A FAIR evaluation of public datasets for stress detection systems

Alvaro Cuno*, Nelly Condori-Fernandez*†‡, Alexis Mendoza*, Wilber Ramos Lovón*
**Departamento de Ingeniería de Sistemas e Informática*
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú
{acunopa,ocondorif,amendoza,wramos}@unsa.edu.pe
†*Universidad de la Coruña*
La Coruña, España
n.condori.fernandez@udc.es
‡Vrije Universiteit Amsterdam, The Netherlands
n.condori-fernandez@vu.nl

***Abstract*—Nowadays, datasets are an essential asset used to train, validate, and test stress detection systems based on machine learning. In this paper, we used two sets of FAIR metrics for evaluating five public datasets for stress detection. Results indicate that all these datasets comply to some extent with the (F)indable, (A)ccessible, and (R)eusable principles, but none with the (I)nteroperable principle. These findings contribute to raising awareness on (i) the need for the FAIRness development and improvement of stress datasets, and (ii) the importance of promoting open science in the affective computing community.**

***Index Terms*—FAIR principles, Stress detection, Datasets.**

I. Introduction

Open Science [1] is the international movement to make any research artifact (i.e., software source code, datasets, algorithms, analysis scripts, manuscripts, tools, and so on) available and accessible to society. One of the common justifications for open science is the growing concern that researchers have expressed about a "reproducibility crisis" in some scientific fields [2], [3], including artificial intelligence [4] and computer science [5]. Open Science seeks to increase the transparency and, thus, **repeatability** (i.e., same team, same experimental setup), **reproducibility** (i.e., different team, same experimental setup) and **replicability** (i.e., different team, different experimental setup) of the scientific process and their results [6].

Several countries have taken the initiative to follow the Open Science movement. For instance, in 2018, the European Commission launched the European Open Science Cloud (EOSC) as a process of making research data in Europe accessible to all researchers under the same terms of use and distribution [7]. This initiative aims to push Europe towards a culture of open research artifacts that are **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. In other words, it aims to satisfy the FAIR principles [8]. In a broader global context, there are similar initiatives such as the USA NIH Data Commons, the Australian Research Data Commons, and the proposed African Open Science Platform [9]. In the South American region, we must highlight the GO FAIR Office's implementation in Brazil [10].

The FAIR principles have been broadly accepted in some scientific communities, such as in bio and natural sciences [11]. However, there is still insufficient implementation in the development of advanced information systems like emotion-aware systems[1]. As this kind of system has received much attention in recent years (e.g., [12], [13], [14]), it is important to determine the degree of FAIRness of datasets used for emotion recognition or detection. Usually, datasets are used to train, validate, and test supervised learning systems [15]; however, very little has been done to explore its findability, accessibility, interoperability, and reusability properties. Therefore, we aim to investigate the following research question: *How FAIR are the public datasets used in human stress detection systems?* To answer this question, five datasets were selected and evaluated employing two sets of FAIR metrics. Our work contributes to the maturity of open science in the affective computing community.

The paper is organized as follows: Section II introduces the FAIR principles, implementations, evaluation tools, and related work. In section III, we present the materials and the method used in this study. Section IV reports the results as well as the threats to the validity of our study. Finally, in Section V, the conclusion and some ideas for future work are presented.

II. Background

A. The FAIR principles

The FAIR foundational principles introduced by Wilkinson et al. [8] provide guidelines to enable digital resources, such as datasets, source code, tools, workflows, and other scientific research artifacts, to become more **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable for humans and computers. Jacobsen et al. [16] describe the core objective of these principles as following:

[1] A type of software-intensive system that is aware of the user's emotions like stress, anger, disgust, fear, happiness, among others.

978-1-7281-8328-2/20/$31.00 ©2020 IEEE

To be Findable:	
F1	(Meta)data are assigned a globally unique and persistent identifier
F2	Data are described with rich metadata (defined by R1 below)
F3	Metadata clearly and explicitly include the identifier of the data they describe
F4	(Meta)data are registered or indexed in a searchable resource
To be Accessible:	
A1	(Meta)data are retrievable by their identifier using a standardised communications protocol
	A1.1 The protocol is open, free, and universally implementable
	A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
A2	Metadata are accessible, even when the data are no longer available
To be Interoperable:	
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2	(Meta)data use vocabularies that follow FAIR principles
I3	(Meta)data include qualified references to other (meta)data
To be Reusable:	
R1	Meta(data) are richly described with a plurality of accurate and relevant attributes
	R1.1. (Meta)data are released with a clear and accessible data usage license
	R1.2. (Meta)data are associated with detailed provenance
	R1.3. (Meta)data meet domain-relevant community standards

TABLE I: The FAIR guiding principles [8].

- **F**indable: Digital resources should be easy to find for both humans and machines. Extensive machine-actionable metadata are essential for the automatic discovery of relevant datasets and services.
- **A**ccessible: Protocols for retrieving digital resources should be made explicit for humans and machines, including well-defined mechanisms to obtain authorization for accessing protected data.
- **I**nteroperable: When two or more digital resources are related to the same topic or entity, it should be possible for machines to merge the information into a richer and unified view.
- **R**eusable: Digital resources should be sufficiently well described for both humans and machines, such that a machine is capable of deciding: (i) if a digital resource should be reused (i.e., is it relevant to the task-at-hand?), (ii) if a digital resource can be reused, and under what conditions (i.e., do I fulfill the conditions of reuse?), and (iii) whom to credit if it is reused.

The four foundational principles are more explicitly and measurably described by the fifteen FAIR guiding principles shown in Table I.

B. FAIR principles implementation

According to Jacobsen et al. [16], there are several alternative routes towards the implementation of the FAIR principles, some specialized for different types of digital resources, for example:

- The FAIR metrics [17]. It comprises an open framework for the creation and publication of metrics and an initial set of 14 FAIR metrics. The metrics are formulated into a series of questions, provided as a questionnaire. The responses to these questionnaires are then manually evaluated.
- The Maturity Indicators [18]. It comprises an automatable evaluation framework composed of fifteen maturity indicators for the objective and quantitative evaluation of the FAIRness level of digital objects. It is defined as the second generation of the FAIR metrics presented in the previous item.
- The EC report on turning FAIR into reality [9]. It offers a survey and analysis of what is needed to implement FAIR in a broad sense, and it provides a set of concrete recommendations and actions for stakeholders in Europe and beyond.
- The "FAIR principles explained" described on the GO FAIR website[2]. It comprises a three-point FAIRification framework that provides practical "how-to" guidance to stakeholders seeking to go FAIR. The three points are Metadata for Machines, FAIR Implementation Profile, and FAIR Data Points. The framework maximizes the reuse of existing resources, maximizes interoperability, and accelerates convergence on standards and technologies supporting FAIR data and services.
- The FAIRshake toolkit [19]. The FAIRshake toolkit was developed to establish community-driven FAIR metrics and rubrics paired with manual and automated FAIR assessments. It enables the systematic assessment of the FAIRness of any digital resource. It contains a database that enlists users, projects, digital resources, metrics, rubrics (i.e., a set of metrics), and assessments. Among the registered rubrics on the toolkit, we can highlight the following: the FAIRshake tool rubric, the FAIRshake dataset rubric, the FAIRshake repository rubric, the FAIRshake JSON-LD rubric, the Repositive Discover datasets rubric, and the FAIRshake universal metrics.
  The FAIRshake toolkit is a full-stack application with a user interface, and it comes with a browser extension and a bookmarklet to enable viewing and submitting assessments from any website. The assessment results are visualized as an insignia representing the FAIR score in a compact grid of squares colored between red and blue.

C. FAIRness evaluation tools

Tools developed for conducting FAIRness evaluations can be categorized into three groups [20]:

- Discrete-answer questionnaire-based evaluations. This approach consists of online self-report questionnaires and checklists, where respondents need to indicate their implementation choice from a predefined set of answers. Questions are grouped according to each principle. The evaluation's output is a sum score, a weighted sum score, or a visual score that is automatically generated after the completion of the evaluation.
- Open-answer questionnaire-based evaluation. This approach is similar to the previous one but with the dif-

[2] https://www.go-fair.org/fair-principles/

ference that there is no predefined set of answers. In this case, respondents need to dig into many different resources to attempt to provide answers. Open-answer questionnaires for FAIRness evaluations have similar limitations to discrete answer category questionnaires, being time-consuming and subject to respondent bias. However, the following three advantages can be attributed to them: (i) answers need to include a statement that evidences the FAIR implementation, e.g., a URI of a metadata record of the adopted standard, (ii) it allows scientific communities to create additional maturity indicators, and (iii) it can be filled in multiple occasions (spreadsheet version).

- Semi-automated evaluation. This approach consists of a web-based metadata harvester that parses metadata of a digital resource. Usually, it requires as input a global unique identifier (GUID) and the definition of maturity indicators. This approach has advantages such as (i) it is automated, which is supposed to reduce respondent bias, and (ii) it ensures transparency as the evaluations are open and the evaluators identifiable. However, the semi-automated approach also has some important limitations: (i) it requires that the resource has some kind of metadata provider available (i.e., and therefore is not useful for projects in phase of development), (ii) it depends on the compatibility between software and metadata provider, (iii) it performs differently when comparing two different identifiers of the same resource.

D. Related work

Trifan et al. [21] presented an overview of the adoption and impact of the FAIR principles in the area of biomedical and life-science research. In particular, they evaluated the FAIR compatibility degree of three biomedical data discovery platforms: MaelstromCatalogue, YummyData, and FAIRsharing. Their findings show a high level of FAIRness achieved by these platforms and an increased concern for enabling data discovery by machines, mainly favored by the rich metadata with which each of these complements the actual data sources.

Berrios et al. [22] evaluated the FAIRness of five open-access ‘omics’ (genomics, transcriptomics, proteomics, and metabolomics) data systems using the 14 FAIRness metrics developed by the GO FAIR Metrics group [23]. The evaluated systems performed the best in the areas of data **F**indability and **A**ccessibility, slightly well in **R**eusability and worst in the area of data **I**nteroperability. Additionally, they proposed two new principles that Big Data system developers, in particular, should consider for maximizing data accessibility.

Lamprecht et al. [24] argued that although the FAIR principles are mostly applied to research data, it is also relevant to analyze their application to research software. They also discussed what makes software different from data concerning applying the FAIR principles and which desired research software characteristics go beyond FAIR. For this reason, they presented an analysis for determining where the existing FAIR principles can directly be applied to software, where they need to be adapted or reinterpreted, and where the definition of

TABLE II: FAIRshake universal metrics

Metric ID	**Name**	**Description**	**P**
FUM-F1	Globally unique identifier	URL to a registered scheme that defines the globally-unique structure of the digital resource identifier.	F
FUM-F2	Persistent identifier	URL to a document that defines the policy of long term support for persisting the identifier.	F
FUM-F3	Machine-readable metadata	URL to a document that contains machine-readable metadata for the digital resource.	F
FUM-F4	Standardized metadata	URI of a registered metadata format in FAIRsharing.	F
FUM-F5	Resource identifier in metadata	The resource identifier that should explicitly appear in the metadata.	F
FUM-F6	Resource discovery through web search	URL, including GET string parameters, that will return a successful search for the subject resource.	F
FUM-A1	Access protocol	URL to an open, free, and standardized access protocol.	A
FUM-A2	Restricted content	URL to a document that specifies the protocol to gain access to restricted content.	A
FUM-A3	Persistence	URL for the persistence policy of the digital resource and its metadata.	A
FUM-I1	Formal language	URL of a formal, shared, and broadly applicable language for the digital resource.	I
FUM-I2	Vocabulary	URL to a FAIR vocabulary used in by the resource.	I
FUM-I3	Linked	A URL to the LinkSet document for the resource.	I
FUM-R1	License	URL to the license that governs the use of the digital resource.	R
FUM-R2	Metadata license	URL to the license that governs the use of the digital resource.	R
FUM-R3	Provenance scheme	URL of a vocabulary used to describe the provenance of the digital resource.	R
FUM-R4	Certificate of compliance	URL to a certified document that the digital resource complies to a community standard.	R

TABLE III: FAIRshake dataset metrics

Metric ID	**Name**	**Description**	**P**
FDM-F1	Identifier	A standardized ID or accession number is used to identify the dataset	F
FDM-F2	Metadata	The dataset is described with metadata using a formal, broadly applicable vocabulary	F
FDM-F3	Repository	The dataset is hosted in an established data repository, if a relevant repository exists	F
FDM-A1	Download	The dataset can be downloaded for free from the repository	A
FDM-R1	Experiment	Information is provided on the experimental methods used to generate the data	R
FDM-R2	Versioning	Version information is provided for the dataset	R
FDM-R3	Contact	Contact information is provided for the creator(s) of the dataset	R
FDM-R4	Citation	Information is provided describing how to cite the dataset	R
FDM-R5	License	Licensing information is provided on the dataset’s landing page	R

additional principles is required. It was found that among the four principles, **I**nteroperability has proven to be the most challenging one.

Berčič et al. [25] analyzed the state of research data in Mathematics. They found that while the mathematical community embraces the notion of open data, the FAIR principles are not yet sufficiently realized. Freely accessible datasets are hard or impossible to reuse because of the missing metadata annotating the raw research data. Therefore, the authors introduce *deep FAIRness* for mathematical research data as an extended set of FAIR requirements, which accommodate the special needs of math datasets.

III. Materials and method

A. Datasets

The stress datasets selected for the evaluation were those five identified by Mahesh et al. [26]:

- DRIVE-DB [27]. It was collected for detecting drivers' overall stress levels. It includes data from 9 subjects. Stress responses were captured while driving on planned routes with varying cognitive load. The driver's video was captured to manually estimate the stress level based on head movements and confirm the cognitive load. It includes physiological modalities such as respiration, electromyogram (EMG), electrocardiogram (ECG), heart rate (HR), and galvanic skin response (GSR), captured in an ambulatory environment.
- SWELL-KW [28]. It was collected for studying the stressful behavior of knowledge workers. It includes data from 25 subjects. Time pressure and email interruptions were used as stressors in an office work scenario. Computer interactions, facial expressions, body postures, and physiological modalities such as ECG and SC were captured. Several subjective self-reports of stress were collected through questionnaires for use as ground truth.
- SUSAS [29]. It is one of three datasets collected with the primary goal of developing robust speech processing algorithms to study the effects of stress and emotion on speech. It includes data from 32 speakers and speech files from four Apache helicopter pilots. Single-word utterances were recorded during aircraft communication and other activities that differed from activities of daily living.
- DDD [30]. It was collected to study driving behaviors under distracting stressors such as cognitive, emotional, sensorimotor, and startling events, often resulting in vehicle accidents. It includes data from 68 subjects and consists of stress response modalities such as heart rate, respiration rate, facial expressions, gaze, and EDA from palm and perinasal areas. Several questionnaires are used to obtain self-reports of task load, cognitive state, and personality type.
- WESAD [31]. It was collected to provide high-quality multimodal data for stress and amusement state (affect) detection. The data was collected from 15 subjects with the Trier Social Stress Test as stress stimuli. Physiological modalities such as ECG, EDA, blood volume pulse (BVP), EMG, respiration, and body temperature were captured along with triaxial acceleration to provide contextual information. A chest-worn device, RespiBAN professional, and a wrist-worn device, Empatica E4, were used for data collection. Four self-reports have been provided along with additional notes wherever available.

B. Metrics

The metrics used for the evaluation were those included in two rubrics of the FAIRshake toolkit [19]:

- FAIRshake universal metrics (FUM): Originally [17], it was composed of fourteen metrics universally applicable to all digital resources in all scholarly domains, but the FAIRshake toolkit added two additional metrics. These sixteen metrics are presented in Table II, where six are associated with the **F**indable principle, three with the **A**ccessible, three with the **I**nteroperable, and four with the **R**eusable.
- FAIRshake dataset metrics (FDM): The FAIRshake toolkit provides a set of nine metrics designed explicitly for datasets evaluation. These metrics are presented in Table III, where three are associated with the **F**indable principle, one with the **A**ccessible, and five with the **R**eusable. No metric is associated with the **I**nteroperable principle. This omission is in accordance with what is established by Wilkinson et al. [8], in the sense that although the principles are related, they are independent and separable, and they can be adhered to in any combination and incrementally.

C. Evaluation

The evaluation was performed using the FAIRshake universal metrics and the FAIRshake dataset metrics presented in section III.B. Each dataset was assessed using only publicly available information. We manually searched for evidence on its webpage and the dataset data. The evidences are detailed in Appendix A and B. Per metric, four possible valid answers are used: yes, no, $yesbut$, and $nobut$. If evidence has been found or not, yes and no are used, respectively. The presence of the suffix but is an indication that a clarification is necessary.

IV. Results

In tables IV and V, we can see the evaluation results using the FAIRshake universal metrics and the FAIRshake dataset metrics, respectively. Next, we will review the results by metric, dataset, and per principle.

A. Compliance per metric

For the evaluation with the FAIRshake universal metrics (see Table IV), we found that those with satisfactory compliance for all datasets are (i) the discovery through web search metric (F6), and (ii) the access protocol metric (A1); whereas those with unsatisfactory compliance are (i) interoperability metrics (I1, I2, I3), (ii) reusability metrics in terms of metadata license (R2), and (iii) compliance certificate (R4).

TABLE IV: Evaluation result using the FAIRshake universal metrics

Metric ID	DRIVE-DB	SWELL-KW	SUSAS	DDD	WESAD
FUM-F1	*yes*	*yes*	*no*	*yes*	*no*
FUM-F2	*no*	*yes*	*no*	*yes*	*no*
FUM-F3	*no*	*yes*	*no*	*no*	*no*
FUM-F4	*no*	*nobut*	*no*	*no*	*no*
FUM-F5	*no*	*yes*	*no*	*no*	*no*
FUM-F6	*yes*	*yes*	*yes*	*yes*	*yes*
FUM-A1	*yes*	*yes*	*yes*	*yes*	*yes*
FUM-A2	*yes*	*yesbut*	*no*	*yes*	*yes*
FUM-A3	*no*	*yes*	*yes*	*yes*	*no*
FUM-I1	*no*	*no*	*no*	*no*	*no*
FUM-I2	*no*	*no*	*no*	*no*	*no*
FUM-I3	*no*	*no*	*no*	*no*	*no*
FUM-R1	*yes*	*yes*	*yes*	*nobut*	*nobut*
FUM-R2	*no*	*no*	*no*	*no*	*no*
FUM-R3	*nobut*	*nobut*	*nobut*	*nobut*	*nobut*
FUM-R4	*no*	*no*	*no*	*no*	*no*

TABLE V: Evaluation result using the FAIRshake dataset metrics

Metric ID	DRIVE-DB	SWELL-KW	SUSAS	DDD	WESAD
FDM-F1	*yes*	*yes*	*no*	*yes*	*no*
FDM-F2	*no*	*yes*	*no*	*no*	*no*
FDM-F3	*yes*	*yes*	*yes*	*yes*	*yes*
FDM-A1	*yes*	*yesbut*	*no*	*yes*	*yes*
FDM-R1	*yes*	*yes*	*yes*	*yes*	*yes*
FDM-R2	*yes*	*nobut*	*not*	*yes*	*not*
FDM-R3	*nobut*	*yes*	*nobut*	*yes*	*yes*
FDM-R4	*yes*	*yes*	*yes*	*yes*	*yes*
FDM-R5	*yes*	*yes*	*yes*	*nobut*	*nobut*

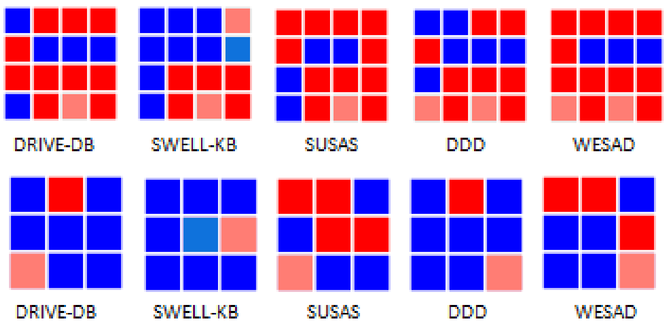

Fig. 1: The FAIR insignia for each dataset using the FAIRshake universal metrics (top) and the FAIRshake dataset metrics (bottom).

Concerning the application of the FAIRshake dataset metrics (see Table V), we observed that all datasets rated very well regarding (i) hosting the dataset in a repository (F3), (ii) the information about the experiment (R1), and (iii) information describing how to cite the dataset (R4) metrics. We can observe that there is no unfulfilled metric for all datasets at the same time. However, we do not forget that FDM does not include any metric for the **I**nteroperable principle.

B. Compliance per dataset

An insignia [19] is used to visualize the compliance level per dataset. It is a bi-dimensional matrix, where each cell that expresses a metric is colored with blue, royal blue, pink, or red. Each color represents a nominal scale with four values: *yes* (satisfactory), *yesbut* (satisfactory but), *no* (unsatisfactory), and *notbut* (unsatisfactory but), respectively.

As shown in Figure 1, for both sets of metrics (FUM and FDM), the SWELL-KB could be considered the FAIRest (since it has fewer red cells than other datasets). For the FUM and FDM metrics, the WESAD dataset and the WESAD dataset, respectively, could be considered the most unFAIR (since it has less blue cells than other datasets).

Comparing both groups of insignias, the evaluations performed with the FUM metrics (top) are less compliant than the FDM metrics (bottom). This situation can be explained due to the different purposes of the two sets of metrics. The FUM aims to be more generic (for all kinds of digital resources), whereas the FDM tends to be more specific (only for datasets).

C. Compliance per principle

Giving that a different number of metrics measures each principle, firstly, the number of obtained positive values (*yes* and *yesbut*) were averaged per principle, and then the corresponding percentages were calculated. Table VI shows these percentages per principle obtained for the corresponding evaluation using FUM (Table IV) and FDM (Table V). For example, the entry “2/6” means that two out of six Findable metrics were positively evaluated, and 33.6% represents the corresponding percentage. Similarly, each cell’s bold values represent the average between the two (non-bold) percentages corresponding to the FUM and FDM metrics.

According to the FUM evaluation, the **A**ccessible principle is the most fulfilled (four out of five datasets comply with it in more than 60%). In the second place was the **F**indable principle (all datasets have a compliance percentage). In the third place was the **R**eusable principle (only three datasets have a compliance percentage). The last place was to the **I**nteroperable principle (no dataset complies with it).

According to the FDM evaluation, the **A**ccessible principle is the most fulfilled (four out of five datasets have 100% compliance). In the second place was the **R**eusable principle (all datasets have more than 50% compliance). In the third place was the **F**indable principle (all datasets have a compliance percentage). The last place was to the **I**nteroperable principle (no dataset complies with it).

Averaging both evaluation results (FUM and FDM), the **A**ccessible principle is the most fulfilled. In the second place is the **F**indable principle, in the third place is the **R**eusable principle, and in the last place is the **I**nteroperable principle.

Further, some significant findings that can be obtained from Table VI are the followings. DRIVE-DB is very **A**ccessible (83%), partially **R**eusable (53%) and **F**indable (49.9%), and nothing **I**nteroperable (0%); SWELL-KW is wholly **A**ccessible (100%), very **F**indable (91.6%), partially **R**eusable (52.5%), and nothing **I**nteroperable (0%); SUSAS

TABLE VI: Compliance level per principle for each dataset.

	FUM|FDM	FUM|FDM	FUM|FDM	FUM|FDM
	F	**A**	**I**	**R**
DRIVE-DB	2/6 | 2/3 33.3%|66.6% **50.0%**	2/3 | 1/1 66%|100% **83%**	0/3 | — 0%|0% **0%**	1/4 | 4/5 25%|80% **53%**
SWELL-KW	5/6 | 3/3 83.3%|100% **91.6%**	3/3 | 1/1 100%|100% **100%**	0/3 | — 0%|0% **0%**	1/4 | 4/5 25%|80% **52.5%**
SUSAS	1/6 | 1/3 16.6%|33.3% **24.9%**	2/3 | 0/1 66.6%|0% **33.3%**	0/3 | — 0%|0% **0%**	1/4 | 3/5 25%|60% **42.5%**
DDD	3/6 | 2/3 50%|66.6% **58.3%**	3/3 | 1/1 100%|100% **100%**	0/3 | — 0%|0% **0%**	0/4 | 4/5 0%|80% **40%**
WESAD	1/6 | 1/3 16.6%|33.3% **25%**	2/3 | 1/1 66.6%|100% **83.3%**	0/3 | — 0%|0% **0%**	0/4 | 3/5 0%|60% **30%**

is partially **R**eusable (42.5%), almost **A**ccessible (33.3%) and **F**indable (24.9%), and nothing **I**nteroperable; DDD is wholly **A**ccessible (100%), both partially **F**indable (58.3%) and **R**eusable (40%), and nothing **I**nteroperable (0%); and, WESAD is pretty **A**ccessible (83.3%), both almost **F**indable (25%) and **R**eusable (30%), and nothing **I**nteroperable (0%).

D. Threats to validity

There are some threats to the validity of this qualitative study's results, which should be considered when interpreting its findings.

- **Construct validity:** The construct validity seeks to verify whether the metrics are measuring what it has been proposed to measure. In effect, this threat has been reduced by means of applying a set of generic metrics (FUM), and another set of specific metrics (FDM) designed explicitly for FAIRness evaluation. Although the FDM does not have metrics associated with the (**I**)nteroperable principle, this has not generated any inconsistency since the FUM evaluation has presented total non-compliance of the interoperability principle.
- **Internal validity:** Internal validity is the extent to which the design of a study supports the conclusion that changes in the independent variable caused any observed differences in the dependent variable [32]. Because this is a non-experimental observational study, the internal validity is low.
- **External validity:** The study took only five public datasets for emotion recognition, so generalization is not possible. The results are only valid for these.
- **Reliability:** The study has been documented in detail so that replicability has been ensured. A researcher should be able to follow the derivation of results and conclusions from this information.

V. Conclusions and future work

Five stress datasets have been evaluated using the FAIRshake universal metrics and the FAIRshake dataset metrics. With both metrics, we found that none of them fully satisfy the FAIR principles. All datasets comply the metrics with some extent respect to, in order of compliance, the **A**ccesible, **F**indable, and **R**eusable principles. However, none of them satisfied with the **I**nteroperable principle. In this context, of the five datasets, SWELL-KW is the FAIRest because it has high compliance ($> 90\%$) on **F** and **A** principles; and SUSAS the worst compliance ($< 50\%$) on all principles; while WESAD, DRIVE-DB, and DDD have intermediate compliance with the **A** principle $> 80\%$. We believe that these findings contribute to (i) creating awareness on the FAIRness limitations of these stress datasets, (ii) the need for its improvement, and (iii) promoting open science in the affective computing community.

A natural progression of this work is to compare the automated evaluation of FAIRness against collections of Maturity Indicator (MI) tests proposed by Wilkinson et al. [18]. Also, we plan to extend the evaluation by analyzing the compliance of the FAIR principles for published datasets' improvement or the development of new ones. Moreover, we consider that it might be interesting to develop specific FAIR metrics to evaluate physiological datasets due to these are not necessarily similar to other types of datasets (e.g., video, audio, image, etc.).

Acknowledgment

A. Cuno, N. Condori-Fernandez, A. Mendoza, and W. Ramos acknowledge financial support from the "Proyecto Concytec - Banco Mundial, Mejoramiento y Ampliación de los Servicios del Sistema Nacional de Ciencia Tecnología e Innovación Tecnológica" 8682-PE, through its executing unit FONDECYT [Contract N° 014-2019-FONDECYT-BM-INC.INV]. Also, this work has been partially supported by Datos 4.0 (TIN2016-78011-C4-1-R) funded by MINECO-AEI/FEDER-UE.

References

[1] The Royal Society, "Science as an open enterprise: open data for open science," 2012.

[2] K. C. Elliott and D. B. Resnik, "Making open science work for science and society," *Environmental health perspectives*, vol. 127, no. 7, p. 075002, 2019.

[3] M. R. Munafò, B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. P. Du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. Ioannidis, "A manifesto for reproducible science," *Nature human behaviour*, vol. 1, no. 1, pp. 1–9, 2017.

[4] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, C. S. Greene, *et al.*, "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, pp. E14–E16, 2020.

[5] J. R. F. Cacho and K. Taghva, "The state of reproducible research in computer science," in *17th International Conference on Information Technology–New Generations (ITNG 2020)*, pp. 519–524, Springer, 2020.

[6] D. M. Fernández, D. Graziotin, S. Wagner, and H. Seibold, "Open Science in Software Engineering," 2019.

[7] P. Budroni, J. Claude-Burgelman, and M. Schouppe, "Architectures of knowledge: the European open science cloud," *ABI Technik*, vol. 39, no. 2, pp. 130–141, 2019.

[8] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, 2016.

[9] S. Collins, F. Genova, N. Harrower, S. Hodson, S. Jones, L. Laaksonen, D. Mietchen, R. Petrauskaitė, and P. Wittenburg, "Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data," 2018.

[10] L. Sales, P. Henning, V. Veiga, M. M. Costa, L. F. Sayão, L. O. B. da Silva Santos, and L. F. Pires, "GO FAIR Brazil: a challenge for brazilian data science," *Data Intelligence*, vol. 2, no. 1-2, pp. 238–245, 2020.

[11] M. van Reisen, M. Stokmans, M. Basajja, A. O. Ong'ayo, C. Kirkpatrick, and B. Mons, "Towards the Tipping Point for FAIR Implementation," *Data Intelligence*, vol. 2, no. 1-2, pp. 264–275, 2020.

[12] M. Feidakis, *A Review of Emotion-Aware Systems for e-Learning in Virtual Environments*, pp. 217–242. 12 2016.

[13] N. Condori-Fernandez, "HAPPYNESS: An Emotion-aware QoS Assurance Framework for Enhancing User Experience," in *Proceedings of the 39th International Conference on Software Engineering Companion*, ICSE-C '17, (Piscataway, NJ, USA), pp. 235–237, IEEE Press, 2017.

[14] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Emma: An emotion-aware wellbeing chatbot," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–7, 2019.

[15] J. Yoon, S. O. Arik, and T. Pfister, "Data Valuation using Reinforcement Learning," in *Proceedings of the 37 th International Conference on Machine Learning*, 2020.

[16] A. Jacobsen et al., "FAIR Principles: Interpretations and Implementation Considerations," *Data Intelligence*, vol. 2, no. 1-2, pp. 10–29, 2020.

[17] M. D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L. O. B. da Silva Santos, and M. Dumontier, "A design framework and exemplar metrics for FAIRness," *Scientific data*, vol. 5, 2018.

[18] M. D. Wilkinson, M. Dumontier, S.-A. Sansone, L. O. B. da Silva Santos, M. Prieto, D. Batista, P. McQuilton, T. Kuhn, P. Rocca-Serra, M. Crosas, *et al.*, "Evaluating FAIR maturity through a scalable, automated, community-governed framework," *Scientific data*, vol. 6, no. 1, pp. 1–12, 2019.

[19] D. J. Clarke, L. Wang, A. Jones, M. L. Wojciechowicz, D. Torre, K. M. Jagodnik, S. L. Jenkins, P. McQuilton, Z. Flamholz, M. C. Silverstein, *et al.*, "FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources," *Cell systems*, vol. 9, no. 5, pp. 417–421, 2019.

[20] R. de Miranda Azevedo and M. Dumontier, "Considerations for the conduction and interpretation of FAIRness evaluations," *Data Intelligence*, vol. 2, pp. 285–292, 2020.

[21] A. Trifan and J. L. Oliveira, "Towards a More Reproducible Biomedical Research Environment: Endorsement and Adoption of the FAIR Principles," in *Biomedical Engineering Systems and Technologies* (A. Roque, A. Tomczyk, E. De Maria, F. Putze, R. Moucek, A. Fred, and H. Gamboa, eds.), (Cham), pp. 453–470, Springer International Publishing, 2020.

[22] D. C. Berrios, A. Beheshti, and S. V. Costes, "FAIRness and usability for open-access omics data systems," in *AMIA Annual Symposium Proceedings*, vol. 2018, p. 232, American Medical Informatics Association, 2018.

[23] GO FAIR Metrics Group, "Fair metrics all," 2018.

[24] A.-L. Lamprecht, L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. Martin Del Pico, V. Dominguez Del Angel, S. van de Sandt, J. Ison, P. A. Martinez, *et al.*, "Towards FAIR principles for research software," *Data Science*, no. Preprint, pp. 1–23, 2019.

[25] K. Berčič, M. Kohlhase, and F. Rabe, "(Deep) FAIR mathematics," *it-Information Technology*, vol. 62, no. 1, pp. 7–17, 2020.

[26] B. Mahesh, E. Prassler, T. Hassan, and J.-U. Garbas, "Requirements for a Reference Dataset for Multimodal Human Stress Detection," in *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 492–498, IEEE, 2019.

[27] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[28] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in *Proceedings of the 16th international conference on multimodal interaction*, pp. 291–298, 2014.

[29] H. J. Steeneken and J. H. Hansen, "Speech under stress conditions: overview of the effect on speech production and on system performance," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 4, pp. 2079–2082, IEEE, 1999.

[30] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis, "A multimodal dataset for various forms of distracted driving," *Scientific data*, vol. 4, p. 170110, 2017.

[31] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 400–408, 2018.

[32] P. Price, R. Jhangiani, I. Chiang, D. Leighton, and C. Cuttler, *Research Methods in Psychology (3rd American Edition).* 2017.

Appendix A
Evidences for evaluating with the FAIRshake universal metrics

- **Globally unique identifier (F1)**: Three datasets have been scored with yes (see the DOI) and two with no:

DRIVE-DB:	https://doi.org/10.13026/C2SG6B	yes
SWELL-KW:	https://doi.org/10.17026/dans-x55-69zp	yes
SUSAS:	Not found	no
DDD:	https://doi.org/10.17605/OSF.IO/C42CN	yes
WESAD:	Not found	no

- **Persistent identifier (F2)**: Three datasets have been scored with yes (see the policies) and two with no:

DRIVE-DB:	https://www.doi.org/doi_handbook/6_Policies.html	yes
SWELL-KW:	https://www.doi.org/doi_handbook/6_Policies.html	yes
SUSAS:	Not found	no
DDD:	https://www.doi.org/doi_handbook/6_Policies.html	yes
WESAD:	Not found	no

- **Machine-readable metadata (F3)**: Only the SWELL-KW have a XML file (https://easy.dans.knaw.nl/ui/resources/easy/export?sid=easy-dataset:58624&format=XML), the other ones do not have a metadata file.
- **Standardized metadata (F4)**: Datasets do not have a registered metadata format in FAIRsharing platform, but the SWELL-KW has a defined format metadata http://easy.dans.knaw.nl/easy/easymetadata/.
- **Resource identifier in metadata (F5)**: Only the SWELL-KW machine-readable metadata contains the resource identifier; the other ones do not have a metadata file.
- **Resource discovery through web search (F6)**: All the datasets are indexed by https://www.google.com/, and found in the first page of the search:

DRIVE-DB:	Search query: stress recognition in automobile drivers	yes
SWELL-KW:	Search query: the swell knowledge work dataset for stress and user modeling research	yes
SUSAS:	Search query: speech under simulated and actual stress	yes
DDD:	Search query: a multimodal dataset for various forms of distracted driving	yes
WESAD:	Search query: wesad: multimodal dataset for wearable stress and affect detection	yes

- **Access protocol (A1)**: All the datasets have been published using the HTTP protocol that is described on https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol.
- **Restricted content (A2)**: Three datasets have been scored with yes, one with no, and one with $yesbut$:

DRIVE-DB:	It is free	*yes*
SWELL-KW:	Downloading for free is possible but only after filling a registration form	*yesbut*
SUSAS:	https://www.ldc.upenn.edu/language-resources/data/obtaining	*no*
DDD:	It is free	*yes*
WESAD:	It is free	*yes*

- **Persistence (A3)**: Three datasets have a persistence policy, but two no.

DRIVE-DB:	Not found	*no*
SWELL-KW:	https://dans.knaw.nl/en/about/organisation-and-policy/policy-and-strategy/preservation-plan-data-archiving-and-networked-services-dans-1	*yes*
SUSAS:	It has a CoreTrustSeal certification, whose 10th requirement establish that the repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.	*yes*
DDD:	https://help.osf.io/hc/en-us/articles/360019737894-FAQs#Backup-Preservation-Policy	*yes*
WESAD:	Not found	*no*

- **Formal language (I1)**: No dataset presents a URL to a formal language for knowledge representation.
- **Vocabulary (I2)**: No dataset presents a URL to a FAIR vocabulary used by the resource.
- **Linked (I3)**: No dataset has a URL to a LinkSet document.
- **License (R1)**: Three datasets have been scored with *yes*, and two with *nobut*:

DRIVE-DB:	https://www.physionet.org/content/drivedb/view-license/1.0.0/	*yes*
SWELL-KW:	https://dans.knaw.nl/en/about/organisation-and-policy/legal-information/DANSLicence.pdf	*yes*
SUSAS:	https://catalog.ldc.upenn.edu/license/ldc-non-members-agreement.pdf	*yes*
DDD:	Not license found but OSF terms and conditions is available	*nobut*
WESAD:	Not license found but a disclaimer is available	*nobut*

- **Metadata license (R2)**: No dataset has a URL to a metadata license.
- **Provenance scheme (R3)**: It was not found a URL of vocabulary used to describe the provenance of the datasets. However, it was found who/what/when produced the dataset (i.e., for citation), and why/how was the dataset produced (i.e., to understand its context and relevance): DRIVE-DB [27], SWELL-KW [28], SUSAS [29], DDD [30], WESAD [31].
- **Certificate of compliance (R4)**: No dataset has a URL to a certified document that the digital resource complies to a community standard.

Appendix B
Evidences for evaluating with the FAIRshake universal metrics

- **Identifier (F1)**: The result is the same as the F1 metric of the FAIRshake universal metrics.
- **Metadata (F2)**: The result is the same as the F2 metric of the FAIRshake universal metrics.
- **Experiment (R1)**: All the datasets have been scored with *yes* because they were presented in scientific papers: DRIVE-DB [27], SWELL-KW [28], SUSAS [29], DDD [30], WESAD [31].
- **Repository (F3)**: All the datasets have been scored with *yes* because are hosted in an established data repository:

DRIVE-DB:	PHYSIONET → https://www.physionet.org/static/published-projects/drivedb/stress-recognition-in-automobile-drivers-1.0.0.zip	*yes*
SWELL-KW:	EASY DANS → https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:58624/tab/2	*yes*
SUSAS:	LDC CATALOG → https://catalog.ldc.upenn.edu/LDC99S78	*yes*
DDD:	OSF → https://osf.io/c42cn/	*yes*
WESAD:	SCIEBO → https://uni-siegen.sciebo.de/s/pYjSgfOVs6Ntahr	*yes*

- **Download (A1)**: Three datasets have been scored with *yes* because can be downloaded for free from the repository, one with *yesbut* and one with *no*:

DRIVE-DB:	https://www.physionet.org/static/published-projects/drivedb/stress-recognition-in-automobile-drivers-1.0.0.zip	*yes*
SWELL-KW:	Downloading for free is possible but only after filling a registration form	*yesbut*
SUSAS:	Downloading for free is not possible because it is necessary to pay fees	*no*
DDD:	https://osf.io/c42cn/files/	*yes*
WESAD:	https://uni-siegen.sciebo.de/s/pYjSgfOVs6Ntahr/download	*yes*

- **Versioning (R2)**: Two datasets have been scored with *yes*, two with *no*, and one with *notbut*:

DRIVE-DB:	Version 1.0.0 (https://www.physionet.org/content/drivedb/1.0.0/)	*yes*
SWELL-KW:	Dataset's version information is not available, however, the EASY DANS platform has a versioning system	*nobut*
SUSAS:	Dataset's version information is not provided.	*not*
DDD:	Version 1.0, moreover the OSF platform has a versioning system (https://help.osf.io/hc/en-us/articles/360019738694-File-Revisions-and-Version-Control)	*yes*
WESAD:	Dataset's version information is not provided	*not*

- **Contact (R3)**: Three datasets have been scored with *yes* because contact information is found for the creator of the dataset, and two with *nobut*:

DRIVE-DB:	Creator's contact information is not found but there is a platform's one: webmaster@physionet.org	*nobut*
SWELL-KW:	wessel.kraaij@tno.nl	*yes*
SUSAS:	Creator's contact information is not found but there is a platform's one: ldc@ldc.upenn.edu	*nobut*
DDD:	ipavlidis@uh.edu	*yes*
WESAD:	kvl@eti.uni-siegen.de	*yes*

- **Citation (R4)**: All the datasets have been scored with *yes* because they can be referenced: DRIVE-DB [27], SWELL-KW [28], SUSAS [29], DDD [30], WESAD [31].
- **License (R5)**: Three datasets have been scored with *yes*, and two with *nobut*:

DRIVE-DB:	https://www.physionet.org/content/drivedb/view-license/1.0.0/	*yes*
SWELL-KW:	https://dans.knaw.nl/en/about/organisation-and-policy/legal-information/DANSLicence.pdf	*yes*
SUSAS:	https://catalog.ldc.upenn.edu/license/ldc-non-members-agreement.pdf	*yes*
DDD:	Not license found but OSF terms and conditions are available	*nobut*
WESAD:	Not license found but a disclaimer is available	*nobut*