A Machine Learning Driven Pipeline for Automated Photoplethysmogram Signal Artifact Detection

Luca Cerny Oliveira
Electrical and Computer Engineering
University of California, Davis
Davis, CA, USA
lcernyo@ucdavis.edu

Zhengfeng Lai
Electrical and Computer Engineering
University of California, Davis
Davis, CA, USA
lzhengfeng@ucdavis.edu

Wenbo Geng
Electrical and Computer Engineering
University of California, Davis
Davis, CA, USA
wgeng@ucdavis.edu

Heather Siefkes
Pediatrics
University of California, Davis
Sacramento, CA, USA
hsiefkes@ucdavis.edu

Chen-Nee Chuah
Electrical and Computer Engineering
University of California, Davis
Davis, CA, USA
chuah@ucdavis.edu

***Abstract*—Recent advances in Critical Congenital Heart Disease (CCHD) research using Photoplethysmography (PPG) signals have yielded an Internet of Things (IoT) based enhanced screening method that performs CCHD detection comparable to SpO2 screening. The use of PPG signals, however, poses a challenge due to its measurements being prone to artifacts. To comprehensively study the most effective way to remove the artifact segments from PPG waveforms, we performed feature engineering and investigated both Machine Learning (ML) and rule based algorithms to identify the optimal method of artifact detection. Our proposed artifact detection system utilizes a 3-stage ML model that incorporates both Gradient Boosting (GB) and Random Forest (RF). The proposed system achieved 84.01% of Intersection over Union (IoU), which is competitive to state-of-the-art artifact detection methods tested on higher resolution PPG.**

***Index Terms*—PPG, CCHD, artifacts, Machine Learning**

I. Introduction

Congenital heart disease (CHD) is the leading cause of birth-defect associated infant illness and death [1]. CHD is also the most common birth defect, affecting nearly 0.8% of all newborns [1]. About 25% of these CHD cases belong to the critical congenital heart disease (CCHD) subset, the most dangerous one [1], [2]. CCHD must be detected as soon as possible because these lesions require surgical or catheter-based intervention soon after birth. If not detected soon after birth, CCHD can lead to preventable poor outcomes, including death [1], [3], [4]. Before the development of oxygen-saturation (SpO2) based CCHD screening, 25% of CCHD-positive newborns would go home undiagnosed [4], [5]. SpO2 screening has helped with earlier diagnosis and diminished sequelae, however an estimated 6.4 deaths due to CCHD occur per 100,000 births in the United States despite mandated screening. [6]

Therefore, efforts aimed at improving postnatal detection of these life-threatening lesions are necessary. To address that, Doshi *et. al* proposed a system that reads dual pulse oximetry data and uploads it online through an Internet of Things (IoT) setup [7]. This system would calculate many of the newborn's health indicators through Photoplethysmography (PPG) pulse oximetry devices and allowed for quick online access of the acquired data. The data from that study was later employed on a Machine Learning (ML) classifier to achieve better CCHD detection metrics than SpO2 screening [8]. Although the preliminary results from [8] were promising, using PPG signals still poses a challenge due to artifacts. An artifact is a period of the measured waveform that was affected by an external factor such as movement, touch or others (refer to Figure 1 and Figure 2). Due to relying on constant contact with the skin, PPG signals are artifact prone. The waveform affected by artifacts cannot be used for medical analysis. This limits the quantity of useful data acquired, especially when the subjects are prone to unpredictable movement, which is the case for infants.

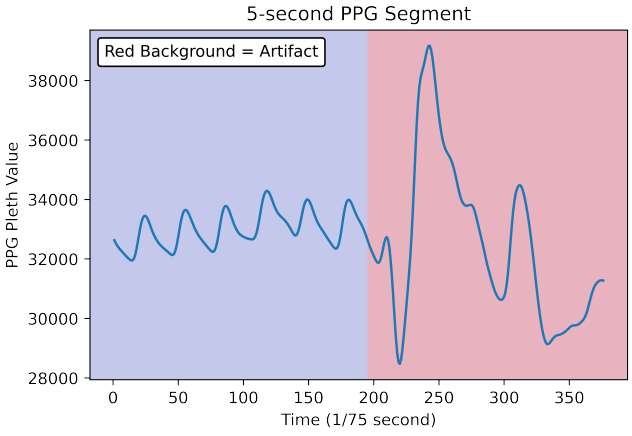

Fig. 1. An example of five second segment of a PPG signal with a normal start followed by a clear artifact.

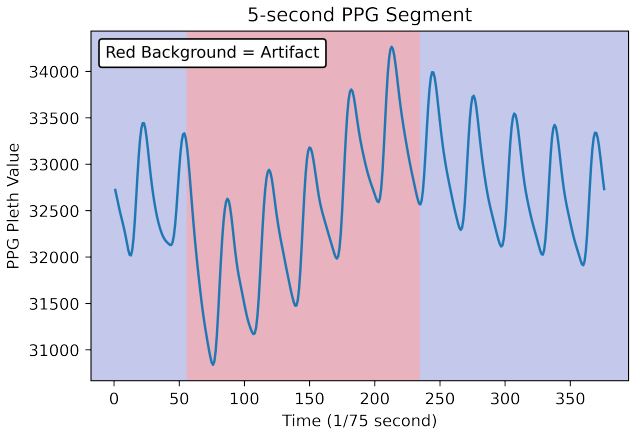

Fig. 2. An example of five second segment of a PPG signal with a normal start, a less clear artifact in the middle, and then a normal ending.

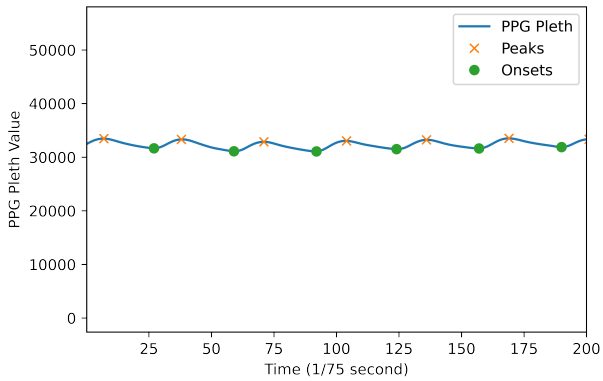

Fig. 3. An example of PPG segment with its onsets and peaks signaled. The beat segments are extracted from onset to onset.

Since the normal PPG signals are highly ordered and homogeneous (see Figure 1), trained annotators can identify and label artifacts. While human annotation is a reliable method to separate the artifact segments, it is time-consuming and labor intensive. Thus, an automated approach to identify PPG artifact is necessary.

The challenge of separating artifacts segments from normal ones in PPG has been tackled by many authors [9]–[14]. They have achieved good results with rule based [9], [10], ML [12] and deep learning methods [11], [13], [14]. However, these studies were done on higher resolution PPG data which was able to display details such as dicrotic notches, which are not found in many pulse oximeters, even those commonly used in medical practice.

With the challenge of extracting artifacts from inexpensive PPG signals in mind, we propose an end-to-end pipeline for automated waveform artifact detection. Such pipeline would receive a raw PPG waveform as input, and return the exact location of where the artifacts are located.

Our contributions can be summarized as follows:

- We combine different ML models to construct a precise framework for the detection of artifacts in PPG signals. To our best knowledge, it is the first time a combination of ML models is used to identify artifacts in PPG signals.
- We evaluate the proposed framework on an artifact detection task using IoT based extracted PPG signals from newborns. Upon evaluation we conclude our model's performance is competitive with state-of-the-art PPG artifact detection ML models [12].

II. Related Work

A. Rule Based methods

Rule based methods extract waveform features and create classification rules based on the values yielded from extracted features. In these methods, the rules are decided by the coder and the waveform must be sliced before analysis. Each rule will decide whether the analyzed segment is normal or artifact, or whether the segment will be analyzed by another rule.

In [9], precise artifact detection was achieved through features extracted on the preprocessed segments only. The features extracted were focused on clinically relevant measurements such as diastolic and systolic phase length, and amount of dicrotic notches. In [10], they achieved precise artifact detection when coupling features based on the preprocessed waveform's segment with features based on a Gaussian fit of said segment. The Gaussian fit would assume the dicrotic notch in each segment is the peak of an independent Gaussian distribution. In all the aforementioned studies, the waveform slicing method chosen was to divide segments based on consecutive onsets (see this type of segmentation in Figure 3). The period from one onset to the next has been called a "pulse" [9], [10], and in this study we will refer to them as beats (see an example of a beat in Figure 4). All features used in previous studies were extracted from those beats. The rules were applied linearly, sometimes utilizing ratios that utilized features from previous beats.

B. Machine Learning methods

In [12], random slices of waveform were taken and certain features were extracted from each slice. These features were then fed to a multitude of ML models and observed an Accuracy of 84% $\pm$ 2.89. The study was also able to assess that Random Forest is the best performing ML model for the specific problem setup analyzed. However, just like the previously seen rule based methods [9], [10] this ML method [12] uses PPG data with higher resolution that is capable of displaying the dicrotic notches of every beat.

III. Methods

A. Data Collection

The PPG data was collected through a real-time data analytic pipeline similar to [7] to improve CCHD detection. The pipeline gathered PPG data through two relatively inexpensive Nonin© WristOx2™ 3150 pulse oximetry medical devices attached to the subject's right hand and one foot. The pulse oximetry devices transmit via Bluetooth to a Pi-Top™ device,

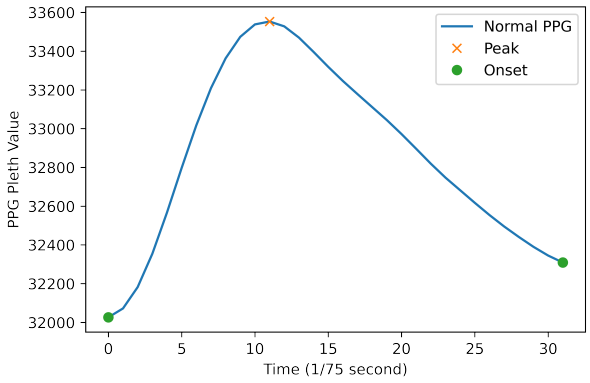

Fig. 4. An example of normal beat segment with its peak and onsets signaled on the graph.

which aggregates the data from both pulse oximetry devices. The Pi-Top serves as the liaison between the medical devices and the internet, making this an IoT setup where clinicians can easily access the data soon after its collection without the necessity of being on site. Using this setup, we were able to harvest PPG waveform, perfusion index (PIx) and other features that are aggregated and calculated in real-time by the Pi-Top™. The data collected is then transferred to a hard-drive and sent to REDCap, an encrypted online database that can be accessed by clinicians and our study's personnel. Due to security and privacy concerns, the Pi-Top™ is not directly connected to the Internet. Since this study focus on the removal of artifacts from PPG waveform, we only kept the de-identified PPG waveform data and discarded all other information such as PIx and SpO2.

B. Subjects

The PPG data (57,659 beats) was acquired from 21 newborns. The patients were enrolled at University of California (UC), Davis, Sutter Medical Center in Sacramento, UC Los Angeles, UC San Francisco, and Cohen Children's Medical Center. From each patient, we recorded at least 5-minute PPG measurements from a foot and hand simultaneously at three different time periods: within 24 hours, 24-48 hours, and after 48 hours following the infant's birth. The data collection efforts for this study yielded 6 hours and 42 minutes of PPG signals.

C. Data Preparation

The acquired PPG signals were annotated by trained observers. The annotators had access to the entire waveform and labeled sections of artifact. We defined artifact free segments, also called normal, as a minimum of 10 consecutive beats without artifact. Given the task of annotating artifacts has inherent subjectivity, we employed two annotators in order to acquire a reliable ground truth. After overlapping their annotations, it was observed that the two annotators disagreed on 12% of the total 6 hours and 42 minutes of annotation, amounting to approximately 48 minutes of PPG signals where annotators disagreed on its classification.

After annotation, we then segmented each waveform into beat segments. The segmentation was done through automatic identification of the waveform's onsets as seen in Figure 3. Every segment between two onsets was extracted, and its peak was automatically detected as seen in Figure 4. There were 57,659 beat segments extracted. Out of these segments: 31,989 segments were labeled as artifact, 19,736 segments were labeled as normal, and 5,934 were labeled as disagreed.

Although there was no preprocessing of our PPG data, we did automatic artifact classification of beat segments larger than 2.4 seconds. A beat of 2.4 seconds indicates the subject's heart rate (HR) is at 25 beats per minute (BPM, see Equation 1), a value we are confident did not occur for newborns during our data collection [15].

$$\text{HR} = \frac{60\ seconds}{\text{Length of Beat Segment in seconds}} \tag{1}$$

D. Feature Extraction

Throughout our experiment we attempted both a ML and a rule based approach. When approaching the feature extraction, we sought to find features that would yield noticeable difference for normal *versus* artifact segments, as well as replicate useful features from previous studies [9], [10], [12].

An example of feature extracted due to noticeable difference between normal and artifact is Averaged Dynamic Time Warping Euclidean distance $\mathcal{E}_{AvDTW}$ (see Algorithm 1). Two examples of features replicated from previous studies are diastolic and systolic phase duration [9], [10]. When analyzing a beat segment we can find a local peak (see Figure 4). The period from the first onset to the peak is the systolic phase, while the period from the peak to the second onset is the diastolic phase.

Other features include: Onset-Amplitude Ratio (OAR, see Equation 2), Amplitude (Amp), beat duration, value at half of systolic phase, peak-mid systolic distance (PMSD, see Equation 3), peak-mid diastolic distance (PMDD, see Equation 4). There were also features that used values from previous neighboring beats such as: Amp ratio (see Equation 5), Diastolic phase ratio (see Equation 6), Systolic phase ratio (see Equation 7), and diastolic/systolic duration ratio (see Equation 8).

$$OAR = \frac{(PPG\ value\ at\ onset_0 + PPG\ value\ at\ onset_1)}{Amp} \tag{2}$$

$$PMSD = \frac{Amp - PPG\ value\ at\ half\ Systolic\ length}{Systolic\ length * 0.5} \tag{3}$$

$$PMDD = \frac{Amp - PPG\ value\ at\ half\ Diastolic\ length}{Diastolic\ length * 0.5} \tag{4}$$

$$Amp\ ratio = \frac{Amp_t}{Amp_{t-1}} \tag{5}$$

$$Diastolic\ Phase\ Ratio = \frac{(Diastolic\ duration)_t}{(Diastolic\ duration)_{t-1}} \quad (6)$$

$$Systolic\ Phase\ Ratio = \frac{(Systolic\ duration)_t}{(Systolic\ duration)_{t-1}} \quad (7)$$

$$Diastolic/Systolic\ duration\ Ratio = \frac{Diastolic\ duration}{Systolic\ duration} \quad (8)$$

E. Artifact Detection Algorithms

a) ***Rule Based****:* When building the rule based method, we followed the trend of previous studies [9], [10] and built a linear set of rules. Meaning each feature is transformed into a rule and if the segment satisfies its conditions, the segment is then analyzed by a different rule. The process goes linearly until all rules have been passed and the segment is classified as normal, or if one rule spots the segment to be an artifact, which then ends the evaluation. Our rule scheme can be seen in Figure 5.

b) ***Machine Learning****:* In order to build the ML classifier, we tested Logistic Regression, Random Forest (RF), Decision Trees and Gradient Boosting (GB). In order to choose which features would be fed to the models, Recursive Feature Elimination (RFE) was employed. When performing RFE, we aimed at picking the features that would yield the highest Sensitivity for artifact detection.

RF and GB were the two best performing models. RF had high Sensitivity (Sens) and GB displayed high Specificity (Spec). Upon combining both models, we were able to achieve a 3-stage ML based classifier that had better performance than using either just RF or GB. The 3-stage ML is described in Figure 6.

F. Performance Evaluation

Artifact detection is a segmentation problem where each part of the waveform could either be an artifact or normal. Since we approach the problem by classifying each beat segment, we thus used the metrics employed in classification problems to evaluate the performance of the different algorithms: Sensitivity, Specificity, and Accuracy (Acc). However, given that the problem still is a segmentation problem at core, we decided to also use Intersection over Union (IoU) score to achieve more comprehensive evaluation.

$$\text{Sensitivity} = \text{TPR} = \frac{TP}{TP + FN} \quad (9)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (10)$$

$$\text{Accuracy} = \frac{TP + TN}{FP + TP + FN + TN} \quad (11)$$

where:

- TP: the number of artifact predicted as artifact
- FP: the number of normal predicted as artifact
- TN: the number of normal predicted as normal
- FN: the number of artifact predicted as normal

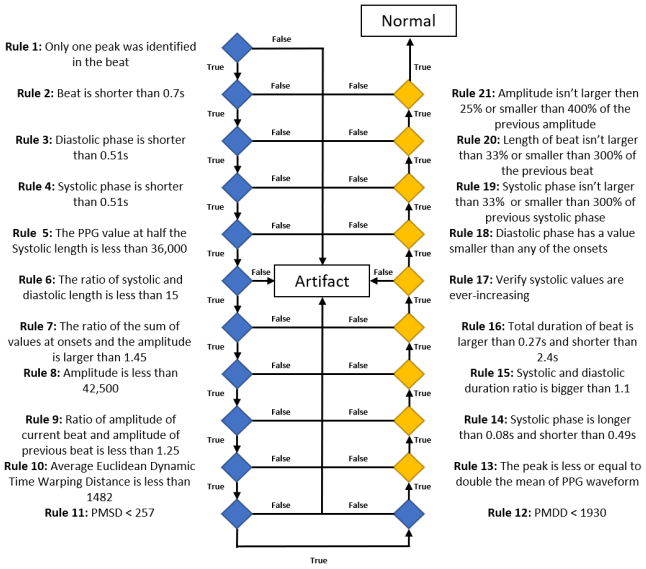

Fig. 5. The rule based algorithm employed by this study. The rules in blue were developed based on our feature engineering. The yellow rules were based on previous studies that also employed rule based artifact detection mechanisms [9], [10]

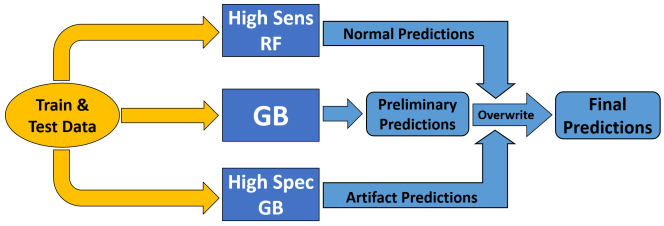

Fig. 6. The proposed 3-Stage ML model. Here, the predictions from the all-around GB model are used as preliminary predictions. Then, if the High Sens RF model predicted any normal (negatives), we overwrite the all-around predictions for those indexes. If the High Spec GB model predicted any artifact (positives), we overwrite the all-around predictions for those indexes. The end result is the final predictions.

IoU score. IoU scores the amount of overlap present between two different measurements. In the example of our artifact detection study, IoU would measure the amount of overlap between the ground truth and the algorithms' prediction. Sensitivity, Specificity, and Accuracy provide reliable information on how many segments were correctly classified. However, these beats, especially the artifact ones, can vary in duration and thus IoU score allows us to calculate what percentage of the total waveform length was correctly segmented.

IV. Results

A. Agreed Beats vs. All Beats

All experiments were done on either all beats or on only agreed beats. When only agreed beats were employed, all beats where the annotators disagreed on its class were removed from both train and test set. When all beats were used, the disagreed beats were considered artifacts.

TABLE I
SUCCESS METRIC SCORES FOR ALL BEATS AND ONLY AGREED BEATS

Beats Used	All Beats				Only Agreed Beats			
Method	Rule Based	RF	GB	3 Stage ML	Rule Based	RF	GB	3 Stage ML
IoU	70.94%	77.20%	78.29%	78.55%	74.01%	81.71%	82.49%	**84.01%**
Accuracy	72.06%	77.79%	78.99%	79.05%	74.98%	81.71%	82.77%	**84.27%**
Specificity	60.15%	55.42%	67.62%	64.68%	60.15%	68.40%	80.67%	**82.18%**
Sensitivity	77.52%	88.22%	84.29%	85.75%	83.31%	**89.65%**	83.95%	85.44%

All Beats evaluation had every single segment extracted from the cases seen and disagreed beats were considered artifacts. Only Agreed Beats evaluation had disagreed beats removed from train and test set before inference.

Algorithm 1 Extraction of Dynamic Time Warping Features

Input: Raw waveform divided into beat segments
Output: obtain Euclidean Dynamic Time Warping Distance $\mathcal{E}_{DTW}$ and Averaged Euclidean Dynamic Time Warping Distance $\mathcal{E}_{AvDTW}$
Initialize: Acquire beat $\mathcal{B}_1$ of length $\mathcal{L}_1$ seconds and the previous beat $\mathcal{B}_0$ of length $\mathcal{L}_0$ seconds. Set total length $\mathcal{L}$ to be $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1$
Feature Extraction: $\mathcal{E}_{DTW} = \mathcal{DTW}(\mathcal{B}_1, \mathcal{B}_0)$ and $\mathcal{E}_{AvDTW} = \mathcal{DTW}(\mathcal{B}_1, \mathcal{B}_0) \div \mathcal{L}$

B. IoU and Accuracy

IoU and Accuracy (Acc) have similar values due to the segments length being fairly stable across all normal and most artifact beat segments. We observed that the ML models significantly outperformed the rule based method as shown in Table I The IoU and Acc metrics taken for disagreed-beats-based ML models performed better than agreed-beats-based rule based methods, which denotes significant superior performance. When comparing different ML models, IoU and Acc are superior for the proposed 3 stage ML model.

TABLE II
HIGH SENS/SPEC MODELS USED IN 3 STAGE ML

Models	Specificity	Sensitivity
RF high Sens	46.76%	95.34%
GB high Spec	93.88%	61.82%

C. Sensitivity and Specificity

The Sensitivity and Specificity of the tested models can be seen both in Table I and Table II. In Table I we see the metrics from the standard RF and GB, as well as for the 3-stage ML model. In Table II we see the metrics from the high Specificity and the high Sensitivity ML models that were later combined with GB to make the 3-stage ML model (see Figure 6). In order to achieve the high Specificity and high Sensitivity ML models, we adjusted the prediction confidence level needed for artifact prediction. The best model all-around was GB. RF had an inherent tendency to achieve high Sensitivity, while GB had an inherent tendency to achieve high Specificity.

V. DISCUSSION

In this study, we investigated the use of rule based and Machine Learning based Artifact Detection algorithm with PPG data from inexpensive IoT based automated collection. We propose a 3-stage Random Forest-Gradient Boosting Artifact Detection model that would balance a high Specificity, a high Sensitivity and an all-around model to achieve high IoU score and accuracy. For our dataset, the 3-stage Machine Learning classifier significantly outperformed the rule based classifiers, achieving 10% higher IoU in both settings where disagreed beats were included in the analysis and when they were removed.

As part of our future work, we aim to increase our dataset and attempt Deep Learning methods. We also seek to employ our 3-stage ML model on a Pi-Top™ so we can perform artifact detection and removal before the PPG data is uploaded to the online encrypted database. A setup with integrated artifact detection and removal would allow for seamless acquisition of important features used for CCHD detection.

ACKNOWLEDGMENT

This work was supported by National Science Foundation (NSF) Harnessing the Data Revolution: Transdisciplinary Research in Principles of Data Science Phase I (HDR: TRIPOD) grant CCF-1934568. The project described was also supported by the National Center for Advancing Translational Sciences, National Institutes of Health (NIH), through grant number UL1 TR001860 and linked award KL2 TR001859, the Eunice Kennedy Shriver National Institute of Child Health & Human Development, NIH, through grant number 1R21HD099239-02, the University of California, Davis (UCD) Artificial Intelligence Seed Grant and UCD Venture Catalyst DIAL Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIH, or UCD.

CONFLICTS OF INTEREST

Chen-Nee Chuah, Zhengfeng Lai, and Heather Siefkes are named as inventors on a patent application "Systems and Methods for Classifying Critical Heart Defects." Heather Siefkes is the founding member of NeoPOSE Inc., a company working to develop devices to detect CCHD.

REFERENCES

[1] M. D. Reller, M. J. Strickland, T. Riehle-Colarusso, W. T. Mahle, and A. Correa, "Prevalence of congenital heart defects in metropolitan atlanta, 1998-2005," *J. Pediatr.*, vol. 153, no. 6, pp. 807–813, 2008.

[2] E. C. Ailes, S. M. Gilboa, M. A. Honein, and M. E. Oster, "Estimated number of infants detected and missed by critical congenital heart defect screening," *Pediatrics*, vol. 135, no. 6, pp. 1000–1008, 2015.

[3] R. I. Koppel, C. M. Druschel, T. Carter, B. E. Goldberg, P. N. Mehta, R. Talwar, and F. Z. Bierman, "Effectiveness of pulse oximetry screening for congenital heart disease in asymptomatic newborns," *Pediatrics*, vol. 111, no. 3, pp. 451–455, 2003.

[4] C. Wren, Z. Reinhardt, and K. Khawaja, "Twenty-year trends in diagnosis of life-threatening neonatal cardiovascular malformations," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, vol. 93, no. 1, pp. F33–F35, 2008.

[5] K. Lannering, M. Bartos, and M. Mellander, "Late diagnosis of coarctation despite prenatal ultrasound and postnatal pulse oximetry," *Pediatrics*, vol. 136, no. 2, pp. e406–e412, 2015.

[6] R. Abouk, S. D. Grosse, E. C. Ailes, and M. E. Oster, "Association of us state implementation of newborn screening policies for critical congenital heart disease with early infant cardiac deaths," *Jama*, vol. 318, no. 21, pp. 2111–2118, 2017.

[7] K. Doshi, G. B. Rehm, P. Vadlaputi, Z. Lai, S. Lakshminrusimha, C.-N. Chuah, and H. M. Siefkes, "A novel system to collect dual pulse oximetry data for critical congenital heart disease screening research," *Journal of Clinical and Translational Science*, vol. 5, no. 1, 2021.

[8] Z. Lai, P. Vadlaputi, D. Tancredi, M. Garg, R. Koppel, M. Goodman, W. Hogan, N. Cresalia, S. Juergensen, E. Manalo, S. Lakshminrusimha, C.-N. Chuah, and H. Siefkes, "Enhanced critical congenital cardiac disease screening by combining interpretable machine learning algorithms," in *2021 43nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021.

[9] C. Fischer, B. Dömer, T. Wibmer, and T. Penzel, "An algorithm for real-time pulse waveform segmentation and artifact detection in photoplethysmograms," *IEEE journal of biomedical and health informatics*, vol. 21, no. 2, pp. 372–381, 2016.

[10] Q. Hu, X. Deng, X. Liu, A. Wang, and C. Yang, "A robust beat-to-beat artifact detection algorithm for pulse wave," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[11] C.-H. Goh, L. K. Tan, N. H. Lovell, S.-C. Ng, M. P. Tan, and E. Lim, "Robust ppg motion artifact detection using a 1-d convolution neural network," *Computer methods and programs in biomedicine*, vol. 196, p. 105596, 2020.

[12] T. Athaya and S. Choi, "Evaluation of different machine learning models for photoplethysmogram signal artifact detection," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2020, pp. 1206–1208.

[13] D. Dao, S. M. Salehizadeh, Y. Noh, J. W. Chong, C. H. Cho, D. McManus, C. E. Darling, Y. Mendelson, and K. H. Chon, "A robust motion artifact detection algorithm for accurate detection of heart rates from photoplethysmographic signals using time–frequency spectral features," *IEEE journal of biomedical and health informatics*, vol. 21, no. 5, pp. 1242–1253, 2016.

[14] J. Torres-Soto and E. A. Ashley, "Multi-task deep learning for cardiac rhythm detection in wearable devices," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.

[15] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. Maconochie, L. Tarassenko, and D. Mant, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies," *The Lancet*, vol. 377, no. 9770, pp. 1011–1018, 2011.