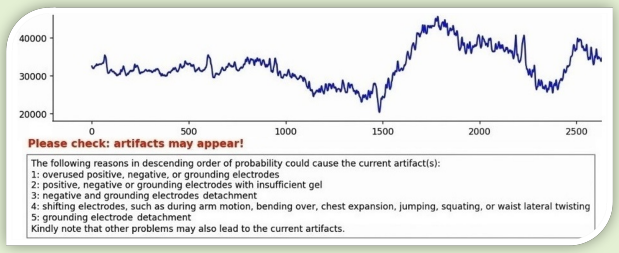
Taxonomy and Real-Time Classification of Artifacts During Biosignal Acquisition: A Starter Study and Dataset of ECG

Hui Liu, *Member, IEEE*, Shiyao Zhang, Hugo Gamboa, *Senior Member, IEEE*, Tingting Xue, Congcong Zhou, *Member, IEEE*, and Tanja Schultz, *Fellow, IEEE*

***Abstract*—This article investigates electrocardiogram (ECG) acquisition artifacts often occurring in experiments due to human negligence or environmental influences, such as electrode detachment, misuse of electrodes, and unanticipated magnetic field interference, which are not easily noticeable by humans or software during acquisition. Such artifacts usually result in useless and irreparable signals; therefore, it would be a great help to research if the problems are detected during the acquisition process to alert experimenters instantly. We put forward a taxonomy of real-time artifacts during ECG acquisition, provide the simulation methods of each category, collect and share a 10-subject data corpus, and investigate machine learning (ML) solutions with a proposal of appropriate handcrafted features that reach an offline recognition rate of 90.89% in a five-best-output person-independent (PI) leave-one-out cross-validation (LOOCV). We also preliminarily validate the real-time applicability of our approach.**



Screenshot of a running demonstration for real-time artifact detection and classification during ECG acquisition: The experimenter is being warned. Model: support vector machine (SVM) with one-tier person-independent training, 8-second window length and no overlap, 11 statistical domain features, and 5-best outputs.

***Index Terms*— Artifact, biosignal, electrocardiogram (ECG), electrocardiography, pattern recognition, real-time system, signal quality.**

Manuscript received 25 June 2023; revised 7 December 2023; accepted 13 January 2024. Date of publication 26 January 2024; date of current version 14 March 2024. The APC was funded by the Open Access Initiative of the University of Bremen and the DFG via SuUB Bremen. The associate editor coordinating the review of this article and approving it for publication was Prof. Jungyoon Kim. *(Corresponding author: Hui Liu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee, College of Biomedical Engineering and Instrument Science, Zhejiang University.

Hui Liu is with the Cognitive Systems Laboratory, University of Bremen, 28359 Bremen, Germany, and also with the Institute for Artificial Intelligence in Medicine, School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: hui.liu@uni-bremen.de).

Shiyao Zhang, Tingting Xue, and Tanja Schultz are with the Cognitive Systems Laboratory, University of Bremen, 28359 Bremen, Germany (e-mail: shiyao.zhang@uni-bremen.de; tingting.xue@uni-bremen.de; tanja.schultz@uni-bremen.de).

Hugo Gamboa is with the Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics (LIBPhys), NOVA School of Science and Technology (Campus de Caparica), 2829-516 Caparica, Portugal (e-mail: h.gamboa@fct.unl.pt).

Congcong Zhou is with the School of Medicine, Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou 310016, China (e-mail: zjdxzcc@zju.edu.cn).

Digital Object Identifier 10.1109/JSEN.2024.3356651

I. Introduction

BIOSIGNALS are increasingly helping human life. In addition to medical applications, biosignals are widely studied and used in areas such as sports assistance, interactive interfaces, smart homes, education, and entertainment. The quality of the obtained signals largely influences various subsequent research and experiments.

Regarding detecting artifacts in real time during the biosignal acquisition process, images and videos, categorized as optical biosignals in biomedical research, can be observed by the naked eye to judge whether they contain artifacts. Similarly, for audios, which can be categorized as acoustic biosignals, it is feasible to check for errors or noise during acquisition from signal observation and accompanying listening.

Optical and acoustic biosignals are mainly recorded by external sensing techniques. In contrast, internal sensing technologies, many of which are associated with wearables, use specific sensors to receive electrical, magnetic, mechanical/kinematic, thermal, and chemical biosignals. Artifacts of electrical biosignals, such as in electrocardiogram

(ECG), electromyogram (EMG), electroencephalogram (EEG), electrocorticogram (ECoG), electrooculogram (EOG), electroretinogram (ERG) electrogastrogram (EGG), electrovaginogram (EVG), electrohysterogram (EHG), and electrodermal activity (EDA), among others, are usually less readable on real-time signal visualization, especially at high sampling rates, meaning that even professionals may not be able to accurately and timely determine from the waveforms of the signals being acquired whether they are of good quality or subject to some interference. Such a situation is exacerbated by the fact that many research tasks involve multimodal biosignal acquisition, that is, simultaneous collection of multiple sensor types (e.g., ECG + EMG), quantities (e.g., four EMGs), and channels (e.g., X-, Y-, and Z-axes of the accelerometer), similar to many human activity data collection tasks [1], [2]. Clear visualization may not be possible when recording a large number of channels at the same time. Even if the graphics technique works, the simultaneous observation of several or even dozens of channels by the collector is challenging. Due to the real-time, continuous, and long-term nature of biosignal visualization, increasing the labor of observation or the personnel number is not feasible to ensure a high-qualitative data corpus. Some researchers have experienced the pain of discovering signal problems in some channels during post-processing that can undo hours of data acquisition effort and put the human, material, and time resources in vain. Not to mention that many users may not be senior signal experts—data acquisition work often occurs at the introductory stage for young scholars. Aiming to provide solutions, real-time intervention in the biosignal acquisition process with the help of machine learning (ML) has become this article's research topic.

Artifacts/errors/problems/noises in biosignal acquisition arise for various sorts of reasons, two of which are quickly addressed or widely researched and not the subject of this work:

1) From hardware, device, and transmission perspectives, errors caused by an insufficient power supply, damaged devices, or unstable wired/wireless transmission, among others, are fatal and cannot be fixed later. Acquisition software and software development kits (SDKs) can alert and avoid such issues during the acquisition by detecting lost connections or receiving error codes.
2) Some unavoidable signal noise/artifacts within reasonable limits are related to device specifications, acquisition surroundings, and biological conditions. Even if they are known to exist, nothing can be done during acquisition. Such problems can be mended at the post-processing stage, such as correction [3], [4], [5], rectification [6], [7], and denoising [8], [9].

This article investigates biosignal acquisition artifacts frequently occurring in experiments due to human negligence or environmental influences, such as electrode detachment, misuse of electrodes, unanticipated magnetic field interference, and signal distortion by human movements, which are not easily noticeable by experimenters or software during acquisition but can be discovered by ML in real time. Such artifacts usually result in useless and irreparable signals; therefore, it would be a great help to research if the problems are detected during the acquisition process, and the experimenters are alerted to them. This work also contributes to diagnosing and tracking medical metrics as a practical aid. Take the continuous positive airway pressure (CPAP) treatment of breathing problems as an example. In many countries, the patient needs to periodically wear biosensors, including ECG, respiration, and blood volume pulse (BVP) sensors, among others, during one night of sleep at home, and the measured signals are returned to the clinic for analysis. Pulling on cables or tangling the CPAP tubing and sensor cables due to sleeping position changes often results in invalid data collection of detachment or strong noise. Thus, the patient must return to the clinic to wear the sensors again for a second acquisition. If, with the patient's permission, some audible/vibration alerts are generated for the user when a critical signal error, such as a dropped electrode, is detected, a significant saving of medical resources (cost, time, and labor) and enhancement of convenience is expected.

We use ECG, a very common biosignal, as the initiating study object of the artifacts during biosignal acquisition. The overall research framework of taxonomy and real-time classification of ECG acquisition artifacts can provide a superior reference value for researching other bioelectrical signals. It can also radiate to more biosignal types, such as inertial biosignals from accelerometers, gyroscopes, or magnetometers. In this work, we

1) propose a taxonomy of ECG acquisition artifacts and their simulation schemes.
2) collect a 10-subject 199.82-min data corpus and make it freely available;
3) investigate lightweight ML models, architectures, handcrafted features, and parameter configurations to achieve offline artifact detection and classification; and
4) validate the method's real-time applicability.

Our pilot work's future development can serve as a feasible plug-and-play aid. Not only can the data collector be immediately alerted when an artifact occurs, but also the types of artifacts that are likely to occur can be prompted to help the experimenter efficiently examine the experimental setup, which facilitates the acquisition of biosignals, as the graphical abstract shows. Although there have been previous works to assess the quality of biosignals, such as ECG [10], [11], EMG [12], [13], EEG [14], [15], and photoplethysmogram (PPG) [16], [17], or to search for normal patterns and anomalies [18], [19], to our knowledge, this work is the first to investigate the artifact detection and classification during biosignal acquisition with a taxonomy proposal and ML solutions.

II. Taxonomy and Simulation

A. Taxonomy of Real-Time Artifacts During Electrocardiography

Leaving aside the easily solvable underlying device problems or the extensively studied allowable noise, artifacts during signal acquisition can be divided into two main categories: technical artifacts caused by the equipment (e.g., electrodes and cables) and biological artifacts generated by the user [20].

TABLE I
TAXONOMY OF REAL-TIME ARTIFACTS DURING ECG ACQUISITION. IN EACH ABBREVIATION, THE FIRST TERM INDICATES THE ARTIFACT CATEGORY (TECHNICAL OR BIOLOGICAL ARTIFACT); THE SECOND POINTS OUT WHERE THE PROBLEM TO BE EXAMINED LIES; THE THIRD SPECIFIES FURTHER DETAILS

TA-*: Technical Artifacts	
Abbreviation	ECG Aquisition Artifact
TA-C-S	**C**able **S**witched: positive and negative
TA-D-P	**D**etachment: **P**ositive electrode
TA-D-N	**D**etachment: **N**egative electrode
TA-D-G	**D**etachment: **G**rounding electrode
TA-D-PN	**D**etachment: **P** and **N** electrodes
TA-D-PG	**D**etachment: **P** and **G** electrodes
TA-D-NG	**D**etachment: **N** and **G** electrodes
TA-E-IG	**E**lectrodes with **I**nsufficient **G**el
TA-E-O	**E**lectrodes **O**verused: **P**, **N**, and **G**
TA-I	Magnetic field **I**nterference
TA-M-P	**P**ostitive electrode on active **M**uscles
TA-M-N	**N**egative electrode on active **M**uscles
TA-M-G	**G**rounding electrode on active **M**uscles
TA-M-PN	**P** and **N** electrodes on active **M**uscles
TA-M-PG	**P** and **G** electrodes on active **M**uscles
TA-M-NG	**N** and **G** electrodes on active **M**uscles
TA-M-PNG	**P**, **N** and **G** electrodes on active **M**uscles
TA-S-A	**S**hifting Electrodes: **A**rms up and down
TA-S-B	**S**hifting Electrodes: **B**end over
TA-S-C	**S**hifting Electrodes: **C**hest expansion
TA-S-J	**S**hifting Electrodes: **J**ump
TA-S-S	**S**hifting Electrodes: **S**quat
TA-S-W	**S**hifting Electrodes: **W**aist lateral twisting
BA-*: Biological Artifacts	
Abbr.	ECG Aquisition Artifact
BA-R-D	**R**espiration: **D**eep Breathing
BA-R-H	**R**espiration: Breath **H**olding
BA-R-R	**R**espiration: **R**apid Breathing

The taxonomy, as listed in Table I, covers the majority of common oversights and errors that a data collector may encounter during biosignal acquisition.

The ECG artifacts defined and their simulation schemes in this article were determined through iterative discussions, tests, and experiments among an international group of experienced biomedical research experts, biosignal instrumentation professors, and senior technical staff of well-known wearable sensor companies (see Acknowledgment).

B. Simulation of Artifacts

Detachment can occur between the electrodes and the skin or between the cables and the electrodes. Signal visualization does not show noticeable differences between these two cases. TA-D-* is simulated by detaching the relevant electrodes from the skin.

In order to simulate the case of insufficient gel (TA-E-IG), we divided the electrode covering film into several equal parts in a scalloped manner and adhered them partially back to the electrodes. Overall, the visualization of the masked signal is close to the original signal, and even many are smoother as if a low-pass filter is applied, especially when the masked part is no larger than 50% (see Fig. 1). We masked up to 87.5% (7/8) of the electrodes since more masking would result in difficulty in affixing the electrodes to the skin. In different

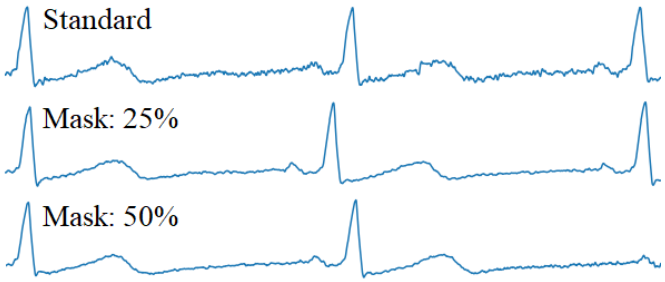


Fig. 1. Example of masking electrodes to simulate the gel deficiency situation. Top: standard electrode; middle: 25% of the gel is covered by a right-angle sector; bottom: 50% of the gel is covered by two right-angle sectors.

subjects' data recordings, the electrodes were shaded with 25%, 50%, 75%, or 87.5% to simulate different gel lack situations (the masking rate is recorded in the file name). We have tested masking each one, each pair, or all of the positive, negative, and grounding electrodes in the preliminary stage, and they caused no significant difference in the signal pattern. Therefore, we only acquired the signals of the masked grounding electrode.

The overused electrodes' situation (TA-E-O) was always simulated after each participant's all recording sessions were finished (about 1.5 h). Even so, the electrodes were repeatedly adhered to and detached from fibers as well as the subject's skin many times for about 20 min to make them almost unstuck.

Magnetic field interference (TA-I) was simulated by subjects holding an energized laptop power transformer at about a fist's distance from the heart.

The left/right sternocleidomastoid muscles (when head turning to the opposite direction) and the quadriceps muscle (when seated) were used to simulate wrongly placing the positive, negative, or grounding electrodes to active muscles (TA-M-*). When simulating the positive electrode sensing active muscles, the electrode was placed on the left sternocleidomastoid muscle, and the subject kept the head turned to the right and vice versa. To simulate both the positive and the negative electrodes connecting to active muscles simultaneously, the subject kept the head tilted upward to activate the sternocleidomastoid muscles on both sides.

TA-S-* may seem at first glance to be classified as biological artifacts since they are produced by human motions. In fact, the essence of the artifacts they cause is the electrodes' drift. In nomenclature, we use shifting (S) instead of drifting to avoid overlapping with the abbreviation of detachment, where S can also be seen as an abbreviation for sports. During signal acquisition, the subject performed the relevant activities at a normal tempo.

C. Impact of Directly Applying Protocol-Based Physical Events as Artifact Classes for ML

In this work, we simulated and modeled artifacts by physical events instead of signal patterns. The advantage is obvious: the outputs are supported in the possible artifact explanation to the user. Nonetheless, using a physical event nomenclature as ML classes instead of grouping similar patterns weakens

the accuracy due to the confusion of some classes. Applying the N-best outputs primitively addressed the problem (see Section IV-G).

Alternatively, we also discussed conducting secondary labeling regarding signal patterns in time and frequency domains, summarizing up to 10 classes, such as baseline drift and significant high-frequency noise. Such categorization cannot correspond to an actual acquisition protocol simulated by a specific physical event; moreover, a recorded signal from physical events may be applied for the model training of multiple signal patterns, for example, by weighting. Better readability and accuracy for pattern recognition are apparent advantages of such an alternative ML process, whereas the classified patterns may lack daily interpretability—what should we suggest to the experimenter for setup adjustment by a baseline drift? If the advice is too much or too general, the work almost degenerates to artifact detection instead of the classification and suggestion that we pursue. In any case, it is another topic to be studied, not related to the taxonomy directly by physical event acquisition in this work.

III. Experimental Device and Dataset

A. Biosignal Acquisition Device

We applied *biosignalsplux Researcher Kit*[1] as the biosignal recording device that provides expandable solutions of hot-swappable sensors. The biosignalsplux hub can simultaneously acquire up to eight channels of signals from arbitrarily selected sensors. In this work, we used one hub channel of the eight to obtain ECG signals at a sampling rate of 1000 Hz and a quantization level of 16 bits. The other seven channel interfaces provide a prospect for future research of multichannel and multimodal biosignal acquisition artifact detection and classification.

The particular model and manufacturer of the equipment may make this article's results not definitely universal. However, following the proposed taxonomy, acquisition process, and experimental procedures, our results should be replicated similarly on other equipment models and facilitate their respective users.

B. Protocol

Each class of TA-containing signals from each subject was acquired continuously for 30 s, while each BA was captured in two sessions of 15 s each, limited by human physiology. Besides, standard ECG signals (STD) were acquired twice from each subject, for 60 s each, once sitting and once standing.

It is impossible to record a signal piece for 15/30/60 s precisely. In order to facilitate the subsequent signal processing, the subject and the device were put into the state specified in the protocol before the signal acquisition started so that signals in all data files were usable from frame 0 onward. Each recording session took longer than the time specified in protocol (15/30/60 s). Our experiments were run based on each signal truncated according to the duration in the protocol (15/30/60 s).

[1]https://www.pluxbiosignals.com/products/researcher-kit (accessed June 25, 2023)

C. Subjects and Ethic

A total of 10 subjects without any known heart-related diseases, four females and six males, aged between 22 and 46 (28.0 Â± 6.9), participated in the data collection events. Each subject's participation time was approximately 2 h, consisting of announcements and precautions, questions and answers, equipment donning and adjustment, software preparation and test runs, data collection along all categories in the protocol, breaks, and device release.

All subjects signed a written informed consent form, and the study was conducted in accordance with Helsinki's World Medical Association (WMA) Declaration [21]. According to the consent form, we only kept the wearable sensor data pseudonymized and did not leave any identification information of the participants. The dataset is shared in an anonymized form.

D. Dataset

The 10-subject h5-formatted dataset we recorded and shared contains 199.82-min (3 h 19 min 49 s) data, of which 163.50 min are truncated according to the protocol described in Section III-B in files of 15, 30, or 60 s as an accompanying set. Each file name contains the artifact abbreviation (see Table I) of the corresponding signals, serving as a label for ground truth.

Some subjects' certain acquisitions contain extra sessions, which we also keep in the dataset in the spirit of using academic resources wisely. It causes a slight imbalance in the data amount between classes.

IV. Offline Study, Evaluation, and Discussion

A. ML Models and Leave-One-Out Cross-Validation

As a proof-of-concept, up-to-date experiments adopted three non-deep ML models, decision tree (DT), support vector machine (SVM), and random forest (RF), with window-based data training and recognition. Although the current models and configurations have achieved acceptable results, sequential modeling or deep learning is expected to improve the recognition rate further, which is one of the subsequent research topics.

This article uses most of the default settings of the three ML models in the *Python* package *scikit-learn* and the following specific parameters based on the validation of iterative experimental parameter tuning.

1) *DT:* $random\_state = 0$.
2) *SVM:* Radial basis function (RBF) kernel with $c = 80$ and $gamma = 250$.
3) *RF:* $min\_samples\_split = 2$ and $n\_estimators = 25$.

All experiments have been conducted using leave-one-out cross-validation (LOOCV) for person-independent (PI) training (see Section IV-F) to comprehensively validate offline PI models, whose overall recognition rate is computed as a macro average by accumulating 10-fold results.

B. Artifact Classes to Recognize

Biological artifacts due to the effect of physical activity on the signal influence the data quality, which is identifiable

by experienced evaluators and can lead to misinterpretations in computer-assisted analysis (e.g., slow eye movements in frontal EEG) [20]; however, on the one hand, they are difficult to avoid, and on the other hand, the resulting signal is still maximally usable by subsequent studies. For example, during sports exercise, non-routine breathing (see BA-R-* in Table I) can affect ECG, which does not mean that the subject should change their exercise habits to accommodate the equipment or even cancel certain sessions.

Although we also simulated and acquired three types of biological artifacts in our dataset, they were difficult to be identified individually and significantly interfered with the effective identification of standard signals and technical artifacts with the existing models. Since biological artifacts are not utterly "wrong" signals and, even being detected in real time, cannot be effectively avoided during the acquisition in most cases [20], the real-time artifact detection and classification proposed in this article do not include biological artifacts for the moment. We still include the acquisition signals of all subjects' three biological artifacts in the shared dataset to seek a better classification in the subsequent work, for example, applying sequential or deep modeling.

The seven TA-M-* and six TA-S-* artifacts exhibit strong confounding within their respective parent categories while also degrading the classification of STDs and other artifacts. Therefore, in this article's proof-of-concept experiments, we grouped these artifacts as TA-M and TA-S, respectively, which is acceptable for reporting artifacts to the experimenter. The tentative scheme should not be the final solution in the future, as some essential differences can still be detected in the signal visualization.

C. Feature Extraction and Selection

Since this proof-of-concept article involves non-deep learning, we extracted features manually utilizing the open-source time-series feature extraction library (TSFEL) [22]. It has been shown in the recent literature that handcrafted features do not necessarily perform worse than deep neural representations in biosignal-based pattern recognition research, for example, human activity recognition [23].

TSFEL offers 60 common and exclusive features. The real-time nature of this study's final application and the planning for future multichannel/multimodal synchronous processing helped us to narrow down the feature selection at the outset. Following the summaries in [18] and [24], we retained only 30 features with low computational consumption in the temporal, statistical, and spectral domains for our initial experiments.

Of particular concern is that using feature sets in the temporal or spectral domain alone, or mixing them with features in the statistical domain, exhibited poor overall recognition rates. Both temporal and spectral domain features have shone in various biosignal research areas, but most research tasks are based on (as far as possible) correctly acquired signals containing advantageous time (e.g., zero-crossings and slop) and frequency (e.g., periodicity) characteristics. On the contrary, most signals studied in this article are exceptional because they are irrational and do not conform to everyday patterns, which may explain the powerlessness of features in temporal and spectral domains in the pilot study.

More desirable results were achieved with statistical domain features alone. The results of the forward or backward greedy selection of 11 statistical domain features, that is, mean, median, standard deviation, mean absolute deviation, variance, root-mean-square, max, min, interquartile range, kurtosis, and skewness [18], cannot decisively distinguish which features are more effective or more redundant. Considering that they are all computationally less expensive, in this article's preliminary experiments, we adopt the entire 11 statistical domain features.

D. Window Length and Overlap Ratio

Window length and overlap ratio are two critical variables in this study. A standard ECG signal has typical waveforms within a reasonable duration. However, various artifacts do not necessarily have the correct ECG waveform and the expected typical duration. Therefore, we do not hypothesize whether a window of 1 or more typical ECG lengths is advantageous for identifying all types of artifacts. In the absence of any referenceable literature on real-time ECG artifact classification, we have at least three reasonable considerations:

1) A window should last at least one typical ECG cycle duration because the standard ECG is one of the classes to be recognized.
2) If performing better, the window length can be longer than that of a typical biosignal-based research task. For example, an online model notifying the user 10 s after an artifact occurs is totally acceptable.
3) Applying solely statistical domain features (see Section IV-C) relaxes the limit of windowing the signal, that is, the window can start from any position of an ECG cycle.

Greedy selections were executed with different ML models and parameter settings. To identify the best combinations, we performed joint selection experiments using heatmaps with 1–10 s of window lengths in a step of 0.5 s and 0%–90% overlap ratios in a step of 10%, as Fig. 2 exemplifies the best-performing SVM model among all three ML models.

No obvious trend on the impact of window length or overlap ratio in the series of greedy selection outcomes of the three models can be summarized. Note that the signals vary greatly across artifact classes; thus, longer/shorter window lengths or overlaps can lead to better recognition of some artifacts, while are worse for others.

E. One-Tier and Hierarchical Modeling

Considering the balance of training data volume and the ease of modeling, the most straightforward architecture is to train the model by placing STD, the class of standard ECG signals, on an equal footing with the various artifact types, which we call a one-tier modeling (1-M), as Fig. 3(bottom) illustrates. In the earlier multiple, PI parameter tuning attempts, STD's recognition rate did not exceed 50%. Understandably, many artifact-containing signals are very similar in visualization to the standard signals, resulting in considerable false negative cases of STD. STD's imperfect recognition rate

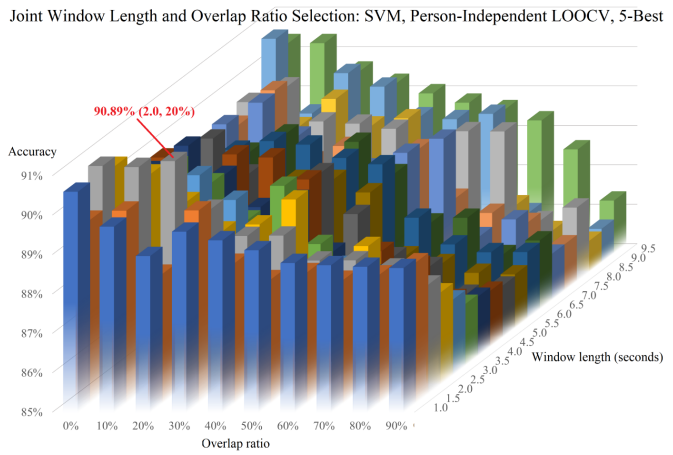


Fig. 2. Instance of the joint window length and overlap ratio selection using heatmap: SVM with 11 statistical domain features, PI LOOCV, 1-M, and five-best outputs.

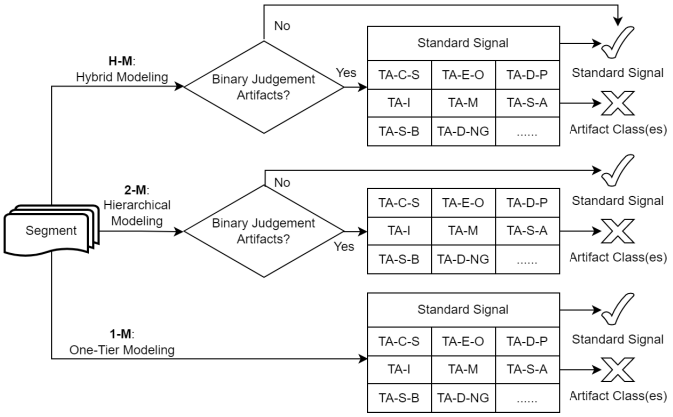


Fig. 3. Flow charts of the 1-M (bottom), the hierarchical modeling (middle), and the hybrid modeling (top).

(or recall) means that there will be a high probability of warning the problem-free signal wrongly, interfering with users' normal acquisition.

To improve STD's accuracy/recall, we designed a hierarchical modeling (2-M), that is, dichotomous recognition for STD/artifact signals first and then further artifact classification for segments judged as artifacts, as Fig. 3(middle) depicts. With 2-M, STD's PI recognition rates raised to over 60% (RF or SVM). The errors should be due to the significant imbalance in the data volume of STD versus TA and the fact that a fuzzy model mixing all artifact types is not generally robust.

A model incorporating 1-M and 2-M was envisioned, called hybrid architecture (H-M), which, in contrast to 2-M, enables that even segments judged as artifacts in the first layer could still be classified as STD in the second layer [see Fig. 3(top)], with a view to further reducing the number of STD's false negative instances.

STD's recognition accuracy with 1-M has been constantly enhanced to reach acceptable values (over 90%) during a huge number of experiments for tuning different technical means,

TABLE II
RECOGNITION RATES OF *N*-BEST LOOCV EXPERIMENTS WITH THE RF AND THE SVM MODELING FOR COMPARING THE PERFORMANCE OF ONE-TIER PURE PI AND SEMI-PI TRAINING WITH 11 STATISTICAL DOMAIN FEATURES. SEMI-PI VALUES INFERIOR TO PI ARE INDICATED IN GRAY

#-best	**RF:** PI	Semi-PI	**SVM:** PI	Semi-PI
1-best	41.11%	40.99%	43.16%	42.59%
2-best	59.00%	60.59%	65.06%	64.89%
3-best	70.83%	71.43%	76.18%	76.46%
4-best	78.03%	77.87%	83.21%	83.22%
5-best	83.53%	83.64%	90.89%	90.98%

such as using *N*-best outputs (see Section IV-G). Furthermore, the data size imbalance between STD and artifact classes unresolvable at the present stage hinders the training of 2-M and H-M. Currently, we focused on 1-M in the offline experiments and the pilot real-time validation. Nonetheless, 2-M and H-M are of great potential practical value, for example, when sequential modeling or deep learning is introduced or when more accurate TA recognition is expected. Moreover, a fuzzy artifact model for dichotomous judgments, as in 2-M and H-M, should allow better identification of unknown/unspecified artifacts than 1-M.

F. PI Training and Semi-PI Training

Many biosignal-based pattern recognition studies and applications take person-dependent (PD) training, which also greatly improves the recognition accuracy; PD training, however, is not the case in this research. It is impracticable that signals containing all artifact types need to be acquired and annotated from the new user beforehand to identify the acquisition artifacts during a new recording. The goal of our approach is that the vast majority of new acquisitions can benefit without barriers, so we trained and evaluated PI models.

With the advancement of various modeling architectures, parameter tuning, and other configurations (see other subsections in Section IV), the PI 1-M model achieves an overall five-best LOOCV accuracy of over 80% (RF) or 90% (SVM) (see Fig. 4 and Table II). To further improve STD's recognition accuracy (for reasons already explained in Section IV-E), we tested applying 50% of the currently recognized subject's STD data (the rest 50% for evaluating STD) together with the STD data pool of other individuals for training the STD model, while all artifact classes remain fully PI, called semi-person-independent (semi-PI) training. This operation is feasible in practice, requiring the experimenter to pre-record the user's signal for a few tens of seconds as correctly as possible.

In the 1-M experiments, Semi-PI can slightly improve the recognition rate of several N-best cases for both RF and SVM, which Table II and Section IV-G analyze. In the binary layer of 2-M and H-M, semi-PI training did not improve the model's performance, which should still be due to the two reasons mentioned in Section IV-D (the data volume imbalance; the blurry overall artifact class).

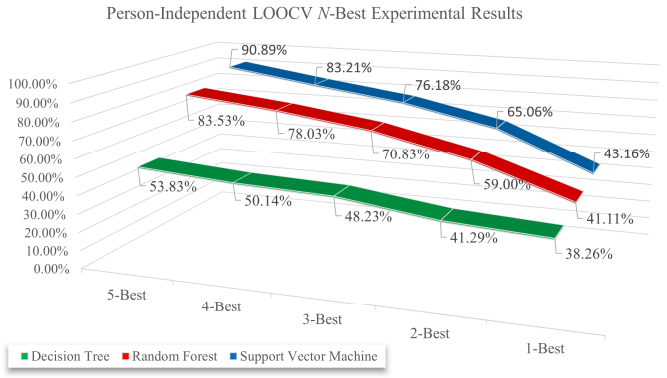


Fig. 4. Recognition rates of one-tier PI *N*-best LOOCV experiments with three non-deep ML models applying 11 statistical domain features. Each accuracy is the top result of the joint window length and overlap ratio selection.

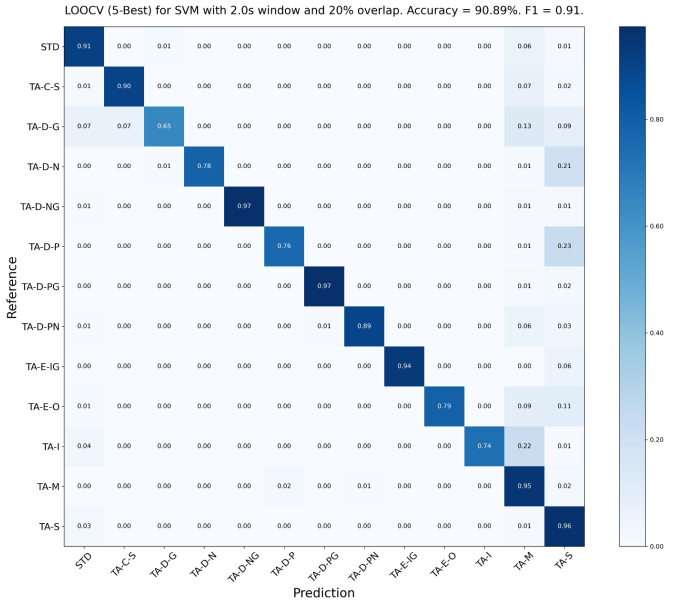


Fig. 5. Confusion matrix of the one-tier SVM modeling with 11 statistical domain features, 2-s window length, 20% overlap ratio, PI LOOCV, and five-best outputs.

A question may arise: Would it be possible to use only individualized STD data combined with PI artifact data? The answer is no, according to even worse experimental performance. We believe that the main reason is still the extreme imbalance in the amount of data. In practice, the size of the newly acquired non-artifact data from the new user is also not comparable to a stored artifact dataset of a large magnitude.

Semi-PI has shown only a limited capability for performance enhancement; it is also inconvenient in practice, though feasible. On the one hand, there is a need for artifact-free acquisition by experimenters or users, some of whom are not biosignal experts; on the other hand, a period of computation time is required for retraining the model before it can be put into use. The PI LOOCV without individualized STD data, even using non-sequential and non-deep ML, has already achieved acceptable results, as Fig. 4, Table II, and Section IV-G elucidate profoundly. The future applications we envision should be pure PI training, with the auxiliary tool generated by our approaches for signal acquisition users "plug-and-play."

G. N-Best Outputs of Recognition Hypotheses

We proposed the *N*-best-output mechanism for this article's research purpose, which does not just give the classification result with the highest probability, but the top-*N* most likely classes in descending order of probability. The fuzziness can be helpful in artifact recognition, for it can improve the recognition accuracy of PI training, where interpersonal normal or artifact-containing signals show significant individual specificity. In practice, *N*-best is also helpful for a scientific data acquisition process—alerting experimenters to multiple possibilities rather than a single artifact and reminding them to examine more adequately.

The following strategy for evaluating the experimental results of *N*-best was adopted. As long as one of the first *N* results is identical to the ground truth, the recognition is taken as a correct case; If all the first *N* results do not match the reference, the one with the highest probability is regarded as the recognition output for calculating the precision, recall, F1-score, accuracy, and confusion matrix statistics.

Fig. 4 depicts the offline recognition rates of the three models using one to five-best outputs and pure PI training, where each result is the best LOOCV recognition rate in the joint window length and overlap ratio selection, as Fig. 2 suggests. The resulting diagram with all best F1-scores is similar, with only minor discrepancies in values that do not influence the judgment.

Table II compares the *N*-best performance of RF and SVM using pure PI or semi-PI training. DT is omitted because of the apparent overall inferior accuracy.

Table II clearly evidences that semi-PI worked even worse by some *N* values on both ML approaches; the improvement, if any, is tiny. Such statistics echo the final arguments in Section IV-F. The current pilot real-time implementation and future in-depth studies should focus on pure PI.

Fig. 5 shows the confusion matrix of the five-best pure PI LOOCV with the highest overall recognition rate (90.89%), using SVM modeling.

Several artifact classes, such as TA-I, TA-E-O, TA-D-P, TA-D-N, and TA-D-G, are the main source of recognition errors. Most of their misattributions are TA-M and TA-S. As already mentioned in Section IV-B, TA-M-* and TA-S-* artifacts are indeed the more difficult ones to deal with. Merging them to two categories, TA-M and TA-S, has quite an impact on the overall model by absorbing plenty of other classes' instances, including STD, as exposed in Fig. 5. The reason should be mainly the unstable fuzzy model mixing many classes. Taking TA-M and TA-S out of training did produce encouraging results (e.g., SVM with 1-s window

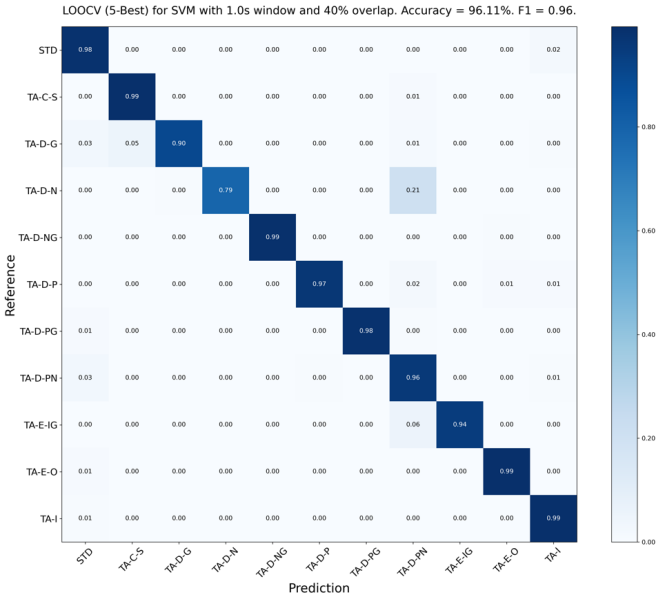


Fig. 6. Confusion matrix of the SVM modeling excluding TA-M and TA-S, with 11 statistical domain features, 1-s window length, 40% overlap ratio, PI LOOCV, and five-best outputs.

length, 40% overlap ratio, and five-best PI LOOCV: 96.11% recognition accuracy and F1-score = 0.96) (see Fig. 6), but we will commit to analyzing and investigating them deeply. We also did not exclude TA-M and TA-S in our pilot real-time validation.

V. From Offline Toward Real Time

Our offline LOOCV experiments can be viewed as pseudo-real-time evaluations since the parameters, such as window length, overlap ratio, features, and N-best, among others, could be directly applicable to the online model. Genuine real time is more appealing, though. We conducted preliminary experiments as an initial attempt to validate the real-time usability of our proposed approach. Using the SDK provided by the acquisition device (see Section III-A), we developed real-time signal acquisition and visualization software and, based on offline PI experimental settings, realized window-based real-time artifact classification, as the graphical abstract demonstrates. Longer windows, for example, 8 s, can be taken to ensure real time, at the expense of a small loss of recognition accuracy. Experimental results reveal that most artifact types can be detected and notified to the user after they are generated, with some aspects for further research and improvement.

1) A shorter window plus a higher overlap rate often performs well in offline models, which occasionally causes real-time delays. A compromise needs further studying.
2) One-best output currently does not achieve good performance (also in the offline model); N-best enhances the accuracy, while the output order or results sometimes change over time.
3) Environment-causing artifacts, such as TA-I, are occasionally poorly recognized in real time while sometimes, in contrast, interfering with other artifacts' recognition. Reducing the sensitivity and improving the accuracy need simultaneous investigation.

VI. Conclusion and Outlooks

We innovated detecting, classifying, and alerting artifacts in real time during biosignal acquisition, for which we proposed a taxonomy of ECG acquisition artifacts and their simulation methods, collected and shared a 10-subject 199.82-min dataset, conducted large-scale offline experiments, and developed a preliminary program to validate our method in real time. Our research provided a perspective that traditional ML methods also work for data with a predominantly error/noise/artifact/problem composition. After all, most research is based on (as much as possible) correct data.

The current real-time procedure verifies that our proposal is achievable from the proof-of-concept perspective. Further research aims for stable and widespread application, which needs to be boosted from two aspects. One is to investigate sounder offline models, for example, through introducing sequential or deep modeling, enriching the dataset (especially the scale of the standard signals), and experimenting with 2-M and H-M. Both hierarchical models can be further augmented with additional layers, such as focusing on classifying artifact subclasses within the parent categories of TA-M or TA-S. The second is to design an intelligent real-time feedback scheme. For example, to cope with the continuously changing N-best results, we consider using artifact category counters that gradually give the most targeted results after several windows. Alternatively, a real-time probability histogram of the entire artifact categories is a worth-implementing illustration.

Our ML study is oriented toward interpretability. The paradigm can be practically extended to studying the real-time artifact classification of other biosignal types.

Acknowledgment

The authors would like to acknowledge Dr. Daniel Osório, Francisco Cachado, Rui Varandas, and Guilherme Ramos' support during conceptualization, discussion, and data acquisition.

Ethical Statement

The study was conducted in accordance with Helsinki's WMA (World Medical Association) Declaration [21]. All participants signed their written informed consent to participate in this study. According to the consent form, we only kept the wearable sensor data pseudonymized and did not leave any identification information of the participants. The dataset is shared in an anonymized form.

Data Availability

The dataset covered in this work can be downloaded at https://www.uni-bremen.de/en/csl/research/sensorder - artifact-classification-during-biosignal-acquisition. The dataset is freely available for non-commercial academic research.

REFERENCES

[1] B. Hu, E. Rouse, and L. Hargrove, "Benchmark datasets for bilateral lower-limb neuromechanical signals from wearable sensors during unassisted locomotion in able-bodied individuals," *Frontiers Robot. AI*, vol. 5, p. 14, Feb. 2018.

[2] H. Liu, Y. Hartmann, and T. Schultz, "CSL-SHARE: A multimodal wearable sensor-based human activity dataset," *Frontiers Comput. Sci.*, vol. 3, p. 90, Oct. 2021.

[3] P. Ritter, R. Becker, C. Graefe, and A. Villringer, "Evaluating gradient artifact correction of EEG data acquired simultaneously with fMRI," *Magn. Reson. Imag.*, vol. 25, no. 6, pp. 923–932, Jul. 2007.

[4] C. Vidaurre, T. H. Sander, and A. Schlögl, "BioSig: The free and open source software library for biomedical signal processing," *Comput. Intell. Neurosci.*, vol. 2011, pp. 1–12, Mar. 2011.

[5] B. Chen, B. Zheng, and W. Sun, "Ultra-resolution spectral correction based on adaptive linear neuron for biomedical signal processing," *Frontiers Public Health*, vol. 9, p. 564, May 2021.

[6] L. J. Myers et al., "Rectification and non-linear pre-processing of EMG signals for cortico-muscular analysis," *J. Neurosci. Methods*, vol. 124, no. 2, pp. 157–165, Apr. 2003.

[7] O. P. Neto and E. A. Christou, "Rectification of the EMG signal impairs the identification of oscillatory input to the muscle," *J. Neurophysiol.*, vol. 103, no. 2, pp. 1093–1103, Feb. 2010.

[8] P. Celka, K. N. Le, and T. R. H. Cutmore, "Noise reduction in rhythmic and multitrial biosignals with applications to event-related potentials," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 7, pp. 1809–1821, Jul. 2008.

[9] M. A. Kabir and C. Shahnaz, "Denoising of ECG signals based on noise reduction algorithms in EMD and wavelet domains," *Biomed. Signal Process. Control*, vol. 7, no. 5, pp. 481–489, Sep. 2012.

[10] D. Hayn, B. Jammerbund, and G. Schreier, "QRS detection based ECG quality assessment," *Physiolog. Meas.*, vol. 33, no. 9, pp. 1449–1461, Sep. 2012.

[11] L. Johannesen and L. Galeotti, "Automatic ECG quality scoring methodology: Mimicking human annotators," *Physiolog. Meas.*, vol. 33, no. 9, pp. 1479–1489, Sep. 2012.

[12] C. Sinderby, L. Lindstrom, and A. E. Grassino, "Automatic assessment of electromyogram quality," *J. Appl. Physiol.*, vol. 79, no. 5, pp. 1803–1815, Nov. 1995.

[13] G. D. Fraser, A. D. C. Chan, J. R. Green, and D. T. MacIsaac, "Automated biosignal quality analysis for electromyography using a one-class support vector machine," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 12, pp. 2919–2930, Dec. 2014.

[14] C.-G. Bénar et al., "Quality of EEG in simultaneous EEG-fMRI for epilepsy," *Clin. Neurophysiol.*, vol. 114, no. 3, pp. 569–580, Mar. 2003.

[15] S. Mohamed, S. Haggag, S. Nahavandi, and O. Haggag, "Towards automated quality assessment measure for EEG signals," *Neurocomputing*, vol. 237, pp. 281–290, May 2017.

[16] C. Orphanidou, "Quality assessment for the photoplethysmogram (PPG)," in *Signal Quality Assessment in Physiological Monitoring*. Cham, Switzerland: Springer, 2018, pp. 41–63.

[17] E. K. Naeini, I. Azimi, A. M. Rahmani, P. Liljeberg, and N. Dutt, "A real-time PPG quality assessment approach for healthcare Internet-of-Things," *Procedia Comput. Sci.*, vol. 151, pp. 551–558, Apr./May 2019.

[18] J. Rodrigues, H. Liu, D. Folgado, D. Belo, T. Schultz, and H. Gamboa, "Feature-based information retrieval of multimodal biosignals with a self-similarity matrix: Focus on automatic segmentation," *Biosensors*, vol. 12, no. 12, p. 1182, Dec. 2022.

[19] D. Folgado et al., "TSSEARCH: Time series subsequence search library," *SoftwareX*, vol. 18, Jun. 2022, Art. no. 101049.

[20] R. Kramme, *Medizintechnik: Verfahren-Systeme-Informationsverarbeitung*. Berlin, Germany: Springer, 2016.

[21] World Medical Association, "World medical association declaration of Helsinki: Ethical principles for medical research involving human subjects," *J. Amer. Med. Assoc.*, vol. 310, no. 20, pp. 2191–2194, 2013.

[22] M. Barandas et al., "TSFEL: Time series feature extraction library," *SoftwareX*, vol. 11, Jan. 2020, Art. no. 100456.

[23] N. Bento et al., "Comparing handcrafted features and deep neural representations for domain generalization in human activity recognition," *Sensors*, vol. 22, no. 19, p. 7324, Sep. 2022.

[24] H. Liu, "Biosignal processing and activity modeling for multimodal human activity recognition," Ph.D. dissertation, Cogn. Syst. Lab, Fac. 3—Math. Comput. Sci., Universität Bremen, Bremen, Germany, 2021.

Hui Liu (Member, IEEE) was born in Shanghai, China, in October 1985. He received the B.Sc. degree in information engineering from Shanghai Jiao Tong University (SJTU), Shanghai, in 2008, the Diplom-Ingenieur degree in electrical engineering from the Technical University of Berlin, Berlin, Germany, in 2009, the M.Sc. degree in communication and information systems from SJTU, in 2011, and the Ph.D. degree from the University of Bremen, Bremen, Germany, in 2021.

He has been a Researcher with the Cognitive Systems Laboratory, University of Bremen, since 2016. His research interests include biosignal processing, human activity recognition, multimodal time-series analysis, and music information retrieval.

Shiyao Zhang was born in Beijing, China, in May 1994. She received the ISPO Certification (Category I Professional) Prosthetist-Orthotist from the International Society for Prosthetics and Orthotics in 2012. She is pursuing the degree in computer science with the University of Bremen, Bremen, Germany.

Her research interests include biosignal acquisition, annotation, and processing.

Hugo Gamboa (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the Instituto Superior Tecnico of University of Lisbon (IST UL), Lisbon, Portugal, in 2007.

He is a Co-Founder and the President of PLUX, a company that develops biosignals monitoring wearable technology. He is currently a Researcher with the Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics (LIBPhys), Faculdade de Ciencias e Tecnologia of NOVA University of Lisbon (FCT NOVA), Lisbon, where he is also a Full Professor with the Physics Department. Since 2014, he has been a Senior Researcher with the Fraunhofer Center, Assistive Information and Communication Solutions (AICOS). His research interests include biosignals processing and instrumentation.

Tingting Xue was born in Shanghai, China, in October 1985. She received the B.Sc. degree in finance from the East China University of Science and Technology, Shanghai, in 2008, and the B.Sc. degree in industrial engineering and management with the direction of information and communication systems from the Technical University of Berlin, Berlin, Germany, in 2020.

Her research interests include biomedical image processing, human activity recognition, and music signal processing.

Congcong Zhou (Member, IEEE) was born in Yiwu, Zhejiang, China. He received the B.S. and Ph.D. degrees in biomedical engineering from Zhejiang University, Hangzhou, China, in 2010 and 2016, respectively.

He is an Assistant Research Fellow with the Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, National Engineering Research Center for Innovation and Application of Minimally Invasive Devices. In recent years, he focused on the research and development of wearable medical equipment and physiological health monitoring systems, including non-contact/wearable human physiological parameter detection sensors or POCT device prototypes, as well as rehabilitation medical devices. He has presided several national, provincial, and enterprise entrusted projects and obtained two National Registration Certificates for Class II Medical Devices authorized by CFDA.

Tanja Schultz (Fellow, IEEE) received the diploma and Ph.D. degrees in informatics from the University of Karlsruhe, Karlsruhe, Germany, in 1995 and 2000, respectively.

She successfully passed the German State Examination for Teachers of Mathematics, Sports, and Educational Science with Heidelberg University, Heidelberg, Germany, in 1990. In 2000, she joined Carnegie Mellon University, Pittsburgh, PA, USA, where she holds a position as a Research Professor with the Language Technologies Institute. From 2007 to 2015, she was a Full Professor with the Department of Informatics, Karlsruhe Institute of Technology (KIT), Karlsruhe, before she accepted an offer from the University of Bremen, Bremen, Germany, in April 2015. Since 2007, she has been directing the Cognitive Systems Laboratory, where her research activities focus on human–machine communication, with a particular emphasis on multilingual speech processing and human–machine interfaces. Together with her team, she investigates the processing, recognition, and interpretation of biosignals, that is, human signals resulting from physical and mental activities, to enable human interaction with machines in a natural way. She has authored more than 450 articles published in books, journals, and proceedings, and is regularly invited as panelist and keynote speaker.

Dr. Schultz serves as a member for numerous conference committees. She is a Fellow of the International Speech Communication Association in 2016, the European Academy of Sciences and Arts in 2017, and the Asia-Pacific Artificial Intelligence Association in 2021. She has been an Editorial Board Member of *Speech Communication* since 2001. She served as a Board Member and an Elected President of the International Speech Communication Association ISCA from 2007 to 2015. She received several awards for her work, such as the FZI award for an outstanding Ph.D. thesis in 2001, the Allen Newell Medal for Research Excellence from Carnegie Mellon in 2002, the ISCA best journal award for her publication on language independent acoustic modeling in 2002 and on Silent Speech Interfaces in 2015, the Plux Wireless Award in 2011 for the development of Airwriting, the Alcatel-Lucent Research Award for Technical Communication in 2012, the Otto-Haxel Award in 2013, and the Google Research Award in 2013 and 2020 as well as several best paper awards. She serves as an Associate Editor for IEEE TRANSACTIONS from 2002 to 2004. She has been an Associate Editor of *ACM Transactions on Asian and Low-Resource Language Information Processing* since 2013.