V5 Detection	V6 Detection
A	818	142 (17.3%)	618(75.6%)	652 (79.7%)	818 (100%)
B	1464	0 (0%)	1099 (75.1%)	1427(97.47%)	1464 (100%)
C	294	0 (0%)	294(100%)	294(100%)	294 (100%)
D	3858	3 ($\sim$ 0%)	3679(95.4%)	3741(96.9%)	3845 (99.6%)
E	2147	0 (0%)	1994(92.8%)	2147(99.9%)	2147 (100%)
F	505	21 (4%)	504 (99.8%)	476 (94.2%)	505 (100%)
G	3239	136 (4%)	1848 (57%)	3199 (98.7%)	3239 (100%)
H	220	0 (0%)	74 (33.6%)	220 (100%)	220 (100%)
Total	12,545	302 (2.4%)	10,110 (80.5%)	12,156 (96.89%)	12,534 (99.89%)

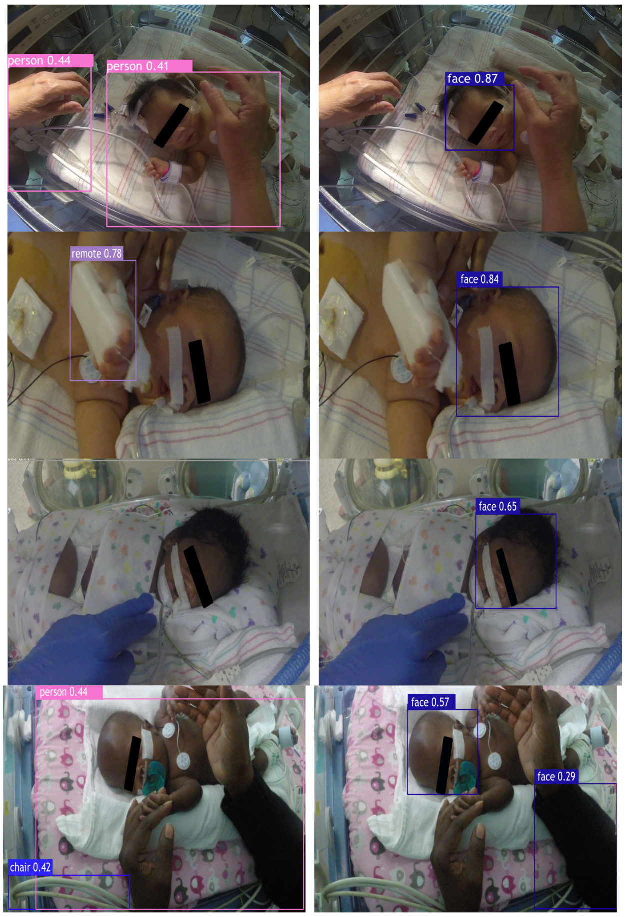

FIGURE 4. Comparison of not trained (Left Column) and trained YOLO output (Right Column).

procedural should yield respectable results on USF-MNPAD-I postoperative test set. An SDC of 0.5 was allowed, while maintaining a threshold confidence score of 0.4. As shown in Table 2, there is notable increase in accuracy between YOLOv3-WIDER FACE and the trained YOLOv5, with trained YOLOv6 achieving near perfect accuracy.

Table 3 shows the performance on YOLOv3 trained on WIDER-FACE, YOLOv5 trained on USF-MNPAD-I and YOLOv6 trained on USF-MNPAD-I, then tested on the full USF-MNPAD-II dataset. As will be analyzed in Section III-D.2, USF-MNPAD-II is a much more complex and challenging dataset than USF-MNPAD-I. Therefore, while YOLOv6 trained on USF-MNPAD-I procedural achieves near perfect accuracy on the USF-MNPAD-I post-operative subset, we see diminished results on USF-MNPAD-II with a mean of 62.7% accuracy, 76.1% AUC. Yolov5 trained on USF-MNPAD-I is also able to achieve a slightly diminished AUC of 64.2%. In comparison however, YOLOv3 trained on WIDER-FACE only achieves 26.4% on USF-MNPAD-II, ACU of 49.5 supporting our overarching assertion that off-shelf-solutions and solutions trained on adult faces are insufficient for neonate face detection. We can see these results additionally respresented as a confusion matrix and a ploted ROC curve in Table 5.

C. PAIN CLASSIFICATION RESULTS

Table 4 shows pain/no-pain classification on the full post-operative USF-MNPAD-I subset, a total of 12,416 images extracted from the 218 video segments. Differing face detection models resulted in differing images passed to the bi-linear CNN for feature extraction, where correctly detected neonates' faces are cropped by the found bounding box. To avoid missing the set LSTM length, any missed detections are supplemented by passing the full original image to the networks. How the different subsets were created is explained more in-depth in Section II-B.3. The metrics from the YOLOv6 subset indicate what the accuracy would be with 100% neonate face detection accuracy as reported in [9] by manual annotation. We see that YOLOv3-Equal (what can be considered our baseline) compared to the superior performing YOLOv5-Equal & YOLOv6 face detectors will result in a significant improvement in both mean Average Precision (mAP) and Area under the ROC Curve (AUC). Specifically, there was an increase of 8.6% mean Average Precision (mAP) and 21.2% Area under the ROC Curve (AUC) between using the baseline YOLOv3-Equal to the superior YOLOv6 face detection.

To back our claim that these results are statistically significant, we used the students paired t-test [51] to compare YOLOv3-Equal and YOLOv5-Equal [40], [44] detector versions which result in a $t - score = 1.4561$, at degree's freedom of 388, $p-value = 0.073087$ indicating significance at $p < 0.10$.

For repeatability reasons, the trained weights used in this paper are available at github.[2]

[2] https://github.com/ja05haus/trained_neonate_face

TABLE 3. Performance of face detection on USF-MNPAD-II with differing face detectors; YOLOv3 (trained on WIDERFACE), YOLOv5 (trained on USF-MNPAD-I), and YOLOv6 (trained on USF-MNPAD-I).

Subject	Video Segments	Accuracy of YOLOv3	Accuracy of YOLOv5	Accuracy of YOLOv6
A	340	6.6 %	80.3%	78.2 %
B	17	0.0 %	2.7%	46.1 %
C	184	63.1 %	70.2%	75.0 %
D	166	12.0 %	74.6%	79.5 %
E	120	0.83 %	83%	50.2 %
F	340	18.5 %	27.4%	52.2 %
G	71	68.5 %	31.3%	38.6 %
H	269	1.5 %	28.2%	59.4 %
I	251	1.6 %	11.5%	48.2 %
J	220	37.3 %	73.4%	64.6 %
K	156	69.2 %	59.3%	75.8 %
L	140	82.1 %	82.1%	79.6 %
M	91	46.2 %	63.5%	57.0%
N	121	24.8 %	26.9%	37.2 %
Total	2,486	26.4%	52.5%	62.7%
	AUC	49.5%	64.2%	67.1%
	Precision	35.3%	62.7%	63.2%
	Recall	49.3%	64.2%	67.1%

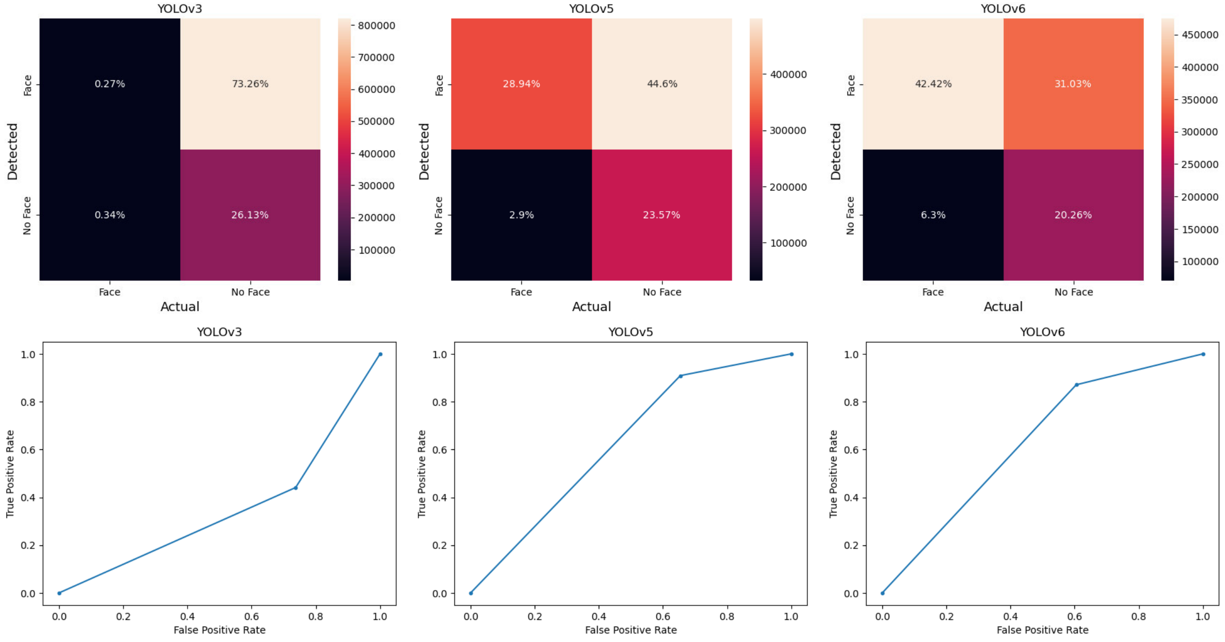

FIGURE 5. This figure includes the confusion matrix as well as the ROC curve for each YOLOv3 (trained on WIDERFACE), YOLOv5 (trained on USF-MNPAD-I), and YOLOv6 (trained on USF-MNPAD-I) while tested on USF-MNPAD-II. These results are also presented numerically in Table 3.

D. ANALYSIS OF RESULTS ACROSS SUBJECTS

1) YOLOv3-WIDER-FACE AND YOLOv5 COMPARISON ON USF-MNPAD-I

Table 2 shows that while the trained YOLOv5 demonstrates a large overall increase in accuracy, there are individual outlier subjects contradictory to these results. Figure 6 shows example images taken from a procedural and a postoperative subject, specifically instances where YOLOv3 was able to detect the neonate face slightly more often than YOLOv5. Conversely, Figure 6 also shows instances where YOLOv5 greatly outperforms YOLOv3. It can be reasoned that a weakness in YOLOv5 presents itself in cases of low resolution, where the neonate's face is small in scale compared to the full image and especially with added extreme pose variations. As demonstrated with the example images in Figure 6 exhibiting where YOLOv5 greatly outperforms YOLOv3, a major strength of YOLOv5 is the ability to detect the neonate's face with large obstructions blocking one or more key facial features.

2) COMPLEXITY OF USF-MNPAD-II

While the images in USF-MNPAD-I pose a variety of issues to current neonate face detectors, such as extreme pose variation and dark pixel values, images in USF-MNPAD-II

TABLE 4. Pain classification accuracy using different face detectors.

Subset	AP	Recall	F1 Score	AUC	FPR	TPR
YOLOv3-SEG16	0.485	0.475	0.404	0.489	0.748	0.777
YOLOv3-SEG32	0.492	0.479	0.448	0.492	0.746	0.732
YOLOv3-Equal	0.623	0.595	0.585	0.607	0.557	0.772
YOLOv5-SEG32	0.656	0.657	0.656	0.655	0.307	0.617
YOLOv5-MM	0.614	0.615	0.614	0.612	0.321	0.546
YOLOv5-Equal	0.689	0.688	0.685	0.683	0.229	0.596
YOLOv6* [9]	0.709	0.695	0.702	0.819	0.500	0.861

TABLE 5. This table presents an analysis of images misclassified by different detectors.

	USF-MNPAD-I Set	Images	Average Image Size	Average API	Average BBox Area
V3 Better	Procedural	130	1080 x 812	109.176	247 x 348
	Postoperative	506	1920 x 1080	105.649	442 x 603
V5 Better	Procedural	525	1080 x 608	104.33	457 x 426
	Postoperative	221	1920 x 1080	106.77	436 x 482
V3 Missed Detection	Postoperative	2435	1822 x 1061	98.01	565 x 593
V5 Missed Detection	Postoperative	389	1861 x 1068	95.9	547 x 667

TABLE 6. This table presents an analysis of images in USF-MNPAD-II.

Subject	Avg Image Size	Avg API of Image	Avg BBox Size	Avg API of BBox
A	1920x1080	91.22	348x303	65.02
B	1920x1080	48.57	274x250	24.11
C	1089x1910	51.37	217x207	79.9
D	1090x1909	61.64	454x363	50.52
E	1080x1920	101.75	248x168	100.5
F	1084x1915	35.76	200x180	43.75
G	1080x1920	49.35	449x497	51.72
H	1080x1920	63.42	268x260	55.25
I	1092x1907	59.13	298x299	51.4
J	1090x1909	79.26	233x262	83.53
K	1094x1905	54.82	257x273	56.48
L	1111x1888	35.05	226x249	73.49
M	1113x1886	58.94	296x288	57.16
N	1097x1902	74.74	205x206	89.86

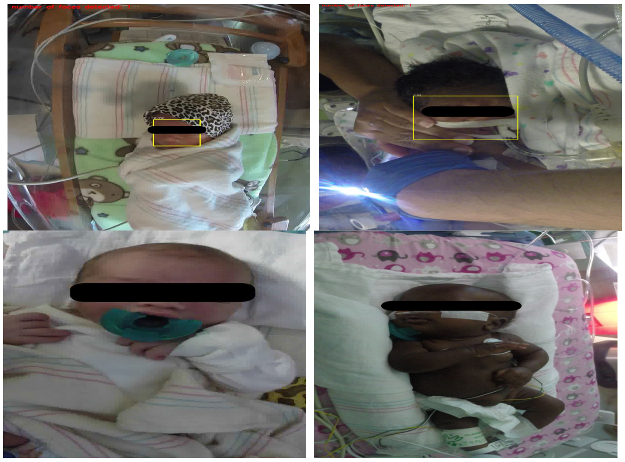

FIGURE 6. Top: Example images of correctly detected neonates face by YOLOv3, incorrectly by YOLOv5. Bottom: Example images of subjects with large increase in accuracy with YOLOv5 detector. Columns: procedural subject example and postoperative subject example, left to right respectively.

contain the same issues exacerbated. Even within the same subject, due to the length of recording, there will be discrepancies between frames in terms of the difficulty of neonate facial recognition. We have given visual examples of various levels of difficulties between subjects in Figure 2. We can observe in Table 6 that most of the subjects have a lower API, in the [40-80] range which indicates that the images are going to be very dark and thus posing a challenge to any face detector. If we look at subject L which is one subject which sees' very low accuracy (reported in Table 3) in face detection results, its API for the input images is as low as 35. During the course of our experimentation, we have determined that even to the human an average pixel intensity of 25 or lower renders images too dark for ground truth labeling, and thus we would not expect a computer algorithm to improve upon the human eye. Further, while the original images have a high resolution, due to the distance of recording many of the bounding box sizes are small in relation to the original image another complication when attempting to accurately detect a neonatal face.

IV. DISCUSSION

An important consideration is how to effectively translate this work into a range of clinical ICU settings, as well as home environments including both affluent and impoverished neighborhoods. To facilitate this transition we have developed a computationally lightweight algorithm with run-times consistent with hardware devices and cameras currently available in the NICU environment. For example, input data from a range of available RGB cameras, including those used by parents to remotely monitor neonates, and output to vital sign monitors capable of displaying algorithm results to clinicians both at the bedside and via remote alarms carried by busy NICU nurses. The emphasis in recent years on increasing accuracy [28], [29] is somewhat limited by testing systems in isolation, leading to a gap between what nurses and doctors can implement into their daily monitoring activities verses what developers envision using computationally heavy prototypes used for technology development. Other practical considerations are that pain assessment and avoidance systems are needed not only by well funded, state-of-the-art facilities with technically well-trained personnel and large equipment budgets, but also by smaller local NICUs whose staffing limitations could be bolstered by the additional support. Bio-medical applications are also most successful when they are not restricted to highly technically trained individuals.

The applications of this work extend beyond the run-time of an accurate neonatal face detector for simple pain/no-pain classification. A substantial body of evidence from human and animal research shows that early pain detection and mitigation with interventions such as non-opioids (e.g., acetominophen, ibuprofen) prior to severe pain onset can also prevent central sensitization (CS), the body's inherent defense strategy for avoidance of further pain. Severe pain and CS, if left unabated, can cause long-term damage to the newborn's developing nervous system [1], [2], [6]. Moreover, the onset of severe pain and CS elevates the potency to achieve pain relief using powerful narcotics, e.g., fentanyl, morphine, which have significantly more potential for side effects and complications, including the need for prolonged stays in the NICU for withdrawal. Thus, automatic face detection of neonatal pain using other modalities could prevent multiple traumas from severe pain, CS and reduce narcotic dependence/withdrawal that can cause long-term neurological damage to neonates during recovery in the NICU.

The challenges associated with collecting data for USF-MNPAD-I [42] highlight the need for a generalized system of accurate neonatal face detection at run-time. The acquisition of video and audio has taken many iterations, even with the most durable and compact camera solutions on the market. Prior to uploading data into deep neural network, human effort is required for supervision of pre-processing that includes ''cleaning the data,'' a time-consuming process that effectively interferes with a systems ability to perform classification at real time. A reliable and accurate neonatal face detector would allow for full automation of these steps.

Many computer vision applications [8], [9], [10], [11], [27], [28], [30], [33] require detection as the first step in the data processing pipeline. This demonstrates the need for an accurate solution, but accuracy must not be sacrificed for efficiency. Currently, many approaches rely on lengthy human-supervised manual annotation for detection. Automating this process through the use of the proposed framework would help begin to bridge the gap between theoretical solutions and solutions to real-life clinical problems.

V. CONCLUSION

We used a proprietary multi-modal facial image dataset from post-surgical neonates in the NICU to develop an automated pain/no-pain classifier. We compared 3 different face detectors, an off-the-shelf model, a second model trained by adult faces, and our customized YOLOv6 models trained on the faces of post-surgical neonate in multiple datasets. Only our automatic face detector trained on the faces of post-surgical infants was able to increase the accuracy of an automated pain/no-pain classifier. We share weights from our trained YOLOv6 models to encourage further enhancements in the accuracy of face detectors and integration of other modalities in automated systems for assessment and prediction of neonatal pain.

REFERENCES

[1] H. Popowicz, W. Medrzycka-Dąbrowska, K. Kwiecień-Jaguś, and A. Kamedulska, ''Knowledge and practices in neonatal pain management of nurses employed in hospitals with different levels of referral—Multicenter study,'' *Healthcare*, vol. 9, no. 1, p. 48, Jan. 2021.

[2] H. Popowicz, K. Kwiecień-Jaguś, J. Olszewska, and W. A. Medrzycla-Dąbrowska, ''Pain scales in neonates receiving mechanical ventilation in neonatal intensive care units—Systematic review,'' *J. Pain Res.*, vol. 3, pp. 1883–1897, Jul. 2020.

[3] A. Llerena, K. Tran, D. Choudhary, J. Hausmann, D. Goldgof, Y. Sun, and S. M. Prescott, ''Neonatal pain assessment: Do we have the right tools?'' *Frontiers Pediatrics*, vol. 10, Feb. 2023, doi: 10.3389/fped.2022.1022751.

[4] S. Brahnam, L. Nanni, and R. Sexton, ''Introduction to neonatal facial pain detection using common and advanced face classification techniques,'' in *Advanced Computational Intelligence Paradigms in Healthcare*. Internation Associaton for the Study of Pain, 2007.

[5] R. V. E. Grunau, C. C. Johnston, and K. D. Craig, ''Neonatal facial and cry responses to invasive and non-invasive procedures,'' *Pain*, vol. 42, no. 3, pp. 295–305, 1990.

[6] N. Bouwmeester, K. Anand, M. van Dijk, W. C. J. Hop, F. Boomsma, and D. Tibboel, ''Hormonal and metabolic stress responses after major surgery in children aged 0–3 years: A double-blind, randomized trial comparing the effects of continuous versus intermittent morphine,'' *Brit. J. Anaesthesia*, vol. 87, no. 3, pp. 309–399, Oct. 2001.

[7] A. Ohlsson and P. S. Shah, ''Paracetamol (acetaminophen) for prevention or treatment of pain in newborns,'' *Cochrane Database Systematic Rev.*, vol. 1, Oct. 2016, doi: 10.1002/14651858.CD011219.pub4.

[8] K. Hoti, P. T. Chivers, and J. D. Hughes, ''Assessing procedural pain in infants: A feasibility study evaluating a point-of-care mobile solution based on automated facial analysis,'' *Lancet Digit. Health*, vol. 3, no. 10, pp. 623–634, Oct. 2021.

[9] M. S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, and Y. Sun, ''Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment,'' *Comput. Biol. Med.*, vol. 129, Feb. 2021, Art. no. 104150.

[10] M. S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, and Y. Sun, ''Multi-channel neural network for assessing neonatal pain from videos,'' presented at the IEEE Int. Conf. Syst., Man Cybern. (SMC), Oct. 2019.

[11] M. S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, and Y. Sun, ''First investigation into the use of deep learning for continuous assessment of neonatal postoperative pain,'' presented at the 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), Nov. 2020.

[12] S. D. Roy, M. K. Bhowmik, P. Saha, and A. K. Ghosh, "An approach for automatic pain detection through facial expression," *Proc. Comput. Sci.*, vol. 84, pp. 99–106, Jan. 2016.

[13] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid, "Automatic detection of pain from facial expressions: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1815–1831, Jun. 2021.

[14] P. D. Barua, N. Baygin, S. Dogan, M. Baygin, N. Arunkumar, H. Fujita, T. Tuncer, R.-S. Tan, E. Palmer, M. M. B. Azizan, N. A. Kadri, and U. R. Acharya, "Automated detection of pain levels using deep feature extraction from shutter blinds-based dynamic-sized horizontal patches with facial images," *Sci. Rep.*, vol. 12, no. 1, p. 17297, Oct. 2022.

[15] A. Sadosky, B. Parsons, C. Schaefer, R. Mann, S. Daniel, S. Nalamachu, B. R. Stacey, E. Nieshoff, M. Tuchman, and A. Anschel, "Economic and humanistic burden of post-trauma and post-surgical neuropathic pain among adults in the United States," *J. Pain Res.*, vol. 17, pp. 459–469, Jun. 2013.

[16] M. Costigan and C. J. Woolf, "Pain: Molecular mechanisms," *J. Pain*, vol. 1, no. 3, pp. 35–44, Sep. 2000.

[17] C. J. Woolf and M.-S. Chong, "Preemptive analgesia—Treating postoperative pain by preventing the establishment of central sensitization," *Anesthesia Analgesia*, vol. 77, no. 2, pp. 362–379, Aug. 1993.

[18] J. L. Y. Cheong, A. C. Burnett, K. Treyvaud, and A. J. Spittle, "Early environment and long-term outcomes of preterm infants," *J. Neural Transmiss.*, vol. 127, no. 1, pp. 1–8, Jan. 2020.

[19] S. A. Ferguson, W. L. Ward, M. G. Paule, R. W. Hall, and K. J. S. Anand, "A pilot study of preemptive morphine analgesia in preterm neonates: Effects on head circumference, social behavior, and response latencies in early childhood," *Neurotoxicol. Teratol.*, vol. 34, no. 1, pp. 47–55, Jan. 2012.

[20] S. Singh, V. Rastogi, and S. Singh, "Pain assessment through facial expression," *Alternative Pain Manage.*, vol. 1, pp. 7–35, Jan. 2020.

[21] G. V. T. D. Silva, M. C. D. M. Barros, J. D. C. A. Soares, L. P. Carlini, T. M. Heiderich, R. N. Orsi, R. D. C. X. Balda, C. E. Thomaz, and R. Guinsburg, "What facial features does the pediatrician look to decide that a newborn is feeling pain?" *Amer. J. Perinatol.*, vol. 40, no. 8, pp. 851–857, Jun. 2021.

[22] J. W. B. Peters, H. M. Koot, R. E. Grunau, J. de Boer, M. J. van Druenen, D. Tibboel, and H. J. Duivenvoorden, "Neonatal facial coding system for assessing postoperative pain in infants: Item reduction is valid and feasible," *Clin. J. Pain*, vol. 19, no. 6, pp. 353–363, Nov. 2003.

[23] K. D. Craig, H. D. Hadjistavropoulos, R. V. E. Grunau, and M. F. Whitfield, "A comparison of two measures of facial activity during pain in the newborn child," *J. Pediatric Psychol.*, vol. 19, no. 3, pp. 305–318, 1994.

[24] M. S. Salekin, P. R. Mouton, G. Zamzmi, R. Patel, D. Goldgof, M. Kneusel, S. L. Elkins, E. Murray, M. E. Coughlin, D. Maguire, T. Ho, and Y. Sun, "Future roles of artificial intelligence in early pain management of newborns," *Paediatric Neonatal Pain*, vol. 3, no. 3, pp. 134–145, Aug. 2021.

[25] M. C. O'Neill, S. A. Kohut, R. P. Riddell, and H. Oster, "Age-related differences in the acute pain facial expression during infancy," *Eur. J. Pain*, vol. 23, no. 9, pp. 1596–1607, Oct. 2019.

[26] R. V. E. Grunau and K. D. Craig, "Pain expression in neonates: Facial action and cry," *Pain*, vol. 28, no. 3, pp. 395–410, Mar. 1987.

[27] M. Awais, C. Chen, X. Long, B. Yin, A. Nawaz, S. F. Abbasi, S. Akbarzadeh, L. Tao, C. Lu, L. Wang, R. M. Aarts, and W. Chen, "Novel framework: Face feature selection algorithm for neonatal facial and related attributes recognition," *IEEE Access*, vol. 8, pp. 59100–59113, 2020.

[28] Y. S. Dosso, K. Greenwood, J. Harrold, and J. R. Green, "RGB-D scene analysis in the NICU," *Comput. Biol. Med.*, vol. 138, Nov. 2021, Art. no. 104873.

[29] Y. S. Dosso, D. Kyrollos, K. J. Greenwood, J. Harrold, and J. R. Green, "NICUface: Robust neonatal face detection in complex NICU scenes," *IEEE Access*, vol. 10, pp. 62893–62909, 2022.

[30] S. Brahnam, L. Nanni, S. Mcmurtrey, A. Lumini, R. Brattin, M. Slack, and T. Barrier, "Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from Gaussian of local descriptors," *Appl. Comput. Informat.*, vol. 19, pp. 122–143, Jul. 2020.

[31] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," presented at the *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.

[32] B. Olmi, C. Manfredi, L. Frassineti, C. Dani, S. Lori, G. Bertini, S. Gabbanini, and A. Lanatà, "Aggregate channel features for newborn face detection in neonatal intensive care units," presented at the 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2022.

[33] E. Grooby, C. Sitaula, S. Ahani, L. Holsti, A. Malhotra, G. A. Dumont, and F. Marzbanrad, "Neonatal face and facial landmark detection from video recordings," 2023, *arXiv:2302.04341*.

[34] P. Barros, N. Churamani, and A. Sciutti, "The FaceChannel: A light-weight deep neural network for facial expression recognition," 2020, *arXiv:2004.08195v1*.

[35] S. Siddiqui, M. Vatsa, and R. Singh, "Face recognition for newborns, toddlers, and pre-school children: A deep learning approach," presented at the 24th Int. Conf. Pattern Recognit. (ICPR), Aug. 2018.

[36] M. Agarwal, "Analysis and evaluation of algorithms for newborn face recognition," Ph.D. dissertation, Indraprastha Inst. Inf. Technol., New Delhi, India, 2017.

[37] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," presented at the Brit. Mach. Vis. Conf., 2015.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[39] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[40] G. Jocher. (Apr. 2020). *YOLOv5 (Version 3.0)*. [Online]. Available: https://github.com/ultralytics/yolov5

[41] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.

[42] M. S. Salekin, G. Zamzmi, J. Hausmann, D. Goldgof, R. Kasturi, M. Kneusel, T. Ashmeade, T. Ho, and Y. Sun, "Multimodal neonatal procedural and postoperative pain assessment dataset," *Data Brief*, vol. 35, Apr. 2021, Art. no. 106796.

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," presented at the *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[44] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[46] Meituan. (2022). *YOLOv6 (Version 0.1.0)*. [Online]. Available: https://github.com/meituan/YOLOv6

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[49] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014.

[50] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *Biologiske Skrifter*, vol. 5, pp. 1–34, Jun. 1948.

[51] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, Mar. 1908.

JACQUELINE HAUSMANN (Student Member, IEEE) received the B.S. degree in computer science from Siena College, Loudonville, NY, USA, and the M.Sc. degree in computer science from the University of South Florida, Tampa, FL, USA, in 2020, where she is currently pursuing the Ph.D. degree in computer science and engineering.

MD SIRAJUS SALEKIN (Member, IEEE) received the M.Sc. degree in computer science and the Ph.D. degree in computer science and engineering from the University of South Florida, Tampa, FL, USA, in 2020 and 2022, respectively. He is currently an Applied Scientist in the IT industry. His research interests include machine learning, machine intelligence, computer vision, natural language processing, and generative AI.

GHADA ZAMZMI (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from the Department of Computer Science and Engineering, University of South Florida, in 2014 and 2018, respectively. She is currently a Research Fellow with the National Library of Medicine, National Institutes of Health. Her research emphasis is on emotion recognition, in particular, pain recognition for infants and individuals with communicative/neurological impairments. Her research interests include medical image analysis, computer vision, machine learning, and emotion recognition.

PETER R. MOUTON (Member, IEEE) received the B.S. degree in chemistry and biology and the Ph.D. degree in neurobiology from the University of South Florida (USF), Tampa, FL, USA, in 1983 and 1990, respectively. From 1988 to 1989, he was a Predoctoral Fellow in neuroscience with the Karolinska Institute, Stockholm, Sweden, for one year. From 1990 to 1992, he was a Postdoctoral Fellow in neurostereology with the University of Copenhagen, Copenhagen, Denmark, and a Postdoctoral Fellow in neuropathology with the Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA, for two-year. From 1994 to 2000, he was a Faculty Member with the Department of Pathology, Johns Hopkins School of Medicine. He is currently a Professor with the Department of Computer Sciences and Engineering, USF. Since March 2003, he has been the CEO and the Chief Scientific Officer of SRC Biosciences, a Maryland-based S-Corporation, with the mission of developing and commercializing biotechnology and related services to support bioscientists and clinicians in the completion of their research projects. He is the author of three books and more than 115 research articles. He holds patents for four inventions, and current or past principal investigator for ten SBIR/STTR grants from the NIH and NSF and four technology development grants from Florida High Tech Corridor Foundation. His research interests include applications of machine learning, computer vision, and deep learning to accelerate research in the fields of neuropathology, neurotoxicology, and prediction and prevention of pain. He is a standing member of the NIH grant review panel for Drug Discovery for Aging, Neuropsychiatric, and Neurologic Disorders. He received the Outstanding Faculty Award from the Office of the President, USF, in March 2022. He was also a recipient of the Ruth M. Salta Award; the Outstanding Alzheimer's Disease Researcher by American Health Assistance Foundation, in 1998; the Faculty Mentor Award from the Office of the Provost at Johns Hopkins University, in 2000; the Nathan Shock Award For Outstanding Aging Research from the Gerontological Society of America, in 2003; and named one of the top 12 technology innovators in the state by *Florida High Tech Magazine*, in 2014. He serves on the editorial board for the *Journal of Chemical Neuroanatomy*.

STEPHANIE PRESCOTT was born in Tampa, FL, USA. She received the B.S. degree in nursing from George Mason University, in 1994, the M.S. degree in neonatal nursing from The University of Alabama at Birmingham, Birmingham, AL, USA, in 2010, and the Ph.D. degree from the University of Virginia, in 2019. She completed postdoctoral training with the Laboratory of Integrative Cancer Immunology, National Cancer Institute, in 2020. She is a board-certified neonatal nurse practitioner with Inova L. J. Murphy Children's Hospital and an Assistant Professor and the Associate Director of the Biobehavioral Laboratory, University of South Florida College of Nursing. She has published most frequently on perinatal microbiota and immune interactions in pregnant women and neonates. Her research interests include neonatal pain and the metabolic, immune, and neurodevelopmental consequences of disruptions to the perinatal microbiota and immune systems.

THAO (TINA) HO is currently a Physician Scientist with the Department of Pediatrics, University of South Florida (USF), and the Physician Lead of the Best Practice Group, Tampa General Hospital NICU. She completed her pediatric residency with the Naval Medical Center, San Diego, CA, USA, in 2008, and neonatal fellowship with USF, in 2014. Her research interests include gut health including gut microbiota in relation to nutrition and diseases in preterm infants supported by the NIH. She has also participated in research and quality improvement projects involving gut microbiome in infants with opioid withdrawal syndrome and neonatal pain monitoring and management.

YU SUN (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Dalian University of Technology, in 1997 and 2000, respectively, and the Ph.D. degree in computer science from The University of Utah, in 2007. From December 2007 to May 2008, he was a Postdoctoral Associate with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. He is currently an Associate Professor with the Department of Computer Science and Engineering, University of South Florida. His research interests include intelligent systems, robotics, virtual reality, and medical applications. He is an Associate Editor of IEEE Transactions on Robotics

DMITRY GOLDGOF (Fellow, IEEE) is currently a Professor with the Department of Computer Science and Engineering, University of South Florida. He is also an Educator and a Scientist working in the area of medical image analysis, image and video processing, computer vision, and bioinformatics. He has graduated 28 Ph.D. and 44 M.S. students, published more than 95 journals and 220 conference papers, 20 book chapters, and edited five books (H-index of 50). He is a fellow of IAPR, AAAS, and AIMBE.