applied sciences

MDPI

Article

fNIRS-QC: Crowd-Sourced Creation of a Dataset and Machine Learning Model for fNIRS Quality Control

Giulio Gabrieli [1,†], Andrea Bizzego [2,†], Michelle Jin Yee Neoh [1] and Gianluca Esposito [1,2,3,*]

1 Psychology Program, Nanyang Technological University, Singapore 639818, Singapore; giulio001@e.ntu.edu.sg (G.G.); MICHELLE008@e.ntu.edu.sg (M.J.Y.N.)
2 Department of Psychology and Cognitive Science, University of Trento, 38068 Trento, Italy; andrea.bizzego@unitn.it
3 Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 308232, Singapore
* Correspondence: gianluca.esposito@ntu.edu.sg
† These authors contributed equally to this work.

Featured Application: Our dataset can be used to train novel Machine Learning and Artificial Intelligence models to automatically identify the quality of fNIRS signals.

Abstract: Despite technological advancements in functional Near Infra-Red Spectroscopy (fNIRS) and a rise in the application of the fNIRS in neuroscience experimental designs, the processing of fNIRS data remains characterized by a high number of heterogeneous approaches, implicating the scientific reproducibility and interpretability of the results. For example, a manual inspection is still necessary to assess the quality and subsequent retention of collected fNIRS signals for analysis. Machine Learning (ML) approaches are well-positioned to provide a unique contribution to fNIRS data processing by automating and standardizing methodological approaches for quality control, where ML models can produce objective and reproducible results. However, any successful ML application is grounded in a high-quality dataset of labeled training data, and unfortunately, no such dataset is currently available for fNIRS signals. In this work, we introduce fNIRS-QC, a platform designed for the crowd-sourced creation of a quality control fNIRS dataset. In particular, we (a) composed a dataset of 4385 fNIRS signals; (b) created a web interface to allow multiple users to manually label the signal quality of 510 10 s fNIRS segments. Finally, (c) a subset of the labeled dataset is used to develop a proof-of-concept ML model to automatically assess the quality of fNIRS signals. The developed ML models can serve as a more objective and efficient quality control check that minimizes error from manual inspection and the need for expertise with signal quality control.

Keywords: fNIRS; machine learning; quality control

Citation: Gabrieli, G.; Bizzego, A.; Neoh, M.J.Y.; Esposito, G. fNIRS-QC: Crowd-Sourced Creation of a Dataset and Machine Learning Model for fNIRS Quality Control. *Appl. Sci.* **2021**, *11*, 9531. https://doi.org/10.3390/app11209531

Academic Editor: Alexander E. Hramov

Received: 12 August 2021
Accepted: 11 October 2021
Published: 14 October 2021

1. Introduction

Functional near-infrared spectroscopy (fNIRS) is a non-invasive neuroimaging modality which allows the detection of cortical brain activity through the use of light in the near-infrared spectrum (650–900 nm). Due to the difference in absorption of oxygenated and deoxygenated hemoglobin, the fNIRS is able to measure the relative changes in the concentrations of oxygenated and deoxygenated hemoglobin which are indicative of cerebral activation and deactivation. In recent years, the use of fNIRS has seen rapid growth in neuroimaging studies [1], gaining traction in fields such as infant neuroimaging [2] and cognitive neuroscience [3].

Despite the burgeoning use of fNIRS, a general consensus or standardization of the best pre-processing practices for the NIRS signal has not been established, unlike other neuroimaging modalities such as functional magnetic resonance imaging (fMRI; see [4,5]). Differences in the use and combination of pre-processing pipelines have been demonstrated to lead to different results in fNIRS studies [6]. Hence, the absence of standardization

in pre-processing methods, analysis tools, and instrumentation can lead to the scarce reproducibility of studies and results, similar to what occurs with other neurophysiological signals (e.g., infant cry [7]).

One key pre-processing step in fNIRS data analysis is the signal quality check of the raw signals for each channel. The presence of a strong cardiac component is one of the main indicators of good optode-scalp coupling characterizing a high-quality fNIRS signal. Noise in the fNIRS signal is typically the result of (i) body or head movements which causes fast spikes or baseline shifts and physiological components such as cardiac, respiratory, and blood pressure components. Usually, a manual visual inspection is conducted to assess signal quality—for example, indicators such as the presence of large motion artifacts, heartbeat oscillations indicative of good coupling between the scalp and optodes. The nature of the manual visual inspection means that signal quality check is dependent on researcher expertise and subjective judgments on a "good" quality signal. Hence, the development of an objective signal quality check can address the issues of experience and subjectivity in the signal quality control step. Machine Learning algorithms have been proven to be effective in supporting researchers' classification of signal quality. Li and colleagues, for example, successfully employed Machine Learning for the automatic quality assessment of pulsatile signals [8] and for multi-level ECG signals [9], while Gabrieli et al. [10] tested the efficacy of different classifiers in the identification of the quality of pupillometry signals.

Currently, a number of algorithms based on morphological characteristics of the fNIRS signal have been proposed for signal quality assessment: (i) Scalp Coupling Index (SCI [11]), (ii) placing headgear optodes efficiently before experimentation (PHOEBE [12]), and (iii) signal quality index (SQI [13]). The SCI and PHOEBE are algorithms that binarily assign signals to "good" or "bad" categories based on the presence of the cardiac component in the signal. The SQI algorithm provides five levels of a quantitative rating of signal quality and was developed based on visual quality assessment markers used by experts in fNIRS.

These algorithms rely on a small number of human-defined signal quality indicators and empirical thresholds. However, deep learning approaches have the key characteristic of automatically extracting high-dimensional features and leverage on non-linear decision functions. They are therefore a promising approach, similar to other studies.

Machine learning involves the training of algorithms with known input-output pairs of the function. For what concerns fNIRS signals, several studies have recently employed different Machine Learning ad Deep Learning techniques to classify signals. Ortega and Faisal [14], for example, employed a deep learning classifier to decode the strength of hand movements in order to develop more accurate Brain-Computer Interfaces (BCI). Similarly, Ma et al. [15] developed a Deep Learning classifier to classify motion imagery of three different hand gestures. Deep Learning models have also been employed to assess and classify the mental workload of difficult tasks, such as driving [16] or memory tasks [17].

Concerning signals' quality estimation, a machine learning version of the SQI (MLSQI [18]) has been developed based on the training dataset described in Sappia et al. [13]. However, the training dataset was collected from only 14 participants and labeled by individuals working at the company that produces the fNIRS recording device used. The limited number of collected signals is a crucial limitation for the efficient application of a Deep Learning approach, which is however intrinsically connected with the fNIRS field for multiple reasons. First, the novelty of the field results in a reduced availability of large-scale fNIRS datasets that can be employed for secondary data analysis or for the development of novel tools and techniques. Secondly, the difficulty in obtaining data labeled by experts of the field, combined with the lack of a ground truth that determines the quality of a signal makes it impossible to obtain large-scale labeling of the quality of fNIRS signals. To overcome these limitations, in this paper, we introduce a crowd-sourced training dataset consisting of 510 10-s segments of single-channel fNIRS signals. Through crowdsourcing, we are able to leverage multiple fNIRS recordings from a wider range of participants. By making use of a web interface, we were able to reach out to more individuals with experience working with fNIRS and tapped onto their expertise in labeling the quality of

the segments through this interface. The labeled dataset is here used to train and test a machine learning model that can identify the quality of a signal, and that can therefore be used to support non-experts of the field that approach the fNIRS signals for the first time or to make pre-processing pipelines more objective, by introducing an objective way to identify segments of signals of high quality that can be used for further analysis.

Aim of This Study

This study aims at improving the quality control step of fNIRS studies by introducing an Artificial Intelligence framework that can support researchers in discriminating between usable and unusable fNIRS segments. Overall, this work brings three main contributions. First, an Open Source web interface that can be used to classify the quality of different signals using a crowd-science approach has been designed and developed. While here we employed it for the collection of fNIRS signals labels, a boilerplate of the platform that can be adapted for other signals or digital objects that require labeling has been made available, allowing other researchers to rapidly deploy citizen science platforms. The second contribution of this study is the creation of a reference dataset of fNIRS signals that can be used by researchers within the field to develop new tools for the preprocessing and analysis of fNIRS data and to train non-experts to discriminate between usable and unusable signals. Third, in this study, a proof-of-concept ML model that can support researchers by automatically assessing the quality of fNIRS signals is presented. The latter is of special interest considering the novelty of the field, the limited number of experts in the visual examination of the quality of fNIRS signals, and the increasing amount of young scholars with previous to no experience in fNIRS signal processing that may need support to evaluate the quality of recorded data. Overall, we believe that the created dataset and developed model favor a more objective and efficient approach to fNIRS quality control.

2. Materials and Methods

2.1. Dataset

The complete fNIRS dataset generated for this study consists of 4385 portions of single-channel fNIRS signals with a duration of 10 s each. In order to obtain a better time and space localization of signals' quality, short portions of signals were selected. To avoid biases introduced by different recording devices, all the signals have been collected using a NIRSport device (NIRx Medical Technologies LLC). This equipment has a scan rate of 7.81 Hz and employs LED emission with source wavelengths of 760 nm and 850 nm.

Signals included within this dataset were drawn from four different studies, and all belong to adult participants. The first study (Mother-Child Synchrony study) involves the simultaneous recording of fNIRS data from mothers and children engaging in a passive video viewing task [19]. Only data from mothers (N = 31, Mean Age = 34.9 $\pm$ 4.16 years) are selected for the dataset used in this work (for details on the experimental procedure, see [19]).

The second study (Father-Child synchrony study) consists of fNIRS recordings of Fathers and Children engaging in both a passive video viewing task and an active play task. Only data from fathers (N = 29, Mean Age = 38.1 $\pm$ 3.67) have been selected for the current dataset [20].

The third study (3-Love study) consists of the data of 69 participants (Mean Age = 21.21 $\pm$ 1.66) [21]. Participants were asked to watch three video clips depicting a couple interacting while baking, eating, and exercising. Before presentation, participants were informed (experimental manipulation) about the status of the couple, being either romantic partners, friends, or siblings.

Finally, the fourth study (Mother-Father synchrony study) consists of the recordings of both mothers and fathers while passively hearing audio stimuli of infants' and adults' vocalizations [22].

A breakdown of the number of signals per study is reported in Table 1.

Table 1. Breakdown of the number of signals per study of origin that have been included in the current dataset.

Study	Number of Signals
Mother-Child Synchrony	836
Father-Child Synchrony	2034
3-Love Study	1035
Mother-Father Synchrony	480
	4385

From the totality of the signals, 510 segments from the Mother-Child synchrony study were randomly selected for the current labeling stage. All the selected segments were resampled at 10 Hz, but no additional preprocessing step was conducted on the raw signals.

2.2. Web Interface

In order to obtain labels for our classifier, a web platform written in HTML5 for the human labeling of signals was designed and deployed on a shared hosting service. The web platform consists of a back-end, where signals and ratings are stored, and a front-end, which allows the users to rate the signals. when users register an account, they are asked to specify their level of expertise, which could be Beginner ("worked on less than 2 datasets (less than 100 fNIRS recordings processed"), Intermediate ("worked on 2 4 datasets (200 fNIRS recordings processed"), or Expert ("worked on more than 4 datasets (more than 200 fNIRS recordings processed)").

One randomly selected signal is presented each time and the user is asked to assign one out of three possible labels: Keep, Keep after correction, or Reject.

Visually, the User Interface presents colored buttons that can be used to rate the signals in the upper part of the screen—the button Reject is in red, the button Keep after correction is in yellow, and the button Keep is in green—followed by two rectangles, one above the other, in which the two waveform components (wavelength 1 and 2) of the signals are visually shown. The platform allows the user to zoom in on signals in order to obtain a closer view of peaks and fluctuations. Finally, on the lower part of the screen, the interface provides details about how many signals the user has rated. A screenshot of the interface is shown in Figure 1. Each user can rate as many signals as are present on the server in an anonymous way, with the only references to the user being an anonymous ID and the expertise of the user.

The web interface was used to collect ratings for a subset of 510 segments of the complete dataset, in order to develop the proof of concept of the automatic quality classification based on Deep Learning.

2.3. Collected Data and Processing

The subset of 510 segments used for the proof-of-concept were selected from the Mother-Child Synchrony study. A total of two thousand four hundred and one (N = 2401) ratings were collected: a breakdown of the ratings by user is reported in Table 2.

Each rating consists of three pieces of information: the label of the signal quality, the self-reported level of expertise of the rater, and the time required for the rater to assign a label to the signal (reaction time). No identifiable or demographic data of the raters are collected.

Overall, the Percent Agreement between self-described Expert fNIRS users is 62.4%, between Intermediate users is 14.4%, and between Beginner users is 39.4%, while the average percent agreement between expert and beginner users is 30.4%, and between expert and intermediate users is 29.3%. The average percent agreement between intermediate and beginner users is 33.9%.

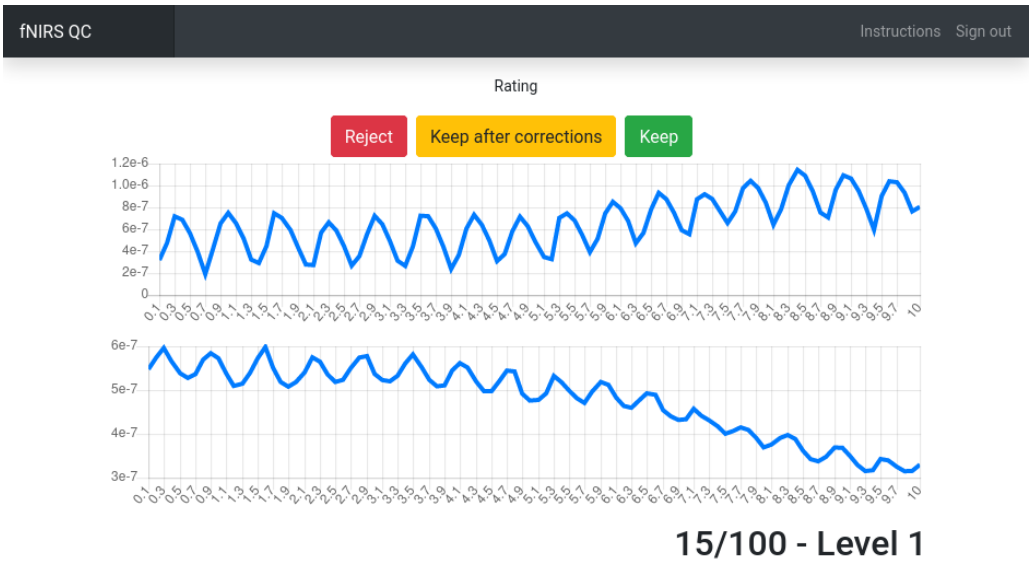

Figure 1. Screenshot of the fNIRS-QC web interface.

Table 2. Breakdown of the ratings by user (N = 9).

User ID	Expertise	Reject	Keep after Correction	Keep	Total
1	Intermediate	153	249	100	502
2	Intermediate	5	7	18	30
3	Expert	52	51	33	136
4	Intermediate	8	81	11	100
5	Intermediate	5	95	0	100
6	Expert	18	6	8	32
7	Intermediate	188	281	33	502
8	Beginner	123	251	125	499
9	Beginner	319	57	124	500
		871	1078	452	2401

The time required for the rating was used to compute the confidence weight (w_c) for each rating. All rating times of each user were assigned to four confidence levels, based on thresholds corresponding to the 25th, 50th, and 75th percentile of the distribution of the rating times. Ratings below the 25th percentile were associated with a high confidence and assigned a $w_c = 1$; similarly, other levels were associated with lower confidence and assigned a $w_c = 0.75$ (25th to 50th percentile), $w_c = 0.5$ (50th to 75th percentile), and $w_c = 0.25$ (greater than 75th percentile). The self-reported expertise was also used to assign and experience weight (w_e) to each user. Users self-reported as "Expert" were assigned a $w_e = 1$, "Intermediate" users were assigned a $w_e = 0.66$ and "Beginner" users were assigned a $w_e = 0.33$

The ratings and the weight were used to compute the class of each segment in the dataset. The three labels correspond to three different quality levels (q) of the signals:

- Accept. The presented segment of a signal has a good quality, that is deemed acceptable by the user. This class was assigned a $q = 3$.
- Keep after correction. The portion is affected by noise or artifacts (e.g., spikes), but after applying appropriate signal processing methods to increase the signal to noise ratio and remove artifacts, the portion can likely be used for further analysis. This class was assigned a $q = 2$.
- Reject. The portion is very noisy or affected by artifacts that cannot be corrected using standard signal processing techniques. This class was assigned a $q = 1$.

Three different methods were adopted to aggregate the ratings from different users and compute a unique quality level Q for each segment. The first (Q_m) method was simply the majority vote, in which the most voted quality level q was assigned.

Other methods were based on weights. First we computed the sum q_k of the weights w_i of the ratings r_i by quality level k (see Equation (1)), where $\hat{w}_i^k = w_i$ if $r_i = k$; $\hat{w}_i^k = 0$ otherwise.

$$q_k = \sum_i \hat{w}_i^k \tag{1}$$

Then we selected the level k corresponding to the maximum q_k.

Thus, we computed the experience-weighted aggregated quality level (Q_e) based on the experience weights and the confidence-weighted aggregated quality level (Q_c) based on the confidence weights.

In case of ties, the lower class was assigned.

A breakdown of the labels assigned to the segment, given the aggregation method is provided in Table 3, while a visual representation of the users' response time by rating is shown in Figure 2.

For the proof of concept DL model, we referred to the Q_m quality levels. Since the classes were highly unbalanced (with only N = 19 segments for the "Keep" class, we focused on the binary classification of "Reject" class versus the others.

2.4. Deep Learning Experiments

The architecture of the Deep Neural Network (DNN) here employed is based on the architecture described by Bizzego et al. [23], and consists of three sequential components: (i) a Convolutional Branch; (ii) a Long Short-Term Memory (LSTM) module; a Fully Connected Head (FCH). The Convolutional Branch consists of four convolutional blocks, each one consisting of a convolutional layer with kernel size set to 3, a batch normalization layer [24], a Rectified Linear Unit [25], and a pooling layer based on maximum, with kernel size 2. Additionally, in the second and third blocks a dropout layer was added to reduce overfitting. In each block, the convolutional layer expands the channels' number. The first layer expands from 2 to 32 channels, while in the subsequent layers the number of channels is duplicated iteratively reaching 256 channels. A pooling layer is then used to compute the average of the convoluted signal at 10-time points, followed by an additional dropout layer.

Table 3. Breakdown of the number of ratings (N = 510) per aggregation and class type.

Rating	Q_m	Q_e	Q_c
Keep	19	53	47
Keep after Correction	324	213	212
Reject	167	244	251

Following the Convolutional Branch is an LSTM Module [26,27], a recursive layer used to leverage the specific properties of sequential data. The Network here employed contains a single-layer LSTM module, with a number of features in the hidden state set to 100.

The DNN was implemented in Python (v. 3.8.10), using the Numpy [28], Pandas [29], and Scikit-learn [30], and Torch [31] packages (Numpy v. 1.19.4, Scikit-learn v. 0.23.2, Pandas v. 1.1.4, Torch v. 1.9.0 + cu102). The network was trained for 1000 epochs, with a batch size of 128. The learning rate is initially fixed to 1, and divided by $\sqrt{10}$ every two epochs. Network's performances are evaluated in terms of accuracy, precision, and recall scores, as well as of the F1 score and Matthew Correlation Coefficient (MCC). While the accuracy of a model —the ratio between correctly classified segments and the total number of segments— is commonly used as the main metric to assess the performances of a model, it has been reported to be biased for dataset with an unbalanced number of labels per class, as in the case of the dataset here presented. In such cases, the accuracy score has been

proved to overestimate the performances of a classifier [10,32,33]. To take into account such a bias, different metrics have been introduced to assess the performances of binary and multiclass classifiers, such as the F1 score and the Matthew Correlation Coefficients. While the first has obtained a higher adoption in the field, the score is not class-independent, thus indicating different scores for binary classifiers accordingly to which class is labeled as positive and which is labeled as negative. Additionally, the F1 score does not take into account segments correctly classified as negative recall scores, therefore not providing a complete and objective evaluation of a model's performance. To overcome the limitations of the F1 score, a new metric named Matthew Correlation Coefficient (MCC), based on a special case of the ϕ correlation has been introduced [34,35]. As compared to the F1 score, MCC has two main advantages. First, all the four categories of the confusion matrix (true positives, false negatives, true negatives, and false positives) are considered in the metric, as well as the ratio between elements of the different classes in the dataset, therefore providing a more balanced performance indicator [36]. Secondly, the metric is class-independent, thus not influenced by the assignment of the positive and negative labels to the different classes. As a result of that, when classes are swapped, the metric does not change, as opposed to the F1 score [37]. Quantitatively, the MCC metric is a value between -1 and $+1$, where a value of -1 is indicative of a discrepancy between predictions and observations, while a coefficient of 1 represents a perfect forecasting capability of the model. As a result, the higher the MCC score, the better a model is performing.

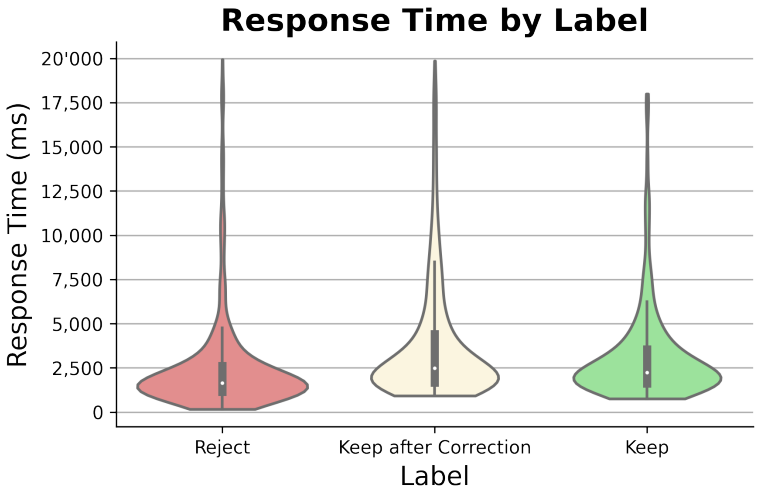

Figure 2. Distribution of users' response time by signals' quality rating.

3. Results

A copy of the segments used in this study, labeled dataset, and pre-trained network are available online on the data repository of this project [38], while the template for the web platform fNIRSQC has been released as an open sourced project under the name `cisciqc` (Citizen Science Quality Control [39]).

The dataset was divided into two partitions: Train (80% of the segments) and Test (20% of the segments). After the training phase, the model reported an Accuracy of 0.70 on the train set (MCC = 0.18). For what concerns the results on the validation set, the network obtained an Accuracy score of 0.63, a Precision score of 0.61, Recall score of 0.95, F1 score of 0.74, and an MCC score equal to 0.25 (Table 4).

Confusion Matrices for the train and test partitions are reported, respectively, in Tables 5 and 6. Overall, the model performs better on the training partition, suggesting a possible over-fitting problem.

Focusing on the confusion matrix, the model seems to wrongly report signals that users labeled as to reject, and therefore not usable signals, as signals that must be kept in subsequent steps of the analysis process.

Table 4. Summary of the results by metrics and partition.

Metric	Partition	
	Train	Test
Accuracy	0.70	0.63
MCC	0.18	0.25
Precision		0.61
Recall		0.95
F1		0.74

Table 5. Confusion matrix for the train partition.

		Predictions	
		Reject	Keep
User labels	Reject	29	92
	Keep	30	257

Table 6. Confusion matrix for the test partition.

		Predictions	
		Reject	Keep
User labels	Reject	10	35
	Keep	3	54

4. Discussion

In this work, we tested the possibility of using a Deep Neural Network model to support researchers in identifying usable and unusable fNIRS signals. First, a web platform for the collection of human labels for fNIRS signals has been designed and implemented, then ratings for 510 segments of fNIRS signals were obtained by raters of different expertise levels. Collected labels were then fed to train a DNN model.

The model's accuracy performances in train and test partitions indicate that the model is learning well on the test partition, but the performance drops on the test partition, suggesting a possible overfitting during the training phase. While the recall score (0.95) indicates that the majority of the relevant elements—which are usable segments—are correctly identified as usable, the precision score (0.61) suggests that a significant amount of signals labeled from the users as unusable are mislabeled by the model, as shown in the Confusion Matrix of the test partition (Table 6).

Imagining a possible implementation in a real research setting, the results here reported suggest that the model can successfully help researchers identifying usable fNIRS segments from a pool of segments that contains both usable or unusable segments, which is the typical case of fNIRS experimental studies. In fact, segments may contain a different type of artifacts, some of which can be corrected, while some are so extensively tied to the signals that require a portion of signals to be discarded. However, the current implementation fails at excluding completely unusable segments, which are labeled by the model so as to include them in further analysis. Presently, the model can still support researchers by reducing the number of segments requiring a manual inspection. The ability of the model to match users' labels for usable signals is also reflected by the F1 score (0.74), while the inability to correctly label non-usable signals with a high degree of precision is highlighted by the MCC score (0.25).

The limits in our model's classification accuracy can be explained by different factors. First, the limited number of segments that have been included in the dataset for this work (N = 510). In fact, the small number of segments may not have been sufficient to cover all the possible combinations of artifacts and noises that can affect the fNIRS signals. However, while a higher number of segments may have been helpful to reduce the bias of the model

toward the segments used for training—with a possible reduction in the accuracy on the train partition and the simultaneous increase of the precision on the test partition—the labeling stage of the segments would have required more extensive resources in terms of time and users. By limiting the number of segments, we have been able to obtain a higher number of ratings per each signal, therefore reducing the impact of a single rater on the overall label used for training, which is evaluated as described in Section 2.3. Future works may try to increase the quality of the predictions, by involving both more users and by adding more signals to the dataset, in order to obtain a bigger and more balanced number of segments, which may cover a higher number of possible cases.

For what concerns the network's performances, another way to improve the classification accuracy of the network is by improving its structure. In this work, we aimed at using a simple network, that has been adapted from a previous work in which the aim was not to classify the usability of a signal, but its nature [40]. The adoption of this simple network has some benefits. First, the linearity of the structure and the simplicity of the architecture allows for easy explainability—which is crucial especially when AI algorithms are used for medical data—and to rapidly modify it, to better adapt it to different scenarios (e.g., rating from more users, a higher number of samples, etc.) and computational resources (e.g., laptop, Cloud clusters, High-Performance Computers, etc.). Currently, the network trains on an average laptop (Intel i7-8565U, 16 GB of RAM) in less than 30 min, and is able to provide predictions within seconds, making it suitable for both offline and online classification of fNIRS segments.

Overall, in this work, we have demonstrated a proof-of-concepts of how a DNN model can be trained and employed to classify the usability of fNIRS signals. The developed model can help researchers estimate the quality of an fNIRS signal segment, and its usability for research purposes in a more objective way, by reducing the subjectivity introduced by a manual inspection stage.

While the performances of the model are not excellent, the limits of the dataset—in terms of number of segments, and of number and expertise of the raters—and of the architecture of the network can help explain the results here reported. Future work should aim at collecting data from a higher number of raters, with different expertise levels, for a higher number of fNIRS segments, and try to use more sophisticated networks designed ad-hoc for this classification task. Moreover, future studies may aim at combining different signals (e.g., fNIRS and EEG) in order to increase the performances of the classification model.

5. Conclusions

In this work, we presented a proof-of-concept for a DNN classifier able to help researchers identifying the quality of fNIRS signals. Moreover, as artifacts of this work, we created an open-source boilerplate for the creation of a citizen science platform for the human labeling of digital elements, called `cisciqc`, and its implementation for the collection of fNIRS signals labels, called `fNIRSQC`, as well as a dataset quality-labeled fNIRS signals that can be used by others to train and test different ML models. Our results demonstrate that a simple network, trained on a small number of signals labeled by users of different expertise levels can successfully help researchers identify high-quality fNIRS signals.

Author Contributions: Conceptualization, G.G., A.B. and G.E.; methodology, G.G., A.B. and G.E.; software, G.G. and A.B.; validation, G.G. and A.B.; formal analysis, G.G. and A.B.; investigation, G.G., A.B. and G.E.; resources, G.G., A.B. and G.E.; data curation, G.G. and A.B.; writing—original draft preparation, G.G., A.B. and M.J.Y.N.; writing—review and editing, G.G., A.B., M.J.Y.N. and G.E.; visualization, G.G.; supervision, G.E.; project administration, G.E.; funding acquisition, G.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by grants from the NAP SUG to GE (M4081597, 2015-2021).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the original studies.

Data Availability Statement: https://doi.org/10.21979/N9/C8VYZG (accessed on 13 October 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yücel, M.A.; Lühmann, A.v.; Scholkmann, F.; Gervain, J.; Dan, I.; Ayaz, H.; Boas, D.; Cooper, R.J.; Culver, J.; Elwell, C.E.; et al. Best practices for fNIRS publications. *Neurophotonics* **2021**, *8*, 012101.
2. Azhari, A.; Truzzi, A.; Neoh, M.J.Y.; Balagtas, J.P.M.; Tan, H.H.; Goh, P.P.; Ang, X.A.; Setoh, P.; Rigo, P.; Bornstein, M.H.; et al. A decade of infant neuroimaging research: What have we learned and where are we going? *Infant Behav. Dev.* **2020**, *58*, 101389. [CrossRef] [PubMed]
3. Pinti, P.; Tachtsidis, I.; Hamilton, A.; Hirsch, J.; Aichelburg, C.; Gilbert, S.; Burgess, P.W. The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Ann. N. Y. Acad. Sci.* **2020**, *1464*, 5. [CrossRef] [PubMed]
4. Strother, S.C. Evaluating fMRI preprocessing pipelines. *IEEE Eng. Med. Biol. Mag.* **2006**, *25*, 27–41. [CrossRef] [PubMed]
5. Poldrack, R.A.; Fletcher, P.C.; Henson, R.N.; Worsley, K.J.; Brett, M.; Nichols, T.E. Guidelines for reporting an fMRI study. *Neuroimage* **2008**, *40*, 409–414. [CrossRef] [PubMed]
6. Hocke, L.M.; Oni, I.K.; Duszynski, C.C.; Corrigan, A.V.; Frederick, B.D.; Dunn, J.F. Automated Processing of fNIRS Data—A visual guide to the pitfalls and consequences. *Algorithms* **2018**, *11*, 67. [CrossRef]
7. Gabrieli, G.; Scapin, G.; Bornstein, M.H.; Esposito, G. Are cry studies replicable? An analysis of participants, procedures, and methods adopted and reported in studies of infant cries. *Acoustics* **2019**, *1*, 866–883. [CrossRef]
8. Li, Q.; Clifford, G.D. Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiol. Meas.* **2012**, *33*, 1491. [CrossRef]
9. Li, Q.; Rajagopalan, C.; Clifford, G.D. A machine learning approach to multi-level ECG signal quality classification. *Comput. Methods Programs Biomed.* **2014**, *117*, 435–447. [CrossRef]
10. Gabrieli, G.; Balagtas, J.P.M.; Esposito, G.; Setoh, P. A Machine Learning approach for the automatic estimation of fixation-time data signals' quality. *Sensors* **2020**, *20*, 6775. [CrossRef]
11. Pollonini, L.; Olds, C.; Abaya, H.; Bortfeld, H.; Beauchamp, M.S.; Oghalai, J.S. Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy. *Hear. Res.* **2014**, *309*, 84–93. [CrossRef]
12. Pollonini, L.; Bortfeld, H.; Oghalai, J.S. PHOEBE: a method for real time mapping of optodes-scalp coupling in functional near-infrared spectroscopy. *Biomed. Opt. Express* **2016**, *7*, 5104–5119. [CrossRef]
13. Sappia, M.S.; Hakimi, N.; Colier, W.N.; Horschig, J.M. Signal quality index: an algorithm for quantitative assessment of functional near infrared spectroscopy signal quality. *Biomed. Opt. Express* **2020**, *11*, 6732–6754. [CrossRef]
14. Ortega, P.; Faisal, A.A. Deep learning multimodal fNIRS and EEG signals for bimanual grip force decoding. *J. Neural Eng.* **2021**, *18*, 0460e6. [CrossRef]
15. Ma, T.; Chen, W.; Li, X.; Xia, Y.; Zhu, X.; He, S. fNIRS Signal Classification Based on Deep Learning in Rock-Paper-Scissors Imagery Task. *Appl. Sci.* **2021**, *11*, 4922. [CrossRef]
16. Liu, R.; Reimer, B.; Song, S.; Mehler, B.; Solovey, E. Unsupervised fNIRS feature extraction with CAE and ESN autoencoder for driver cognitive load classification. *J. Neural Eng.* **2021**, *18*, 036002. [CrossRef] [PubMed]
17. Saikia, M.J.; Brunyéa, T.T. K-means clustering for unsupervised participant grouping from fNIRS brain signal in working memory task. In Proceedings of the Optical Techniques in Neurosurgery, Neurophotonics, and Optogenetics, International Society for Optics and Photonics, Online, 6–12 March 2021; Volume 11629, p. 116292M.
18. Sappia, M.S.; Hakimi, N.; Svinkunaite, L.; Alderliesten, T.; Horschig, J.M.; Colier, W.N. fNIRS signal quality estimation by means of a machine learning algorithm trained on morphological and temporal features. Biophotonics in Exercise Science, Sports Medicine, Health Monitoring Technologies, and Wearables II. *Int. Soc. Opt. Photonics* **2021**, *11638*, 116380F.
19. Azhari, A.; Leck, W.; Gabrieli, G.; Bizzego, A.; Rigo, P.; Setoh, P.; Bornstein, M.; Esposito, G. Parenting stress undermines mother-child brain-to-brain synchrony: A hyperscanning study. *Sci. Rep.* **2019**, *9*, 11407. [CrossRef] [PubMed]
20. Azhari, A.; Bizzego, A.; Esposito, G. Father-child dyads exhibit unique inter-subject synchronisation during co-viewing of animation video stimuli. *Soc. Neurosci.* **2021**, *16*, 522–533. [CrossRef]
21. Azhari, A.; Rigo, P.; Tan, P.Y.; Neoh, M.J.Y.; Esposito, G. Viewing Romantic and Friendship Interactions Activate Prefrontal Regions in Persons With High Openness Personality Trait. *Front. Psychol.* **2020**, *11*, 490. [CrossRef] [PubMed]
22. Azhari, A.; Lim, M.; Bizzego, A.; Gabrieli, G.; Bornstein, M.H.; Esposito, G. Physical presence of spouse enhances brain-to-brain synchrony in co-parenting couples. *Sci. Rep.* **2020**, *10*, 7569. [CrossRef] [PubMed]
23. Bizzego, A.; Gabrieli, G.; Esposito, G. Deep Neural Networks and Transfer Learning on a Multivariate Physiological Signal Dataset. *Bioengineering* **2021**, *8*, 35. [CrossRef] [PubMed]
24. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.

25. Shang, W.; Sohn, K.; Almeida, D.; Lee, H. Understanding and improving convolutional neural networks via concatenated rectified linear units. In Proceedings of the ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 2217–2225.
26. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
27. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Piscataway, NJ, USA, 26–31 May 2013; pp. 6645–6649.
28. Van den Walt, S.; Colbert, S.C.; Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30. [CrossRef]
29. McKinney, W. Pandas: A foundational Python library for data analysis and statistics. In Proceedings of the Workshop Python for High Performance and Scientific Computing (SC11), Seattle, WA, USA, 18 November 2011; Volume 14. Available online: https://www.dlr.de/sc/portaldata/15/resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf (accessed on 13 October 2021).
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
31. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037.
32. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1015–1021.
33. Bekkar, M.; Djemaa, H.K.; Alitouche, T.A. Evaluation measures for models assessment over imbalanced data sets. *J. Inf. Eng. Appl.* **2013**, *3*, 7633.
34. Guilford, J.P. *Psychometric Methods*; McGraw-Hill: Washington, DC, USA, 1954.
35. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
36. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C.A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412–424. [CrossRef]
37. Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **2004**, *28*, 367–374. [CrossRef]
38. Gabrieli, G.; Bizzego, A.; Esposito, G. Replication Data for: fNIRS-QC: Crowd-Sourced Creation of a Dataset and Machine Learning Model for fNIRS Quality Control. 2021. Available online: https://researchdata.ntu.edu.sg/dataset.xhtml?persistentId=doi:10.21979/N9/C8VYZG (accessed on 13 October 2021) [CrossRef]
39. Gabrieli, G. sanlab-ntu/cisciqc: 0.0.1.1. 2021. Available online: https://zenodo.org/record/5163238 (accessed on 13 October 2021).
40. Bizzego, A.; Gabrieli, G.; Furlanello, C.; Esposito, G. Comparison of wearable and clinical devices for acquisition of peripheral nervous system signals. *Sensors* **2020**, *20*, 6778. [CrossRef]