Serious games for a technology-enhanced early screening of handwriting difficulties

Linda Greta Dui
Dept. Electronics, Information and Bioengineering
Politecnico di Milano
Milan, Italy
lindagreta.dui@polimi.it

Chiara Piazzalunga
Dept. Electronics, Information and Bioengineering
Politecnico di Milano
Milan, Italy
chiara.piazzalunga@polimi.it

Simone Toffoli
Dept. Electronics, Information and Bioengineering
Politecnico di Milano
Milan, Italy
simone.toffoli@polimi.it

Stefania Fontolan
Dept. Medicine and Surgery
University of Insubria
Varese, Italy
stefania.fontolan@gmail.com

Sandro Franceschini
Dept. Medicine and Surgery
University of Insubria
Varese, Italy
sandro.franceschini @gmail.com

Marisa Bortolozzo
Dept. Medicine and Surgery
University of Insubria
Varese, Italy
marisa.bortolozzo@googlemail.com

Nunzio Alberto Borghese
Dept. Computer Science
University of Milan
Milan, Italy
alberto.borghese@unimi.it

Cristiano Termine
Dept. Medicine and Surgery
University of Insubria
Varese, Italy
cristiano.termine@uninsubria.it

Simona Ferrante
Dept. Electronics, Information and Bioengineering
Politecnico di Milano
Milan, Italy
simona.ferrante@polimi.it

***Abstract*—Early screening of handwriting difficulties is key to start remediation activities that help distinguishing between a simple delay and dysgraphia. Technology is fundamental in this process, as also claimed by guidelines for dysgraphia diagnosis: it allows to implement artificial intelligence techniques to help in the discrimination of the difficulty. To this end, a serious game was leveraged to assess handwriting laws altered in dysgraphia starting from symbols drawing. 66 first and second graders were longitudinally tested both with the serious game and with a handwriting proficiency test. Objective features computed from the game were tested to understand if they significantly differed between children at risk and not at risk of dysgraphia, according to a standardized clinical test used to assess handwriting. Then, machine learning models were leveraged to predict the risk and understand the areas of difficulty. On average, 62% of the features significantly differ between risk levels for first graders, whilst only 35% for second graders, thus revealing a better sensitivity in younger children. This is encouraging for an early observation. As for machine learning, a Logistic classifier was able to predict risk with an area under the precision-recall curve of 0.84 for the risk class and 0.98 for the non-risk class. The results of this study could be a valid help for an artificial intelligence-enhanced screening of dysgraphia.**

Keywords—Dysgraphia, Early Screening, Serious Games.

I. INTRODUCTION

Starting the learning path can be challenging, particularly for children who struggle to acquire basic knowledge. If these difficulties are not addressed, they can lead to frustration, low self-esteem, depression, and school abandonment [1]. Some children simply need more time; others may never reach their peers' level, nonetheless any remediation attempt, thus being affected by specific learning disabilities (SLDs) [2]. Discriminating between a delay and a disability requires the diagnosis from a child neuropsychiatrist (CNP), that is possible only when the corresponding skill is acquired, i.e., the end of second grade. School closures that occurred between 2020 and 2022 due to COVID-19 pandemic caused an additional delay in learning, and a further postponement of the diagnosis has been proposed [3].

While waiting for a diagnosis, it is school's responsibility to help these children. Indeed, it has been recognized that early training could be beneficial for improving delays [4], whilst it is not effective with children with SLDs. Hence, responsiveness to structured training could help reducing the number of children addressed to CNP, thus avoiding overwhelming the healthcare system. To this end, school-level preventive initiatives have been proposed: at the beginning of the school year, teachers are supposed to observe signs of potential delay; during the year, they propose activities to strengthen weak abilities; at the end of the year, a follow-up evaluation assesses improvements. If a very severe situation is detected or if the delay is persistent after the training, a visit with CNP is envisaged with priority.

The main drawbacks of the process are that schools lack specialized figures to understand if a child deserves more attention, and that the evaluation is time-consuming and subjective. Recently, novel Italian guidelines on SLDs diagnosis acknowledged that the evaluation of a difficulty would be more effective if paired with digital tools capable of offering an objective description of the problem [3], such as, tablets and smart ink pens [5]. International scientific research is moving to fill this gap (e.g., [6,7]), but there is still the need of objective tools that could be used by teachers without the help of an expert.

To this extent, the Play-Draw-Write tablet app has been devised [8]. It is a collection of serious games that investigate characteristics of handwriting production known to be altered in dysgraphia, but leveraging symbols drawing, as they do not require that handwriting is mastered. Specifically, game design was based on exercises on isochrony and homothety (i.e., adjusting speed to size to keep execution time approximately constant), and speed-accuracy trade off (SAT, i.e., adjusting movement time to task difficulty). These laws are investigated by asking children to copy a square and a sequence of a circle, a vertical line and a reversed U at different sizes (spontaneous, big and small) and to steer tunnels shaped as squares and as letters ("ELE" in cursive) of

Research supported by H2020 grant N.101016112 ESSENCE: Empathic platform to personally monitor, Stimulate, enrich, and aSsist Elders aNd Children in their Environment.

different length and width (14 levels, 5 combinations of indices of difficulty [8]). The app allows also to collect a rich set of objective features that characterize gesture production, such as, kinematics, fluidity, pen angles, tip pressure, spatial organization, and execution errors [9]. Such objective parameters are altered in children affected by dysgraphia [6,10]. The predictive capabilities were tested on 241 preschoolers, 73 of them with a learning delay. A convolutional neural network reached 0.75 sensitivity and 0.76 specificity when playing on an iPad with Apple pencil [9]. Given the promising results on kindergartners, the study proceeded by longitudinally assessing these children up to second grade.

However, the app was still a prototype that needed to be tested for quasi-unsupervised administration, as it would be at school with teachers. Additionally, it was necessary to test for the effectiveness and generalizability of artificial intelligence techniques leveraged to predict the risk of dysgraphia. This scenario has been reproduced in the context of the H2020 project ESSENCE (Empathic platform to personally monitor, Stimulate, enrich, and aSsist Elders aNd Children in their Environment https://cordis.europa.eu/project/id/101016112/it). In the project, the whole process of observation, training and even clinical consultation have been translated into a digital one, and the Play-Draw-Write app replaced teachers' observation of handwriting delay.

In this context, the aim of the study was (1) to understand if the preliminary results achieved with the Play-Draw-Write app could be generalized to (a) older ages (first and second grade), (b) different technology (tablet and pen), and (c) quasi-unsupervised administration (without an experimenter), and (2) to provide artificial intelligence-based classification of children at risk of an handwriting delay.

II. Methods

A. Game refinement for self-assessment

The first version of Play-Draw-Write [8] was designed to be administered by an experimenter that explained the exercises, checked that they were performed correctly, and downloaded locally saved files. It was deployed on iPads with Apple pencil, that allowed sample frequency up to 240 Hz. To propose the game at school, it was necessary to implement different refinements, with an iterative co-design that involved children who provided feedback on the modifications.

As for games explanation, (1) a guide character replaced the experimenter, both to illustrate the rules and to provide feedback, (2) the tunnel tutorial could be watched whenever needed, without losing the progress. The most common errors in tunnels steering have been automatically recognized: (1) lifting the pen before the end, (2) steering in the wrong direction, (3) spending too much time outside the borders, or (4) drawing with the finger. Navigation between scenes was improved. Data transfer to a server was implemented via RESTful APIs. In case a Wi-Fi connection was not detected at the startup, the guide character asks for switching it on. At the end of each level, if data cannot be sent to the server, they are saved on the device and an uploading attempt is performed again at the following startup. Finally, the games have been deployed on Samsung S6 lite with S-pen, that are cheaper devices, more suitable for school adoption. The same data collected in [9] were stored on the server (i.e., pen position, pressure and tilt), but at 60 Hz.

Game development was performed in Unity 2018.3.2f1.

B. Participants and protocol

To reach the objective of the work, during the ESSENCE field testing, 21 first graders and 45 second graders were recruited and observed at three time points, in January 2022 (TP1), in May 2022 (TP2) and in October 2022 (TP3), as reported in Fig. 1A. For each time point, children executed a handwriting test and played with the Play-Draw-Write app.

The handwriting ability was assessed using standardized clinical tests selected from the Battery for the evaluation of handwriting and orthographic competence, second edition (BVSCO-2) [11]. They were administered and graded by CNPs. First, children were asked to repeat a sequence of "LE" in cursive for one minute; second, they were asked to write numbers as words for one minute. Both tests were provided with normative data, to compute a z-score and allow a comparison between grades and time points.

As for the Play-Draw-Write app, children played the games described in [8], that are Copy Game - Square (CGSq) and Copy Game - Sequence (CGSe) in three modalities: spontaneous, big and small; Tunnel Game - Square (TGS) and Tunnel Game - Word (TGW), see a tutorial, make a trial and steer tunnels at 14 different combinations of amplitude and width. The games were administered in a quiet room at school, to allow children hearing and following the vocal instructions under the supervision of a teacher.

The protocol was approved by Politecnico di Milano Ethical Committee nr. 04/2021. Data collected during this study are openly available in Zenodo at http://doi.org/10.5281/zenodo.8208391 [12].

C. Data analysis

The risk of dysgraphia was computed at each time point from the BVSCO-2 results considering z-scores below the -2 threshold [11]. TP3 only was considered to define the risk label, as it is supposed to be more reliable and closer to the time when the diagnosis can be performed.

From the Play-Draw-Write app, a total of 369 features were computed for each time point from the four games, as in [9]. For each time point, collinear features (Spearman's correlation greater than 0.9) were discarded. An exploratory analysis was conducted to investigate if the features at the tree time points were sensitive to the risk of dysgraphia at TP3, to

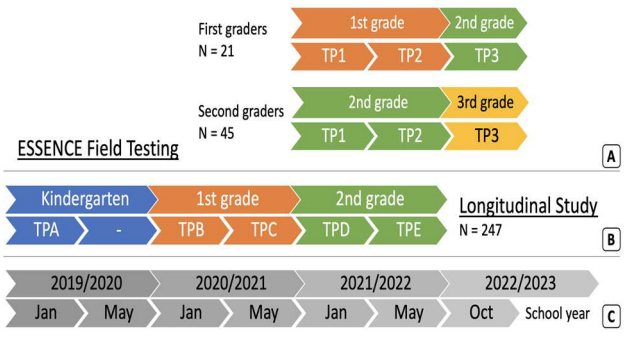

Figure 1. Panel A: ESSENCE field testing structure. Panel B: Longitudinal study structure. Panel C: chronological arrangement of time points. Colors mark the correspondence between time points in terms of school grade: as for ESSENCE first graders, TP1-3 correspond to TPB-D; as for ESSENCE second graders, TP1-2 correpond to TPD-E.

understand if an anticipation of the screening was possible. Depending on feature normality (Lilliefors test) features were leveraged as dependent variables for a Mann-Whitney test considering the risk as independent variable for each time point, or an ANOVA for mixed effects considering time as within-subject factor and risk as between-subject factor. Statistical significance was set to 5%. The procedure was repeated separately for first and second grades.

To help in the observation of a learning delay, the risk computed from the BVSCO-2 test should be automatically predicted. If trying to apply a machine learning model, data coming from the ESSENCE field testing alone would suffer from the "curse of dimensionality", as 66 subjects were enrolled in respect to 369 features. Moreover, considering the fast development during early childhood that affects also drawing features [13], it could be necessary to build separate models for first and second graders, thus further worsening the dimensionality problem. To this end, it could be useful to leverage the wide but slightly different dataset coming from the previous Longitudinal data collection to (1) select a subsample of features and (2) allow training on more samples. Only ESSENCE TP1 and TP2 could be fully compared to Longitudinal data, as shown in Fig. 1. Then, it was decided to leverage TP1 and TP2 games execution to predict the risk in TP3, in a preventive perspective. Fig. 2 schematizes the procedure. Given that the acquisition system was different, it was not possible to train models on longitudinal data and directly test them on ESSENCE data without complex calibration (e.g., pressure encoding differed). We opted for an approach that is close to what is performed in clinical practice to assess risk: we leveraged features from the Longitudinal study as "normative" data, separately for each grade and time point, and standardized the ESSENCE data to achieve z-scores that are grade-independent. In this way, we were able to achieve a dataset that comprised both first and second graders, without introducing the age as a confounding effect. As for training, leave-one-out was chosen to train on as much children as possible, and to test the performance on each of them. This is a suitable technique when it is necessary to assess single individuals, as in this application. Performance were then compared to the dummy classifier (Baseline), i.e., assigning all the predictions to the most represented (no-risk) class. Finally, prediction errors of the best model were analyzed to understand potential improvements.

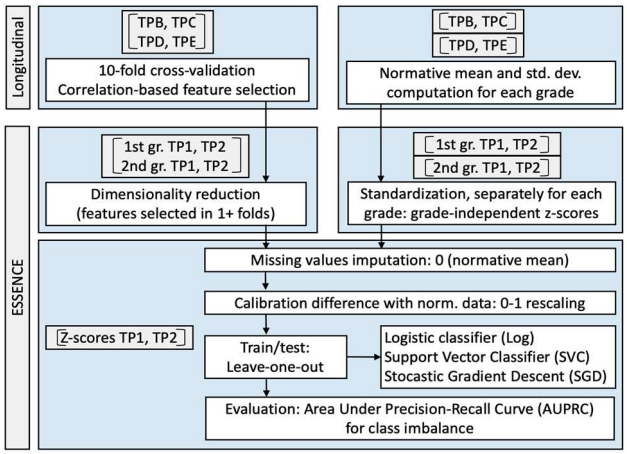

Figure 2. ESSENCE data preparation leveraging Longitudinal data for preprocessing, and models building. For each step, data arrangement is shown in brackets, referring to Fig. 1 for time point definition.

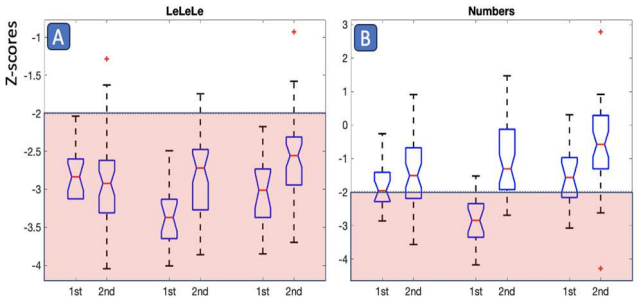

Figure 3. BVSCO-2 tests to assess handwriting abilities. Red line: median; notch: 95% of the median; boxes: quartiles; red crosses: outliers.

Feature extraction and statistics were performed in Matlab R2020a, machine learning models were built in Weka 3.8.6.

III. Results

18 first graders (11 males, 7 females) and 42 second graders (19 males, 23 females) completed the evaluation.

The results achieved in the BVSCO-2 test are reported in Fig. 3, divided by test (Panels A and B) and grade (1st and 2nd). The red-shadowed area highlights the risk zone, i.e., z-score < -2. As most children were considered at risk at TP3 in the "LE" test, "Number" test only was considered for the analysis. In total, 12 children were considered at risk (R), that are, 43% of first graders and 20% of second graders.

The percentage of features that significantly differed between risk levels are reported in Table I. As for machine learning models, 19 features were selected from TP2 and 14 features from TP1. The most important positions were shared between CGSe and TGW. Table II reports the results on the test set. As Log was the best performing model (bold in the Table), the analysis of prediction errors proceeded on this model only. None of the first graders was mistakenly assigned to the wrong class. Two second graders were false positives (FPs), four were false negatives (FNs). As for FPs, children were not at risk on the Number test of the BVSCO-2, but they were for the LE exercise (z-score -3.21 and -2.88). One of the FNs was only slightly under the risk threshold (-2.07). Another FN was not assessed at TP1, thus having all the features replaced by zeros, i.e., the normative mean, thus being similar to a "normal" subject.

IV. Discussion

In this work, a technology-based solution to enhance the early observation of the risk of dysgraphia has been presented. A serious game that investigates potential alterations in graphical gesture production was leveraged [8]. It was administered to 60 children from first and second grade, together with a handwriting proficiency test, at 3 time points.

TABLE I. Percentage of features that significantly differed according to risk for each game, grade and time point

Game	TP1		TP2		TP3	
	1st	*2nd*	*1st*	*2nd*	*1st*	*2nd*
Copy Square	64%	33%	59%	34%	59%	34%
Copy Sequence	61%	32%	62%	38%	61%	38%
Tunnel Square	66%	27%	68%	27%	66%	27%
Tunnel Word	47%	39%	65%	45%	66%	42%

TABLE II. MACHINE LEARNING RESULTS ON THE TEST SET

Models	AUPRC R	AUPRC NR	AUPRC weighted
Log	**0.84**	**0.98**	**0.95**
SVC	0.61	0.93	0.87
SGD	0.60	0.92	0.85
Baseline	0.20	0.80	0.68

The first step was to assess children on their handwriting abilities. A standardized test, the BVSCO-2, was leveraged. As also stated in novel guidelines for dysgraphia diagnosis [3], handwriting proficiency is low in children who experienced COVID-19 school closures, as those recruited in this work. Indeed, most children showed very bad performance, in comparison with the peers the normative BVSCO-2 data were computed on. This is particularly true for first graders, that could not train with pre-graphical activities during kindergarten years. However, the situation improved in time, till reaching a percentage of children at risk comparable with literature [1].

Statistical testing on gesture-related features demonstrated that most of them significantly differed between children at risk and not at risk of dysgraphia for first graders. Instead, the same features computed on second graders reached significance less times. Indeed, serious games were designed for early screening, when pre-graphical activities are preponderant, and the games could be more challenging. The discriminative power since a precocious time point is a strength of the games. However, as children skills evolve, exercises complexity should increase to keep being a valid predictor of risk.

The results from the machine learning models suggest the same concept, as prediction errors regarded second graders only. Anyhow, the results of the model are really good, given the reduced sample size, as a Logistic classifier allowed to discriminate children at risk with an AUPRC of 0.84 and children not at risk with an AUPRC of 0.98. The results are even more impressive if considering that most prediction errors could be justified: the model captures problems that are not completely described by a single handwriting test.

These results were possible thanks to the availability of similar data from a previous Longitudinal study. Even if a direct application of the models built on such data was not possible, they were leveraged for a grade-specific standardization to widen the sample and improve the models. The proposed approach recalls different domain adaptation techniques reported in literature, but with the advantage of being similar to an approach commonly used in clinical practice, thus preserving explainability. Indeed, other authors suggest to transform the features, but this would mine the interpretation [14], whilst other techniques [15] are scarcely beneficial when source and target domains are too similar [16].

To conclude, Play-Draw-Write can be a valid aid to help in a technology-enhanced screening for dysgraphia from an early stage, even without experts' supervision. Thanks to its usage in the ESSENCE project, an AI-powered screening of handwriting delays could be possible.

ACKNOWLEDGMENT

We would like to thank "Ministero dell'Istruzione, Ufficio Scolastico Regionale per la Lombardia Ufficio XIV – Varese; children, teachers, parents and school dean from I.C.S Gavirate; and the ESSENCE Consortium.

REFERENCES

[1] Chung PJ, Patel DR, Nizami I. Disorder of written expression and dysgraphia: definition, diagnosis, and management. Transl Pediatr 2020;9. https://doi.org/10.21037/tp.2019.11.01.

[2] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). 2013.

[3] Salute M della. Linea Guida sulla gestione dei Disturbi Specifici dell'Apprendimento 2021.

[4] Kadar M, Wan Yunus F, Tan E, Chai SC, Razaob@Razab NA, Mohamat Kasim DH. A systematic review of occupational therapy intervention for handwriting skills in 4–6 year old children. Aust Occup Ther J 2020;67. https://doi.org/10.1111/1440-1630.12626.

[5] Dui LG, Calogero E, Malavolti M, Termine C, Matteucci M, Ferrante S. Digital Tools for Handwriting Proficiency Evaluation in Children, 2021. https://doi.org/10.1109/bhi50953.2021.9508539.

[6] Mekyska J, Faundez-Zanuy M, Mzourek Z, Galaz Z, Smekal Z, Rosenblum S. Identification and Rating of Developmental Dysgraphia by Handwriting Analysis. IEEE Trans Human-Machine Syst 2017;47:235–48. https://doi.org/10.1109/THMS.2016.2586605.

[7] Asselborn T, Gargot T, Kidziński Ł, Johal W, Cohen D, Jolly C, et al. Automated human-level diagnosis of dysgraphia using a consumer tablet. Npj Digit Med 2018;1. https://doi.org/10.1038/s41746-018-0049-x.

[8] Dui LG, Lunardini F, Termine C, Matteucci M, Stucchi NA, Borghese NA, et al. A tablet app for handwriting skill screening at the preliteracy stage: Instrument validation study. JMIR Serious Games 2020;8. https://doi.org/10.2196/20126.

[9] Dui LG, Lomurno E, Lunardini F, Termine C, Campi A, Matteucci M, et al. Identification and characterization of learning weakness from drawing analysis at the pre-literacy stage. Sci Rep 2022;12:21624.

[10] Pagliarini E, Guasti MT, Toneatto C, Granocchio E, Riva F, Sarti D, et al. Dyslexic children fail to comply with the rhythmic constraints of handwriting. Hum Mov Sci 2015;42:161–82. https://doi.org/10.1016/j.humov.2015.04.012.

[11] Cornoldi C, Re AM, Tressoldi PE. Battery for the Clinical Evaluation of Handwriting and Orthographic Competence. Batteria per la Valutazione Clinica della Scrittura e delle Competenze Ortografiche nella Scuola dell'Obbligo (BVSCO-2). 2nd ed. Giunti Psychometrics; 2013.

[12] Ferrante S. ESSENCE Field Testing - DATASET 2023. https://doi.org/10.5281/zenodo.8208391.

[13] Dui LG, Toffoli S, Speziale C, Termine C, Matteucci M, Ferrante S. Can Free Drawing Anticipate Handwriting Difficulties? A Longitudinal Study. BHI-BSN 2022 - IEEE-EMBS Int. Conf. Biomed. Heal. Informatics IEEE-EMBS Int. Conf. Wearable Implant. Body Sens. Networks, Symp. Proc., 2022. https://doi.org/10.1109/BHI56158.2022.9926884.

[14] Sun B, Feng J, Saenko K. Return of frustratingly easy domain adaptation. 30th AAAI Conf. Artif. Intell. AAAI 2016, 2016. https://doi.org/10.1609/aaai.v30i1.10306.

[15] Daumé H. Frustratingly easy domain adaptation. ACL 2007 - Proc. 45th Annu. Meet. Assoc. Comput. Linguist., 2007.

[16] Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big Data 2016;3. https://doi.org/10.1186/s40537-016-0043-6.