Depth-based Patient Monitoring in the NICU with Non-Ideal Camera Placement

By

Zein Hajj-Ali

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Master of Applied Science

In

Electrical and Computer Engineering
with Specialization in Data Science

Ottawa-Carleton Institute for Electrical and Computer Engineering

Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
January 2023

Abstract

Depth cameras can improve the performance of patient monitoring systems without the introduction of multiple sensors in the NICU. A method was developed to correct non-ideal camera placement. The mean absolute percentage error of the method tested on 28 patients was 5.58 for camera angles up to 38.58° away from the optimal camera placement. An ROI selection method was developed and tested for the use of extracting a respiratory rate signal. The ROI selection method was found to have an average Sørensen–Dice coefficient of 0.62 and Jaccard index of 0.46. The signal was compared to a simpler method resulting in an improvement to the percentage of acceptable estimates. An intervention detection method was developed using a vision transformer model, and the performance was compared to the state-of-the-art in the field. The best model was found to achieve a sensitivity of 85.6%, precision of 89.8%, and F1-Score of 87.6%.

Extended Abstract

Depth cameras can improve non-contact patient monitoring systems in the Neonatal Intensive Care Unit (NICU). Camera placement is secondary to equipment used for patient care; therefore, a method was developed to correct for non-ideal camera placement. The mean absolute percentage error (MAPE) of the perspective transformation method of correcting the viewpoint of the camera was tested on 28 patients was found to be 5.58 for camera angles up to 38.58° away from the optimal camera placement. Since depth data can be more privacy-preserving than RGB or RGB-D data, Region-of-Interest (ROI) selection using depth cameras can be enable the automatic blurring of identifiable features. An ROI selection method was developed and tested for the use of extracting a respiratory rate signal. The ROI selection method was evaluated against manually selected ROIs and found to result in an average Sørensen–Dice coefficient of 0.62 and Jaccard index of 0.46. The signal extracted from the automatically selected ROI was compared to a simpler method resulting in an improvement to the percentage of acceptable estimates, where the mean absolute error is less than 5 breaths per minute (3.60% to 13.47% in the frequency domain and 6.12% to 8.97% in the time domain). Clinical interventions and routine care in the NICU can disrupt the process of data collection, and commonly need to be excluded from recording when studying patients in the NICU. Detecting these periods automatically can decrease the time needed for hand-annotating segments of recordings and may further be used for intervention classification in the future. An intervention detection method based solely on depth data was developed using a vision transformer model. Multiple variables were investigated, and the performance was compared to the state-of-the-art in the field. The best performing model was utilized ~85M trainable parameters and was trained and evaluated on data that had been

perspective transformed and HHA encoded and was found to achieve a sensitivity of 85.6%, precision of 89.8%, and F1-Score of 87.6%.

Acknowledgements

I would first like to thank my supervisor Prof. James Green for his support, guidance, and encouragement throughout my graduate studies. His expertise has been invaluable in helping me develop my research skills and complete my thesis project. I am glad to have been part of the Carleton University Biomedical Collaboratory (CU-BIC), run by Prof. Green, where I was surrounded by supportive peers and labmates conducting important research.

Most importantly, I would like to thank my parents, Nouhad and Bassam, and my sister Jana, for their unwavering support and motivation. This thesis would not have been possible without their love.

This study was supported by the Natural Sciences and Engineering Research Council of Canada and the IBM Centre for Advanced Studies (CAS).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANOVA	Analysis of Variance
BPM	Breaths per minute
CHEO	Children's Hospital of Eastern Ontario
CNN	Convolutional Neural Network
HHA	Horizontal disparity, Height above the ground, Angle between the local surface normal and the inferred gravity direction
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCC	Matthew's Correlation Coefficient
NICU	Neonatal Intensive Care Unit
NLP	Natural Language Processing
NN	Neural Network
PAE	Percentage of Acceptable Estimates
PMDI	Patient Monitor Data Import
ReLU	Rectified Linear Unit
RGB	Red-Green-Blue
RGB-D	Red-Green-Blue-Depth
ROI	Region-Of-Interest
rPPG	remote Photoplethysmography
RR	Respiratory Rate
ViT	Vision Transformer

1 Introduction

1.1 Introduction

This chapter presents the motivation for this thesis the problem statement, summarizes the research contributions, and outlines the organization of the thesis.

1.2 Motivation

Newborn patients admitted to the NICU require continuous monitoring and round-the-clock care. This typically involves several wired sensors attached to the patient's skin which are susceptible to motion artifacts and may interfere with clinical and parental care. Furthermore, wired sensors can irritate sensitive skin, which can be exacerbated by the need for removal and reapplication due to medical interventions.

Previous studies have focused on a range of technologies for the non-intrusive non-contact monitoring of NICU patients. These include methods using RGB cameras [1]–[4] ultrawideband radar [5], [6], and pressure-sensitive mats [7].

When compared to 2D images from regular RGB cameras, images from RGB-D (Depth) cameras can increase understanding of the scene being investigated. The third dimension allows for easier background removal and foreground segmentation, while also providing information about the topography of the scene. Using the depth channel of the frame and discarding the RGB channels can increase the privacy of subjects in the scene without sacrificing the ability to understand and measure aspects of the scene.

1.3 Problem Statement

Non-contact monitoring of patients often requires the definition of a region-of-interest (ROI) as a first step. The target biosignal is then estimated from measurements taken across the ROI. For example, Eulerian video magnification has been applied to multi-modal video (RGB, depth, and near-infrared) to estimate heart rate based on time-varying intensity within an ROI [8]. In previous studies, ROI selection for neonates has been done by hand using manual ROI selection tools. Some studies have also investigated developing automated methods to speed up the process. Usually, this is done by looking at regions where a patient's skin can be seen in RGB images. Due to the nature of real-world NICU environments, varying levels and colours of light can be seen over the course of a day. Patients in the NICU may also need to be covered with a quilt or blanket, leaving very few regions of skin exposed. A combination of these factors may cause a deterioration in the performance of RGB-based methods. This thesis develops and evaluates automated ROI detection algorithms, designed for depth data.

Depth-based methods may circumvent the previously mentioned issues, though not without introducing new challenges. If a depth camera is not placed in an ideal location, or is calibrated incorrectly for the scene, the angles at which the scene is viewed can cause a conventional/deterministic method to underperform. The non-ideal placement of a depth camera is likely in an NICU environment, since camera placement must come as a secondary priority to patient care and should not interfere with other equipment in the environment. Figure 1 illustrates one such example; when recording video from an enclosed isolette (incubator), the camera must be placed off-centre, due to the presence of phototherapy equipment in the middle of the plexiglass cover. This results in a non-ideal camera placement such that the image plane is not aligned with the plane of the patient's bed. This thesis

explores the use of perspective transforms to address the misalignment between the camera and patient planes.

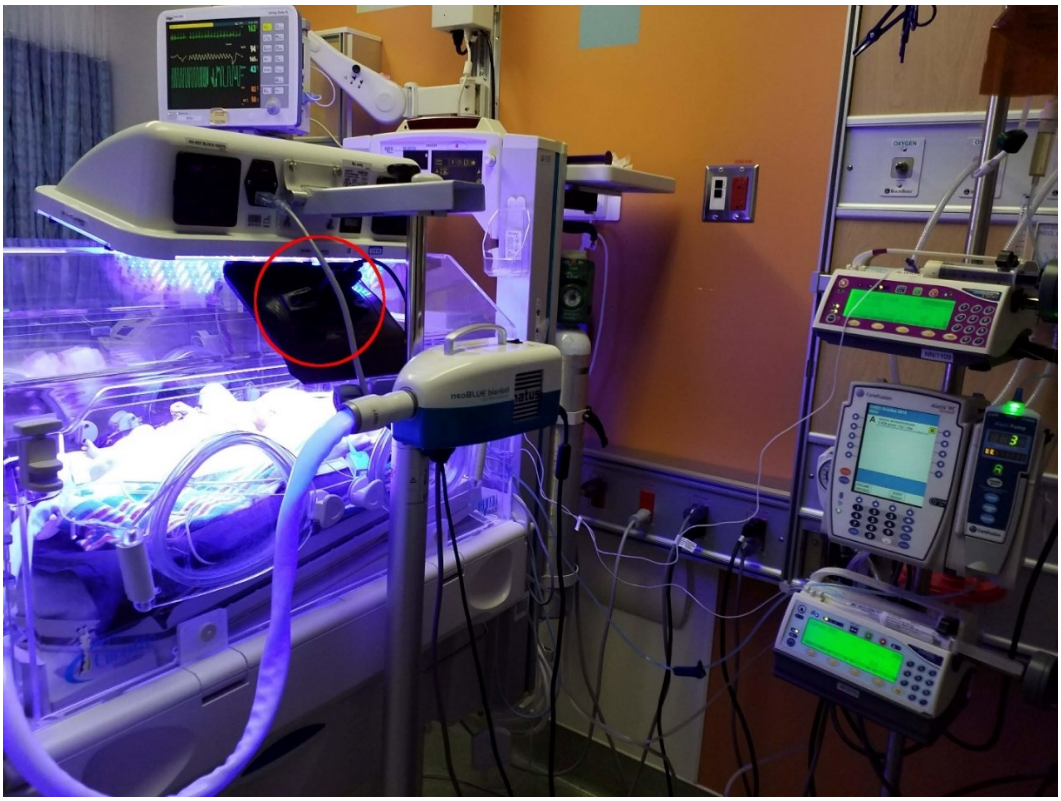


Figure 1: Patient undergoing phototherapy in the NICU. RGB-D camera highlighted with red circle

Training a neural network requires large amounts of labelled data. This can be hard to come by for niche use cases, such as intervention detection in the NICU. Transfer learning can be leveraged in cases such as these, where a model is pre-trained on large amounts of data for a more general task, like image classification, before being fine-tuned on a smaller task-specific dataset. Pre-trained models for image classification are usually pre-trained on large amounts of labelled RGB (colour) images, and the feature extraction layers may not be easily transferable to 1-channel depth images. This thesis investigates how to encode depth information using three channels, such that deep learning models trained to use 3-channel RGB data can be effectively applied to 1D depth data.

Further, a larger portion of a patient's time in the NICU is spent without clinical intervention than with. In the context of collecting data for the purpose of training an intervention detection model, this means that many more hours of non-intervention data are recorded when compared to the periods of intervention. This leads to a situation of class imbalance, where any machine learning or computer vision model trained on such data will tend to under-predict the rare class (i.e., a period of clinical intervention) and over-predict the dominant class (i.e., periods without ongoing intervention). This thesis explores the use of simulated patient intervention data to augment the minority class and address class imbalance when training deep learning models for automated clinical intervention detection.

1.4 Summary of Contributions

This thesis presents advancements to the field of non-contact neonatal patient monitoring using the depth modality of RGB-D cameras. Solving the challenges faced through the development of these methods in a dynamic manner results in methods that can be used for future research into non-contact patient monitoring in the NICU. The major contributions are highlighted in the section below:

1. Developed a method for transforming the viewing perspective of a depth camera to correct for non-ideal camera placement. The method was tested on a number of patients with different camera positions and its impact on automatic ROI detection and respiration rate estimation is assessed. This is described in Chapter 4 of this thesis.
2. Built an automatic depth-based ROI selection method that does not rely on the appearance of skin regions in the scene. The method was built on top of the perspective transformation method to make use of its view plane correction. The ROI selection and perspective transformation method pipeline was tested to show improvements against a simpler method for use in respiratory rate estimation from depth. This is described in Chapter 5.

3. Trained a vision transformer (ViT) deep learning model to detect periods of clinical or routine care intervention from single depth frames. The model outperforms the state-of-the-art in the field, while being more privacy-preserving than an RGB-based model. This contribution is described in Chapter 6.
4. Investigated the use of an alternative encoding method for depth data and its effects on the performance of the intervention detection model. This is also described in Chapter 6.
5. Investigated the use of simulated data to correct the high class imbalance in the dataset. Evaluated the effects of the simulated data on the intervention detection model. Finally, this is also described in Chapter 6.

1.5 Organization of Thesis

This thesis consists of 7 chapters. In Chapter 2, background information on RGB-D cameras, geometric transformations, and deep learning methods (including vision transformers) in presented. In addition, literature review of HHA encoding for depth data, non-contact vital sign monitoring, intervention detection in the NICU, and depth-based respiratory rate estimation and ROI selection is outlined. Chapter 3 describes the configuration of the data collection from patients in the NICU. Chapter 4 presents the method for correcting the non-ideal camera placement by utilizing perspective transformation, and the evaluation of results from patient recordings. Chapter 5 introduces the ROI selection method built on the methods described in Chapter 4, and presents the results of evaluating the newly build pipeline on two respiratory estimation methods. In Chapter 6, an intervention detection model is presented. The dataset used for training and testing is described, and the results are discussed in the same chapter. Chapter 7 presents a summary of contributions and provides recommendations for future work. The methods described in Section 4.2 and Chapter 5 contain content from a paper published in MeMeA 2022 [9], of which I was the lead author.

2 Background & Literature Review

This chapter will present background information and literature review of depth cameras and depth data encoding, convolutional neural networks, and vital sign monitoring in the NICU including respiratory rate estimation, ROI selection, and intervention detection.

2.1 Measuring Depth Using RGB-D Cameras

A standard (RGB) camera outputs an image of three channels, each corresponding to the intensity of a colour in the scene, hence R for the red channel, G for the green channel, and B for the blue channel. Each pixel in the resulting image is displayed as a specific colour and combining a large number of the pixels will make an image of the scene. A Depth or RGB-D camera adds another channel to this image. In this case the fourth channel will correspond to the distance away from the camera, measured at each pixel location. Such depth images are often displayed in grey scale, where brightness corresponds to distance, or using a colour map, such as shown in Figure 2. Figure 3 shows the RGB image output from the RGB-D camera on the left, and the depth image on the right.

Figure 2: Colour map key for depth images

Figure 3: RGB and depth image of the same scene taken using the Intel RealSense SR300

Three types of technologies are mainly used to record distance: stereo cameras, time-of-flight, and coded light. Stereo cameras place two sensors a set distance apart from each other. Images are taken from each, and the depth information is calculated by comparing the images while knowing the exact distance between the sensors. Time-of-flight cameras include an emitter and a sensor. Light is emitted from the camera and the time taken to reflect back to the sensor is used to calculate the distance. Coded light cameras (like the Intel RealSense SR300 [10]) project patterned infrared light on to the scene using an IR module in the camera itself, then calculates the shape and depth of the scene by looking at the distortion of the light when it is reflected back to the camera. Limitations of this technology include susceptibility to distortion from other noise in the scene, ambient infrared light overwhelming the projected pattern, and the distance the projected infrared light can cover due to the power of the emitter.

2.2 Geometric Transformations

Geometric transformations can be used to manipulate the content of a (2D) image or (3D) shape. This is done by remapping each pixel or point in the image to another through the use of some mapping function [11]. By changing the arrangement of the pixels while preserving the relative relationship between them, we can extract useful information about the image that might otherwise be unknown. Simpler transformations such as scaling, translation, reflection, and shear are used regularly when working with images, not least in medical imaging [12]. These affine transformations preserve the parallelism of lines. Similarity transformations, a subset of affine transformations including scaling, translation, and reflection, retain the angles and ratios of the distances between points [13].

More complex than Similarity transformations, Perspective transformations change the distance between points. By changing the perspective, objects may appear closer or farther away from the viewpoint. To this end, Perspective transformations will subject lines to

foreshortening. This means that lines (or the distances between points) that are further away from the center of the projection are proportionally shortened. When applied to 3D matrices that are then projected onto a 2D frame, the transformation will alter the perspective that the 3D shape is seen from.

When working with depth information from a RGB-D camera as described in Section 2.1, we are able to de-project the 2D depth frame matrix into a set of points in 3D space called a point cloud [14] (as in Figure 4). It is then possible to apply a perspective transformation operation on the 3-dimensional points (resulting in Figure 5) before projecting the points back into a 2D depth frame matrix to form a new image. Perspective transformations in 3 dimensions are used throughout this thesis. The parameters of the transform can be estimated using a set of registration points. The perspective transformation operation is formally defined in Chapter 4.

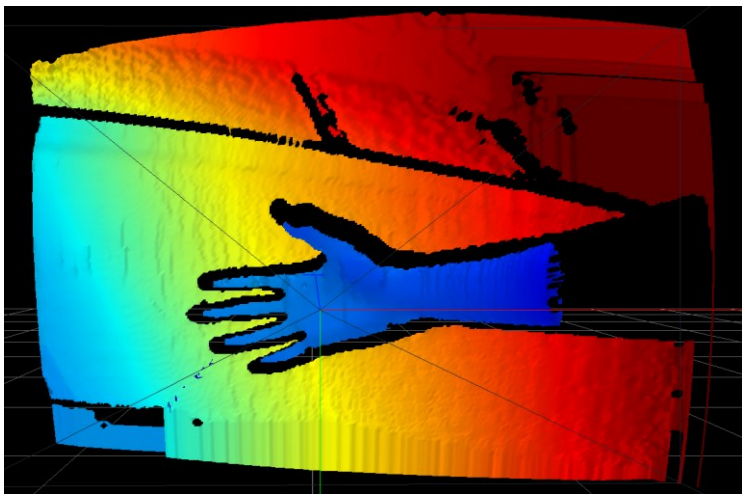


Figure 4: Original point cloud representation

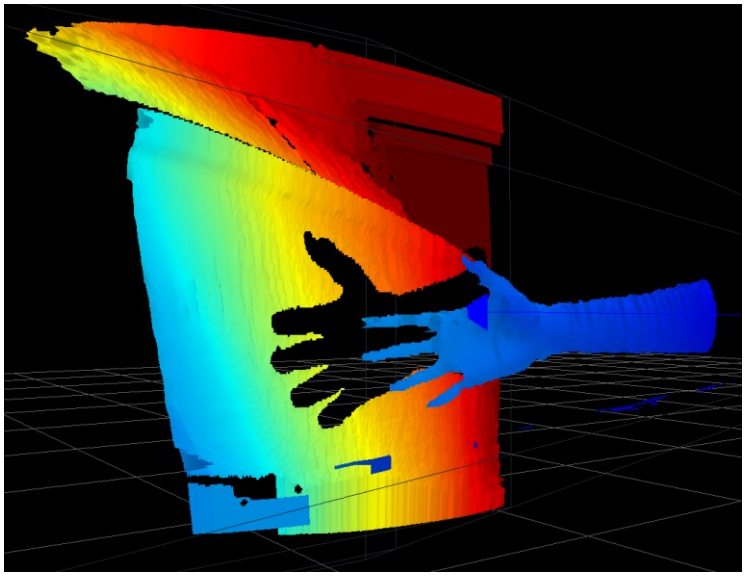


Figure 5: Transformed point cloud representation

2.3 Introduction to Deep Learning and Convolutional Neural Networks

Neural Networks (NNs) are a subset of Machine Learning. They mimic the human brain in the way that neurons transmit signals to each other. A neuron receives a weighted sum of input signals, processes the sum using a non-linear function, and outputs the resulting signal to the next layer of neurons in the network. The connections between neurons are called edges. Neurons and edges both have weights that can be adjusted during training. The weights affect the resulting output of the neurons.

Neurons are organized into sets of layers. Layers may have different functions and transformations that they perform on the input data. In a Feedforward Neural Network, the signals travel from the first layer (the input layer), through any middle layers (called hidden layers), and out through the output layer.

Training a NN consists of finding the error between the predicted output (found by forward-propagating the input through the network) and the target output (or ground truth). This error is found using a loss function. The error propagates backwards through the network

during training, following backpropagation. The NN then uses a learning rule or (optimizer) to update the weights of the edges by an amount determined by the error using stochastic gradient descent. Repeating this optimization process for many training epochs will bring the network's predictions closer to the target output. Training Neural Networks with more than 2 or 3 hidden layers is called Deep Learning.

Convolutional Neural Networks (CNNs) are a subset of neural networks that can reduce the number of learnable parameters in the network while maintaining a high performance [15]. They are commonly used for image, speech, and audio signal inputs as they tend to perform significantly better than prior technologies [16]. CNN typically contain three different types of layers: Convolutional Layers, Pooling Layers, and Fully-Connected Layers. As the inputs pass through the layers, features in the inputs (like colors, edges, and blobs) can be identified during training, and these learned features can be used by subsequent network layers for object identification and/or semantic segmentation.

Convolutional layers take an input matrix, for instance an image, and check to see if a certain feature (represented by a filter) is found. The filter is a matrix (often as small as 3x3) with weights to be applied to each section of the input image. This is done by calculating the dot product of the section of the input image and the chosen filter (i.e., convolution). After a section is convolved, the filter is then applied to the next section in the image, chosen by moving across the image by a given 'stride'. A Rectified Linear Unit (ReLU) function (or similar function) is applied after every convolution operation to introduce nonlinear behaviour. Figure 6 shows a representation of a convolution operation on an image represented as a 5x5 matrix. The weights in the filter of the convolutional layer can be updated after each forward pass of the network, in the same way as is done for any other neuron.

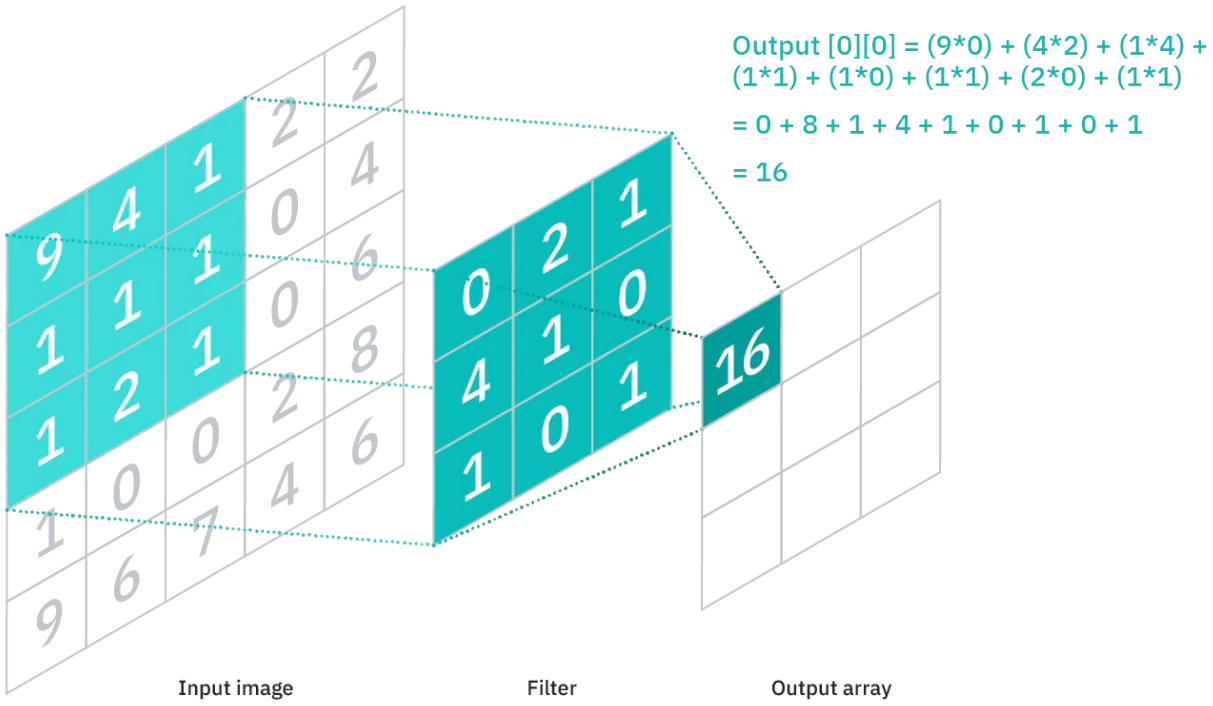


Figure 6: Diagram demonstrating a convolution operation on an image matrix (reproduced from [16]).

Pooling layers reduce the number of dimensions of the network at a node. This is done by applying an aggregation operation (rather than convolving with a filter) to each position in the input data. Two examples are Max Pooling, which returns the pixel with the maximum value, and Average Pooling, which returns the average values of the pixels. Pooling typically decreases the required training time through dimensionality reduction and enhances the invariance to small distortions in the inputs [16], [17].

Fully-connected layers connect each pixel from the input of the layer to each output node using weighted connections. The weights are trainable, and a SoftMax function is usually applied before the output. Fully-connected layers are used to get predictions and classifications from the features extracted in previous convolutional and pooling layers.

2.4 Introduction to Vision Transformers

The transformer architecture (Figure 7), first proposed in [18], is another subset of deep learning. It relies on an 'attention' [19] mechanism to learn the embeddings and relationships of a given set of inputs through an encoder structure before returning a set of output probabilities through the decoder structure. The architecture was first used for machine translation tasks, and are currently considered to be the state-of-the-art in the field of natural language processing (NLP) [20]–[22].

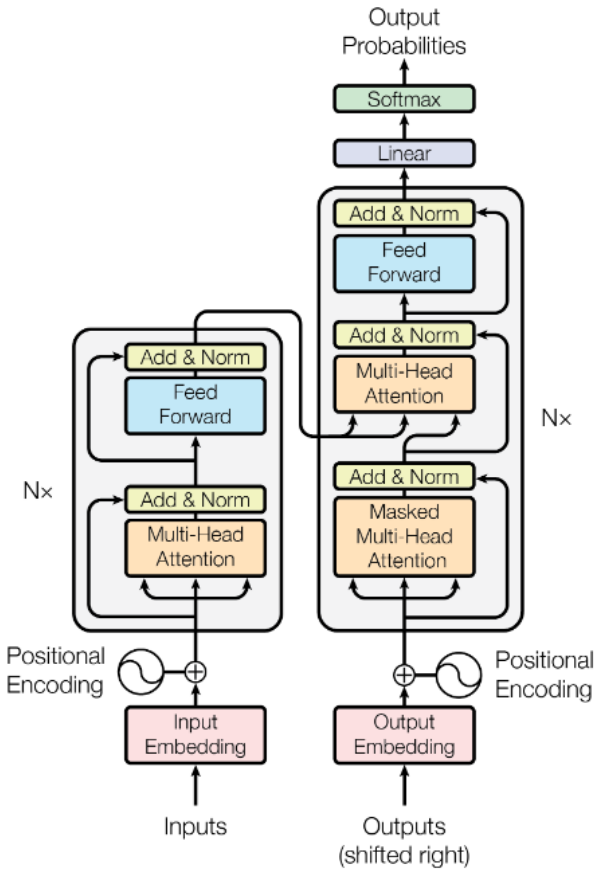


Figure 7: Transformer architecture (reproduced from [18])

Transformers are composed of stacked transformer layers, which are themselves composed of attention and feedforward (i.e., a neural network with 1 hidden layer) sublayers. Transformers receive a set of inputs (e.g., words or tokens in NLP) called a sequence. At each

transformer layer, each element in the sequence is linearly projected into three different vectors, a query (Q), a key (K), and a value (V), by applying three separate learnable weight vectors. The attention mechanism can be thought of as a dictionary mapping between the query of the input and an output through a key-value pairing. Attention weights are found by obtaining the normalized dot-product of the query and the key as in Equation 1. The outputs are then computed as the attention-weighted sum of the value vectors. Multi-headed attention refers to repeating the attention computation multiple times (once for each 'attention head') by splitting up each Q, K, and V into smaller vectors.

Equation 1

$$Attention(Q, K, V) = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V$$

where, d_k is the dimensionality of K.

The encoder portion of the network is used to extract contextualized features of the input sequence, while the decoder portion attempts to generate an output sequence from the encoded features. Both portions make use of the attention mechanism described previously. The encoder adds the positional encoding to the embeddings of each input word, since the positions of words or elements in the input sequence can be important to the output of the model, before processing the input sequence via multiple transformer layers. The final output of the encoder is a contextualized version of the input sequence. The decoder portion takes as inputs the embeddings of the previously generated outputs and adds the positional embeddings. Then, multiple transformer layers process these inputs before forming predictions; crucially, a cross-attention sublayer is inserted between the multi-headed self-attention and feedforward sublayers. Cross-attention allows for the decoder's predictions to be conditioned on the input sequence that was processed by the encoder.

More recently, Dosovitskiy *et al.* [23] presented a modified version of the architecture proposed by Vaswani *et al.* [18], for use in image classification tasks. Rather than tasking the model to learn the relationships between every pixel of an image by feeding an input image into the model as one long sequence of pixels, the vision transformer (ViT) model divides each input image into a number of non-overlapping patches. These patches are flattened into vectors of pixel values and used as the input to the transformer encoder, where each patch can be thought of as a single token in the original transformer model. The vision transformer does away with the decoder portion of the original transformer architecture, utilizing a fully connected head layer after the encoder for the task of image classification. Variations and extensions of this model have had success in image segmentation (with the addition of either transformer-based or multi-layer perceptron decoder portion) [24]–[26], object detection [27], and video action recognition [28].

2.5 Transfer Learning and Fine-Tuning a Deep Learning Model

When training a deep learning model, large amounts of data and compute resources are needed. For this reason, a transfer learning is usually employed to pretrain models on large datasets prior to fine-tuning the model to perform specific tasks with smaller training datasets. Transfer learning is a machine learning technique that involves training a model on a large dataset, before transferring most of the learned parameters over to a new model to be fine-tuned for a related but more specialized task. For image classification, many models have been trained on the ImageNet dataset [29] consisting of ~14 million annotated images in 1000 classes. When published publicly, the weights of these pretrained models can be imported and used for feature extraction in a new model. The pretrained model learns the features of the original dataset through multiple layers, before being fine-tuned on a training set from the smaller dataset.

The pretrained models chosen for transfer learning are normally within the same or similar domains (image classification, object detection, semantic segmentation, etc.). Transfer learning has been shown to improve the average accuracy of CNN models [30] as well as ViT [23] for image classification.

2.6 HHA Encoding for Depth Frames

CNN and transformer architectures have demonstrated success in image classification and semantic segmentation (as mentioned in sections 2.3 and 2.4). These networks are generally trained on large amounts of labelled 3-channel RGB data. HHA encoding is a method of encoding depth data using three channels for each pixel rather than just the 1-channel of depth [31]. An example illustrating the three channels resulting from HHA encoding a depth image can be seen in Figure 8. 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 pretrained on RGB data and finetuned with labelled HHA encoded depth data when compared to fine-tuning on regular 1-channel depth or disparity data. Gupta *et al.* suggests 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 [31]. The authors verify this by fine-tuning a CNN originally trained for object detection and semantic segmentation from RGB images [32].

The horizontal disparity can be calculated from the depth by using Equation 2 [33]:

Equation 2

$$Disparity = \frac{(FocalLength \times Baseline)}{Depth}$$

, where the Focal Length and Baseline are found from the camera's intrinsic matrix [34]. The height above the ground and the angle between the surface normal and inferred gravity

direction can be found using the algorithms presented in [35] and implemented in [36] and [37]. The algorithms require the point cloud representation of the depth image as well as the camera matrix. The direction of gravity is first estimated by finding the direction which is the most aligned to surface normal directions as possible. The Y-axis is initially selected as the direction of gravity before iteratively refining the guess by looking at looking at local surface normal and optimizing for the vector that is most aligned and most orthogonal to the normals. The height above ground can then be found by rotating the pointcloud of the data to the horizontal direction, taking each point and subtract the smallest *y* coordinate from its *y* coordinate value [38]. The angle between the surface normal and the gravity direction can be found from the difference in the respective vectors. The values in each of the channels are also mapped to the range of 0-255 (i.e., an 8-bit value).

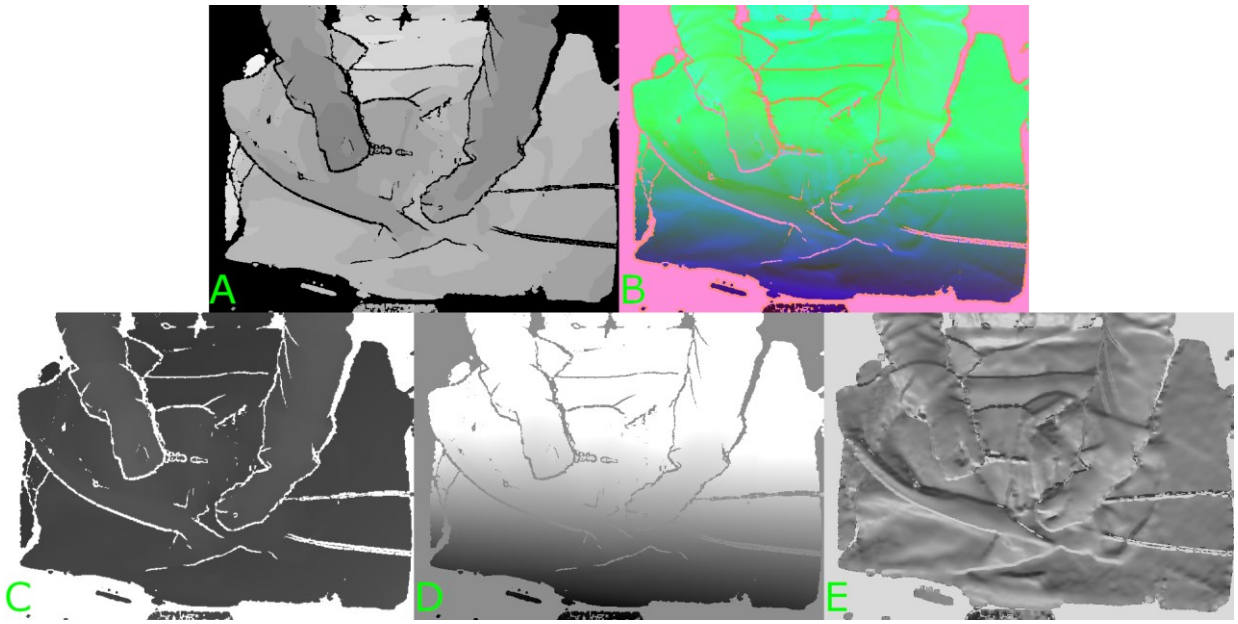


Figure 8: Example of HHA encoding a depth image. A) Original depth image. B) 3-channel HHA encoded image. C) First channel of B (H). D) Second channel of B (H). E) Third channel of B (A).

2.7 Non-Contact Vital Sign Monitoring

Patients in the NICU require continuous monitoring of vital signs and round-the-clock care. Typically, this involves sensors attached to the patient's skin using plastic tape and/or

adhesive material. The application and removal of these sensors may cause irritation and 'skin trauma' that may disrupt the skin barrier function [39]. For this reason, the field of non-contact monitoring of vital signs has investigated different modalities for the use in the NICU.

Wallace *et al.* in [40] studied the application of remote photoplethysmography (rPPG) [41] and Eulerian Video Magnification [42], methods dealing with magnifying slight changes of color or motion in a recorded scene, for non-contact estimation of heart rate and respiratory rate from a person's skin regions. The authors found that although the methods were initially developed to monitor blood flow and estimate heart rate, they were able to extend such methods to estimate of other vital signs such as respiratory rate and blood pressure. The use of these methods with RGB cameras is not always feasible, as a patient's skin may be covered by blankets or clothing, and variations in lighting conditions may decrease their performance as well.

Abbas *et al.* [43] applied an infrared thermography-based method for neonatal patients for the detection of respiratory rate. Kim *et al.* [5] and Lee *et al.* [6] both investigate the use of impulse-radio ultra-wideband radar in the NICU for respiratory rate and, in the case of [6], heart rate of neonatal patients. Bekele *et al.* [7] used a pressure sensitive mat placed in a crib in the NICU under the patient to estimate respiratory rate. Kyrollos *et al.* in [44] also use a pressure sensitive mat to improve false alarm detection in the NICU.

Khanam *et al.* [45] explored the use of a high quality video camera for the estimation of heart rate and respiratory rate of neonates. The region-of-interest (ROI) was found using skin detection based on skin color. The heart rate was then estimated using a color-based method, while the respiratory rate was estimated from the apparent motion in the videos. Fernando *et al.* [46] also investigates the use of camera-based methods for vital sign estimation. The authors placed a camera in view of the patients face for pulse rate estimation, and another in

view of the patient's torso for respiratory rate estimation. The work was done primarily to study the feasibility of using these methods with a wearable camera. Jorge *et al.* [47] studied the use of a camera-based method for the detection of 'cessation of breathing events' in neonates, concluding that using video cameras in conjunction with traditional NICU monitors could help reduce false alarms when detecting these cessation of breathing events during periods of motion.

A 2019 paper by Villarroel *et al.* [48], considered to be the current state-of-the-art in the field of non-contact vital sign monitoring of neonatal patients, studied patients in incubator beds and recorded RGB data using video cameras. Videos were recorded by cutting a hole in the plexiglass top of the incubator, such that the camera had an uninterrupted view of the patient. Modifying the bed in this way is not always feasible, and the reflection artifacts caused by the plexiglass surface can have a detrimental effect on the methods used for the non-contact estimation of vital signs [49]. Although Villarroel's results were promising, finding a mean absolute error (MAE) of 3.5 breaths per minute over 82% of the recordings, the authors did find that their RGB-based method encountered some errors during periods of low light or when shadows were cast over the patient.

2.8 Depth-Based Respiratory Rate Estimation

A 2021 review study summarized and compared technologies for non-contact respiratory rate monitoring of neonatal patients [50]. That study identified three semi-automated depth-based methods. Eastwood-Sutherland *et al.* [51] presented a noncontact respiratory monitoring method, using a Microsoft Kinect camera, and demonstrated its effectiveness on an infant mannequin. Cenci *et al.* [52] derived respiratory rate by calculating structural chest wall motions using a camera positioned directly above an infant lying in an infant-warmer. They found that a depth-based method could be suitably used indoors with poor lighting when compared to methods based on using RGB data. Rehouma *et al.* [53] explored the use of a

custom-built 3D imaging system for monitoring the respiration of pediatric patients. Using two Kinect v2 sensors placed at different angles around the patient, they were able to build a 3D representation of the region of interest (ROI). Rehouma *et al.* then estimated the respiratory rate from the change in volume of the found ROI.

The method presented by Cenci *et al.* used a camera placed directly above the patient's crib, with a view plane that was parallel to the surface of the crib [52]. This is not always possible when monitoring patients in the NICU, as patients may need to be placed in different bed types. In some situations, placing a patient in an incubator or crib incorporating an overhead warmer makes it difficult to place a camera directly above the patient due to integral lighting and heating equipment.

2.9 Depth Based ROI Selection

A number of systems for ROI selection for neonates have been explored; many have done so using manual ROI selection [3], [52], [54] while others have developed automatic or semi-automatic methods. Villarroel *et al.* presented a CNN to detect regions where the patient's skin can be seen in RGB images [48]. The method had difficulty segmenting the ROI of smaller skin regions and, clearly, can not be relied on when a patient is covered with a blanket or quilt. Eastwood-Sutherland *et al.* proposed a method based on detecting the change in luminance, though the method was only tested on a infant mannequin rather than in a real-world NICU setting [51]. A method presented by Rehouma *et al.* uses depth data collected by two sensors to built a point cloud representation and extract the volume of the relevant cuboids [53]. The method was tested on adults and pediatric patients and requires more resources than the method presented in this study.

Yu *et al.* in [55] presented a method for selecting an adult subject's torso ROI using a depth camera for the purpose of extracting a respiratory rate signal. Although they used a Microsoft

Kinect as the depth camera, they did not make use of the skeletal tracking system, as bedding and blankets can cause interference. The camera was placed on the wall above the head of the patient at a known orientation and height. The depth data collected was then transformed (Section 2.2) so that plane of the subject's bed appeared to be parallel to the view plane of the camera. Cross-sections of the frame where taken at multiple depth thresholds, and connected-component analysis was used to find regions where the target ROI may be.

2.10 Intervention Detection in the NICU

During a patient's stay at the NICU, they might experience multiple periods of clinical intervention. A nurse or other medical practitioner may be required to intervene and tend to the needs of the patient at a point in time. This includes intervals of bottle-feeding, diaper changing, adjusting the NICU monitor sensors on the patient's skin, or re-fitting the respirator. These periods of intervention are most commonly excluded from analysis when studying novel techniques of monitoring neonates in the NICU ([9], [56]). However, studies by Villarroel *et al.* [48] and Souley Dosso *et al.* [2], [57] attempt to detect these periods of intervention and, in the case of [2], classify a subset of them (bottle-feeding interventions).

Souley Dosso *et al.* [57] uses the VGG-16 CNN model introduced in [58] as the feature extractor for their method of intervention detection. They attempted using frames from the RGB channels and depth channel separately, as well as multiple forms of multi-modal (RGB and depth) fusion, resulting in similar performance between the RGB and RGB-D fusion models and significantly lower performance of the depth-based model. This thesis presents a method for intervention detection from single 1-channel depth frames using a ViT architecture and compares the results to methods using deep learning architectures as well as conventional (non-deep learning) methods.

2.11 Summary

Methods have been explored for the non-contact monitoring of patient vital signs. RGB-based methods may be the current state-of-the-art in the field, though they introduce disadvantages not shared by depth-based methods. The leading method using RGB cameras involves cutting a hole in the bed if the patient is in an incubator [48]. Though the authors took care to test that the presence of this hole does not interfere with the function of the incubator bed, modifying the medical equipment in this way is not always feasible.

The method presented by Cenci *et al.* [52], using an RGB-D camera, assumed that the position of the camera would be directly above the patient's bed. When monitoring neonatal patients in the NICU, different bed types may be used, therefore this assumption may not hold. For these reasons we discussed the concept of using geometric transformations on the depth data to adjust the perspective the scene is viewed from.

ROI selection has primarily been done manually, though methods have been studied using RGB cameras to detect skin regions [48], [51]. Rehouma *et al.* [53] introduced a method using multiple depth cameras in conjunction. Yu *et al.* [55] introduced a method for ROI selection from non-ideal camera placement, though the method relied on knowing the camera's location in space, and the method was tested on adult subjects. This thesis explores the use of one lower-cost RGB-D camera without the need for other specialized equipment for ROI segmentation from depth.

Detecting periods of intervention in the NICU has not been very widely explored. The state-of-the-art in the field utilizes an RGB-D fusion model built on the VGG-16 convolution neural network [57]. This thesis explores the use of ViT and only depth data to achieve this same task, but with improved accuracy.

3 Data Acquisition

Data were collected from neonatal patients in the NICU at the Children's Hospital of Eastern Ontario (CHEO) following approval by the appropriate research ethics boards. The data was collected as part of a larger research initiative to develop non-contact patient monitoring methods and technologies. The dataset can not be released publicly due to the restrictions set by the research ethics board.

Figure 9 shows an example of the setup in the NICU environment. An RGB-D camera was placed above or around the patient's bed. The gold standard respiratory rate signals of the patients were recorded from the hospital patient monitors. 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.

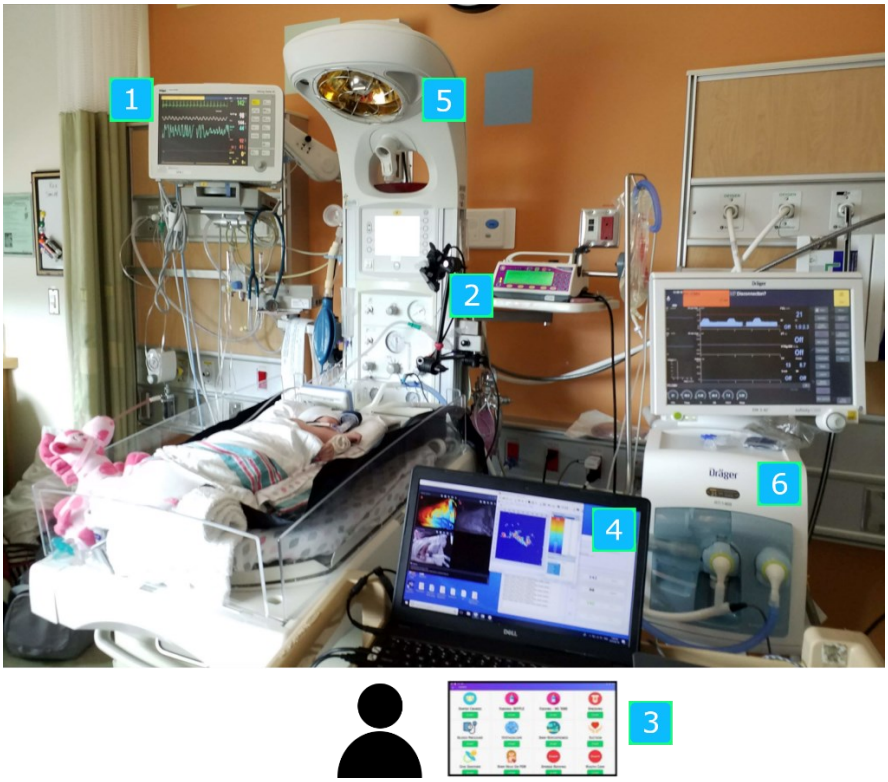


Figure 9: Overview of equipment setup: 1. Patient monitor, 2. RGB-D camera, 3. Bedside annotation application, 4. Data acquisition laptop, 5. Neonatal bed (overhead warmer), 6. Ventilator

3.1 Patient Monitor

The patient monitor used at the hospital was a Draegar Infinity Delta patient monitor. Custom Patient Monitor Data Import (PMDI) software, developed for the project, was used to import the gold standard respiratory rate data from the serial port on the monitor [59].

3.2 RGB-D Camera

The RGB-D camera used for the project was the Intel RealSense SR300 (Figure 10). The camera was chosen due to its small size and affordability. Recordings were captured at a resolution of 640x480 pixels at 30 frames per second. The cameras were placed such that the view planes were at non-uniform 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.

Figure 10: Intel RealSense SR300 with examples of RGB and color-mapped Depth patient data

3.3 Bedside Annotation Application

A custom mobile application was used to record annotations of events during the study. These events included clinical interventions, physiological events, alarms, and routine care. The annotations collected were used to find appropriate sections of the patient recordings for study in the upcoming sections of this thesis. The annotations were also used to label the 'periods of intervention' data used in Chapter 6.

3.4 Data Acquisition Laptop

The metadata along with the color and depth frames collected from the RGB-D camera reached multiple gigabytes per minute of recording. Therefore, data collected from the camera and patient monitor were saved to a dedicated data acquisition laptop with a 2 TB high-speed solid-state drive (SSD) before being transferred to a secure network attached storage (NAS) device at Carleton University.

3.5 Collection of Simulated Data

After the initial data collection stage, some simulated data was also collected. A neonatal mannequin (StandInBaby) was used along with the same RGB-D camera from the initial data collection (Intel RealSense SR300) to capture 600 still depth images (illustrated in Figure 11). This was done in an effort to further balance the number of datapoints of the non-intervention/intervention set of images. 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 the future (Figure 12).

Figure 11: StandInBaby [60] neonatal mannequin on the left; example simulated data collection scene on the right

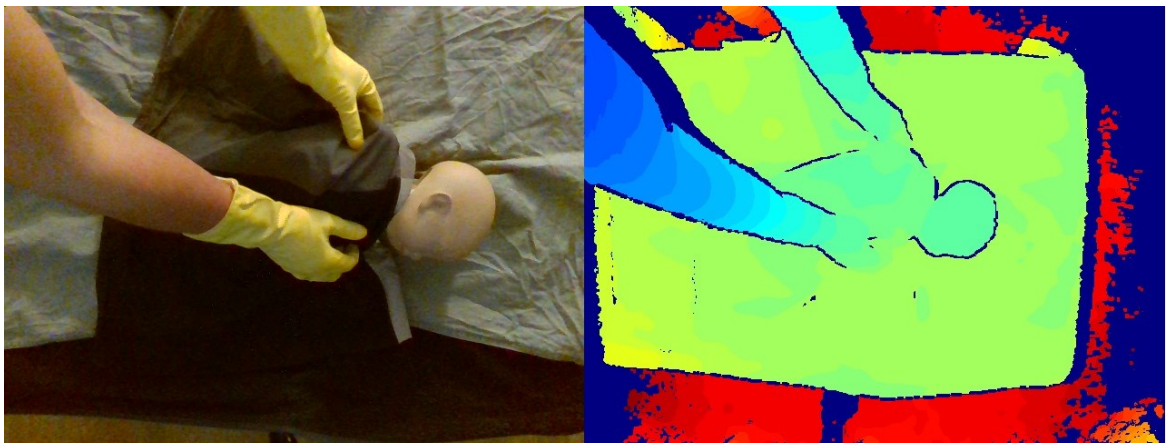


Figure 12: Example of simulated data. RGB image shown on the left, color-mapped depth image shown on the right

4 Perspective Transformation from Non-Uniform Camera Placement

4.1 Introduction

As has been mentioned in Section 1.3, it can be difficult to fit additional monitoring equipment into the NICU environment. Retrofitting a depth sensor (camera) to an NICU bed can lead to non-optimal camera placement, as in Figure 14 and Figure 15, where the camera may not be directly overhead of the patient and may be rotated with respect to the patient plane. This can be especially true in cases where the patient is undergoing phototherapy (Figure 13). Camera placement is secondary to patient care and must not interfere with other equipment nor clinical care or interventions.

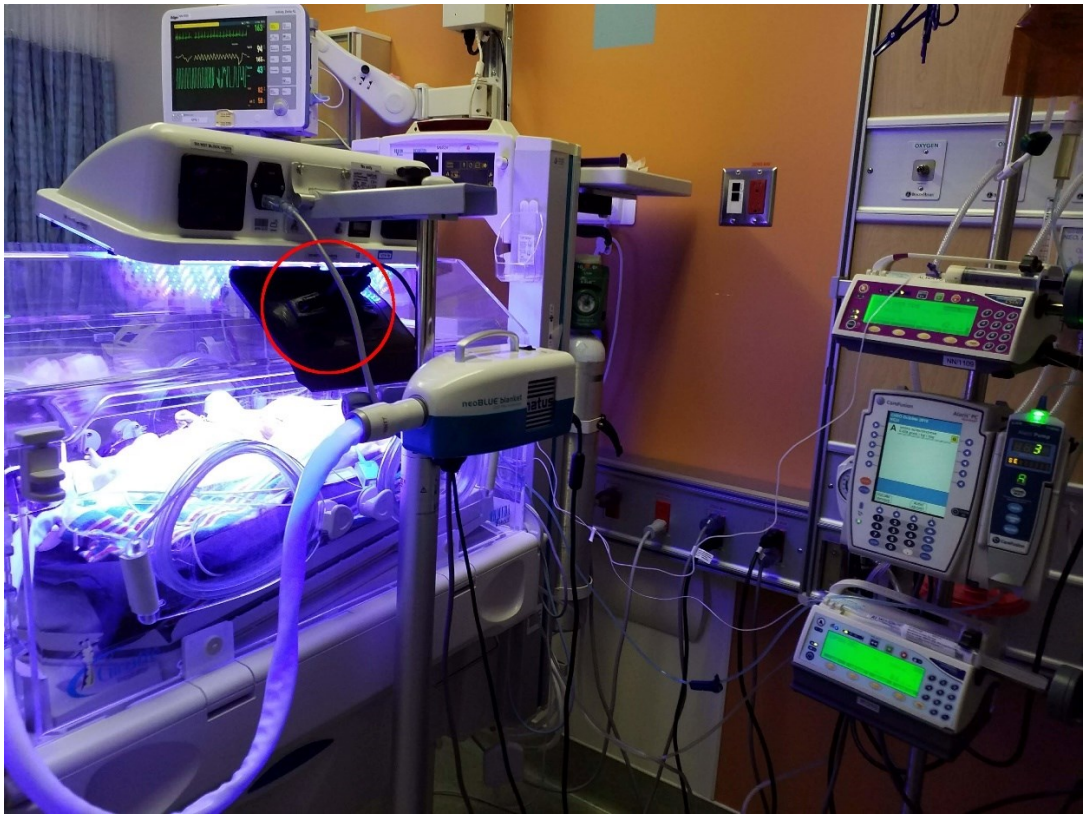


Figure 13: Patient undergoing phototherapy in the NICU. RGB-D camera highlighted with red circle (Reproduced from Figure 1)

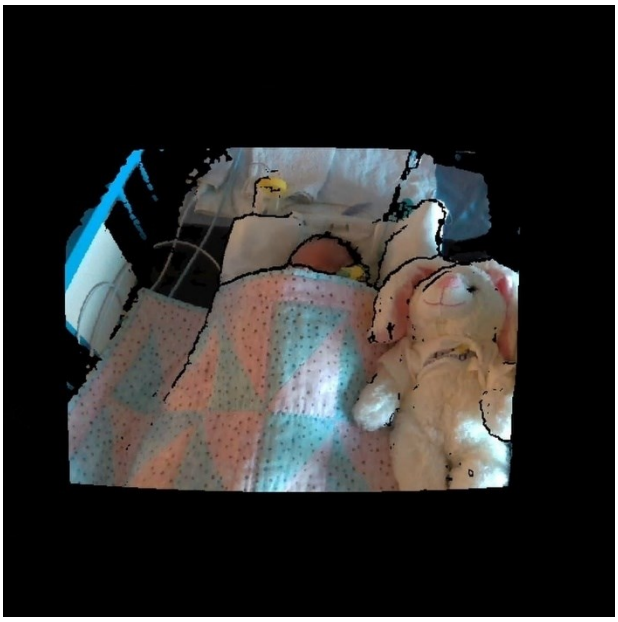


Figure 14: RGB image of patient in the NICU with non-optimal camera placement

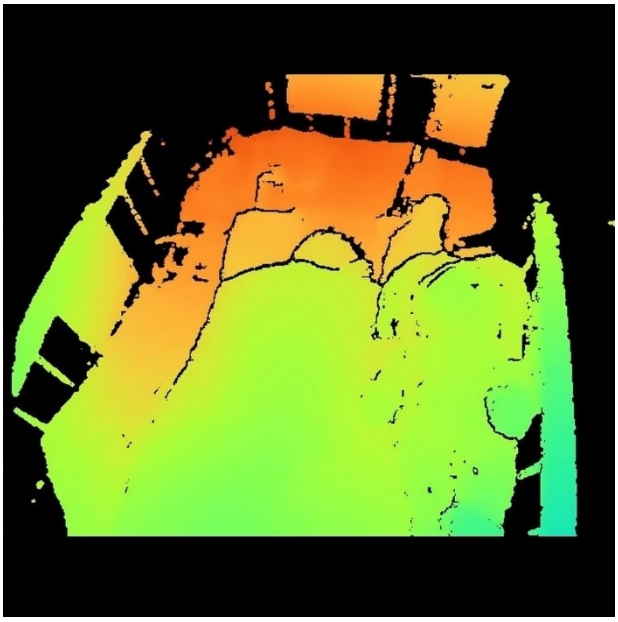


Figure 15: Depth image of a patient in the NICU with non-optimal camera placement showing greater distance to the far end of the bed

To account for the non-uniform placement of a depth camera in or around the patient's bed, the recorded depth video must be transformed (Figure 16). The view plane of the depth video must be shifted to appear parallel to the patient's bed. This results in a depth frame where

the pixels corresponding to the patient's bed are all approximately the same depth away from the camera (Figure 17).

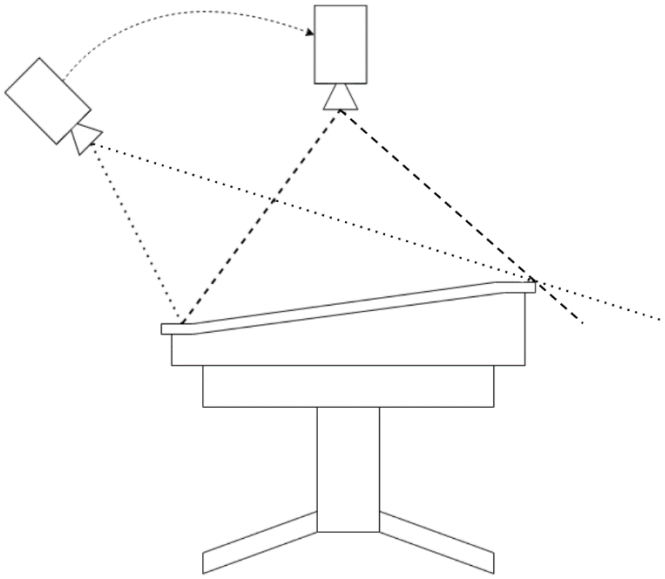


Figure 16: NICU bed with two different camera perspectives.

4.2 Methods

The required results were achieved by calculating a rotation matrix using three user-selected points on the surface of the bedding (Equation 7). Each pixel in a depth frame (Figure 15) is de-projected into a point cloud using its position in the frame, the depth from the camera, and elements of the camera's intrinsic matrix. Equation 3 finds the normal vector (N0) of the plane defined by three user-selected points (A, B, C). Equation 4 finds a normal vector (N1) parallel to the camera's view plane (where i, j, and k denote the x, y, and z axes respectively). The rotation axis (u) and angle (θ) can then be found by solving for a rotation that aligns N0 to N1 using Equation 5 and Equation 6 respectively. The rotation matrix can then be calculated from (u) and (θ) as in Equation 7. The 3D point cloud representation of the full depth frame is rotated using the rotation matrix before being projected back into an array of depth pixels (Figure 17). Following this process results in some pixels lacking depth information due to the nature of the rotation; thus, a dilation operation is applied to the frame to impute these missing depth values. To fit and test the perspective transform, the user was asked to select

four points in the RGB image that are on the bed's surface (i.e., four points that should be equidistant from the camera if it were placed directly overhead). A rotation matrix was calculated from each combination of three points and tested on the fourth in a 'leave-one-out' test. In an ideal transformation, the fourth point will have the same depth as each of the three other points since they would be found on the same plane (the bed). The rotation matrix that resulted in the highest agreement between the four points was selected for use.

Equation 3

$$N_0 = \frac{vAB \times vAC}{\|vAB \times vAC\|}$$

Equation 4

$$N_1 = 0i + 0j - 1k$$

Equation 5

$$u = \frac{N_0 \times N_1}{\|N_0 \times N_1\|}$$

Equation 6

$$\theta = \cos^{-1}(N_0 \cdot N_1)$$

Equation 7

$$R = \cos(\theta)\, I + \sin(\theta)\, [u]_\times + (1 - \cos(\theta))(u \otimes u)$$

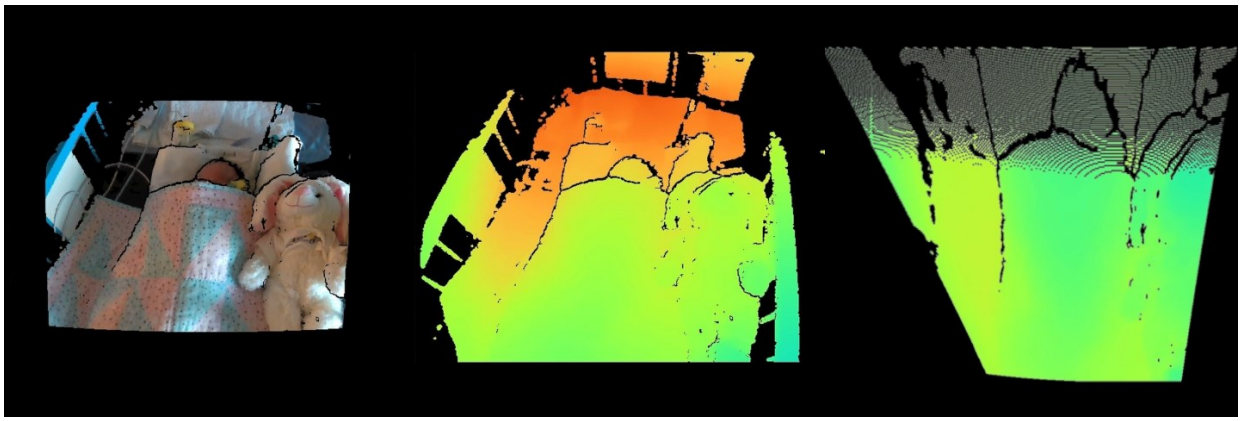


Figure 17: RGB image of patient on the left, original depth image of patient showing non-optimal camera placement in the middle, and corrected depth image of patient with uniform depth values after perspective transformation on the right

4.3 Results & Discussion

The algorithm was tested on 28 patients in various situations, camera angles, blanket coverage, and patient pose. The error of the fourth point chosen during the selection of the rotation matrix was used to evaluate the perspective transformation process. The mean absolute percentage error (MAPE) over all 28 patients tested was found to be 5.58% and the MAPE of each patient separately can be found in Figure 18. Although we do not know the exact requirement for the MAPE, our test explored camera angles up to 38.58° away from the optimal angle. We can conclude that the algorithm is robust to varying camera angles. Some patient recordings were restarted or interrupted due to clinical care or parent time, and the patient and camera were often repositioned in these instances. These restarts represent a second opportunity to evaluate the perspective transformation for that patient. Figure 19 shows a scatter plot of the estimated camera angle found when calculating the transformation matrix vs. the absolute percentage error measured using the fourth point of that transformation. In the analysis, it was expected to show a positive correlation between the two variables, as depth cameras tend to have a higher accuracy when the subject is closer. However, as can be seen in Figure 19, this trend did not materialise. This may be due to human error when selecting the four points on the bed, or because estimated angles were used rather than the actual angles at which the cameras were placed.

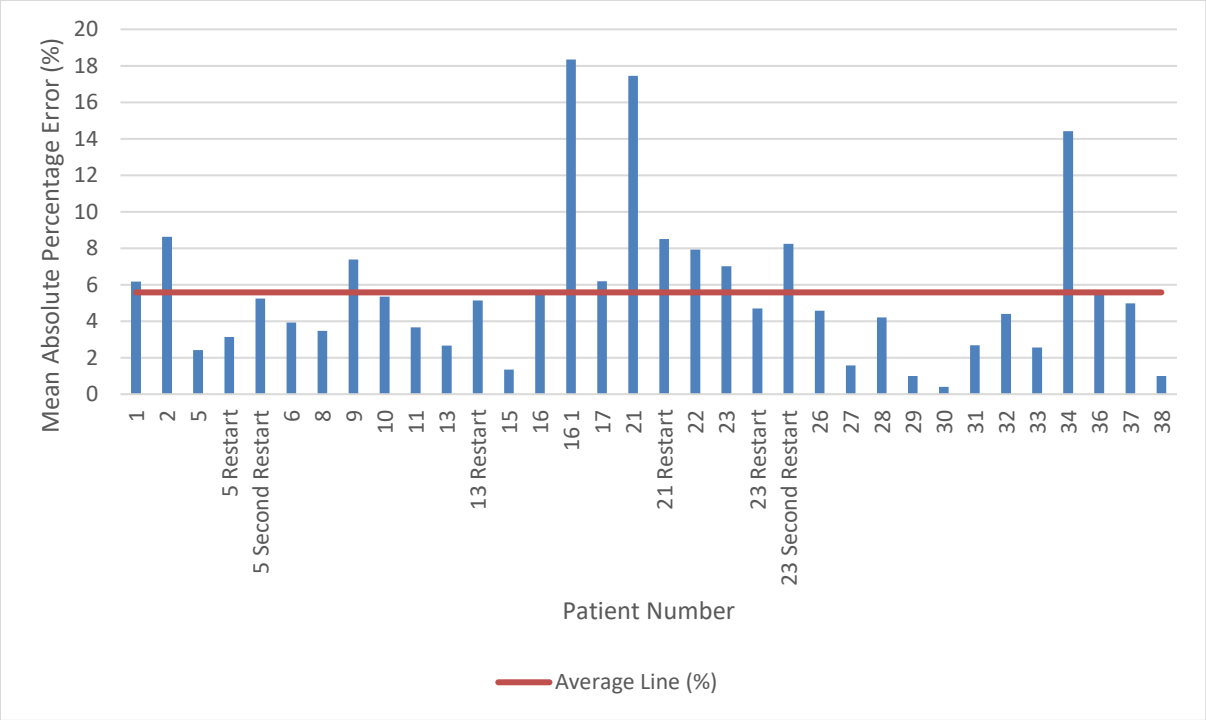


Figure 18: Mean Absolute Percentage Errors (MAPE) of all patients when three calibration points are used to fit the transform and a fourth point is used to evaluate the transformed depth value.

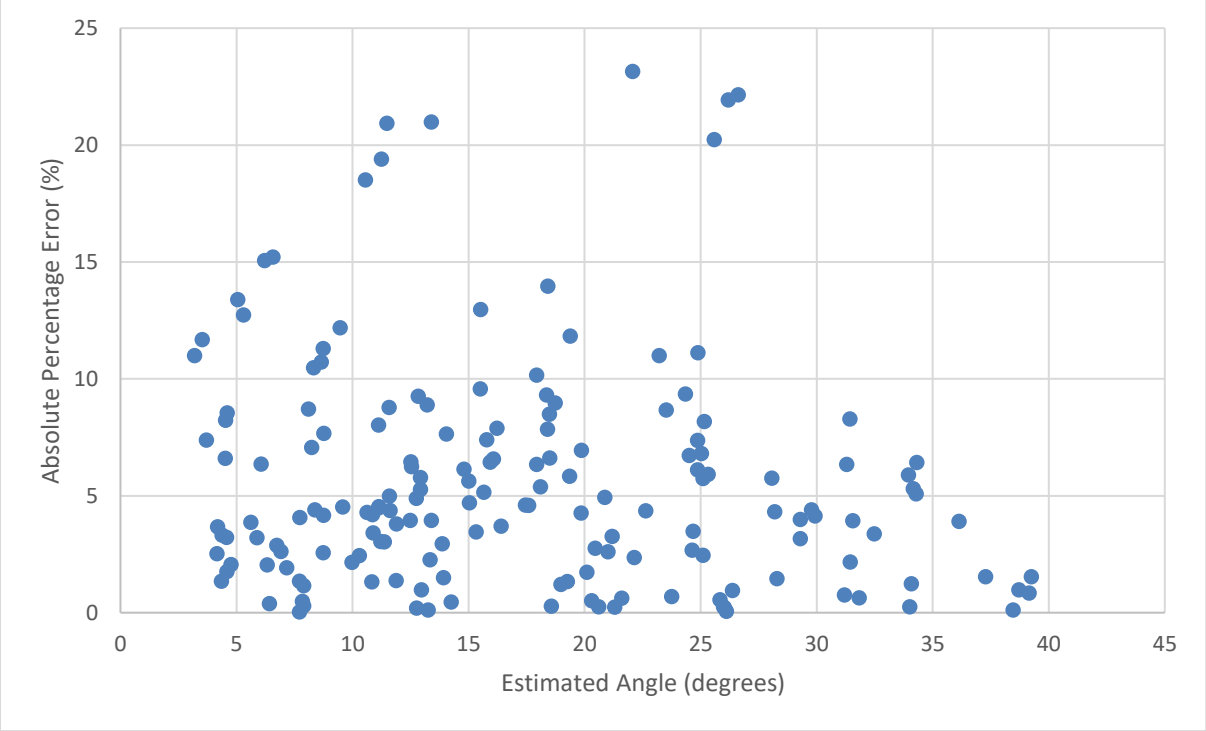


Figure 19: Graph of estimated angle when calculating rotation matrix vs absolute percentage error of the fourth point when compared to the three calibration points

4.4 Conclusions

By evaluating the perspective transformation on a number of patients in different scenes, it was shown to consistently correct the plane of the bed to be at a uniform distance away from the camera. Applying the perspective transformation on the patient data is in the hope of improving downstream tasks including region-of-interest selection, respiratory rate estimation, and intervention detection. Although we do not know the actual requirements for MAPE when applying the perspective transform, the results seem to verify the capability of the method, and the practical impact of applying the perspective transformation will be evaluated in the following chapters on two such downstream tasks. Further work can be done to increase robustness of the method to wrinkles and anomalies in the scene, including allowing the user to select a region rather than a point when selecting the initial points to calculate the rotation matrix. By averaging the depth over the selected region and using the centroid of the selection as the 2D coordinates, we may be able to neutralize the effect of any ripples or wrinkles in the bedding on the depth of the selected points.

5 Respiratory Rate Estimation Pipeline

5.1 Introduction

Using the perspective transformation method presented in the previous chapter, we can correct the angle at which the patient is seen by the camera. We here leverage such corrected depth maps to build an automatic region-of-interest segmentation method and subsequently estimate respiration rate. This ROI selection method is built such that it does not depend on segmentation of image regions exhibiting skin tone pixel colours. Such a method is important to address varying skin tones and cases when a patient is covered by a blanket or quilt. Once the ROI is selected, we develop a pre-processing pipeline to improve non-contact respiratory rate estimation when working with the depth channel alone. The method adapts and extends the methods described in [55] that did not study newborn patients nor non-uniform camera angles. We evaluate this pre-processing pipeline by implementing an RR estimation method using the pipeline and comparing its performance against an RR estimation method without the perspective transformation driven ROI selection using the actual neonatal patient data collected from the patient monitor as the ground truth.

5.2 Methods

5.2.1 Region-Of-Interest Selection

Region of interest selection was automated by examining successive cross-sections of the transformed depth frame. The point with the greatest depth of the three calibration points selected during the perspective transformation process is used to threshold the depth frame and filter out the majority of the patient's bed from the scene. The remainder of the frame is then cross sectioned into twenty slices between the thresholded depth and the point closest to the camera (least observed depth). A contour finding algorithm [61], [62] is then used to detect contours enclosing unfiltered data.

The head is found first, under the assumption that it will form a semi-spherical shape. The resulting contours are used to find semi-spherical shapes in the scene by iterating through each of the twenty slices (from the bed depth to the highest point in the scene) and building sets of contours with a certain level of circularity that contain smaller contours within them. The semi-sphere with the most circular contours in its top-most depth slice is chosen as the shape corresponding to the head.

The torso region is then chosen by building subsets of contours with a lower circularity threshold and taking the one with the largest area, on the condition that no part of the contour crosses into the selected head semi-spherical region. Figure 21 and Figure 22 show the visual output of the automated ROI selection process, with Figure 20 as the RGB reference. Two concentric contours of the selected head semi-sphere can be seen outlined in blue, and two concentric torso cuboids in red.

Figure 20: Reference colour image

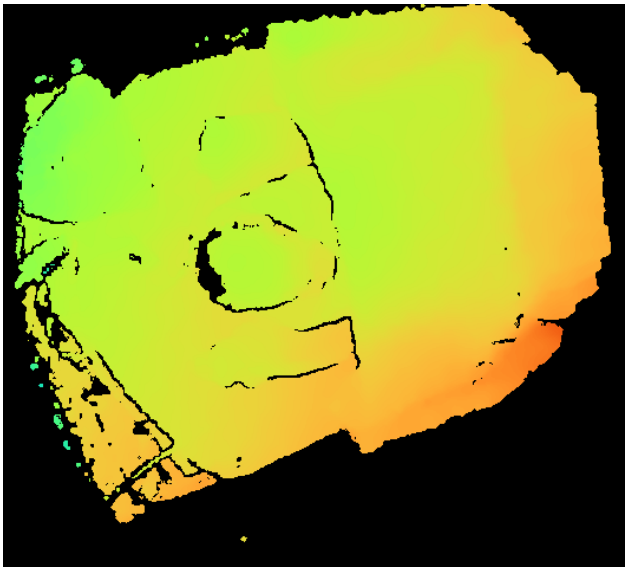


Figure 21: Original depth image

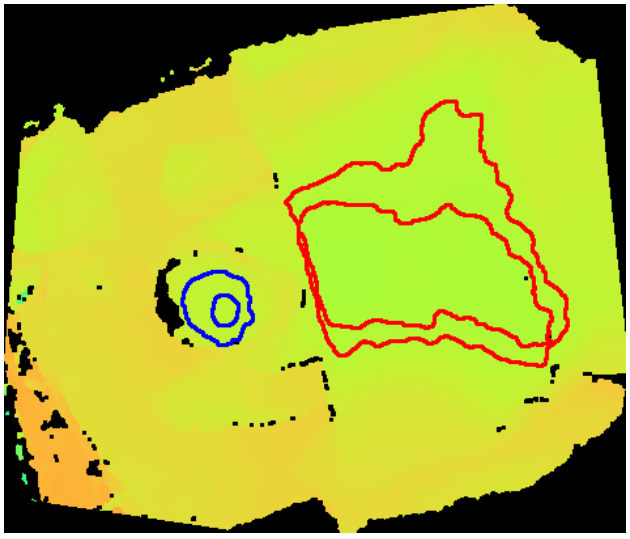


Figure 22: Depth image after perspective transformation with automatically selected ROI semi-sphere and cuboid illustrated

A number of anthropomorphic checks are used to improve the method's performance. The contours are accepted or rejected based on criteria looking at the minimum/maximum area and the degree of circularity (Equation 8). The head semi-sphere's largest contour needed to have a minimum area of 300 pixels and a maximum of 30000. This range was chosen to account for a wide range of ROI sizes in recordings, since the patient was not always a set distance away from the camera. The threshold is expected to generalize to other datasets

when expressed as a percentage of the overall image area. The contour's circularity was also limited to be greater than 0.50. The torso cuboid's largest contour needed to have an area larger than that of the head, and no maximum area was imposed. The distance between the closest torso contour point and head contour point was also checked to make sure that it is less than the radius of an ellipse that was found to best fit to the head contour. These criteria were found empirically by testing on some separate patient data. Due to this, the ability of the method to generalize to different datasets needs to be tested in the future.

Equation 8

$$circularity = 4\pi \frac{area}{perimeter^2}$$

5.2.2 Respiratory Rate Estimation Methods

Two algorithms for estimating respiratory rate were chosen to demonstrate the effectiveness of the perspective transformation and ROI selection pre-processing stages. The methods were adapted from the work of Bekele [63]. A single signal over each of the investigated time segments was first derived from the depth frames. The signal was comprised of the mean depth of the torso ROIs for each of the frames in the recordings over time. A band-pass filter in the form of a second-order Butterworth filter is then applied to the signal with cut-off frequencies of 0.35 Hz and 1.80 Hz. The filter is applied in order to eliminate any high or low frequency signal artifacts, using a passband that covers the range of neonatal respiratory rates where 0.35-1.80Hz corresponds to 21- 108bpm. The filter also removes the mean loading of the signal, de-emphasizing the effect of relatively unchanging pixels in the depth frame.

The RR estimation method in the time domain involves extracting peaks and calculating the period of the signal. Equation 9 shows the formula used for computing the respiratory rate in breaths per minute (bpm), where n is the number of peaks found, Fs is the sample rate of

the average depth, and last and first are the sample numbers of the last peak found and first peak found respectively.

Equation 9

$$\mathrm{RR} = \frac{(\mathrm{n} - 1)\mathrm{Fs}}{\mathrm{last} - \mathrm{first}}$$

The second method estimates the RR in the frequency domain by finding the power spectral density of the signal and selecting the frequency with the largest power contribution. It is assumed that the largest power contribution is attributable to the breathing signal, since the sections of the recordings that were selected had minimal movement and other factors affecting the scene. Future work will include more robust filtering to remove low frequency motion artifacts. The formula used for computing the RR from the power spectrum of the signal (P_{xx}) and the frequency with the highest power contribution (fp) can be seen in Equation 10.

Equation 10

$$RR = f_p \times 60\, bpm,\ where\ f_p = argmax_f P_{xx}$$

5.3 Results

5.3.1 Region-Of-Interest Selection

To evaluate the accuracy of the automated ROI selection algorithm, I manually determined a gold-standard ROI that corresponds to the patient's head and torso regions from the RGB image data taken at the same time as the depth frames. The method was tested on six frames for each patient with varying levels of blanket coverage, camera angles, and patient pose. The Sørensen–Dice coefficient [64], [65] (Equation 11) and Jaccard index [66] (Equation 12) were used to evaluate the performance of the ROI selection algorithm. Both metrics quantify the union over the intersection, where X and Y refer to Boolean masks of the automatically segmented frames and the manually segmented frames respectively (as illustrated in Figure 23). An average Sørensen–Dice coefficient of 0.62 and Jaccard index of 0.46 were found for

the torso ROIs over all four patients. The results for each patient individually are found in Table 1. Table 2 presents the Sørensen–Dice coefficient and Jaccard index of the head ROIs for the same frames. Patients 1-4 in Table 1 correspond to patient 6, 13, 26, and 37 from Figure 18 in Chapter 4.

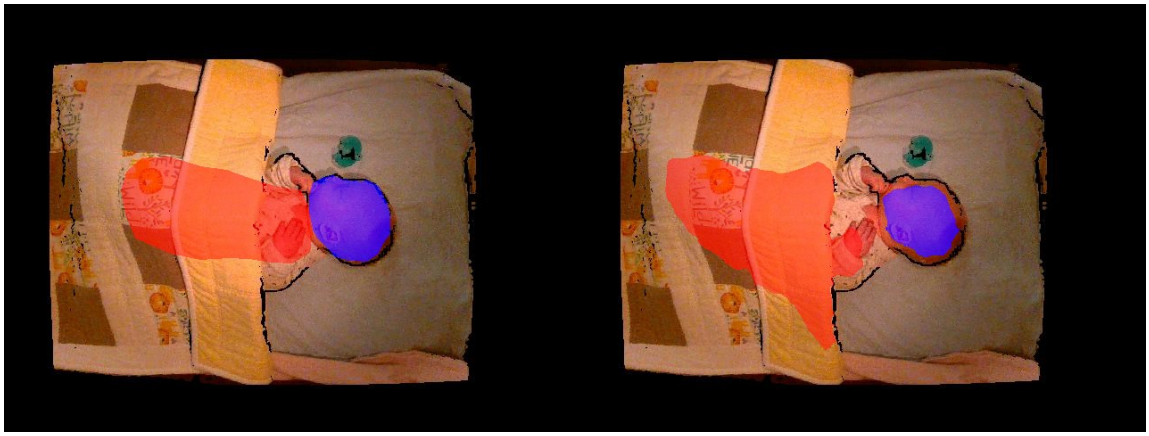


Figure 23: Example ROI selection evaluation masks. Manually selected head and torso ROI masks in blue and red respectively on the left, automatically estimated head and torso ROI masks in blue and red respectively on the right.

Equation 11

$$Sørensen\text{–}Dice = \frac{2|X \cap Y|}{|X| + |Y|}$$

Equation 12

$$Jaccard = \frac{|X \cap Y|}{|X \cup Y|}$$

Table 1: Automatic torso ROI selection method performance evaluated against manually selected ROI ground truth.

Patient	Sørensen–Dice	Jaccard Index
1	0.6826	0.5193
2	0.6108	0.4548
3	0.6133	0.4500
4	0.5968	0.4292

Table 2: Automatic head ROI selection method performance evaluated against manually selected ROI ground truth.

Patient	Sørensen–Dice	Jaccard Index
1	0.293103	0.261381
2	0.366793	0.29949
3	0.237261	0.18899
4	0.199708	0.142768

5.3.2 Respiratory Rate Estimation Comparative Performance

The pre-processing pipeline was tested on data recorded from four different patients. In all four cases, the depth-sensing camera was placed in a sub-optimal position, with non-uniform rotation, angle, and translation with respect to the patient's bed. The cameras were placed at ~5-28 degrees away from the optimal position. For two of the patients, the cameras were repositioned during the recording due to clinical interventions. One of the patients was recorded in a dimly lit environment and another in an environment with the lights off. All the patients were clothed during the majority of the recordings, and three were covered with a blanket or quilt at different points during the recordings.

The time- and frequency-domain respiratory rate estimation methods were applied to the signal derived from the average depth of the selected region of interest, as well as on the unaltered depth frame as a baseline test. A five-minute portion of each patient's recording was chosen for the evaluation. The five-minute segments were chosen to exclude periods of high patient movement, medical interventions, and obstructed camera views. Non-overlapping sliding windows of 10 seconds were used for evaluation. The percentage of acceptable estimates (PAE) was defined as the proportion of the RR estimates that resulted in a mean absolute error of 5 bpm or less, as used in [21]. An improvement in the percentage

of acceptable estimates can be seen when using either the time domain method or the frequency domain method for RR estimation. The full results can be seen in Table 3 and Table 4. A substantial improvement in the PAE (3.60% to 13.47% in the frequency domain and 6.12% to 8.97% in the time domain) can be seen.

Table 3: Percentage of acceptable respiratory rate estimates (MAE < 5 bpm) using the frequency domain method over a window of 10 seconds.

Patient	PAE When Estimating Over the Whole Frame	PAE When Estimating over the Segmented ROI
1	2.86%	22.86%
2	0.0%	0.0%
3	11.54%	15.38%
4	0.0%	15.63%

Table 4: Percentage of acceptable respiratory rate estimates (MAE < 5 bpm) using the time domain method over a window of 10 seconds.

Patient	PAE When Estimating Over the Whole Frame	PAE When Estimating over the Segmented ROI
1	0.0%	0.0%
2	5.71%	17.14%
3	0.0%	0.0%
4	18.75%	18.75%

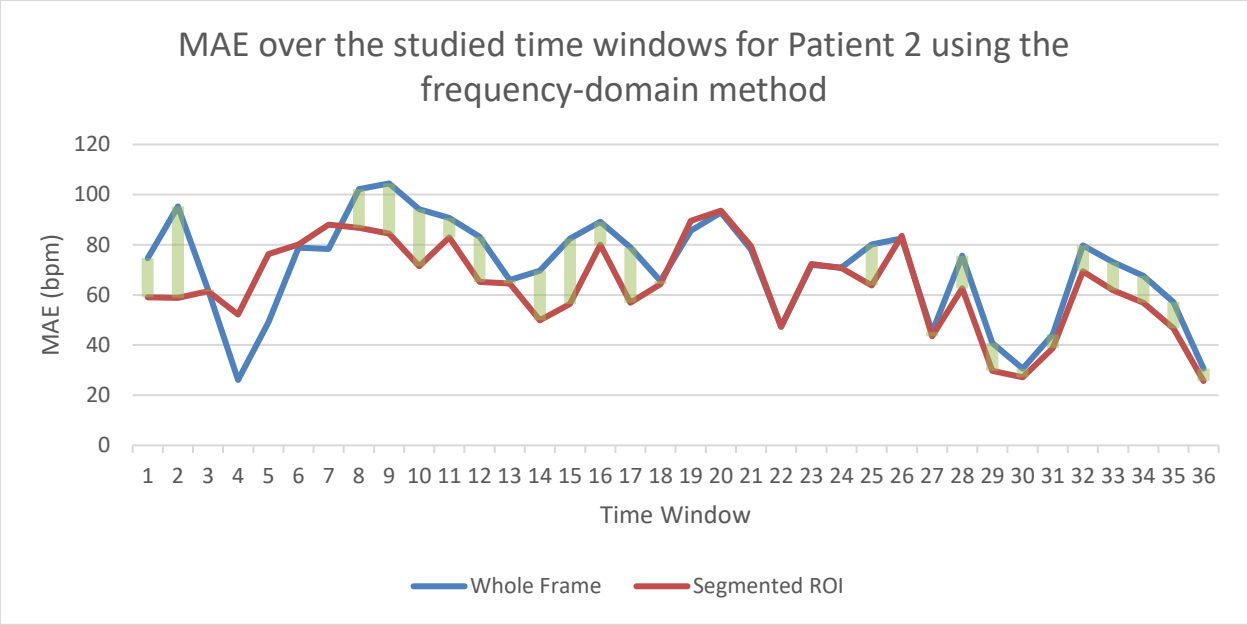


Figure 24: The mean absolute error of patient 2 respiratory rate estimated using the frequency-domain method from the whole frame in blue, and from the segmented ROI in red. Green bars note the improvement in absolute error when using the segmented ROI for estimation.

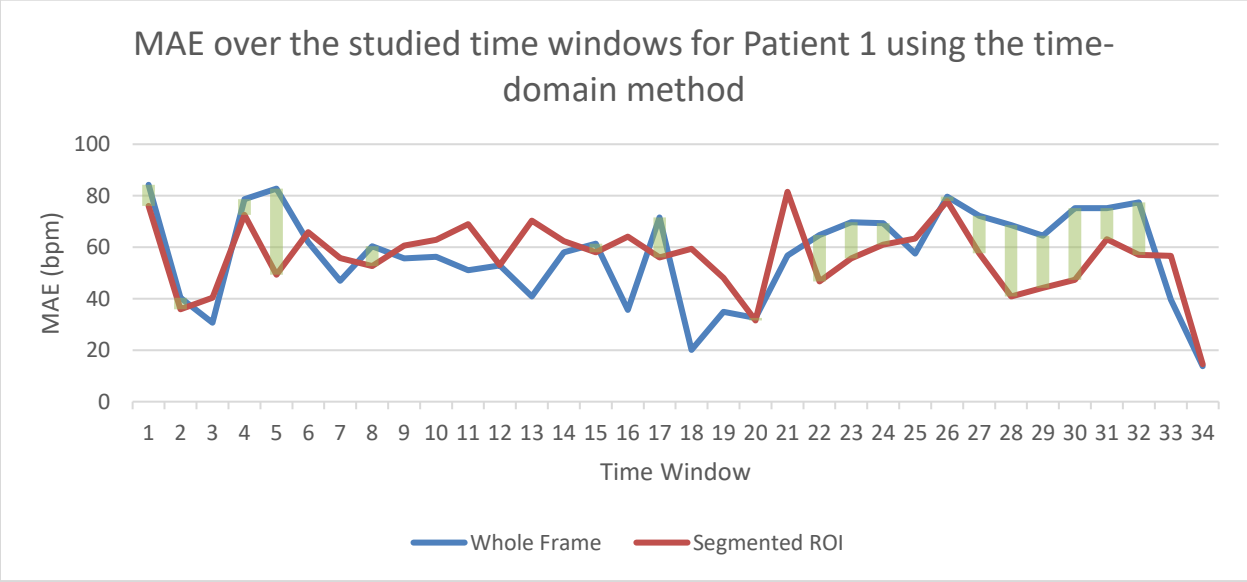


Figure 25: The mean absolute error of patient 1 respiratory rate estimated using the time-domain method from the whole frame in blue, and from the segmented ROI in red. Green bars note the improvement in absolute error when using the segmented ROI for estimation.

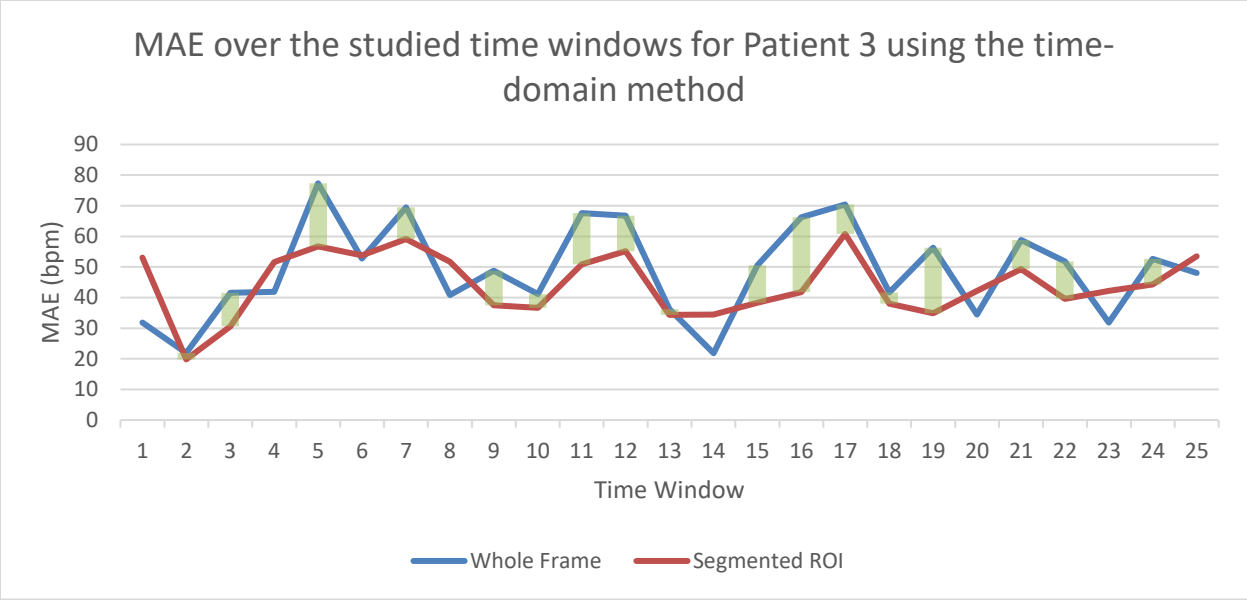


Figure 26: The mean absolute error of patient 3 respiratory rate estimated using the time-domain method from the whole frame in blue, and from the segmented ROI in red. Green bars note the improvement in absolute error when using the segmented ROI for estimation.

It can be seen in Table 3 and Table 4 that the PAE of the RR estimates from the majority of the tested patient's recordings increased when utilizing the pipeline, except three outliers that were stagnant at 0% PAE. When applying the frequency-domain method on the recording from Patient 2, we can see from Figure 24 that, although the PAE did not change when the pipeline is used (as the MAE is consistently above 5 bpm), there is an improvement in MAE over most segments of the recording. The same can be said when applying the time-domain method on Patient 1 (Figure 25) and Patient 3 (Figure 26). The remaining MAE figures can be found in Appendix A. Although the results of each of the patients differed, their mean absolute percentage error when calculating the rotation matrix to perform the perspective transformation are all below the average. Patients 1-4 resulted in MAPE of 3.92%, 2.66%, 4.58%, and 4.97% respectively with angles ranging from 6.6° to 25.6°.

Although the PAE either improved or stagnated, and the graphs of the MAE over the time windows in the recordings show a general improvement, the absolute errors after applying

the pipeline are not acceptable for use in a medical setting. Future work can include improving the ROI selection method and verifying that the improved ROI will increase the performance of the RR estimation algorithms (by comparing against RR estimates from a gold standard ROI). Further, the PAE may also improve with the use of different depth-based RR estimation algorithms.

Although the performance is not as high as that of some other RR estimation methods utilizing RGB data, using depth cameras presents several advantages. Depth-based methods do not require the environment to be brightly lit (which is important when caring for premature infants who require a low-stimulus environment) and depth data can be more privacy-preserving than RGB video. Further work is expected to increase depth-based RR estimation to the same accuracy as RGB-based RR estimation.

5.4 Conclusions

The ROI selection method was tested on patients with varying levels of blanket coverage, camera angles, and patient pose. These factors did not seem to have a substantial impact, as can be seen from the results of the similarity tests in Table 3. The results in Table 3 & Table 4 show that there does not appear to be a correlation between the accuracy of the head ROI and the torso ROI. This might be because the although the head ROI is found first, it is only used to eliminate some contours from being candidates for the torso. Therefore, if the head ROI's general position is found, the exact shape and it's similarity to the manually selected head ROI are not key to finding the torso ROI. The ROI method may underperform for some patients when other objects are found in the scene. An example of this can be seen in Figure 27.

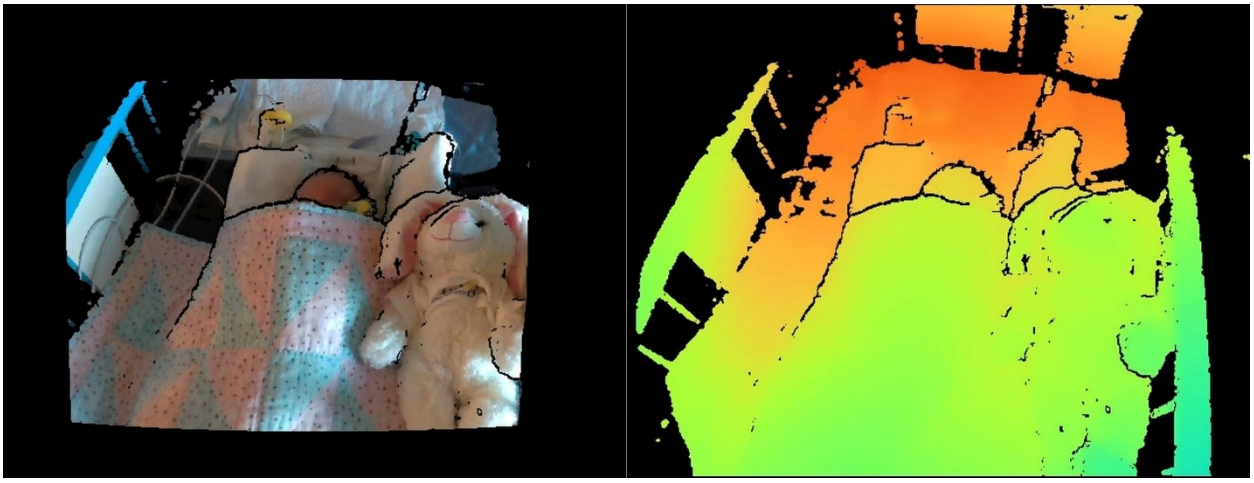


Figure 27: Example of a more challenging scene for ROI selection. Patient is in the middle of the scene; stuffed animal can be seen to the patient's right.

The performance of the respiratory rate estimation algorithms was shown to improve with the use of the perspective transformation and ROI selection pipeline. This was found to be applicable to both the time-domain and frequency-domain methods. At the same time, the percentage of acceptable estimates for both methods were not found to be useful for practical applications. The exploration of different RR estimation methods as well as improving the performance of the ROI selection method may increase the percentage of acceptable estimates further.

6 Depth-Based Intervention Detection

6.1 Introduction

As mentioned in Section 2.10, a patient may experience multiple periods of clinical intervention or routine care throughout their time at the NICU. These interventions can include a nurse or other practitioner reaching into the scene to replace sensors, take readings, change a diaper, feed the patient, or otherwise move the patient. In this chapter, we develop a model to detect these periods of intervention. Detecting such interventions is useful for a number of reasons. For example, when estimating vital signs, estimation may be paused or change sensor modalities during interventions, or patient monitor alarms may be silenced automatically during interventions since a clinician is already attending to the patient. Lastly, 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 occur frequently in the NICU.

6.1.1 Intervention Detection Dataset

Still images are extracted from the patient recordings every 30 seconds and labelled as either 'Intervention' (positive) or 'No Intervention' (negative), using the annotations collected as described in Chapter 3. This resulted in 14,892 images in total, 1,260 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 28). We compare the results of the intervention detection method proposed in Section 6.2 with the baseline methods described in Section 6.3.

Figure 28: Example frames of 'No Intervention' on the left and 'Intervention' on the right

The difficulty of intervention detection from depth data can sometimes be misrepresented. Looking at Figure 28, one would assume that the difference in the depth frame between the nurse's hands and the patient/bed would be apparent. In Figure 29, an intervention frame can be seen that is more challenging to classify by looking only at the depth channel (on the right). If the nurse or other medical practitioner's hands are lower or on the patient's bed, the difference in depth can be small enough to require more advanced methods than merely comparing the average depth over the whole scene between frames.

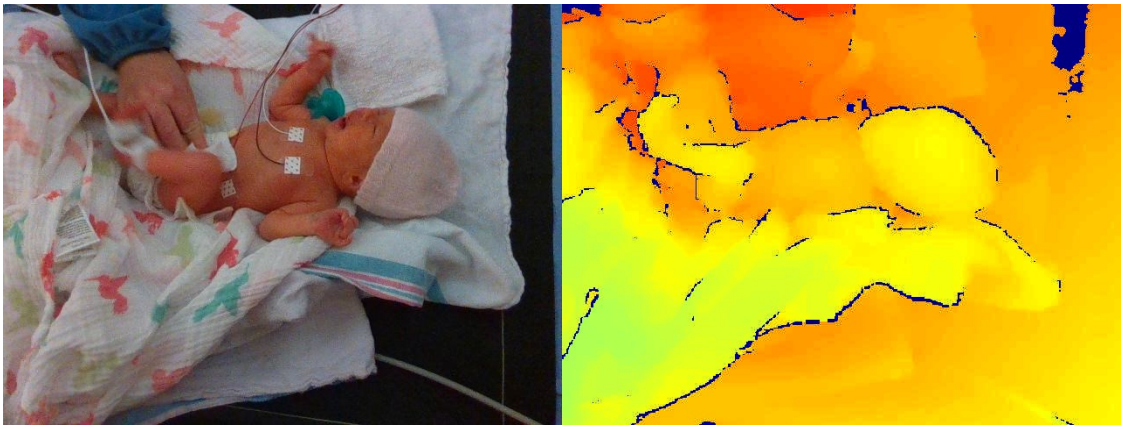


Figure 29: Example of more difficult 'Intervention' class frame. RGB image on the left, corresponding depth frame on the right

6.2 Proposed Method

Vision Transformers (ViT) have demonstrated a high capability in image classification tasks since their introduction in [23]. We propose the use of a ViT pre-trained on the ImageNet dataset [29] and fine-tuned on a subset of our own set of 14,892 images. The model architectures were implemented using the PyTorch Image Models library [67]. Two model sizes with similar architecture but different numbers of trainable parameters were chosen, 'vit_tiny_patch16_224' and 'vit_base_patch_16_224', with ~5.4 M parameters and ~85 M parameters, respectively. Each of the models takes as input images with a resolution of 224x224 pixels and divides them into 16x16 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.

The models were trained with a mini-batch size of 16 and learning rate of 0.01 for a maximum of 15 epochs. Stochastic gradient descent with a momentum of 0.9 was chosen as the optimizer. The models were evaluated using 5-fold cross-validation, repeated 5 times. Each of the input images were resized to 224x224 pixels (changing the aspect ratio from 4:3 to 1:1) and the training sets were also randomly rotated (between 0 and 360°) and flipped (horizontally and vertically).

Along with the size of the model, the effect of three other variables on the performance of the models were also explored. These variables are described in the upcoming Sections 6.2.1, 6.2.2, and 6.2.3 and a summary can be seen in Table 5.

6.2.1 Simulated Data

Since the data collected from the NICU contains longer periods 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 described previously (Section 3.5). 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.

6.2.2 Perspective Transformation

The effect of perspective transformation (presented in Chapter 4) on the performance of the models was explored. The perspective transformation process was shown to facilitate the use of a rule-based ROI selection algorithm (Chapter 0). 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 the four registration points separately 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 this transformed data as the input. Models were trained with and without perspective transform to investigate its effect on intervention detection accuracy.

6.2.3 HHA Encoding

ViT are not typically trained from scratch for specific image classification tasks. Rather, ViT models are typically pre-trained on large datasets using self-labelled techniques, such as masked auto-encoding (MAE) [68]. Pre-trained ViT 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 [69] and other groups. As these models have been pre-trained on 3-channel RGB images, there is latitude as to how the 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.

HHA encoding is a method of encoding 1-channel depth data in 3-channels. 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). The process for HHA encoding depth images was fully described in Section 2.6.

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, B were summed into a single channel.

Table 5: Summary of Vision Transformer experiments

Experiment	Model Size	Simulated Data	Perspective Transformation	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

6.3 Baseline Methods

The models were compared against the best-performing intervention detection model proposed by Souley Dosso *et al.* in [57]. Specifically, the model chosen as the baseline was the multi-modal RGB-D fusion model exhibiting a high average sensitivity, specificity, and accuracy over the 5-fold cross-validation. Their exclusively depth-based model was also

chosen for comparison, since it depends on the same modality as the proposed model, although it resulted in lower scores over all performance metrics tested. The models were built on the VGG-16 [58] convolutional neural network, pre-trained on the ImageNet dataset [29], and finetuned on the same intervention detection dataset described in Section 6.1.1. The input layer for the depth-based model was stripped of two of its three channels, leaving the pre-trained weights from one channel for fine-tuning. Multiple RGB-D models were tested, taking the best performing results. Early fusion, middle fusion, late fusion, and image fusion were all used. Early fusion is when the RGB and depth images are fed separately to the model and merging occurs after the first convolutional layer. Late fusion describes the merging of RGB and depth data after the final convolutional layer. Merging in middle fusion occurs after any other convolutional layer in the network. Image fusion was defined as fusing the data before inputting it to the model (Figure 30). The same cross-validation splits were used to enable direct comparisons between all models. Table 6 shows a summary of the metrics of the relevant comparison models reproduced from [57].

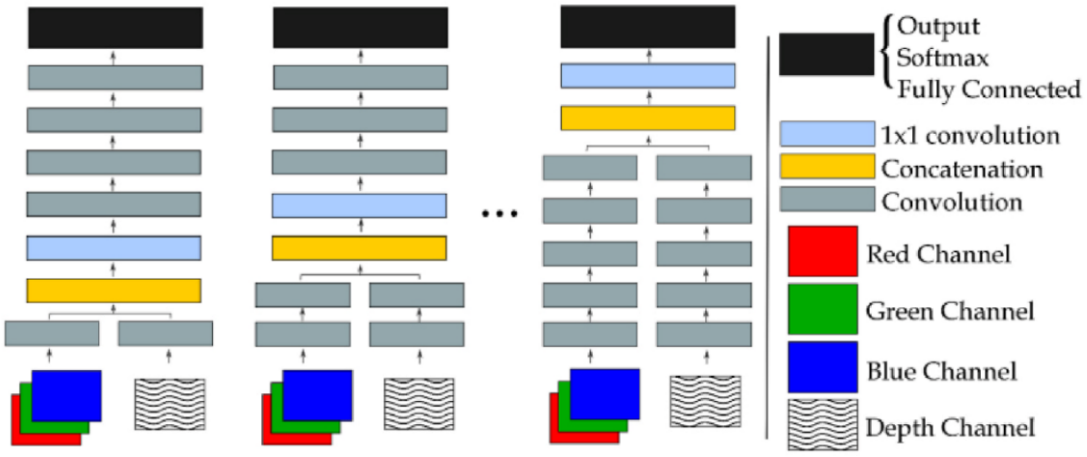


Figure 30: Architecture of baseline RGB-D fusion models (reproduced from [57])

Table 6: Summary of results from baseline comparison models

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%

In addition to the original RGB-D fusion and depth-based models presented in [57], the depth-based models were also evaluated using the variables outlined in Sections 6.2.1, 6.2.2, and 6.2.3. This enabled direct comparisons between the vision transformer models and the depth-based VGG-16 models for each of the variables tested.

6.4 Results and Discussion

Each of the models was evaluated using 5-fold cross-validation repeated five times. Each fold contained data from unique patients, leaving data from 5 or 6 patients as the test set each time. The frames were extracted at the same time points in the videos as the data used in [57] to enable direct performance comparisons against the chosen baseline models. For experiments where simulated data was 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) (Equation 13-Equation 18). An Analysis of Variance (ANOVA) test was run on the results from the proposed models to determine the statistical significance of the effects of the different variables. As a secondary test, the results of the repetitions of each fold were collapsed into each other by calculating the average of each metric, before running another ANOVA test. This meant that the number of records was reduced from 400 (5 folds x 5 repetitions x 16 combinations of variables) down to 80 (Average of the repetitions of the 5 folds x 16 combinations of variables).

Equation 13

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

Equation 14

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

Equation 15

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

Equation 16

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

Equation 17

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

Equation 18

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

6.4.1 Comparison Between Baseline Models and Proposed Models

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 shows an improvement over all tested metrics except sensitivity, where it shows a slight decrease. The 'base' ViT model shows a further improvement over all metrics. Results are summarized in Table 7 and Figure 31. To determine whether the improvement in results observed when moving from the 'tiny' model to the 'base' model is statistically significant, we performed an Analysis of Variance (ANOVA) test over each of the metrics. A p-value of less than 0.05 indicated a statistically significant difference in all resulting metrics from the two model sizes. For the secondary ANOVA test

after collapsing the repetitions of each fold, the model size was found to have a statistically significant effect on only on MCC.

Table 7: 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%

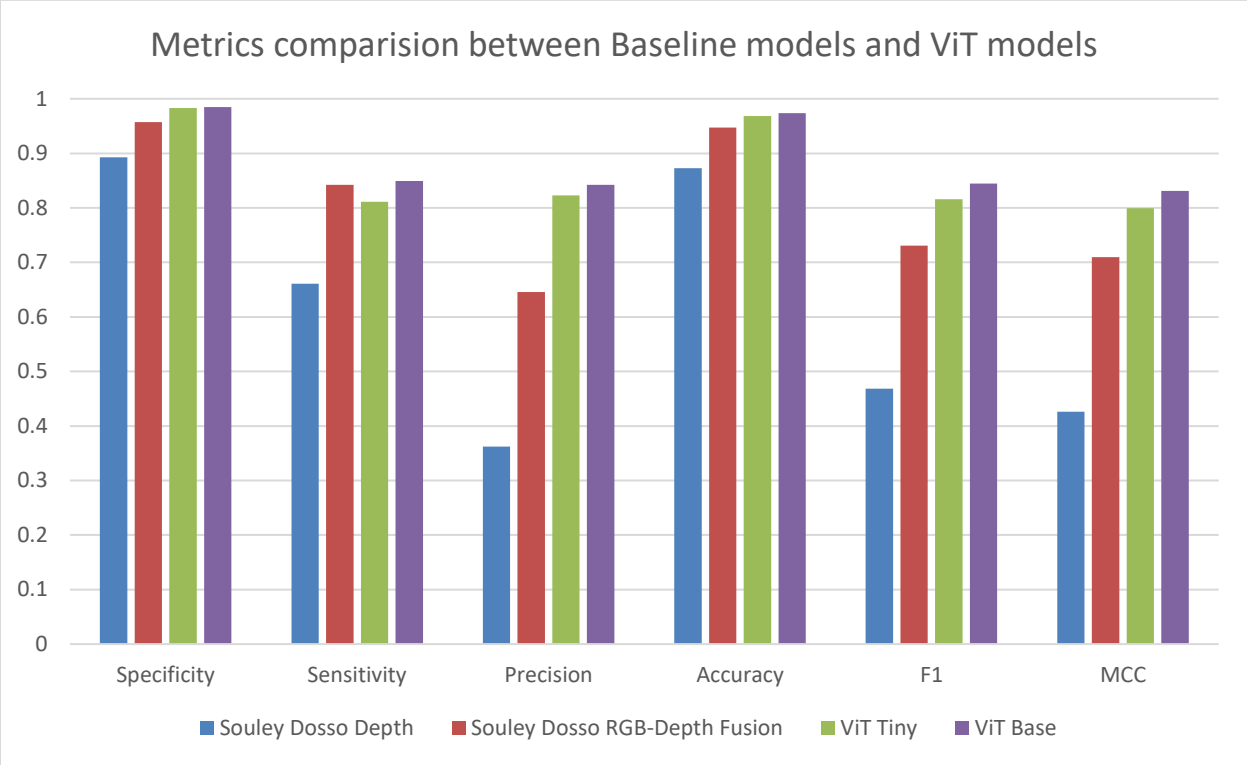


Figure 31: Specificity, sensitivity, precision, accuracy, F1-score, and MCC for Baseline models, ViT Tiny, and ViT Base

6.4.2 Effect of Simulated Data on Model Performance

After observing the higher performance of the ViT models on the same data as the baseline models, the tests were repeated after adding the simulated data into the training folds. These

results are shown in Table 8 and Figure 32. Relative to the results in Table 7, the performance decreased with the addition of the simulated data over all metrics, except sensitivity. Whereas the decrease in performance was shown to be statistically significant, the increase in sensitivity was not. This result was unexpected as the addition of the simulated data decreased the imbalance in the training dataset. The decrease in performance could be attributed to the possibility that the simulated data might not accurately represent the positive class as found in the NICU dataset. The secondary ANOVA test did not find a statistically significant difference in the results when utilizing the simulated data. This may be due to the lack of a sufficient number of simulated data points. A larger set of simulated data that further balances the classes may have a larger impact on the results of the models. When looking at the results of the VGG-16 model, the addition of simulated data showed an increase in specificity and accuracy and a decrease in all other metrics compared to the original VGG-16 model.

Table 8: Summary of results from 'tiny' and 'base' vision transformer models with simulated data

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
ViT Tiny with Simulated Data	97.92%	80.32%	78.43%	96.43%	79.28%	77.39%
ViT Base with Simulated Data	97.70%	86.38%	77.96%	96.76%	81.78%	80.24%
VGG-16 with Simulated Data	98.07%	39.19%	65.45%	93.09%	48.94%	47.26%

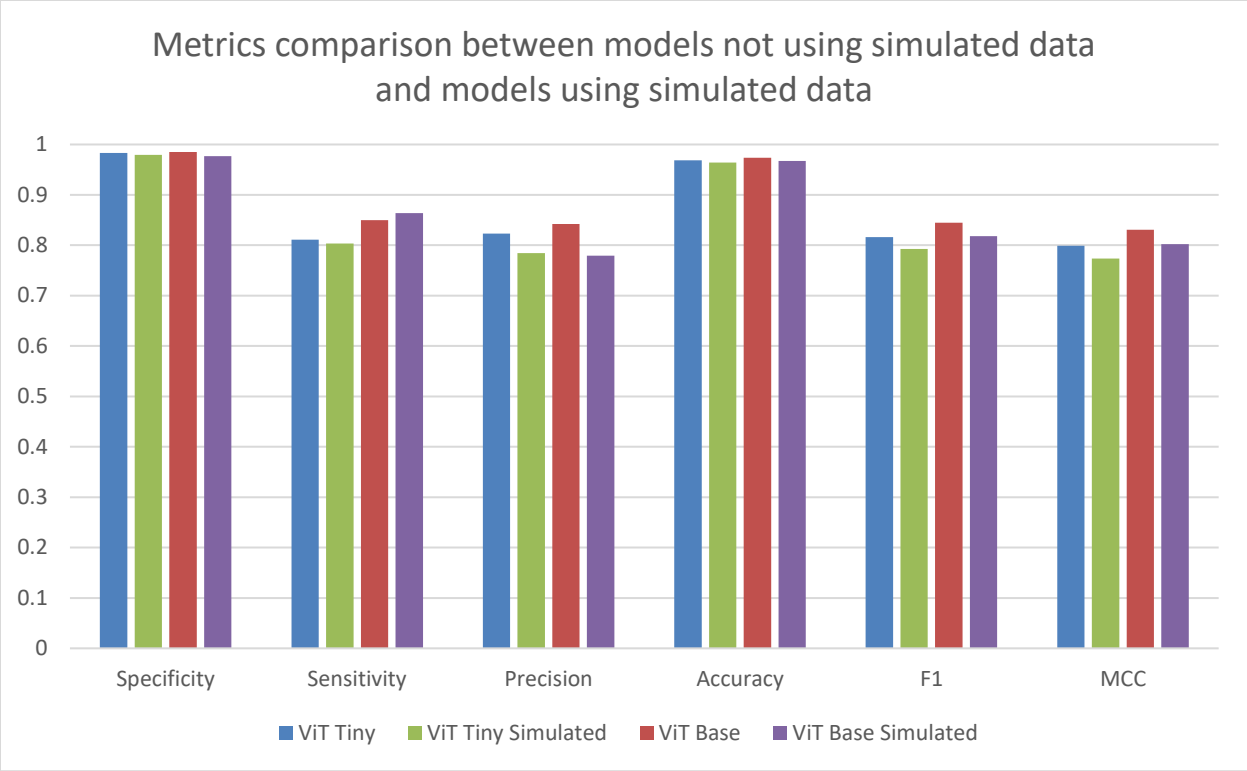


Figure 32: Specificity, sensitivity, precision, accuracy, F1-score, and MCC for models not using simulated data and models using simulated data.

6.4.3 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 9 and Figure 33). The ANOVA test found that the specificity of the models experienced a statistically significant increase due to the application of perspective transformation. All other changes in performance were not statistically significant and are therefore within the margin of error/chance. The secondary ANOVA did not find any statistically significant effect occurring from preprocessing the images using perspective transformation. The VGG-16 model showed improvements in over all metrics, except sensitivity. The difference in the trend of results between the ViT models and the VGG-16 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 9: Summary of results from 'tiny' and 'base' vision transformer models with perspective transformed data

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
ViT Tiny with Perspective Transformed Data	98.83%	76.27%	85.90%	96.92%	80.72%	79.26%
ViT Base with Perspective Transformed Data	98.72%	82.16%	85.66%	97.32%	83.81%	82.41%
VGG-16 with Perspective Transformed Data	93.21%	56.48%	44.05%	90.10%	49.17%	44.42%

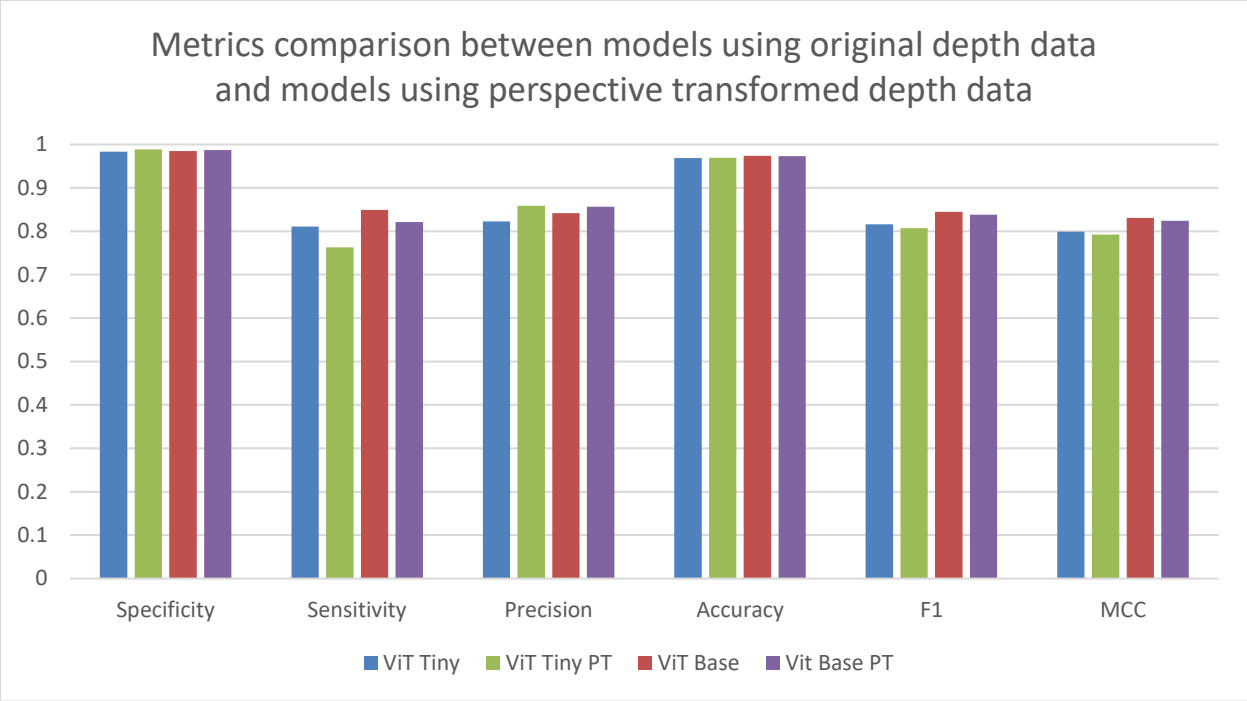


Figure 33: Specificity, sensitivity, precision, accuracy, F1-score, and MCC for models using original depth data and models using perspective transformed data.

6.4.4 Effect of HHA Encoding on Model Performance

As seen in Figure 34, when comparing the performance of the models using the HHA encoded depth data against that of the models using the original 1-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 10). The effect of HHA encoding the data used to train and evaluate the models was found to be statistically significant for all metrics when applying the ANOVA test. The secondary ANOVA test revealed a statistically significant effect on the precision and MCC of the models. The improvement in the model's performance was expected, as the model was pre-trained on 3-channel RGB images before transferring the weights. Although the VGG-16 model was pretrained 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 VGG-16 model using HHA encoded data showed improvements to specificity, accuracy, and

precision and a detrimental effect on sensitivity, F1-score, and MCC. This was unexpected, as HHA encoded depth images have been shown to increase the performance of CNNs pretrained in this way [32].

Table 10: Summary of results from 'tiny' and 'base' vision transformer models with HHA encoded data

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
ViT Tiny with HHA Encoded Data	98.68%	80.62%	82.96%	97.39%	81.60%	80.30%
ViT Base with HHA Encoded Data	98.64%	83.59%	85.08%	97.37%	84.25%	82.86%
VGG-16 with HHA Encoded Data	96.81%	18.95%	40.42%	90.22%	24.36%	22.14%

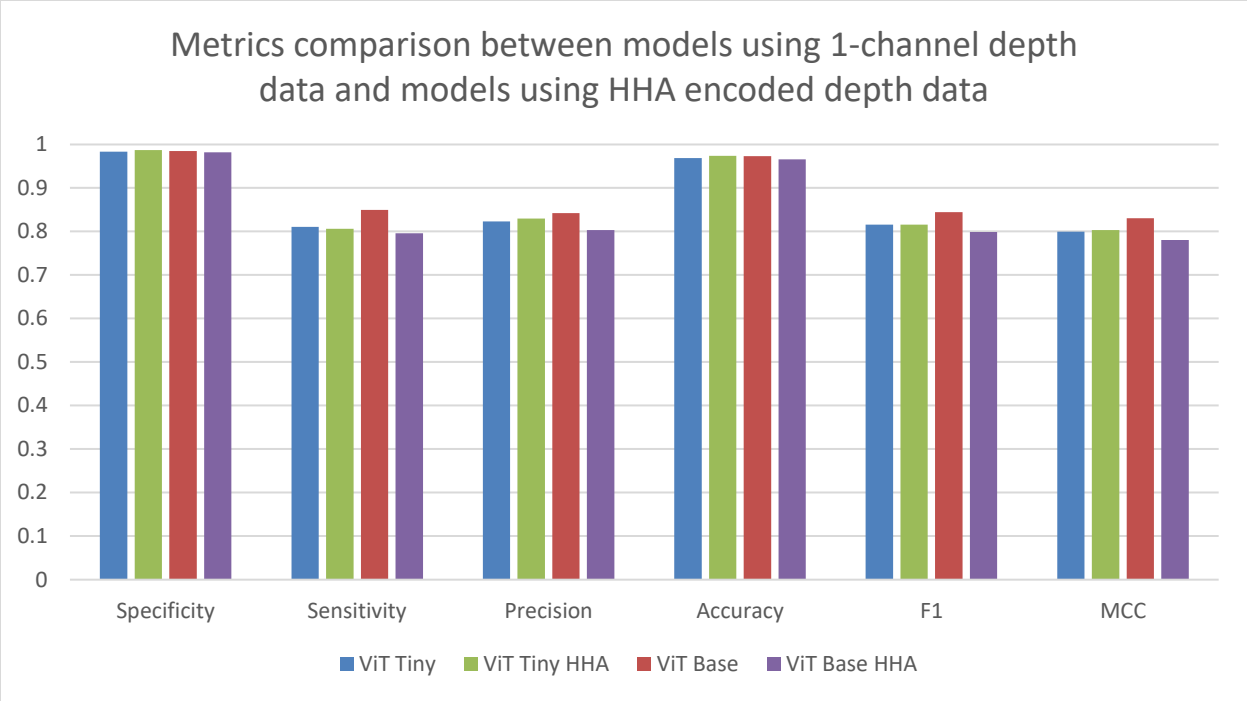


Figure 34: Specificity, sensitivity, precision, accuracy, F1-score, and MCC for models using 1-channel depth data and models using HHA encoded depth data.

6.4.5 Effects of Multiple Variables on Model Performance

The previous four sections outlined the four independent variables applied to the models separately (i.e., model size, simulated data, perspective transform, 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 (not including each variable separately). The results of the remaining models not shown previously can be found in Table 11. An n-way ANOVA was run, where n is the number of independent variables (4). One of the resulting ANOVA tables can be seen in Table 13. It can be seen that a combination of model size, encoding type, use of perspective transformation, and use of simulated data had a statistically significant effect on the specificity and precision of the models. The full ANOVA tables for all metrics can be found in Appendix B. A combination of type of encoding and perspective transformation application also had a statistically significant effect on all other metrics (sensitivity, accuracy, F1-score, MCC). Figure 35 to Figure 40 display the metrics for each of the models with and without a combination of variables. Unexpectedly, when running

the secondary ANOVA test, no combination of variables was found to have a statistically significant effect on the performance of the models (Appendix C). This may be due to certain variables that have a positive and negative effect counteracting each other when acting in conjunction. The effects of combinations of variables on the performance of the VGG-16 model was also investigated. The results of the remaining VGG-16 models not shown previously can be seen in Table 12. The best performing VGG-16 model utilizes the simulated data as well as HHA encoding. It shows an improvement over all metrics except sensitivity, where it has a detrimental effect.

Table 11: 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 with Perspective Transformed Simulated Data	97.08%	77.27%	71.91%	95.41%	74.18%	71.92%
ViT Base with Perspective Transformed Simulated Data	98.65%	81.79%	84.93%	97.22%	83.29%	81.82%
ViT Tiny with HHA Encoded Simulated Data	97.62%	81.52%	77.50%	96.26%	78.94%	77.24%
ViT Base with HHA Encoded Simulated Data	99.04%	78.86%	88.66%	97.34%	83.13%	82.06%
ViT Tiny with HHA Encoded Perspective Transformed Data	98.79%	84.16%	86.74%	97.55%	85.38%	84.09%
ViT Base with HHA Encoded Perspective Transformed Data	**99.10%**	**85.59%**	**89.76%**	**97.95%**	**87.62%**	**86.54%**
ViT Tiny with HHA Encoded Perspective Transformed Simulated Data	98.90%	82.41%	87.39%	97.50%	84.81%	83.51%
ViT Base with HHA Encoded Perspective Transformed Simulated Data	98.99%	85.35%	88.64%	97.84%	86.95%	85.80%

Table 12: Summary of results from VGG-16 models with combinations of studied variables

Model	Specificity	Sensitivity	Precision	Accuracy	F1-Score	MCC
VGG-16 with Perspective Transformed Simulated Data	98.83%	36.59%	74.76%	93.56%	48.88%	49.36%
VGG-16 with HHA Encoded Simulated Data	98.89%	45.78%	79.24%	94.39%	57.98%	57.63%
VGG-16 with HHA Encoded Perspective Transformed Data	98.97%	1.90%	20.00%	90.76%	3.32%	3.05%
VGG-16 with HHA Encoded Perspective Transformed Simulated Data	99.15%	33.35%	78.71%	93.50%	46.74%	48.57%

Table 13: N-Way ANOVA table for precision as a representative example. P-values < alpha (where alpha = 0.05) highlighted in green to indicate positive statistical significance and red to indicate negative statistical significance

Effect	DF	Sum of Squares	Mean Squares	F-Value	P-Value
Model Size	1	0.084583	0.084583	8.956442	0.002944
HHA Encoding	1	0.154052	0.154052	16.31235	6.49E-05
Perspective Transformation	1	0.016422	0.016422	1.738941	0.188059
Simulated Data	1	0.087886	0.087886	9.306104	0.002442
Model Size : HHA Encoding	1	3.11E-07	3.11E-07	3.30E-05	0.995421
Model Size : Perspective Transformation	1	0.001904	0.001904	0.201584	0.653698
Model Size : Simulated Data	1	0.034983	0.034983	3.704346	0.055009
HHA Encoding : Perspective Transformation	1	0.017125	0.017125	1.813363	0.1789
HHA Encoding : Simulated Data	1	0.071431	0.071431	7.563694	0.006237
Perspective Transformation : Simulated Data	1	0.002798	0.002798	0.296266	0.586549
Model Size : HHA Encoding :Perspective Transformation	1	0.028676	0.028676	3.036499	0.082212
Model Size : HHA Encoding : Simulated Data	1	0.001112	0.001112	0.11772	0.731708
Model Size : Perspective Transformation : Simulated Data	1	0.000461	0.000461	0.048809	0.825267
HHA Encoding : Perspective Transformation : Simulated Data	1	6.58E-06	6.58E-06	0.000697	0.978955
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	0.052747	0.052747	5.585333	0.018609
Residual	384	3.626446	0.009444		

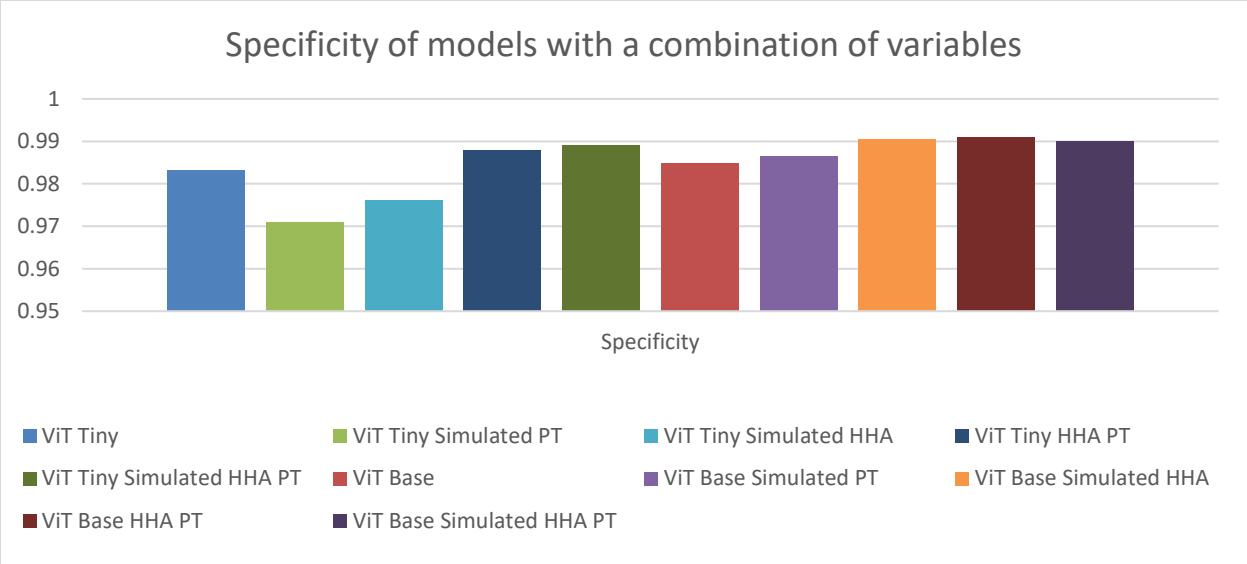


Figure 35: Specificity of models with a combination of variables.

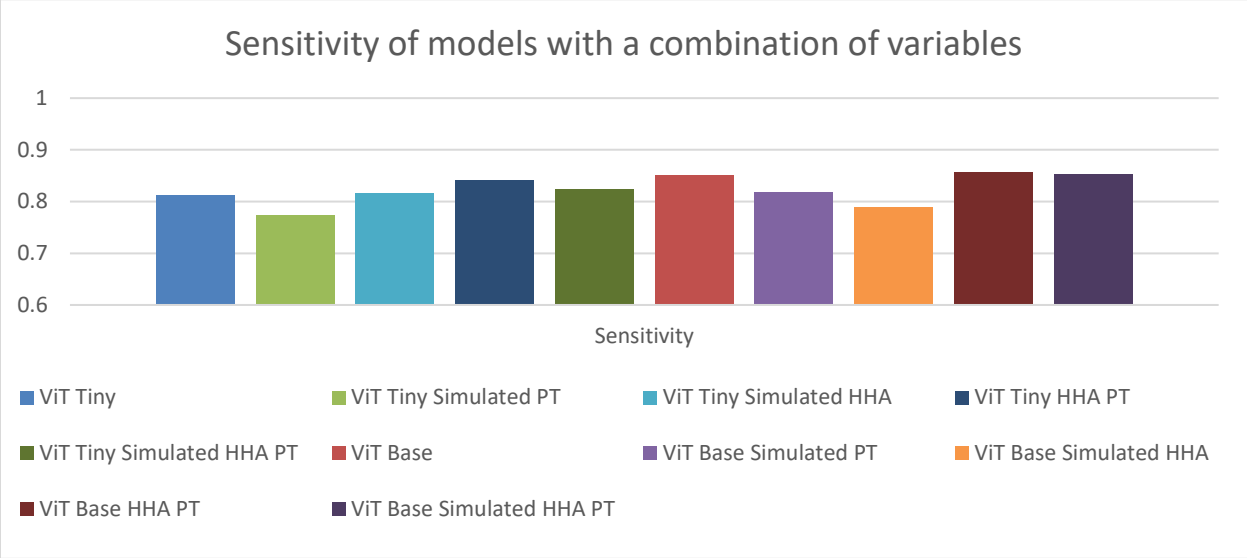


Figure 36: Sensitivity of models with a combination of variables.

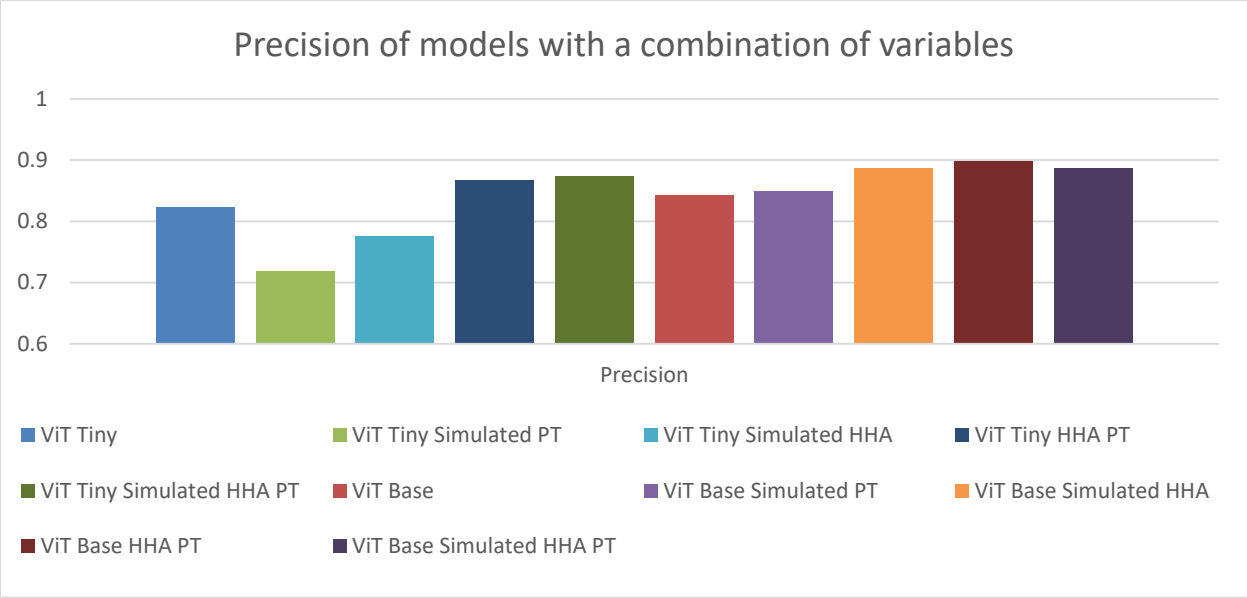


Figure 37: Precision of models with a combination of variables.

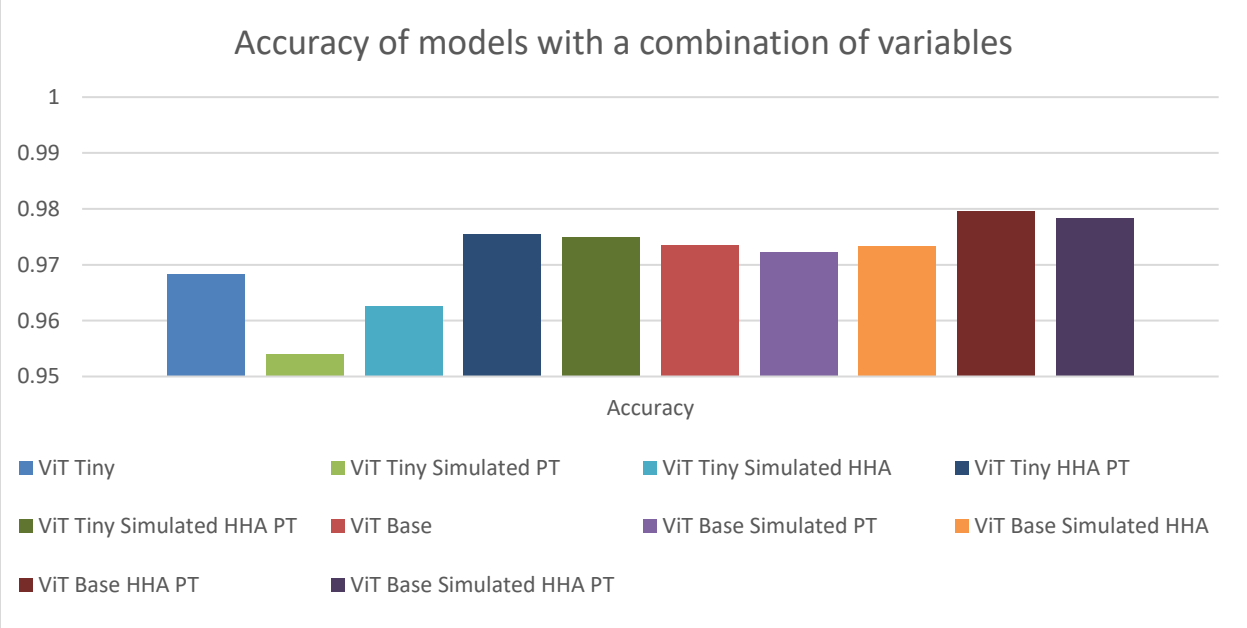


Figure 38: Accuracy of models with a combination of variables.

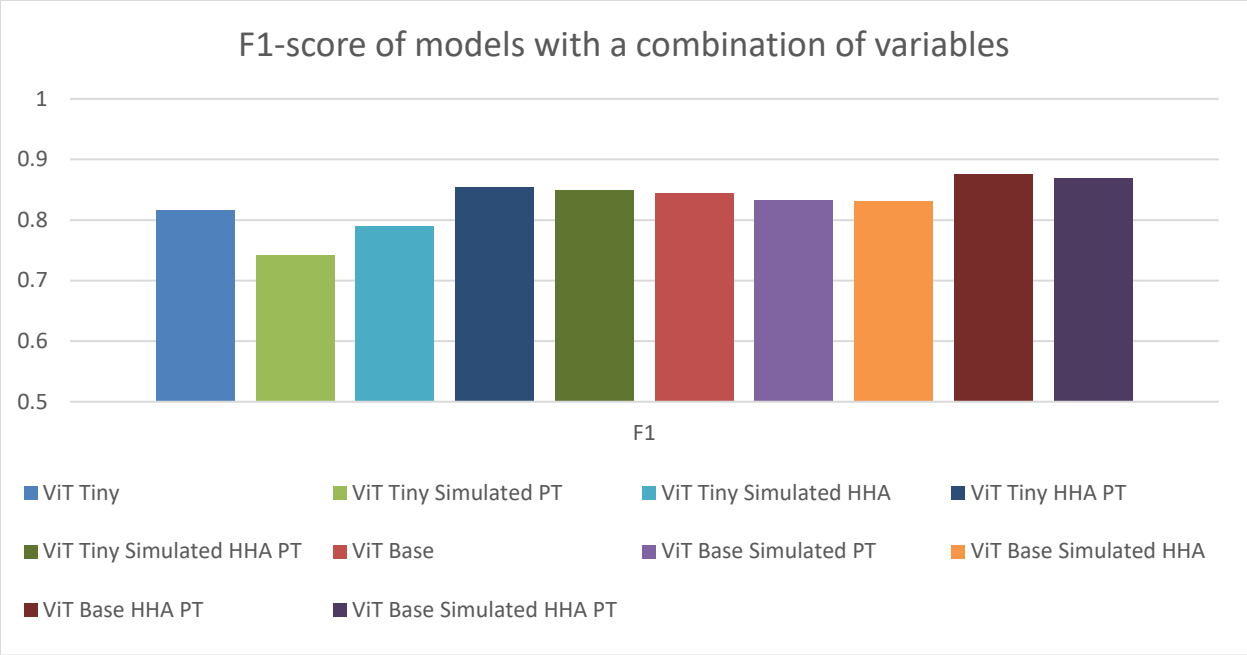


Figure 39: F1-score of models with a combination of variables.

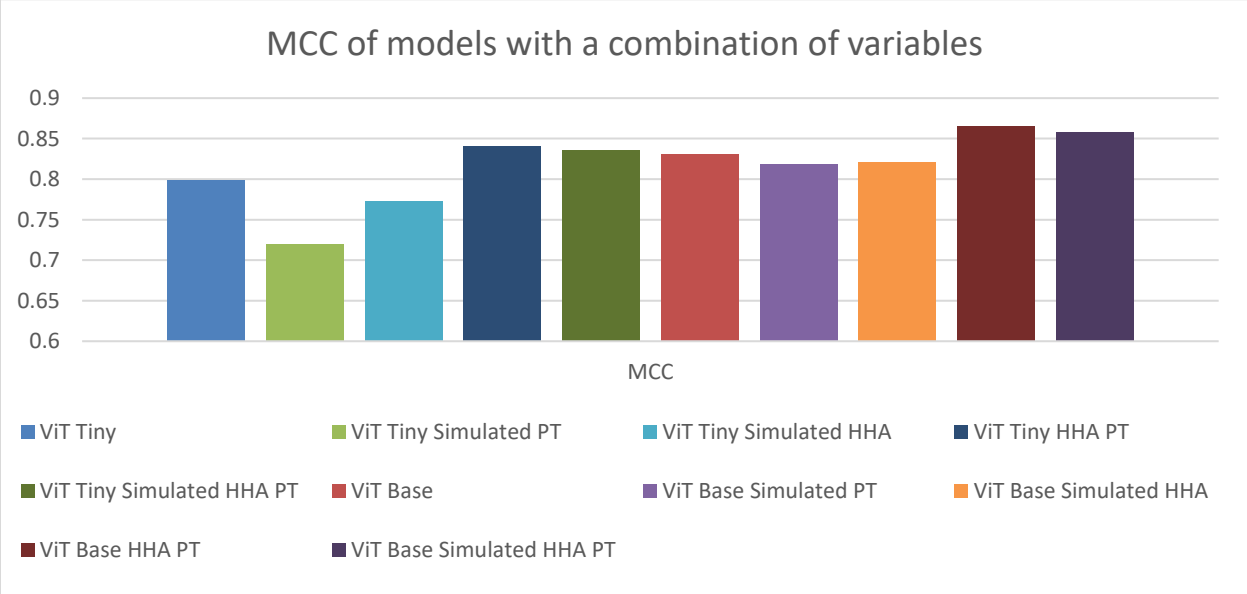


Figure 40: MCC of models with a combination of variables.

6.5 Conclusions

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, while a model 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 '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 '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 the depth data after applying the perspective transformation process. This model was overall the highest performing model. 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.

7 Thesis Summary and Future Recommendations

7.1 Summary

The objective this thesis set out to achieve was to develop and improve non-contact neonatal patient monitoring methods using the depth modality. ROI selection and respiratory rate estimation for neonates in the NICU has previously primarily focused on RGB-based methods. Depth methods that have been studied in the past have assumed the positioning of the camera to be constant or known, or otherwise relied on specialized equipment and setup. Periods of intervention during clinical studies of non-contact patient monitoring were typically manually determined and discarded. Methods laid out in this thesis account for non-ideal camera placement for ROI selection and RR estimation. A model was also trained to accurately detect periods of intervention during recording.

Data were collected from patients in the NICU to develop non-contact patient monitoring methods and technologies. The data comprised of RGB-D recordings of patients during their time in the NICU, respiratory rate signals of the patients taken from the hospital's patient monitors, and annotations of events and interventions during the recordings taken on a bedside annotation application.

A method was developed for transforming the depth data taken from cameras placed at non-ideal angles and positions. The method was test on 28 patients with estimated camera angles of 5.46-38.58 degrees away from the optimal position looking straight down at the bed. The mean absolute percentage error was found to be 5.58% over all patients, ranging from 0.40% to 18.34%. The results show a capability for correcting the plane of the bed to be at a uniform distance away from the camera.

An automated depth-based ROI selection method was developed by building on the perspective transformation method. ROI selection was automated after perspective transformation by taking cross-sections of the depth image and building candidate regions using contour-finding algorithm. The method was evaluated on frames from 4 patients with varying poses, camera angles, and levels of blanket coverage, resulting in an average Sørensen–Dice coefficient of 0.62 and Jaccard index of 0.46.

Perspective transformation and the automated ROI selection method were then evaluated by building a pipeline for respiratory rate estimation. The evaluation investigated the use of the pipeline in conjunction with a time- and frequency-domain respiratory rate estimation method. Use of the pipeline improved the percentage of acceptable estimates overall (6.12% to 8.97% in the time domain and 3.60% to 13.47% in the frequency domain), though the pipeline had no effect on the results for two patients when using the time-domain RR estimation method and one patient when using the frequency-domain RR estimation method.

A deep learning model was trained to detect moments of clinical intervention in recorded scenes from the NICU. A VIT was chosen for this task, and the effects of adding simulated data, applying perspective transformation, and HHA encoding were investigated. The larger (Base) vision transformer model using perspective transformed HHA encoded data was found to outperform the baseline VGG-16 based models overall, with 85.6% sensitivity, 89.8% precision, and 87.6% F1-score. Surprisingly, the use of the simulated data was found to have a slightly detrimental effect on the performance of the models, though this was not found to be statistically significant.

7.2 Conclusions

A perspective transformation method was built and tested, achieving a consistent correction such that the viewpoint of the camera appears to be directly above the patient's bed facing

downwards . The transformation method appears to be robust to varying angles of the camera. Through testing the ROI selection method, the impact of the varying blanket coverage, camera angles, and patient pose was found to be minimal. The exact requirements for the mean absolute error of the perspective transform are not known, so an improvement in downstream tasks, such as respiration rate, was computed as a reflection of the effectiveness of the transformation. The PAE of the respiratory rate estimation methods was found to improve after the application of the pipeline, though the impact of the ROI selection and perspective transformation methods separately are not known. Comparing the performance of the respiratory rate estimation methods when utilizing the pipeline against the same methods using a gold standard ROI over the recording may better represent the impact of the pipeline.

The models built for intervention detection were found to outperform the baseline state-of-the-art models. Of the tested variables that were expected to affect the performance of the models, only model size and the use of HHA encoding had a significantly positive impact. The best performing model was found to be the larger 'base' vision transformer using HHA transformed data after applying the perspective transformation.

7.3 Recommendations for future work

The work in this thesis has provided a pre-processing pipeline that can be used in conjunction with various RR estimation methods. The thesis has also presented the use of vision transformers on depth data for classification of periods of intervention in the NICU. The following subsections propose some recommendation for future work.

7.3.1 Improving Region-Of-Interest Selection

The ROI selection method was tested on our limited dataset. Though it is expected to generalize for a wide range of ROI sizes and scenes, further testing on a larger dataset is

needed to confirm this. To further study the performance of the method, a direct comparison to an RGB based method can be used on a single set of data. Increasing the robustness of the perspective transformation process may also improve the performance of the ROI selection method. By allowing the user to select a region in the scene rather than single points for calculating the rotation matrix, we can make the method more resistant to wrinkles and other factors that can make the surface of the bed seem uneven. Additionally, implementing some post-processing steps after the initial ROI contours are found may further improve the torso ROI segmentation process. For example, since the general shape of the torso ROI is approximately known, steps could be taken to fit this shape onto the found contour. This might include modeling the torso ROI as an ellipse at the centroid of the chosen contour and aligning it's major and minor axes. Finally, the positions and orientations of the candidate regions-of-interest may be screened to differentiate and correct for false positives where the method may detect a doll or other item in the bed rather than the patient.

7.3.2 Investigating Alternative Respiratory Rate Estimation Methods

Although the perspective transformation and ROI selection pipeline was shown to improve the performance of the respiratory rate estimation methods, the total percentage of acceptable estimates was still underperforming the state of the art. It is suspected that applying the pipeline as a preprocessing step before estimating RR using a method built specifically for use with depth video will show similar improvements and a higher PAE. Another contribution could be applying the RR estimation methods to the gold-standard ROI over the same time-periods in the recording. This would verify the proposition that finding the ideal ROI automatically is a useful step toward automatic non-contact RR estimation. Further, disregarding more low frequency artifacts in the scene may be achieved through more robust filtering methods during signal extraction from the video segments. Finally, one should consider a Bland-Altman analysis to evaluate the difference in RR estimation measurements between the depth-based methods and the 'gold-standard' of the patient monitor. Although the patient monitor is

designated as the 'gold-standard' in this study, the agreement of the methods must be studied, since the output of patient monitor necessarily the absolute truth.

7.3.3 Studying the Effect of Other Variables on Intervention Detection

This thesis studied the effect of multiple variables (separately and in conjunction) on the performance of intervention detection models. The models trained after the addition of simulated data into the training sets seemed to perform worse overall than their counterparts without the simulated data. This was unexpected, as it was thought that correcting for the class imbalance might improve some metrics. Future study can consist of the inclusion of more simulated data for the underrepresented class, studying the effects of a completely balanced training dataset. Another avenue of research may be to train and test vision transformers with varying numbers and sizes of patches. Perspective transformation was shown to increase the performance of the CNN architecture but not the ViT. By splitting the input images into smaller patches, the enhanced features of the transformed images may have a greater effect on the performance of the model. Other variables that can be explored are using multiple successive frames to predict the probability of intervention using a known image as a negative baseline or studying the use of larger models (utilizing more trainable parameters).

The use of depth video for intervention detection may also be explored, since looking at a series of frames using a 3D convolution (with a dimension being the changes in the scene over time) may affect the performance of the model. Since the movements of the clinician or other practitioner in the scene will likely be obvious and more exaggerated than those of the patient, the temporal information might be useful for this application.

7.3.4 Classification of Periods of Intervention

A model was built to binarily classify whether a frame occurred during a period of intervention or otherwise. Extending this classification to include the type of intervention being undertaken is a logical next step. This would require manually categorizing the frames taken from the NICU recordings.

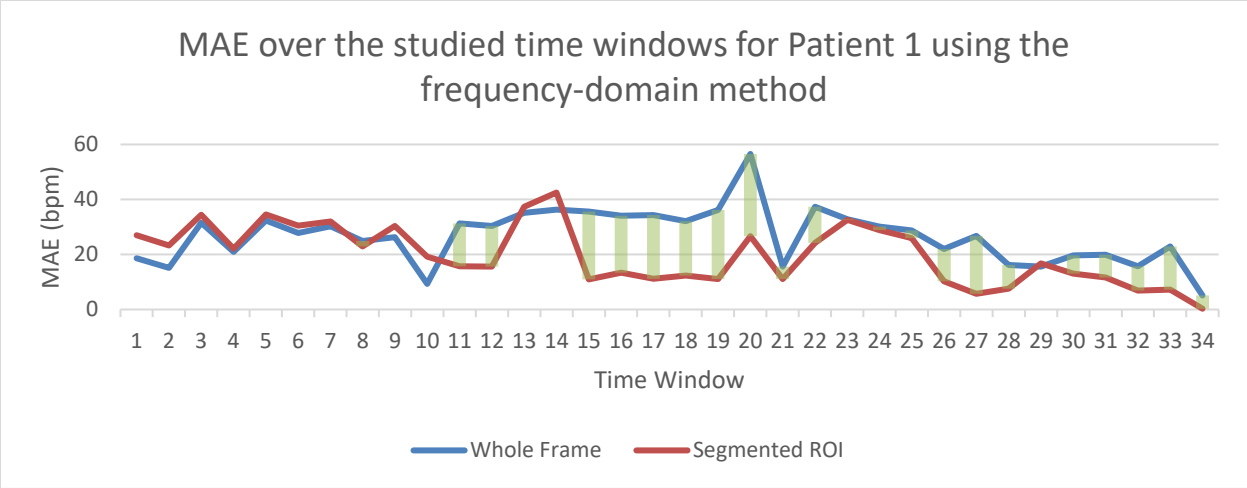
7.3.5 Semantic Segmentation of Intervention Frames

As shown in this thesis, vision transformers have a useful application for non-contact patient monitoring. Though the periods of intervention have been classified, further development can lead to semantic segmentation of frames of intervention, labeling the practitioners hands or other equipment being used in the scene. Extending the use of vision transformers for semantic segmentation has been explored previously [24]–[26], and it is expected to accomplish the task in this case as well. The 'Intervention' class frames can be labelled pixelwise, either manually or by applying a color-based semantic segmentation model on the simulated data collected as described in Chapter 3.

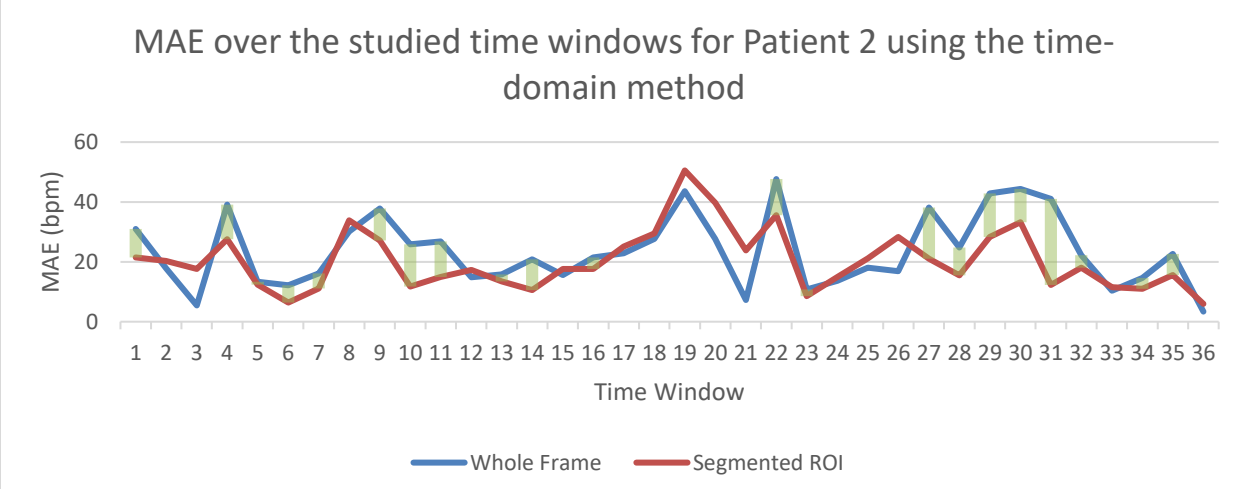
Appendix A: Additional Plots of MAE for Chapter 5

Appendix A includes the remaining plots of the mean absolute error for the frequency- and time-domain methods for each of the patients tested in Chapter 5.

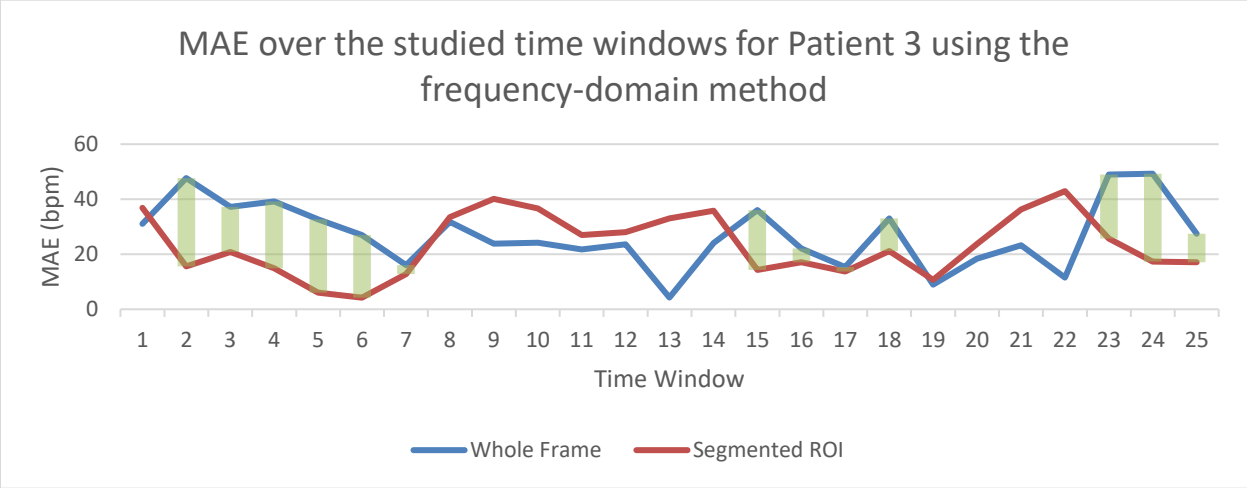
A.1 The MAE of Patient 1 RR Estimated Using the Frequency-Domain Method



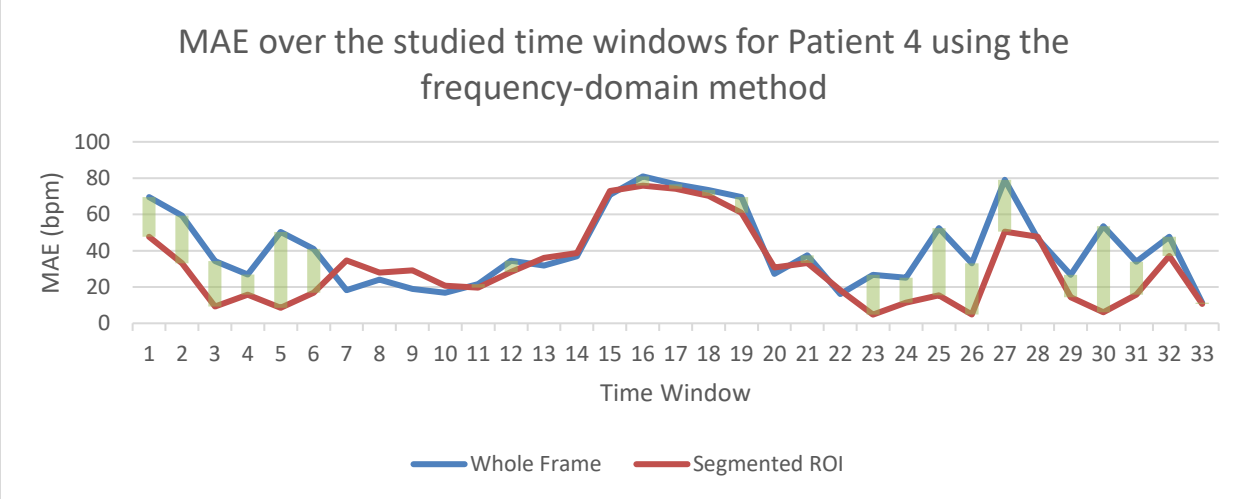
A.2 The MAE of Patient 2 RR Estimated Using the Time-Domain Method



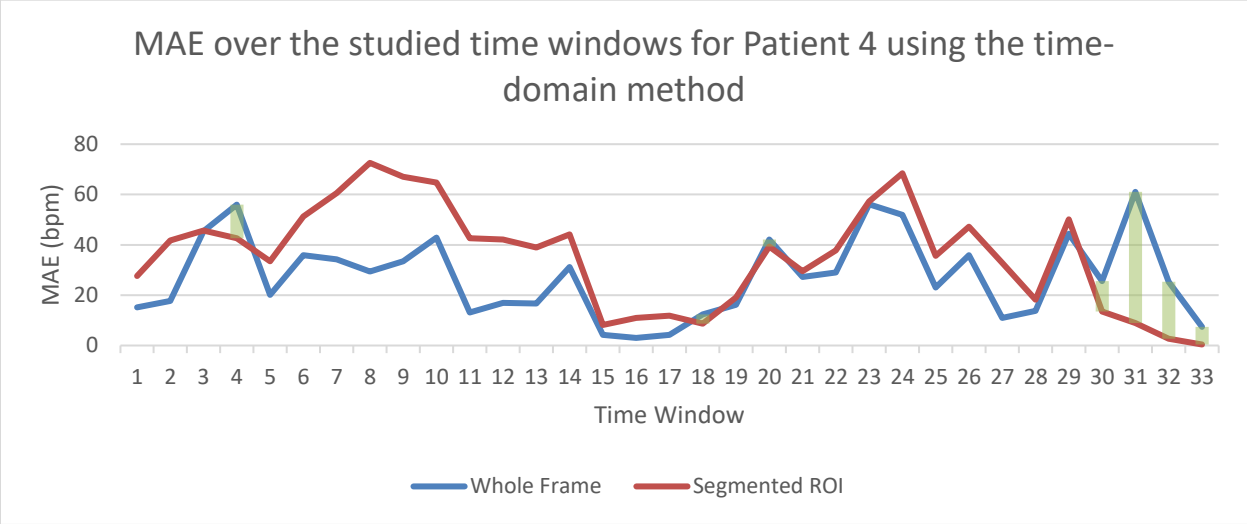
A.3 The MAE of Patient 3 RR Estimated Using the Frequency-Domain Method



A.4 The MAE of Patient 4 RR Estimated Using the Frequency-Domain Method



A.5 The MAE of Patient 4 RR Estimated Using the Time-Domain Method



Appendix B: Additional N-Way ANOVA Tables for Chapter 6

Appendix B includes the remaining n-way ANOVA tables that the results in Chapter 5.4 were interpreted from. This consists of ANOVA tables calculating the statistical significance of each of the studied variables on sensitivity, specificity, accuracy, f1-score, and MCC.

B.1 N-Way ANOVA Table for Specificity

Effect	DF	Sum of Squares	Mean Squares	F-Value	P-Value
Model Size	1	0.001828	0.001828	4.916969	0.027179
HHA Encoding	1	0.002199	0.002199	5.915685	0.015463
Perspective Transformation	1	0.001578	0.001578	4.244858	0.040043
Simulated Data	1	0.001618	0.001618	4.352545	0.037613
Model Size : HHA Encoding	1	0.000104	0.000104	0.28099	0.59636
Model Size : Perspective Transformation	1	6.50E-06	6.50E-06	0.017479	0.894889
Model Size : Simulated Data	1	0.000754	0.000754	2.027339	0.155303
HHA Encoding : Perspective Transformation	1	0.000209	0.000209	0.562767	0.453607
HHA Encoding : Simulated Data	1	0.000985	0.000985	2.648765	0.104451
Perspective Transformation : Simulated Data	1	9.43E-06	9.43E-06	0.025372	0.873526
Model Size : HHA Encoding :Perspective Transformation	1	0.001193	0.001193	3.210188	0.073968
Model Size : HHA Encoding : Simulated Data	1	5.21E-06	5.21E-06	0.01402	0.905807
Model Size : Perspective Transformation : Simulated Data	1	7.64E-05	7.64E-05	0.205529	0.650551
HHA Encoding : Perspective Transformation : Simulated Data	1	0.000163	0.000163	0.438767	0.508116
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	0.002045	0.002045	5.49965	0.019529
Residual	384	0.142763	0.000372		

B.2 N-Way ANOVA Table for Sensitivity

Effect	DF	Sum of Squares	Mean Squares	F-Value	P-Value
Model Size	1	0.089944	0.089944	9.105086	0.002719
HHA Encoding	1	0.039752	0.039752	4.024137	0.045554
Perspective Transformation	1	0.005826	0.005826	0.589768	0.442981
Simulated Data	1	0.005741	0.005741	0.581192	0.446314
Model Size : HHA Encoding	1	0.036692	0.036692	3.714334	0.054685
Model Size : Perspective Transformation	1	0.005525	0.005525	0.559316	0.454994
Model Size : Simulated Data	1	0.001522	0.001522	0.154053	0.69491
HHA Encoding : Perspective Transformation	1	0.097202	0.097202	9.839878	0.001839
HHA Encoding : Simulated Data	1	0.008806	0.008806	0.891411	0.345689
Perspective Transformation : Simulated Data	1	0.001767	0.001767	0.178884	0.672571
Model Size : HHA Encoding :Perspective Transformation	1	0.003746	0.003746	0.379226	0.538383
Model Size : HHA Encoding : Simulated Data	1	0.004549	0.004549	0.460452	0.497822
Model Size : Perspective Transformation : Simulated Data	1	0.001645	0.001645	0.166525	0.683446
HHA Encoding : Perspective Transformation : Simulated Data	1	0.000971	0.000971	0.098253	0.754106
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	0.016374	0.016374	1.657582	0.198706
Residual	384	3.793312	0.009878		

B.3 N-Way ANOVA Table for Accuracy

Effect	DF	Sum of Squares	Mean Squares	F-Value	P-Value
Model Size	1	0.004243	0.004243	11.0491	0.000973
HHA Encoding	1	0.003027	0.003027	7.882593	0.005246
Perspective Transformation	1	0.001192	0.001192	3.103431	0.078923
Simulated Data	1	0.001778	0.001778	4.629684	0.032046
Model Size : HHA Encoding	1	9.61E-05	9.61E-05	0.250371	0.617101
Model Size : Perspective Transformation	1	5.89E-05	5.89E-05	0.153471	0.695458
Model Size : Simulated Data	1	0.000463	0.000463	1.205815	0.272851
HHA Encoding : Perspective Transformation	1	0.001875	0.001875	4.882085	0.027726
HHA Encoding : Simulated Data	1	0.000407	0.000407	1.059505	0.303976
Perspective Transformation : Simulated Data	1	2.22E-07	2.22E-07	0.000578	0.980829
Model Size : HHA Encoding :Perspective Transformation	1	0.00066	0.00066	1.718465	0.190674
Model Size : HHA Encoding : Simulated Data	1	4.55E-05	4.55E-05	0.118601	0.730745
Model Size : Perspective Transformation : Simulated Data	1	0.000135	0.000135	0.352415	0.553099
HHA Encoding : Perspective Transformation : Simulated Data	1	0.000215	0.000215	0.561041	0.4543
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	0.00089	0.00089	2.317628	0.128738
Residual	384	0.147445	0.000384		

B.4 N-Way ANOVA Table for F1-Score

Effect	DF	Sum of Squares	Mean Squares	F-Value	P-Value
Model Size	1	0.087045	0.087045	11.94513	0.000609
HHA Encoding	1	0.08441	0.08441	11.5835	0.000735
Perspective Transformation	1	0.006058	0.006058	0.831386	0.362444
Simulated Data	1	0.056296	0.056296	7.72548	0.005712
Model Size : HHA Encoding	1	0.015659	0.015659	2.148856	0.143495
Model Size : Perspective Transformation	1	0.008374	0.008374	1.149103	0.28441
Model Size : Simulated Data	1	0.003593	0.003593	0.492998	0.483018
HHA Encoding : Perspective Transformation	1	0.071341	0.071341	9.790126	0.001889
HHA Encoding : Simulated Data	1	0.003305	0.003305	0.453516	0.501075
Perspective Transformation : Simulated Data	1	0.001305	0.001305	0.179063	0.672417
Model Size : HHA Encoding :Perspective Transformation	1	0.002358	0.002358	0.323542	0.569819
Model Size : HHA Encoding : Simulated Data	1	0.005871	0.005871	0.805708	0.369955
Model Size : Perspective Transformation : Simulated Data	1	0.003411	0.003411	0.468059	0.494294
HHA Encoding : Perspective Transformation : Simulated Data	1	0.003674	0.003674	0.504198	0.478092
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	0.002267	0.002267	0.311056	0.577358
Residual	384	2.798231	0.007287		

B.5 N-Way ANOVA Table for MCC

Effect	DF	Sum of Squares	Mean Squares	F-Value	P-Value
Model Size	1	0.106892	0.106892	14.82945	0.000138
HHA Encoding	1	0.102389	0.102389	14.2047	0.00019
Perspective Transformation	1	0.004688	0.004688	0.65032	0.420497
Simulated Data	1	0.053505	0.053505	7.422893	0.006734
Model Size : HHA Encoding	1	0.013652	0.013652	1.893964	0.169558
Model Size : Perspective Transformation	1	0.00625	0.00625	0.86714	0.352333
Model Size : Simulated Data	1	0.005773	0.005773	0.800958	0.371368
HHA Encoding : Perspective Transformation	1	0.071005	0.071005	9.850791	0.001829
HHA Encoding : Simulated Data	1	0.006736	0.006736	0.934466	0.334313
Perspective Transformation : Simulated Data	1	0.000187	0.000187	0.025954	0.872098
Model Size : HHA Encoding :Perspective Transformation	1	0.003722	0.003722	0.516323	0.472852
Model Size : HHA Encoding : Simulated Data	1	0.004726	0.004726	0.655617	0.418613
Model Size : Perspective Transformation : Simulated Data	1	0.00269	0.00269	0.373151	0.541653
HHA Encoding : Perspective Transformation : Simulated Data	1	0.002069	0.002069	0.287049	0.592428
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	0.003746	0.003746	0.519692	0.471412
Residual	384	2.767904	0.007208		

Appendix C: N-Way ANOVA Tables after Collapsing Repetitions in Chapter 6

Appendix C consists of the secondary n-way ANOVA tables that the results in Chapter 5.4 after collapsing the repetitions were interpreted from. This includes ANOVA tables calculating the statistical significance of each of the studied variables on specificity, sensitivity, precision, accuracy, f1-score, and MCC.

C.1 N-Way ANOVA Table for Specificity after Collapsing Repetitions

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
Perspective Transformation	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 : Perspective Transformation	1	0.012997	0.012997	0.006625	0.935383
Model Size : Simulated Data	1	1.507441	1.507441	0.768399	0.383991
HHA Encoding : Perspective Transformation	1	0.418449	0.418449	0.213299	0.64576
HHA Encoding : Simulated Data	1	1.969506	1.969506	1.003931	0.320134
Perspective Transformation : Simulated Data	1	0.018866	0.018866	0.009617	0.922188
Model Size : HHA Encoding :Perspective Transformation	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 : Perspective Transformation : Simulated Data	1	0.152822	0.152822	0.077899	0.781065
HHA Encoding : Perspective Transformation : Simulated Data	1	0.326248	0.326248	0.166301	0.684782
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	4.0893	4.0893	2.08447	0.153681
Residual	64	125.5548	1.961794		

C.2 N-Way ANOVA Table for Sensitivity after Collapsing Repetitions

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
Perspective Transformation	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 : Perspective Transformation	1	11.05032	11.05032	0.206125	0.651358
Model Size : Simulated Data	1	3.043599	3.043599	0.056773	0.812432
HHA Encoding : Perspective Transformation	1	194.4048	194.4048	3.626291	0.06137
HHA Encoding : Simulated Data	1	17.61146	17.61146	0.328512	0.568545
Perspective Transformation : Simulated Data	1	3.53418	3.53418	0.065924	0.79819
Model Size : HHA Encoding :Perspective Transformation	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 : Perspective Transformation : Simulated Data	1	3.290016	3.290016	0.06137	0.805137
HHA Encoding : Perspective Transformation : Simulated Data	1	1.941173	1.941173	0.036209	0.849686
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	32.74856	32.74856	0.610869	0.437343
Residual	64	3431.029	53.60982		

C.3 N-Way ANOVA Table for Precision after Collapsing Repetitions

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
Perspective Transformation	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 : Perspective Transformation	1	3.807472	3.807472	0.075388	0.784532
Model Size : Simulated Data	1	69.96673	69.96673	1.385336	0.243553
HHA Encoding : Perspective Transformation	1	34.25033	34.25033	0.678154	0.413281
HHA Encoding : Simulated Data	1	142.8611	142.8611	2.828639	0.097471
Perspective Transformation : Simulated Data	1	5.595797	5.595797	0.110796	0.740327
Model Size : HHA Encoding :Perspective Transformation	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 : Perspective Transformation : Simulated Data	1	0.921884	0.921884	0.018253	0.892953
HHA Encoding : Perspective Transformation : Simulated Data	1	0.01316	0.01316	0.000261	0.987171
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	105.4943	105.4943	2.08878	0.153263
Residual	64	3232.336	50.50525		

C.4 N-Way ANOVA Table for Accuracy after Collapsing Repetitions

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
Perspective Transformation	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 : Perspective Transformation	1	0.117856	0.117856	0.042577	0.837179
Model Size : Simulated Data	1	0.925994	0.925994	0.334524	0.565037
HHA Encoding : Perspective Transformation	1	3.749151	3.749151	1.354416	0.248828
HHA Encoding : Simulated Data	1	0.813637	0.813637	0.293934	0.589593
Perspective Transformation : Simulated Data	1	0.000444	0.000444	0.00016	0.989934
Model Size : HHA Encoding :Perspective Transformation	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 : Perspective Transformation : Simulated Data	1	0.270634	0.270634	0.097769	0.75554
HHA Encoding : Perspective Transformation : Simulated Data	1	0.430846	0.430846	0.155647	0.694508
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	1.7798	1.7798	0.642969	0.425605
Residual	64	177.158	2.768094		

C.5 N-Way ANOVA Table for F1-Score after Collapsing Repetitions

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
Perspective Transformation	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 : Perspective Transformation	1	16.74717	16.74717	0.35398	0.553966
Model Size : Simulated Data	1	7.185012	7.185012	0.151868	0.698051
HHA Encoding : Perspective Transformation	1	142.6825	142.6825	3.015839	0.087266
HHA Encoding : Simulated Data	1	6.6096	6.6096	0.139705	0.70981
Perspective Transformation : Simulated Data	1	2.609685	2.609685	0.05516	0.815066
Model Size : HHA Encoding :Perspective Transformation	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 : Perspective Transformation : Simulated Data	1	6.821549	6.821549	0.144185	0.705412
HHA Encoding : Perspective Transformation : Simulated Data	1	7.348248	7.348248	0.155318	0.694814
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	4.533364	4.533364	0.09582	0.75791
Residual	64	3027.906	47.31104		

C.6 N-Way ANOVA Table for MCC after Collapsing Repetitions

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
Perspective Transformation	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 : Perspective Transformation	1	12.50084	12.50084	0.245439	0.622002
Model Size : Simulated Data	1	11.54674	11.54674	0.226706	0.635599
HHA Encoding : Perspective Transformation	1	142.0106	142.0106	2.788204	0.099845
HHA Encoding : Simulated Data	1	13.47142	13.47142	0.264495	0.60882
Perspective Transformation : Simulated Data	1	0.374156	0.374156	0.007346	0.931965
Model Size : HHA Encoding :Perspective Transformation	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 : Perspective Transformation : Simulated Data	1	5.379411	5.379411	0.105618	0.746249
HHA Encoding : Perspective Transformation : Simulated Data	1	4.13815	4.13815	0.081247	0.776534
Model Size : HHA Encoding : Perspective Transformation : Simulated Data	1	7.49196	7.49196	0.147095	0.702598
Residual	64	3259.69	50.93265		

References

[1] Y. S. Dosso, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, "Video-Based Neonatal Motion Detection," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2020, pp. 6135–6138. doi: 10.1109/EMBC44109.2020.9175354.

[2] Y. S. Dosso, K. Greenwood, J. Harrold, and J. R. Green, "Bottle-Feeding Intervention Detection in the NICU," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Nov. 2021, pp. 1814–1819. doi: 10.1109/EMBC46164.2021.9631105.

[3] N. Koolen *et al.*, "Automated Respiration Detection from Neonatal Video Data:," in *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, Lisbon, Portugal, 2015, pp. 164–169. doi: 10.5220/0005187901640169.

[4] R. Janssen, W. Wang, A. Moço, and G. de Haan, "Video-based respiration monitoring with automatic region of interest detection," *Physiol. Meas.*, vol. 37, no. 1, pp. 100–114, Dec. 2015, doi: 10.1088/0967-3334/37/1/100.

[5] J. D. Kim *et al.*, "Non-contact respiration monitoring using impulse radio ultrawideband radar in neonates," *R. Soc. Open Sci.*, vol. 6, no. 6, p. 190149, Jun. 2019, doi: 10.1098/rsos.190149.

[6] W. H. Lee *et al.*, "Feasibility of non-contact cardiorespiratory monitoring using impulse-radio ultra-wideband radar in the neonatal intensive care unit," *PLOS ONE*, vol. 15, no. 12, p. e0243939, Dec. 2020, doi: 10.1371/journal.pone.0243939.

[7] A. Bekele *et al.*, "Real-time Neonatal Respiratory Rate Estimation using a Pressure-Sensitive Mat," in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Jun. 2018, pp. 1–5. doi: 10.1109/MeMeA.2018.8438682.

[8] Y. S. Dosso, A. Bekele, and J. R. Green, "Eulerian Magnification of Multi-Modal RGB-D Video for Heart Rate Estimation," in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Jun. 2018, pp. 1–6. doi: 10.1109/MeMeA.2018.8438741.

[9] Z. Hajj-Ali, K. Greenwood, J. Harrold, and J. R. Green, "Towards Depth-based Respiratory Rate Estimation with Arbitrary Camera Placement," in *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Jun. 2022, pp. 1–6. doi: 10.1109/MeMeA54994.2022.9856449.

[10] I. RealSense, "Beginner's guide to depth (Updated)," *Intel® RealSense™ Depth and Tracking Cameras*, Jul. 16, 2019. https://www.intelrealsense.com/beginners-guide-to-depth/ (accessed Nov. 02, 2022).

[11] G. Wolberg, "Geometric Transformation Techniques for Digital Images: A Survey," 1988, doi: 10.7916/D8TH8VRW.

[12] G. Dougherty, *Digital Image Processing for Medical Applications*. Cambridge University Press, 2009.

[13] D. G. Kyrollos, R. Hassan, Y. S. Dosso, and J. R. Green, "Fusing Pressure-Sensitive Mat Data with Video through Multi-Modal Registration," in *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, May 2021, pp. 1–6. doi: 10.1109/I2MTC50364.2021.9459886.

[14] Intel Corporation, "IntelRealSense/librealsense: Intel® RealSense™ SDK," 2018. https://github.com/IntelRealSense/librealsense (accessed Nov. 08, 2022).

[15] A. Burkov, *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019.

[16] IBM Cloud Education, "What are Convolutional Neural Networks?," *IBM Cloud Learn Hub*, Oct. 20, 2020. https://www.ibm.com/cloud/learn/convolutional-neural-networks (accessed Nov. 06, 2022).

[17] D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed Pooling for Convolutional Neural Networks," in *Rough Sets and Knowledge Technology*, Cham, 2014, pp. 364–375. doi: 10.1007/978-3-319-11740-9_34.
[18] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Dec. 05, 2017. doi: 10.48550/arXiv.1706.03762.
[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate." arXiv, May 19, 2016. doi: 10.48550/arXiv.1409.0473.
[20] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," p. 12.
[21] T. B. Brown *et al.*, "Language Models are Few-Shot Learners." arXiv, Jul. 22, 2020. doi: 10.48550/arXiv.2005.14165.
[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.
[23] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021. doi: 10.48550/arXiv.2010.11929.
[24] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7262–7272. Accessed: Nov. 14, 2022. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Strudel_Segmenter_Transformer_for_Semantic_Segmentation_ICCV_2021_paper.html
[25] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 12077–12090. Accessed: Nov. 14, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html
[26] S. Zheng *et al.*, "Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890. Accessed: Nov. 14, 2022. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Zheng_Rethinking_Semantic_Segmentation_From_a_Sequence-to-Sequence_Perspective_With_Transformers_CVPR_2021_paper.html
[27] Y. Fang *et al.*, "You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection," in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 26183–26197. Accessed: Nov. 14, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/dc912a253d1e9ba40e2c597ed2376640-Abstract.html
[28] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video Transformer Network," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3163–3172. Accessed: Nov. 14, 2022. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021W/CVEU/html/Neimark_Video_Transformer_Network_ICCVW_2021_paper.html
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[30] M. Hussain, J. J. Bird, and D. R. Faria, "A Study on CNN Transfer Learning for Image Classification," in *Advances in Computational Intelligence Systems*, Cham, 2019, pp. 191–202. doi: 10.1007/978-3-319-97982-3_16.
[31] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning Rich Features from RGB-D Images for Object Detection and Segmentation," in *Computer Vision – ECCV 2014*, Cham, 2014, pp. 345–360. doi: 10.1007/978-3-319-10584-0_23.

[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 580–587. doi: 10.1109/CVPR.2014.81.
[33] "OpenCV: Depth Map from Stereo Images." https://docs.opencv.org/4.x/dd/d53/tutorial_py_depthmap.html (accessed Oct. 18, 2022).
[34] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge: Cambridge University Press, 2004. doi: 10.1017/CBO9780511811685.
[35] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 564–571. Accessed: Sep. 01, 2022. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2013/html/Gupta_Perceptual_Organization_and_2013_CVPR_paper.html
[36] X. Chen, "Depth2HHA." Mar. 10, 2022. Accessed: Oct. 18, 2022. [Online]. Available: https://github.com/charlesCXK/Depth2HHA
[37] S. Gupta, "s-gupta/rcnn-depth." Oct. 14, 2022. Accessed: Oct. 18, 2022. [Online]. Available: https://github.com/s-gupta/rcnn-depth/blob/7a7baf7dcccc6fdf6be7c13d16828064d89dff4e/rcnn/saveHHA.m
[38] F. Tan, Z. Xia, Y. Ma, and X. Feng, "3D Sensor Based Pedestrian Detection by Integrating Improved HHA Encoding and Two-Branch Feature Fusion," *Remote Sens.*, vol. 14, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/rs14030645.
[39] C. H. Lund, L. B. Nonato, J. M. Kuller, L. S. Franck, C. Cullander, and D. K. Durand, "Disruption of barrier function in neonatal skin associated with adhesive removal," *J. Pediatr.*, vol. 131, no. 3, pp. 367–372, Sep. 1997, doi: 10.1016/S0022-3476(97)80060-1.
[40] B. Wallace, L. Y. Kassab, A. Law, R. Goubran, and F. Knoefel, "Contactless Remote Assessment of Heart Rate and Respiration Rate Using Video Magnification," *IEEE Instrum. Meas. Mag.*, vol. 25, no. 1, pp. 20–27, Feb. 2022, doi: 10.1109/MIM.2022.9693458.
[41] G. de Haan and V. Jeanne, "Robust Pulse Rate From Chrominance-Based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013, doi: 10.1109/TBME.2013.2266196.
[42] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, p. 65:1-65:8, Jul. 2012, doi: 10.1145/2185520.2185561.
[43] A. K. Abbas, K. Heimann, K. Jergus, T. Orlikowsky, and S. Leonhardt, "Neonatal non-contact respiratory monitoring based on real-time infrared thermography," *Biomed. Eng. OnLine*, vol. 10, no. 1, p. 93, Oct. 2011, doi: 10.1186/1475-925X-10-93.
[44] D. G. Kyrollos, K. Greenwood, J. Harrold, and J. R. Green, "Detection of False Alarms in the NICU Using Pressure Sensitive Mat," in *2021 IEEE Sensors Applications Symposium (SAS)*, Aug. 2021, pp. 1–5. doi: 10.1109/SAS51076.2021.9530191.
[45] F.-T.-Z. Khanam, A. Al-Naji, A. G. Perera, K. Gibson, and J. Chahl, "Non-contact automatic vital signs monitoring of neonates in NICU using video camera imaging," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 0, no. 0, pp. 1–8, May 2022, doi: 10.1080/21681163.2022.2069598.
[46] S. Fernando *et al.*, "Feasibility of Contactless Pulse Rate Monitoring of Neonates using Google Glass," in *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, Brussels, BEL, Dec. 2015, pp. 198–201. doi: 10.4108/eai.14-10-2015.2261589.
[47] J. Jorge *et al.*, "Non-Contact Monitoring of Respiration in the Neonatal Intensive Care Unit," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, May 2017, pp. 286–293. doi: 10.1109/FG.2017.44.

[48] M. Villarroel *et al.*, "Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit," *Npj Digit. Med.*, vol. 2, no. 1, Art. no. 1, Dec. 2019, doi: 10.1038/s41746-019-0199-5.
[49] Y. S. Dosso, R. Selzler, K. Greenwood, J. Harrold, and J. R. Green, "RGB-D Sensor Application for Non-Contact Neonatal Monitoring," in *2021 IEEE Sensors Applications Symposium (SAS)*, Aug. 2021, pp. 1–6. doi: 10.1109/SAS51076.2021.9530044.
[50] L. Maurya, P. Kaur, D. Chawla, and P. Mahapatra, "Non-contact breathing rate monitoring in newborns: A review," *Comput. Biol. Med.*, vol. 132, p. 104321, May 2021, doi: 10.1016/j.compbiomed.2021.104321.
[51] C. Eastwood-Sutherland, T. J. Gale, P. A. Dargaville, and K. Wheeler, "Elements of vision based respiratory monitoring," in *2015 8th Biomedical Engineering International Conference (BMEiCON)*, Nov. 2015, pp. 1–5. doi: 10.1109/BMEiCON.2015.7399536.
[52] A. Cenci, D. Liciotti, E. Frontoni, A. Mancini, and P. Zingaretti, "Non-Contact Monitoring of Preterm Infants Using RGB-D Camera," presented at the ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Jan. 2016. doi: 10.1115/DETC2015-46309.
[53] H. Rehouma, R. Noumeir, W. Bouachir, P. Jouvet, and S. Essouri, "3D imaging system for respiratory monitoring in pediatric intensive care environment," *Comput. Med. Imaging Graph.*, vol. 70, pp. 17–28, Dec. 2018, doi: 10.1016/j.compmedimag.2018.09.006.
[54] L. Antognoli, P. Marchionni, S. Spinsante, S. Nobile, V. P. Carnielli, and L. Scalise, "Enanced video heart rate and respiratory rate evaluation: standard multiparameter monitor vs clinical confrontation in newborn patients," in *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Jun. 2019, pp. 1–5. doi: 10.1109/MeMeA.2019.8802147.
[55] M.-C. Yu, H. Wu, J.-L. Liou, M.-S. Lee, and Y.-P. Hung, "BREATH AND POSITION MONITORING DURING SLEEPING WITH A DEPTH CAMERA," in *Proceedings of the International Conference on Health Informatics - HEALTHINF, (BIOSTEC 2012)*, 2012, pp. 12–22. doi: 10.5220/0003702000120022.
[56] S. Orlandi *et al.*, "Detection of Atypical and Typical Infant Movements using Computer-based Video Analysis," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2018, pp. 3598–3601. doi: 10.1109/EMBC.2018.8513078.
[57] Y. Souley Dosso, K. Greenwood, J. Harrold, and J. R. Green, "RGB-D scene analysis in the NICU," *Comput. Biol. Med.*, vol. 138, p. 104873, Nov. 2021, doi: 10.1016/j.compbiomed.2021.104873.
[58] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015. doi: 10.48550/arXiv.1409.1556.
[59] M. Hozayen, S. Nizami, A. Bekele, and K. Dick, "Developing a Real-Time Patient Monitor Data Import System," p. 8, 2018.
[60] "StandInBaby® | Single," *StandInBaby®*. https://www.standinbaby.com/product/standinbaby_single/ (accessed Nov. 21, 2022).
[61] G. Bradski, "The OpenCV Library," *Dr Dobbs J. Softw. Tools*, 2000.
[62] S. Suzuki and K. be, "Topological structural analysis of digitized binary images by border following," *Comput. Vis. Graph. Image Process.*, vol. 30, no. 1, pp. 32–46, Apr. 1985, doi: 10.1016/0734-189X(85)90016-7.
[63] A. B. Bekele, "Neonatal Respiratory Rate Monitoring Using a Pressure-Sensitive Mat," Text, Carleton University, 2019. Accessed: Feb. 04, 2022. [Online]. Available: https://curve.carleton.ca/2b48d44a-5bd3-48e3-9148-bf635a49111d
[64] T. Sørenson, *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. I kommission hos E. Munksgaard, 1948.

[65] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945, doi: 10.2307/1932409.
[66] P. Jaccard, "The Distribution of the Flora in the Alpine Zone.1," *New Phytol.*, vol. 11, no. 2, pp. 37–50, 1912, doi: 10.1111/j.1469-8137.1912.tb05611.x.
[67] R. Wightman, "PyTorch Image Models," *GitHub repository*. GitHub, 2019. doi: 10.5281/zenodo.4414861.
[68] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009. Accessed: Nov. 30, 2022. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html
[69] "Vision Transformer and MLP-Mixer Architectures." Google Research, Nov. 25, 2022. Accessed: Nov. 25, 2022. [Online]. Available: https://github.com/google-research/vision_transformer