Comparative Analysis of Binary and Multiclass Activity Recognition in High-Quality Newborn Resuscitation Videos

Jorge García-Torres*[1], Øyvind Meinich-Bache[1,2], Sara Brunner[2], Siren Rettedal[3,4], Amalie Kibsgaard[5], and Kjersti Engan†[1]

[1]Dept. Electrical Engineering and Computer Science, University of Stavanger, Norway
[2]Laerdal Medical AS, Stavanger, Norway
[3]Dept. of Pediatrics, Stavanger University Hospital, Norway
[4]Faculty of Health Sciences, University of Stavanger, Norway
[5]Dept. of Research, Stavanger University Hospital, Norway

Abstract

Globally, 3-10% of newborns do not breathe spontaneously at birth and need resuscitation. Prompt initiation of resuscitative interventions such as tactile stimulation and positive pressure ventilation can reduce neonatal mortality and morbidity associated with birth asphyxia. Automated video analysis of resuscitation episodes may be beneficial for evaluation and debriefing purposes. In this work, a dataset of 220 newborn resuscitation videos collected at the Stavanger University Hospital (Norway) is used to develop NBT-I3D, a deep neural network pipeline to automatically recognize resuscitation activities. To assess the task, both binary and multiclass networks have undergone training, allowing for a comparison of the two approaches. Results obtained for binary classification show a mean precision and recall of 84.76% and 80.92%, respectively. For multiclass, a mean precision and recall of 72.26% and 74.80% are reported.

1 Introduction

Neonates are at the highest risk of dying during and shortly after birth. Of the 2.4 million infants that died in the first month of life in 2020, about 1 million died within the first 24 hours [1]. The majority of these deaths occur in low and medium-income countries, but even high-income countries face significant human and economic costs from neonatal birth-related complications.

Birth asphyxia, defined as the failure to establish breathing at birth, is one of the leading causes of neonatal morbidity and mortality [2]. Early initiation of resuscitation interventions within the first minute after birth (called the golden minute) can reduce the risk of death and long-term damage related to birth asphyxia [3]. Video analysis of newborn resuscitation episodes has shown promise for evaluation and training purposes [4, 5], but it is time-consuming and raises privacy concerns. Therefore, implementing artificial intelligence to automate this analysis would offer substantial benefits and could lead to the development of tools for newborn resuscitation research, debriefing, quality improvement, and real-time support.

Figure 1. Example of PPV (left) and CPAP (right). A T-piece resuscitator is a manually operated, flow-driven device used to deliver controlled breathing support at specified pressures. For PPV, the valve of the T-piece device should be intermittently blocked with a finger 30-60 times per minute, while for CPAP it is just held in position.

According to resuscitation guidelines [6, 7], critical resuscitative interventions for non-breathing newborns include positive pressure ventilation (PPV), tactile stimulation, suction, and chest compressions. PPV should be initiated within one minute after birth if the newborn is breathing insufficiently or the heart rate is lower than 100 beats per minute [8]. Suctioning is only recommended if the airway is obstructed [7]. Continuous positive airway pressure (CPAP) can be used for spontaneously breathing newborns to prevent airway collapse and improve breathing [9]. When analyzing resuscitation videos, distinguishing between PPV and CPAP can be challenging since the differences are mostly subtle finger movements (blocking a valve or squeezing a bag), as illustrated in Figure 1. In high-income countries, respiratory support is mostly delivered by fixed pressure devices or self-inflating bags [10]. In low-resource settings, self-inflating bags are mostly used.

Video-based Activity Recognition typically involves capturing spatial and temporal information from video frames [11]. Most datasets for human

*Corresponding Author: jorge.garcia-torres@uis.no
†Corresponding Author: kjersti.engan@uis.no

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.

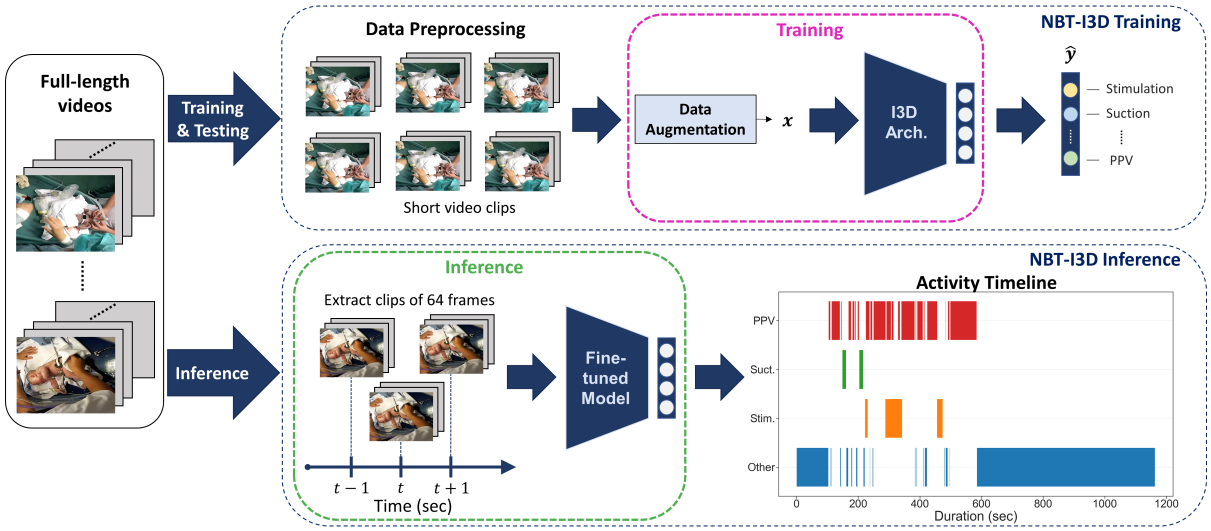

Figure 2. Methodology overview. *Training and Testing*: Full-length videos are used to extract short video clips containing the specific resuscitation activities. Data augmentation is applied to the RGB frames from video clips during training. For every input $\mathbf{x}$ to our model, we generate a one-hot output vector $\hat{\mathbf{y}}$ with the predicted activity. *Inference*: Sliding windows of 64 frames with a stride of 1 second (25 frames) are streamed into the model. The predictions are represented in a full timeline description.

activity recognition focus on single, easily recognizable activities. The state-of-the-art in activity recognition previously relied on 3D Convolutional Neural Network (CNN) [12] until vision transformers emerged [13]. Nevertheless, vision transformers often require extensive data to overcome the inherent inductive bias associated with CNNs [14]. In video-based neonatal resuscitation activity recognition, progress has been constrained due to the limited amount of available data. Previous studies employed techniques like combining a CNN for region detection and support vector machines for frame classification in Guo et al.[15]. Meinich-Bache et al. [16] introduced ORAA-net, a two-step neural network that integrated Object Detection with Region Proposal and Activity Recognition to identify activities in low-quality newborn resuscitation videos. However, this system required the estimation of optical flow and a high computational cost. More recently, Urdal et al. [17] combined 3D CNN with physiological signals for stimulation detection in low-quality videos.

NewbornTime project [18] is an interdisciplinary collaboration that seeks to enhance newborn care by employing artificial intelligence for activity and event recognition in video during and after birth. The objective of the project is to develop NewbornTimeline, a system capable of automatically generating a resuscitation timeline, including the time of birth and resuscitation activities. This tool is expected to facilitate the analysis of extensive newborn resuscitation videos, contributing to a deeper understanding of optimal newborn resuscitation practices. Previous work in this project has investigated the challenges of using thermal imaging during birth scenarios [19].

In this paper, we present NewbornTime-I3D (NBT-I3D), a pipeline that benefits from high-quality video recordings from a high-income country setting for detecting newborn resuscitation activities. We propose a more straightforward approach than Meinich-Bache et al. [16] as we eliminate object detection and optical flow processing, inputting the entire RGB video frame into the model. As a preliminary study towards developing a multi-activity NewbornTimeline, we present results for two simpler approaches in activity recognition: binary classification (activity or no activity) and multiclass classification. We concentrate on the most important resuscitation activities: *PPV*, *Stimulation*, *Suction*, and *CPAP*.

2 Data Material

The dataset contains 220 newborn resuscitation episodes that were collected at the Stavanger University Hospital in Norway, using cameras mounted over the resuscitation stations. The videos were edited to mainly cover the time period when a newborn was treated at the resuscitation station, with a duration ranging from 2-50 minutes and a median duration of 16 minutes. The resuscitation episodes were collected as part of a randomized controlled trial, and not primarily intended for video recognition. Therefore, there are variations between the videos in the position of the newborn with respect to the camera, and in the lighting. In addition, several healthcare providers may be applying different interventions at the same time, and obstructions to the view happen relatively frequently. All these factors

Activity	Time (hours)
CPAP	30.95
PPV	7.06
Stimulation	2.25
Suction	0.86

Table 1. Summary of the cumulative duration recorded for the main activities included in this study.

contribute to what we define as noise in our dataset. Video resolution is 720×1280 with a frame rate of 25 frames per second (fps).

The annotation of resuscitation episodes was carried out by a medical doctor using ELAN 5.8 (The Language Archive, Nijmegen, The Netherlands) [20]. For each activity considered to be of clinical interest that occurred in the videos, the starting time, the ending time, and the total duration of the activity were annotated. While segments of the video with complete occlusions were omitted from labeling, labels may be imprecise due to the aforementioned noise existing in the videos. A total of 21 activity labels are included in the 220 videos, including if the newborn is present at the resuscitation station. All activities are labeled so that a hands-off label can be inferred from the absence of all other labels when the newborn is at the resuscitation station. In this work, we will focus on the main therapeutic interventions: *PPV*, *Stimulation*, and *Suction*. We also include *CPAP* due to the visual similarity with *PPV*. The remaining labels are grouped together. Table 1 shows a summary of the cumulative duration recorded for each activity.

3 Methods

An overview of the proposed NBT-I3D pipeline is illustrated in Figure 2. It is designed to make the activity recognition process simple and efficient. It benefits from high-quality video recordings to create a one-step system that directly works with the original RGB video frames.

NBT-I3D uses the Inflated 3D CNN (I3D) architecture introduced by Carreira and Zisserman [12]. They proposed to combine RGB video frames and optical flow for activity recognition, demonstrating that 3D CNNs can benefit from pre-trained 2D CNNs and that transfer learning is also highly efficient in this task. Our pipeline omits the optical flow estimation for efficiency, using RGB frames only. Given the limited size of our dataset, we decided not to employ vision transformers.

3.1 Data Preprocessing

Data preprocessing consists of trimming the resuscitation videos into several short video clips and assigning the corresponding label to each video clip according to the manual annotations (see Figure 3).

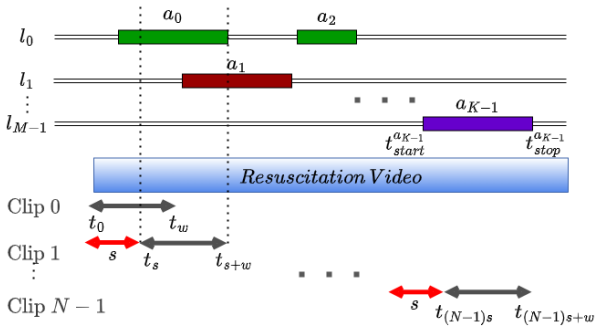

Figure 3. Representation of the data preprocessing. For Clip 1, we define the start and end points as t_s and t_{s+w}. We then assess the overlap between the clip and the time range of each annotation, defined by start and end points $t^{a_k}_{start}$ and $t^{a_k}_{stop}$. If the overlap meets the T_{min} criterion, the clip is labeled with the label associated with the annotation. In this example, Clip 1 overlaps with two annotations, a_0 and a_1. Assuming a 50% of T_{min}, both annotations meet it, and labels l_0 and l_1 are assigned to Clip 1.

To extract samples from a resuscitation video, we begin by defining the set of activity labels l_m contained in the dataset, where m ranges from 0 to $M-1$, and M represents the total number of labels. We also define the annotated sequences a_k with $k = 0, 1, ..., K-1$, where K denotes the total number of annotations in a resuscitation video. Each a_k has a starting time $t^{a_k}_{start}$ and a stopping time $t^{a_k}_{stop}$ associated with it, indicating the time in the video when an activity is occurring. We set a moving window of size w seconds, containing $fps \times w$ frames, and with a stride of s seconds ($fps \times s$ frames). Based on the values of w and s, we determine the total number of clips N that can be created from the resuscitation video. Subsequently, we identify the activities that are occurring within each video clip n by evaluating the time overlap $O_{w,s}(n,k)$ between the video clip time interval $[t_{ns}, t_{ns+w}]$ with respect to each annotation a_k:

$$O_{w,s}(n,k) = [t_{ns}, t_{ns+w}] \cap [t^{a_k}_{start}, t^{a_k}_{stop}] \quad (1)$$

where $0 \leq O_{w,s}(n,k) \leq w$. We can establish a label matrix $L_{w,s}(n,m)$ with dimensions $N \times M$ that stores the time overlap between video clips and activity labels. Each row and column corresponds to a video clip and specific activity, respectively:

$$L_{w,s}(n,m) = \sum_{k}^{K-1} O_{w,s}(n,k) I(a_k, l_m) \quad (2)$$

The indicator function I equals 1 if $a_k = l_m$ and 0 otherwise. We then set a minimum time of activity T_{min} that must be contained in the video clip to assign that activity label to the clip. Observing the label matrix L allows us to assess if a specific activity meets the T_{min} criterion, is partially occurring

(resulting in a "partial" label), or is completely absent in each video clip. Multiple labels and partial labels can be assigned to the same clip.

3.2 Subset Extraction

In each experiment, we selectively utilize video clips based on the designated Activities of Interest (AoI) we want to focus on. We extract video clips containing these AoI, even if they are only partially labeled. However, clips with partial AoI labels are not considered during the training process.

If there are multiple AoI, we encounter a disparity in the number of instances per activity, and to address this, we can balance our subset. This is achieved by downsampling the instances per activity to match the number of instances in the least dominant AoI. The downsampling is done randomly while ensuring that the number of instances per video is evenly distributed to introduce more variation in our AoI subset.

Additionally, since our dataset encompasses a wider range of activities, we can create an "Other" class comprising video clips that do not belong to the specific AoI designated for each particular experiment, including non-AoI video clips that are partially labeled. We extract samples from this non-AoI set until it contains the same number of instances as the most dominant AoI. This sampling process is again random but aims to distribute the number of instances per non-AoI uniformly, leading to increased variation within the "Other" class.

3.3 NBT-I3D Pipeline

In NBT-I3D, we maintain the original I3D architecture for the RGB pipeline but we omit the optical flow pipeline. We adjust the input size to match our video resolution while retaining the same number of input frames, i.e., 64. For the output, we replace the top layer with a new top layer with randomized weights and the right number of output classes according to the experiments. We do transfer learning from ImageNet and Kinetics, fine-tuning on newborn resuscitation data with no frozen layers.

Cross-entropy [21] is used as the loss function ($\mathcal{L}_{CE}(\mathbf{y}, \hat{\mathbf{y}})$), where $\mathbf{y}$ is the true label, and $\hat{\mathbf{y}} = f(\mathbf{x}, \theta)$ is the predicted label as a function of the input vector $\mathbf{x}$ of size $64 \times 256 \times 456 \times 3$ and the set of parameters of the model θ. In case of unbalanced data, we estimate the inverted class weight w_c for each class c and use a weighted cross-entropy loss function $\mathcal{L}_{WCE}$ as follows:

$$\mathcal{L}_{WCE}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} w_c y_c \log \hat{y}_c \qquad (3)$$

where C is the number of classes. With balanced data, w_c is 1 for all classes, and $\mathcal{L}_{WCE} = \mathcal{L}_{CE}$.

We train for a total of 20 epochs, setting a batch size of 6 and early stopping. Stochastic Gradient Descent with a learning rate of 0.001 and momentum of 0.9 is utilized for the optimizer. A Tesla V100 GPU with 32 GB is used. We assess precision and recall as evaluation metrics for binary experiments and macro-average values for multiclass [22]. In addition, we use Matthew Correlation coefficient (MCC) both for binary and multiclass cases [23].

For training and testing, the full-length videos are first preprocessed in order to generate short video clips (see Section 3.1). Video clip frames are resized to 256×456. We use a 3-sec window size (75 frames) with 1.5 sec of stride and consider a minimum amount of activity occurring in the video clip of 50% of the clip length (1.5 sec) to assign the activity label. A total amount of 141299 video clips are generated from the 220 videos. Thereafter, video clips containing specific resuscitation activities are extracted as explained in Section 3.2.

Data augmentation is performed during training. First, a small temporal cropping is applied, randomly picking 64 consecutive frames among the 75 frames in the clip. Then, we apply random left-right flipping consistently for all frames of each video clip. During testing and validation, we utilize the 64 frames centered in the clip.

To perform inference, we run the complete resuscitation video through the model and create a timeline. This involves resizing the whole video spatially to 256×456 and capturing segments of 64 consecutive frames centered around each second in the video. This method enables us to predict the activity corresponding to each specific second.

4 Experiments

We define the test set by manually selecting video clips from 20 videos, ensuring a label distribution similar to our dataset. The remaining video clips are allocated in the training and validation sets. This division is performed separately for each experiment, guaranteeing an 85%/15% split for each AoI, even if different videos are used. AoI samples from the same video are exclusively assigned to either the training or validation set, avoiding any information leakage. A total of 10 models were generated in the experiments summarized in Table 2.

The experiments are arranged in three groups. In *PPV Recognition*, we aim to recognize *PPV* from other activities. We then evaluate the influence of *CPAP* in case of not being considered in the "Other" class or as a combined label with *PPV*. Finally, we try to distinguish between *PPV* and *CPAP*. In *Other Binary Recognition*, we aim to recognize *Stimulation* and *Suction* from other activities. In *Multiclass Recognition*, we train multiclass models to detect *PPV*, *Stimulation*, and *Suction*. We try with bal-

Exp.	Classes	Balanced
PPV Recognition		
PPV-1	PPV vs. Other + CPAP	Yes
PPV-2	PPV vs. Other	Yes
PPV-3	PPV + CPAP vs. Other	Yes
PPV-4	PPV vs. CPAP	Yes
Other Binary Recognition		
Stim-1	Stim. vs. Other + CPAP	Yes
Suct-1	Suct. vs. Other + CPAP	Yes
Multiclass Recognition		
Multi-1	PPV, Stim., Suct.	Yes
Multi-2	PPV, Stim., Suct.	No
Multi-3	PPV, Stim., Suct., Other + CPAP	Yes
Multi-4	PPV, Stim., Suct., Other + CPAP	No

Table 2. List of experiments arranged in three groups. *Balanced* indicates if the training and test sets used in the experiment are balanced. If not, all available data is used.

anced/unbalanced data and with/without adding the "Other" class.

5 Results and Discussion

The results for the binary classification problems, Exp. PPV-x, Stim-1, and Suct-1 are presented in Table 3 together with results reported by Meinich-Bache et al. [16] on similar binary experiments. As shown, NBT-I3D exhibits a promising performance in detecting the main resuscitation activities. This indicates that using a larger database with higher-quality recordings is essential to achieve improvement while reducing the complexity of the system. It is crucial to underline that the dataset used in [16] is noticeably distinct. As stated in their work, data was collected from a low-resource setting at Haydom Lutheran Hospital, Tanzania, using low-quality cameras with variable frame rates, view angles, distances to the resuscitation station, and light conditions. Subsequently, object detection, frame interpolation, and optical flow were used to overcome this variability. Additionally, manual devices were employed for resuscitation activities like *PPV* or *Suction* in [16]. Our dataset was collected in a more controlled setting and using electronic devices. Therefore, a direct performance comparison is not feasible, but ORAA-net's results serve as a valuable reference.

In our results, we can observe the influence of the activity *CPAP* when attempting to predict *PPV*. Due to the visual similarity between these two activities (see Figure 1), we expect a higher number of *PPV* samples to be misclassified as *CPAP* when comparing Exp. PPV-1 (with *CPAP*) and PPV-2 (without *CPAP*). However, the problem with excluding *CPAP* is that it may lead to a significant number of false *PPV* predictions during inference. We explore grouping both activities into the same category in Exp. PPV-3, resulting in nearly perfect performance. However, it does come at the cost

	Experiment	Prec.*	Rec.*	MCC
ORAA [16]	PPV RGB	81.70	83.57	-
	PPV RGB+Flow	88.64	88.34	-
	Stim. RGB	75.48	73.13	-
	Stim. RGB+Flow	79.41	78.15	-
	Suct. RGB	44.96	59.81	-
	Suct. RGB+Flow	56.01	65.61	-
NBT-I3D	PPV-1	89.99	83.81	0.747
	PPV-2	94.34	90.40	0.856
	PPV-3	96.75	94.21	0.894
	PPV-4	85.36	83.96	0.725
	Stim-1	90.08	84.65	0.767
	Suct-1	69.86	66.81	0.516

Table 3. Performance of NBT-I3D on binary classification problems using RGB frames solely. Results provided by ORAA-net [16] on similar experiments are also included. *AoI are used as positive labels.

Experiment	Prec.*	Rec.*	MCC
Multi-1	81.64	81.47	0.723
Multi-2	81.71	81.65	0.792
Multi-3	73.00	71.33	0.623
Multi-4	72.26	74.80	0.665

Table 4. Performance of NBT-I3D on multiclass experiments. *Macro average is used.

of losing the distinction between two important activities. An attempt to classify only between *PPV* and *CPAP* is made in Exp. PPV-4, depicting that the model somewhat struggles to distinguish these activities accurately. Even so, precision and recall values in 0.84-0.85 are promising.

In Exp. Stim-1, NBT-I3D provides positive results when detecting *Stimulation*. The influence of other similar activities, such as *Dry*, may negatively affect the performance of the model. Similarly to Exp. PPV-3, grouping *Stimulation* and *Dry* into the same category can be explored in the future. Regarding Exp. Suct-1, the model also finds the detection of *Suction* challenging. This might be due to the small amount of video clips containing this activity.

For multiclass, Table 4 demonstrates consistent precision and recall values among experiments featuring matching class counts (Exp. Multi-1 and Multi-2, Exp. Multi-3 and Multi-4), regardless of data balance. A closer examination of the confusion matrices reveals that experiments with a balanced dataset exhibit more equitable metrics across all classes. In contrast, experiments with an unbalanced dataset expose the impact of the data distribution, even if a weighted loss function is used. To compare both approaches, MCC offers a comprehensive overview of the models' performance, highlighting the models using a weighted loss function.

To compare binary and multiclass recognition, we run inference on the same test video example utilizing both approaches. We use binary models from Exp. PPV-1, Stim-1, and Suct-1, and the

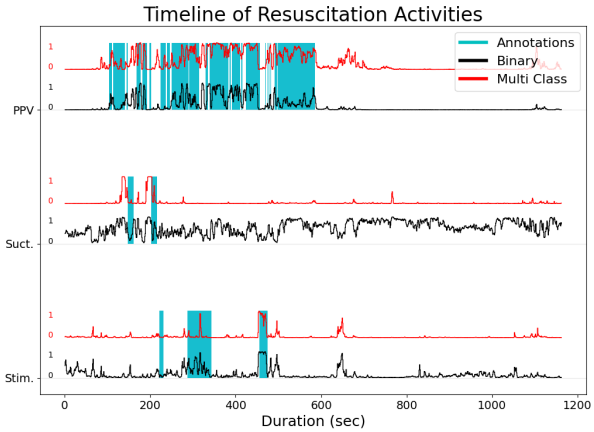

Figure 4. Illustration of the activity timeline generated from a video example. In blue, the original manual annotations containing *PPV*, *Stimulation*, and *Suction* activities. In black and red, the probability distribution of the predicted activities by the binary and the multiclass approaches, respectively.

multiclass model from Exp. Multi-4. As presented in Figure 4, for this episode, performance is similar in *PPV* and *Stimulation* for both approaches. However, as mentioned earlier, the binary model encounters challenges with *Suction*. The benefit of using multiple binary models is that we can detect multiple activities happening simultaneously, while the multiclass model can only predict one activity at a time.

Since our approach does not rely on additional steps and instead uses the whole video frame along with activity labels, we are confident in the adaptability of our methodology to diverse hospital settings and camera resolutions. However, ensuring minimum quality standards in camera configurations is essential to avoid spatial distortions or frame loss during video capture. The main challenge to the generalization of our method is the significant disparity that exists in intervention protocols and the medical devices employed during newborn resuscitation events across hospitals, particularly comparing low-income and high-income countries. While activities that are applied similarly across hospitals could be detectable for our system, fine-tuning the model with specific data from the target hospital is a potential solution for better adaptation to a new environment.

6 Conclusion

In this study, we demonstrate that high-quality videos enable us to create an effective and straightforward pipeline for newborn resuscitation activity recognition. NBT-I3D shows promising results in both binary detection and multiclass classification, with a mean precision and recall of 84.76% and 80.92% for binary detection and 72.26% and 74.80% for multiclass classification.

In future work, we will explore the multilabel approach to improve activity timelines and enhance system versatility. Our investigation will also extend to the detection of additional relevant activities within the dataset. Furthermore, we plan to leverage a newly acquired dataset from the same Norwegian hospital and incorporate simulation data for underrepresented classes. Finally, we aim to develop interpretable algorithms for video-based activity recognition to improve transparency and explainability.

Acknowledgements

The NewbornTime project is funded by the Norwegian Research Council (NRC), project number 320968. Additional funding has been provided by Helse Vest, Fondation Idella, and Helse Campus, Universitetet i Stavanger.

The study has been approved by the Regional Ethical Committee, Region West, Norway (REK-Vest), REK number: 222455. The project has been recommended by Sikt - Norwegian Agency for Shared Services in Education and Research, formerly known as NSD, number 816989. Informed consent was obtained from all mothers involved in the study.

We would like to express our gratitude to all the mothers, the health care providers, and the mercantile personnel who made this study possible. We would also like to express our gratitude to the other contributors involved in the NewbornTime project, with a special acknowledgment to Vilde Kolstad for their dedicated efforts in annotating the videos.

References

[1] W. H. Organization. *Newborn Mortality*. https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report-2021. [Online; accessed 21-August-2023]. 2022.

[2] W. H. Organization. *Perinatal Asphyxia*. https://www.who.int/teams/maternal-newborn-child-adolescent-health-and-ageing/newborn-health/perinatal-asphyxia. [Online; accessed 24-August-2023].

[3] H. L. Ersdal, E. Mduma, E. Svensen, and J. M. Perlman. "Early initiation of basic resuscitation interventions including face mask ventilation may reduce birth asphyxia related mortality in low-income countries: a prospective descriptive observational study". In: *Resuscitation* 83.7 (2012), pp. 869–873. DOI: https://doi.org/10.1016/j.resuscitation.2011.12.011.

[4] C. Skåre, A. M. Boldingh, J. Kramer-Johansen, T. E. Calisch, B. Nakstad, V. Nadkarni, T. M. Olasveengen, and D. E. Niles. "Video performance-debriefings and ventilation-refreshers improve quality of neonatal resuscitation". In: *Resuscitation* 132 (2018), pp. 140–146. DOI: https://doi.org/10.1016/j.resuscitation.2018.07.013.

[5] M. Heydarzadeh, A. Mousavi, S. Azizi, A. Hamedi, and S. S. Alavi. "Impact of Video-recorded Debriefing and Neonatal Resuscitation Program Workshops on Short-term Outcomes and Quality of Neonatal Resuscitation." In: *Iranian Journal of Neonatology* 11.2 (2020). DOI: https://doi.org/10.22038/ijn.2020.40999.1673.

[6] J. Madar, C. C. Roehr, S. Ainsworth, H. Ersdal, C. Morley, M. Ruediger, C. Skåre, T. Szczapa, A. Te Pas, D. Trevisanuto, et al. "European Resuscitation Council Guidelines 2021: Newborn resuscitation and support of transition of infants at birth". In: *Resuscitation* 161 (2021), pp. 291–326. DOI: https://doi.org/10.1016/j.resuscitation.2021.02.014.

[7] M. H. Wyckoff, R. Greif, P. T. Morley, K.-C. Ng, T. M. Olasveengen, E. M. Singletary, J. Soar, A. Cheng, I. R. Drennan, H. G. Liley, et al. "2022 International consensus on cardiopulmonary resuscitation and emergency cardiovascular care science with treatment recommendations: Summary from the basic life support; advanced life support; pediatric life support; neonatal life support; education, implementation, and teams; and first aid task forces". In: *Pediatrics* 151.2 (2023), e2022060463. DOI: https://doi.org/10.1542/peds.2022-060463.

[8] D. E. Niles, C. Cines, E. Insley, E. E. Foglia, O. U. Elci, C. Skåre, T. Olasveengen, A. Ades, M. Posencheg, V. M. Nadkarni, et al. "Incidence and characteristics of positive pressure ventilation delivered to newborns in a US tertiary academic hospital". In: *Resuscitation* 115 (2017), pp. 102–109. DOI: https://doi.org/10.1016/j.resuscitation.2017.03.035.

[9] C. C. Claassen and M. L. Strand. "Understanding the risks and benefits of delivery room CPAP for term infants". In: *Pediatrics* 144.3 (2019). DOI: https://doi.org/10.1542/peds.2019-1720.

[10] S. Tribolet, N. Hennuy, and V. Rigo. "Ventilation devices for neonatal resuscitation at birth: A systematic review and meta-analysis". In: *Resuscitation* (2023), p. 109681. DOI: https://doi.org/10.1016/j.resuscitation.2022.109681.

[11] G. Saleem, U. I. Bajwa, and R. H. Raza. "Toward human activity recognition: a survey". In: *Neural Computing and Applications* 35.5 (2023), pp. 4145–4182. DOI: https://doi.org/10.1007/s00521-022-07937-4.

[12] J. Carreira and A. Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308. DOI: https://doi.org/10.1109/CVPR.2017.502.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020). DOI: https://doi.org/10.48550/arXiv.2010.11929.

[14] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. "Transformers in vision: A survey". In: *ACM computing surveys (CSUR)* 54.10s (2022), pp. 1–41. DOI: https://doi.org/10.1145/3505244.

[15] Y. Guo, J. Wrammert, K. Singh, K. Ashish, K. Bradford, and A. Krishnamurthy. "Automatic analysis of neonatal video data to evaluate resuscitation performance". In: *2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*. IEEE. 2016, pp. 1–6. DOI: https://doi.org/10.1109/ICCABS.2016.7802775.

[16] Ø. Meinich-Bache, S. L. Austnes, K. Engan, I. Austvoll, T. Eftestøl, H. Myklebust, S. Kusulla, H. Kidanto, and H. Ersdal. "Activity recognition from newborn resuscitation videos". In: *IEEE journal of biomedical and health informatics* 24.11 (2020), pp. 3258–3267. DOI: https://doi.org/10.1109/JBHI.2020.2978252.

[17] J. Urdal, K. Engan, T. Eftestøl, Øyvind Meinich-Bache, I. A. Haug, P. F. Mdoe, E. Mduma, L. B. Yarrot, H. Kidanto, and H. Ersdal. "Automatic prediction of therapeutic activities during newborn resuscitation combining video and signal data". In: *Biomedical Signal Processing and Control* 86 (2023), p. 105290. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2023.105290.

[18] K. Engan, Ø. Meinich-Bache, S. Brunner, H. Myklebust, C. Rong, J. García-Torres, H. L. Ersdal, A. Johannessen, H. M. Pike, and S. Rettedal. "Newborn Time-improved newborn care based on video and artificial intelligence-study protocol". In: *BMC Digital Health* 1.1

(2023), pp. 1–11. DOI: https://doi.org/10.1186/s44247-023-00010-7.

[19] J. García-Torres, Ø. Meinich-Bache, S. Brunner, A. Johannessen, S. Rettedal, and K. Engan. "Towards using Thermal Cameras in Birth Detection". In: *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE. 2022, pp. 1–5. DOI: https://doi.org/10.1109/IVMSP54334.2022.9816177.

[20] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. "ELAN: A professional framework for multimodality research". In: *5th international conference on language resources and evaluation (LREC 2006)*. 2006, pp. 1556–1559. URL: https://archive.mpi.nl/tla/elan.

[21] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. Springer, 2006. DOI: https://doi.org/10.1117/1.2819119.

[22] Z. C. Lipton, C. Elkan, and B. Naryanaswamy. "Optimal thresholding of classifiers to maximize F1 measure". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*. Springer. 2014, pp. 225–239. DOI: https://doi.org/10.1007/978-3-662-44851-9_15.

[23] S. Boughorbel, F. Jarray, and M. El-Anbari. "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". In: *PloS one* 12.6 (2017), e0177678.