Research Article

Maria Rauschenberger* and Ricardo Baeza-Yates

How to Handle Health-Related Small Imbalanced Data in Machine Learning?

https://doi.org/10.1515/icom-2020-0018

Abstract: When discussing interpretable machine learning results, researchers need to compare them and check for reliability, especially for health-related data. The reason is the negative impact of wrong results on a person, such as in wrong prediction of cancer, incorrect assessment of the COVID-19 pandemic situation, or missing early screening of dyslexia. Often only small data exists for these complex interdisciplinary research projects. Hence, it is essential that this type of research understands different methodologies and mindsets such as the *Design Science Methodology*, *Human-Centered Design* or *Data Science* approaches to ensure interpretable and reliable results. Therefore, we present various recommendations and design considerations for experiments that help to avoid over-fitting and biased interpretation of results when having small imbalanced data related to health. We also present two very different use cases: early screening of dyslexia and event prediction in multiple sclerosis.

Keywords: Machine Learning, Human-Centered Design, HCD, interactive systems, health, small data, imbalanced data, over-fitting, variances, interpretable results, guidelines

ACM CCS: Computing methodologies → Machine learning, Computing methodologies → Crossvalidation, Human-centered computing → Human computer interaction (HCI), Social and professional topics → People with disabilities

1 Introduction

Independently of the source of data, we need to understand our machine learning (ML) results to compare them and validate their reliability. Wrong (*e.g.,* wrong interpreted, compared) results on critical domains such as in health data can cause an immense individual or social harm (*e.g.,* the wrong prediction of a pandemic) and therefore, must be avoided.

In this context, we talk about *big data* and *small data*, which depend on the research context, profession, or mindset. We usually use the term "*big data*" in terms of size, but other key characteristics are usually missing such as variety and velocity [5, 16]. The choice of algorithm depends on the size, quality, and nature of the data set, as well as the available computational time, the urgency of the task, and the research question. In some cases, small data is preferable to big data because it can simplify the analysis [5, 16]. In some circumstances, this leads to more reliable data, lower costs, and faster results. In other cases, only small data is available, *e.g.,* in data collections related to health since each participant (*e.g.,* patient) is costly in terms of time and resources. This is the case when participants are difficult to contact due to technical restrictions (*e.g.,* no Internet) or data collecting is still ongoing, but results are urgently needed as in the COVID-19 pandemic. Some researchers prefer to wait until they have larger data sets, but this means people waiting with less hope for help. Therefore, researchers have to make the best of limited data sets and avoid over-fitting, being aware of issues such as *imbalanced data, variance, biases, heterogeneity of participants, or evaluation metrics*.

In this work we address the main criteria to avoid overfitting and taking care of imbalanced data sets related to health from a previous research experience with different small data sets related to early and universal screening of dyslexia [32, 34, 39]. We mean by universal screening of dyslexia a language-independent screening. Our main contribution is a list of recommendation when using small imbalanced data for ML predictions. We also suggest an approach for collecting data from online experiments with interactive systems to control, understand and analyze the

***Corresponding author: Maria Rauschenberger,** Max Planck Institute for Software Systems, Saarbrücken, Germany; and University of Applied Science, Emden, Germany, e-mail: rauschenberger@mpi-sws.org, ORCID: https://orcid.org/0000-0001-5722-576X
Ricardo Baeza-Yates, Khoury College of Computer Sciences, Northeastern University, Silicon Valley, CA, USA, e-mail: rbaeza@acm.org, ORCID: https://orcid.org/0000-0003-3208-9778

data. Our use cases show the complexity of interpreting machine learning results in different domains or contexts. We do not claim completeness and we see our proposal as a starting point for further recommendations or guidelines.[1] We focus mainly in newly designed interactive systems but consider also existing systems and data sets.

The rest of the paper is organized as follows: Section 2 gives the background and related work. Section 3 explains our approach to collect data from interactive systems with Design Science Research Methodology (DSRM) and Human-Centered Design (HCD), while Section 4 describes the general considerations of a research design. In Section 5 we propose our guidelines for small imbalanced data with machine learning while we show in Section 6 design considerations for different uses cases as well as examples. We finish with conclusions and future work in Section 7.

2 Background and Related Work

Interdisciplinary research projects require a standardized approach, like the *Design Science (DS) Research Methodology (DSRM)* [31], to compare results with different methodologies or mindset. A standardized approach for the design of software products, like the *Human-Centered Design (HCD)* [26], is needed to ensure the quality of the software by setting the focus on the users' needs. We explain these approaches, field of work, and advantages briefly to stress the context of our hybrid approach. Since the HCD and DSRM methods are not so well known in Machine Learning, next, we explain the basics of them to understand also the similarities of each method.

2.1 Design Science Research Methodology

The *Design Science Research Methodology (DSRM)* supports the standardization of design science, for example, to design systems for humans. The DSRM provides a flexible and adaptable framework to make research understandable within and between disciplines [31]. The core elements of DSRM have their origins in human-centered computing and are complementary to the human-centered design framework [25, 26]. DSRM suggests the following six steps to carry out research: *problem identification and motivation, the definition of the objectives for a solution, design and development, demonstration, evaluation, and communication.*

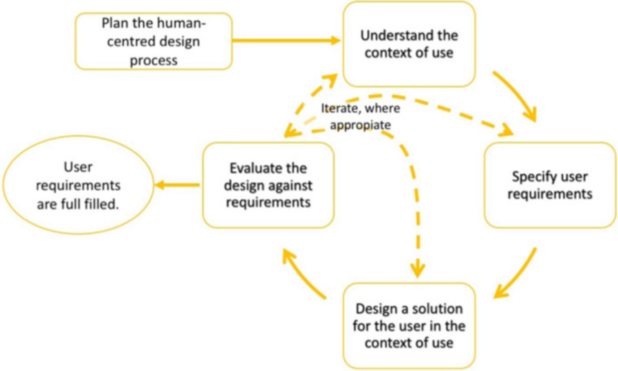
Plan the human-centred design process
Understand the context of use
Iterate, where appropiate
Specify user requirements
Evaluate the design against requirements
User requirements are full filled.
Design a solution for the user in the context of use

Figure 1: Activities of the human-centered design process adapted from [27].

The information system design theory can be considered to be similar to social science or theory-building as a class of research, e. g., to describe the prescriptive theory how a design process or a social science user study is conducted [54]. However, designing systems was not and is still not always regarded to be as valuable research as "*solid-state physics or stochastic processes*" [49]. One of the essential attributes for design science is a system that targets a new problem or an unsolved or otherwise important topic for research (quoted after [31] and [22]). If research is structured in the six steps of DSRM, a reviewer can quickly analyze it by evaluating its contribution and quality. Besides, authors do not have to justify a research paradigm for system design in each new thesis or article. Recently, DSRM and machine learning are combined for the prediction of dyslexia, or developing standards for machine learning projects in business management [24].

2.2 Human-Centered Design

The *Human-Centered Design* (HCD) framework [26] is a well-known methodology to design interactive systems that takes the whole design process into account and can be used in various areas: health related applications [2, 20, 35, 51], remote applications (Internet of things) [45], social awareness [53], or mobile applications [2, 36]. With HCD, designers focus on the users' needs when developing an interactive system to improve *usability* and *user experience*.

The HCD is an iterative design process (see Figure 1). HCD is used to designing a user-centric explainable AI frameworks [55] or to understand HCI researchs to develop

[1] Our template for self-reporting small data with our guidelines is available at https://github.com/Rauschii/smalldataguidelines

explainable systems [1]. Usually, the process starts with the planning of the HCD approach itself. After that, the (often interdisciplinary) design team members (*e.g.*, UX designers, programmers, visual designers, project managers or scrum masters) define and understand the context of use (*e.g.*, at work in an open office space). Next, user requirements are specified and can result in a description of the user requirements or a *persona* to communicate the typical user's needs to, *e.g.*, the design team [6]. Subsequently, the system or technological solution is designed with the defined scope from the context of use and user requirements. Depending on the skills or the iterative approach, the designing phase can produce a (high- or low-fidelity) prototype or product as an artifact [6]. A low-fidelity prototype, such as a paper prototype, or a high-fidelity prototype, such as an interactive designed interface, can be used for an iterative evaluation of the design results with users [3].

Ideally, the process finishes when the evaluation results reach the expectations of the user requirements. Otherwise, depending on the goal of the design approach and the evaluation results, a new iteration starts either at understanding the context of use, specifying the user requirements, or re-designing the solution.

Early and iterative testing with the user in the context of use is a core element of the HCD and researcher observe users' behavior to avoid unintentional use of the interactive system. This is especially true for new and innovative products, as both the scope of the context of use and the user requirements are not yet clear and must be explored. Early and iterative tests results often in tiny or small data and the analysis of such data can help making design decisions.

There are various methods and artifacts which can be included in the design approach depending, (*e.g.*, on the resources, goals, context of use, or users) to observe, measure and explore users' behavior. Typical user evaluations are, for example, the five-user study [30], the User Experience Questionnaire (UEQ) [23, 37, 40], observations, interviews, or the think-aloud protocol [11]. Recently, HCD and ML have been combined such that HCI researchers can design explainable systems [1]. Methods can be combined to get quantitative and/or qualitative feedback, and the most common sample size at the Computer Human Interaction Conference (CHI) in 2014 was 12 participants [12]. With small testing groups ($n < 10 - 15$) [12], mainly qualitative feedback is obtained with (semi-structured) interviews, think-aloud protocol, or observations. Taking into account the guidelines for conducting questionnaires by rules of thumb, like the UEQ could be applied from 30 participants to obtain quantitative results [41].

2.3 Related Work

The time has passed since machine learning algorithms have produced results without the need for explanation. This has changed due to a lack of control and the need for explanation in tasks affecting people like disease prediction, job candidate selection, or risk of committing a crime act again [1]. These automatic decisions could impact humans life negatively due to biases in the data set and need to be made transparent [4].

Recently, explainable user-centric frameworks [55] have been published to establish an approach across fields. But data size or imbalanced data are not mentioned as well as how to avoid over-fitting which makes a difference when using machine learning models.

As in the beginning of machine learning, today small data is used by models in spite of the focus in big data [16]. But the challenge to avoid over-fitting remains [5] and rises with imbalanced data or data with high variances. Avoiding over-fitting in health care scenarios is especially important as wrong or over interpretation of results can have major negative impacts on individuals. Current research is focusing either on collecting more data [44], develop new algorithms and metrics [13, 19], or over- and under-sampling [19]. But to the best of our knowledge, a standard approach for the analysis of a small imbalanced data sets with variances when doing machine learning classification or prediction in health is missing.

Hence, we propose some guidelines based in our previous research to consider the context given above. This is very important as most institutions in the world will never have big data [5].

3 Collecting and Analyzing Data

An interdisciplinary research project requires a standardized approach to allow other researchers to evaluate and interpret results. Therefore, we combine the *Design Science (DS) Research Methodology (DSRM)* [31] with the *Human-Centered Design (HCD)* [26] to collect data for new interactive systems.

Researchers combine methodologies or approaches and need to evaluate results from other disciplines. Combining discipline techniques is a challenge because of different terms, methods, or communication within each discipline. For example, the same term, such as *experiments*, can have a different interpretation in *data science* versus *human computer interaction (HCI)* approaches. In HCI, *experiments* mainly refer to user studies with humans,

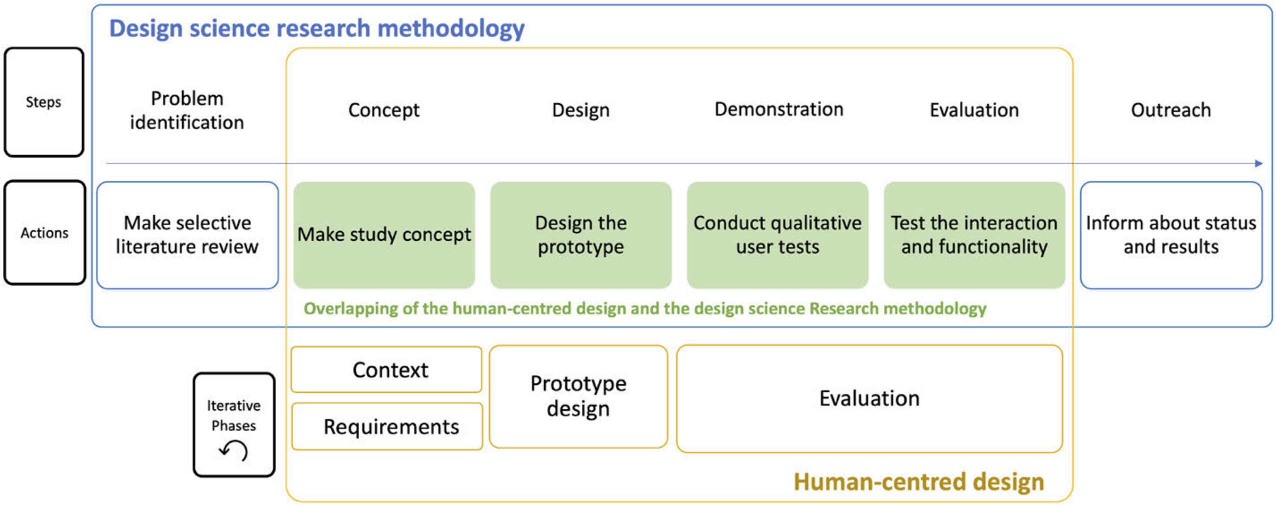
Design science research methodology
Steps
Problem identification
Concept
Design
Demonstration
Evaluation
Outreach
Actions
Make selective literature review
Make study concept
Design the prototype
Conduct qualitative user tests
Test the interaction and functionality
Inform about status and results
Overlapping of the human-centred design and the design science Research methodology
Iterative Phases
Context
Requirements
Prototype design
Evaluation
Human-centred design

Figure 2: Integration of the *Human-centered Design* in the *Design Science Research Methodology.*

whereas in data science, experiments refer to running algorithms on data sets. HCD is not well known in the machine learning community but provides methods to solve current machine learning challenges, such as how to avoid collecting bias in data sets from interactive systems.

We combine HCD and DSRM because of their similarities and advantages as explained before in Section 2. Most common ways in big data analysis are existing interactive systems or already collected data sets which we describe very briefly.

3.1 New Interactive Systems

It is a challenge to collect machine learning data sets with interactive systems since the system is designed not only for the users' requirements but also for the underlying research purpose. The DSRM provides the research methodology to integrate the research requirements while HCD focuses on the design of the interactive systems with the user.

Here we show how we combined the six DSRM steps with the *Actions* and match them with the four HCD phases (see Figure 2, black boxes). The blue boxes are only related to the DSRM, while the green boxes are also related to the HCD approach. The four green boxes match the four HCD phases (see Figure 2, yellow boxes). Next, we describe each of the six DSRM steps following Figure 2 with an example from our previous research on early screening of dyslexia with a web game using machine learning [32, 34, 39].

First, we do a selective literature review to identify the problem, *e. g.*, there is the need of early, easy and language-independent screening of dyslexia. This results in a concept, *e. g.*, for targeting the language-independent screening of dyslexia using games and machine learning. We then describe how we design and implement the content and the prototypes as well as how we test the interaction and functionality to evaluate our solution.

In our example, we designed our interactive prototypes to conduct online experiments with participants with dyslexia using the *human-centered design* [26].

With the HCD, we focus on the participant and the participant's supervisor (*e. g.*, parent/legal guardian/teacher/therapist) as well as on the context of use when developing the prototype for the online experiments to measure differences between children with and without dyslexia.

The user requirements and context of use define the content for the prototypes, which we design iteratively with the knowledge of experts. In this case, the interactive system has an integration of *game elements* to apply the concept of *gamification* [42].

Furthermore, HCD enhances the design, usability, and user experience of our prototype by avoiding external factors that could unintentionally influence the collected data. In particular, the early and iterative testing of the prototypes helps to prevent unintended interactions from participants or their supervisors.

Example iterations are internal feedback loops of human-computer interaction experts or user tests (*e. g.*, five-user test). For instance, we discovered that to interact with a tablet, children touch quickly multiple times. Because of the web implementation technique used, a double click on a web application to *zoom* in, which was not good in a tablet. Therefore, we controlled the layout setting for mobile devices to avoid the *zoom*-effect on tablets, which caused interruptions during the game [32]. The evaluation requires the collection of remote data with the experimental design to use the dependent measures for statistical analysis and prediction with machine learning classifiers.

When taking into account participants with a learning disorder, in our case, participants with dyslexia, we need to address their needs [38] in the design of the application and the experiment as well as consider the ethical aspects [7]. As dyslexia is connected to nine genetic markers and reading ability is highly hereditary [14], we support readability for participants' supervisors (who could be parents) with a large font size (minimum 18 points) [43].

3.2 Existing Interactive Systems or Data Sets

In big data analysis, it is very common to validate new algorithms with existing data sets and to compare to a baseline algorithm. The details how the data from this existing data sets has been collected or prepared is mainly unknown. Mainly, the annotation and legend is provided as a description, *e. g.*, https://www.kaggle.com/datasets. This could lead to false interpretation as the circumstances are not clear. Example could be time duration, time pressure, missing values, clean values lead to new values or data entries have been excluded but are valuable for the interpretation. As a result, data sets can be more biased because if missing entries or information. Small data has the same effects and less data means more influence of missing information. In small critical data, by which we mean personal data, health-related or even clinical data, this data sets are protected and only shared with a limited group of people. But some data is shared publicly, *e. g.*, https://www.kaggle.com/ronitf/heart-disease-uci. As mentioned before, descriptions are very limited and the advantages to collect your own data set are the knowledge researchers get over their own data sets. This knowledge should be passed on to other researchers when the data sets are made public. The circumstances when the data set has been collected is important for the analysis. For example, during a pandemic the information shared in social media might increase as most people are at home using a internet device. In a few years from now, researchers might not be aware of these circumstances anymore and the information needs to be passed on when analysis the collect social media data sets, *e. g.*, screening time and health in 2020 *vs.* 2030.

Researchers should provide and be aware of the surrounding information regarding society when collecting and analysis data from existing systems or data sets.

Table 1: Example aspects that can have a big influence on small data analysis.

Human-Centered Design	Data Science/ Machine Learning
• Different terms, *e. g., experiments*	• ≠ *experiments*
• (tiny) Small data	• Big data
• Iterative design	• Data set knowledge
• DIN, ISO, various methods	• No global standards
• Balanced	• Imbalanced
• Multiple Testing; Bonferroni	• Multiple Testing
…	…

3.3 Main Challenges Combining HCD and ML

As described before, combining discipline-specific techniques can be a challenge due to very simple and easy to solve issues such as, *e. g.*, same terms different meaning. We would like to raise awareness for certain aspects as this aspects can have a bigger impact on the results of small data (see Table 1).

We should consider that even the term small data is probably interpreted differently, as small data in data science might reference to 15.000 data points [56] for image analysis while HCD reference to around 200 or much less participants.

The focus in the HCD is the iterative design of a interactive systems which can be a website or a voice assistant. Also, it is an approach which includes the persona and context for the evaluation. Data scientist mainly get a data set and focus on the analysis of the data without traditional the context of how this data has been collected.

4 Research Design Considerations

A quasi-experimental study helps to collect dependent variables from an interactive system, which we use as features for the machine learning models later. In this way, there is control over certain variables such as participant attributes, which then assigns participants to either the control or the experimental group [17]. An example of such an attribute could be whether or not one has a dyslexia diagnosis.

In a *within-subject design*, all participants take part in all study conditions, *e. g.*, tasks or game rounds. When applying a *within-subject* design, the conditions need to be randomized to avoid *systematic or order effects* produced by order of the conditions. These unwanted effects can be

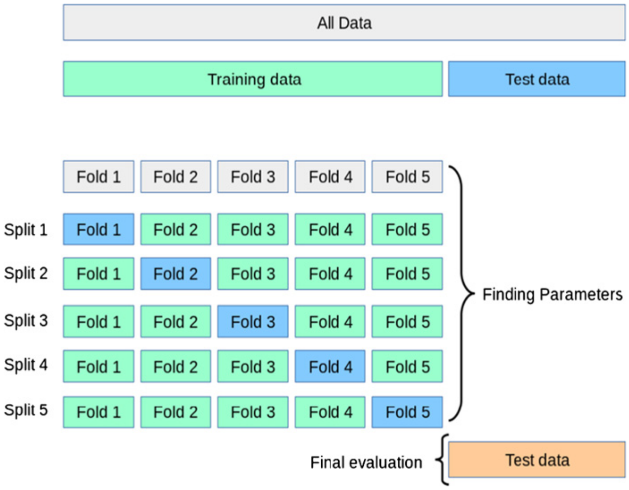

Figure 3: Cross-validation approach [46].

avoided by counterbalancing the order of the conditions, for example with Latin Squares [17].

The advantage of a *repeated-measures design* in a *within-subject design* is that participants can engage in multiple conditions [17]. When participant attributes such as age or gender are similar in different groups, a repeated-measures design is more likely to reveal the effects caused by the dependent variable of the experiment.

When conducting a *within-subject* design with a repeated measures design, and assuming a non-normal and non-homogeneous distribution for independent participant groups, a non-parametric statistical test is needed, such as the *Mann-Whitney-Wilcoxon Test* [17]. As for psychology in HCD, multi-variable testing must be addressed to avoid having significance by chance. This can be achieved by using a method such as *Bonferroni-Correction* and having a clear hypothesis.

Dependent measures are used to find, for example, differences between variables [17], while features are used as input for the classifiers to recognize patterns [8]. Machine learning is a data-driven approach in which the data is explored with different algorithms to minimize the objective function [15]. In the following we refer to the implementation of the Scikit-learn library (version 0.21.2) if not stated otherwise [48]. Although a hypothesis is followed, optimizing the model parameters (multiple testing) is not generally considered problematic (as it is in HCD) unless we are *over-fitting*, as stated by Dietterich in 1995:

> *"Indeed, if we work too hard to find the very best fit to the training data, there is a risk that we will fit the noise in the data by memorizing various peculiarities of the training data rather than finding a general predictive rule. This phenomenon is usually called over-fitting."* [15]

If enough data is available, common practice *holds out* (that is separating data for training, test or validation) a percentage to evaluate the model and to avoid over-fitting, *e.g.*, a test data set of 40 % of the data [46]. A validation set (holding out another percentage of the data) can be used to, say, evaluate different input parameters of the classifiers to optimize results [46], *e.g.*, accuracy or F1-score. Holding out part of the data is only possible if a sufficient amount of data is available. As models trained on small data are prone to develop over-fitting due to the small sample and feature selection [28], cross-validation with k-folds can be used to avoid over-fitting when optimizing the classifier parameters (see Figure 3). In such cases, the data is split into training and test data sets. A model is trained using $k - 1$ subsets (typically 5-folds or 10-folds) and evaluated using the missing fold as test data [46]. This is repeated k times until all folds have been used as test data, taking the average as final result. It is recommended that one hold out a test data set while using cross-validation when optimizing input parameters of the classifiers [46]. However, small data sets with high variances are not discussed.

So far, mainly in data science we talk about big and small data [5] and small data can be still 15,000 data points in, *e.g.*, image analysis with labeled data [56]. In HCD we may have less than 200 data points, as explained before. Hence, it depends on the domain and context what is considered small and we suggested to consider talking about *tiny* data in the case of data under a certain threshold, *e.g.*, 200 subjects or data points. We should consider to distinguish small and tiny data as they need to be analyzed differently, *i.e.*, tiny data cannot be separated in neither a test or validation set to do data science multiple testing and/or parameter optimization.

Model-evaluation implementations for cross-validation from Scikit-learn, such as the *cross val score* function, use scoring parameters for the quantification of the quality of the predictions [47]. For example, with the parameter *balanced accuracy* imbalanced data sets are evaluated. The parameter *precision* describes the classifiers ability "*not to label as positive a sample that is negative*" [47]. Whereas the parameter *recall* "*is the ability of the classifier to find all the positive samples*" [47]. As it is unlikely to have a high precision and high recall, the *F1-score* (also called F-measure) is a "*weighted harmonic mean of the precision and recall*" [47]. Scikit-learn library suggests different implementations for computing the metrics (*e.g.*, recall, F1-score) and the confusion matrix [46]. The reason is that the *metric function* reports over all (cross-validation) fold, whereas the *confusion matrix function* returns the probabilities from different models.

5 Proposed Recommendations

Based in our previous research [32, 34, 39] and our workshop [33] we propose the main criteria that should be considered when applying machine learning classification for small data related to health. We present an overview of the criteria to avoid over-fitting in the following:

Precise data set In the best-case scenario, no missing values, and participants' attributes are similarly represented, *e.g.*, age, gender, language.

Biases Data sets having biases are very likely, in health data gender or age biases are even normal. Many factors determine the quality of the data set, and we recommend accepting the existence of possible biases and start with the "awareness of its existence" [4].

Hypothesis Use a hypothesis from the experimental design, confirm with existing literature, and pre-evaluated with, *e.g.*, statistical analysis of the dependent variables to avoid significance or high predictions by chance.

Domain knowledge Interdisciplinary research topics need a deeper understand of each discipline and the controversy when analyzing or training the data, *e.g.*, (multiple testing) HCD *vs.* Data Science or domain experts such as learn-therapist for dyslexia or virologists.

Data set knowledge Knowledge about how the data has been collected to identify external factors, *e.g.*, duration, society status, pandemic situation, technical issues.

Simplified prediction Depending on the research question and certainty of correlations, a binary classification instead of multiple classifications is beneficial to avoid external factors and understand results better.

Feature Selection Feature selection is essential in any machine learning model. However, for small data, the dependent variables from the experimental design can address the danger of selecting incorrect features [28] by taking into account previous knowledge. Therefore, pre-evaluate dependent variables with traditional statistical analysis and then use the dependent variables as input for the classifiers [8].

Optimizing input parameters Do not optimize input parameters unless data sets can hold out test and validation sets. Hold out tests and cross-validation are proposed by Scikit-learn 0.21.2 documentation to evaluate the changes [46] and to avoid biases [52].

Variances When imbalanced data show high variances, we recommend not to use over-sampling as the added data will not represent the class variances. We recommend not under-sampling data sets with high variances when data sets are already minimal and would reduce it to $n < 100$. The smaller the data set, the more likely it is to produce the unwanted over-fitting.

Over- and under-sampling Over- and under-sampling can be considered when data sets have small variances.

Imbalanced Data Address this problem with models made for imbalanced data (*e.g.*, Random Forest with class weights) or appropriate metrics (*e.g., balanced accuracy*).

Missing or wrong values Missing data can be imputed by using interpolation/extrapolation or via predictive models. Wrong values can be verified with domain knowledge or data set knowledge and then excluded or updated (if possible).

These criteria are the starting point of machine learning solutions on health-related small data analysis as this can have a significant impact on specific individuals and the society.

6 Two Use Cases

Beside the considerations to design research studies in an interdisciplinary project and the criteria explained before, there are for each domain, data size and task different possibilities to explore the data, *e.g.*, image analysis [56], game measure analysis [34, 44], or eye-tracking data [29]. Here we describe two possible approaches in different data types and domains to raise awareness of possible pitfalls for future use cases.

Small or tiny data can have the advantage of being precise by which we mean, *e.g.*, less/no missing data, specific target group. Data sets can be collected from social media (*e.g.*, Twitter, Reddit,..), medical studies (*e.g.*, MRT, dyslexia diagnose) or user studies (*e.g.*, HCD user evaluation). As mentioned before, small or tiny data could answer specific questions and reduce the complexity of the analysis, by simplifying the prediction, *e.g.*, binary classification. Missing out influencing factors is a limitation which is why tiny data set experiments should follow a hypothesis as in HCD.

In big data analysis most probably there is missing data which is then interpolated or predicted (average to assume potential user behavior) with the already existing values from other, *e.g.*, subjects. In small or tiny data (especially with variances), dealing with missing data entries is a challenge due to uncertainty of the information value. This means, we cannot predict the missing value as this might add biases.

We follow with two use cases to explain different data issues, possible pitfalls and possible solutions to gain more reliable machine learning results.

6.1 Early Dyslexia Screening

We explain further our approach to avoid over-fitting and overly interpret machine learning results with the following use case: finding a person with dyslexia to achieve early and universal screening of dyslexia [32, 34]. The prototype is a game with a visual and auditory part. The content is related to indicators that showed significant differences in lab studies among children with and without dyslexia. First, the legal guardian answered the background questionnaire (*e.g.*, age, official dyslexia diagnoses yes/maybe/no), and then children played the web game once. Dependent variables have been derived from previous literature and then matched to game measures (*e.g.*, number of total clicks, duration time). A demo presentation of the game is available at https://youtu.be/P8fXMZBXZNM.

The example data set has 313 participants, with 116 participants with dyslexia and 197 participants without dyslexia, the control group (imbalance of 37 % *vs.* 63 %, respectively).

A precise data set helps to avoid external factors and reveals biases within the data sets due to missing data or missing participants. For example, one class is represented by one feature (*e.g.*, language) due to missing participants from that language, and therefore the model predicts a person with dyslexia mainly by the feature language, which is not a dependent variable for dyslexia.

Although dyslexia is more a spectrum than a binary classification, we rely on current diagnostic tools [50] such as the DRT [21] to select our participants' groups. Therefore, a simple binary classification is representative although dyslexia is not binary. The current indicators of dyslexia require the children to have minimal linguistic knowledge, such as phonological awareness, to measure reading or writing mistakes. These linguistic indicators for dyslexia in diagnostic tools are probably stronger as language-independent indicators because a person with dyslexia shows a varying severity of deficits in more than one area [9]. Additionally, in this use case, participants call raised awareness from parents who suspected their child of having dyslexia but did not have an official diagnosis. We, therefore, decided for precise data set on children who have a formal diagnosis and show no sign of dyslexia (control group) to avoid external factors and focus on cases with probably more substantial differences in behavior.

Notably, in a new area with no published comparison, a valid and clear hypothesis derived from existing literature confirms that the measures taken are connected to the hypothesis. While a data-driven approach is exploitative and depends on the data itself, we propose to follow a hypothesis to not over-interpret anomalies for small data analysis. We agree that anomalies can help to find new research directions but should not be taken as facts and instead explore them to find the origin as for the example of one class represented by one feature (see above). This is also connected to *Which features to collect and analyze?* as this could mean having correlations by chance due to the small data set or selected participants with features similar to the multi-variable testing in HCD. As far as we know, there is no similar *Bonferroni-Correction* for machine learning in small data.

We propose to use different kinds of features (input parameters) depending on different hypotheses derived from literature. For example, at this point, the central two theories are that dyslexia is related to auditory and visual perception. We, therefore, also separated our features for different machine learning test related to auditory or visual to evaluate if one of the theories is more valid. This approach is taking advantage of the machine learning techniques without over-interpretation results and, at the same time, takes into account previous knowledge of the target research area with hypotheses as done in HCD.

At this point, we could not find a *rules of thumb* or literature recommendation when to over- or under-sample a data set. Also, no approach for variances within a data set and over- or under-sampling are discussed. We propose to not over- and under-sample for data sets having high variances.

When comparing machine learning classification results, the metrics for comparison should not be only (balanced) accuracy as this describes mainly the accuracy of the model and does not focus on the screening of dyslexia. Obtaining both high precision and high recall is unlikely, which is why researchers reported the F1-score (the weighted average between precision and recall) for dyslexia to compare the model's results [34]. However, as in this case false positives are much more harmful than false negatives (that is, missing a person with dyslexia), we should focus on the dyslexia class recall.

6.2 Multiple Sclerosis

Here we explain another use case with the focus on predicting the next events and treatments for Multiple Sclerosis (MS) [18]. Typically, MS starts with symptoms such as limbs numbness or reduced vision for days or weeks that usually improve partially or completely. These events are followed by quiet periods of disease remission that can last months or even years until they relapse. The main challenges are the low frequency of those affected, the long periods to gather meaningful events, and the size of some data sets.

Multiple Sclerosis (MS) is a degenerative nervous disease where intervention at the right stage and time is crucial to slowdown the process. At the same time, there is a high variance among patients and then personalized health care is preferable. So, today, clinical decisions are based in different parameters such as age, disability milestones, retinal atrophy, brain lesion activity and load. Collecting enough data is difficult as is not a common disease, where less than two people per million will suffer it (that is, 5 orders of magnitude less frequent than the previous use case). Hence, here the approach to collect data has to be collaborative. For example, the Sys4MS European project was able to have a cohort of 322 patients coming from four different hospitals in as many countries: Germany, Italy, Norway and Spain [10, 57]. This means on one hand that we are sure these participants are affected of MS but the life style, history or regions are heterogeneous. Additional parameters can complicate the prediction of events because, *e. g.,* interpolating data, already, adds noise due to the variance among participants. Therefore, we always need to collect more individual data over a longer period.

In addition, collecting this data needs time as the progress of the disease takes years and hence you may need to follow patients for several years to have a meaningful number of events. During this period, very different types of data are collected at fixed intervals (months), such as clinical (demographics, drugs usage, relapses, disability level, etc.), imaging (brain MRI and retinal OCT scans), and multi-modal omics data. The latter include cytomics (blood samples), proteomics (mass spectrometry) and genomics (DNA analysis using half a million genetic markers). For example, for the disability level there are no standards and there exists close to ten different scales and tests. Hence collecting this data is costly in terms of time and personnel (*e. g.*, brain and retinal scans, or DNA analysis). In addition, due to the complexity of some of these techniques involved, the data will have imprecision. Additionally, not all participants do all tests which lead to missing data, as they come from different hospitals. This data completeness heterogeneity between participants adds an additional level of complexity.

The imbalance here is different to the previous use case as is also hard to have healthy patients (baseline) that are willing to go through this lengthy data collection process. Indeed, the final data set has only 98 people in the control group (23 %). Originally, we had 100 features but we selected only the ones that had at least 80 % correlation with each of the predictive labels. We also consider only features with at least 80 % of their values and similarly patients with values for all the features. As a result, we also have feature imbalance with 24 features in the clinical data, 5 in the imaging data and 26 in the omics data. Worse, due to the collaborative nature of the data collection and the completeness constraint, the number of patients available for each of the 9 prediction problems drastically varies, as not all of them had all the features and/or the training labels, particularly the omics features. Hence, the disease cases varied from 29 to 259 while the control cases varied from 4 to 78. This resulted in imbalances of as low as 7 % in the control group. Therefore, we could even talk about tiny data instead of small data.

Machine learning predictors for the severity of the disease as well as to predict different events to guide personalized treatments were trained with this data by using random forests [18], which was the best of the techniques we tried. To take care of the imbalance, we used weighted classes and measures. Parameters were not optimized to avoid over-fitting. The recall obtained were over 80 % for seven of predictive tasks and even over 90 % for three of them. One surprising result is that in 5 of the tasks, using just the clinical data gave the best result, showing that possibly the rest of the data was not relevant or noisy. For other 3 tasks, adding the image data was just a bit better. Finally, only in one task the omics data made a difference and a significant one. However, this last comparison is not really fair, as most of the time the data sets having the omics data were much smaller (typically one third of the size and also were the most imbalanced). Hence, when predictive results suggest that certain tests are better than others, they may have different priorities to obtain better data sets.

7 Conclusion and Future Directions

We propose the first step towards guidelines when exploring health-related small better *tiny* imbalanced data sets with various criteria. We show two use cases and reveal opportunities to discuss and develop this further with, *e. g.*, new machine learning classification for imbalanced data considering small data. We explained analytical decisions for each use case depending on the relevant criteria.

Our proposed guidelines are a starting point and need to be adapted for each use case. Therefore, we provide a template for researchers to follow them for their projects available at https://github.com/Rauschii/smalldataguidelines. Additionally, we encourage other researchers to update the use case collection in the template with their own projects for further analysis.

Future work will explore the limits of small data analysis with machine learning techniques, existing metrics, and models, as well as approaches from other disciplines to verify the machine learning results.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems – CHI'18*. ACM Press, New York, New York, USA, 1–18. https://doi.org/10.1145/3173574.3174156.

[2] Muneeb Imtiaz Ahmad and Suleman Shahid. 2015. Design and Evaluation of Mobile Learning Applications for Autistic Children in Pakistan. In *INTERACT* (Lecture Notes in Computer Science, Vol. 9296), Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Springer International Publishing, Cham, 436–444. https://doi.org/10.1007/978-3-319-22701-6.

[3] Jonathan Arnowitz, Michael Arent, and Nevin Berger. 2007. *Effective Prototyping for Software Makers*. Morgan Kaufmann, unknown. 584 pages. https://www.oreilly.com/library/view/effective-prototypingfor/9780120885688/.

[4] Ricardo Baeza-Yates. Bias on the web. *Commun. ACM* 61, 6 (may 2018), 54–61. https://doi.org/10.1145/3209581.

[5] Ricardo Baeza-Yates. 2018. *BIG, small or Right Data: Which is the proper focus?* https://www.kdnuggets.com/2018/10/big-small-right-data.html [Online, accessed 22-July-2019].

[6] Protima Banerjee. 2004. *About Face 2.0: The Essentials of Interaction Design*. Vol. 3. Wiley Publishing, Inc., USA, 223–225. https://doi.org/10.1057/palgrave.ivs.9500066.

[7] Gerd Berget and Andrew MacFarlane. 2019. Experimental Methods in IIR. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval – CHIIR'19*. ACM Press, New York, New York, USA, 93–101. https://doi.org/10.1145/3295750.3298939.

[8] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Vol. 1. Springer Science+Business Media, LLC, Singapore, 1–738. http://cds.cern.ch/record/998831/files/9780387310732_TOC.pdf.

[9] Donald W. Black, Jon E. Grant, and American Psychiatric Association. 2016. *DSM-5 guidebook: The essential companion to the Diagnostic and statistical manual of mental disorders, fifth edition* (5th edition ed.). American Psychiatric Association, London. 543 pages. https://www.appi.org/dsm-5_guidebook.

[10] S. Bos, I. Brorson, E.A. Hogestol, J. Saez-Rodriguez, A. Uccelli, F. Paul, P. Villoslada, H.F. Harbo, and T. Berge. Sys4MS: Multiple sclerosis genetic burden score in a systems biology study of MS patients from four countries. *European Journal of Neurology* 26 (2019), 159.

[11] Henning Brau and Florian Sarodnick. 2006. *Methoden der Usability Evaluation (Methods of Usability Evaluation)* (2. ed.). Verlag Hans Huber, Bern. 251 pages. http://d-nb.info/1003981860, http://www.amazon.com/Methoden-Usability-Evaluation-Henning-Brau/dp/3456842007.

[12] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *CHI'16*. ACM, San Jose California, USA, 981–992. https://doi.org/10.1145/2858036.2858498.

[13] André M. Carrington, Paul W. Fieguth, Hammad Qazi, Andreas Holzinger, Helen H. Chen, Franz Mayr, and Douglas G. Manuel. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making* 20, 1 (2020), 1–12. https://doi.org/10.1186/s12911-019-1014-6.

[14] Greig De Zubicaray and Niels Olaf Schiller. 2018. *The Oxford handbook of neurolinguistics*. Oxford University Press, New York, NY. https://www.worldcat.org/title/oxford-handbook-ofneurolinguistics/oclc/1043957419&referer=brief_results.

[15] Tom Dietterich. Overfitting and undercomputing in machine learning. *Comput. Surveys* 27, 3 (sep 1995), 326–327. https://doi.org/10.1145/212094.212114.

[16] Julian J. Faraway and Nicole H. Augustin. When small data beats big data. *Statistics & Probability Letters* 136 (may 2018), 142–145. https://doi.org/10.1016/j.spl.2018.02.031.

[17] Andy P. Field and Graham Hole. 2003. *How to design and report experiments*. SAGE Publications, London. 384 pages.

[18] Ana Freire, Magi Andorra, Irati Zubizarreta, Nicole Kerlero de Rosbo, Steffan R. Bos, Melanie Rinas, Einar A. Høgestøl, Sigrid A. de Rodez Benavent, Tone Berge, Priscilla Bäcker-Koduah, Federico Ivaldi, Maria Cellerino, Matteo Pardini, Gemma Vila, Irene Pulido-Valdeolivas, Elena H. Martinez-Lapiscina, Alex Brandt, Julio Saez-Rodriguez, Friedemann Paul, Hanne F. Harbo, Antonio Uccelli, Ricardo Baeza-Yates, and Pablo Villoslada. to appear. Precision medicine in MS: a multi-omics, imaging, and machine learning approach to predict disease severity.

[19] Koichi Fujiwara, Yukun Huang, Kentaro Hori, Kenichi Nishioji, Masao Kobayashi, Mai Kamaguchi, and Manabu Kano. Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis. *Frontiers in Public Health* 8 (may 2020), 178. https://doi.org/10.3389/fpubh.2020.00178.

[20] Ombretta Gaggi, Giorgia Galiazzo, Claudio Palazzi, Andrea Facoetti, and Sandro Franceschini. 2012. A serious game for predicting the risk of developmental dyslexia in pre-readers children. In *2012 21st International Conference on Computer Communications and Networks, ICCCN 2012 – Proceedings*. IEEE, Munich, Germany, 1–5. https://doi.org/10.1109/ICCCN.2012.6289249.

[21] Martin Grund, Carl Ludwig Naumann, and Gerhard Haug. 2004. *Diagnostischer Rechtschreibtest für 5. Klassen: DRT 5 (Diagnostic spelling test for fifth grade: DRT 5)* (2., aktual ed.). Beltz Test, Göttingen. https://www.testzentrale.de/shop/diagnostischer-rechtschreibtest-fuer-5-klassen.html.

[22] Alan Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design Science in Information Systems Research.

MIS Quarterly 28, 1 (2004), 75. https://doi.org/10.2307/25148625.

[23] Andreas Hinderks, Martin Schrepp, Maria Rauschenberger, Siegfried Olschner, and Jörg Thomaschewski. 2012. Konstruktion eines Fragebogens für jugendliche Personen zur Messung der User Experience (Construction of a questionnaire for young people to measure user experience). In *Usability Professionals Konferenz 2012*. German UPA e.V., Stuttgart, UPA, Stuttgart, 78–83.

[24] Steven A. Hoozemans. 2020. *Machine Learning with care: Introducing a Machine Learning Project Method*. 129 pages. https://repository.tudelft.nl/islandora/object/uuid:6be8ea7b-2a87-45d9-aaa8-c82ff28d56c2.

[25] Robert R. Huffman, Axel Roesler, and Brian M. Moon. What is design in the context of human-centered computing? *IEEE Intelligent Systems* 19, 4 (2004), 89–95. https://doi.org/10.1109/MIS.2004.36.

[26] ISO/TC 159/SC 4 Ergonomics of human-system interaction. 2010. Part 210: Human-centred design for interactive systems. In *Ergonomics of human-system interaction*. Vol. 1. International Organization for Standardization (ISO), Brussels, 32. https://www.iso.org/standard/52075.html.

[27] ISO/TC 159/SC 4 Ergonomics of human-system interaction. 2018. *ISO 9241-11, Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts*. 2018 pages. https://www.iso.org/standard/63500.html, https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

[28] Anil Jain and Douglas Zongker. Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 2 (1997), 153–158. https://doi.org/10.1109/34.574797.

[29] Anuradha Kar. MLGaze: Machine Learning-Based Analysis of Gaze Error Patterns in Consumer Eye Tracking Systems. *Vision (Switzerland)* 4, 2 (may 2020), 1–34. https://doi.org/10.3390/vision4020025, arXiv:2005.03795.

[30] Jakob Nielsen. Why You Only Need to Test with 5 Users. *Jakob Nielsens Alertbox* 19 (sep 2000), 1–4. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/, http://www.useit.com/alertbox/20000319.html [Online, accessed 11-July-2019].

[31] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems* 24, 8 (2007), 45–78. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.535.7773&rep=rep1&type=pdf.

[32] Maria Rauschenberger. 2019. *Early screening of dyslexia using a languageindependent content game and machine learning*. Ph.D. Dissertation. Universitat Pompeu Fabra. https://doi.org/10.13140/RG.2.2.27740.95363.

[33] Maria Rauschenberger and Ricardo Baeza-Yates. 2020. Recommendations to Handle Health-related Small Imbalanced Data in Machine Learning. In *Mensch und Computer 2020 – Workshopband (Human and Computer 2020 – Workshop proceedings)*, Bernhard Christian Hansen and Nürnberger Andreas Preim (Ed.). Gesellschaft für Informatik e.V., Bonn, 1–7. https://doi.org/10.18420/muc2020-ws111-333.

[34] Maria Rauschenberger, Ricardo Baeza-Yates, and Luz Rello. 2020. Screening Risk of Dyslexia through a Web-Game using Language-Independent Content and Machine Learning. In *W4a'2020*. ACM Press, Taipei, 1–12. https://doi.org/10.1145/3371300.3383342.

[35] Maria Rauschenberger, Silke Füchsel, Luz Rello, Clara Bayarri, and Jörg Thomaschewski. 2015. Exercises for German-Speaking Children with Dyslexia. In *Human-Computer Interaction – INTERACT 2015*. Springer, Bamberg, Germany, 445–452.

[36] Maria Rauschenberger, Christian Lins, Noelle Rousselle, Sebastian Fudickar, and Andreas Hain. 2019. A Tablet Puzzle to Target Dyslexia Screening in Pre-Readers. In *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good – GOODTECHS*. ACM, Valencia, 155–159.

[37] Maria Rauschenberger, Siegfried Olschner, Manuel Perez Cota, Martin Schrepp, and Jörg Thomaschewski. 2012. Measurement of user experience: A Spanish Language Version of the User Experience Questionnaire (UEQ). In *Sistemas Y Tecnologias De Informacion*, A. Rocha, J.A. CalvoManzano, L.P. Reis, and M.P. Cota (Eds.). IEEE, Madrid, Spain, 471–476.

[38] Maria Rauschenberger, Luz Rello, and Ricardo Baeza-Yates. 2019. Technologies for Dyslexia. In *Web Accessibility Book* (2nd ed.), Yeliz Yesilada and Simon Harper (Eds.). Vol. 1. Springer-Verlag London, London, 603–627. https://doi.org/10.1007/978-1-4471-7440-0.

[39] Maria Rauschenberger, Luz Rello, Ricardo Baeza-Yates, and Jeffrey P. Bigham. 2018. Towards language independent detection of dyslexia with a web-based game. In *W4A'18: The Internet of Accessible Things*. ACM, Lyon, France, 4–6. https://doi.org/10.1145/3192714.3192816.

[40] Maria Rauschenberger, Martin Schrepp, Manuel Perez Cota, Siegfried Olschner, and Jörg Thomaschewski. Efficient Measurement of the User Experience of Interactive Products. How to use the User Experience Questionnaire (UEQ). Example: Spanish Language. *International Journal of Artificial Intelligence and Interactive Multimedia (IJIMAI)* 2, 1 (2013), 39–45. http://www.ijimai.org/journal/sites/default/files/files/2013/03/ijimai20132_15_pdf_35685.pdf.

[41] Maria Rauschenberger, Martin Schrepp, and Jörg Thomaschewski. 2013. User Experience mit Fragebögen messen – Durchführung und Auswertung am Beispiel des UEQ (Measuring User Experience with Questionnaires–Execution and Evaluation using the Example of the UEQ). In *Usability Professionals Konferenz 2013*. German UPA eV, Bremen, 72–76.

[42] Maria Rauschenberger, Andreas Willems, Menno Ternieden, and Jörg Thomaschewski. Towards the use of gamification frameworks in learning environments. *Journal of Interactive Learning Research* 30, 2 (2019), 147–165. https://www.aace.org/pubs/jilr/, http://www.learntechlib.org/c/JILR/.

[43] Luz Rello and Ricardo Baeza-Yates. 2013. Good fonts for dyslexia. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'13)*. ACM, New York, NY, USA, 14. https://doi.org/10.1145/2513383.2513447.

[44] Luz Rello, Enrique Romero, Maria Rauschenberger, Abdullah Ali, Kristin Williams, Jeffrey P. Bigham, and Nancy Cushen White. 2018. Screening Dyslexia for English Using HCI Measures and Machine Learning. In *Proceedings of the 2018*

International Conference on Digital Health – DH'18. ACM Press, New York, New York, USA, 80–84. https://doi.org/10.1145/3194658.3194675.

[45] Claire Rowland and Martin Charlier. 2015. *User Experience Design for the Internet of Things*. O'Reilly Media, Inc., Boston, 1–37.

[46] Scikit-learn. 2019. 3.1. Cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html [Online, accessed 17-June-2019].

[47] Scikit-learn. 2019. 3.3. Model evaluation: quantifying the quality of predictions. https://scikit-learn.org/stable/modules/model_evaluation.html#scoring-parameter [Online, accessed 23-July-2019].

[48] Scikit-learn Developers. 2019. Scikit-learn Documentation. https://scikit-learn.org/stable/documentation.html [Online, accessed 20-June-2019].

[49] Herbert A. Simon. 1997. *The sciences of the artificial, (third edition)*. Vol. 3. MIT Press, London, England. 130 pages. https://doi.org/10.1016/S0898-1221(97)82941-0.

[50] Claudia Steinbrink and Thomas Lachmann. 2014. *Lese-Rechtschreibstörung (Dyslexia)*. Springer Berlin Heidelberg, Berlin. https://doi.org/10.1007/978-3-642-41842-6.

[51] Lieven Van den Audenaeren, Véronique Celis, Vero Van den Abeele, Luc Geurts, Jelle Husson, Pol Ghesquière, Jan Wouters, Leen Loyez, and Ann Goeleven. 2013. DYSL-X: Design of a tablet game for early risk detection of dyslexia in preschoolers. In *Games for Health*. Springer Fachmedien Wiesbaden, Wiesbaden, 257–266. https://doi.org/10.1007/978-3-658-02897-8_20.

[52] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7 (feb 2006), 91. https://doi.org/10.1186/1471-2105-7-91.

[53] Torben Wallbaum, Maria Rauschenberger, Janko Timmermann, Wilko Heuten, and Susanne C.J. Boll. 2018. Exploring Social Awareness. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems – CHI'18*. ACM Press, New York, New York, USA, 1–10. https://doi.org/10.1145/3170427.3174365.

[54] Joseph G. Walls, George R. Widmeyer, and Omar A. El Sawy. Building an information system design theory for vigilant EIS. *Information Systems Research* 3, 1 (1992), 36–59. https://doi.org/10.1287/isre.3.1.36.

[55] Danding Wang, Qian Yang, Ashraf Abdul, Brian Y. Lim, and United States. 2019. Designing Theory-Driven User-Centric Explainable AI. In *CHI'19*. ACM, Glasgow, Scotland, UK, 1–15.

[56] Huaxiu Yao, Xiaowei Jia, Vipin Kumar, and Zhenhui Li. 2020. Learning with Small Data, 3539–3540. https://doi.org/10.1145/3394486.3406466, arXiv:1910.00201.

[57] I. Zubizarreta, F. Ivaldi, M. Rinas, E. Hogestol, S. Bos, T. Berge, P. Koduah, M. Cellerino, M. Pardini, G. Vila, et al. The Sys4MS project: personalizing health care in multiple sclerosis using systems medicine tools. *Multiple Sclerosis Journal* 24 (2018), 459.

Bionotes

Maria Rauschenberger
Max Planck Institute for Software Systems, Saarbrücken, Germany
University of Applied Science, Emden, Germany
rauschenberger@mpi-sws.org

Maria Rauschenberger is Professor for Digital Media at the University of Applied Science in Emden/Leer. Before she was a Post-Doc at the Max-Planck Institute for Software Systems in Saarbrücken, research associate at the OFFIS – Institute for Information Technology in Oldenburg and Product Owner at MSP Medien Systempartner in Bremen/Oldenburg. Maria did her Ph.D. at *Universitat Pompeu Fabra* in the Department of Information and Communication Technologies under the supervision of Luz Rello and Ricardo Baeza-Yates since early 2016, graduating in 2019 with the highest outcome: Excellent Cum Laude. Her thesis focused on the design of a language-independent content game for early detection of children with dyslexia. She mastered the challenges of user data collection as well as of small data analysis for interaction data using machine learning and shows innovative and solid approaches. Such a tool will help children with dyslexia to overcome their future reading and writing problems by early screening. All this work has been awarded every year (three years in a row) in Germany with a special scholarship (fem:talent) as well as with the prestigious German Reading 2017 award and recently with the 2nd place of the Helene-Lange-Preis.

Ricardo Baeza-Yates
Khoury College of Computer Sciences, Northeastern University, Silicon Valley, CA, USA
rbaeza@acm.org

Ricardo Baeza-Yates is since 2017 the Director of Data Science Programs at Northeastern University, Silicon Valley campus, and part-time professor at University Pompeu Fabra in Barcelona, Spain; as well as at University of Chile. Before, he was VP of Research at Yahoo Labs, based in Barcelona, Spain, and later in Sunnyvale, California, from 2006 to 2016. He is co-author of the best-seller Modern Information Retrieval textbook published by Addison-Wesley in 1999 and 2011 (2nd edition), that won the ASIST 2012 Book of the Year award. From 2002 to 2004 he was elected to the Board of Governors of the IEEE Computer Society and between 2012 and 2016 was elected for the ACM Council. In 2009 he was named ACM Fellow and in 2011 IEEE Fellow, among other awards and distinctions. Finally, in 2018 he obtained the National Spanish Award for Applied Research in Computing. He obtained a Ph.D. in CS from the University of Waterloo, Canada, in 1989, and his areas of expertise are web search and data mining, information retrieval, data science and algorithms in general. He has over 40 thousand citations in Google Scholar with an h-index of over 80.