Identifying Variation in the Newborn Life Support Procedure: An Automated Method

Alfian Tan

Resilience Engineering Research Group, University of Nottingham, United Kingdom. E-mail: alfian.tan@nottingham.ac.uk,
Industrial Engineering Department, Parahyangan Catholic University, Indonesia. E-mail: alfian.tan@unpar.ac.id

Rasa Remenyte-Prescott

Resilience Engineering Research Group, University of Nottingham, United Kingdom. E-mail: r.remenyte-prescott@nottingham.ac.uk

Joy Egede

Computer Vision Lab, School of Computer Science, University of Nottingham, United Kingdom, E-mail: joy.egede@nottingham.ac.uk

Michel Valstar

Computer Vision Lab, School of Computer Science, University of Nottingham, United Kingdom, E-mail: michel.valstar@nottingham.ac.uk
Blueskeye AI, E-mail: michel@blueskeye.com

Don Sharkey

Centre for Perinatal Research, School of Medicine, University of Nottingham, United Kingdom. E-mail: don.sharkey@nottingham.ac.uk

This research is conducted for developing an automated method to recognize variations in the Newborn Life Support (NLS) procedure. Compliance with the NLS standard guideline is essential to prevent any adverse consequences for the newborn. Video recordings of resuscitation are frequently used in research to identify types of variations and understand how to minimize unwanted ones. Despite their benefits, it takes a significant amount of time and human resources to manually evaluate the procedure from videos. Therefore, an automated method could help. In this study, a variation recognition based on an action recognition technique is built. In the first step, automatic object segmentation is performed on every NLS action image. In the second stage, a number of features involving the proportion of medical objects availability and their movement, as well as association among actions are extracted and fed into machine learning models. The results show that the strategy of considering actions' associations and preliminary prediction of actions succeeded in improving the model performance. However, the whole recognition system still works fairly, and it is only for the wet towel removal step in the procedure, yet it has been useful to inform the adherence of the recorded procedure to the NLS guideline. This study is an initial work that will advance toward the integration of automated variation recognition with reliability modeling work on the NLS procedure.

Keywords: Newborn Life Support, Image Segmentation, Action Recognition, Reliability in Healthcare

1. Introduction

The Newborn Life Support procedure is an evidence-based protocol to resuscitate and stabilize a compromised newborn. Around 5-10% of newborns do not spontaneously breathe when they are born so basic life support is needed (Finer and Rich 2010). Based on the NLS guideline (Fawke et al. 2021), the support includes thermal care, stimulation, inflation breaths, and ventilation while continuously observing the response of the baby. Based on the response, further escalation involving chest compressions and emergency drug administration may be considered.

Accurate and in-time action is essential to recover babies' condition. However, a

Proceedings of the 33rd European Safety and Reliability Conference
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio

ISBN: 981-973-0000-00-0 :: doi: 10.3850/981-973-0000-00-0 esrel2023-paper

considerable error rate of 15%-28% is still found (Yamada, Yaeger, and Halamek 2015). Based on (Sawyer, Lee, and Aziz 2018), it is realized that optimal teamwork and communication are critical in addition to individual clinical skills. To improve the NLS performance; action study and technology development are continuously pursued. For instance, novel heart rate monitoring devices have been developed (Henry et al. 2021). Beneficial use of Artificial Intelligence (AI) has also been reported (Kwok et al. 2022) such as for disease and response prediction, as well as infant's gestational age estimation.

As part of NLS research, video recordings are often used to study variations in the NLS procedure. Unfortunately, manually evaluating the procedure from videos needs a significant amount of time. Therefore, AI technology for recognizing variations in the NLS procedure would be helpful to deal with this issue. In this research, an action recognition system is developed to study NLS variation.

According to (Kong and Fu 2022), an action recognition method can be categorized into shallow and deep approaches. The latter can automatically define the action features and do the action prediction by itself. A large number of datasets is usually needed for this approach. Unfortunately, limited data is also common in this area of work. Therefore, a shallow approach is still a useful option considering that it is easier to train, and usually performs well with small datasets (Kong and Fu 2022).

The work of (Wang et al. 2017) and (Smith et al. 2019) deal with small datasets in action recognition, and they use the Deep Learning approach. The first work applies an internal transfer learning strategy, while the second work divides the action recognition into two steps which firstly simplify the input data to make the subsequent action classification easier. A shallow approach was also considered in the second work since an alternative representation of the input data is created for the first part of the method. These 2 papers show that the deep approach can still be used for small dataset problems, while the combination of the two approaches will also be beneficial to consider.

In our work, the action recognition task deals with a small amount of data. There are 23 NLS video recordings used for this study. A combination of shallow and deep approaches is considered to utilize the benefit of these two approaches. The concept of the 2-step approach of (Smith et al. 2019) is also applied to our study. In the first step, the input data is simplified by applying image segmentation. In the second step, a number of features are extracted from the segmentation results and used as inputs for the action prediction. In addition, since there is possibly more than one action happening simultaneously at any time during the NLS procedure, the relationship among actions is considered in the learning process in this study.

The rest of the paper is organized as follows: Section 2 briefly describes the video recordings used in this study. Section 3 explains the image segmentation method. Section 4 describes how action prediction models are developed. Section 5 shows the result and presents a discussion of this study. Section 6 summarizes the results and shares avenues for future work.

2. NLS Video Recordings and Data Ethics

The 23 videos used in this study are top-view action recordings performed at a speed of 30 frames per second with a resolution of 1280 x 720 pixels. These videos come from the observations conducted by (Henry et al. 2021), as well as those used by (Smith et al. 2019). Ethical approval (NHS Health Research Authority Yorkshire & The Humber-Sheffield Research Ethics Committee 15/YH/0522) and parental consent to collect and use the video recordings for research purposes have been obtained as can be referred to (Henry et al. 2021). It includes the use of the dataset for further related research.

3. Image Segmentation

3.1. *Objects for Segmentation Process*

In this step, a list of 18 medical objects (see Table 1) is defined. These objects are determined based on observations in the video recordings and literature (Sawyer, Lee, and Aziz 2018; Fawke et al. 2021). The first 15 objects are chosen based on the consideration that they can uniquely represent action categories in the NLS procedure. The last 3 objects (object 16-18) are added to a supporting category that will be used to gain

information about how the medical equipment is used during the NLS procedure.

3.2. *Segmentation Model*

The U-net model (Ronneberger, Fischer, and Brox 2015) is trained to do object segmentation. Model input is an image of 128 x 128 pixels with an RGB color channel, while the output is a segmented image of 128 x 128 pixels with 1 channel of pixel labels that correspond to a set of objects under study. This method is a type of supervised learning strategy so that annotated images containing object labels are needed.

Table 1. NLS Medical Objects

No	Object	Act Category
1	Blue towel	Thermal care
2	Hat	
3	Plastic bag	
4	Dry towel	
5	Suction catheter 1	Airway clearing
6	Suction catheter 2	
7	Stethoscope	Auscultation
8	Electric patches	
9	SpO_2 monitor	
10	Pipes	Inflation and ventilation
11	Syringe (Feeding Tube)	
12	T-piece	
13	Laryngoscope	Intubation
14	Endotracheal tube	
15	Endotracheal securing tape	
16	Baby	Supporting Object
17	Gloves (hand)	
18	Arm	

Due to the small number of domain-specific data, the U-net model is firstly pre-trained on a larger and more domain-general, publicly available COCO Dataset (Tsung-Yi Lin et al. 2014). The model is subsequently fine-tuned on our domain-specific data. COCO Dataset 2017 version is used with 90 classes of objects in the training process. A total of 10000 images are used from the COCO dataset for this stage.

A total of 1141 video frames are extracted from NLS videos for the fine-tuning step. Ideally, all these images need to be annotated/labeled. However, since each image needs a long manual annotation time (around 20-30 minutes), a cooperative learning strategy (Wagner et al. 2018) is used to save time.

Using this strategy, the dataset is randomly divided into 8 groups, containing a total sample ranging from 112-189 images per group. The training starts by using the first group of datasets. The best model of the first training stage is used to do the segmentation on the second group of datasets. The segmentation results are evaluated, and label corrections are made for poorly segmented images. Finally, after getting a proper annotation of all images, the dataset from the two groups is used to retrain the U-net model. This procedure is repeated until no further improvement in the performance can be obtained. The best model in every stage is obtained by experimenting with the hyperparameters of the model and its loss function. It focuses on the Intersection over Union ($0 \leq IoU \leq 1$) indicator, in which a higher IoU signifies a better performance of the segmentation process.

4. Action Recognition

4.1. *Dataset Generation*

The second step of the action recognition system aims to recognize NLS actions based on segmented images. However, instead of learning from a single image, the model is designed to analyze a group (volume) of sequential segmented action frames. A number of overlapping volume samples containing 150 frames each are generated. It results in a volume sample with a dimension of $128 \times 128 \times 150$. Sixteen and four NLS videos are allocated to a group of population for training and validation dataset generation, respectively. The rest three videos are used for testing the trained action recognition system.

There are 8 prioritized NLS actions and 1 unknown action category on which the action recognition model is trained. Other actions besides the 8 NLS categories will be considered as the unknown class. All these 8 + 1 action categories are (1). Inflation breath, (2). Tracheal Intubation, (3). Heart rate assessment, (4). Ventilation, (5). Removing wet towel, (6). Baby head position, (7). Covering the baby with dry towels, (8). Covering the baby with a polythene bag, and (9). Unknown category.

Every generated volume sample is paired up with information regarding the presence of each of the 9 actions in that particular action scene (volume). This information becomes the data labels for the supervised action recognition learning process.

4.2. *Machine Learning Algorithm*

In order to do the action recognition, traditional Machine Learning (ML) algorithms are utilized. Since the first step of data processing has used a Deep Learning (DL) method to produce a simplified image, it is more useful if we can define our more interpretable features from this simplified image so that it can give us more intuitive information for the action recognition learning process. These handcrafted features can be fed into the algorithms for them to learn the action types. This ML approach can also reduce the required end-to-end computing resources of the prediction process since it usually deals with a much fewer number of parameters, compared to DL models.

4.3. *Feature Definition and Extraction*

There are 3 defined features to be extracted from every volume sample. The proportion of a medical object's appearance in a clinical scene is used to characterize an action due to its association with NLS actions. To get more information on how the devices are used, the relative distances between medical objects and the baby, as well as between objects and clinical staff's hands, are extracted from each image/frame in the volume sample. The correlation statistic between the 150 relative distances obtained for every object in each volume sample and the time order of the frames is then calculated. A negative correlation may indicate that a particular object becomes closer either to the baby or to the hand as time advances, while a positive correlation means the opposite. A more detailed explanation of this feature extraction can be found below:

1. *Proportion of every medical object*
   Every segmented frame in a volume sample of 150 images has information about identified objects in it. It is indicated by different colors in a segmented image (see example in Fig. 1). The proportion of an object's availability is calculated as a fraction of a particular object's appearance in the 150 video frames.
2. *Correlation between time order and distance of a medical object to the baby.*
   The distance is calculated between 2 objects' centroids. The centroid of every object is computed by averaging all pixel locations of the object. Fig 1 illustrates centroids of baby (■) and T-piece (■). The Euclidean distance is used to calculate the distance. There would be 150 distance values for every object in one volume sample. This set of values, from the first frame until the 150th frame, will be correlated to its corresponding enumerated frame order of 1 to 150. The Pearson correlation coefficient is used here.
3. *Correlation between time order and the closest distance of an object to hand.*
   The distance is calculated by searching for the closest pixel location of the referenced object to another object centroid. This method is chosen since there may be more than one hand appearing in the resuscitation area. Since the segmentation method doesn't differentiate between hands with different values (colors), it would not be correct to represent all different hands with only one centroid value. Therefore, the algorithm will compute distances between a medical object's centroid and all identified pixel locations of hands and choose the minimum value as its relative distance. A correlation statistic is calculated afterward.

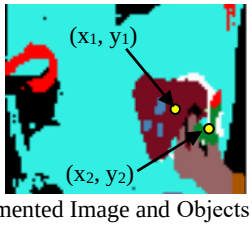

Fig 1. Segmented Image and Objects' Centroids

4.4. *Action Association Analysis*

This analysis is conducted for every action pair in the population of the training dataset. Table 2 shows an example of the frequencies of actions in the dataset. Binary values of 0 (negative signal) and 1 (positive signal) refer to the absence and the presence of an action, respectively. The proportion refers to the conditional proportion of an action's presence, given either the presence or the absence of another member of the action pair. For example, in Table 2, the proportion of positive signal of action 2 given the absence of action 1 is equal to 1.33%.

Based on Table 2, it can be observed that there is never a sample with both action 1 and action 2 happening at the same time. It indicates that these two actions may have a relationship to

consider. This condition can be clinically justified because the inflation breath (action 1) is usually part of the first resuscitation actions given to a baby, while tracheal intubation (action 2) is part of advanced actions considered when the baby doesn't respond to respiratory support procedures following the inflation breath. Table 3 shows the relationships for all action combinations based on the analysis. It is a non-symmetrical table with a one-way interpretation of the row action that affects the status of the column action.

Table 2. Action 1 vs Action 2

	Action 2		
Action 1	**0**	**1**	**Proportion**
0	11234	152	1.33%
1	176	0	0%
Proportion	1.54%	0%	

Table 3. Actions Relationship

Act	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1		v	v	v	v	v	v	v	v
2	v		v	v	v	v	v	v	v
3	v	v		v	v	v	v	v	v
4	v	v	v		v	v	v		v
5	v	v	v	v		v		v	v
6	v	v	v	v	v		v	v	
7	v	v	v	v		v		v	v
8	v	v	v		v	v	v		
9	v	v	v	v	v	v	v	v	

*v = action association exists

4.5. *Learning Strategy*

To deal with multilabel action recognition in this study, an approach of optimizing independent binary classifiers (0: absence/negative, 1: presence/positive) for each action is applied. However, action relationships are introduced in the learning process. For each action, eight binary classifier algorithms are trained, and the best-performing algorithm is selected. The algorithms include Logistic Regression (LogReg), Linear Support Vector Machine (LinSVM), Random Forest (RandForest), Adaptive Boosting (AdaBoost) (Hastie et al. 2009), Gradient Boosting (GradBoost) (Friedman 2002), XGBoost (Chen and Guestrin 2016), LightGBM (LitGBM) (Ke et al. 2017), and Categorical Boosting (CatBoost) (Dorogush, Ershov, and Gulin 2018). These classifiers will predict an action's presence by returning either a binary value (0 or 1) or a confidence score (0,1).

Table 4 shows the number of training and validation datasets for every action as well as for every binary signal in each action. The training set arrangement considers the balance between the negative and the positive signal instances to minimize bias during the learning process. On the contrary, the validation set is proportionally sampled following the real proportion of positive and negative signals in the population so that it can represent the real situation. A ratio of 75%:25% of the number of training and validation datasets is applied.

Table 4. Action Composition Dataset

	Training		**Validation**	
Act	**0**	**1**	**0**	**1**
1	267	176	138	15
2	227	152	119	10
3	2992	2066	1351	252
4	1893	1297	908	103
5	154	103	81	8
6	82	53	42	4
7	292	200	154	8
8	129	80	66	4
9	2407	3481	290	1574

The learning process will go through 2 stages. In the first stage, every action model will be trained only on the 3 types of features of volume samples. The best model of every action is determined based on the average F1 score and the minimum F1 score between the negative and the positive signal class. In the second stage (Stage 2a), action association is introduced by applying all the best models obtained in the first stage to the training dataset of targeted action. The confidence scores of the positive signal class of every associated action from these models are taken and used as additional information along with the 3 types of initial data features to train action classifiers in the second stage. Fig 2 illustrates how this second stage is performed for action 1. It considers the 8 actions associated with action 1 (i.e., act 2 to 9). The confidence score of each associated action will be concatenated to the three types of action 1 data features to become the final feature set to train the binary classifier algorithms of the action. The eight types of algorithms are also used for this purpose and the best one is chosen.

An alternative learning strategy for stage 2 (Stage 2b) is explored. It will not only consider

the associated actions' predictions but also the preliminary prediction of the targeted action from stage 1. In the end, the best prediction model for every action is selected by evaluating all models obtained from stage 1 and stage 2.

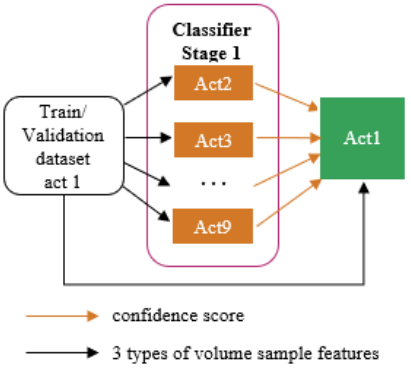

Fig 2. Training Strategy: Stage 2a for Action 1

5. Results and Discussion

5.1. *Image Segmentation Performance*

Based on our experiment, the best performance that the image segmentation model can achieve is an average IoU of 63.95% (range: 30.8% - 91.7%) on the validation dataset. Among the 19 object categories, there are 13 (68.42%) categories with IoU greater than 0.5. This performance is achieved by training the U-net model with Focal Loss ($\gamma = 9$) function (T -Y. Lin et al. 2020) for 3250 iterations (epochs) and fine-tuned by freezing the parameter values of the first half part of U-net layers (decoding block) that are already pre-trained on the COCO dataset. Adam optimizer with a learning rate decay of exp (-1/650) is used to optimize the model parameter. The gamma (γ) parameter in the Focal loss function aims to deal with imbalanced sample classes. An example of segmentation results with this best model can be seen in Fig 3, where the Ground Truth (GT) for the segmented image is put in the middle.

Despite the effort to reduce the imbalanced learning of object classes occupying larger pixel areas in the image, poor performance for small object segmentation can still be observed. The lowest performance is for the endotracheal tube with an IoU of 30.8%. Increasing the gamma value does not necessarily work, in fact, based on our experiment, it may worsen the performance. Introducing object class weight to overcome this issue also does not give any better results. Hence, a different learning strategy is needed to deal with a problem where the imbalance is significantly severe. This poor performance can also be caused by the characteristics of the objects. For example, in the original image, the endotracheal tube is observed to have a light blue color which is the same base color as the glove (hand). This characteristic along with the tiny size of the object could make the segmentation difficult.

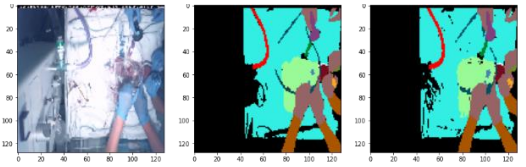

Fig 3. Segmentation Model Output
Real (Left), GT (Middle), Prediction (Right)

5.2. *Action Recognition Performance*

Different ML algorithms and 3 combinations of hyperparameter values are applied to find the best model for every action. Briefly explained, the 3 hyperparameter configurations are differentiated by the combinations of the number of iterations, learning rate, randomization factor of features and samples, as well as the penalty values to reduce an overfitting issue. Table 5 shows the best minimum F1 score coming either from the positive or negative signal class for every action across all training stages.

Six one-sided paired t-tests ($\mu_{difference} > 0$) on the mean difference of the minimum F1 score between stage 2 and stage 1 are conducted with a significance level of 0.05. The score difference of Stage 2a-Stage 1 and Stage 2b-Stage1 for every hyperparameter setting is computed. All the statistical tests indicate significant positive differences with P-values ranging from 3.2×10^{-7} to 4.3×10^{-2}. It confirms the benefit of introducing the action association into the learning process. However, in the end, the action recognition system is built by practically choosing action prediction models with the best score across all stages and experiments as can be seen in Table 5.

The average minimum F1 score of all these best models is 51%. Although improvements can be obtained, 56% of action models still have less than 50% minimum F1 score. Unequal opportunity for the models to comprehensively

learn the variation of each binary class is suspected to be the culprit of this performance. The number of instances of a member of the 2 binary classes in the dataset population that is often much lower than its counterpart becomes the limitation to generating more training samples for the counterpart class. The importance of a balanced training sample between class categories to avoid learning bias constrains the training sample generation.

Table 5. The Best Models Across Stages

Act	Model	Score	Stage
1	XGBoost	48%	2b
2	RandForest	57%	2a
3	LogReg	39%	2b
4	RandForest	60%	1
5	XGBoost	89%	2a
6	XGBoost	22%	2b
7	LogReg	28%	2b
8	AdaBoost	67%	2b
9	CatBoost	48%	2a

5.3. *NLS Action Reliability Monitoring*

The developed action recognition system is applied to 3 NLS recording videos to see how this system performs in a full action video. It results in an average F1 score of 57% and an average IoU of 51% across videos. The IoU indicator is also used for the evaluation because, in the end, the purpose of the action recognition is to accurately segment the videos into blocks of actions.

Based on this result and the action segment resulting from the current action recognition system, it can be concluded that the system still needs further improvement. The prediction of action segments is still very noisy, and confusion frequently happens among actions, such as between inflation breath, ventilation, and unknown action category relating to free flow oxygen support. These 3 actions are visually similar and only differentiated by how the airflow is controlled through the relief valve.

However, what is still considered useful in this current system is its ability to appropriately predict the presence of the wet towel removal step which is always performed at the beginning of the resuscitation procedure in all test set videos. Moreover, the predicted action segment of this action category always overlaps the ground truth. Based on the test set videos, it is observed that the model still makes false positive signals in the absence of this towel removal step in the later segments of an NLS video, but it never fails to correctly predict the existence of this important action which is part of the initial thermal care in the NLS procedure.

Based on the observation of (Yamada, Yaeger, and Halamek 2015), the late removal of wet linen/towel is one of the most common errors found. Failure or late removal of this wet towel will cause difficulty in maintaining the normal body temperature of the baby. Unidentified wet towel removal in the first few minutes of the resuscitation before administering the respiratory support may give a signal of inappropriate and risky NLS protocol. Only looking at how this initial thermal care is performed; the current action recognition system can appropriately inform 100% adherence of these 3 recorded NLS procedures to the NLS guideline. Such a system will be useful to inform the likelihood of errors or deviations of the NLS procedure that can further be used for reliability modeling and analysis of this procedure, such as the ones in (Tan et al. 2023), which is the next step of this study.

6. Conclusion and Future Works

In this research, an action recognition system is developed to automatically recognize variations in the NLS procedure. A combination of deep and traditional machine learning approaches, as well as the inclusion of action associations and the preliminary prediction of a targeted action has been proven to be beneficial to improve the system's performance. In the context of the initial thermal care procedure of wet towel removal, the current action recognition system has shown its potential to automatically evaluate how the NLS procedure is performed. However, further improvements are needed.

Future work can consider the time dimension of an action. It means the action recognition task can be initiated by learning whether an action is likely to happen in the initial or in the later stages of the procedure. It can be followed by a more detailed relevant action class prediction. Integration of this automated variation recognition system into the reliability modeling and analysis is the next step of this research.

Acknowledgment

I would like to thank Dr. Thomas Smith (Blueskeye AI) for sharing his research and data. Alfian Tan receives funding from the Indonesia Endowment Funds for Education (LPDP).

References

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94.

Dorogush, Anna Veronika, Vasily Ershov, and Andrey Gulin. 2018. "CatBoost: Gradient Boosting with Categorical Features Support." *ArXiv Preprint ArXiv:1810.11363*.

Fawke, Joe, Sean Ainsworth, Adam Benson Clarke, T'ng Chang, Andy Coleman, Karen Cooper, Jonathan Cusack, et al. 2021. *Newborn Life Support*. Edited by Sean Ainsworth and Joe Fawke. 5th ed. London, England: Resuscitation Council UK.

Finer, N, and W Rich. 2010. "Neonatal Resuscitation for the Preterm Infant: Evidence versus Practice." *Journal of Perinatology* 30 (1): S57–66. https://doi.org/10.1038/jp.2010.115.

Friedman, Jerome H. 2002. "Stochastic Gradient Boosting." *Computational Statistics & Data Analysis* 38 (4): 367–78. https://doi.org/10.1016/S0167-9473(01)00065-2.

Hastie, Trevor, Saharon Rosset, Ji Zhu, and Hui Zou. 2009. "Multi-Class Adaboost." *Statistics and Its Interface* 2 (3): 349–60.

Henry, Caroline, Lara Shipley, Carole Ward, Siavash Mirahmadi, Chong Liu, Steve Morgan, John Crowe, James Carpenter, Barrie Hayes-Gill, and Don Sharkey. 2021. "Accurate Neonatal Heart Rate Monitoring Using a New Wireless, Cap Mounted Device." *Acta Paediatrica* 110 (1): 72–78. https://doi.org/10.1111/apa.15303.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems* 30.

Kong, Yu, and Yun Fu. 2022. "Human Action Recognition and Prediction: A Survey." *International Journal of Computer Vision* 130 (5): 1366–1401.

Kwok, T'ng Chang, Caroline Henry, Sina Saffaran, Marisse Meeus, Declan Bates, David Van Laere, Geraldine Boylan, James P Boardman, and Don Sharkey. 2022. "Application and Potential of Artificial Intelligence in Neonatal Medicine." *Seminars in Fetal and Neonatal Medicine* 27 (5): 101346. https://doi.org/10.1016/j.siny.2022.101346.

Lin, T -Y., P Goyal, R Girshick, K He, and P Dollár. 2020. "Focal Loss for Dense Object Detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2): 318–27. https://doi.org/10.1109/TPAMI.2018.2858826.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. "Microsoft COCO: Common Objects in Context." In *Computer Vision – ECCV 2014*, edited by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, 740–55. Cham: Springer International Publishing.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by Nassir Navab, Joachim Hornegger, William M Wells, and Alejandro F Frangi, 234–41. Cham: Springer International Publishing.

Sawyer, Taylor, Henry C Lee, and Khalid Aziz. 2018. "Anticipation and Preparation for Every Delivery Room Resuscitation." *Seminars in Fetal & Neonatal Medicine* 23 (5): 312–20. https://doi.org/10.1016/j.siny.2018.06.004.

Smith, Thomas J, Michel Valstar, Don Sharkey, and John Crowe. 2019. "Clinical Scene Segmentation with Tiny Datasets." In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0.

Tan, Alfian, Rasa Remenyte-Prescott, Michel Valstar, and Don Sharkey. 2023. "A Modelling Approach to Studying Variations in Newborn Life Support Procedure." *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, May, 1748006X231173595. https://doi.org/10.1177/1748006X231173595.

Wagner, Johannes, Tobias Baur, Yue Zhang, Michel F Valstar, Björn Schuller, and Elisabeth André. 2018. "Applying Cooperative Machine Learning to Speed up the Annotation of Social Signals in Large Multi-Modal Corpora." *ArXiv Preprint ArXiv:1802.02565*.

Wang, T, Y Chen, M Zhang, J Chen, and H Snoussi. 2017. "Internal Transfer Learning for Improving Performance in Human Action Recognition for Small Datasets." *IEEE Access* 5: 17627–33. https://doi.org/10.1109/ACCESS.2017.2746095.

Yamada, Nicole K, Kimberly A Yaeger, and Louis P Halamek. 2015. "Analysis and Classification of Errors Made by Teams during Neonatal Resuscitation." *Resuscitation* 96: 109–13.