sensors

IMPACT FACTOR 3.4

Indexed in: PubMed

CITESCORE 7.3

Article

Depth-Based Intervention Detection in the Neonatal Intensive Care Unit Using Vision Transformers

Zein Hajj-Ali, Yasmina Souley Dosso, Kim Greenwood, JoAnn Harrold and James R. Green

Special Issue
Machine Learning and Image-Based Smart Sensing and Applications

Edited by
Prof. Dr. Ran-Zan Wang and Dr. Shang-Kuan Chen

MDPI

https://doi.org/10.3390/s24237753

sensors

MDPI

Article

Depth-Based Intervention Detection in the Neonatal Intensive Care Unit Using Vision Transformers

Zein Hajj-Ali [1,*], Yasmina Souley Dosso [1], Kim Greenwood [2], JoAnn Harrold [3] and James R. Green [1]

[1] Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada; yasminasouleydosso@cmail.carleton.ca (Y.S.D.); jrgreen@sce.carleton.ca (J.R.G.)
[2] Clinical Engineering, Children's Hospital of Eastern Ontario, Ottawa, ON K1H 8L1, Canada; greenwood@cheo.on.ca
[3] Neonatology, Children's Hospital of Eastern Ontario, Ottawa, ON K1H 8L1, Canada; jharrold@cheo.on.ca
* Correspondence: zeinhajjali@sce.carleton.ca

Abstract: Depth cameras can provide an effective, noncontact, and privacy-preserving means to monitor patients in the Neonatal Intensive Care Unit (NICU). Clinical interventions and routine care events can disrupt video-based patient monitoring. Automatically detecting these periods can decrease the time required for hand-annotating recordings, which is needed for system development. Moreover, the automatic detection can be used in the future for real-time or retrospective intervention event classification. An intervention detection method based solely on depth data was developed using a vision transformer (ViT) model utilizing real-world data from patients in the NICU. Multiple design parameters were investigated, including encoding of depth data and perspective transform to account for nonoptimal camera placement. The best-performing model utilized ∼85 M trainable parameters, leveraged both perspective transform and HHA (Horizontal disparity, Height above ground, and Angle with gravity) encoding, and achieved a sensitivity of 85.6%, a precision of 89.8%, and an F1-Score of 87.6%.

Keywords: depth camera; neonatal patient monitoring; NICU; transformer; vision transformer; ViT; intervention detection

check for updates

Citation: Hajj-Ali, Z.; Dosso, Y.S.; Greenwood, K.; Harrold, J.; Green, J.R. Depth-Based Intervention Detection in the Neonatal Intensive Care Unit Using Vision Transformers. *Sensors* **2024**, *24*, 7753. https://doi.org/10.3390/s24237753

Academic Editors: Ran-Zan Wang and Shang-Kuan Chen

Received: 26 August 2024
Revised: 26 November 2024
Accepted: 2 December 2024
Published: 4 December 2024

1. Introduction

The Neonatal Intensive Care Unit (NICU) provides critical care for the most vulnerable newborn patients. Such patients are characterized by precarious health and require continuous monitoring. Such continuous monitoring in the NICU typically involves sensors attached to the patient's skin, which are susceptible to motion artifacts and may interfere with both clinical and parental care. The wired sensors can irritate sensitive skin, with frequent removal and reapplication sometimes required during medical interventions. This motivates the development of robust video-based noncontact patient monitoring [1–3].

A patient may experience multiple periods of clinical intervention or routine care throughout their time in the NICU. These interventions can include a clinician or parent reaching into the scene to replace sensors, take readings, change a diaper, feed the patient, or otherwise move the patient. These periods of intervention are often excluded from analysis when studying novel noncontact techniques of monitoring neonates in the NICU (e.g., [4,5]). However, studies by Villarroel et al. [3] and Souley Dosso et al. [2,6] attempt to detect these periods of intervention and, in the case of [6], classify a subset of them (bottle-feeding interventions).

Deep learning has led to dramatic advancements in computer vision, which have translated into new forms of noncontact patient monitoring [1]. Souley Dosso et al. [2] used the VGG-16 CNN model introduced in ref. [7] as the feature extractor for their method of intervention detection. They examined several forms of multi-modal (RGB and depth) fusion, resulting in similar performance between the RGB and RGB-D fusion models while observing significantly lower performance for depth-only models.

In this paper, we develop a model to detect periods of clinical or routine care intervention using only depth-based images, as this modality is more privacy-preserving than RGB or RGB-D images. Detecting such interventions is useful for several reasons. For example, when estimating vital signs, estimation may be paused or patient monitor alarms may be silenced automatically during interventions since a clinician is already attending to the patient. Detecting interventions is a step towards classifying interventions, which may ultimately lead to automated charting of patient care. Furthermore, by creating an intervention detection system based strictly on depth data, detection will be robust to changes in lighting, which can occur frequently in the NICU due to clinical interventions and parent time and regularly throughout the day for some premature patients to promote sleep and support development. This paper focuses on utilizing depth video alone for intervention detection, building on the preliminary results reported in the following thesis [8]. Note that portions of this manuscript previously appeared in the following thesis [9].

This study leverages vision transformers (ViTs) [10], which have been shown to outperform CNNs for image classification in several application areas. The ViT divides each input image into a number of nonoverlapping patches which are flattened into vectors of pixel values and used as the input to the transformer's encoder. The ViT culminates in a fully connected head layer for the task of image classification. Variations and extensions of this model have had success in image segmentation, object detection, and video action recognition [11–13].

When training a deep learning model, large amounts of data and compute resources are needed. For this reason, transfer learning is usually employed, where models are pre-trained on large datasets prior to fine-tuning the model to perform specific tasks with smaller training datasets. For image classification, several convolutional neural network (CNN) and vision transformer (ViT) models are available that have been pre-trained on the ImageNet dataset [14] consisting of ∼14 million annotated images from 1000 classes. Given a downstream task, pre-trained models are normally chosen from the same or similar domains (e.g., RGB image classification, object detection, semantic segmentation). Transfer learning has been shown to improve the average accuracy of CNN models [15] as well as ViT [10] for image classification.

Pre-trained image classification models are generally trained on large amounts of labelled three-channel RGB data. HHA encoding is a method of encoding depth data using three channels for each pixel rather than just the one channel of depth [16]. An example illustrating the three channels resulting from HHA encoding of a depth image can be seen in Figure 1. The three channels correspond to the horizontal disparity (H), the height above the ground (H), and the angle the pixel's local surface normal makes with the inferred gravity direction (A). This has been shown to improve the performance of a network pre-trained on RGB data and fine-tuned with labelled HHA-encoded depth data when compared to fine-tuning on regular one-channel depth or disparity data. Gupta et al. suggest that this is because the disparity and angle channels may show edges that correspond to object boundaries that can be seen in the RGB images of the same scene [16]. The authors verify this by fine-tuning a CNN originally trained for object detection and semantic segmentation from RGB images [17].

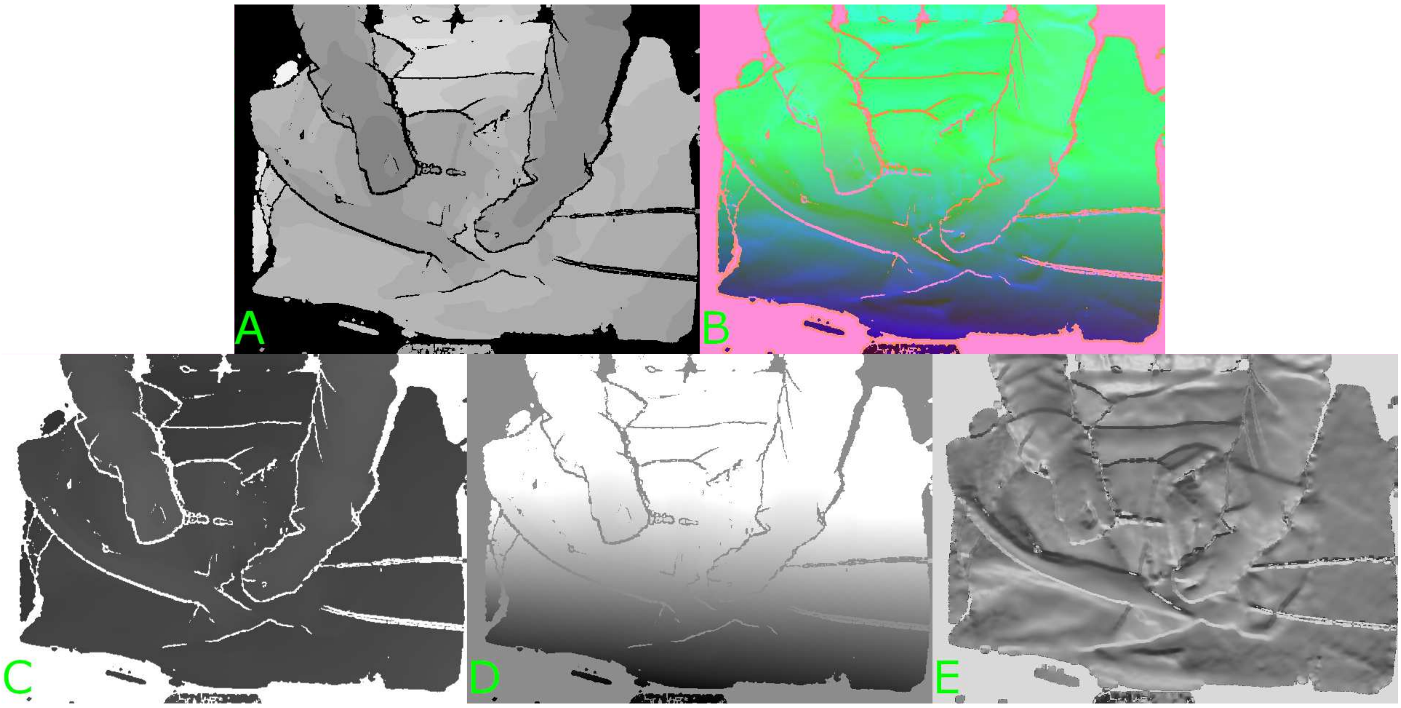

Figure 1. Example of HHA encoding of a depth image. (**A**) Original depth image. (**B**) Three-channel HHA-encoded image. (**C**) First channel (H). (**D**) Second channel (H). (**E**) Third channel (A).

The horizontal disparity can be calculated from depth data by using Equation (1) [18]:

$$Disparity = \frac{Focal\ Length \times Baseline}{Depth} \tag{1}$$

where the *Focal Length* and *Baseline* are found from the camera's intrinsic matrix [19]. The height above the ground and the angle between the surface normal and inferred gravity direction can be found using the algorithms presented in ref. [20] and implemented in refs. [21,22]. The algorithms require the point cloud representation of the depth image as well as the camera matrix. The direction of gravity is estimated by finding the direction that is best aligned to surface normals, under the assumption that most surfaces in the scene are horizontal. The direction of gravity is initialized to the camera's Y-axis before iteratively refining the estimated direction by examining local surface normals in the depth data. The height above the ground can then be found by rotating the point cloud of the data to the horizontal direction then subtracting the smallest Y-coordinate value in the scene [23]. The angle between the surface normal and the gravity direction can be found from the difference in the respective vectors. Finally, the values in each of the channels are normalized to the range of 0–255 (i.e., an 8-bit value).

Despite the notable advancements in noncontact monitoring of patients in the NICU, there remains a critical gap in the literature concerning the automatic detection of clinical interventions and routine care events, particularly using depth data. Many existing studies [2,3,6] have leveraged RGB (colour) or multi-modal RGB-D image data for such detection. Chaichulee et al. achieved excellent accuracy when detecting clinical interventions using RGB video [24]. However, RGB (and RGB-D) video may be considered intrusive and is sensitive to ambient lighting. Therefore, in this study, we specifically focus on models restricted to privacy-preserving depth images rather than RGB (or RGB-D).

The contributions of this study are as follows: First, we propose an intervention detection method based solely on depth data, thereby increasing robustness to lighting changes and maintaining patient privacy. The method utilizes a vision transformer (ViT) model to interpret the depth data, an approach not previously explored for this application in the NICU setting. We also investigate several design parameters such as the encoding of depth data and the application of perspective transform to account for varying camera placement, hence offering a versatile solution suitable for different NICU environments. Finally, we evaluate our model using real-world NICU data, demonstrating its practical utility and efficacy. Utilizing real-world data for the evaluation is critical in this case, as a simulated environment may not accurately reflect the range of challenges that arise in a complex clinical environment. An example of several challenging scenarios can be seen in Figure 2. Our results not only confirm the feasibility of the proposed approach but also set the stage for future work in automatic classification of interventions and eventually automated charting of patient care.

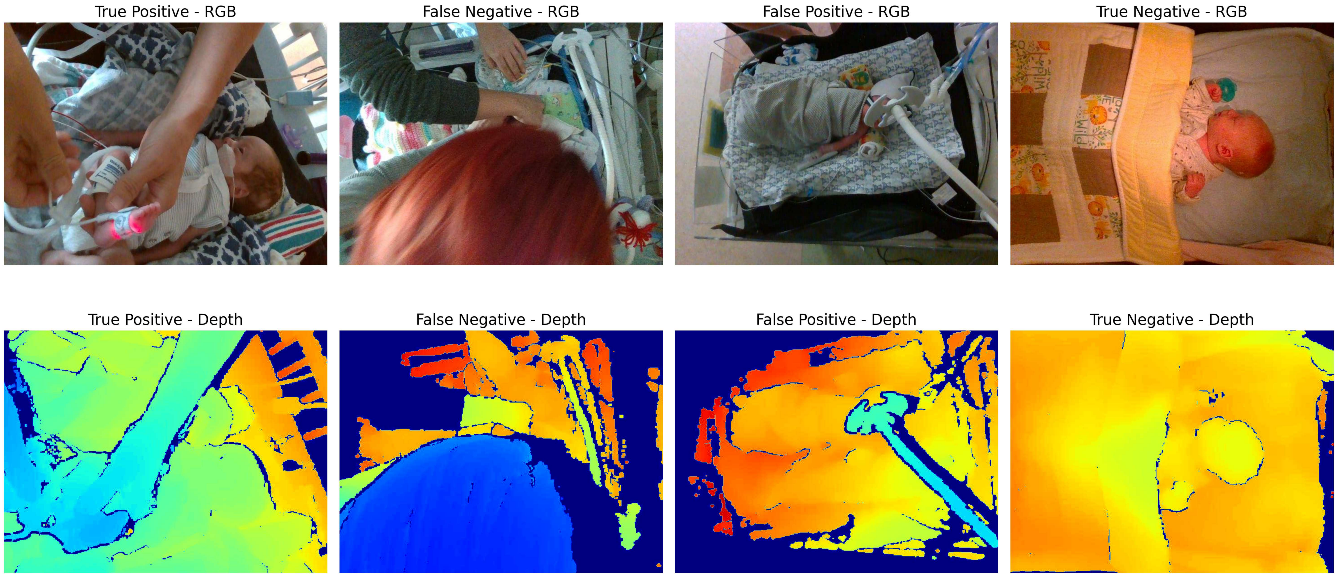

Figure 2. Examples of scenarios where different types of models may underperform. True positive: Image is clear, with distinction between patient and clinician's hands. False negative: Occlusion of a large portion of the scene by the clinician's head may confuse depth-based methods (since the clinician's hands appear to be farther away than the main area of the scene). False positive: Ventilator equipment may be confused for the arm of the clinician performing an intervention. True negative: Patient in crib with no occlusions, angle of camera is top-down which may simplify intervention classification. Note that RGB images are shown here for illustration purposes only; intervention detection models require only depth images.

2. Materials and Methods

2.1. Data Collection

To support our study, we collected two types of data: clinical data from neonatal patients and simulated data from a neonatal manikin. In the following subsections, we describe the data collection process for each dataset, including details on the data collection setup, data processing, and class labelling.

2.1.1. NICU Data Collection

Data were collected from 27 neonatal patients in the NICU at the Children's Hospital of Eastern Ontario (CHEO) following approval by the research ethics boards from the hospital and Carleton University. The data were collected as part of a larger research initiative to develop multi-modal noncontact patient monitoring methods and technologies. The dataset cannot be released publicly due to the restrictions set by the research ethics board.

Figure 3 shows an example of the setup in the NICU environment. An RGB-D camera (Intel RealSense SR300, Santa Clara, CA, USA) was placed above or beside the patient's bed. The camera was chosen due to its small size, affordability, and suitable depth range to capture patients at a close distance. Recordings were captured at a resolution of 640 × 480 pixels at 30 frames per second. The cameras were placed such that the view planes were at nonuniform angles relative to the plane of the bed. The SR300 captures depth information using the coded-light method; using a combination of an IR projector and IR camera sensor to generate a depth pixel frame. The camera also includes a separate RGB camera sensor that can be used in conjunction with the depth stream to form an RGB-D image. Note that in the present study, all proposed methods use only privacy-preserving depth image data.

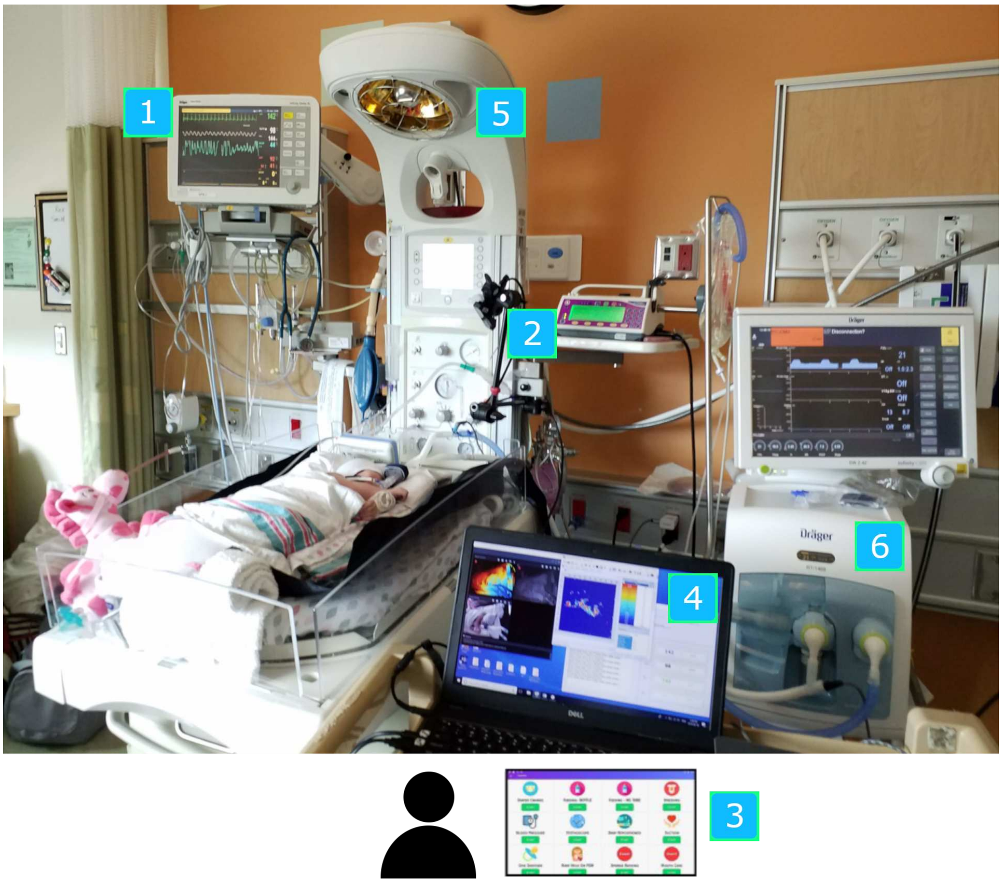

Figure 3. Overview of equipment setup: 1. Patient monitor. 2. RGB-D camera. 3. Bedside annotation application. 4. Data acquisition laptop. 5. Neonatal bed (overhead warmer). 6. Ventilator.

The gold standard respiratory rate signals of the patients were recorded from the bedside patient monitor (Draegar Infinity Delta). Custom Patient Monitor Data Import (PMDI) software Version 1.0, developed for the project, was used to import the data from the serial port on the monitor [25]. A bedside annotation application was used to annotate events (clinical interventions, etc.) in real time. All data from the camera and patient monitor were saved on a data acquisition laptop.

Still images were extracted from the patient recordings every 30 s and labelled as either 'Intervention' (positive) or 'No Intervention' (negative). This resulted in 14,892 images in total, with 1260 in the positive class and 13,632 in the negative class (a class imbalance of 10.8:1 in favour of the negative class). The 'Intervention' class comprised images where a nurse or other practitioner was reaching into the camera's view to tend to the patient, while the 'No Intervention' class included only the patient (Figure 4).

The difficulty of intervention detection from depth data can sometimes be misrepresented. Looking at Figure 4, one would assume that the difference in the depth frame between the nurse's hands and the patient/bed would be apparent; however, the task is often more difficult. In Figure 5, an intervention frame can be seen that is more challenging to classify by looking only at the depth channel (on the right). If the caregiver's hands are near or at the height of the patient's bed, the difference in depth can be sufficiently small to require more advanced methods. This is demonstrated by including a baseline approach in the present study.

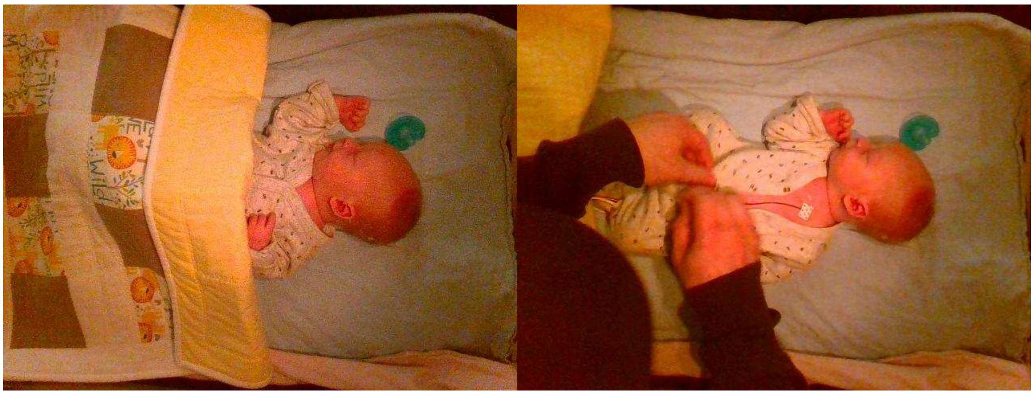

Figure 4. Example frames of 'No Intervention' (**left**) and 'Intervention' (**right**).

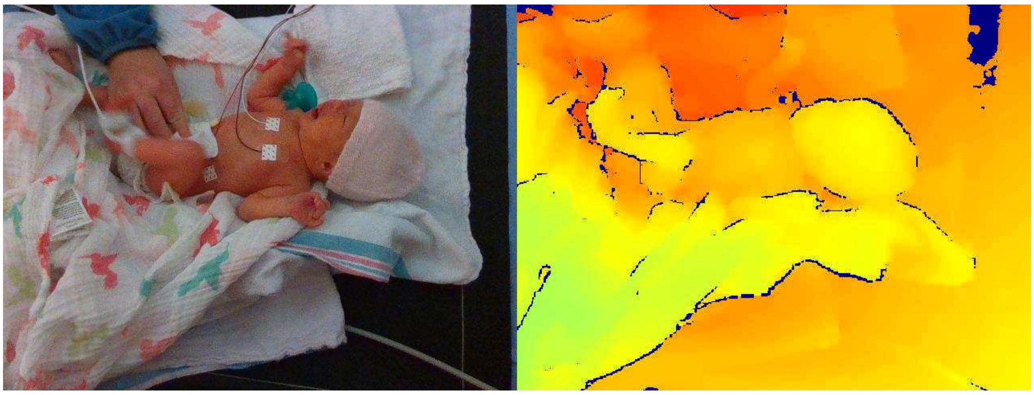

Figure 5. Example of more difficult 'Intervention' class frame with both RGB image (**left**) and corresponding depth frame (**right**).

2.1.2. Simulated Data Collection

After the initial data collection stage, additional simulated data were collected to partially address the class imbalance between nonintervention/intervention frames in the clinical data. A neonatal manikin (StandInBaby [26]) was placed on simulated clinical bedding, and the RealSense SR300 RGB-D camera was used to capture 600 depth images, as illustrated in Figure 6. A camera arm was used to place the camera at 5 different angles relative to the plane of the bed. Yellow gloves were worn during data collection to facilitate the use of the collected data in image segmentation studies in future studies by providing a consistent colour reference for the hands (Figure 7).

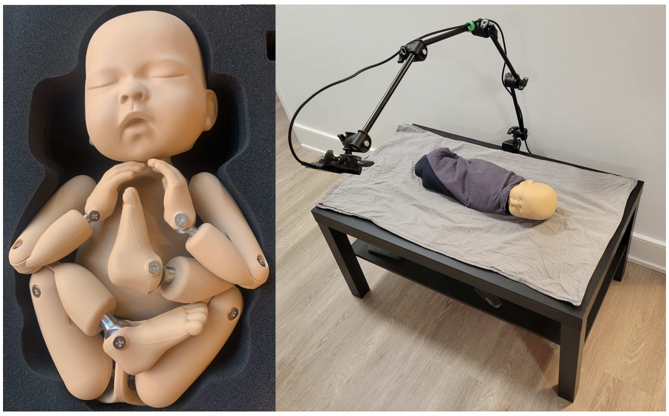

Figure 6. StandInBaby neonatal manikin on the left; example simulated data collection scene on the right.

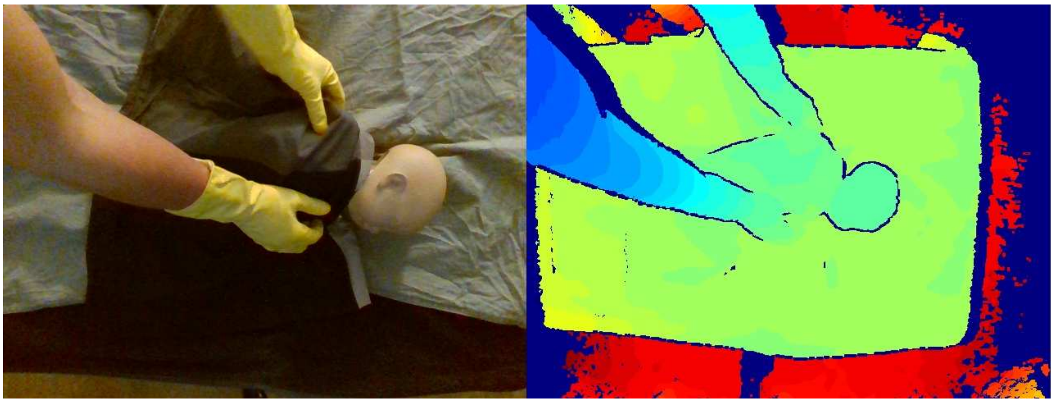

Figure 7. Example of simulated data in RGB (**left**) and colour-mapped depth image (**right**).

2.2. Proposed Method

Vision transformers have demonstrated excellent accuracy in image classification tasks since their introduction in ref. [10]. We propose the use of a ViT pre-trained on the ImageNet dataset [14] and fine-tuned on a subset of our own set of 14,892 depth images. The model architectures were implemented using the PyTorch Image Models library [27]. Two model sizes with similar architecture but different numbers of trainable parameters were chosen, 'vit_tiny_patch_16_224' and 'vit_base_patch_16_224', with ∼5.4 M parameters and ∼85 M parameters, respectively. Each of the models accepts input images with a resolution of 224 × 224 pixels and divides them into 16 × 16 patches for embedding. The difference in the number of trainable parameters comes from an increase in the dimensions of the hidden embedding layer and the number of heads in the attention mechanism when moving from the 'tiny' model to the 'base' model.

Performance evaluation: Throughout this study, we have used a repeated five-fold cross-validation approach, where the dataset of 27 patients was divided into five distinct "folds". For each combination of system design parameters, five models were trained and evaluated, and the average performance across the five models was computed. Within each of the five folds, a classification model was trained on four folds (approximately 21 patients), while the remaining 5–6 patients *not used to train the model* were used to evaluate the model. In this way, all patient data were used to both train and evaluate models but never the same model. This entire process was repeated five times, with different patients assigned to each fold in each repetition. The mean across the five repetitions was reported as the final performance metric.

Hyperparameters: The training of the models utilized a mini-batch size of 16 and a learning rate set at 0.01 over a maximum of 15 epochs. Visual inspection of preliminary learning curves indicated no substantial reduction in validation loss beyond 15 epochs.

Stochastic gradient descent was employed as the optimizer, with a momentum of 0.9. The input images were each resized to dimensions of 224 × 224 pixels, which altered their aspect ratio from 4:3 to 1:1. Finally, random rotations (ranging from 0° to 360°) and horizontal/vertical flips were applied to the images in the training sets.

Along with the size of the model, the effect of three other system design parameters on model performance was also explored. These parameters are described in the upcoming Sections 2.2.1–2.2.3, and a summary can be seen in Table 1. A visual representation of the data flow and utilization of the proposed system design parameters can be seen in Figure 8.

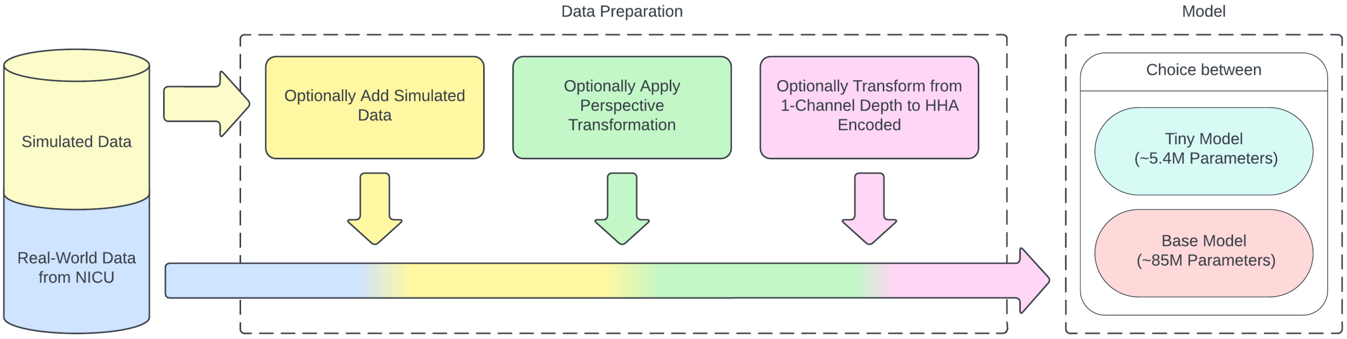
Data Preparation
Model
Simulated Data
Real-World Data from NICU
Optionally Add Simulated Data
Optionally Apply Perspective Transformation
Optionally Transform from 1-Channel Depth to HHA Encoded
Choice between
Tiny Model (~5.4M Parameters)
Base Model (~85M Parameters)

Figure 8. Flow of data and addition of proposed system design parameters.

Table 1. Summary of vision transformer experiments.

Experiment	Model Size	Simulated Data	PT	Encoding
1	Tiny	Unused	Unused	1-channel depth
2	Tiny	Unused	Unused	HHA
3	Tiny	Unused	Applied	1-channel depth
4	Tiny	Unused	Applied	HHA
5	Tiny	Added	Unused	1-channel depth
6	Tiny	Added	Unused	HHA
7	Tiny	Added	Applied	1-channel depth
8	Tiny	Added	Applied	HHA
9	Base	Unused	Unused	1-channel depth
10	Base	Unused	Unused	HHA
11	Base	Unused	Applied	1-channel depth
12	Base	Unused	Applied	HHA
13	Base	Added	Unused	1-channel depth
14	Base	Added	Unused	HHA
15	Base	Added	Applied	1-channel depth
16	Base	Added	Applied	HHA

2.2.1. Simulated Data

Since the data collected from the NICU contain more instances without interventions than those with, the resulting labelled data had a high class imbalance of 10.8:1 in favour of the negative (no-intervention) class. To help correct for this imbalance, simulated intervention data were collected as previously described (Section 2.1). These data comprised 600 images of simulated interventions that were added to the positive class, bringing the class imbalance down to approximately 7.3:1. Both model sizes were trained without the addition of the simulated data, and then the process was repeated with the inclusion of the simulated data in each training fold (i.e., simulated data were used for training but not for testing).

2.2.2. Perspective Transformation

The effect of a perspective transformation (PT) algorithm on the performance of the models was explored. As discussed in ref. [5], we previously demonstrated that perspective transformation can account for nonoptimal depth camera placement relative to the patient. In that study, perspective transform was shown to improve an ROI selection algorithm for subsequent respiration rate estimation. Based on those results, it was thought that applying the transformation to the data used to build the ViT-based intervention detection model might also improve its performance. The patient data collected from the NICU and the simulated data were transformed by manually selecting four registration points in the plane of the bedding for each new recording. The rotation matrix was found and applied to all frames extracted from the same recording. The experiments were then re-run using these transformed data as the input. Models were trained and tested with and without perspective transform to investigate its effect on intervention detection accuracy.

2.2.3. HHA Encoding

ViTs are not typically trained from scratch for specific image classification tasks. Rather, ViT models are typically pre-trained on large datasets using self-supervised learning techniques, such as masked auto-encoding (MAE) [28]. Pre-trained ViTs are then fine-tuned for specific tasks through the addition of a task-specific prediction head. Such pre-training of ViT requires a large amount of data and extensive compute resources. Some ViT models pre-trained on large image datasets, such as ImageNet, have been released publicly by researchers at Google Research [29] and other groups. As these models have been pre-trained on 3-channel RGB images, there is latitude as to how the single channel of depth data should be mapped to a 3-channel input. The effect of HHA encoding on the performance of the proposed intervention detection model was investigated.

Each of the datasets described previously was transformed to be HHA-encoded, and the experiments were re-run. Models were trained with and without HHA encoding to investigate its effect on intervention detection accuracy. Models trained without HHA encoding were modified to accept 1-channel images as inputs. The pre-trained input layer weights from each of the 3 channels normally used for R, G, and B were summed into a single channel.

2.3. Baseline Methods

The models explored in this study were compared against the best-performing CNN-based intervention detection model proposed by Souley Dosso et al. in ref. [2]. Specifically, the model chosen for comparison was the multi-modal RGB-D fusion model, which used a VGG-16 CNN architecture [7] and was pre-trained on the ImageNet dataset [14] and fine-tuned on the intervention detection dataset described in Section 2.1.

Additionally, the exclusively depth-based model from Souley Dosso's study was included for comparison given its shared reliance on depth modality, though it resulted in lower performance metrics overall. For this model, the VGG-16 input layer was modified by removing two of its three input channels, allowing the pre-trained weights to be fine-tuned on the single depth channel. Further, a conventional (rules-based) method was also evaluated as an alternative baseline for comparison. The method consists of designating a known nonintervention frame for each patient recording and calculating the mean squared error of each of the rest of the frames.

As a final baseline model for comparison with our depth-based models, the RGB-D and depth-based models presented in ref. [2] were also re-trained and evaluated using the design parameters outlined in Sections 2.2.1–2.2.3. This enabled direct comparisons between the depth-based vision transformer models proposed here and Souley Dosso's depth-based CNN models for each of the design variables explored in this study.

3. Results

Each of the models was evaluated using 5-fold cross validation repeated five times. Each fold contained data from unique patients, leaving data from five or six patients as the test set each time. The frames were extracted at the same time points in the videos as the data used in ref. [2] to enable direct performance comparisons against the chosen baseline models. For experiments where simulated data were used, the simulated frames were added to the training set in each fold. The metrics used to evaluate the models were specificity, sensitivity, precision, accuracy, F1-score, and Matthew's correlation coefficient (MCC) (2)–(7). Analysis of Variance (ANOVA) was run on the results from the proposed models to determine the statistical significance of the effects for each of the design parameters. This was performed by collapsing the results of the repetitions of each fold by calculating the average of each metric before running the ANOVA test. This meant that the number of records was reduced from 400 (5 folds × 5 repetitions × 16 combinations of variables) down to 80 (averages of the repetitions of the 5 folds × 16 combinations of variables). The full ANOVA test results can be found in Appendix A.

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$F1\text{-}Score = \frac{2TP}{2TP + FP + FN} \tag{6}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

3.1. Baseline Model Performance Evaluation

Each of the baseline models described in Section 2.3 was assessed using 5-fold cross validation. The multi-modal RGB-D fusion model by Souley Dosso et al. achieved high average sensitivity, specificity, and accuracy, consistently outperforming the exclusively depth-based baseline model across all performance metrics. The cross-validation splits were held constant across models for direct comparison. The rules-based baseline was evaluated by fitting a logistic regression model on the data using the same 5-fold cross-validation splits. The ROC curve of this method can be seen in Figure 9. Table 2 shows a summary of the metrics of the relevant comparison models reproduced from [2].

Table 2. Summary of results from baseline CNN models used for comparison.

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
RGB-D fusion	**95.70 %**	**84.25%**	**64.54%**	**94.73%**	**73.06%**	**70.98%**
Depth-based	89.25%	66.11%	36.24%	87.29%	46.82%	42.64%
Rules-based	34.97%	43.08%	5.56%	35.63%	9.84%	−12.46%

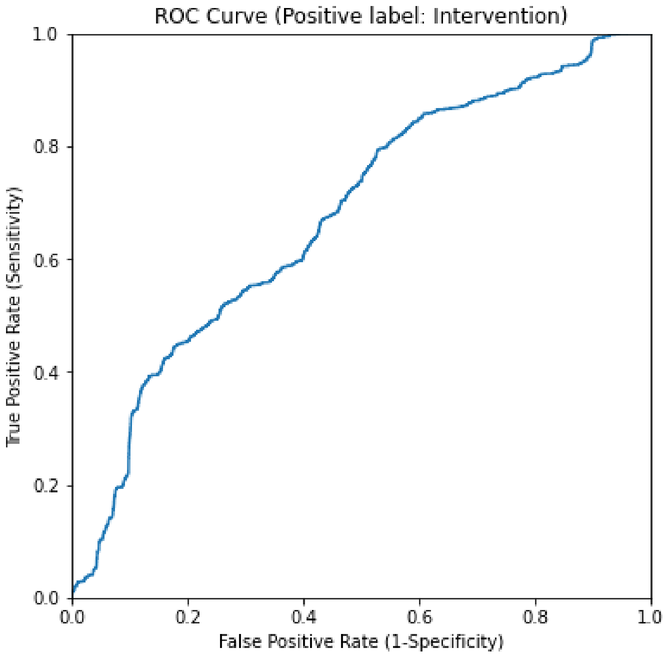

Figure 9. ROC curve for rules-based baseline method.

3.2. Comparison Between Baseline Models and Proposed Model

Initially, we compared the results of the depth-based ViT models to those of the baseline depth and RGB-D fusion models. The 'tiny' ViT model showed an improvement over all tested metrics except sensitivity, where it showed a slight decrease. The 'base' ViT model showed a further improvement over all metrics. Results are summarized in Table 3 and Figure 10. To determine whether the improvement in results observed when moving from the 'tiny' model to the 'base' model is statistically significant, we performed ANOVA over each of the performance metrics. A p-value of less than 0.05 indicated a statistically significant difference in the MCC score when changing the model size.

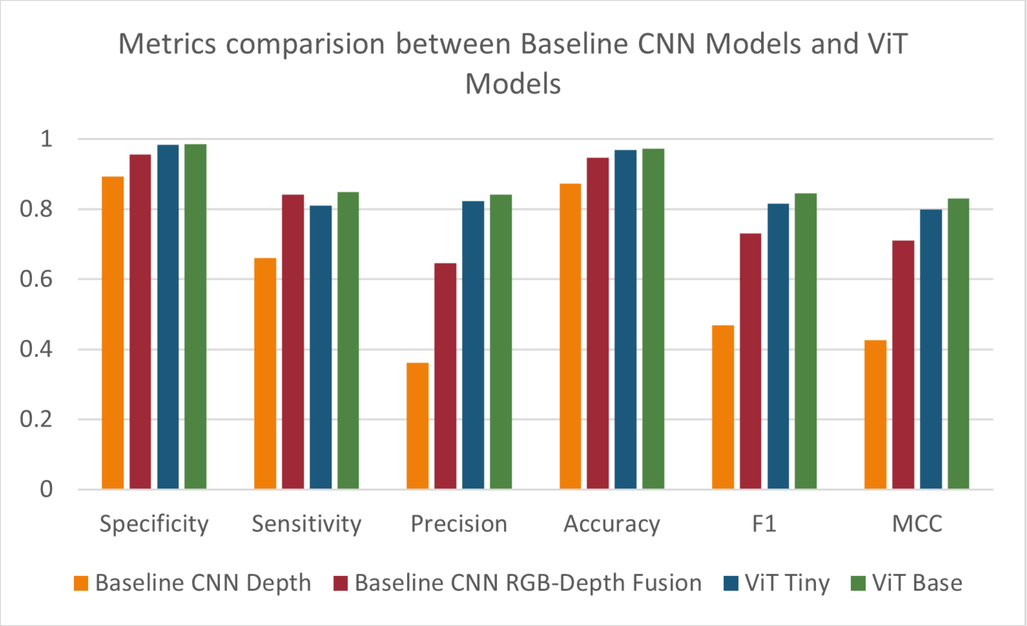

Figure 10. Specificity, sensitivity, precision, accuracy, F1-score, and MCC for baseline models, ViT Tiny, and ViT Base.

Table 3. Summary of results from 'tiny' and 'base' vision transformer models.

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
ViT Tiny	98.33%	81.10%	82.29%	96.84%	81.61%	79.93%
ViT Base	**98.50%**	**84.95%**	**84.20%**	**97.35%**	**84.47%**	**83.09%**

3.3. Effect of Simulated Data on Model Performance

After observing the improved performance of the ViT models on the same data as the baseline models, the tests were repeated with the addition of simulated data into the training folds. These results are shown in Table 4 and Figure 11. Relative to the results in Table 3, the performance decreased with the addition of the simulated data over all metrics except sensitivity. The ANOVA test did not find a statistically significant difference in the results when utilizing the simulated data. This outcome was unexpected, as the addition of the simulated data partially addressed the class imbalance in the training dataset. The decrease in performance could be attributed to domain differences between the simulated and clinical data collection environments or to the difference in class imbalance between the training (7.3:1) and test (10.8:1) datasets. Although efforts were made to recreate the environment when collecting the simulated data, many factors could contribute to the resulting performance, like differences in lighting or mismatch of camera angles between the data collected from the NICU and the simulated data. It is also possible that a larger and more diverse set of simulated data may have a beneficial impact on the models. When looking at the results of the depth-based baseline CNN model, the addition of simulated data showed an increase in specificity and accuracy and a decrease in all other metrics compared to the original depth-based baseline CNN model.

Table 4. Summary of results from 'tiny' vision transformer, 'base' vision transformer, and depth-based CNN models with simulated data.

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
ViT Tiny Simulated	97.92%	80.32%	**78.43%**	96.43%	79.28%	77.39%
ViT Base Simulated	97.70%	**86.38%**	77.96%	**96.76%**	**81.78%**	**80.24%**
Depth-based CNN Simulated	**98.07%**	39.19%	65.45%	93.09%	48.94%	47.26%

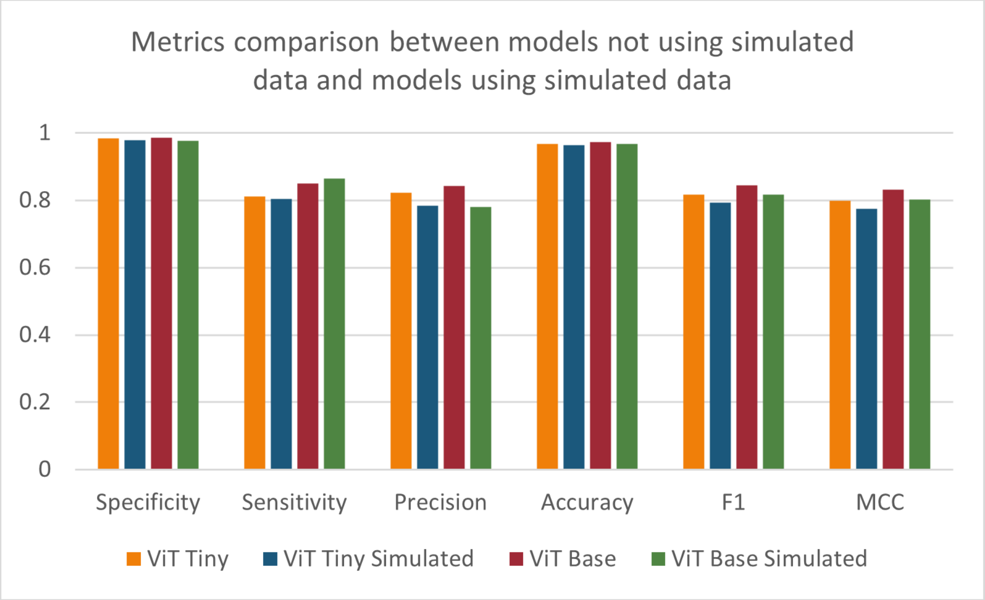

Figure 11. Specificity, sensitivity, precision, accuracy, F1-score, and MCC for ViT models trained with and without supplemental simulated data.

3.4. Effect of Perspective Transformation on Model Performance

When repeating the cross validation after applying the perspective transformation process, no pattern of significant increases or decreases in performance could be found (see Table 5 and Figure 12). The ANOVA test did not find any statistically significant effect resulting from pre-processing the images using perspective transformation. The depth-based baseline CNN model showed improvements in most metrics except sensitivity. The difference in the trend of results between the ViT models and the depth-based baseline CNN models may be due to the way each architecture handles images. A CNN uses convolutional operations to learn the patterns of edges and corners in an image, and these features may be enhanced when the perspective of the image is altered. Vision transformers may not benefit in the same way from the enhancement of these features due to the way the ViT splits the input image into patches that are then encoded.

Table 5. Summary of results from 'tiny' vision transformer, 'base' vision transformer, and depth-based CNN models with perspective-transformed data.

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
ViT Tiny PT	**98.83%**	76.27%	**85.90%**	96.92%	80.72%	79.26%
ViT Base PT	98.72%	**82.16%**	85.66%	**97.32%**	**83.81%**	**82.41%**
Depth-based CNN PT	93.21%	56.48%	44.05%	90.10%	49.17%	44.42%

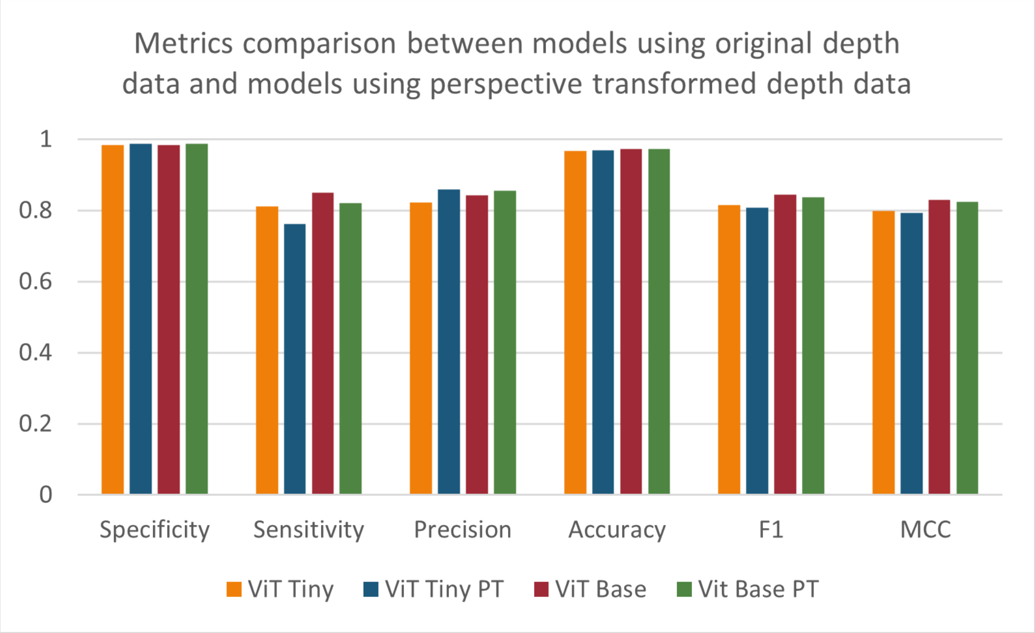

Figure 12. Specificity, sensitivity, precision, accuracy, F1-score, and MCC for models using original depth data and models using perspective-transformed data.

3.5. Effect of HHA Encoding on Model Performance

As seen in Figure 13, when comparing the performance of the models using the HHA-encoded depth data against that of the models using the original one-channel depth data, a decrease across all metrics can be seen for the larger-sized 'base' vision transformer. However, the smaller 'tiny' vision transformer model was shown to improve its specificity, precision, accuracy, and MCC scores, with a stagnant F1-score and a slight decrease in its sensitivity (Table 6). The effect of HHA encoding of the data used to train and evaluate the models was found to be statistically significant for the precision and MCC metrics when applying the ANOVA test. The improvement in the model's performance was expected, as the model was pre-trained on three-channel RGB images before transferring the weights. Although the depth-based baseline CNN model was pre-trained on the same dataset as the vision transformer models, there was a decrease in the metrics most relevant to the imbalanced dataset being investigated. The depth-based baseline CNN model using HHA-encoded data showed improvements to specificity, accuracy, and precision and a detrimental effect on sensitivity, F1-score, and MCC. These results were surprising since previous studies, such as Gupta et al. [16], have demonstrated that HHA-encoded depth images generally increase the effectiveness of similarly pre-trained CNNs. The deviations observed might stem from the intricate nature of the scenes and the specific conditions of the NICU setting.

Table 6. Summary of results from 'tiny' vision transformer, 'base' vision transformer, and depth-based CNN models with HHA-encoded data.

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
ViT Tiny HHA	**98.68%**	80.62%	82.96%	**97.39%**	81.60%	80.30%
ViT Base HHA	98.64%	**83.59%**	**85.08%**	97.37%	**84.25%**	**82.86%**
Depth-based CNN HHA	96.81%	18.95%	40.42%	90.22%	24.36%	22.14%

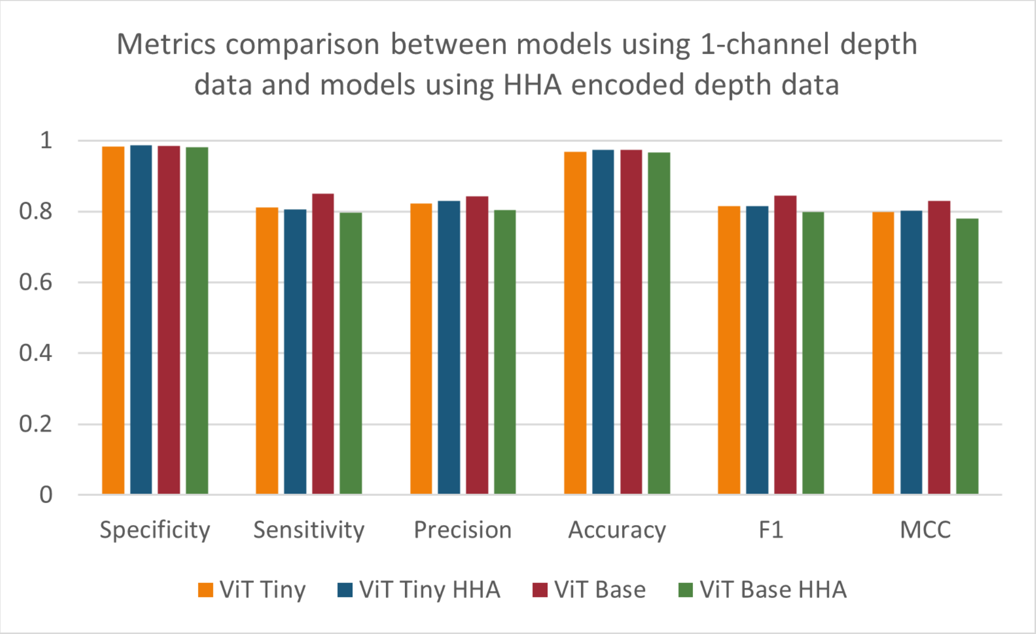

Figure 13. Specificity, sensitivity, precision, accuracy, F1-score, and MCC for models using 1-channel depth data and models using HHA-encoded depth data.

3.6. Effect of Multiple Variables on Model Performance

The previous four sections outlined the four design parameters applied to the models separately (i.e., model size, simulated data, perspective transform, and HHA encoding). All combinations of the variables were then tested to evaluate their performance and determine the ideal model. This resulted in 11 different combinations of variables (in addition to each variable separately). The results of the remaining models not shown previously can be found in Table 7. An n-way ANOVA was conducted, where $n = 4$ is the number of independent variables. Unexpectedly, it can be seen that no combination of variables was found to have a statistically significant effect on the performance of the models. This may be due to certain variables that have a positive and negative effect counteracting each other when acting in conjunction. In addition, the models may be approaching a performance ceiling as the metrics approach a maximum value that can be achieved with the available input information. Figures 14 and 15 display the metrics for each of the models with and without a combination of variables. The effects of combinations of design variables on the performance of the depth-based baseline CNN model were also investigated. The results of the remaining baseline CNN models not shown previously can be seen in Table 8. The best-performing baseline CNN model utilizes the simulated data as well as HHA encoding. It shows an improvement across all metrics except sensitivity, where the performance decreased. The confusion matrix for the best-performing model can be found in Table 9.

Table 7. Summary of results from 'tiny' and 'base' vision transformer models with combinations of studied variables.

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
ViT Tiny Simulated PT	97.08%	77.27%	71.91%	95.41%	74.18%	71.92%
ViT Base Simulated PT	98.65%	81.79%	84.93%	97.22%	83.29%	81.82%
ViT Tiny Simulated HHA	97.62%	81.52%	77.50%	96.26%	78.94%	77.24%
ViT Base Simulated HHA	99.04%	78.86%	88.66%	97.34%	83.13%	82.06%
ViT Tiny HHA PT	98.79%	84.16%	86.74%	97.55%	85.38%	84.09%
ViT Base HHA PT	**99.10%**	**85.59%**	**89.76%**	**97.95%**	**87.62%**	**86.54%**
ViT Tiny Simulated HHA PT	98.90%	82.41%	87.39%	97.50%	84.81%	83.51%
ViT Base Simulated HHA PT	98.99%	85.35%	88.64%	97.84%	86.95%	85.80%

Table 8. Summary of results from depth-based baseline CNN models with combinations of studied variables.

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
Depth-based CNN Simulated PT	98.83%	36.59%	74.76%	93.56%	48.88%	49.36%
Depth-based CNN Simulated HHA	98.89%	**45.78%**	**79.24%**	**94.39%**	**57.98%**	**57.63%**
Depth-based CNN HHA PT	98.97%	1.90%	20.00%	90.76%	3.32%	3.05%
Depth-based CNN Simulated HHA PT	**99.15%**	33.35%	78.71%	93.50%	46.74%	48.57%

Table 9. Confusion matrix for ViT Base HHA PT.

	Predicted	
	Positive	**Negative**
Actual positive	1083 (TP)	177 (FN)
Actual negative	110 (FP)	13,522 (TN)

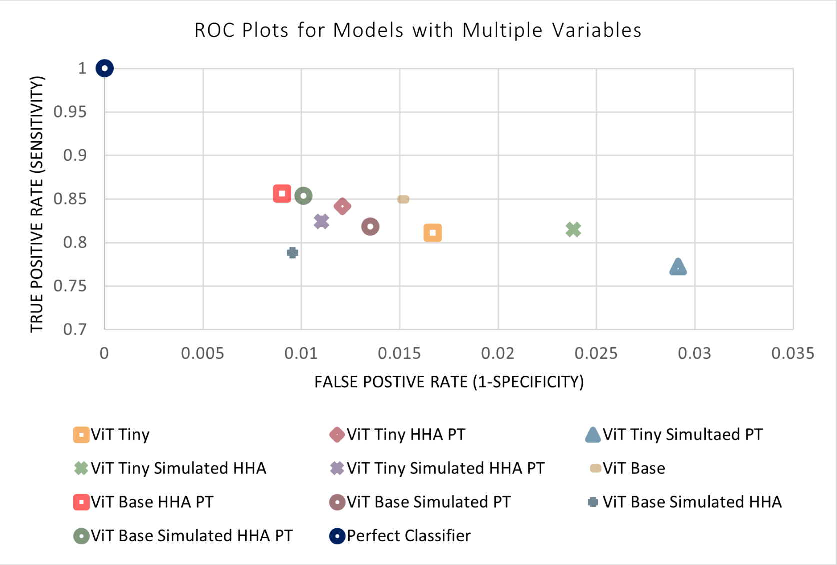

Figure 14. ROC plots of models with a combination of variables.

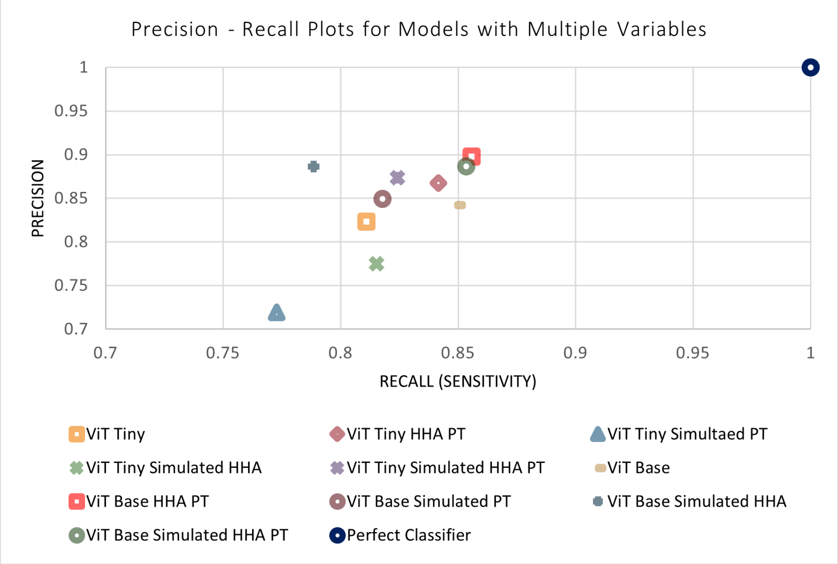

Figure 15. Precision–recall plots of models with a combination of variables.

4. Discussion

Detecting periods of intervention from a recording presents a number of issues depending on the modality used. RGB video suffers from a decrease in performance during periods of lower light or lighting changes. Models utilizing depth data may be tricked by a nurse's hands being the same depth away from the camera as the patient or near the patient's bed. The difference in difficulty of identifying the period of intervention from depth can be seen between Figures 4 and 5. The 'base'-size vision transformer trained here for the task of intervention detection outperformed the baseline (state-of-the-art) models over all metrics, while the sensitivity of the 'tiny' vision transformer was only slightly outperformed by the RGB-D fusion baseline model. When exploring variables that might affect the performance of the models, one of the models trained was a 'base' vision transformer that took advantage of HHA encoding of the depth data after applying the perspective transformation process. This model was overall the highest performing model, and its associated confusion matrix can be seen in Table 9. Model size and the encoding type of the depth data were found to have statistically significant effects on the performance of the models, where the 'base' model size and HHA encoding were advantageous. Based on the results of our study, we recommend using a vision transformer model with a larger number of trainable parameters applied on depth data that takes advantage of the perspective transformation process outlined in ref. [5] and HHA encoding of depth data, since this approach was found to have the greatest performance in detecting periods of intervention compared to other models tested.

The methods developed here examined individual representative frames from each intervention event, sampled every 30 s, which is in line with the state of the art in RGB-based intervention detection [24]. Our ability to classify the representative frame is expected to reflect the performance of the model when applied to all frames within a continuous period of intervention. We did examine a single period of intervention at greater temporal resolution. For this experiment, we extracted each frame of a 90 s period (2695 frames in total). The period began with no intervention. An intervention (vital sign check and re-swaddling) started after 48 s and continued until the end of the 90 s period. The model (trained on patients different from the test patient) was applied to all 2695 frames, and this performance was compared with the performance estimated from the 14,892 representative frames, originally extracted at 1 frame per 30 s. The resulting performance metrics (Sn = 96.25%, Sp = 100%, Acc = 98.26%, F1 = 98.09%) were equivalent to the performance metrics observed when using the representative frames, validating our approach of evaluating models using representative frames sampled at 1 frame per 30 s. Future work will examine the accuracy with which the precise start and end of each intervention can be determined by the proposed methods. This will be a somewhat nebulous task, since even a human annotator will have difficulty determining the precise start and end points of an intervention (e.g., is the start of the intervention when the clinician's hands are first visible in frame or when the clinician first makes contact with the patient, etc.).

Future Work

This paper studied the effect of multiple design parameters (separately and in conjunction) on the performance of ViT clinical intervention detection models. While the use of perspective transform and HHA encoding was found to be beneficial, supplementing the training data with simulated patient care scenes alone did not improve model performance. This outcome was unexpected, as it was thought that correcting for the class imbalance would improve classification accuracy. Although the best-performing model did not include the use of simulated data, the second best model overall did utilize it (ViT Base Simulated HHA PT). This suggests that simulated data may have promise when used in conjunction with other design parameters, and future research could explore the benefits of incorporating more diverse simulated data from a variety of care settings. Researchers could also consider repeating the experiment with simulated data that are captured in an environment that is more comparable to real-world NICU environments. Since the use of HHA encoding had a detrimental effect on some of the baseline depth-based CNN model's metrics, a possible explanation for the lack of benefit from HHA-encoded data is that the hyperparameter search space may have been insufficient to fine-tune the model and fully leverage these new data. Future work should expand on this search space to re-examine the potential benefit from HHA encoding and consider identifying or training a foundation model pre-trained on HHA-encoded data. Another avenue of research may be to train and test ViT with patches of varying numbers and sizes to study their effect on the performance change that occurs when using perspective transformation. Perspective transformation was shown to increase the performance of the CNN architecture, though the improvement to the ViT model's performance was not consistent. Future research may also look at background subtraction techniques, where a reference frame containing only the patient is used to highlight differences in depth during an intervention. Lastly, this study examined single-frame depth data; future research will extend this work to consider depth video, since the movements of a clinician in the scene will likely differ from those of the patient. ViT models have recently been extended to RGB video [30,31], and 3D CNN models [32] have also shown great promise for this type of analysis.

Author Contributions: Conceptualization, Z.H.-A., Y.S.D., J.H. and J.R.G.; software, Z.H.-A. and Y.S.D.; investigation, Z.H.-A. and J.R.G.; resources, K.G., J.H. and J.R.G.; data curation, Z.H.-A. and Y.S.D.; writing—original draft preparation, Z.H.-A. and J.R.G.; writing—review and editing, Y.S.D., K.G., J.H. and J.R.G.; supervision, J.R.G.; project administration J.H. and J.R.G.; funding acquisition, J.H. and J.R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the IBM Center for Advanced Studies and Natural Sciences and Engineering Research Council of Canada grant number CRDPJ 543940-19.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Carleton University (Project # 107193) and the Children's Hospital of Eastern Ontario (REB # 17/76X).

Informed Consent Statement: All subjects gave their informed consent for inclusion before they participated in the study.

Data Availability Statement: The patient data cannot be shared due to Research Ethics Board constraints. The simulated manikin data can be obtained by contacting the authors.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

NICU	Neonatal Intensive Care Unit
RGB	Red, green, and blue
ViT	Vision transformer
CNN	Convolutional neural network
HHA	Horizontal disparity, Height above ground, and Angle with gravity
CHEO	Children's Hospital of Eastern Ontario
RGB-D	Red, green, blue, and depth
PMDI	Patient Monitor Data Import
IR	Infrared
M	Million
ROI	Region of interest
PT	Perspective transformation
MAE	Masked auto-encoding
ROC	Receiver operating characteristic
MCC	Mathew's correlation coefficient

Appendix A. N-Way ANOVA Tables

Table A1. N-way ANOVA table for specificity.

Effect	DF	Sum of Squares	Mean Squares	F-Value	*p*-Value
Model Size	1	3.656044	3.656044	1.863623	0.176987
HHA encoding	1	4.398646	4.398646	2.242155	0.13921
PT	1	3.156292	3.156292	1.60888	0.209244
Simulated data	1	3.236363	3.236363	1.649696	0.20363
Model size: HHA encoding	1	0.208932	0.208932	0.1065	0.745229
Model size: PT	1	0.012997	0.012997	0.006625	0.935383
Model size: simulated data	1	1.507441	1.507441	0.768399	0.383991
HHA encoding: PT	1	0.418449	0.418449	0.213299	0.64576
HHA encoding: simulated data	1	1.969506	1.969506	1.003931	0.320134
PT: simulated data	1	0.018866	0.018866	0.009617	0.922188
Model size: HHA encoding: PT	1	2.386956	2.386956	1.216721	0.274136
Model size: HHA encoding: simulated data	1	0.010425	0.010425	0.005314	0.942115
Model size: PT: simulated data	1	0.152822	0.152822	0.077899	0.781065
HHA encoding: PT: simulated data	1	0.326248	0.326248	0.166301	0.684782
Model size: HHA encoding: PT: simulated data	1	4.0893	4.0893	2.08447	0.153681
Residual	64	125.5548	1.961794		

Table A2. N-way ANOVA table for sensitivity.

Effect	DF	Sum of Squares	Mean Squares	F-Value	*p*-Value
Model size	1	179.8877	179.8877	3.355498	0.071635
HHA encoding	1	79.50419	79.50419	1.483015	0.227774
PT	1	11.65195	11.65195	0.217347	0.642652
Simulated data	1	11.48251	11.48251	0.214187	0.645075
Model size: HHA encoding	1	73.38347	73.38347	1.368844	0.246349
Model size: PT	1	11.05032	11.05032	0.206125	0.651358
Model size: simulated data	1	3.043599	3.043599	0.056773	0.812432
HHA encoding: PT	1	194.4048	194.4048	3.626291	0.06137
HHA encoding: simulated data	1	17.61146	17.61146	0.328512	0.568545

Table A2. *Cont.*

Effect	DF	Sum of Squares	Mean Squares	F-Value	p-Value
PT: simulated data	1	3.53418	3.53418	0.065924	0.79819
Model size: HHA encoding: PT	1	7.4923	7.4923	0.139756	0.709759
Model size: HHA encoding: simulated data	1	9.097069	9.097069	0.16969	0.681764
Model size: PT: simulated data	1	3.290016	3.290016	0.06137	0.805137
HHA encoding: PT: simulated data	1	1.941173	1.941173	0.036209	0.849686
Model size: HHA encoding: PT: simulated data	1	32.74856	32.74856	0.610869	0.437343
Residual	64	3431.029	53.60982		

Table A3. N-way ANOVA table for precision.

Effect	DF	Sum of Squares	Mean Squares	F-Value	p-Value
Model size	1	169.167	169.167	3.349493	0.071883
HHA encoding	1	308.1034	308.1034	6.100424	0.016192
PT	1	32.84467	32.84467	0.650322	0.422984
Simulated data	1	175.7713	175.7713	3.480258	0.06669
Model size: HHA encoding	1	0.000623	0.000623	1.23E-05	0.997209
Model size: PT	1	3.807472	3.807472	0.075388	0.784532
Model size: simulated data	1	69.96673	69.96673	1.385336	0.243553
HHA encoding: PT	1	34.25033	34.25033	0.678154	0.413281
HHA encoding: simulated data	1	142.8611	142.8611	2.828639	0.097471
PT: simulated data	1	5.595797	5.595797	0.110796	0.740327
Model size: HHA encoding: PT	1	57.35261	57.35261	1.135577	0.290593
Model size: HHA encoding: simulated data	1	2.223465	2.223465	0.044024	0.834475
Model size: PT: simulated data	1	0.921884	0.921884	0.018253	0.892953
HHA encoding: PT: simulated data	1	0.01316	0.01316	0.000261	0.987171
Model size: HHA encoding: PT: simulated data	1	105.4943	105.4943	2.08878	0.153263
Residual	64	3232.336	50.50525		

Table A4. N-way ANOVA table for accuracy.

Effect	DF	Sum of Squares	Mean Squares	F-Value	p-Value
Model size	1	8.485055	8.485055	3.065306	0.084772
HHA encoding	1	6.053363	6.053363	2.186834	0.1441
PT	1	2.38325	2.38325	0.860972	0.356953
Simulated data	1	3.555322	3.555322	1.284394	0.261311
Model size: HHA encoding	1	0.19227	0.19227	0.069459	0.792972
Model size: PT	1	0.117856	0.117856	0.042577	0.837179
Model size: simulated data	1	0.925994	0.925994	0.334524	0.565037
HHA encoding: PT	1	3.749151	3.749151	1.354416	0.248828
HHA encoding: simulated data	1	0.813637	0.813637	0.293934	0.589593
PT: simulated data	1	0.000444	0.000444	0.00016	0.989934
Model size: HHA encoding: PT	1	1.319679	1.319679	0.476746	0.492396
Model size: HHA encoding: simulated data	1	0.091078	0.091078	0.032903	0.856633
Model size: PT: simulated data	1	0.270634	0.270634	0.097769	0.75554
HHA encoding: PT: simulated data	1	0.430846	0.430846	0.155647	0.694508
Model size: HHA encoding: PT: simulated data	1	1.7798	1.7798	0.642969	0.425605
Residual	64	177.158	2.768094		

Table A5. N-way ANOVA table for F1-score.

Effect	DF	Sum of Squares	Mean Squares	F-Value	*p*-Value
Model size	1	174.0898	174.0898	3.679687	0.059543
HHA encoding	1	168.8194	168.8194	3.568288	0.063426
PT	1	12.11672	12.11672	0.256108	0.614546
Simulated data	1	112.5921	112.5921	2.379827	0.127842
Model size: HHA encoding	1	31.31768	31.31768	0.661953	0.418888
Model size: PT	1	16.74717	16.74717	0.35398	0.553966
Model size: simulated data	1	7.185012	7.185012	0.151868	0.698051
HHA encoding: PT	1	142.6825	142.6825	3.015839	0.087266
HHA encoding: simulated data	1	6.6096	6.6096	0.139705	0.70981
PT: simulated data	1	2.609685	2.609685	0.05516	0.815066
Model size: HHA encoding: PT	1	4.715344	4.715344	0.099667	0.753257
Model size: HHA encoding: simulated data	1	11.74248	11.74248	0.248198	0.620055
Model size: PT: simulated data	1	6.821549	6.821549	0.144185	0.705412
HHA encoding: PT: simulated data	1	7.348248	7.348248	0.155318	0.694814
Model size: HHA encoding: PT: simulated data	1	4.533364	4.533364	0.09582	0.75791
Residual	64	3027.906	47.31104		

Table A6. N-way ANOVA table for MCC.

Effect	DF	Sum of Squares	Mean Squares	F-Value	*p*-Value
Model size	1	213.7838	213.7838	4.197383	0.044592
HHA encoding	1	204.7773	204.7773	4.020551	0.049182
PT	1	9.375126	9.375126	0.184069	0.66934
Simulated data	1	107.0097	107.0097	2.101003	0.152085
Model size: HHA encoding	1	27.3037	27.3037	0.536075	0.466737
Model size: PT	1	12.50084	12.50084	0.245439	0.622002
Model size: simulated data	1	11.54674	11.54674	0.226706	0.635599
HHA encoding: PT	1	142.0106	142.0106	2.788204	0.099845
HHA encoding: simulated data	1	13.47142	13.47142	0.264495	0.60882
PT: simulated data	1	0.374156	0.374156	0.007346	0.931965
Model size: HHA encoding: PT	1	7.443392	7.443392	0.146142	0.703517
Model size: HHA encoding: simulated data	1	9.451489	9.451489	0.185568	0.668078
Model size: PT: simulated data	1	5.379411	5.379411	0.105618	0.746249
HHA encoding: PT: simulated data	1	4.13815	4.13815	0.081247	0.776534
Model size: HHA encoding: PT: simulated data	1	7.49196	7.49196	0.147095	0.702598
Residual	64	3259.69	50.93265		

References

1. Zhang, X.; Hu, M.; Zhang, Y.; Zhai, G.; Zhang, X.P. Recent progress of optical imaging approaches for noncontact physiological signal measurement: A review. *Adv. Intell. Syst.* **2023**, *5*, 2200345. [CrossRef]
2. Souley Dosso, Y.; Greenwood, K.; Harrold, J.; Green, J.R. RGB-D Scene Analysis in the NICU. *Comput. Biol. Med.* **2021**, *138*, 104873. [CrossRef] [PubMed]
3. Villarroel, M.; Chaichulee, S.; Jorge, J.; Davis, S.; Green, G.; Arteta, C.; Zisserman, A.; McCormick, K.; Watkinson, P.; Tarassenko, L. Non-Contact Physiological Monitoring of Preterm Infants in the Neonatal Intensive Care Unit. *npj Digit. Med.* **2019**, *2*, 1–18. [CrossRef] [PubMed]
4. Orlandi, S.; Raghuram, K.; Smith, C.R.; Mansueto, D.; Church, P.; Shah, V.; Luther, M.; Chau, T. Detection of Atypical and Typical Infant Movements Using Computer-based Video Analysis. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 3598–3601. [CrossRef]
5. Hajj-Ali, Z.; Greenwood, K.; Harrold, J.; Green, J.R. Towards Depth-based Respiratory Rate Estimation with Arbitrary Camera Placement. In Proceedings of the 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Messina, Italy, 22–24 June 2022; pp. 1–6. [CrossRef]

6. Souley Dosso, Y.; Greenwood, K.; Harrold, J.; Green, J.R. Bottle-Feeding Intervention Detection in the NICU. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Guadalajara, Mexico, 1–5 November 2021; pp. 1814–1819. [CrossRef]
7. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
8. Souley Dosso, Y. Machine Vision for Patient Monitoring in the Neonatal Intensive Care Unit. Ph.D. Thesis, Philosophy, Biomedical Engineering, Carleton University, Ottawa, ON, Canada, 2022. [CrossRef]
9. Hajj-Ali, Z. Depth-Based Patient Monitoring in the NICU with Non-Ideal Camera Placement. Master's Thesis, Applied Science, Carleton University, Ottawa, ON, Canada, 2023. [CrossRef]
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2010**, arXiv:2010.11929.
11. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *ACM Comput. Surv.* **2022**, *54*, 200:1–200:41. [CrossRef]
12. Jamil, S.; Jalil Piran, M.; Kwon, O.J. A Comprehensive Survey of Transformers for Computer Vision. *Drones* **2023**, *7*, 287. [CrossRef]
13. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2023**, *45*, 87–110. [CrossRef] [PubMed]
14. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
15. Hussain, M.; Bird, J.J.; Faria, D.R. A Study on CNN Transfer Learning for Image Classification. In *Proceedings of the Advances in Computational Intelligence Systems, Portsmouth, UK, 4–6 September 2019*; Lotfi, A., Bouchachia, H., Gegov, A., Langensiepen, C., McGinnity, M., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 191–202. [CrossRef]
16. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In *Proceedings of the Computer Vision — ECCV, Zurich, Switzerland, 6–12 September 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 345–360. [CrossRef]
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
18. OpenCV: Depth Map from Stereo Images. Available online: https://docs.opencv.org/4.x/dd/d53/tutorial_py_depthmap.html (accessed on 1 August 2024).
19. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2004. [CrossRef]
20. Gupta, S.; Arbeláez, P.; Malik, J. Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 564–571. [CrossRef]
21. Chen, X. Depth2HHA. 2018. Available online: https://github.com/charlesCXK/Depth2HHA (accessed on 1 August 2024).
22. Gupta, S. S-Gupta/Rcnn-Depth. 2020. Available online: https://github.com/s-gupta/rcnn-depth/blob/7a7baf7dcccc6fdf6be7c13d16828064d89dff4e/rcnn/saveHHA.m (accessed on 1 August 2024).
23. Tan, F.; Xia, Z.; Ma, Y.; Feng, X. 3D Sensor Based Pedestrian Detection by Integrating Improved HHA Encoding and Two-Branch Feature Fusion. *Remote Sens.* **2022**, *14*, 645. [CrossRef]
24. Chaichulee, S.; Villarroel, M.; Jorge, J.; Arteta, C.; McCormick, K.; Zisserman, A.; Tarassenko, L. Cardio-respiratory signal extraction from video camera data for continuous non-contact vital sign monitoring using deep learning. *Physiol. Meas.* **2019**, *40*, 115001. [CrossRef] [PubMed]
25. Hozayen, M.; Nizami, S.; Bekele, A.; Dick, K. Developing a Real-Time Patient Monitor Data Import System. In Proceedings of the National Conference On Undergraduate Research (NCUR), Edmond, OK, USA, 4–7 April 2018; p. 8.
26. StandInBaby® | Single. Available online: https://www.standinbaby.com/product/standinbaby_single/ (accessed on 1 August 2024).
27. Wightman, R. PyTorch Image Models. 2019. Available online: https://github.com/rwightman/pytorch-image-models (accessed on 1 August 2024).
28. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15979–15988. [CrossRef]
29. Vision Transformer and MLP-Mixer Architectures. 2022. Available online: https://github.com/google-research/vision_transformer (accessed on 1 August 2024).
30. Neimark, D.; Bar, O.; Zohar, M.; Asselmann, D. Video Transformer Network. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Virtual, 11–17 October 2021; pp. 3156–3165. [CrossRef]

31. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lucic, M.; Schmid, C. ViViT: A Video Vision Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Los Alamitos, CA, USA, 10–17 October 2021; pp. 6816–6826. [CrossRef]
32. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Arzani, M.M.; Yousefzadeh, R.; Van Gool, L. Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification. *arXiv* **2017**, arXiv:1711.08200.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.