IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 May 2024, accepted 6 June 2024, date of publication 10 June 2024, date of current version 17 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3411774

RESEARCH ARTICLE

Label Propagation Techniques for Artifact Detection in Imbalanced Classes Using Photoplethysmogram Signals

CLARA MACABIAU[1], THANH-DUNG LE[1], (Member, IEEE), KÉVIN ALBERT[2], MANA SHAHRIARI[2], PHILIPPE JOUVET[2], AND RITA NOUMEIR[1], (Member, IEEE)
[1]Biomedical Information Processing Laboratory, École de Technologie Supérieure, Montréal, QC H3C 1K3, Canada
[2]CHU Sainte-Justine Research Center, CHU Sainte-Justine Hospital, Université de Montréal, Montréal, QC H3T 1J4, Canada

Corresponding author: Clara Macabiau (clara.macabiau.1@ens.etsmtl.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC), and in part by the Fonds de recherche en santé du Québec (FRQS).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Review Ethic Board of Sainte-Justine Hospital.

ABSTRACT This study aimed to investigate the application of label propagation techniques to propagate labels among photoplethysmogram (PPG) signals, particularly in imbalanced class scenarios and limited data availability scenarios, where clean PPG samples are significantly outnumbered by artifact-contaminated samples. We investigated a dataset comprising PPG recordings from 1571 patients, wherein approximately 82% of the samples were identified as clean, while the remaining 18% were contaminated by artifacts. Our research compares the performance of supervised classifiers, such as conventional classifiers and neural networks (Multi-Layer Perceptron (MLP), Transformers, Fully Convolutional Network (FCN)), with the semi-supervised Label Propagation (LP) algorithm for artifact classification in PPG signals. The results indicate that the LP algorithm achieves a precision of 91%, a recall of 90%, and an F1 score of 90% for the ''artifacts'' class, showcasing its effectiveness in annotating a medical dataset, even in cases where clean samples are rare. Although the K-Nearest Neighbors (KNN) supervised model demonstrated good results with a precision of 89%, a recall of 95%, and an F1 score of 92%, the semi-supervised algorithm excels in artifact detection. In the case of imbalanced and limited pediatric intensive care environment data, the semi-supervised LP algorithm is promising for artifact detection in PPG signals. The results of this study are important for improving the accuracy of PPG-based health monitoring, particularly in situations in which motion artifacts pose challenges to data interpretation.

INDEX TERMS Motion artifacts, imbalanced classes, label propagation algorithm, machine learning classifiers, photoplethysmogram (PPG) signals.

I. INTRODUCTION

Machine learning, a sub-field of artificial intelligence [1], has emerged as a transformative technology in various domains, including healthcare. With its ability to analyze large amounts of data [2], it has the potential to improve healthcare outcomes, help doctors make better decisions [3], and revolutionize medical research with models that aim to predict injuries [4], detect heart disease earlier [5] and mortality [6]. Additionally, machine learning algorithms can contribute to drug discovery and development, optimizing drug efficacy and predicting potential adverse reactions [7]. Machine learning can extract all the necessary information from various types of healthcare data, such as electronic medical records [8], medical images, and physiological signals.

The associate editor coordinating the review of this manuscript and approving it for publication was Ganesh Naik.

Despite its potential, the integration of machine learning into healthcare comes with challenges and considerations. Privacy and ethical implications must be taken into account [9]. The data acquired must respect patient privacy and confidentiality and also require standardization and centralized collection for ease of management and consistency, ensuring harmonization [10]. One major concern is the availability of high-quality data for training and testing these algorithms [11]. To evaluate the performance of the algorithms implemented, it is necessary to have access to a ground truth. Accessing ground truth for evaluating algorithms is challenging, often requiring expert input and large, complete datasets, particularly due to class imbalances in medical data. This further complicates model training, necessitating rebalancing while preserving medical value, with erroneous, missing, or imprecise data exacerbated by artifacts from patient motion or clinical interventions posing additional obstacles to accurate predictions.

During a patient's stay in the hospital, it is important to constantly monitor vital signs. One of these vital signals is the PPG signal, which is frequently captured during different types of movements, introducing motion noise and interfering with the accuracy of the signals. This noise is irregular and causes high-amplitude fluctuations within the PPG signals [12]. Motion artifacts can result in the pulse oximeter either misinterpreting movement as the actual signal or masking the true signal with unwanted interference, leading to incorrect readings, false alarms, and missed important alarms [13]. The main objective of this work is to detect motion artifacts in PPG signals obtained from the Pediatric Intensive Care Unit (PICU) database of the CHU Sainte-Justine Hospital (CHUSJ). The cleaned PPG signals will be used to construct clinical decision systems (CDSS) at CHUSJ's PICU. Specifically, annotated signals will be used in screening and identifying various health-related concerns in children. For example, changes in blood pressure in children are significant indicators for identifying patients who require immediate care and admission to the PICU. Invasive methods, like catheter insertion for continuous blood pressure monitoring, offer precise real-time data but come with significant risks such as bleeding and infection [14]. On the other hand, conventional cuff-based measurements, though less invasive, provide only intermittent readings and may not capture sudden clinical changes effectively. Therefore, predicting blood pressure from PPG waveforms has emerged as a successful approach [15] for comprehensive CDSS applications.

This study contributes to the field in three main ways. Firstly, we compare resampling methods commonly used in medical data analysis to address the imbalance between clean PPG samples and artifact-contaminated ones. Secondly, we validate the efficacy of the LP algorithm for motion artifact detection within PPG signals, offering insights into its performance in scenarios with limited labeled data. Lastly, we present a detailed performance comparison between traditional supervised algorithms and the semi-supervised LP approach, highlighting the advantages of leveraging unlabeled data in artifact classification tasks.

II. RELATED WORK

Numerous methods have already been developed to detect motion artifacts in PPG signals. First, the traditional methods are easy to implement. In [16], the authors used statistical analysis to compare the values of three statistics calculated for each pulse of the PPG signal to determine which pulses are noisy. This method will be used in the labeling step for the rest of the project. Adaptive filtering is another method of artifact detection [17]. The adaptive filter uses an algorithm that continuously updates its coefficients to obtain an error signal as close as possible to the original PPG signal. Both approaches have the advantage of being easy to implement but are notably sensitive to empirical thresholds. Among other popular methods, the wavelet transform uses cascaded high-pass and low-pass filters to obtain the desired signal decomposition. Once the signal has been decomposed, the coefficients are analyzed to identify any artifacts [18]. Empirical mode decomposition, like the wavelet transform, is a time-frequency analysis of the signal [19]. When these modes are obtained, the objective is to calculate the instantaneous frequency for each mode to detect modes that have a frequency close to the harmonics of a PPG signal and modes characteristic of motion artifacts. So, these methods have the advantage of being fast and simple, which is useful, but when used alone, they have limited adaptability and may not work as well with complex movements or unexpected scenarios. A summary table with the review activity is presented in Table 1. Therefore, we decided to use a combination of signal processing algorithms for the preprocessing part and machine learning models.

Regarding machine learning models, in [20], the authors explore using semi-supervised models to classify temporal data. These models are based on a graphical approach like the LP algorithm. The algorithm's results are evaluated on different datasets of varying lengths, including ECG (electrocardiogram) signal data. The results show that semi-supervised models are accurate for classifying time series data. However, these algorithms have not been applied to artifact detection. Semi-supervised learning is widely used as a classification algorithm in cases where not all data is annotated. Active learning is also a powerful semi-supervised classification method that has proven effective for temporal data [21]. In our scenario, the LP algorithm is effective because the availability of labeled data is limited, and there is a large amount of unlabeled data [22]. Considering this information, the LP algorithm was implemented for this project.

In the first part of the project, the LP algorithm is used for data annotation. First, an expert annotated a small proportion of data, and a statistical analysis algorithm was

TABLE 1. Summary table of the literature review.

Authors	Title	Year	Source	Findings
Q. Wang et al.	Artifact reduction based on empirical mode decomposition (EMD) in photoplethysmography for pulse rate detection	2010	2010 Annual International Conference of the IEEE Engineering in Medicine and Biology	Empirical Mode Decomposition and Hilbert transform combined give good results for the decomposition of PPG signals and the reduction of motion artifacts
G. Joseph et al.	Photoplethysmogram (PPG) signal analysis and wavelet denoising	2014	2014 Annual International Conference on Emerging Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD)	Unwanted PPG signal interference is successfully removed using wavelet transform while preserving the signal information
Z. Xu and K. Funaya	Time series analysis with graph-based semi-supervised learning	2015	2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)	The use of a new probabilistic semi-supervised method combining different graph constructions and distance techniques gives better results on different types of real data
S. Hanyu and C. Xiaohui	Motion artifact detection and reduction in PPG signals based on statistics analysis	2017	29th Chinese Control And Decision Conference (CCDC)	The use of statistical thresholds on corrupted PPG segments correlated with high-quality segments effectively removes motion artifacts
C.-C. Wu et al.	An implementation of motion artifacts elimination for PPG signal processing based on recursive least squares adaptive filter	2017	2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)	Adaptive approach to remove motion artifacts, using DC Remover method and Recursive Least Squares adaptive filter
Y. Shin et al.	Coherence-based label propagation over time series for accelerated active learning	2021	International Conference on Learning Representations	A new active learning method for time series, called TCLP, improves classification accuracy when a very small number of data points are already labeled
D. Bünger et al.	An empirical study of graph-based approaches for semi-supervised time series classification	2022	Frontiers in Applied Mathematics and Statistics, vol. 7	Comparison of different distance measures in the implementation of graph-based models, including some semi-supervised models, in classifying binary time series datasets

used to validate the annotations. Then, the LP annotates all data using only a small proportion of previously annotated data. Our medical data are unbalanced, with around 80% of pulses free of artifacts and only 20% with artifacts. This means that to have an accurate labeling algorithm, a rebalancing of the classes in the training part needs to be done. Several methods are available for this: oversampling, undersampling, and both oversampling and undersampling. It must be remembered that medical data is being worked with, so sampling methods must make medical sense, whether by randomly duplicating data or by removing it. Medical data involves intricate relationships among data elements, such as patient demographics, medical history, symptoms, diagnoses, treatments, and outcomes [23].

Another aim of this project is to compare classifiers to the LP algorithm, used as a classifier, to accurately detect artifacts. In health care, classifiers are a real help in decision-making [24]. The spectrum of classifiers is very wide: from traditional classifiers like KNN, Support Vector Machine (SVM), Decision Tree (DT), and Naive Bayes classifier (NB) [25], to classifiers using neural networks, such as MLP or Transformers. A comparison of the results of each type of classifier with the semi-supervised LP algorithm will be presented. The effectiveness of these two streams is analyzed by the experimental results (in section V) from the comparative analysis of semi-supervised LP (with KNN kernel) and fully-supervised learning, including conventional machine learning classifiers (KNN, Support Vector Classification (SVC), DT, Random Forest (RF), GaussianNB, MultinominalNB, and Logistic Regression (LR)), MLP, and Transformers. Then, the best classification method will be presented, followed by a conclusion on artifact detection.

The paper is structured as follows. In section III, data characteristics, preprocessing, methodology, labeling, and classification are introduced. In section IV, the implementation of experiments is presented. Section V is used to evaluate the results with different metrics and present a comparison of experimental result tables. In section VII, the results are interpreted, and the limitations are discussed.

III. MATERIAL AND METHODS

This study was conducted following ethical approval from the research ethics board at CHUSJ (protocol number 2023-4556, accepted January 18, 2023). The detailed workflow of the various work stages is shown in Fig. 1. Specifically, the workflow of the proposed method for detecting motion artifacts in PPG signals begins with the input of a 30-second PPG signal. This signal undergoes data preprocessing, which includes filtering, segmentation, resampling, and normalization to prepare the data for analysis. Next, a labeling and classification step is performed using a label propagation algorithm to identify and classify segments of the signal. Finally, the process outputs artifact detection, highlighting the portions of the signal affected by motion artifacts.

A. DATA COLLECTION

This project aims to detect motion artifacts in PPG signals. The eligible study population includes all children aged 0 to 18 years, admitted between September 2018 and July 2022 inclusive, for whom electrocardiogram (ECG), PPG, and arterial blood pressure (ABP) waveform records are available. In this population, specific exclusion criteria have been established to avoid bias. Data collected beyond the fourth day of hospital stay will be disregarded to prevent potential bias from a few patients who may have prolonged stays with arterial lines. Patients on extracorporeal membrane oxygenation (ECMO) treatment will also be excluded from the analysis. Furthermore, if a patient is readmitted to the PICU multiple times, only data from the first stay will be analyzed.

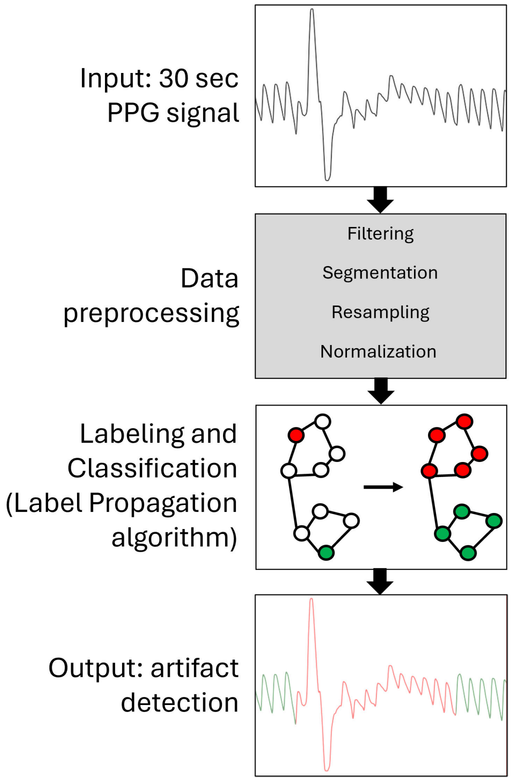


FIGURE 1. Workflow of the proposed method for detecting motion artifacts in PPG signals.

A PPG signal is recorded using a sensor called the pulse oximeter. This device is placed on a patient's skin, for children, on a fingertip or earlobe. A PPG sensor emits light into the skin, partially absorbed by the blood vessels. Changes in blood flow during the cardiac cycle cause variations in light absorption. The sensor detects the reflected light, measuring its intensity modulated by blood volume changes. This varying intensity is converted into an electrical signal, creating the PPG waveform. Blood pressure signals are recorded using an invasive and continuous method, i.e., the catheter, and a non-invasive and discontinuous method, i.e., the blood pressure cuff. ECG is continuously recorded by placing electrodes on the patient's chest. The Sainte-Justine University Hospital PICU utilizes a high-resolution research database (HRDB) [26], [27] that has been approved by the ethical committee. The HRDB links biomedical signals extracted from the different devices, displayed through patient monitors, to the electronic patient record continuously throughout their stay in the unit [28].

Between 2018 and 2022, 1571 patients met the inclusion criteria. For each patient, four physiological signals were extracted: ECG, PPG, blood pressure from the catheter, and blood pressure from the cuff. Each signal was extracted over 96 hours (4 days). Signal values are grouped together in a table with the date and time of acquisition. For the PPG signal, 640 values are acquired every 5 seconds, corresponding to a sampling frequency of 128 Hz. For blood pressure and ECG signals, 2560 values are acquired every 5 seconds, with a sampling frequency of 512 Hz. For the duration of the extraction, a fixed 30-second window of PPG signals will be used for further processing.

B. PREPROCESSING

The raw PPG signal is preprocessed to increase its quality, remove unwanted noise, and make it more suitable for subsequent processing steps [29]. The different steps are described below:

1) **Filtering**: each signal window is filtered using a band-pass Butterworth filter; the cut-off frequencies are 0.5 and 5 Hz, corresponding to a heart rate between 30 and 300 bpm. A forward-backward filtering is used to avoid phase distortions. The objective is to remove baseline wander and high-frequency noise.
2) **Pulse segmentation**: a function to find all local minima by comparing samples is used. The aim is to divide the preprocessed PPG signal into smaller segments or windows to detect the artifacts present for each pulse. In our case, a segment is a pulse. The size of each segment may vary depending on the characteristics of the PPG signal and the specific application of the signal pulses. A pulse is considered to lie between two minima.
3) **Resampling**: the duration of a cardiac cycle for children is between 0.3 and 1 second. A pulse represents a cardiac cycle. Therefore, not all pulses have the same number of samples. Each pulse is uniformly oversampled in time to contain 256 samples, corresponding to a heart cycle of 1s. A linear interpolation

function [30] is used to create the missing points for each pulse. Linear interpolation is favored for signals due to its simplicity, computational efficiency, and ability to estimate values between known data points. It maintains signal continuity and linearity, making it suitable for signals with relatively smooth and linear variations.

4) **Normalization**: the data are normalized to have a unit variance and zero mean. This normalization ensures that all features or variables in the data have the same scale, preventing certain features from dominating the learning process simply because they have larger numerical values.

5) **Data transformation**: each PPG pulse, essentially a waveform representing blood volume changes over time, can be represented as a data point in a column containing 256 values. These values are equally spaced points obtained using step 3 of the preprocessing. At the end of preprocessing, a vector of 256 points is obtained, representing a pulse of the PPG signal. The number of vectors depends on the number of pulses. This method allows us to work with PPG data in a structured manner suitable for various applications, from statistical analysis to machine learning.

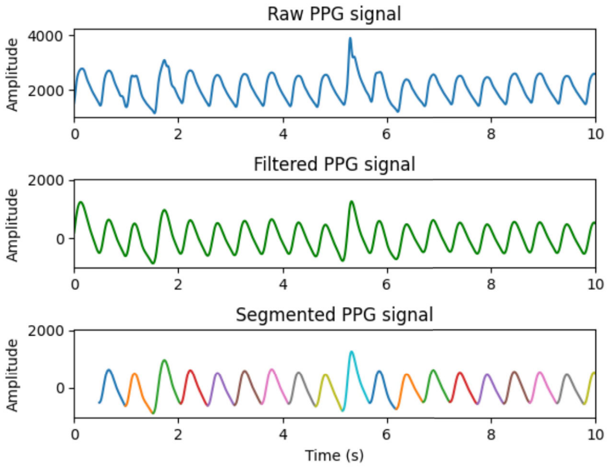


FIGURE 2. Example of a 10s segment of a 30s raw PPG signal in the top image, filtered signal in the middle image, and segmented signal in the bottom image.

Fig. 2 shows the first 10 seconds of a raw PPG signal, when the signal has been filtered, and finally when the pulses have been segmented. The effect of the bandpass filter can be seen in the second figure. The filter has smoothed the signal by removing the extreme frequency components. The signal waveform is preserved, and the filter does not introduce resonances or significant ripples in the desired frequency range. Note that the first pulse has not been segmented. This is because the function could not detect the two minima that make up a pulse and, therefore, could not segment it. The signal does not start at the first low point of the pulse.

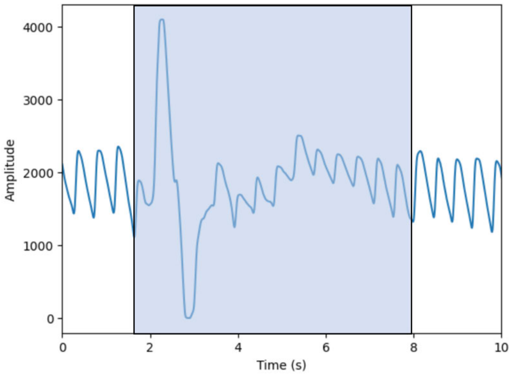


FIGURE 3. Example of a 10s segment of a 30s raw PPG signal. Inside the blue box are all the pulses containing motion artifacts.

C. DATASET ANNOTATION

First, after preprocessing the data, the aim is to build a ground truth for future evaluation of the classification algorithms. To this end, one human expert annotated 10% of the database. Annotation is visual, comparing pulses with each other and binary classifying each pulse as good or as containing artifacts. To avoid involving another human expert, in consideration of time and specialist resources, we implemented an automated algorithm to handle additional annotations. This algorithm, acting as a surrogate expert, was developed to recheck the entirety of the 10% annotated data by the human expert, identifying similarities in the process. Employing a statistical approach, the algorithm determines if the values of a given pulse lie within standard parameters or deviate from the norm. We subsequently cross-validated the algorithm's annotations against those from the professional expert to determine the algorithm's accuracy. It was decided to annotate a maximum of 10% of the database and then use the LP algorithm to annotate the rest of the data.

1) EXPERT LABELING

PPG signals already segmented are presented to the expert. By analyzing each pulse, the expert classifies each pulse as artifact or artifact-free. A pulse is defined as artifact-free if its morphology is typical, i.e., if its characteristics - amplitude, width, shape - are the same as those of adjacent signals. A pulse is defined with artifacts if its characteristics differ from those of adjacent pulses (see fig. 3). To recheck the annotations of one expert, an algorithm that acts as a second expert was set up, allowing all impulses to be reannotated to see similarities. This algorithm uses a statistical approach to assess whether the statistical values of a pulse are normal or outside the norm.

2) STATISTICAL ANALYSIS

For each cardiac cycle, which corresponds to a pulse, if the waveform is similar, statistics such as skewness, standard deviation (std), and kurtosis are approximately constant for each cycle. It is, therefore, possible to detect motion artifacts by using the value of these statistics to differentiate a pulse without artifacts from a pulse with motion artifacts [16]. Skewness indicates the degree of asymmetry in the probability distribution of a random variable around its mean. It can take on positive, zero, negative, or undefined values, reflecting the shape and symmetry of the distribution. Kurtosis is the sharpened of the peak of a frequency-distribution curve, and standard deviation reflects the dispersion degree of a data set. If X is considered a variable with μ and σ, the mean and standard deviation, respectively, statistical values are calculated as follows:

$$\mathrm{Kurt}[X] = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{E\left[(X-\mu)^4\right]}{\sigma^4} \tag{1}$$

$$\mathrm{Skew}[X] = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{E\left[(X-\mu)^3\right]}{\sigma^3} \tag{2}$$

$$\mathrm{std}[X] = \sqrt{E\left[(X-\mu)^2\right]} \tag{3}$$

These values are calculated for each pulse of a signal. So, in our case, the variable X represents a vector of all the samples in a pulse. If the shape of the cycle changes, then these statistical values will no longer be constant. To be able to detect outliers, thresholds that detect skewness, kurtosis, and standard deviation values that are not normal, i.e. values for artifact-free cycles, were set up. For this reason, the distribution of each of these three statistics over a pulse can be estimated using a normal distribution [31]. The aim is to reduce the risk of a pulse being incorrectly annotated. To do this, a wide confidence interval is taken to ensure that the probability that the value of the corresponding statistic is not unnecessarily rejected. If X is considered to be a variable that can be approximated by a normal distribution $\mathcal{N}(\mu, \sigma^2)$, the probability that this variable lies within the chosen confidence interval can be written as follows:

$$\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545 \tag{4}$$

After several experiments, this 95% confidence interval gives the best results, as it reduces the risk of poor detection. So, lower and upper thresholds can be defined as follows:

$$th_l = \mu - 2\sigma \tag{5}$$

$$th_u = \mu + 2\sigma \tag{6}$$

The mean and standard deviation are calculated for each statistic measured by taking the set of values for each pulse of a signal. A waveform segment is classified as containing motion artifacts to effectively detect motion artifacts if at least one of the three statistics falls outside the defined thresholds. The result of this first step is a small proportion of the annotated dataset, with a binary value for each pulse: pulse with artifact or without artifact. The annotations given by the algorithm are then compared with the expert's annotations and found to have 80% similarity. After examining the annotations with the expert, the function chosen to segment the pulses did not always correctly segment a pulse that was formed by a distinct diastole and systole curve. In a cardiac cycle, diastole is the relaxation phase when the heart fills with blood, and systole is the contraction phase when the heart pumps blood out to the body or lungs. In this case, two pulses were detected instead of one. This segmentation error partly explains the 20% difference in annotation between the expert and the algorithm. The percentage of similarity is considered high enough to validate the expert's annotations.

3) IMBALANCED DATASET

The two classes of annotated data are unevenly distributed. The annotation includes many more pulses without motion artifacts, approximately 80% and 20% of pulses with motion artifacts. For accurate results with the algorithms, the data needs to be resampled. The complex characteristics of our clinical data, such as small training sizes, many features, and correlations between the features, make the task more complicated. Understanding the interconnectedness of these variables is crucial for accurate analysis and prediction [23]. Oversampling and undersampling methods are the most frequently used. Under-sampling reduces the majority of class examples, achieving a balanced dataset, with random under-sampling (RUS) being a well-known method. However, under-sampling may lead to the loss of valuable information from the majority class. On the other hand, over-sampling increases the minority class examples. Random over-sampling (ROS) replicates existing minority examples, but it may result in overfitting. Synthetic minority over-sampling technique (SMOTE) generates artificial minority examples by interpolating between selected examples and their nearest neighbors. Modifications such as adaptive synthetic sampling (ADASYN) adjust the number of artificial minority examples based on the density of majority examples surrounding the original minority example [32]. Also, it is concluded that there is no clear winner between oversampling and undersampling to compensate for the class imbalance if factors such as class distribution, class prevalence, and features correlations in medical decision-making [23] are not taken into consideration. In the section V, the different results obtained with the sampling methods will be presented to conclude on the best method for our study.

4) LABEL PROPAGATION

The LP algorithm is an iterative algorithm that assigns labels to unlabeled data points by propagating labels through the dataset. It was first presented in an article published in 2002 by X. Zhu and Z. Ghahramani, entitled "Learning from labeled and unlabeled data with label propagation" [33]. In graph-based semi-supervised learning methods, a graph where each node is represented by a vector of features is created. The edges between nodes are weighted based on how similar the features are. When the weights of the edges are high, it means that the connected nodes are likely to have the same label. This idea is based on the assumption that samples close to each other in the graph are part of the same group or category [34]. At the start of the algorithm, only a small proportion of the data is already labeled, corresponding here to the proportion of data annotated in the previous step. In our case, considering that we have around 51 pulses per signal and that we have annotated 10% of the entire database of 1571 signals, we therefore have 8000 pulses, and thus 8000 nodes in the graph. This algorithm is based on the hypothesis that if two nodes are connected, they carry a similarity. Usually, the Euclidean distance between nodes is calculated to establish the graph. Depending on the kernels chosen for the algorithm's operation, this distance measurement may be different. Consider the following notations:

$$u : \text{number of unlabeled points}$$
$$l : \text{number of labeled points}$$
$$k : \text{number of classes}$$

In the final state, this algorithm aims to look at all the probabilities a node has of belonging to a certain class and take the largest. Y a matrix with rows containing the probabilities that a node belongs to a certain class is considered. This matrix Y is a $N \times k$ matrix where $N = l + u$. Also considered T, a $N \times N$ probability transition matrix. This matrix T is obtained by calculating the degree matrix (D) and the adjacency matrix (A). It defines the probability of jumping from one node to another in t steps. This number t can tend towards infinity [35]. The matrix Y contains two sub-matrices: Y_l and Y_u, respectively, for the known and unknown labels. The same applies to the T matrix, which contains 4 sub-matrices:

- T_{ll}: probability to get from labeled nodes to labeled nodes. This matrix will be an identity matrix.
- T_{lu}: probability of getting from labeled to unlabelled nodes. This will be a zero matrix because labelled nodes are absorbing states, it means you are in a self-loop and can't move in any direction.
- T_{ul} and T_{uu}: probability to get from unlabelled nodes to labeled and unlabelled nodes, respectively.

Consider $\hat{Y}$, the probability matrix of annotations obtained in the final state. The matrix T is set to the infinite power, and Y_0 represents the initial annotations of the nodes. The equation for the final stage of this algorithm can be expressed as:

$$\hat{Y} = T^{t \to \infty} Y_0 \tag{7}$$

In 7, the matrix T is set to the power t with t tending to infinity. It can be written as:

$$\lim_{t \to \infty} T^t = \begin{bmatrix} I & 0 \\ \left(\sum_{t=0}^{\infty} T_{uu}^t\right) T_{ul} & T_{uu}^{\infty} \end{bmatrix} \tag{8}$$

The sum between the brackets is similar, when t tends to infinity, to a geometric series that has an argument that is less than 1 in modulus. And if T_{uu} is multiplied by itself a large number of times, knowing that the values are less than 1, it will become very close to 0. Therefore, a conclusion on the limit of the transition matrix for a very large number of steps is:

$$\lim_{t \to \infty} T^t = \begin{bmatrix} I & 0 \\ (I - T_{uu})^{-1} T_{ul} & 0 \end{bmatrix} \tag{9}$$

The equation 7 can, therefore, be rewritten:

$$\begin{bmatrix} \hat{Y}_l \\ \hat{Y}_u \end{bmatrix} = \begin{bmatrix} I & 0 \\ (I - T_{uu})^{-1} T_{ul} & 0 \end{bmatrix} \begin{bmatrix} Y_{l0} \\ Y_{u0} \end{bmatrix} \tag{10}$$

For unknown labels, the following formula can be written:

$$\hat{Y}_u = (I - T_{uu})^{-1} T_{ul} Y_{l0} \tag{11}$$

This matrix contains the new labels and is the output of the algorithm. To sum up, the various stages of the algorithm can be summarized as follows:

1) Creation of a graph with nodes labeled and unlabeled.
2) Calculation of the probability transition matrix T. This matrix is linked to the degree matrix D, a diagonal matrix where each diagonal element corresponds to the sum of edge weights connected to that node. Also linked to the adjacency matrix A, it is a square matrix where each row and column corresponds to a node, and the value at the intersection indicates whether there's an edge (value 1, otherwise 0) connecting those nodes. The formula is: $T = D^{-1} \cdot A$. This matrix is the same throughout the algorithm.
3) Calculation of the new labels for each t iteration:

$$Y^{t+1} = T^t Y^t \tag{12}$$

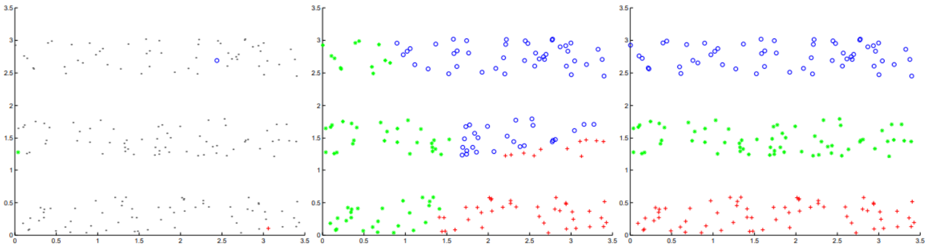
4) Repeat step 3 until convergence.

A concrete example of how the LP algorithm works is shown in figure 4. This example is based on a sample of synthetic data where each of the three classes is represented in a band. The KNN algorithm is unaware of the band structure and fails to propagate labels efficiently. The LP model, on the other hand, recognizes this structure and uses it to its advantage to group labels.

While performing label propagation, groups of closely linked nodes quickly reach a consensus on a single label, causing many labels to vanish. Only a few labels remain after propagation. When nodes end up with the same label after convergence, it signifies that they are part of the same group.

D. CLASSIFICATION

Once the ground truth has been established, the aim is to classify the pulses and compare the results obtained with the annotations. Machine learning classifiers are used for classification. These automatic algorithms categorize data into the two classes of our problem. They operate as mathematical models, utilizing statistical analysis and optimization techniques to detect patterns within the data. By identifying these patterns, classifiers can assign each instance to a specific class or category. There are a wide variety of traditional classifiers, both supervised and unsupervised. Supervised classifiers have been chosen to be



(a) The data (b) kNN (c) Label propagation

FIGURE 4. **Comparison of label propagation between KNN and LP model with the "3 Bands dataset". (a) 3 initial annotated points (3 classes represented in green, red, and blue) and 178 non-annotated points (b) annotated dataset with KNN (c) with LP. From [33].**

utilized to process medical data, which is also temporal data. Here are 4 examples [25]:

1) **KNN**: this is a supervised method where k represents the number of neighbors. For classification, when given a new input data point, the algorithm identifies the k nearest neighbors from the training dataset based on their feature similarity. The class label of the majority of these k neighbors is then assigned to the new data point.
2) **SVM**: in SVM, data points are mapped as vectors within a high-dimensional space. The algorithm aims to identify the optimal hyperplane that distinctly categorizes the classes. In binary classification, a hyperplane can be considered as a boundary delineating two distinct data classes. While numerous hyperplanes might achieve this separation, the algorithm selects the one that provides the most effective separation. For a specific classification purpose, as is the case for this project, we have subsequently used SVC, a type of SVM specialized for classifications.
3) **DT**: Each internal node represents a feature or attribute, and each branch represents a decision rule based on that feature. The leaf nodes of the tree represent the final class label or predicted value. During training, each value is separated based on the attribute. When making predictions, new data points traverse the decision tree by following the decision rules at each node until reaching a leaf node, which then provides the predicted class label or value.
4) **NB classifier**: this classifier uses probability to predict whether an input will fit into a certain category. It builds a statistical model based on these probabilities. Naive Bayes calculates the likelihood of the data point belonging to each class using the previously estimated probabilities.

Traditional classifiers have big advantages for small or medium datasets that require simpler or linear models. They have few layers in their architecture; conversely, deep learning (MLP and Transformers) architectures comprise multiple layers of neural networks. Deep architectures take advantage of unsupervised pre-training at the layer level, which facilitates efficient tuning of the deep networks and enables them to extract intricate structures from input data. These extracted features at higher levels contribute to improved predictions and overall performance [36]. For classification, MLP and Transformers are neural networks classifiers:

- **MLP**: it consists of multiple layers of nodes (neurons) that are interconnected through weighted connections. MLP employs a feedforward mechanism, where information flows from the input layer through the hidden layers to the output layer. Each node in the network applies an activation function to the weighted sum of its inputs to produce an output. Through a process called backpropagation, the MLP classifier adjusts the weights to minimize the error between predicted and actual labels during training.
- **Transformers**: it relies on the attention mechanism. The attention-mechanism looks at an input sequence and decides at each step which other parts of the sequence are important. A Transformer is an architecture for transforming one sequence into another one with the help of two parts (Encoder and Decoder).

The objective is to apply all these classifiers to the PPG signal pulses so that a comparison of the classifiers on our medical data can be built. In addition to being compared with each other, these classifiers will also be compared with the LP semi-supervised algorithm, which annotates our database and classifies artifacts in PPG signals.

IV. EXPERIMENTAL IMPLEMENTATION

First, as a reminder, in the LP algorithm, the two input matrices are the annotation matrix, a binary vector, and a matrix containing the features for each pulse. Each pulse represents a node in the algorithm's graph. For the choice of features, the signal from a temporal perspective has been considered. Therefore, an input matrix for the algorithm of size 256 samples × the number of pulses can be obtained.

Different metrics have been chosen to evaluate our results. The negative state (0) is a pulse without motion artifacts, whereas the positive state (1) is a pulse with motion artifacts. All these measures are based on the evaluation of false negatives (FN), pulse with artifact incorrectly identified as a clean pulse, false positives (FP), clean pulse incorrectly identified as a pulse with artifact, true negatives (TN), clean pulse correctly identified as a clean pulse, and true positives (TP), pulse with artifact correctly identified as a pulse with artifact. The following metrics are defined:

- **Confusion Matrix**: a table with two rows and two columns that reports the number of true positives, false negatives, false positives, and true negatives.
- **Precision, Recall, and F1**: these three scores give a more general idea of how the algorithm works, rather than just looking at the algorithm's accuracy, which can be biased in certain situations. They are defined as follows:

$$\text{Precision} = \frac{TP}{TP+FP}$$
$$\text{Recall} = \frac{TP}{TP+FN}$$
$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Matthews Correlation Coefficient (MCC), Cohen's Kappa Coefficient, and Critical Success Index (CSI)**: MCC is particularly useful in the case of binary classification, where the two classes are unbalanced. It varies between 0 and 1. CSI is also known as the Threat Score. A CSI of 1 indicates perfect prediction, while a score of 0 indicates no successful predictions beyond random chance. Kappa Coefficient (κ) is stronger than accuracy; it ranges from -1 to 1.

$$\text{mcc} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$
$$\text{csi} = \frac{TP}{TP+FN+FP}$$
$$\kappa = \frac{2 \times (TP \times TN - FP \times FN)}{(TP+FP)(FP+TN)+(TP+FN)(FN+TN)}$$

- **AUROC/AUC**: the AUROC or AUC (Area Under the Receiver Operating Characteristic curve) represents the probability that the model correctly ranks a randomly chosen positive instance higher than a randomly chosen negative instance. The ROC curve is created by plotting the TP rate against the FP rate.

Using these metrics, the hyperparameters of our model for the LP algorithm need to be defined. The first parameters defined are the parameters of the function used. The choice of the kernel is between KNN or RBF (radial basis function); depending on this choice, two other associated parameters could be modified. The maximum number of iterations and the algorithm's convergence tolerance remained the default values: 1000 iterations and 10^{-3} for the convergence threshold. The first parameter to be defined was the choice of kernel. For this, a cross-validation was carried out on the data. This involves dividing the data into several parts and then running the two algorithms using different values for the parameters on each part, keeping one part aside for performance testing. Then a calculation of the average performance can be done over all the test parts for each value and choose the one that gives the best performance. A KNN kernel with a number of neighbors of 7 has been chosen. The table 2 summarizes the parameters chosen for the algorithm. The data are separated as follows: 70% training and 30% testing. The data from the training part are redivided evenly to obtain 50% of unlabeled data and 50 % of labeled data.

V. RESULTS AND DISCUSSION

Different proportions of the dataset were tried for annotation to optimize the Lable Propagation algorithm and achieve the best performance on automatic labeling. The aim is to annotate as few pulses as possible. 2.5%, 5%, 7.5%, and 10% of the dataset were annotated, given that the entire database contains 1571 signals, and the proportion that gave the best results was evaluated. For each proportion of the dataset, the precision, recall, and F1 values were analyzed to decide. The results for class 1 (class "pulse with artifacts") of these metrics are presented in the table 4. Because the data are imbalanced, the results for class 0 (class "pulse without artifacts") remain consistently good and don't change much with different parameters. The best results are obtained for a proportion of 5% of the dataset. Indeed, as the proportion of annotated data increases, the distribution of classes becomes even more disparate. For 2.5% there are 17.3% of pulses with artifacts, and for 5%, the proportion of pulses with artifacts is 18.1%. As the size of the annotated dataset increases, for 7.5% there are 16.4% of pulses with artifacts. For 10% of the dataset, 17.7% of pulses contain artifacts. All these values are summarized in the table 3. If the classes are more unbalanced, this may influence the algorithm, which will have greater difficulty in finding a constant pattern for propagating the labels. In the case of 10%, the proportion of pulses with artifacts is high, but the number of annotated pulses increases, and this may induce new data that is less representative of the overall data distribution, leading to poor generalization on unseen data.

The various resampling methods presented in section III-C3 were applied. The results are shown in table 5. First, we can see that despite the class imbalance, the algorithm manages to detect the artifacts for the training part and that the scores are correct (96% precision, 82% recall, and 89% F1). However, when we apply a resampling method, the results between the scores are more balanced. This results in a more robust algorithm. The difference in results between undersampling and oversampling can be explained by the fact that undersampling will reduce the number of majority, which leads to loss of data and loss of information from this data. On the contrary, oversampling increases the number of values in the minority class, providing more data. In our case, SMOTE is the best oversampling method. SMOTE selects a minority class instance and identifies its k-nearest neighbors in the feature space. It then creates new synthetic examples

TABLE 2. Summary table containing the label propagation parameters.

Model	Kernel	Number of neighbors	Number of iterations	Convergence threshold
Label Propagation	KNN	7	1000	10^{-3}

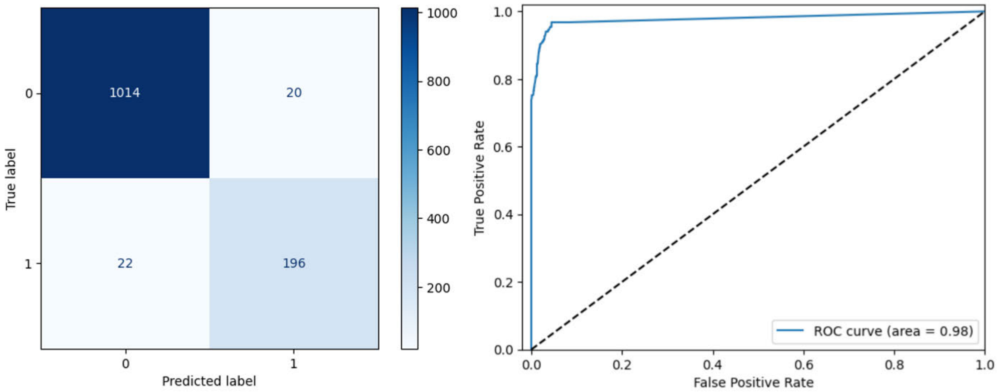


FIGURE 5. Confusion matrix and ROC curve for the LP algorithm with a KNN kernel with 7 neighbors, an oversampling method SMOTE, and 5% of the dataset already labeled.

TABLE 3. Summary table containing the labeling portion and the imbalance rate.

Dataset proportion	Artifacts (%)	Non-artifacts (%)
2.5%	17.3	82.7
5%	18.1	81.9
7.5%	16.4	83.6
10%	17.7	82.3

TABLE 4. Results for the class "with artifacts" for different proportions of the dataset.

Dataset proportion	Precision	Recall	F1
2.5%	0.89	0.88	0.89
5%	0.91	0.90	0.90
7.5%	0.84	0.88	0.86
10%	0.83	0.90	0.86

TABLE 5. Results for the class "with artifacts" for different sampling methods.

Sampling method	Precision	Recall	F1
None	0.96	0.82	0.89
RUS	0.87	0.90	0.88
ROS	0.91	0.87	0.89
SMOTE	0.91	0.90	0.90
ADASYN	0.88	0.91	0.90
ROS+RUS	0.89	0.91	0.90

along the line segments connecting the selected instance and its neighbors. By introducing these synthetic examples, SMOTE effectively increases the size of the minority class, making it comparable to the majority class and improving the performance of classifiers in handling imbalanced datasets. However, it is important to consider that resampling can potentially cause issues such as overfitting. It is important to monitor the model's performance after oversampling to detect any signs of overfitting or other potential issues. Our model showed no signs of overfitting, and resampling was very important for training a well-balanced classifier in the case of the imbalanced dataset.

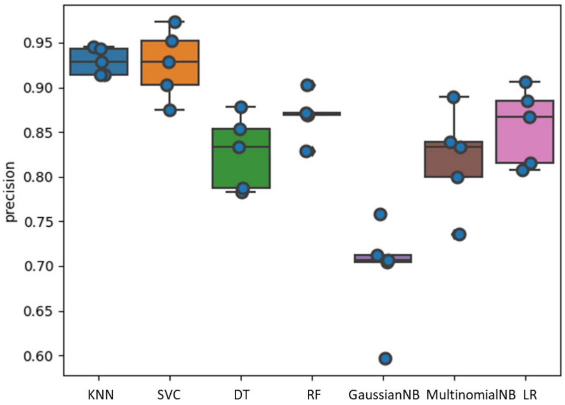


FIGURE 6. Precision evaluation for different traditional Machine Learning classifiers (in axis order: KNN, SVC, Decision Tree, Random Forest, Gaussian NB, Multinomial NB, Logistic Regression).

Given the correct sampling method and the appropriate proportion of the dataset to be selected, the results of the LP algorithm were evaluated. The confusion matrix is shown on the left and the ROC curve on the right, on the Fig. 5. The ROC curve is plotted for each decision threshold. In the case of the LP algorithm, this represents the probability assigned to each instance for each class. For a 5% dataset, the number of pulses for the validation part is 1252. 1034 belong to the "without artifacts" category, and 218 belong to the "with artifacts" category. The number of true positives and true negatives is higher than the number of false positives or false negatives. This indicates

TABLE 6. A comparison of the performance of different classifiers for the "with artifact" class. The oversampling method chosen is SMOTE for the LP model and ADASYN for the other models. In addition, 5% of the dataset is already annotated.

Model	Precision	Recall	F1	MCC	Kappa	CSI	AUROC
LP	0.91	0.90	0.90	0.89	0.89	0.82	0.98
KNN	0.89	0.95	0.92	0.90	0.90	0.85	0.98
MLP	0.76	0.97	0.85	0.83	0.82	0.74	0.99
Transformer	0.85	0.86	0.85	0.82	0.82	0.74	0.97
FCN	0.86	0.83	0.84	0.82	0.82	0.74	0.95

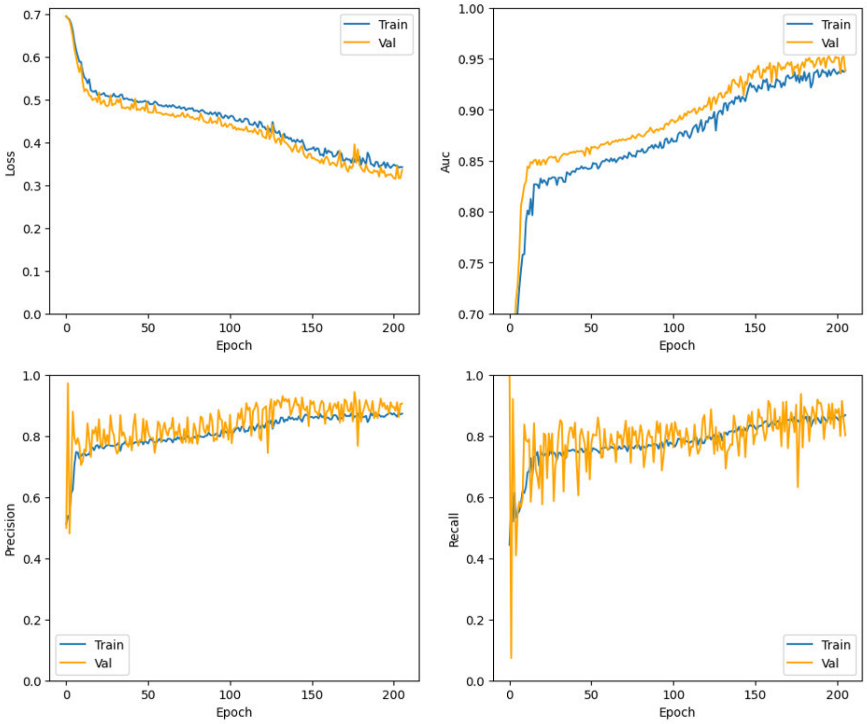


FIGURE 7. The comparison of CNN performance during training (Train) and validation (Val) for loss, AUROC, precision, and recall.

that the algorithm can understand and apply the model to unlabeled signals. However, the number of pulses detected as clean but containing artifacts (false negatives) is higher than the reverse (false positives). The LP algorithm uses neighborhood information to propagate labels through the data network. This means that labels for samples close in feature space tend to be similar. However, in the case of pulses containing artifacts, these artifacts may be similar to certain features of other clean pulses, leading to incorrect label propagation. As a result, pulses containing artifacts may be incorrectly labeled as clean by the LP algorithm, leading to a higher number of falsely classified pulses. Misclassified pulses are always an important problem in the medical field. This can lead to false alarms if the pulse is not a clean pulse or to misdetections. False alarms force hospital staff to make emergency visits due to outliers. These situations are exhausting and not necessary as an additional burden on caregivers. Presented in Table 6, the three additional scores, MCC, Kappa, and CSI, support this explanation. Indeed, these three values must be close to 1 to indicate a good algorithm prediction. The closer the score is to 0, the more random the algorithm's prediction. We note that the MCC, CSI, and Kappa values demonstrate a strong prediction of the LP algorithm. These scores allow us to validate the algorithm's correct performance. The AUROC is 0.98. The closer the AUROC is to 1, the better the model's performance. A high AUROC indicates that our model can distinguish between positive and negative classes.

After evaluating the LP algorithm, the performance of the different types of classifiers, presented in section III-D, were assessed. First, dealing with the imbalanced classes by oversampling using the ADASYN algorithm. Cross-validation was employed to ensure accurate model prediction and assess the reliability of the machine-learning algorithms. The results of the 5-fold cross-validation on different classifiers are presented in Fig. 6 shows a precision comparison using a box plot. Each blue dot represents the performance of an individual fold in the cross-validation. The figure indicates that KNN and SVC (with a kernel of 'rbf') are the top-performing classifiers, with median precision rates above 90%. However, KNN shows a slightly better and more consistent performance than SVC. This observation is highlighted by the broader range of variability for SVC, as indicated by the whiskers on the box plot, compared to KNN. To sum up, KNN and SVC are the top classifiers, but KNN is the more reliable and stable solution. KNN is also the best classifier compared to classifiers that use neural networks such as MLP and Transformers. Table 6 shows the different performances of the neural networks classifiers: MLP classifier, Transformer, Fully Convolutional Network (FCN) VS KNN classifier. MLP consists of 3 hidden layers with 500 neurons for each hidden layer. Its macro average accuracy (calculates the accuracy for each class individually and then computes the average accuracy across all classes) is 0.88, compared with 0.94 for KNN. In our case, using a complex model like MLP could lead to overfitting, as the model may have a high capacity relative to the amount of data available. In addition, training an MLP can be computationally expensive, especially with larger architectures and limited computational resources. Transformers, especially large ones like BERT (Bidirectional Encoder Representations from Transformers), have a high computational complexity and require significant computational resources for training and inference. Like the MLP classifier, Transformers works best on larger datasets because it needs a lot of data for the training part. Otherwise, the model has a greater capacity than the limited data, and the risk is overfitting. Generally speaking, in the medical field, Transformers excel in natural language processing tasks [37]. They can learn complex relationships and patterns within the text, making them suitable for medical text classification and understanding tasks.

Additionally, The LP algorithm demonstrates consistent and balanced performance in precision and recall, as evidenced by the experiment results in Fig. 5 and Table 6. The confusion matrix in Figure 5 shows that the LP algorithm with a KNN kernel (7 neighbors) achieves high accuracy, correctly classifying the majority of positive and negative cases, leading to a precision of 0.91 and a recall of 0.90. The ROC curve with an area of 0.98 further indicates the model's ability to distinguish between classes. Table 6 reinforces these findings by comparing the LP algorithm to other classifiers, where LP maintains competitive precision and recall values while achieving high F1 (0.90) and MCC (0.89) scores. These results highlight that the LP algorithm, combined with SMOTE for oversampling and leveraging 5% of already labeled data, effectively balances precision and recall, ensuring robust performance in classifying the "with artifact" class.

For the last classifier, experiments were conducted with an FCN model. FCN is a neural network architecture for semantic segmentation, producing dense pixel-wise predictions. It consists of convolutional layers without fully connected layers, enabling it to handle images of any size and preserve spatial information. Using FCN for time series classification involves adapting the fully convolutional architecture to process one-dimensional time series data. Instead of working with two-dimensional images, the FCN is applied to sequences of data points. The temporal convolutional layers capture temporal patterns and dependencies in the time series, and the decoding path with transposed convolutional layers helps to produce dense predictions for each data point in the sequence, enabling accurate time series classification [38]. One key benefit is that FCN eliminates manual feature engineering, as they can directly learn relevant features from raw time series data. This streamlines the classification process and saves time and effort in designing handcrafted features. Additionally, FCN enables end-to-end learning, optimizing feature representations and classification jointly, which can lead to improved performance. The flexibility of FCN with input size allows them to handle time series data of varying lengths without requiring resizing or padding, making them suitable for irregularly sized data. Moreover, FCN produces dense predictions for each time step, capturing fine-grained temporal patterns and enhancing the informativeness of classification results. Experimentally, during FCN training, it is evident that the process takes longer than other approaches. However, its performance is not comparable to those methods, mainly due to its lower accuracy.

For training MLP, FCN, and Transformer, we use the binary cross-entropy loss as follows:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(Y_i \times log(\hat{Y}_i) + (1 - Y_i) \times log(1 - \hat{Y}_i)) \tag{13}$$

We use the Adam optimizer and early stopping to deal with the overfitting. We use GridSearchCV to fine-tune the hyper-parameters, balancing the best combination and computation time. Only certain hyper-parameters typically affect a neural network's accuracy, specifically the number of hidden layers, nodes in each hidden layer, and the learning rate [39]. By focusing on these, grid search effectively optimizes all parameters simultaneously, allowing for quick model training. Grid search also offers straightforward parallelization and flexible resource allocation, which other approaches lack [40].

Fig. 7 provides a comprehensive view of the model's performance over time. The improvements in metrics like loss, AUC, precision, and recall suggest that the model is learning and improving its performance with each epoch. The consistent trends between training and validation data indicate that the model is generalizing well and not overfitting significantly. However, we can see the fluctuation between precision and recall; it can be confirmed that FCN cannot deal with the imbalanced classes from the nature of the data.

So, compared with previous studies, the artifact classification algorithm we have implemented has the advantage of having a faster execution time on large volumes of data compared to EMD or wavelet denoising, for example. It exploits the intrinsic relationships between the data rather than decomposing each signal individually. It also has the advantage of being easily generalizable to other signals since no additional parameters are required.

VI. LIMITATIONS AND FUTURE WORK

This study delved into utilizing semi-supervised LP methods for artifact classification within PPG signals, especially in scenarios characterized by imbalanced class distributions. The study showed us that our model is sensitive to data volume, and its improvement is limited as data volume increases. One future objective is to improve our model, particularly in feature detection. To augment the capability of our model, we can add some steps in the preprocessing part. In section III-C2, the segmentation problem has already been mentioned. First, adaptive filtering techniques can attenuate artifacts without affecting the signal. Signal quality can be improved through noise reduction methods, such as singular value decomposition (SVD). Alternative segmentation approaches can be employed to enhance the efficiency of the statistical analysis algorithm. Peak or minimum detection can be improved by employing derivative-based algorithms. A CNN model can also detect peaks, known for its pattern recognition capabilities [41]. Implementing alternative segmentation approaches and employing a CNN model for peak detection in PPG signals may face challenges in parameter tuning, validation, and network architecture design. However, careful optimization and validation of these approaches can improve the algorithm's accuracy and reliability.

Exploring data augmentation can also be a method to tackle the model's sensitivity to data volume. The authors in [42] found that using data augmentation allowed them to better handle the unbalanced class problem for binary or multiclass classification. These authors also draw a parallel with ensemble learning methods, which are new hybrid methods that are more robust against unbalanced data. In [43] and [44], the authors put into practice the use of RUSBoost for epilepsy seizure detection and schizotypy classification. Compared to classical models, the ensemble model performed very well, making it a candidate for feature classification. Investigating ensemble learning models may help to better handle class imbalances and classification tasks. Exploring self-supervised or unsupervised learning methods could be considered for future work to address the labeling challenge we encountered without relying on manual annotations.

VII. CONCLUSION

This study explored applying semi-supervised LP methods for artifact classification in PPG signals, addressing challenges posed by imbalanced class distributions. Comparative analysis with traditional supervised learning algorithms, MLP, and Transformer-based models demonstrated the superior performance of the LP classifier. This algorithm can enhance the overall quality of PPG signals in artifact classification by dynamically adapting to the specific characteristics of the dataset. It becomes more adaptable to variations in PPG signals caused by different types of motion artifacts. The improved balance between precision and recall indicates more robust classifier performance, which is critical for real-life medical applications.

Overall, this model holds promise for enhancing healthcare monitoring systems, with potential applications in ECG and arterial blood pressure signal analysis.

ACKNOWLEDGMENT

The clinical data used in this study were generously provided by the Research Center of the CHU Sainte-Justine Hospital, Université de Montréal. The authors thank Dr. Sally Al Omar, Dr. Michael Sauthier, and Edem Tiassou for their data support of this research and Kévin Albert for annotating the data.

REFERENCES

[1] J. M. Helm, A. M. Swiergosz, H. S. Haeberle, J. M. Karnuta, J. L. Schaffer, V. E. Krebs, A. I. Spitzer, and P. N. Ramkumar, "Machine learning and artificial intelligence: Definitions, applications, and future directions," *Current Rev. Musculoskeletal Med.*, vol. 13, no. 1, pp. 69–76, Feb. 2020.

[2] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: Management, analysis and future prospects," *J. Big Data*, vol. 6, no. 1, pp. 1–12, Dec. 2019.

[3] M. Greco, P. F. Caruso, and M. Cecconi, "Artificial intelligence in the intensive care unit," *Seminars Respiratory Crit. Care Med.*, vol. 42, no. 1, pp. 2–9, Feb. 2021.

[4] L. N. Sanchez-Pinto and R. G. Khemani, "Development of a prediction model of early acute kidney injury in critically ill children using electronic health record data," *Pediatric Crit. Care Med.*, vol. 17, no. 6, pp. 508–515, 2016.

[5] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 2, pp. 361–370, Mar. 2017.

[6] L. Huang, A. L. Shea, H. Qian, A. Masurkar, H. Deng, and D. Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," *J. Biomed. Informat.*, vol. 99, Nov. 2019, Art. no. 103291.

[7] S. Dara, S. Dhamercherla, S. S. Jadav, C. Babu, and M. J. Ahsan, "Machine learning in drug discovery: A review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 1947–1999, 2022.

[8] L. V. Ho, D. Ledbetter, M. Aczon, and R. Wetzel, "The dependence of machine learning on electronic medical record quality," in *Proc. AMIA Symp.*, 2018, pp. 883–891.

[9] K. Y. Ngiam and I. W. Khor, "Big data and machine learning algorithms for health-care delivery," *Lancet Oncol.*, vol. 20, no. 5, pp. e262–e273, May 2019.

[10] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine learning and decision support in critical care," *Proc. IEEE*, vol. 104, no. 2, pp. 444–466, Feb. 2016.

[11] H. Habehh and S. Gohel, "Machine learning in healthcare," *Current Genomics*, vol. 22, no. 4, pp. 291–300, 2016.

[12] D. Pollreisz and N. TaheriNejad, "Detection and removal of motion artifacts in PPG signals," *Mobile Netw. Appl.*, vol. 27, no. 2, pp. 728–738, Apr. 2022.

[13] M. T. Petterson, V. L. Begnoche, and J. M. Graybeal, "The effect of motion on pulse oximetry and its clinical significance," *Anesthesia Analgesia*, vol. 105, no. 6, pp. S78–S84, 2007.

[14] S.-H. Kim, M. Lilot, K. S. Sidhu, J. Rinehart, Z. Yu, C. Canales, and M. Cannesson, "Accuracy and precision of continuous noninvasive arterial pressure monitoring compared with invasive arterial pressure," *Anesthesiology*, vol. 120, no. 5, pp. 1080–1097, May 2014.

[15] B. L. Hill, N. Rakocz, Á. Rudas, J. N. Chiang, S. Wang, I. Hofer, M. Cannesson, and E. Halperin, "Imputation of the continuous arterial line blood pressure waveform from non-invasive measurements using deep learning," *Sci. Rep.*, vol. 11, no. 1, p. 15755, Aug. 2021.

[16] S. Hanyu and C. Xiaohui, "Motion artifact detection and reduction in PPG signals based on statistics analysis," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, May 2017, pp. 3114–3119.

[17] C.-C. Wu, I.-W. Chen, and W.-C. Fang, "An implementation of motion artifacts elimination for PPG signal processing based on recursive least squares adaptive filter," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2017, pp. 1–4.

[18] G. Joseph, A. Joseph, G. Titus, R. M. Thomas, and D. Jose, "Photoplethysmogram (PPG) signal analysis and wavelet de-noising," in *Proc. Annu. Int. Conf. Emerg. Res. Areas, Magn., Mach. Drives*, Jul. 2014, pp. 1–5.

[19] Q. Wang, P. Yang, and Y. Zhang, "Artifact reduction based on empirical mode decomposition (EMD) in photoplethysmography for pulse rate detection," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 959–962.

[20] Z. Xu and K. Funaya, "Time series analysis with graph-based semi-supervised learning," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2015, pp. 1–6.

[21] Y. Shin, S. Yoon, S. Kim, H. Song, J. G. Lee, and B. S. Lee, "Coherence-based label propagation over time series for accelerated active learning," in *Proc. Int. Conf. Learn. Represent.*, Oct. 2021, pp. 1–20.

[22] D. Bünger, M. Gondos, L. Peroche, and M. Stoll, "An empirical study of graph-based approaches for semi-supervised time series classification," *Frontiers Appl. Math. Statist.*, vol. 7, pp. 1–18, Jan. 2022.

[23] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Netw.*, vol. 21, nos. 2–3, pp. 427–436, Mar. 2008.

[24] D. E. Robbins, V. P. Gurupur, and J. Tanik, "Information architecture of a clinical decision support system," in *Proc. IEEE Southeastcon*, Mar. 2011, pp. 374–378.

[25] A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain, P. Fergus, M. Al-Jumaily, and N. Radi, "Applied machine learning classifiers for medical applications: Clarifying the behavioural patterns using a variety of datasets," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Sep. 2015, pp. 228–232.

[26] D. Brossier, R. E. Taani, M. Sauthier, N. Roumeliotis, G. Emeriaud, and P. Jouvet, "Creating a high-frequency electronic database in the PICU: The perpetual patient," *Pediatric Crit. Care Med.*, vol. 19, no. 4, pp. 189–198, 2018.

[27] N. Roumeliotis, G. Parisien, S. Charette, E. Arpin, F. Brunet, and P. Jouvet, "Reorganizing care with the implementation of electronic medical records: A time-motion study in the PICU," *Pediatric Crit. care Med.*, vol. 19, no. 4, pp. 172–179, 2018.

[28] A. Mathieu, M. Sauthier, P. Jouvet, G. Emeriaud, and D. Brossier, "Validation process of a high-resolution database in a paediatric intensive care unit—Describing the perpetual patient's validation," *J. Eval. Clin. Pract.*, vol. 27, no. 2, pp. 316–324, 2021.

[29] P. K. Lim, S.-C. Ng, N. H. Lovell, Y. P. Yu, M. P. Tan, D. McCombie, E. Lim, and S. J. Redmond, "Adaptive template matching of photoplethysmogram pulses to detect motion artefact," *Physiological Meas.*, vol. 39, no. 10, Oct. 2018, Art. no. 105005.

[30] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiological Meas.*, vol. 33, no. 9, pp. 1491–1501, Sep. 2012.

[31] R. Krishnan, B. Natarajan, and S. Warren, "Analysis and detection of motion artifact in photoplethysmographic data using higher order statistics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 613–616.

[32] K. Fujiwara, Y. Huang, K. Hori, K. Nishioji, M. Kobayashi, M. Kamaguchi, and M. Kano, "Over-and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis," *Frontiers Public Health*, vol. 8, p. 178, May 2020.

[33] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," School Comput. Sci., Tech. Rep. CMU-CALD-02-107, 2002.

[34] Z. Song, X. Yang, Z. Xu, and I. King, "Graph-based semi-supervised learning: A comprehensive review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 1–21, Nov. 2022.

[35] Z. Bodó and L. Csató, "A note on label propagation for semi-supervised learning," *Acta Universitatis Sapientiae, Inf.*, vol. 7, no. 1, pp. 18–30, Jun. 2015.

[36] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings Bioinf.*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018.

[37] V. Yogarajan, J. Montiel, T. Smith, and B. Pfahringer, "Transformers for multi-label classification of medical text: An empirical comparison," in *Proc. Int. Conf. Artif. Intell. Med.*, 2021, pp. 114–123.

[38] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1578–1585.

[39] G. Luo, "A review of automatic selection methods for machine learning algorithms and hyper-parameter values," *Netw. Model. Anal. Health Informat. Bioinf.*, vol. 5, no. 1, pp. 1–16, Dec. 2016.

[40] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," 2020, *arXiv:2003.05689*.

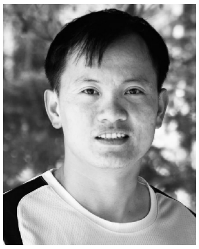
[41] K. Kazemi, J. Laitala, I. Azimi, P. Liljeberg, and A. M. Rahmani, "Robust PPG peak detection using dilated convolutional neural networks," *Sensors*, vol. 22, no. 16, p. 6054, Aug. 2022.

[42] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst. Appl.*, vol. 244, Jun. 2024, Art. no. 122778.

[43] M. Shen, P. Wen, B. Song, and Y. Li, "An EEG based real-time epilepsy seizure detection approach using discrete wavelet transform and machine learning methods," *Biomed. Signal Process. Control*, vol. 77, Aug. 2022, Art. no. 103820.

[44] A. Zandbagleh, S. Mirzakuchaki, M. R. Daliri, A. Sumich, J. D. Anderson, and S. Sanei, "Graph-based analysis of EEG for schizotypy classification applying flicker ganzfeld stimulation," *Schizophrenia*, vol. 9, no. 1, p. 64, Sep. 2023.

CLARA MACABIAU received the dual degree in Canada. After three years at the ENSEEIHT Engineering School, Toulouse, specializing in electronics, electrical energy and automation (EEEA). She is currently pursuing the master's degree in electrical engineering with the École de Technologie Supérieure, Montréal. Her master's project focused on the detection of artifacts in photoplethysmography signals from children admitted to pediatric intensive care at CHU Sainte-Justine. Her research interests include signal processing, machine learning, and electronics.



THANH-DUNG LE (Member, IEEE) received the B.Eng. degree in mechatronics engineering from Can Tho University, Vietnam, the M.Eng. degree in electrical engineering from Jeju National University, South Korea, and the Ph.D. degree in biomedical engineering from École de Technologie Supérieure (ÉTS), Canada. He is currently a Postdoctoral Fellow with the Biomedical Information Processing Laboratory, ÉTS. His research interests include applied machine learning approaches for biomedical informatics problems. Before that, he joined Institut National de la Recherche Scientifique, Canada, where he researched classification theory and machine learning with healthcare applications. He received the Merit Doctoral Scholarship from Le Fonds de Recherche du Quebec Nature et Technologies. He also received the NSERC-PERSWADE Fellowship from Canada and the Graduate Scholarship from Korean National Research Foundation, South Korea.

KÉVIN ALBERT received the degree from the EUSES School of Health and Sport, Girona, Spain, in 2018. He is currently pursuing the master's degree in biomedical engineering program with Université de Montréal. He has been with the Clinical Decision Support System (CDSS) Laboratory, Pediatric Intensive Care Unit, Sainte-Justine Hospital, Montréal, Canada, since May 2023, under the supervision of Prof. P. Jouvet. He developed clinical expertise in the field of function rehabilitation after neuro-traumatic injury (France) and in cardio-respiratory rehabilitation (Switzerland). He is a Physiotherapist. His research interests include the application of new technologies of support care system tools with artificial intelligence, especially in ventilatory support. His research program is supported by the Sainte-Justine Hospital and Quebec Respiratory Health Research Networks (QRHNs).

MANA SHAHRIARI received the bachelor's degree in electrical engineering, the master's degree in artificial intelligence, and the Ph.D. degree in electrical engineering. She is currently a Postdoctoral Researcher with CHU Sainte-Justine Research Centre, affiliated with Université de Montréal. She is also an Artificial Intelligence (AI) Researcher passionate about employing AI to address practical and real-world challenges. Her research interests include signal processing (including time-series analysis), image processing and computer vision, machine learning, deep learning, and statistical analysis of data.

PHILIPPE JOUVET received the M.D. degree, the M.D. degree (specialty) in pediatrics, and the M.D. degree (subspecialty) in intensive care from Paris V University, Paris, France, in 1989, 1989, and 1990, respectively, and the Ph.D. degree in pathophysiology of human nutrition and metabolism from Paris VII University, Paris, in 2001. He joined the Pediatric Intensive Care Unit, Sainte Justine Hospital, Université de Montréal, Montréal, QC, Canada, in 2004. He is currently the Deputy Director of the Research Center and the Scientific Director of the Health Technology Assessment Unit, Sainte Justine Hospital, Université de Montréal. He has a Salary Award for Research from Quebec Public Research Agency (FRQS). He currently conducts a research program on computerized decision support systems for health providers. His research program is supported by several grants from the Sainte-Justine Hospital, Quebec Ministry of Health, FRQS, Canadian Institutes of Health Research (CIHR), and the Natural Sciences and Engineering Research Council (NSERC). He has published more than 160 articles in peer-reviewed journals. He gave more than 120 lectures at national and international congresses.

RITA NOUMEIR (Member, IEEE) received the master's and Ph.D. degrees in biomedical engineering from École Polytechnique de Montréal. She is currently a Full Professor with the Department of Electrical Engineering, École de Technologie Superieure (ÉTS), Montréal. She has extensively worked in healthcare information technology and image processing. She has also provided consulting services in large-scale software architecture, healthcare interoperability, workflow analysis, and technology assessment for several international software and medical companies, including Canada Health Infoway. Her research interest includes applying artificial intelligence methods to create decision support systems.