Recommendations to Handle Health-related Small Imbalanced Data in Machine Learning

Maria Rauschenberger
rauschenberger@mpi-sws.org
Max Planck Institute for Software Systems
Saarbrücken, Germany

Ricardo Baeza-Yates
rbaeza@acm.org
Khoury College of Computer Sciences
Northeastern University at SV, USA

ABSTRACT

When discussing interpretable machine learning results, researchers need to compare results and reflect on reliable results, especially for health-related data. The reason is the negative impact of wrong results on a person, such as in missing early screening of dyslexia or wrong prediction of cancer. We present nine criteria that help avoiding over-fitting and biased interpretation of results when having small imbalanced data related to health. We present a use case of early screening of dyslexia with an imbalanced data set using machine learning classification to explain design decisions and discuss issues for further research.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Cross-validation*; • **Human-centered computing** → **Human computer interaction (HCI)**; • **Social and professional topics** → *People with disabilities.*

KEYWORDS

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

1 INTRODUCTION

Independently of the source of data, we need to understand our machine learning results. 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 characteristics are usually missing such as variaty and velocity [4, 14]. 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 [4, 14]. 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. Therefore, researchers have to make the best of a limited data set and avoid over-fitting, being aware of issues such as *small data, imbalanced data, variance, biases, heterogeneity of participants, or evaluation metrics*. We address the main criteria to avoid over-fitting and taking care of imbalanced data sets related to health from a previous research project with different small data sets related to early and universal screening of dyslexia [28, 29, 35]. Our main contribution is a list of nine criteria when exploring small imbalanced data for machine learning predictions. We also suggest an approach for collecting data from online experiments with interactive systems to control, understand and analyze the data. We do not claim completeness and we see our proposal as a starting point for further recommendations or guidelines.[1]

The rest of the paper is organized as follows: Section 2 covers related work while Section 4 gives the background and explains our approach to collect data from interactive systems with DSRM and HCD. Section 5 describes the general considerations of a research design. In Section 6 we propose our guidelines for small imbalanced data with machine learning and in Section 7 we give an use case. We finish with conclusions and future work in Section 8.

2 RELATED WORK

As in the beginning of machine learning, today small data is used by machine learning models in spite of the focus in big data [14]. But the challenge to avoid over-fitting remains [4] 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 [40], develop new algorithms and metrics [11, 16] or over- and under-sampling [16]. But to the best of our knowledge, a standard approach for the analysis of a small imbalanced data set 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 when analyzing small data with machine learning classification. This is very important as most institutions in the world will never have big data [4].

3 BACKGROUND

Interdisciplinary research projects require a standardized approach, like the *Design Science (DS) Research Methodology (DSRM)* [26], to compare results with different methodologies or mindset. A standardized approach for the design of software products, like the

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

MuC'20 Workshops, Magdeburg, Deutschland

https://doi.org/10.18420/muc2020-ws111-333

Human-Centered Design (HCD) [22], 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 similarity of each method.

3.1 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 [26]. Since the early 1990s, design science is integrated into information systems and provides with DSRM a methodology to justify system design research (quoted after [26]). The core elements of DSRM have their origins in human-centered (British English human-centred) computing and are complementary to the human-centered design framework [21, 22]. 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.* In the first step, researchers describe the problem and the motivation for the technological solution. The level of detail depends on the complexity of the problem. Next, the goals and functionality of the solution are stated, taking into account the information from the previous step and quantifying the solution. In step three, a technological solution is designed and implemented with the proposed architecture or functionality. In the next two steps, the technological solution is presented and evaluated to compare with the goals set at step two. In the last step, researchers "*communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences, such as practicing professionals, when appropriate*" [26].

The information system design theory can be considered to be similar to social science or theory-building [50]. However, designing systems was not and is still not always regarded to be as valuable research as "*solid-state physics or stochastic processes*" [45]. 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 [26] and [19]). 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.

3.2 Human-Centered Design

The *Human-Centered Design* (HCD) framework [22] is a well-known methodology to design interactive systems that takes the whole design process into account and can be used in various areas: enterprise software [27, 31], health related applications [1, 17, 30, 47], remote applications (Internet of things) [41], social awareness [49], or mobile applications [1, 32]). With HCD, designers focus on the user when developing an interactive system to improve *usability* and *user experience*.

The two main terms to describe and quantify the methods for HCD are *usability* and *user experience* (UX). How well a user interacts for a certain goal or task in a specific context is called *usability*

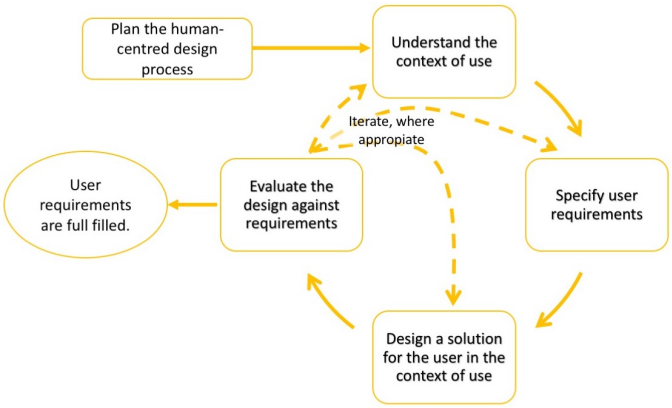

Figure 1: Activities of the human-centered design process adapted from [23].

[23]. This means a certain type of user (for example, a student) wants to do a specific task (for example, writing an email to her/his professor on her/his computer from home). The level of detail for the task description can depend on the design resources (*i.e.,* time or personnel) or design goal (*i.e.,* proof of concept or product). The main focus is on the user achieving the task effectively, efficiently, and achieve satisfaction. *User experience* incorporates usability and advances the concept of interaction through the perception and responses of the user as well as the "*emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors and accomplishments that occur before, during and after use*" [22].

The HCD is an iterative design process (see Figure 1). 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 [5]. 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 [5]. 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 [2].

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.

There are various methods and artifacts which can be included in the design approach depending, (*e.g.*, on the resources, goals,

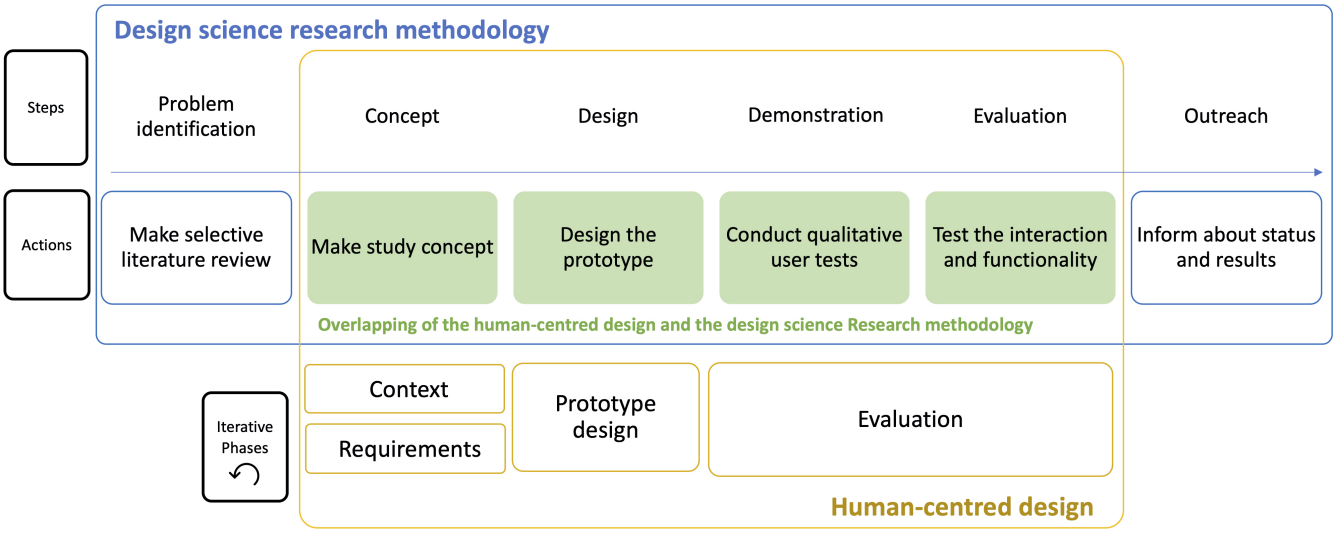

Figure 2: Integration of the *Human-centered Design* in the *Design Science Research Methodology*.

context of use, or users) to observe, measure and explore users' behavior. Evaluation methods are, for example, the five-user study [25], the User Experience Questionnaire (UEQ) [20, 33, 36], observations, interviews, or the think-aloud protocol [9]. 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 [10]. With small testing groups ($n < 10-15$) [10], 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 [37].

4 COLLECTING AND ANALYZING DATA FROM INTERACTIVE SYSTEMS

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)* [26] with the *Human-Centered Design (HCD)* [22]. 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, 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 3.

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 underlining 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 [28, 29, 35].

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* [22]. The *human-centered design* complements the *design science research methodology* with a focus on the participants and provides various guidelines, methods, and artifacts for the design of a prototype.

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 iteratively designed with the knowledge of experts. In this case, the interactive system has an integration of *game elements* to apply the concept of *gamification* [38]. 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 we used, a double click on a web application generally *zooms* in, which was not intended 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 [28]. 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 [34] in the design of the application and the experiment as well as consider the ethical aspects [6]. As dyslexia is connected to nine genetic markers and reading ability is highly hereditary [12], we support readability for participants' supervisors (who could be parents) with a large font size (minimum 18 points) [39].

5 GENERAL CONSIDERATIONS OF RESEARCH DESIGN

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 [15]. 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 avoided by counterbalancing the order of the conditions, for example with Latin Squares [15].

The advantage of a *repeated-measures design* in a *within-subject design* is that participants can engage in multiple conditions [15]. 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* [15].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 [15], while features are used as input for the classifiers to recognize patterns [7]. Machine learning is a data-driven approach in which the data is explored with different algorithms to minimize the objective function [13]. In the following we refer to the implementation of the Scikit-learn library (version 0.21.2)

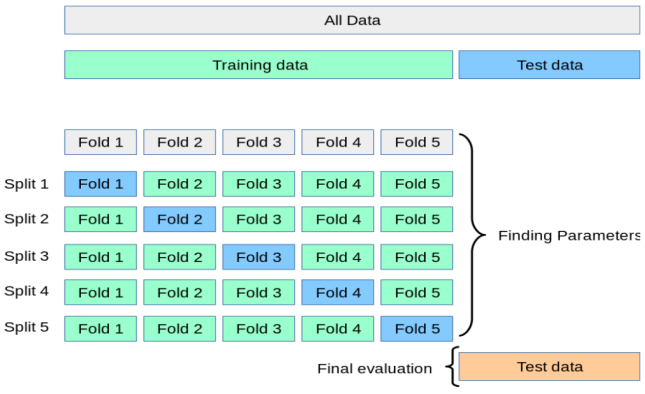

Figure 3: Approach of cross-validation from [42].

if not stated otherwise [44]. 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* (also written as *overfitting*), 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 overfitting.*" [13]

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 [42]. 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 [42], *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 [24], 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 [42]. 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 [42]. However, small data sets with high variances are not discussed.

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 [43]. 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*" [43]. Whereas the parameter *recall* "*is the ability of the classifier to find all the positive samples*" [43]. 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*" [43]. Scikit-learn library suggests different implementations for computing the metrics (*e.g.,*

recall, F1-score) and the confusion matrix [42]. The reason is that the *metric function* reports over all (cross-validation) fold, whereas the *confusion matrix function* returns the probabilities from different models.

6 PROPOSED RECOMMENDATIONS FOR HEALTH-RELATED SMALL DATA SETS

Based in our previous research [28, 29, 35] 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"[3].

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.

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 [24] 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 [7].

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 [42] and to avoid biases [48].

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

This nine criteria is the starting point of machine learning solutions on health-related small data analysis as this can have a significant impact on specific individuals.

7 USE CASE: 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 [28, 29]. 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 [46] such as the DRT [18] 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 [8]. 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 [29]. 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.

8 CONCLUSION AND FUTURE DIRECTIONS

We propose the first step towards guidelines when exploring health-related small imbalanced data sets with nine criteria. We show a use case and reveal opportunities to discuss and develop this further with, *e.g.*, new machine learning classification for imbalanced data considering small data.

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.

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] 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

[2] 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-prototyping-for/9780120885688/

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

[4] 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].

[5] Protima Banerjee. 2004. *About Face 2.0: The Essentials of Interaction Design: Alan Cooper and Robert Reimann Published by John Wiley & Sons, 2003, 576 pp, ISBN 0764526413.* Vol. 3. Wiley Publishing, Inc., USA. 223–225 pages. https://doi.org/10.1057/palgrave.ivs.9500066

[6] 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

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

[8] 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

[9] 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/1003981860http://www.amazon.com/Methoden-Usability-Evaluation-Henning-Brau/dp/3456842007

[10] 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

[11] André M. Carrington, Paul W. Fieguth, Hammad Qazi, Andreas Holzinger, Helen H. Chen, Franz Mayr, and Douglas G. Manuel. 2020. 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

[12] 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-of-neurolinguistics/oclc/1043957419{&}referer=brief{_}results

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

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

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

[16] Koichi Fujiwara, Yukun Huang, Kentaro Hori, Kenichi Nishioji, Masao Kobayashi, Mai Kamaguchi, and Manabu Kano. 2020. 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

[17] 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

[18] 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

[19] Alan Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. Design Science in Information Systems Research. *MIS Quarterly* 28, 1 (2004), 75. https://doi.org/10.2307/25148625

[20] 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.

[21] Robert R. Huffman, Axel Roesler, and Brian M. Moon. 2004. 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

[22] 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

[23] 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.htmlhttps://www.iso.org/obp/ui/{#}iso:std:iso:9241:-11:ed-2:v1:en

[24] Anil Jain and Douglas Zongker. 1997. 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

[25] Jakob Nielsen. 2000. Why You Only Need to Test with 5 Users. http://www.useit.com/alertbox/20000319.html. *Jakob Nielsens Alertbox* 19, September 23 (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].

[26] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. 2007. 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

[27] Maria Rauschenberger. 2015. Entwicklung von Designentwürfen zur Unterstützung von Hafenmanövern für Lotsen mittels der Hierarchischen Aufgabenanalyse. (Deriving designs in harbour manoeuvre for harbour pilots with the hierarchical task analysis.). https://doi.org/10.18420/muc2016-up-0131

[28] Maria Rauschenberger. 2021. *Early screening of dyslexia using a language-independent content game and machine learning*. Ph.D. Dissertation. Universitat Pompeu Fabra. https://doi.org/10.13140/RG.2.2.27740.95363

[29] 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

[30] 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.

[31] Maria Rauschenberger, Andreas Hinderks, and Jörg Thomaschewski. 2011. Benutzererlebnis bei Unternehmenssoftware: Ein Praxisbericht über die Umsetzung attraktiver Unternehmenssoftware. (Enterprise Software User Experience: A real-world report on how enterprise software can be made attractive.). In *Usability Professionals Konferenz 2011*, Vol. 1. German UPA e.V., Stuttgart, UPA, Stuttgart, 154–158.

[32] 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.

[33] 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.

[34] Maria Rauschenberger, Luz Rello, and Ricardo Baeza-Yates. 2019. Technologies for Dyslexia. In *Web Accessibility Book* (2 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

[35] 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

[36] Maria Rauschenberger, Martin Schrepp, Manuel Perez Cota, Siegfried Olschner, and Jörg Thomaschewski. 2013. 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

[37] 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 *In Usability Professionals Konferenz 2013*. German UPA eV, Bremen, 72–76.

[38] Maria Rauschenberger, Andreas Willems, Menno Ternieden, and Jörg Thomaschewski. 2019. 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/

[39] 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

[40] 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

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

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

[43] 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].

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

[45] 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

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

[47] Lieven Van den Audenaeren, Véronique Celis, Vero Vanden 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

[48] Sudhir Varma and Richard Simon. 2006. 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

[49] 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

[50] Joseph G. Walls, George R. Widmeyer, and Omar A. El Sawy. 1992. 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