Evaluating Artifact Detection Algorithms for the Arterial Blood Pressure Waveform Acquired from the Intensive Care Unit: A PRECICECAP Informatics Approach

Tony K. Okeke, Manil Shrestha, Ethan Moyer, Karen G. Hirsch, Teresa L. May, Zihuai He, Richard Moberg, and Jonathan Elmer

***Abstract*—Objective: This study aims to develop and evaluate automated methods for detecting artifacts in arterial blood pressure waveforms from intensive care unit monitoring, with the goal of improving data quality for both clinical interpretation and machine learning applications. Methods: We analyzed data from a prospective, multicenter study including intensive care unit-monitored patients resuscitated from cardiac arrest. We developed and evaluated 17 different models for automated artifact detection in arterial blood pressure waveforms. Using a two-stage annotation process, we first asked research assistants identify regions containing artifacts, then had clinical experts annotate individual pulses within these regions. We evaluated six different approaches to artifact detection: rule-based heuristics, traditional machine learning with handcrafted features, deep learning architectures, transfer learning with pretrained image classifiers, patient-specific model fine-tuning, and ensemble methods. Results: We included data from 35 patients. On an independent test set of 1,423 pulses from five patients, our best-performing model, an ensemble combining ResNet-18 and EfficientNet-B0, achieved 89.0% accuracy, 91.1% sensitivity, and 86.0% specificity. Conclusion: Our results demonstrate the feasibility of automated artifact detection in arterial blood pressure waveforms and suggest that ensemble approaches combining multiple deep learning models may be optimal for this task.**

***Index Terms*—Artifact detection, arterial blood pressure, machine learning, deep learning, transfer learning.**

I. Introduction

PATIENTS treated in the intensive care unit (ICU) often undergo continuous cardiopulmonary monitoring. The resulting waveform data are a rich source of information for clinicians and scientists. Arterial blood pressure (ABP) tracings are a specific example where the waveform characteristics can provide critical information about cardiovascular status (e.g., preload, contractility, and afterload) beyond the blood pressure itself [1]. ICU waveform data are prone to artifacts caused by patient movement, electromagnetic interference, sensor malfunctions, medical interventions, noise from nearby devices, and other issues [2]. This can limit the ability of researchers to use this data to interpret real-time monitoring data. Artifacts can also adversely affect the performance of artificial intelligence and machine learning (AI/ML) models trained on such physiologic data [3], [4]. The development of automated artifact detection algorithms requires careful development and has the potential to facilitate clinical care and enhance the rigor and reproducibility of research.

The PREcision Care In Cardiac arrEst - ICECAP (PRECICECAP) study [5] collects high-resolution multiparametric data, including continuous cardiopulmonary waveforms such as ABP, from patients enrolled in the prospective Influence of Cooling duration on Efficacy in Cardiac Arrest Patients (ICECAP) trial [6]. PRECICECAP aims to discover novel phenotypes of treatment responsiveness by applying AI/ML to inputs including cardiopulmonary waveform data. We aimed to develop and evaluate automated approaches to artifact detection in ABP waveforms with the ultimate goal of developing an automated data curation pipeline to facilitate AI/ML readiness [5].

II. Methods

A. Setting and Data Sources

We performed a secondary analysis of a prospective observational cohort study. The ICECAP trial is a prospective randomized, controlled trial that aims to identify the optimal duration of hypothermic temperature control after resuscitation from out-of-hospital cardiac arrest (OHCA) [6]. Briefly, after informed consent from a legally authorized representative,

Manuscript submitted November 19, 2024. This work was supported in part by the National Institute of Neurological Disorders and Stroke (NINDS) under Award R01NS119825 and its supplements (01S1, 03S1). This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Advarra Institutional Review Board. Corresponding author: Jonathan Elmer.

Tony K. Okeke, Ethan Moyer and Richard Moberg are with Moberg Analytics, Inc., Philadelphia, PA, USA.

Manil Shrestha is with Drexel University, Philadelphia, PA, USA.

Karen G. Hirsch is with the Department of Neurology, Stanford University, Palo Alto, CA, USA.

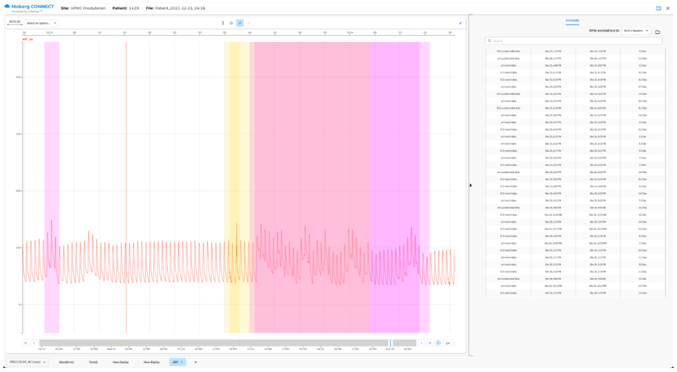
Teresa L. May is with the Department of Critical Care Services, Neuroscience Institute, Maine Medical Center, Portland, ME, USA.

Zihuai He is with the Department of Neurology and the Quantitative Sciences Unit, Department of Medicine at Stanford University, Stanford, CA, USA.

Jonathan Elmer is with the Department of Emergency Medicine, Critical Care Medicine and Neurology, University of Pittsburgh, Pittsburgh, PA, USA (e-mail: elmerjp@upmc.edu).

ICECAP enrolls comatose survivors of OHCA who have a clinical indication for hypothermic temperature control to 33°C and reach a core temperature of less than 34°C within four hours of cardiac arrest. Participants are randomized to one of ten possible durations of hypothermic temperature control ranging from 6 to 72 hours in a Bayesian response-adaptive allocation ratio. All trial procedures and the scope of this secondary analysis were approved by the Advarra Institutional Review Board.

The PRECICECAP ancillary study collects high-resolution multiparametric physiological waveforms and other data. PRECICECAP aims to discover novel phenotypes that predict treatment responsiveness and/or long-term recovery [5]. Thirteen high-enrolling sites participating in ICECAP also participate in PRECICECAP, and all ICECAP participants treated at these sites are eligible for co-enrollment. A Moberg CNS Monitor (Natus, USA) records continuous physiological data from the ICU clinical monitor, including ABP, at the native sampling frequency of the clinical monitor (typically 125Hz) from ICU admission through death or rewarming. After data collection is complete, research staff upload recorded data to Moberg Cloud Platform (MCP) (Moberg Analytics, Philadelphia, PA). This cloud software platform is designed for the storage, harmonization, visualization, and annotation of neurocritical care data. ICU-acquired waveform data is securely stored on IBM Cloud, using Docker for containerization and Simple Storage Service (S3). Upon upload, data is converted into the Hierarchical Data Format version 5 (HDF5), optimizing data integrity, performance, and compatibility with downstream analytical tools. MCP supports customizable annotation workflows to meet specific research needs, including options for manual or automated signal quality assessment, blinded multiple annotation, and flags for secondary review (Figure 1).



Site: UPMC Presbyterian Patient: 1429 File: Patient_2021-12-21_04:16

Fig. 1: Screenshot of the Data Review interface in Moberg Cloud Platform showing ABP waveform data with annotations from multiple independent reviewers (left panel). Colored regions indicate different types of signal annotations: purple regions represent regions of artifact labeled by independent viewer number 1, yellow regions indicate regions of artifact labeled by independent reviewer number 2, and red regions indicate overlapped regions of labeled artifact by both reviewers. Clicking on these artifacts allows the user to edit their properties, such as their title, description, labels, and associated measurements. The right panel displays the annotation timeline with timestamps and reviewer identifiers. Clicking on these rows will jump the user to where the annotation is in time.

B. Data Preprocessing and Annotation

For this analysis, we randomly sampled 35 participants across seven enrolling sites from the PRECICECAP cohort. To develop a ground truth reference standard for artifact identification, we used a two-stage approach. First, a PRECICECAP principal investigator (PI) trained two research assistants (RAs), both of whom had previous clinical experience as paramedics, to review the recorded ABP data to identify regions that contained artifacts. To establish baseline agreement, the PI and the RAs achieved greater than 95% concordance on artifact events within a 12-hour segment of ABP data. During the first stage of annotation, the RAs independently double-annotated the ABP data, identifying regions that were likely to contain artifacts. To account for the expected variability in the precise start and stop times of annotated artifacts and the differences in annotation of multiple artifacts in close succession (Figure 1), we focused on regions in which both RAs agreed artifact was present or absent. Annotations made while training the RAs were excluded from the analysis.

Next, we performed pulse-level segmentation of the ABP data to enable more granular annotations of artifacts. Our pulse detection algorithm first applied Gaussian smoothing to reduce signal noise, then identified pulse boundaries by detecting troughs in the inverted signal using Scipy's peak detection algorithm [7]. The algorithm identifies the local maxima by comparing each sample point with its neighboring values, ensuring detected peaks are at least the specified minimum distance apart. To ensure physiologically plausible pulse durations, we enforced a minimum distance of 0.46 seconds between troughs (equivalent to a maximum heart rate of about 130 beats per minute, which was empirically determined to optimize pulse segmentation while exceeding the 99.9th percentile of observed heart rates during monitoring).

Based on characteristics noted during first-stage RA annotation, we extracted segments of different lengths from the ABP signal: two-second segments from regions marked as artifacts to capture their typically brief and transient nature, and twenty-second segments from non-artifact regions to ensure signal stability across multiple cardiac cycles. Within these segments, we applied our pulse segmentation algorithm to identify individual pulses. Pulses extracted from segments marked as artifact by both RAs were included in the set of likely artifact pulses, while those from regions marked as artifact-free by both RAs were included in the set of likely non-artifact pulses.

During the second stage of annotation, we randomly split the 35 patients into a training cohort with 30 patients and a testing cohort with 5 patients. For each patient in the training cohort, we randomly sampled 15 likely artifact pulses and 15 likely non-artifact pulses (with one patient having only 14 available artifact pulses in the annotated segments). In total, this resulted in 899 pulses for the training dataset. These pulses were subsequently annotated by two PRECICECAP PIs using a 4-point ordinal scale: 1 = perfect waveform; 2

= very good waveform with slight artifact; 3 = substantial artifact but recognizable physiologic properties; and, 4 = no recognizable physiologic properties. The PIs were blinded to the initial annotations made by the RAs. Each PI annotated the pulses independently. In cases where there was disagreement between the two sets of annotations, the PIs met to discuss the disagreement and reach a consensus. The PIs separately identified pulses where our segmentation algorithm failed to properly isolate discrete individual pulses (e.g., capturing partial pulses or multiple pulses in one segment), and these cases were excluded from the training dataset. Finally, we binned the annotations into binary classes, where pulses annotated as 3 or 4 were classified as artifacts (positive class), and those annotated as 1 or 2 were classified as non-artifacts (negative class). These binned labels were used for subsequent model training.

The models developed in this study are intended for use in identifying and removing artifacts from ABP waveforms in the broader PRECICECAP cohort, which includes hundreds of hours of neuromonitoring data per patient. While conventional machine learning studies typically use larger training sets, our primary concern was ensuring robust evaluation under realistic clinical conditions. Therefore, for each of the 5 patients in the testing cohort, we randomly sampled approximately 300 pulses from the set of pulses extracted from their annotated segments, yielding a total of 1,573 pulses. Unlike the training dataset, we maintained the natural distribution of artifacts in these samples rather than enforcing class balance. These pulses were independently annotated as either artifact or non-artifact by two PIs, with disagreements resolved through discussion.

To create a separate validation dataset, we applied stratified random sampling to select 15 artifact and 15 non-artifact pulses from each patient in the testing cohort (150 pulses total) for use in hyperparameter tuning. This balanced sampling strategy matched our approach in creating the training dataset. The remaining 1,423 annotated pulses formed our independent test dataset. Figure 2 provides a comprehensive visualization of our data curation process, from initial study population through the final distribution of pulses across training, validation, and testing datasets.

C. Approaches to Artifact Detection

We implemented and tested eight approaches to artifact detection, one based on simple heuristics that align with clinical intuition, three using existing machine learning methods developed for similar purposes, two applying transfer learning to pre-trained image classifiers, one where models were fine-tuned on patient-specific data, and one where model results were ensembled. We identified candidate methods by performing a scoping review of the literature, selecting methods based on their applicability to high-resolution physiological data and their ability to achieve greater than 80% accuracy in previous implementations.

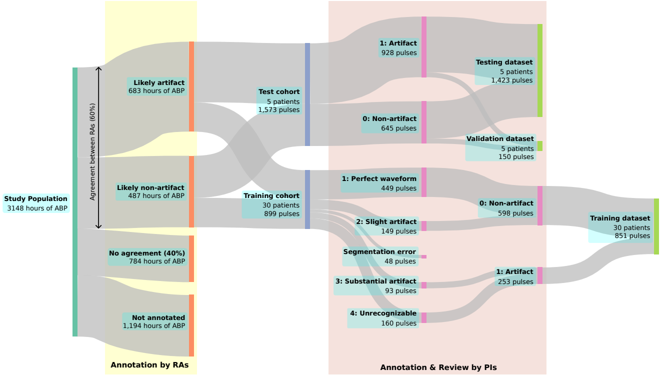


Fig. 2: Sankey diagram illustrating the data curation workflow. The process begins with 3148 hours of ABP data from the study population (left), flows through the initial annotation phase by research assistants (RA) showing agreement rates, and continues through the detailed pulse-level review by principal investigators (PI). The diagram shows how the data was ultimately distributed across training (30 patients, 851 pulses), validation (5 patients, 150 pulses), and testing (5 patients, 1,423 pulses) datasets, with detailed breakdowns of artifact classifications at each stage.

Heuristic Approach: To establish a baseline for artifact detection, we implemented a heuristic-based classification approach. This baseline utilized a combination of clinically informed rules derived from discussions between the PIs and RAs. These heuristics were designed to capture the characteristics of "good" ABP pulses based on clinical understanding. We implemented four heuristics to classify pulses as either artifact or non-artifact. First, we established a minimum value threshold of 10 mmHg, as even a pulseless patient maintains a mean systemic filling pressure around this value. Second, we set a maximum range threshold of 200 mmHg for the difference between maximum and minimum values, as larger pulse pressures strain physiological credulity. Third, we implemented a minimum range threshold of 15 mmHg, as pulses with smaller ranges likely represent dampened or artificial signals. Finally, we set a peak count threshold of 5, where signals with more peaks were classified as artifacts, as physiological ABP pulses typically display a single primary peak with a single dicrotic notch. To optimize these values, we performed a grid search on the training dataset, selecting values that balanced both accuracy and sensitivity. We evaluated model performance using both the initial and optimized thresholds on the independent test dataset. While this rule-based approach provides interpretable results and is computationally efficient, its performance is heavily dependent on the chosen thresholds and may not generalize will across different patient populations or monitoring conditions.

Handcrafted Feature Engineering (HCFE): In this approach, we compute various statistical features from the ABP pulses and use them to train supervised machine learning models for classification [8]. For each pulse, we compute a set of 38 features (including periodic, temporal domain, spectral domain, Poincare, and beat-to-beat difference features). We used these features to train Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN) models, as described in the original publication [8]. We also included Random Forest (RF), and XGBoost (XGB) models due to their proven effectiveness in handling complex feature interactions and demonstrated success in similar biomedical signal classification tasks [9], [10]. We performed hyperparameter optimization for each model using grid search with 5-fold cross-validation on the training dataset. We selected the hyperparameters that balanced sensitivity and overall accuracy across cross-validation folds. Although this approach requires careful feature selection and extensive hyperparameter tuning, it provides a robust foundation for artifact detection while maintaining computational efficiency and interpretability. A complete description of the feature engineering process and hyperparameter optimization can be found in the supplementary material.

DeepClean: In this approach, we train a self-supervised convolutional variational autoencoder (VAE) to learn a latent representation of ABP pulses and reconstruct the pulses from this latent representation [11]. The VAE minimizes the mean squared error (MSE) between input and reconstructed pulses, training only on non-artifact pulses from the training dataset under the assumption that artifact-laden signals would exhibit greater reconstruction error. We tuned the model architecture and hyperparameters by minimizing reconstruction error on the validation set. During inference, we compute the reconstruction error for each pulse. A pulse is classified as an artifact if its error exceeds the 90th percentile of reconstruction errors from non-artifact training pulses. This unsupervised approach offers the advantage of not requiring artifact annotations for training, but its performance may be limited by the inherent variability in physiological waveform morphology and the challenge of selecting an appropriate reconstruction error threshold. A complete description of the model architecture and hyperparameters can be found in the supplementary material.

SCAE-CNN: This approach combines a stacked convolutional autoencoder (SCAE) and a convolutional neural network (CNN) [12]. The ABP pulses are first normalized and converted to raw 64 × 64 images, as described in the original paper. The SCAE transforms these images into representative images that contain more generalized feature information from the raw signal. The representative images are then used as input to a CNN for classification. We trained the SCAE and CNN end-to-end on the training dataset using a combined loss function that balanced reconstruction (MSE) and classification (cross-entropy) objectives. We tuned the model architecture and hyperparameters using the validation dataset. This architecture provides an efficient approach to learning both reconstructive and discriminative features from the ABP waveforms, though careful tuning of the combined loss function may be needed to balance these competing objectives. A complete description of the model architecture, image conversion process, and hyperparameters can be found in the supplementary material.

Pre-trained Image Classifiers: Inspired by the image-based approach of SCAE-CNN, we explored the use of transfer learning with pre-trained image classification models. Transfer learning allows leveraging generalized models trained on large datasets and applying them to new tasks with smaller datasets [13]. We selected two popular pre-trained image classification models that have been extensively used in the literature for transfer learning: ResNet18 [14] and EfficientNetB0 [15]. Following the approach of Lee et al. [12], we converted ABP pulses to 224 × 224 pixel images. We then fine-tuned these models on our training dataset to classify pulses as artifact or non-artifact using cross-entropy loss. For both models, we replaced the final classification layer with a new layer containing two outputs (artifact and non-artifact). We evaluated two fine-tuning strategies: (1) fine-tuning just the final classification layer while keeping the convolutional layers frozen (CL FT), and (2) fine-tuning the entire model (Full FT). This allowed us to compare the effectiveness of adapting the entire network versus leveraging the pre-trained features directly. We tuned the hyperparameters, including learning rate and batch size, using the validation dataset. This approach leverages the power of large-scale pre-trained models while requiring relatively few labeled examples, though careful consideration must be given to the fine-tuning strategy to optimize performance for the specific task of artifact detection. The complete model architectures and fine-tuning process are detailed in the supplementary material.

Patient-specific Fine-tuning (PS FT): Due to significant between-patient variability in ABP pulse morphology, we explored the use of patient-specific fine-tuning as a proof-of-concept experiment. In real-world settings, we expect that researchers will be able to annotate a small set of pulses for each patient, allowing for the adaptation of pre-trained models to patient-specific ABP characteristics. Applying this fine-tuned model to subsequent data quality annotation can eliminate the need to review many hours of data, even though a handful of patient-specific pulses must still be selected manually. We investigated this approach using the SCAE-CNN, ResNet-18, and EfficientNet-B0 models, which showed promising performance in our initial experiments. We first trained the models on the complete training set as described above, then fine-tuned them using 30 annotated pulses from each of the 5 test patients (150 pulses total) from the validation set. This sample size was chosen to represent a realistic number of pulses that could be manually annotated in clinical practice. While this approach requires additional annotation effort per patient, it offers the potential for improved performance by adapting to individual patient characteristics and monitoring conditions. A full description of the fine-tuning process can be found in the supplementary material.

Ensemble models: During model development, we observed that individual models often struggled to achieve high sensitivity and high specificity simultaneously, instead tending to trade off between these metrics. This presents a challenge for clinical applications, where we need to both identify artifacts reliably (high sensitivity) and avoid incorrectly flagging valid pulses as artifacts (high specificity). To address this limitation, we conducted an experiment to combine models with complementary strengths. We selected the most sensitive model (ResNet-18 with full fine-tuning) and paired it with the model showing the best overall performance across other metrics (EfficientNet-B0 with full fine-tuning). We then combined the probability outputs of these two models using a weighted average to produce a final classification probability. This approach leverages the strengths of both models while maintaining a balance between sensitivity and specificity. The results of this ensemble approach are detailed in the results section.

D. Model Evaluation

We evaluated model performance using accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic curve (AU-ROC). These metrics were chosen to assess both overall classification performance and the models' ability to balance artifact detection (sensitivity) with preservation of valid pulses (specificity). We ran inference on the test dataset using each of the models described above. For models producing probability outputs (all except the heuristic and DeepClean models), we generated ROC curves to visualize the trade-off between sensitivity and specificity across different classification thresholds. To facilitate comparison between models, we report these metrics for each model in Table I. Confusion matrices were created for each model to visualize performance and are shown in the supplementary material.

To assess model performance at clinically relevant operating levels, we analyzed the specificity achieved by each model at fixed sensitivity thresholds of 90%, 95%, and 99%. The motivation for this approach is that artifact is likely to affect a minority of overall available data but have substantial adverse consequences of AI/ML models. Therefore, high sensitivity identifying artifacts may be desirable, even at the expense of substantially lower specificity. We determined these values from the ROC curve data for the models that produce probability outputs.

E. Implementation Details

All implementations were carried out using Python 3.9.19. We utilized sci-kit-learn (v1.5.1) for traditional machine learning model implementations, cross-validation, and computation of performance metrics [16]. We implemented deep learning models using PyTorch (v2.4.0) based on the architectures provided in the original publications [17]. We used Weights & Biases to track training runs and store model checkpoints [18]. We performed signal processing and general data management using SciPy (v1.10.1) and NumPy (v1.20.3) [7], [19]. We utilized the cns-utils library (Moberg Analytics, Philadelphia, PA) to process high-resolution ABP waveforms. Model training and inference were performed on a single NVIDIA RTX 3090 GPU. A detailed description of each method, including specific modifications made to the original algorithms to suit our research objectives, is provided in the supplementary material. The trained weights for our best performing models are available for research use through Moberg Analytics the Open Source Models repository.

III. Results

The 35 patients included in our sample underwent a median of 107.98 [IQR 68.93 to 137.65] hours of continuous ABP monitoring resulting in a total of 3148.39 hours of ABP waveform data, of which a median of 5.83 [IQR 3.03 to 14.06] hours per patient were flagged by both RA annotators as artifact. Agreement between the two RAs in first-round annotation was 60%. Pulse-level agreement between the two PI annotators in second-round annotation was 90%, so in the remaining 10% the final annotation resulted from consensus discussion.

TABLE I: Evaluation metrics (accuracy, sensitivity, specificity, AU-ROC) on the testing dataset for all 17 models. CL FT indicates fine-tuning of only the classification layer, Full FT indicates fine-tuning of the entire model, and PS FT indicates patient-specific fine-tuning. AU-ROC values are not reported for heuristic and DeepClean models as they produce binary classifications rather than probability outputs. The best scores for each metric are highlighted in bold. Metrics range from 0 to 1, with 1 indicating perfect performance.

Model	Accuracy	Sensitivity	Specificity	AU-ROC
Heuristics (Initial)	0.407	0.684	0.019	-
Heuristics (Tuned)	0.721	0.632	0.846	-
SVM	0.714	0.653	0.799	0.778
Decision Tree	0.800	0.723	0.909	0.816
K-Nearest Neighbors	0.668	0.533	0.856	0.721
Random Forest	0.826	0.767	0.909	0.927
XGBoost	0.833	0.775	0.916	0.920
DeepClean	0.734	0.622	0.892	-
SCAE-CNN	0.843	0.779	0.932	0.908
ResNet-18 (CL FT)	0.866	0.810	0.944	0.937
ResNet-18 (Full FT)	0.781	**0.952**	0.542	0.932
EfficientNet-B0 (CL FT)	0.829	0.735	**0.961**	0.928
EfficientNet-B0 (Full FT)	0.878	0.892	0.860	0.932
SCAE-CNN (PS FT)	0.870	0.876	0.861	0.921
ResNet-18 (PS FT)	0.878	0.866	0.895	0.941
EfficientNet-B0 (PS FT)	0.888	0.865	0.919	0.944
Ensemble	**0.890**	0.911	0.860	**0.952**

Our baseline heuristic approach demonstrated the importance of threshold optimization for artifact detection. Through grid search optimization across thresholds (minimum value: 5 - 50mmHg, maximum range: 50 - 300mmHg, minimum range: 5 - 50mmHg, and peak count: 1 - 10 peaks), we significantly improved the model's performance. The optimal parameters identified were a minimum threshold of 40 mmHg, maximum range of 150 mmHg, minimum range of 30 mmHg, and peak count of 1. As shown in Table I, the initial implementation achieved modest performance (ACC: 0.407, SEN: 0.684, SPE: 0.019), but optimization significantly improved overall accuracy to 0.721 while maintaining a better balance between sensitivity (0.632) and specificity (0.846).

Traditional machine learning methods using handcrafted features showed strong performance across all metrics (Table I). Among these approaches, ensemble methods demonstrated superior performance, with Random Forest and XGBoost achieving accuracies of 0.826 and 0.833, respectively. Both models maintained high specificity (>0.90) while achieving reasonable sensitivity (>0.75). The ROC curves (Figure 3A) illustrate the robust performance of these models across different classification thresholds, with AU-ROC values of 0.927 and 0.920 for Random Forest and XGBoost, respectively. The cross-validation results obtained during hyperparameter optimization are reported in the supplementary material.

The DeepClean VAE demonstrated moderate performance (ACC: 0.734, SEN: 0.622, SPE: 0.892; Table I), falling short of other deep learning methods. While the model achieved good specificity, suggesting reliable identification of clean signals, its lower sensitivity indicates challenges in detecting

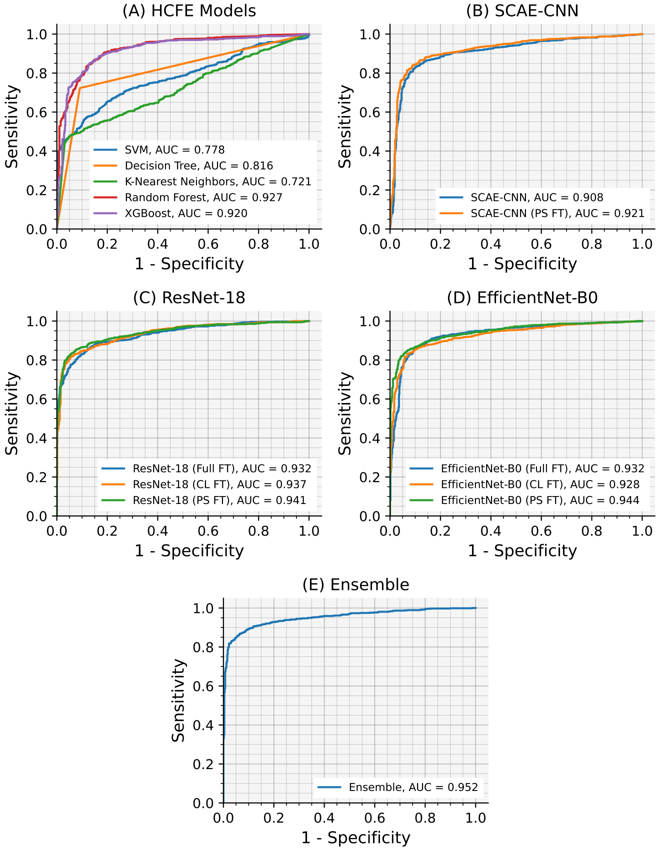


Fig. 3: Receiver Operating Characteristic (ROC) curves comparing model performance. (A) Traditional machine learning models using handcrafted features, including Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), Random Forest, and XGBoost. (B) SCAE-CNN model performance shows the baseline model and patient-specific fine-tuning (PS FT). (C) ResNet-18 performance under different fine-tuning strategies: classification layer only (CL FT), full model fine-tuning (Full FT), and patient-specific fine-tuning (PS FT). (D) EfficientNet-B0 performance under the same three fine-tuning strategies. (E) Performance of the ensemble model combining ResNet-18 and EfficientNet-B0. Area Under the Curve (AUC) values are shown in the legend for each model.

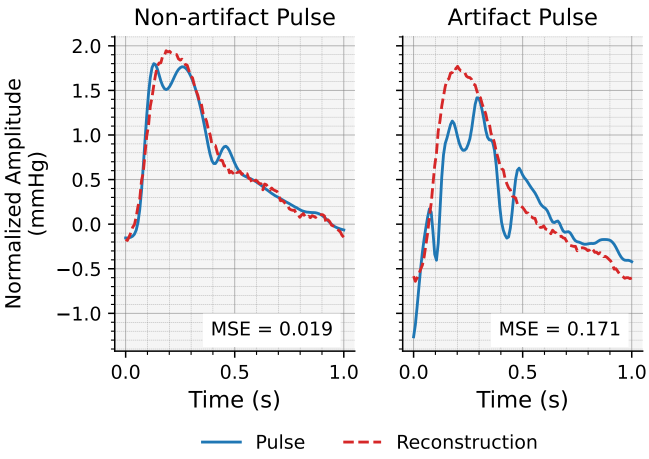


Fig. 4: Example reconstructions from the DeepClean variational autoencoder (VAE) showing normalized arterial blood pressure pulses (solid blue line) and their corresponding reconstructions (dashed red line). The non-artifact pulse (left) shows excellent reconstruction with low mean squared error (MSE = 0.017), while the artifact pulse (right) exhibits poor reconstruction with higher error (MSE = 0.387), demonstrating the model's ability to discriminate between clean and artifactual signals based on reconstruction quality.

subtle artifacts. The model exhibited similar performance when evaluated on the validation dataset. Figure 4 illustrates this behavior through example reconstructions showing a non-artifact pulse (left) that is accurately reconstructed with low error (MSE = 0.017) and an artifact pulse (right) that produces higher reconstruction error (MSE = 0.387) due to its irregular morphology. The model's ability to accurately reconstruct normal ABP morphology while struggling with artifactual pulses explains its bias towards specificity over sensitivity, as it more readily identifies clearly clean signals than subtle artifacts.

The SCAE-CNN architecture achieved strong baseline performance when evaluated on the testing dataset (ACC: 0.843, SEN: 0.779, SPE: 0.932; Table I), surpassing traditional machine learning approaches (Figure 5). The ROC curve in Figure 3B shows the strong discriminative ability of the model. Validation results (Table II) demonstrate consistent performance across datasets, suggesting good generalization capabilities. After performing patient-specific fine-tuning, we reevaluated the model's performance on the test set and observed improvements across accuracy (+3.20%), sensitivity (+12.45%), and AU-ROC (+1.43%), at the cost of some specificity (-1.43%).

Transfer learning approaches using pre-trained image classifiers showed promising results with different fine-tuning strategies (Table I). ResNet-18 achieved the highest sensitivity (0.952) when fully fine-tuned, though at the cost of lower specificity (0.542) and accuracy (0.781). In contrast, EfficientNet-B0 demonstrated more balanced performance with full fine-tuning (ACC: 0.878, SEN: 0.892, SPE: 0.860). Notably, fine-tuning only the classification layers (CL FT) produced competitive results, with both models achieving high specificity (ResNet-18: 0.944, EfficientNet-B0: 0.961) while maintaining moderate sensitivity. The performance of these models on the validation dataset is reported in Table II. We further improved these models through patient-specific fine-tuning of the fully fine-tuned models, with EfficientNet-B0 showing the strongest results (ACC: 0.888, SEN: 0.865, SPE: 0.919) followed closely by ResNet-18 (ACC: 0.878, SEN: 0.866, SPE: 0.895). The ROC curves (Figure 3C & 3D) illustrate the strong discriminative ability of these models across all fine-tuning strategies, with AU-ROC values consistently above 0.925.

Given the complementary strengths observed in the transfer learning models, we developed an ensemble approach combining the fully fine-tuned ResNet-18 and EfficientNet-B0 models. We implemented a simple weighted averaging strategy, combining the probability outputs from both models

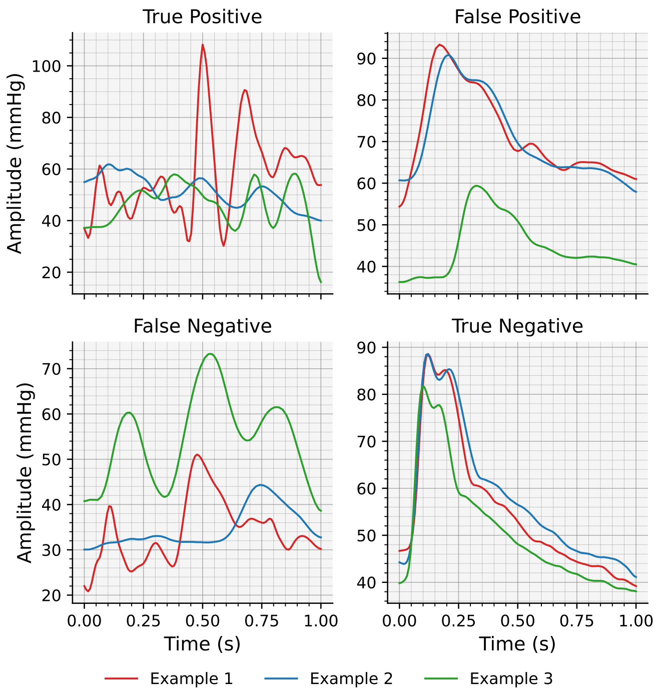


Fig. 5: Representative arterial blood pressure pulses illustrating the four possible classification outcomes arranged in a confusion matrix format. Each panel shows three example pulses (red, blue, and green dashed) for each category. True Positives (top left) show clearly artifactual waveforms with non-physiologic morphology. False Positives (top right) demonstrate normal physiologic waveforms that were incorrectly classified as artifacts. False Negatives (bottom left) display artifacts that were missed by the models. Note the morphology of these False Negatives may reflect good quality signal that was improperly segmented to include partial or multiple pulse contours. True Negatives (bottom right) show clean physiologic waveforms that were correctly identified as non-artifacts. Time is shown in seconds, and amplitude is in mmHg.

TABLE II: :Evaluation metrics (accuracy, sensitivity, specificity, AU-ROC) on the validation set for deep learning models that were tuned using the validation set. AU-ROC values are not reported for DeepClean as it produces binary classifications rather than probability outputs. The best scores for each metric are highlighted in bold. Metrics range from 0 to 1, with 1 indicating perfect performance.

Model	Accuracy	Sensitivity	Specificity	AU-ROC
DeepClean	0.713	0.619	0.887	-
SCAE-CNN	0.840	0.804	0.906	0.918
ResNet-18 (CL FT)	0.867	0.825	**0.943**	0.958
ResNet-18 (Full FT)	0.847	**1.000**	0.566	**0.970**
EfficientNet-B0 (CL FT)	0.833	0.773	**0.943**	0.956
EfficientNet-B0 (Full FT)	**0.913**	0.959	0.830	0.928

TABLE III: Specificity scores for each of the models at different thresholds of sensitivity (90%, 95%, and 99%). Only models that produce probability outputs are included. Specificity values range from 0 to 1, with higher values indicating better performance at maintaining true negative classifications while achieving the target sensitivity. The best specificity scores at each sensitivity level are highlighted in bold.

Model	90% Sensitivity	95% Sensitivity	99% Sensitivity
SVM	0.269	0.199	0.003
Decision Tree	0.000	0.000	0.000
K-Nearest Neighbors	0.209	0.000	0.000
Random Forest	0.814	**0.662**	0.201
XGBoost	0.804	0.655	0.100
SCAE-CNN	0.758	0.475	0.115
ResNet-18 (CL FT)	0.770	0.625	0.216
ResNet-18 (Full FT)	0.758	0.562	**0.274**
EfficientNet-B0 (CL FT)	0.772	0.562	0.160
EfficientNet-B0 (Full FT)	0.845	0.625	0.196
SCAE-CNN (PS FT)	0.787	0.532	0.103
ResNet-18 (PS FT)	0.814	0.606	0.191
EfficientNet-B0 (PS FT)	0.840	0.613	0.206
Ensemble	**0.883**	0.650	0.218

with weights of 0.40 and 0.60 for ResNet-18 and EfficientNet-B0, respectively. The final classification was determined using a threshold of 0.9 on the weighted probability for the artifact class. This ensemble approach achieved the best overall performance in our evaluation (ACC: 0.890, SEN: 0.911, SPE: 0.860; Table I), effectively balancing the high sensitivity of ResNet-18 with the more balanced performance of EfficientNet-B0. The ROC curve for the ensemble model (Figure 3E) demonstrates its superior discriminative ability, achieving the highest AU-ROC (0.952) among all approaches tested.

To evaluate model performance at clinically relevant operating points, we analyzed the specificity achieved by each model when maintaining a high sensitivity threshold (Table III). At 90% sensitivity, the ensemble model maintained the highest specificity (0.883), followed by EfficientNet-B0 with both full fine-tuning (0.845) and patient-specific fine-tuning (0.840). When increasing the sensitivity threshold to 95%, Random Forest achieved the best specificity (0.662), with XGBoost (0.655) and the ensemble model (0.650) performing similarly. At the most stringent threshold of 99% sensitivity, all models showed substantial degradation in specificity, with the fully fine-tuned ResNet-18 achieving the highest specificity (0.274). Notably, simpler models like Decision Tree and K-Nearest-Neighbors did not maintain any specificity at these higher sensitivity thresholds (> 95%).

IV. Discussion

This study found varying performance across different approaches to artifact detection in ABP waveforms from ICU patients. While traditional machine learning methods achieved good overall accuracy, deep learning approaches - particularly ensemble methods combining ResNet-18 and EfficientNet-B0 - demonstrated superior performance with accuracy up to 89.0%. This is the first study to systematically evaluate and compare multiple artifact detection strategies on a common dataset of ICU-derived waveforms, providing important insights into the relative strengths and limitations of each approach.

Preparing high-resolution clinical data for AI/ML is complex, time-consuming and requires significant human effort. Beyond the costs associated with data preparation, many advanced analytical tools such as neural networks are data-sensitive during training, making them fragile in performance and interpretation of results [20], [21]. In plain language, this means that arbitrary differences during data preparation (for example, subjective judgments about waveform quality or selection of waveform segments for model inclusion) can result in substantially different models after training. Taken together, these considerations threaten the feasibility, rigor and reproducibility of AI/ML models based on real-world high-resolution ICU-derived data.

Our results highlight the limitations of manual review of waveform data for the presence of non-physiologic artifact. Despite considerable training and prior clinical experience interpreting these data, agreement between raters in our initial round of annotation was only 60%. Agreement on our second round review increased to 90%. This may be due in part to task-specific (i.e., review of individual pulses may be a simpler cognitive task than review of continuous, unsegmented waveforms) or because the first-round review meant second-round annotation was of signal that had previously been reviewed and determined to be likely artifact or artifact-free (i.e., fewer ambiguous data may have been included). Regardless, from the perspective of data curation, initial annotation quality was unlikely to reflect an adequate "ground truth" against which data could be selected for subsequent model inclusion.

We tested a variety of analytical tools for ABP waveform annotation, selected after a scoping review of the literature. Most models had excellent discrimination overall. As a second, domain-relevant performance metric, we evaluated specificity for artifact detection at high sensitivity. The motivation for this metric was that ML from ICU-acquired data often analyze relatively short data segments (e.g., 5 minutes of data selected hourly). Since it is unlikely for all or most of any given hour of data to be artifact, we accept some cost in specificity (i.e., acceptable data are falsely labeled as artifact) to ensure all artifacts are correctly identified. We acknowledge this decision may affect the performance of subsequent models. For example, at a highly sensitive threshold non-artifact data measured during periods of severe physiologic abnormality (e.g., severe hypotension) may be labeled as artifact, thus preventing the inclusion of the clinically important epochs in subsequent models. In future work, thresholds for binarizing the output of models trained to predict artifact could be tuned jointly with subsequent predictive algorithms.

We note several design choices in our approach that affect interpretation of the results. First, our choice to use balanced classes in the training dataset, while not reflecting the natural distribution of artifacts in clinical practice, was motivated by the need to ensure robust learning of both artifact and non-artifact features. This approach is common in ML when dealing with imbalanced datasets [22], though it can potentially bias model expectations. We addressed this limitation by evaluating model performance on a test dataset that maintained the natural distribution of artifacts, providing a more realistic assessment of real-world performance. Second, in our development of heuristic models we selected initial thresholds for decision rules based on clinical intuition. However, because these initial heuristics performed poorly we refined these thresholds by searching for better performing cutoffs with improvement in overall model performance. This improvement validates the clinical intuition underlying our heuristic rules while highlighting the need for more sophisticated approaches. Finally, in several models we performed patient-specific fine tuning, in which we presented the trained model with a small number of pulses from test set cases. In essence, this step served to "teach" the model the normal pulse morphology for each test set case, thereby improving model performance. A similar strategy of referencing ML models to patients' individual baselines has proven effective in other ICU-based predictive tasks that utilize waveform data [23]. In practice, this approach would allow a workflow whereby a small number of normal pulses (e.g., ten pulses per patient) could be manually selected with minimal effort for each future patient, after which a fine-tuned model could automatically annotate the remainder of those patients' recordings.

Our work has several important limitations to consider. The premise of our analyses was that automated artifact detection could improve the rigor and reproducibility of subsequent modeling. However, while we demonstrated the feasibility of artifact detection we did not quantify the impact of this approach on subsequent models. This is an area of ongoing investigation by our group and others. Second, while we evaluated models at specific sensitivity thresholds, the optimal balance between sensitivity and specificity likely varies across different research applications. Rather than favoring a single approach, researchers should carefully consider their specific needs when selecting detection thresholds and methods. Finally, a fundamental question remains regarding the degree of artifact detection necessary for developing robust ML models from continuous waveform data. It is possible that less stringent approaches to artifact detection may be sufficient, or that some applications may be resilient to artifact contamination. Future work should investigate the relationship between artifact detection stringency and downstream model performance to establish evidence-based guidelines for waveform data preparation.

V. Conclusion

This study demonstrates the feasibility of automated artifact detection in arterial blood pressure waveforms from ICU patients, with ensemble deep learning methods achieving up to 89.0% accuracy. Our systematic comparison of multiple approaches reveals that while traditional machine learning methods can achieve good performance, deep learning methods—particularly those leveraging transfer learning and patient-specific fine-tuning—offer superior results. The trade-offs between sensitivity and specificity highlight the importance of selecting appropriate operating thresholds based on specific clinical or research needs.

Future work should focus on quantifying the impact of automated artifact detection on downstream machine learning tasks and establishing evidence-based guidelines for waveform data preparation. Additionally, investigating the generalizability of these methods across different patient populations and monitoring conditions will be crucial for clinical implementation. Our findings suggest that a practical approach combining pre-trained models with minimal patient-specific annotation could provide an efficient workflow for automated quality assessment of physiological waveforms in clinical research.

References

[1] M. P. Mulder, M. Broomé, D. W. Donker, and B. E. Westerhof, "Distinct morphologies of arterial waveforms reveal preload-, contractility-, and afterload-deficient hemodynamic instability: An in silico simulation study," *Physiological Reports*, vol. 10, no. 7, e15242, Apr. 2022. DOI: 10.14814/phy2.15242. (visited on 10/25/2024).

[2] D. M. Mirvis, A. S. Berson, A. L. Goldberger, L. S. Green, J. J. Heger, T. Hinohara, J. Insel, M. W. Krucoff, A. Moncrief, and R. H. Selvester, "Instrumentation and practice standards for electrocardiographic monitoring in special care units. a report for health professionals by a task force of the council on clinical cardiology, american heart association," *Circulation*, vol. 79, no. 2, pp. 464–471, Feb. 1989, ISSN: 0009-7322. DOI: 10.1161/01.cir.79.2.464.

[3] H. Kim, S.-B. Lee, Y. Son, M. Czosnyka, and D.-J. Kim, "Hemodynamic instability and cardiovascular events after traumatic brain injury predict outcome after artifact removal with deep belief network analysis," *Journal of Neurosurgical Anesthesiology*, vol. 30, no. 4, pp. 347–353, Oct. 2018, ISSN: 1537-1921. DOI: 10.1097/ANA.0000000000000462.

[4] M. Blount, C. McGregor, A. James, D. Sow, R. Kamaleswaran, S. Tuuha, J. Percival, and N. Percival, "On the integration of an artifact system and a real-time healthcare analytics system," in *Proceedings of the 1st ACM International Health Informatics Symposium*, ser. IHI '10, New York, NY, USA: Association for Computing Machinery, Nov. 2010, pp. 647–655, ISBN: 978-1-4503-0030-8. DOI: 10.1145/1882992.1883094. (visited on 10/23/2024).

[5] J. Elmer, Z. He, T. May, E. Osborn, R. Moberg, S. Kemp, J. Stover, E. Moyer, R. G. Geocadin, K. G. Hirsch, and PRECICECAP Study Team, "Precision care in cardiac arrest: Icecap (precicecap) study protocol and informatics approach," *Neurocritical Care*, vol. 37, no. Suppl 2, pp. 237–247, Aug. 2022, ISSN: 1556-0961. DOI: 10.1007/s12028-022-01464-9.

[6] W. J. Meurer, F. F. Schmitzberger, S. Yeatts, V. Ramakrishnan, B. Abella, T. Aufderheide, W. Barsan, J. Benoit, S. Berry, J. Black, N. Bozeman, K. Broglio, J. Brown, K. Brown, N. Carlozzi, A. Caveney, S.-M. Cho, H. Chung-Esaki, R. Clevenger, R. Conwit, R. Cooper, V. Crudo, M. Daya, D. Harney, C. Hsu, N. J. Johnson, I. Khan, S. Khosla, P. Kline, A. Kratz, P. Kudenchuk, R. J. Lewis, C. Madiyal, S. Meyer, J. Mosier, M. Mouammar, M. Neth, B. O'Neil, J. Paxton, S. Perez, S. Perman, C. Sozener, M. Speers, A. Spiteri, V. Stevenson, K. Sunthankar, J. Tonna, S. Youngquist, R. Geocadin, and R. Silbergleit, "Influence of cooling duration on efficacy in cardiac arrest patients (icecap): Study protocol for a multicenter, randomized, adaptive allocation clinical trial to identify the optimal duration of induced hypothermia for neuroprotection in comatose, adult survivors of after out-of-hospital cardiac arrest," *Trials*, vol. 25, p. 502, Jul. 2024, ISSN: 1745-6215. DOI: 10.1186/s13063-024-08280-w. (visited on 10/16/2024).

[7] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, "Scipy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2. (visited on 10/11/2024).

[8] T. Pereira, K. Gadhoumi, M. Ma, X. Liu, R. Xiao, R. A. Colorado, K. J. Keenan, K. Meisel, and X. Hu, "A supervised approach to robust photoplethysmography quality assessment," *IEEE journal of biomedical and health informatics*, vol. 24, no. 3, pp. 649–657, Mar. 2020, ISSN: 2168-2208. DOI: 10.1109/JBHI.2019.2909065.

[9] T. Chen and C. Guestrin, *Xgboost: A scalable tree boosting system*, Jun. 2016. DOI: 10.48550/arXiv.1603.02754. arXiv: 1603.02754. (visited on 10/24/2024).

[10] A. Tiwari and A. Chaturvedi, "A multiclass eeg signal classification model using spatial feature extraction and xgboost algorithm," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4169–4175, Nov. 2019. DOI: 10.1109/IROS40897.2019.8967868. (visited on 10/24/2024).

[11] T. Edinburgh, P. Smielewski, M. Czosnyka, S. J. Eglen, and A. Ercole, *Deepclean – self-supervised artefact rejection for intensive care waveform data using deep generative learning*, Jan. 2020. DOI: 10.48550/arXiv.1908.03129. arXiv: 1908.03129 [cs, eess, stat]. (visited on 10/07/2024).

[12] S.-B. Lee, H. Kim, Y.-T. Kim, F. A. Zeiler, P. Smielewski, M. Czosnyka, and D.-J. Kim, "Artifact removal from neurophysiological signals: Impact on intracranial and arterial pressure monitoring in traumatic brain injury," *Journal of Neurosurgery*, vol. 132, no. 6, pp. 1952–1960, Jun. 2020, ISSN: 1933-0693. DOI: 10.3171/2019.2.JNS182260.

[13] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, *A comprehensive survey on transfer learning*, Jun. 2020. arXiv: 1911.02685. (visited on 10/24/2024).

[14] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, Dec. 2015. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385. (visited on 10/11/2024).

[15] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, Sep. 2020. DOI: 10.48550/arXiv.1905.11946. arXiv: 1905.11946. (visited on 10/11/2024).

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, ISSN: 1533-7928. (visited on 10/11/2024).

[17] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *Pytorch: An imperative style, high-performance deep learning library*, Dec. 2019. DOI: 10.48550/arXiv.1912.01703. arXiv: 1912.01703. (visited on 10/11/2024).

[18] L. Biewald, *Experiment tracking with weights and biases*, 2020.

[19] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2. (visited on 10/11/2024).

[20] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, *Sensitivity and generalization in neural networks: An empirical study*, Jun. 2018. DOI: 10.48550/arXiv.1802.08760. arXiv: 1802.08760. (visited on 11/05/2024).

[21] A. Ghorbani, A. Abid, and J. Zou, *Interpretation of neural networks is fragile*, Nov. 2018. DOI: 10.48550/arXiv.1710.10547. arXiv: 1710.10547. (visited on 11/05/2024).

[22] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, Mar. 2019, ISSN: 2196-1115. DOI: 10.1186/s40537-019-0192-5. (visited on 11/05/2024).

[23] M. R. Pinsky, A. Wertz, G. Clermont, and A. Dubrawski, "Parsimony of hemodynamic monitoring data sufficient for the detection of hemorrhage," *Anesthesia and Analgesia*, vol. 130, no. 5, pp. 1176–1187, May 2020, ISSN: 1526-7598. DOI: 10.1213/ANE.0000000000004564.