A Machine Learning Perspective on fNIRS Signal Quality Control Approaches

Andrea Bizzego, Michelle Neoh, Giulio Gabrieli, and Gianluca Esposito

***Abstract*— Despite a rise in the use of functional Near Infra-Red Spectroscopy (fNIRS) to study neural systems, fNIRS signal processing is not standardized and is highly affected by empirical and manual procedures. At the beginning of any signal processing procedure, Signal Quality Control (SQC) is critical to prevent errors and unreliable results. In fNIRS analysis, SQC currently relies on applying empirical thresholds to handcrafted Signal Quality Indicators (SQIs). In this study, we use a dataset of fNIRS signals (N = 1,340) recorded from 67 subjects, and manually label the signal quality of a subset of segments (N = 548) to investigate the pitfalls of current practices while exploring the opportunities provided by Deep Learning approaches. We show that SQIs statistically discriminate signals with bad quality, but the identification by means of empirical thresholds lacks sensitivity. Alternatively to manual thresholding, conventional machine learning models based on the SQIs have been proven more accurate, with end-to-end approaches, based on Convolutional Neural Networks, capable of further improving the performance. The proposed approach, based on machine learning, represents a more objective SQC for fNIRS and moves towards the use of fully automated and standardized procedures.**

***Index Terms*— Deep learning, functional near infrared spectroscopy, machine learning, signal quality control.**

I. Introduction

THE adoption of functional Near-InfraRed Spectroscopy (fNIRS) has seen rapid growth in neuroimaging studies in recent years [1], particularly in fields such as infant neuroimaging [2] and cognitive neuroscience [3]. fNIRS is a non-invasive neuroimaging technique which detects the activity of cortical brain regions through the use of near-infrared light. Specifically, fNIRS measures the relative changes in the concentrations of oxygenated and deoxygenated hemoglobin, indicative of cerebral activation and deactivation, relying on the different light absorption [3].

Notwithstanding the broad adoption of fNIRS, there is no general consensus on the best fNIRS signal processing approaches [4], [5], with the use of different combinations of processing steps reported leading to different study outcomes [6].

One key pre-processing step in fNIRS data analysis is the Signal Quality Control (SQC) of the raw signals to remove low quality signals from the downstream analysis [1], [7]. Since a reference ground-truth for the quality of fNIRS signals is not available, the baseline is derived from human assessments performed by visual inspection of the signals. However, visual inspection makes the SQC dependent on researcher expertise and subjective judgments about what is expected to be observed in a "good" quality signal. Current approaches tend to avoid the use of visual inspection as the only SQC method, although it is often used to validate the results of the fNIRS signal processing pipelines (e.g. as in [8] and [9]). Finally, visual inspection is expected to have a key role for the creation of reference datasets, where the quality label is derived from human assessments [10].

SQC currently relies on the computation of Signal Quality Indicators (SQIs), based on a number of algorithms that aim at quantifying morphological characteristics of the fNIRS signal. The decision about which signals to remove from the downstream analysis is based on empirical fixed thresholds applied to the SQIs. Several SQIs have been proposed, for instance Scalp Coupling (SC) and Scalp Coupling Power (SCP) [11], Coefficient of Variation (CV) and the Coefficient of Variation of the Wavelengths (CVW) [12], Signal Quality Index [13], association with cardiac signals [14] and others.

Alternatively to thresholding, Machine Learning (ML) algorithms have been adopted on a wide range of physiological signals to classify signal quality [15]. Li and colleagues, for example, developed an automatic quality assessment method for pulsatile signals [16] and for ECG signals [17], while Gabrieli and colleagues [18] considered different ML classifiers to identify the quality of pupillometry signals. Regarding fNIRS signals, Sappia and colleagues suggested the Signal Quality Index [13], and then developed a ML algorithm based

Manuscript received 2 May 2022; revised 14 July 2022 and 3 August 2022; accepted 7 August 2022. Date of publication 11 August 2022; date of current version 19 August 2022. This work was supported in part by the Italian Ministry of University and Research through the Excellence Department Grant Awarded to the Department of Psychology and Cognitive Science, University of Trento, Italy, and in part by the European Union–FSE-REACT-EU, PON Research and Innovation 2014–2020 under Grant DM1062/2021. *(Corresponding author: Andrea Bizzego.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Nanyang Technological University Psychology Ethics Committee under Approval No. PSY-IRB-2020-007.

Andrea Bizzego and Gianluca Esposito are with the Department of Psychology and Cognitive Science, University of Trento, 38068 Trento, Italy (e-mail: andrea.bizzego@unitn.it).

Michelle Neoh is with the Psychology Program, Nanyang Technological University, Singapore 639818.

Giulio Gabrieli is with the Psychology Program, Nanyang Technological University, Singapore 639818, and also with the Neuroscience and Behaviour Laboratory, Italian Institute of Technology, 00161 Rome, Italy.

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2022.3198110, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2022.3198110

on the Signal Quality Index [19] which achieved promising results. However, the Signal Quality Index was developed and tested on a very limited sample size (N=123 for the development, N=40 for the evaluation) and data were collected within an environment in which potential conflict of interest was present.

Besides conventional ML approaches, deep learning methods, based on the use of Artificial Neural Networks (ANN) are nowadays being applied in a growing number of fields, typically improving the results achieved with conventional ML approaches [20], [21]. The adoption of ANNs in applications based on medical data is rapidly growing, with a wide range of applications [22], [23], [24], [25], [26]. Convolutional Neural Networks (CNNs) are a family of ANNs that rely on the use of a number of subsequent non-linear filtering units (layers). CNNs enable the creation of end-to-end models, since the raw data (e.g.: images or signals) are directly used as input, with no need of computing hand-crafted features: the hierarchical structure of the CNNs allows obtaining high-level features [27], thus transforming input data into a multi-dimensional representation useful to solve the classification task [28]. This is a key difference from conventional ML methods, which are based on relational data, where the features are manually defined by the user, based on a priori information.

Regarding the application of deep learning approaches to fNIRS signals, only one study addressed the classification of the signal quality [10], while other examples addressed task and gesture recognition for Brain-Computer-Interaction applications [29], [30]. The study of Gabrieli and colleagues [10] aimed at using a CNN based to classify the quality of 510 short fNIRS portions. Notably, the quality labels were collected by means of a web interface that experts used to rate the corpus of fNIRS signals. Their study was the first to demonstrate the use of CNNs for the classification of fNIRS signal quality, and used the Matthew Correlation Coefficient (MCC) to measure the classification performance. The proposed CNN achieved a performance of MCC=0.18 on the subset of data used for training and MCC=0.25 on the subset used for testing.

While ML approaches have shown promising results for addressing the SQC for fNIRS data, they have not been thoroughly researched and the literature on this topic is still sparse.

In this study, we conducted a detailed investigation of several aspects involved in the SQC of fNIRS data. First, we analyzed the role of human subjective evaluations, measuring the consensus of four different raters. Second, we tested the appropriateness of 5 of the most used hand-crafted SQI: first statistically, then by evaluating the performance of SQC based on SQI thresholds. Third, we explored the potential of using conventional machine learning and deep learning approaches for SQC; in particular, we assessed the performance of two conventional ML models trained on hand-crafted SQIs and of a CNN trained on raw signals. Finally, fourth, we applied model inspection techniques to explore the possibility of extracting knowledge from trained models, aiming at providing practical guidelines for the implementation of the SQC and optimization of the data acquisition settings.

II. Materials and Methods

A. Dataset

Data were collected during an experiment aimed at assessing the differential brain response in males and females to dialogues with sexist comments. The experiment was approved by the ethics committee of Nanyang Technological University Psychology Program (PSY-IRB-2020-007).

Eight experimental vignettes of hypothetical situations were constructed, each one lasting 50 seconds. The vignettes presented four scenarios in which a protagonist received sexist comments from four different partners, with two types of comments for each scenario (praise and criticism). The experiment involved 67 participants (38 females). Participants had to read all eight experimental vignettes; after reading, they were presented with a set of questions to measure their emotional responses towards each vignette.

During the experiment, fNIRS signals were collected to measure the activation of the dorsolateral prefrontal cortex. Signals were collected using a NIRS device (NIRSport, NIRx Medical Technologies LLC, Glen Head, NY, USA) equipped with a cap which mounted 8 light emitting diodes (760-850nm) and 7 photo-diodes detectors, composing a setup with 20 multi-distant channels (sampling rate: 7.81 Hz). Data were recorded using the NIRStar Software 15.0. In this study, we focused on the raw data collected by the photo-diodes for each channel, each composed of two signals, one for each wavelength. Signals were resampled at 10 Hz by cubic spline interpolation.

In total, the dataset included 1,340 channels signals (20 channels x 67 subjects). In order to proceed to the manual labeling of the signal quality, we restricted our sample to 548 segments relative to the vignette presentation (50 s length), which were randomly selected. A hierarchical random selection was performed, first by randomly selecting the condition, then the scenario, and finally the channel. The selected subset included data from 64 different subjects; the number of segments included for each subject was between 3 and 15 (Median=8, Mean=8.6, SD=2.8); the number of segments included for each channel was between 18 and 41 (Median=27, Mean=27.4, SD=5.2).

Similar to what was done in [10], 4 trained experts manually rated the quality of each segment, based on images of the fNIRS signals. Images had a size of 1500 × 1000 pixels, and a temporal resolution 23.6 pixels/s (see a scaled example of an image in Figure 1). The raters were asked to label the quality of each segment as Good (i.e.: the signal can be used) or Bad (i.e.: the signal should be discarded). The ratings of the 4 raters were aggregated by majority vote, to obtain the final label of each portion. In the case of ties, the segment was considered as having a Bad signal quality.

The dataset with 548 labeled portions was then randomly split into two separate partitions, to test the generalizability of each SQC approach: 75% (N = 411) of the portions were assigned to the Train partition, the remaining 25% (N = 137) was assigned to the Test partition. The proportion of signals of Good class was 66.2% on Train and 62.0% on Test.

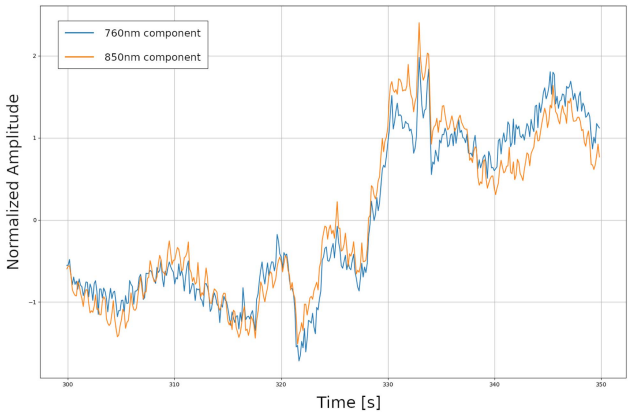

Fig. 1. Example of a 50 seconds length portion of fNIRS data, with the signals associated with the two wavelengths (760nm and 850nm), which was used to manually rate the quality of the signal.

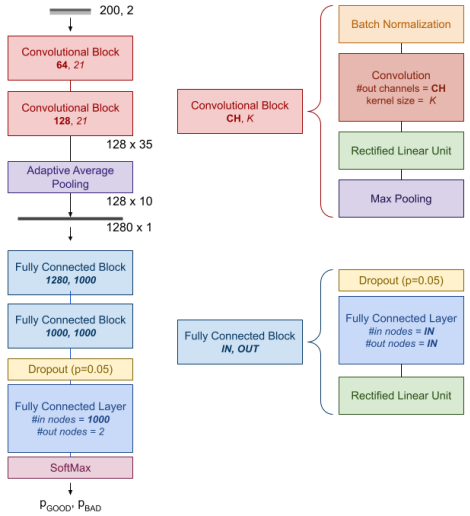

Fig. 2. Diagram of the convolutional neural network, composed of a sequence of convolutional blocks, followed by fully connected blocks.

B. Signal Quality Indicators

Among the most used SQIs for fNIRS signals, 4 were selected and used in this study: Scalp Coupling (SC) and Scalp Coupling Power (SCP) [31], Coefficient of Variation (CV) and the Coefficient of Variation of the Wavelengths (CVW) [12]. For each SQI, the literature also defines thresholds that are typically applied in automated pipelines to categorize the quality of signals. Typically, good quality signals are expected to have: $SC > 0.7$, $SCP > 0.1$, $CV < 7.5$, and $CVW < 5$.

Additionally, we also computed the Cardiac Power (CP), which, similarly to the SC, aims at quantifying the presence of the cardiac components. Starting from the filtered fNIRS signals (bandpass filter: [0.83 - 2.5] Hz), we estimated the cardiac frequency (f_c) as the frequency with highest power in the range 0.83-2.5 Hz. Then we computed the CP as the ratio between the power in the $f_c - 0.2$ - $f_c + 0.2$ Hz band and the $f_c - 0.5$ - $f_c + 0.5$ Hz band. Signals of the Good class were expected to have: $CP \geq 0.5$.

C. Machine Learning

Two conventional Machine Learning approaches have been adopted to classify the quality of the fNIRS signals: in the first, we tested two standard models based on the five SQIs: a Support Vector Machine with linear kernel (SVM) and a Random Forest (RF); in the second, we used an end-to-end Convolutional Neural Network (CNN) which was directly applied to the raw signals.

To train the standard models, we first optimized the model parameters: the regularization parameter C (C: 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000) and the number of trees n (n: 1, 5, 10, 50, 100, 250), for the SVM and RF respectively. To prevent overfitting, the following parameters of the RF model were also calibrated: the maximum depth of the trees was set to 3 and the minimum number of samples for leaf nodes was set to 10.

The optimization was based on a traditional 10 × 5-fold Cross Validation scheme [32]. The Train partition was randomly split into 5 folds: all folds except one were used to train the model which was then evaluated on the left-out fold. The procedure was iterated on the 5 folds, then repeated 10 times, shuffling the data before each repetition.

The performance of the model for each value of the model parameter was estimated by bootstrapping the distribution of the Matthew Correlation Coefficient (MCC) scores on the left-out fold at each iteration. The MCC was computed as follows:

$$MCC = \frac{(T_B T_G - F_B F_G)}{\sqrt{(T_B + F_B)(F_G + T_G)(T_B + F_G)(F_B + T_G)}} \quad (1)$$

where T_B and T_G are the number of segments correctly assigned to the Bad and Good signal quality class respectively, and F_B and F_G are the number of segments wrongly assigned to the Bad and Good signal quality class respectively.

The value of the model parameter with the higher MCC was selected as the optimal value. The final model was trained on the whole Train partition, using the optimal value of the model parameter.

The architecture of the CNN here employed is inspired by the architecture introduced by Bizzego and colleagues [33]. The original architecture, trained on the dataset of fNIRS signals (from the Train partition), achieved an MCC of.648 and.622 on the Train and Test partitions respectively. We conducted additional experiments to evaluate alternative network architectures. In particular, we tested different solutions, which differed in the number of convolutional blocks (from 2 to 4), output channels (from 16 to 256) and kernel sizes (from 3 to 21). Results of the architecture optimization procedure are reported in Supplementary Material (Table S1). The architecture which achieved the best results was the one composed of two sequential components: (i) a Convolutional Branch, and (ii) a Fully Connected Head (see Figure 2).

The Convolutional Branch consists of two convolutional blocks, each one composed of a 1-dimensional convolutional layer with the kernel size set to 21, a 1-dimensional batch normalization layer [34], a Rectified Linear Unit (ReLU) activation layer [35], and, finally, a 1-dimensional pooling

layer based on maximum, with the kernel size set to 2. The two blocks use a different number of channels; the first block has 2 input channels (the signals of the two wavelengths) and 64 output channels, the second block has 64 input channels and 128 output channels. The final layer of the Convolutional Branch is an adaptive averaging pooling layer, used to compute the average of the convolution operations at 10-time points.

The Fully Connected Head consists of three linear layers, each one preceded by a dropout layer (dropout probability set to 0.05), and followed by an activation layer. The first linear layer has 1280 input nodes and 1000 output nodes, and is followed by a ReLU activation layer; similarly, the second linear layer has 1000 input and output nodes and is followed by a ReLU activation layer. The final linear layer has 1000 input nodes and 2 output nodes, followed by a SoftMax activation layer, used to compute the class probability.

The input of the network is a 20 seconds length portion of a fNIRS signal. During training, the 20 seconds portion is randomly selected within the 50 seconds corresponding to the vignette presentation; during the evaluation, the 20 seconds portion correspond to the central part of the vignette presentation. The random selection of the portion during training is used to perform a data augmentation: each time the signal is used for training, a different portion is selected. This procedure adds a certain amount of stochastic variability to the input of the network, thus reducing the risk of overfitting.

The training was performed with back-propagation to minimize the Weighted Cross-Entropy between the true and predicted class [36]. Since the Bad and Good classes had a different sample size, weights were set to 0.66 for the Good class and to 0.34 for the Bad class.

The network was trained for 200 epochs on signals from the Train partition, with random batches of size 64, an Adadelta optimizer [37] and an initial learning rate of 0.01. In each epoch, signals in the Train partition were randomly shuffled and grouped into batches of 64 signals. Each batch was processed by the network to output the predicted class probability, which was then compared with the true class. Prediction error was computed using the Weighted Cross-Entropy. The error is then back-propagated to and weights of the networks are optimized using the Adadelta algorithm.

D. Analysis Plan

The study was divided into four separate analyses, one for each separate aim.

Analysis 1: Role of Human Subjective Evaluations on the Signal Quality: Since a ground truth about signal quality of fNIRS is missing, the reference is based on human subjective evaluations. All subsequent efforts to define efficient SQI and their thresholds, or to develop automated approaches rely on the capability of humans to provide a reliable assessment.

We quantified this capability in terms of consensus between the four raters, and in terms of performance of each rater. The consensus was quantified based on the two-way random effects model average Intraclass Correlation (ICC) [38]. The performance of each rater was quantified by computing the MCC score between the ratings of the rater and those from the other raters (aggregated by majority vote).

Analysis 2: Validity of Hand-Crafted SQI: We then focused on assessing the appropriateness of the 5 SQI. First, a two-tailed Mann-Whitney test was performed for each SQI, to assess if there are statistical differences between Good and Bad signals. We expected all SQI to show a significant result ($\alpha < 0.05$).

Additionally, we evaluated the Spearman correlation between the SQIs. We expected a high correlation ($\rho > 0.7$) for the SC, SCP and CP indicators, as they target the quantification of the cardiac components in the raw fNIRS signals. Second, we performed the SQC based on the value of the SQIs and their thresholds, which is the most common approach to identify signals with good and bad quality in current research practices. In practice, for each fNIRS segment, we tested that each SQI's value lied within the intervals associated with good quality, and considered the segment as having a good quality if all five SQIs had values that were within their respective intervals.

Analysis 3: Machine Learning Approaches: Finally we explored the use of Machine Learning approaches as an alternative to current SQC practices. Both the SVM model and the CNN were trained on data from the Train partition.

To allow an objective comparison between the different SQC methods (human raters, SQI thresholding, SVM model, and CNN) the classification performance was always computed on both the Train and Test partitions. Specifically, we used bootstrapping to generate the overall MCC with 90% Confidence Intervals (90%CI). In the bootstrap procedure, we randomly selected 25% of the samples, with replacement, and computed the MCC score on the selected subset; then repeated the procedure 1000 times. The overall MCC with 90%CI were computed as the 50^{th}, 5^{th} and 95^{th} percentiles respectively, of the generated distribution of MCC scores.

From the experimental point of view, data are a precious resource, which is often collected with high costs and efforts. Therefore, it is important that the SQC avoids rejecting data with good quality, which would represent a waste of resources that could instead be used for the study. On the other hand, the SQC should avoid that bad quality data contaminate the downstream analysis. Therefore, while the MCC gives a general indication of how well the SQC method performs, we also computed the Sensitivity and the Precision scores, which give additional insights. In our study, Sensitivity is defined as the ratio between the number of correctly classified Good segments and the total number of Good segments, indicating of how well each SQC method is able to avoid wasting Good quality data. Precision is the ratio between the number of correctly classified Good segments and the total number of segments classified with Good quality, indicating how well each SQC is able to avoid the contamination of the analysis with Bad quality data.

The distributions of MCC values generated by bootstrap from the Test partition were used to compare the performance of each SQC approach (thresholding of SQI values, SVM model, RF model, and DL model), using pair-wise t-tests. The Bonferroni correction was applied to correct for the multiple comparisons.

Analysis 4: Model Inspection: Aiming at providing practical guidelines for the implementation of the SQC and optimization of the data acquisition settings, we applied two model inspection techniques to extract knowledge from the trained ML models. For the two conventional ML models (SVM and RF), we computed the ranking of the SQIs, to obtain information about which SQIs are the most important for the prediction of the signal quality. For the DL model, we performed an unsupervised exploration of the output nodes of the Convolutional Branch.

To obtain the ranking of the SQIs, we first computed the permutation importance of the SQIs based on the trained model and data from the Test partition. The permutation importance of a SQI was computed as the decrease in the MCC score after the values of such SQI have been randomly shuffled [39]. In our implementation, the permutation importance is computed 30 times, then averaged, to determine the ranking of the SQIs.

The unsupervised exploration of 1280 output nodes of the Convolutional Branch was performed using the Uniform Manifold Approximation and Projection (UMAP) multidimensional projection method [40], [41]. Specifically, a two dimensional UMAP was applied, to facilitate the visualization of the results of the projection. We then aimed at investigating the main differences between the low-quality signals, to identify key diagnostic patterns that could suggest strategies to improve the signal quality during the experimental setup. We applied a K-means clustering on the UMAP projection, selecting the optimal number of clusters using the elbow method based on the sum of squared distances between each datapoint and its closest centroid. We then qualitatively analysed the main characteristics of each cluster in terms of signal patterns.

E. Data and Code Availability

The analyses performed in this study were implemented in Python (v. 3.8.10). The Machine Learning pipelines were built using the Numpy ([42], v. 1.19.4), Pandas ([43], v. 1.1.4), scikit-learn ([44], v. 0.23.2) and pyTorch ([45], v. 1.9.0+cu102) packages.

The UMAP and clustering pipelines were built using the umap-learn ([46], v. 0.5.3), scikit-learn ([44], v. 0.23.2), and yellowbrick ([47], v. 1.4) packages. Data used for this study and code to replicate the analysis are available at: https://gitlab.com/abp-san-public/fnirs-qsi-ml

III. Results

A. Consensus and Performance of Human Raters

The consensus between human raters was ICC = 0.774 ($p < .001$), which is considered acceptable, although optimal values are typically above 0.9 [48]. The overall MCCs for the four raters were in the range [0.629 - 0.759] on Train and [0.672 - 0.802] on Test (see Table I).

Notably, raters appear to maximize either Sensitivity or Precision (Figure 3). Rater1 maximized Sensitivity over Precision, achieving a Precision of 0.865 on Train and of 0.840 on Test (90%CI: [0.80-0.92] and [0.71-0.96] respectively), and a Sensitivity of 1.000 on both Train and Test. Similarly, Rater3 achieved a Precision of 0.840 on Train and of 0.800 on Test (90%CI: [0.77-0.90] and [0.67-0.92] respectively), and a Sensitivity of 1.000 on both Train and Test. On the opposite, Rater2 maximized Precision over Sensitivity, achieving a Precision 0.979 on Train and 1.000 on Test (90%CI on Train: [0.93-1.00]), and a Sensitivity of 0.706 on Train and 0.750 on Test (90%CI: [0.61-0.79] and [0.57-0.89] respectively).

TABLE I
MATTHEW CORRELATION COEFFICIENT (MCC) SCORES FOR THE DIFFERENT RATERS AND SIGNAL QUALITY CONTROL METHODS, ON TRAIN AND TEST, WITH 90% CONFIDENCE INTERVALS (90%CI) ESTIMATED BY BOOTSTRAPPING. SVM: SUPPORT VECTOR MACHINE, RF: RANDOM FOREST, CNN: CONVOLUTIONAL NEURAL NETWORK

		MCC [90% CI]			
Input data	Method	Train		Test	
Signals	Rater1	0.759	[0.66 - 0.86]	0.751	[0.57 - 0.93]
Signals	Rater2	0.629	[0.52 - 0.73]	0.734	[0.57 - 0.89]
Signals	Rater3	0.703	[0.60 - 0.81]	0.672	[0.50 - 0.86]
Signals	Rater4	0.715	[0.59 - 0.83]	0.802	[0.62 - 0.94]
SQIs	Thresholding	0.476	[0.38 - 0.58]	0.533	[0.38 - 0.70]
SQIs	SVM	0.671	[0.53 - 0.79]	0.717	[0.51 - 0.89]
SQIs	RF	0.712	[0.59 - 0.82]	0.722	[0.52 - 0.89]
Signals	CNN	0.726	[0.60 - 0.84]	0.757	[0.53 - 0.94]

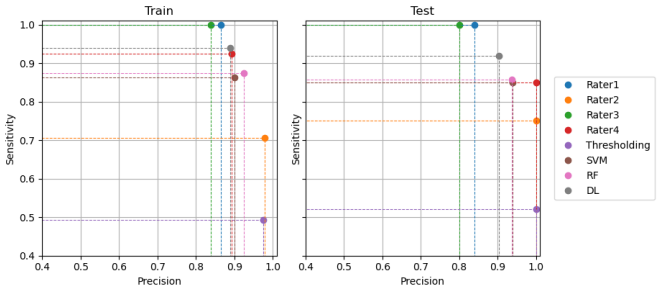

Fig. 3. Sensitivity and Precision scores for the different raters and signal quality control methods, on train and test.

TABLE II
MEDIAN VALUES FOR BAD AND GOOD QUALITY SIGNALS FOR THE FIVE SIGNAL QUALITY INDICATORS CONSIDERED IN THIS STUDY, WITH RESULTS OF THE MANN-WHITNEY TEST

	Median		Mann-Whitney test	
Signal Quality Indicator	Bad	Good	U	p
Scalp Coupling	0.045	0.869	8722	<0.001
Scalp Coupling Power	0.021	0.311	8168	<0.001
Cardiac Power	0.412	0.7	13792	<0.001
Wavelenght Coefficient of Variation	1.016	0.426	46652	<0.001
Coefficient of Variation	7.397	1.62	60121	<0.001

B. Validation of Hand-Crafted SQI

Results from Mann-Whitney Tests (Table II) indicate that the distribution of SQIs is significantly different for signals with Good or Bad signal quality.

In addition, all SQIs are significantly correlated (Table III, with the highest correlation observed between SC and SCP ($\rho = 0.80$, $p < .001$), the second being the correlation between SCP and CP ($\rho = 0.72$, $p < .001$).

TABLE III
VALUES OF THE SPEARMAN CORRELATION BETWEEN THE FIVE SIGNAL QUALITY INDICATORS CONSIDERED IN THIS STUDY ***: $p < .001$

	SC	SCP	CV	CVW
SCP	0.80***	-		
CV	-0.52***	-0.65***	-	
CVW	-0.21***	-0.25***	0.53***	-
CP	0.57***	0.72***	-0.57***	-0.15***

TABLE IV
CONFUSION MATRIX OF THE SQC BASED ON THRESHOLDING OF SQI. VALUES IN BOLD INDICATE THE NUMBER OF SEGMENTS WITH GOOD QUALITY THAT WOULD BE REJECTED

	Predicted			
	Train		Test	
Class	Bad	Good	Bad	Good
Bad	136	3	52	0
Good	**138**	134	**41**	44

The SQC method based on SQI thresholds (Table I and Figure 3) achieved an MCC of 0.476 on Train and 0.533 on Test (90%CI: [0.38-0.58] and [0.38-0.70] respectively), showing high Precision (Train: 0.976, 90%CI: [0.93-1.00]; Test 1.00), but low Sensitivity (Train: 0.493, 90%CI: [0.40-0.60]; Test 0.520, 90%CI: [0.33-0.69]). With this method, many signals with good quality would not be used (Table IV).

C. Machine Learning

The performance achieved by the conventional ML models was comparable. The SVM model (optimal C = 100) based on the SQI achieved an MCC of 0.671 on Train and 0.717 on Test (90%CI: [0.53-0.79] and [0.51-0.89] respectively), in line with the performance of human raters (Table I).

The RF model (optimal number of trees = 100) based on the SQI achieved an MCC of 0.712 on Train and 0.722 on Test (90%CI: [0.59-0.82] and [0.52-0.89] respectively). The results on both the Train and Test partitions are comparable with the SVM model (Table I).

The application of DL models allowed a further improvement. The CNN applied on the raw signals achieved an MCC of 0.726 on Train and 0.757 on test (90%CI: [0.60 - 0.84] and [0.53 - 0.94] respectively). The CNN achieved the better performance among all ML approaches, corresponding to a Precision of 0.890 (90%CI: [0.53-0.79]) on Train and 0.900 (90%CI: [0.77-1.00]) on Test, and a Sensitivity of 0.938 (90%CI: [0.88-0.97]) on Train and 0.917 (90%CI: [0.81-1.00]) on Test (Table V and Figure 3).

Performances achieved with the CNN (M=.752, SD=.121) resulted significantly better than both the conventional ML models: SVM (M=.716, SD=.126, t(1998)=6.37, $p < .001$) and RF (M=.713, SD=.127, t(1998)=7.15, $p < .001$). No difference was found between the SVM and the RF approach (t(1998)=−1.68, $p = .092$). In turn, the SVM model achieved significantly better performances than the thresholding of the value of the SQIs (M=.537, SD=.094, t(1998)=35.28, $p < .001$).

TABLE V
CONFUSION MATRIX OF THE SQC BASED ON THE CONVOLUTION NEURAL NETWORK APPLIED ON RAW SIGNALS

	Predicted			
	Train		Test	
Class	Bad	Good	Bad	Good
Bad	107	32	43	9
Good	18	254	7	78

D. Model Inspection

The rankings of the SQIs based on the permutation importance was slightly different for the two models. For the SVM model, the ranking was, in order: SC, CV, CP, SCP, and CVW. For the RF model, the ranking was, in order: CV, SC, CVW, SCP and CP. Notably, the two top features were the same for both models: SC and CV. This would suggest that both an indicator of the cardiac component and a global indicator of the signal variability are required to asses the signal quality. Future implementations of tools and algorithms for the automated or semi-automated identification of the signal quality should then consider these two key SQIs.

The two dimensional embedding of the 1280 nodes computed with the UMAP (Figure 4A) shows a clear separation between the two classes, with the low-quality signals more tightly clustered. We then applied the K-means clustering algorithm, with a target number of clusters equal to three, which resulted in the optimal number of clusters according to the elbow method.

The first cluster (Figure 4B) mainly includes low quality signals that overlap with the group of good quality signals. By observing some randomly chosen examples of signals belonging to this cluster (Figure 5), we recognize that these signals have in general a good quality except for the presence of spikes or drops, probably due to movement, the presence of which is hard to foresee during the setup of the instrumentation.

The second and third cluster appear to split the group of low quality signals, with the second group "topologically" nearer to the good-quality cluster. The two clusters seems to differ in terms of magnitude of the noise and distance between the average of the two components. In addition, we note that the cardiac component and some other signal components might, in some cases, be recognized in signals from the second cluster: thus suggesting that the reason of the low quality is mainly a poor coupling between the optodes and the scalp, or an interference from external sources of light. On the opposite, it is hard to identify any component in signals in the third cluster except for white noise, thus suggesting that the setup of the optodes should be thoroughly revised. In general, these two clusters suggest a trajectory in the UMAP associated with a decrease of signal quality, as long as we move far from the group of good-quality signal.

IV. DISCUSSION

The results from the analysis of consensus between human raters and their performances highlight that the outcomes of visual inspection of the signals are highly dependent on subjective evaluations. The consensus between the four raters

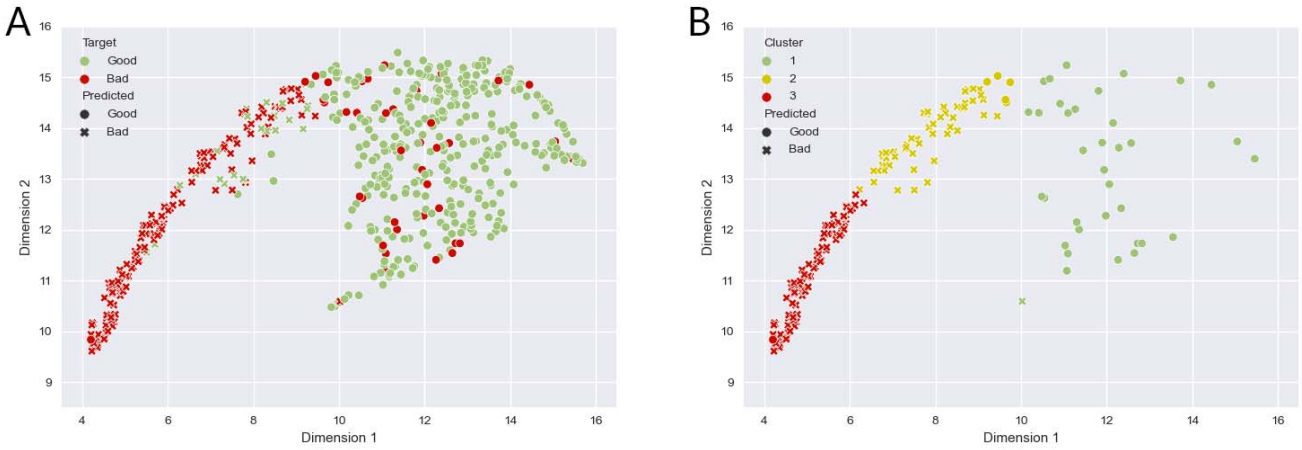

Fig. 4. Results of the Uniform Manifold Approximation and Projection algorithm. The marker indicates the class assigned by the Convolutional Neural Network: circle: Good class, Cross: Bad class. A: Results for the whole dataset. The color of the datapoints indicates the true class: in red the Bad quality signals, in green the good quality signals. B: Results of the K-means clustering of Bad quality signals, colored by cluster.

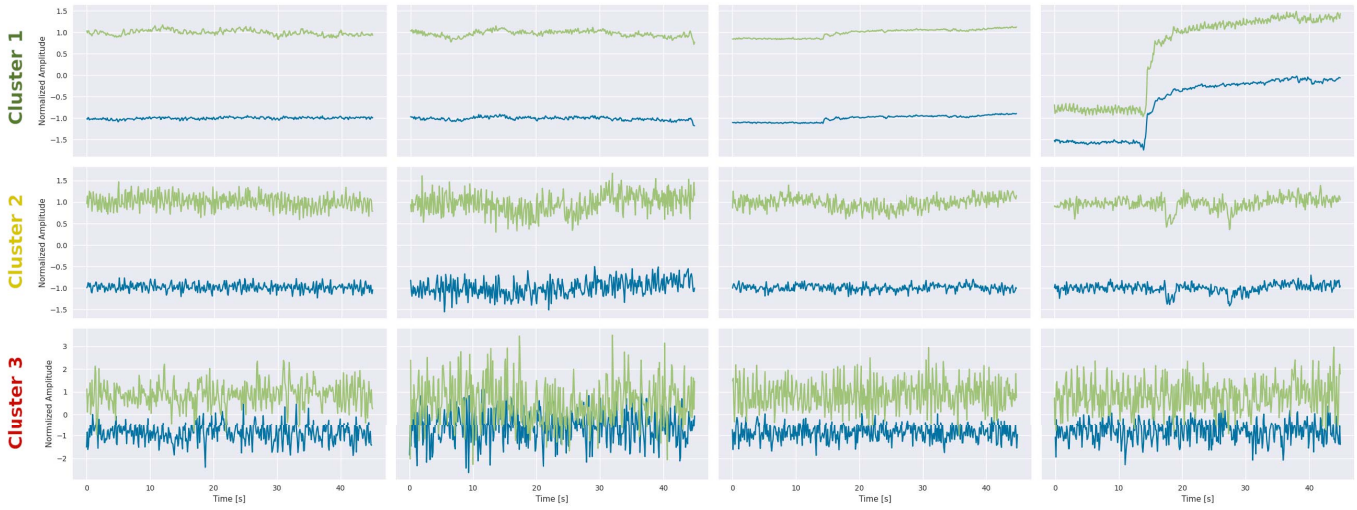

Fig. 5. Examples of signals randomly selected from each cluster. Blue: 760nm component; Green: 850nm component. Top: signals from cluster 1; Middle: signals from cluster 2; Bottom: signals from cluster 3.

(ICC=0.774) was acceptable - although not optimal [48], but the individual performances showed high variability. It must be noted that the raters involved in this study belong to the same research group: it can be expected that all shared a knowledge base about the quality of fNIRS data. Future research should investigate the consensus between raters coming from different laboratories, and research initiatives such as the many labs [49] or large scale studies would be needed to provide a better insight. These results also highlight the urgency of providing the research community with more reference datasets (e.g. [50]) that could be used for both training new scholars and researchers and evaluating and comparing new algorithms and procedures [51]. The creation of such a dataset would require a shared effort, as contributions from multiple centers are needed to overcome the limitation due to the scarce consensus between raters.

As an alternative to visual inspection, hand-crafted SQI are still commonly used in the research practice. The advantage of SQI over visual inspection, is that they are not influenced by subjective evaluations. However, the results from this study highlight that if used in combination with threshold-based rejection approaches, they have low sensitivity toward Good quality signal: they tend to be much too conservative, thus increasing the costs of the data acquisition, due to the rejection of many usable signals. In this study, we considered five SQI and proved that all are statistically valid in discriminating between Good and Bad quality signals. Typically, SQI-based SQC procedures consider three or less SQI only [8], [12], [31]: we note that the choice of SQI to consider, and respective thresholds, might be, again, influenced by subjective preferences or empirical rules, thus representing another source of variability in the fNIRS signal processing pipeline [5].

Although future research could focus on identifying an optimal combination of SQIs and on the determination of more appropriate thresholds, our study showed that the full potential of SQI could be exploited if they are used in

combination with machine learning approaches. The use of a SVM or RF instead of manually-defined thresholds allowed the improvement of the performance of the SQC; in particular, in terms of Sensitivity, which increased from 0.493/0.520 on Train/Test for the threshold-based method, to 0.864/0.850 on Train/Test for the SVM. We note, however, that the limited sample size ($N_{train} = 411$) and relatively shallow architecture might have prevented us from fully exploiting the potential of ANN approaches.

Finally, we showed that ANNs are a promising technique to additionally improve the performance of automated SQC procedures: the CNN achieved the best performance among all SQC methods considered in this study (MCC=0.726/0.757 on Train/Test). Although targeting a different dataset, this study improves over previous results reported by Gabrieli and colleagues [10] which achieved a MCC=0.18/0.25 on Train/test. The main advantages of DL methods are that, as shown in other fields, (a) they are more effective predictive models in general, and (b) they can be applied using an end-to-end approach. The end-to-end approach requires no additional signal processing steps, thus simplifying the signal processing pipeline, while removing a potential source of variability.

An open issue with DL approaches is the explainability and interpretability of models [52], [53]. Explainability in the context of SQC, could be useful as a diagnostic tool during the setting up of the experimental equipment. In our study, we implemented two model inspection techniques that allow some insight in the internal functioning of the models, focusing on the internal representation of the features. However, further research is needed to extract from the model the information to discriminate between possible causes of Bad quality data (e.g.: physiological noise, sub-optimal scalp-sensor coupling, technical issues), and provide practical indications.

V. Conclusion

This study investigated several aspects involved in the SQC of fNIRS data, aiming at identifying the open issues and opportunities toward the development of fully automated and reliable SQC procedures.

We highlighted the role of subjectivity in the assessment of the quality of the signals based on visual inspection, measuring the consensus and performance of four different raters. Then we evaluated the use of hand-crafted SQI for the classification of signal quality, showing the superiority of machine learning models (SVM) over threshold-based approaches. Finally, we explored the potential of using DL approaches, using a CNN that was directly applies on raw signals.

Overall, this study highlighted that the main limitation towards automated SQC is probably the lack of consensus between human raters. In fact, the CNN achieved a performance that was in line with that of human raters; furthermore, some raters showed lower 90%CI ranges than the CNN itself. This study suggests that the computational methods already available are appropriate to define reliable SQC procedures, and the main obstacle towards this development seems to be the lack of reference dataset with high consensus labels. The efforts of the scientific community should therefore be directed towards the creation of a common knowledge base and shared resources, under the principle of Open Science.

References

[1] M. A. Yücel *et al.*, "Best practices for fNIRS publications," *Neurophotonics*, vol. 8, no. 1, Jan. 2021, Art. no. 012101.

[2] A. Azhari *et al.*, "A decade of infant neuroimaging research: What have we learned and where are we going?" *Infant Behav. Develop.*, vol. 58, Feb. 2020, Art. no. 101389.

[3] P. Pinti *et al.*, "The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience," *Ann. New York Acad. Sci.*, vol. 1464, no. 1, pp. 5–29, 2020.

[4] P. Pinti, F. Scholkmann, A. Hamilton, P. Burgess, and I. Tachtsidis, "Current status and issues regarding pre-processing of fNIRS neuroimaging data: An investigation of diverse signal filtering methods within a general linear model framework," *Frontiers Hum. Neurosci.*, vol. 12, p. 505, Jan. 2019.

[5] A. Bizzego, J. P. M. Balagtas, and G. Esposito, "Commentary: Current status and issues regarding pre-processing of fNIRS neuroimaging data: An investigation of diverse signal filtering methods within a general linear model framework," *Frontiers Hum. Neurosci.*, vol. 14, p. 247, Jul. 2020.

[6] L. Hocke, I. Oni, C. Duszynski, A. Corrigan, B. Frederick, and J. Dunn, "Automated processing of fNIRS data—A visual guide to the pitfalls and consequences," *Algorithms*, vol. 11, no. 5, p. 67, May 2018.

[7] F. Orihuela-Espina, D. R. Leff, D. R. C. James, A. W. Darzi, and G. Z. Yang, "Quality control and assurance in functional near infrared spectroscopy (fNIRS) experimentation," *Phys. Med. Biol.*, vol. 55, no. 13, p. 3701, Jun. 2010.

[8] A. Azhari *et al.*, "Parenting stress undermines mother-child brain-to-brain synchrony: A hyperscanning study," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.

[9] A. Azhari, M. Lim, A. Bizzego, G. Gabrieli, M. H. Bornstein, and G. Esposito, "Physical presence of spouse enhances brain-to-brain synchrony in co-parenting couples," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, Dec. 2020.

[10] G. Gabrieli, A. Bizzego, M. J. Y. Neoh, and G. Esposito, "FNIRS-QC: Crowd-sourced creation of a dataset and machine learning model for fNIRS quality control," *Appl. Sci.*, vol. 11, no. 20, p. 9531, Oct. 2021.

[11] L. Pollonini, C. Olds, H. Abaya, H. Bortfeld, M. S. Beauchamp, and J. S. Oghalai, "Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy," *Hearing Res.*, vol. 309, pp. 84–93, Mar. 2014.

[12] S. Lloyd-Fox, A. Blasi, A. Volein, N. Everdell, C. E. Elwell, and M. H. Johnson, "Social perception in infancy: A near infrared spectroscopy study," *Child Develop.*, vol. 80, no. 4, pp. 986–999, Jul. 2009.

[13] M. S. Sappia, N. Hakimi, W. N. Colier, and J. M. Horschig, "Signal quality index: An algorithm for quantitative assessment of functional near infrared spectroscopy signal quality," *Biomed. Opt. Exp.*, vol. 11, no. 11, pp. 6732–6754, 2020.

[14] R. B. Govindan, A. N. Massaro, and A. du Plessis, "Ensuring signal quality of cerebral near infrared spectroscopy during continuous longterm monitoring," *J. Neurosci. Methods*, vol. 309, pp. 147–152, Nov. 2018.

[15] M. S. Zaman and B. I. Morshed, "Estimating reliability of signal quality of physiological data from data statistics itself for real-time wearables," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 5967–5970.

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

[17] Q. Li, C. Rajagopalan, and G. D. Clifford, "A machine learning approach to multi-level ECG signal quality classification," *Comput. Methods Programs Biomed.*, vol. 117, no. 3, pp. 435–447, Dec. 2014.

[18] G. Gabrieli, J. P. M. Balagtas, G. Esposito, and P. Setoh, "A machine learning approach for the automatic estimation of fixation-time data Signals' quality," *Sensors*, vol. 20, no. 23, p. 6775, Nov. 2020.

[19] M. S. Sappia, N. Hakimi, L. Svinkunaite, T. Alderliesten, J. M. Horschig, and N. W. Colier, "FNIRS signal quality estimation by means of a machine learning algorithm trained on morphological and temporal features," *Proc. SPIE*, vol. 11638, pp. 29–39, Mar. 2021.

[20] A. Manzalini, "Towards a quantum field theory for optical artificial intelligence," *Ann. Emerg. Technol. Comput.*, vol. 3, no. 3, pp. 281–2516, Jul. 2019.

[21] C. Sanchez-Sanchez, D. Izzo, and D. Hennes, "Learning the optimal state-feedback using deep networks," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2016, pp. 1–8.

[22] A. Bizzego *et al.*, "Integrating deep and radiomics features in cancer bioimaging," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Jul. 2019, pp. 1–8.

[23] H.-H. Tseng, L. Wei, S. Cui, Y. Luo, R. K. T. Haken, and I. El Naqa, "Machine learning and imaging informatics in oncology," *Oncology*, vol. 23, pp. 1–19, Nov. 2018.

[24] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Med.*, vol. 25, no. 1, p. 44, 2019.

[25] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[26] P. Mobadersany *et al.*, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 13, pp. E2970–E2979, Mar. 2018.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, Feb. 2015.

[28] A. K. Mukhopadhyay and S. Samui, "An experimental study on upper limb position invariant EMG signal classification based on deep neural network," *Biomed. Signal Process. Control*, vol. 55, Jan. 2020, Art. no. 101669.

[29] J. Hennrich, C. Herff, D. Heger, and T. Schultz, "Investigating deep learning for fNIRS based BCI," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 2844–2847.

[30] K. Khalil, U. Asgher, and Y. Ayaz, "Novel fNIRS study on homogeneous symmetric feature-based transfer learning for brain–computer interface," *Sci. Rep.*, vol. 12, no. 1, pp. 1–12, Dec. 2022.

[31] L. Pollonini, H. Bortfeld, and J. S. Oghalai, "Phoebe: A method for real time mapping of optodes-scalp coupling in functional near-infrared spectroscopy," *Biomed. Opt. Exp.*, vol. 7, no. 12, pp. 5104–5119, 2016.

[32] A. Bizzego *et al.*, "Predictors of contemporary under-5 child mortality in low-and middle-income countries: A machine learning approach," *Int. J. Environ. Res. Public Health*, vol. 18, no. 3, p. 1315, 2021.

[33] A. Bizzego, G. Gabrieli, and G. Esposito, "Deep neural networks and transfer learning on a multivariate physiological signal dataset," *Bioengineering*, vol. 8, no. 3, p. 35, Mar. 2021.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[35] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2217–2225.

[36] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 8778–8788.

[37] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*.

[38] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, p. 420, Mar. 1979.

[39] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[40] E. Becht *et al.*, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnol.*, vol. 37, no. 1, pp. 38–44, Jan. 2019.

[41] A. Bizzego *et al.*, "Evaluating reproducibility of AI algorithms in digital pathology with DAPPER," *PLOS Comput. Biol.*, vol. 15, no. 3, Mar. 2019, Art. no. e1006269.

[42] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: A structure for efficient numerical computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, 2011.

[43] W. McKinney *et al.*, "Pandas: A foundational Python library for data analysis and statistics," *Python High Perform. Sci. Comput.*, vol. 14, no. 9, pp. 1–9, 2011.

[44] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[45] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.

[46] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[47] B. Bengfort and R. Bilbro, "Yellowbrick: Visualizing the scikit-learn model selection process," *J. Open Source Softw.*, vol. 4, no. 35, p. 1075, Mar. 2019.

[48] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropractic Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016.

[49] R. A. Klein *et al.*, "Many labs 2: Investigating variation in replicability across samples and settings," *Adv. Methods Practices Psychol. Sci.*, vol. 1, no. 4, pp. 443–490, 2018.

[50] J. Shin, A. von Lühmann, D.-W. Kim, J. Mehnert, H.-J. Hwang, and K.-R. Müller, "Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset," *Sci. Data*, vol. 5, no. 1, pp. 1–16, Dec. 2018.

[51] R. A. Poldrack and K. J. Gorgolewski, "Making big data open: Data sharing in neuroimaging," *Nature Neurosci.*, vol. 17, no. 11, pp. 1510–1517, 2014.

[52] G. Ras, N. Xie, M. Van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *J. Artif. Intell. Res.*, vol. 73, pp. 329–397, Jan. 2022.

[53] G. Riccardo *et al.*, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.