Towards a robust and accurate screening tool for dyslexia with data augmentation using GANs

Thomais Asvestopoulou*†, Victoria Manousaki*†, Antonis Psistakis*†, Erjona Nikolli*†,
Vassilios Andreadakis ‡, Ioannis M. Aslanides§, Yannis Pantazis¶, Ioannis Smyrnakis†‡ and Maria Papadopouli*†‖
*Department of Computer Science, University of Crete, Heraklion, Greece
†Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece
‡Optotech Ltd., Heraklion, Greece
§Emmetropia Eye Institute, Heraklion, Greece
¶Institute of Applied and Computational Mathematics, Foundation for Research and Technology-Hellas, Heraklion, Greece

***Abstract*—Eye movements during text reading can provide insights about reading disorders. We developed the DysLexML, a screening tool for developmental dyslexia, based on various ML algorithms that analyze gaze points recorded via eye-tracking during silent reading of children. We comparatively evaluated its performance using measurements collected from *two systematic field studies* with 221 participants in total. This work presents DysLexML and its performance. It identifies the features with prominent predictive power and performs dimensionality reduction. Specifically, it achieves its best performance using linear SVM, with an accuracy of 97% and 84% respectively, using a small feature set. We show that DysLexML is also robust in the presence of noise. These encouraging results set the basis for developing screening tools in less controlled, larger-scale environments, with inexpensive eye-trackers, potentially reaching a larger population for early intervention. Unlike other related studies, DysLexML achieves the aforementioned performance by employing only a small number of selected features, that have been identified with prominent predictive power. Finally, we developed a new data augmentation/substitution technique based on GANs for generating synthetic data similar to the original distributions.**

I. Introduction

Dyslexics manifest significant and persistent reading difficulties [1], which often involve difficulty in reading due to word decoding (relating sounds with written phrases, i.e., graphemes to phonemes) [2]. Early intervention can be effective in alleviating the symptoms of the disability. However, screening large populations of children is rather time-consuming and expensive [3]. For example, a differential diagnosis of dyslexia can take up to 14 months [4]. It has been known that the eye movements during text reading can be particularly revealing [5]–[9]. For example, dyslexics exhibit more aberrant eye movements than normal readers at the same age level [1], although it is unlikely that the primary cause of dyslexia is erratic eye movements. *Fixations*, i.e., maintaining the visual gaze on a single location, and *saccadic movements*, i.e., quick simultaneous movements of the eyes between fixations, are important characteristics for screening dyslexia. Readers with developmental dyslexia generate different eye movements than typical readers during text reading: longer and more frequent fixations, shorter saccade lengths, more backward refixations than typical readers [6]–[8], [10]. Furthermore, readers with dyslexia have difficulty in reading long words, lower skipping rate of short words, and high gaze duration (total fixation duration) on many words [4]. Nonetheless, it is still an open question whether it is possible to build a screening tool that can reliably identify readers who may be of high risk for dyslexia by analyzing these distinctive oculomotor patterns collected during reading and can be robust under noise.

This work develops DysLexML, a screening tool for dyslexia, that employs various ML classifiers, such as SVM, Naïve Bayes, and comparatively evaluates their performance using data obtained from two systematic field studies. The first study (RADAR [4]) involved 69 native Greek speakers, children, 32 of which were diagnosed as dyslexic. The second field study involved 152 participants, 72 of which were diagnosed as dyslexic. The diagnosis was performed by the official governmental agency for diagnosing learning and reading difficulties in Greece.

To examine the robustness of DysLexML, we assessed its accuracy under various fixation position noise levels, introduced by the eye-tracking technology or the small screen size (e.g., the small size of the text when a mobile device is used). DysLexML can achieve high accuracy and is robust in the presence of noise. It performs dimensionality reduction, achieving the aforementioned performance using *only a small number of features*, namely the mean and median saccade length, the number of short forward movements, and the number of multiply fixated words for the first dataset, and the number of fixations, the median fixation duration, the median length of medium forward movements, the number of short forward movements, the number of multiply fixated

This work has been partially funded from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT) under grant agreement No. 2285, the Erasmus+ International Mobility between University of Crete and Harvard Medical School 2017-1-EL01-KA107-035639, the Marie Curie RISE NHQWAVE project under grant agreement No. 4500, and the Human Resources Development, Education and Life Lifelong Learning for the implementation of the European Social Fund and the Youth Employment Initiative.

‖Contact author: Maria Papadopouli (mgp@ics.forth.gr)

words, and age for the second dataset. Finally, to address the lack of large-size datasets and privacy requirements, we developed a methodology based on generative models for generating realistic synthetic datasets. Such synthetic data can be used either as additional samples to assist the classifiers to perform better or to completely replace the original data when sensitive information cannot be shared with others (privately or publicly).

The innovative contributions of this work include (i) the analysis of the robustness in the presence of noise, (ii) the identification of features with large discriminating power, (iii) the high accuracy using only a small set of features, (iv) the comparative evaluation with other screening tools/algorithms, and (v) a data augmentation/substitution technique based on generative adversarial networks (GANs), which is, to the best of our knowledge, one of the first published approaches using small datasets that achieves training convergence. This is the first attempt of data augmentation in the context of dyslexia.

The paper overviews briefly the small field study in Section II. Section III presents the DysLexML and Section IV evaluates its performance for both datasets. The system's sensitivity to noise is examined in Section V. Section VI describes the procedure of creating new synthetic datasets with the use of generative models. Section VII gives insights in other work based on dyslexia analysis and prediction, while Section VIII summarizes our key findings and future work plans.

II. First field study (small population)

The first field study was performed in Greece and included 69 children, 32 of which were diagnosed as dyslexic by the official governmental agency for diagnosing learning and reading difficulties in Greece. Participants age span is between 8.5 and 12.5 years old [4]. The children were instructed to read two passages, at their own pace. Both texts were written by a special education teacher in Greek. The first passage (*baseline* text) consists of 181 words, many of which multi-syllable. A second passage, simpler than the first one, targeting to younger participants, was also given to the subjects. It included 143 words, mostly of one or two syllables. It was also emphasized that the purpose was to understand the text in order to answer five comprehension questions at the end. The experimental procedure consisted of recording the eye movements of the participants, while they were silently reading the texts in front of a computer monitor.

A custom-made eye-tracker, developed by Medotics AG was employed [4]. It consists of two steady cameras that can record images up to 60Hz with a resolution of 1600×1200 pixels. Cameras are positioned between the screen and the participant with a viewing field from down towards the participant's face. While the participant performs a reading task, the cameras record the participant's face. The images extracted are then used to detect pupil and corneal reflection coordinates. Based on the collected raw gazing measurements, the fixations were identified according to a dispersion algorithm [11]. More information about the field study, the inclusion and exclusion criteria, texts, and data collection, can be found in [4]. The dataset includes for each fixation, its x- and y-axis coordinates, its starting and termination time, as well as the Region of Interest (ROI) (i.e., word) the subject is looking at.

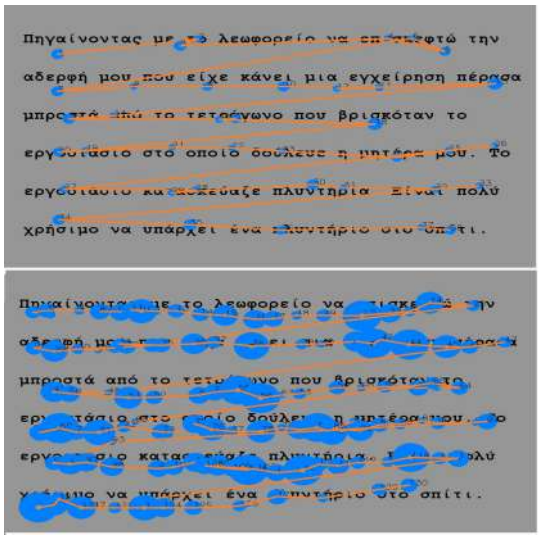

Fig. 1. Reading "path" from a typical reader (top) and from a reader with dyslexia (bottom). The blue circles are the fixations and the orange lines the saccadic movements. The larger the circle, the longer the fixation (Figure appeared in [4]).

III. DysLexML system

The three main modules of the DysLexML algorithm are (i) the feature extraction, (ii) the identification of dominant features (feature selection), and (iii) its classifiers, that employ these dominant features. DysLexML extracts *general (non-word-specific)* features and *word-specific* ones that take into account the word the subject is looking at. Examples of non-word specific features are the number of fixations on the screen, mean and median duration of fixations, and features related to saccades, such as the mean and median length of saccades, i.e., the Euclidean distance between consecutive fixations, and characterization of the types of eye movements. DysLexML creates a feature vector of 35 features in total.

People with reading difficulties tend to perform back and forth movements (saccades) on the text line as they proceed as a result of difficulty to focus or understand [6]. Thus, the identification of such movements and definition of features based on them can provide valuable information about the dyslexic population. Typical readers tend to perform medium to large movements (saccades) in terms of length, while readers with reading difficulties "generate" many choppy movements [10]. A movement is labeled as *short*, if the Euclidean distance between its consecutive fixations is less than 100 pixels (about 5 letters in the text). Most of the movements occur within words. Given that the line of text was about 900 pixels long, the threshold for *medium to long movements* was set to be 400 pixels (about half a line). With this threshold a *medium backward movement* includes re-reads of small groups of words but not of entire phrases. That is, short movements are

of less than 100 pixels, long ones are of more than 400 pixels, and medium movements in the range between 100 and 400 pixels. Change-of-line movements have been excluded from both forward and backward movement sets. We also derive information about the number of visits of each word, namely the number of words that were not visited at all (skipped) and number of words that were visited more than once during the text reading.

To identify the features with the most predictive power, we employed the least absolute shrinkage and selection operator (LASSO) [12], [13], a particular case of penalized least squares regression with L1-penalty function. LASSO finds the minimum of the residual sum of squares, subject to the sum of the absolute value of the coefficients being less than a constant. The LASSO estimate can be defined by:

$$\beta^{LASSO} = \arg\min_{\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad (1)$$

In practice, as λ gets higher, less features are taken into account. Specifically, the parameter λ in LASSO regression is estimated using 5-fold cross validation. Two values of λ were examined, namely the λ_{minMSE} that corresponds to the minimum mean cross-validation error MSE (vertical dotted line in Fig. 2) and λ_{1SE} which is one standard error of the mean higher than λ_{minMSE} (vertical solid line). The purpose of the addition of the 1SE is to reduce the number of regression coefficients, while the mean square error remains close enough (1SE) to λ_{minMSE}. Both variations were considered in our analysis.

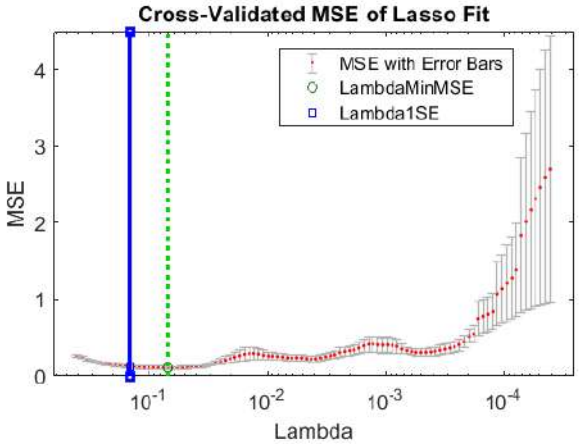

Fig. 2. Cross-Validated MSE of LASSO fit for the baseline text.

DysLexML builds classifiers based on SVM, Naïve Bayes, and K-means. The SVM with a linear kernel performs better than the ones with Gaussian or Polynomial, so only the performance of the linear kernel is reported here. The K-Means-based classifier was built as follows: the subjects of the training set are clustered using k-means, and a label is assigned to each cluster based on the most frequent label within that cluster. The distance of the test subject from the centroid of each cluster was estimated using the Euclidean distance. The classifier reports the label of the cluster whose centroid has the shortest distance from the test data.

IV. Performance analysis

Using the first dataset. DysLexML first employs the LASSO regression five-fold cross-validation to identify the dominant features. Based on the dominant features, it then applies various classification algorithms. To evaluate its performance, we used the Leave One Out Cross Validation (LOOCV), an appropriate choice given the relatively small size of the dataset. Given that there were subjects with missing values in the word specific features, we filled in the missing values with the median of the corresponding feature values of the training set.

TABLE I
Classification performance of the first dataset, including all subjects, treating the missing values. The first column corresponds to the baseline text, while the second column to the easier text.

Classifier	LOOCV accuracy	
K-means (k=2) , LASSO (λ_{minMSE})	86.95	89.39
K-means (k=3) , LASSO (λ_{minMSE})	91.30	84.84
K-means (k=4) , LASSO (λ_{minMSE})	81.15	84.84
K-means (k=2) , LASSO (λ_{1SE})	89.85	78.78
K-means (k=3) , LASSO (λ_{1SE})	86.95	84.84
K-means (k=4) , LASSO (λ_{1SE})	89.85	83.33
Linear SVM, LASSO (λ_{minMSE})	94.20	80.30
Linear SVM, LASSO (λ_{1SE})	97.10	87.87
Linear SVM, without feature selection	85.50	81.81
Naïve Bayes, LASSO (λ_{minMSE})	91.30	86.36
Naïve Bayes, LASSO (λ_{1SE})	92.75	84.84

DysLexML can accurately perform classification. DysLexML, with SVM and LASSO (λ_{1SE}), exhibits an accuracy of 97.10 % for the baseline text (Table I), while for the easy text, with K-means (of k equal to 2) and LASSO (λ_{minMSE}), it reports 89.39% correct classification.

The exclusion of the word-specific features from the feature vector results to a lower average accuracy for the baseline (difficult) text. The performance remains the same in the case of the easier text, indicating that the word-specific features are not useful when the text is not challenging for the reader.

The dominant features (as selected by the LASSO) for both texts are the mean saccade length and number of short forward movements. In the case of the baseline (difficult) text, the additional dominant features are the median saccade length, and the number of multiply fixated words. Prior research has also reported the important role of these dominant features identified by LASSO (as discussed in Section I). The distributions of the mean and the median saccade length for both populations are significantly different (as shown in Fig. 3), which explains their presence as separate dominant features.

Diversity in dyslexic population. Not all cases of dyslexia are equally severe. Note that in the experiment, the subjects were instructed to not rush their reading and understand the text in order to answer some comprehension questions at the end. This may have prolonged the reading sessions even for typical

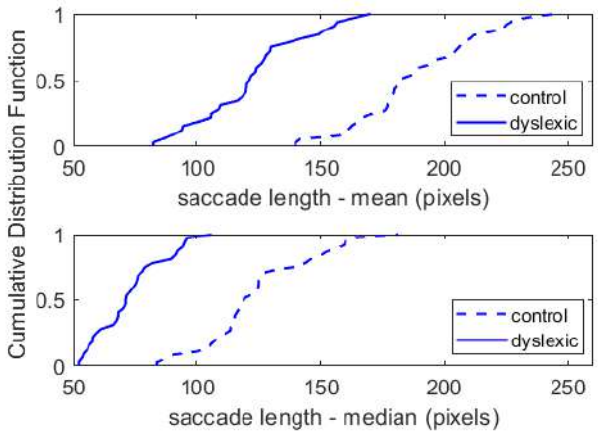

Fig. 3. Empirical CDF of saccade length features for the first dataset.

readers. The number of short forward movements was expected to play a prominent role. Dyslexics have been reported to perform more and shorter saccades during reading, in their attempt to decode the text [7]. The number of short forward movements of the dyslexic population has large variance, as shown in the upper part of Fig. 4. 50% of the dyslexic subjects have more than twice total short forward movements than the control population.

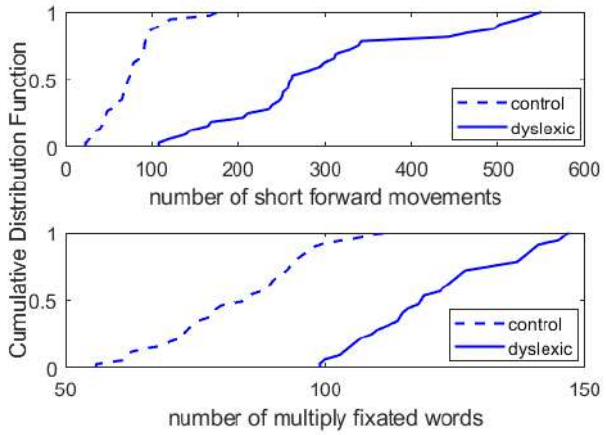

Fig. 4. Empirical CDF of number of short forward movements (top) and number of multiply fixated words, i.e. the words that have been fixated more than once during the reading session (bottom) for the first dataset.

Dyslexics tend to revisit words more often, especially those that are long or difficult to read [14]. 90% of the typical readers have less than 100 words fixated more than once, while this is the starting value for the dyslexic subjects (Fig. 4 (bottom)). This illustrates the value of the word specific analysis of the eye-tracking study.

Second field study with commercial eye-tracker and larger population. In the second systematic field study, Tobii 4C, a commercial inexpensive eye-tracker was employed. Unlike the eye-tracker in the first study that operated at 60Hz, Tobii 4C has sampling frequency of 90Hz. The use of chin rest is not necessary anymore, hence the participants can freely move their head making the procedure completely non-invasive. The field study included 152 participants in total, 72 of which were diagnosed as dyslexics. All the participants were native Greek speakers and the age span this time was larger and varied from 7 to 16 years old. The participants were reading *silently* the difficult text that was also used in the previous study. For the evaluation of DysLexML with the dataset obtained from the second field study we applied the same methodology as before.

Validation with the model built from the first study. We first employed the SVM model built on the first dataset to assess DysLexML with the second dataset, collected from the larger scale field study. The model's accuracy of 75.55% indicates that the datasets differ. The differences in the two eye trackers, the higher age variance of the second dataset compared to the first one, and the absence of chin rest resulting to a less controlled environment may cause substantial differences. Thus we repeated the analysis for identifying the dominant features for the second dataset and assess the impact of the age on the classification.

Age is a dominant feature. Age was expected to be among the dominant features, as the reading skills vary with age. Different eye movement patterns are expected from elementary school readers compared to high school children. When age is used as an extra feature in the analysis, the accuracy is improved in most cases (Table II). Age is selected as a dominant feature at every iteration of the cross-validation procedure contributing to an increase of about 12% in classification accuracy (Table II 72.36% vs 84.21%), achieving the best score again with the use of linear SVM model. Age was also reported as an important feature in [15], where the participants' age range was wider.

The number of multiply fixated words and the number of short forward movements remain dominant features. The number of multiply fixated words and number of short forward movements remain in the list of important features, as in the first study (Fig. 5). Moreover, the number of fixations, the median fixation duration, the median length of medium forward eye movements, and the age of the participant are identified as important for the classification.

Increased number of fixations and fixation duration in dyslexics. The reading skills can also affect the number of fixations and the fixation duration. The number of fixations has also been identified as a dominant feature in other studies [4]; Dyslexics make more choppy eye movements, resulting to a larger number of fixations than regular readers (Fig. 6 (top)). Moreover their fixations last more compared to the fixations of the control population (Fig. 6 (middle)), as also appears in Fig. 1.

V. Robustness to Noise

To examine the robustness of DysLexML, we evaluated its performance in the presence of noise in the form of *small displacements of the fixation points*. The noise follows

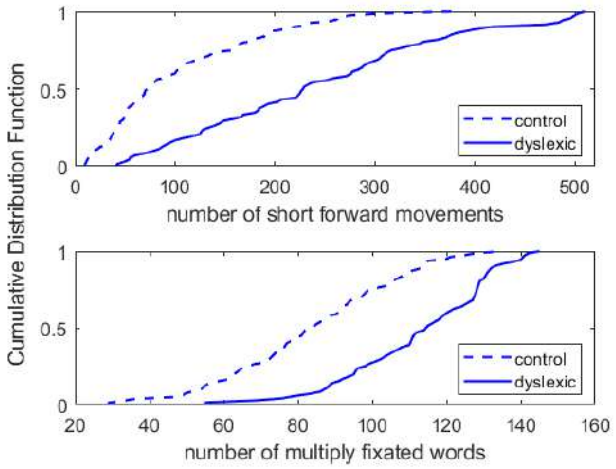

Fig. 5. Empirical CDF of number of short forward movements (top) and number of multiply fixated words (bottom) for the second dataset.

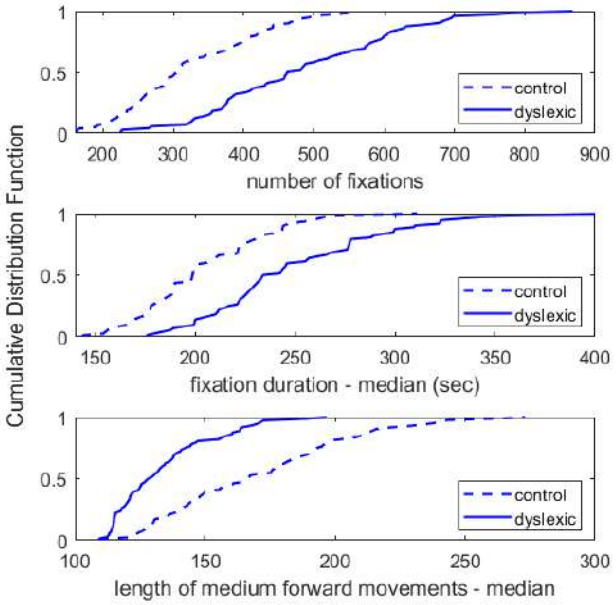

Fig. 6. Empirical CDF of the number of fixations (top), the median fixation duration (middle) and median length of medium movements (bottom) for the second dataset.

a Gaussian distribution with mean value equal to zero and standard deviation varying from 10 to 100 pixels (with a step size of 10). For small displacement, only the saccadic movement features changed. However, large displacements result to changes in the word-specific features as well. For each subject, the noise was added and the new feature vectors were generated. Note that the "shifted" eye-movements result to different feature vectors. The DysLexML was then evaluated for this new dataset. Specifically, for each σ, we generated 10 synthetic datasets. We then trained a linear SVM model using

TABLE II
CLASSIFICATION PERFORMANCE OF THE SECOND DATASET. THE FIRST COLUMN CORRESPONDS TO FEATURE SET WITHOUT THE USE OF AGE, WHILE THE SECOND COLUMN TO A FEATURE SET THAT CONTAINS AGE IN IT.

Classifier	LOOCV accuracy	
	without age	with age
K-means (k=2) , LASSO (λ_{minMSE})	71.71	69.07
K-means (k=3) , LASSO (λ_{minMSE})	63.15	69.07
K-means (k=4) , LASSO (λ_{minMSE})	65.13	70.39
K-means (k=2) , LASSO (λ_{1SE})	72.36	68.42
K-means (k=3) , LASSO (λ_{1SE})	55.26	70.39
K-means (k=4) , LASSO (λ_{1SE})	71.71	72.36
Linear SVM, LASSO (λ_{minMSE})	69.07	82.89
Linear SVM, LASSO (λ_{1SE})	67.76	84.21
Naïve Bayes, LASSO (λ_{minMSE})	69.73	71.05
Naïve Bayes, LASSO (λ_{1SE})	72.36	71.05

the dominant features that were reported by LASSO using the *original* datasets. For testing, we employed the 10 synthetic datasets with noise for each given σ. Only the children that did not have missing values from the first dataset were considered in this analysis. Fig. 7 presents the acquired results for both original datasets.

The model based on the first small field study exhibits a robust performance for relatively small noise levels, up to σ of 30 pixels, which corresponds to a displacement of something more than a character on the x-axis and 1/3 of the line on the y-axis. However, as the noise level increases, the accuracy drops significantly. Similar trend persists for the model built using the data of the second field study (larger dataset). Specifically it exhibits a robust performance for relatively small noise levels (up to σ of 20 pixels), which corresponds to displacement of something less than a character on the x-axis and 1/4 of the line on the y-axis. However, as in the case of the small dataset, as the noise level increases, the accuracy drops significantly. Through SVM, that behaves well under generalization, DysLexML addresses the noise in the fixation coordinates in a robust manner.

VI. SYNTHETIC DATA GENERATION USING GANS

User field studies, like the presented ones, are often time-consuming and cost-demanding. The small sample size of datasets makes the use of sophisticated and complex classification models prohibited due to overfitting issues. Moreover, collected measurements may contain sensitive personal information. To address these issues, we developed a data augmentation technique based on generative model for generating synthetic data that follow the same distributional characteristics as the original datasets. The aim is to create a *new surrogate* dataset, that can be used instead of the original one, as it follows the same distribution.

In recent years, novel generative modeling approaches based on neural networks have produced impressive results in generating realistic samples from complex and unknown distributions. Two popular families of generative models are Generative Adversarial Nets (GANs) [16] and Variational AutoEncoders (VAEs) [17]. Here, we focus on GANs which

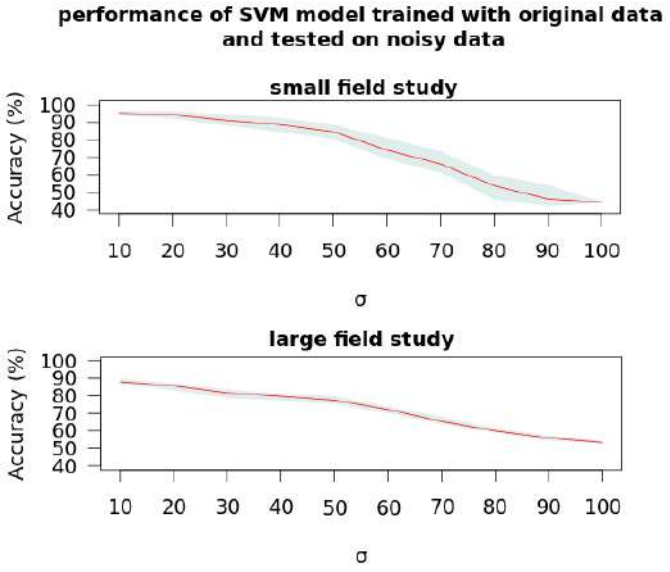

Fig. 7. Performance of DysLexML under noisy fixation positions. Training SVM linear model with original data of first field study (top) and larger scale study (bottom). The testing was performed on noisy data (10 datasets for each σ value, σ in [10, 100] with stepsize 10). The solid red line indicates the mean accuracy over the 10 noisy synthetic datasets, while the gray area represents the range between the lowest and the maximum accuracy achieved at each noise level.

consist of two powerful neural networks: the Generator and the Discriminator. The Generator shapes random noise to look like real data while the Discriminator decides if a given sample is real or fake. The pair of neural nets compete with each other forming a two-player zero-sum game. Both nets are simultaneously trained to achieve the optimum equilibrium of the game. At equilibrium, the Generator produces samples from the real distribution, while the Discriminator cannot distinguish between the original and the fake samples. An important extension of GAN is the *conditional GAN* [18], where the input noise vector is concatenated with a condition vector in order to produce samples from the conditional distributions. Thus, a practitioner is able to handle the sampling from a family of distributions based on the condition vector. We utilize conditional GANs, where we condition on the disease state to generate new samples from both classes: dyslexic cases and controls.

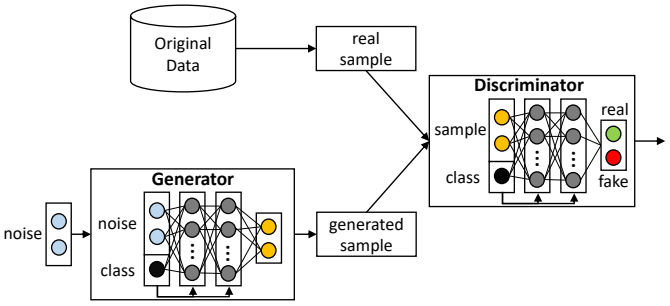

Fig. 8. Conditional GAN architecture.

The main challenge in training (conditional) GANs is the limited number of real data (152 samples in total). To alleviate this, we choose to train, evaluate, and validate GANs' performance on a reduced dimensional dataset. The reduced dataset consists of the five most significant features as selected by LASSO (λ_{1SE} approach). We trained the GAN on this new dataset and performed a moderate hyper-parameter tuning for GAN parameters. We ended up with the following setting (Fig. 8): Both Discriminator and Generator consist of a two-layer neural network with 32 hidden units each. The input noise is uniform with dimension 10, while the condition vector is not only concatenated in the input layer but also in the hidden layers. ReLU is employed as an activation function for the hidden layers. The output layer of the Discriminator has sigmoid activation, while no nonlinearity is applied to the output layer of the Generator. The training is performed using stochastic gradient descent with minibatch size of 128. We utilize Adam as an optimizer with learning rate 0.00001.

GAN-generated data are almost indistinguishable from the real data. The comparison between the real and synthetic distributions was performed using several metrics (Fig. 9). Specifically, Fig. 9 (top) shows the ECDF of the number of fixations for both control class (dashed lines) and dyslexic class (solid lines). The ECDFs of GAN-generated data (red lines) are almost indistinguishable from the ECDFs of the real data (blue lines). Similar results are obtained for the other four dominant features. Apart from the marginal distributions, we compare the overall similarity between the generated and real distributions using the Maximum Mean Discrepancy (MMD) distance [19]. MMD takes into account not only the marginals but also the correlations and the shape of the statistical distributions. Fig. 9 (bottom) shows the average MMD for both classes. The gradual decrease of MMD implies that the GAN indeed learns the real distributions. After 15K iterations and until the end of training, MMD fluctuates around the very small value 0.01.

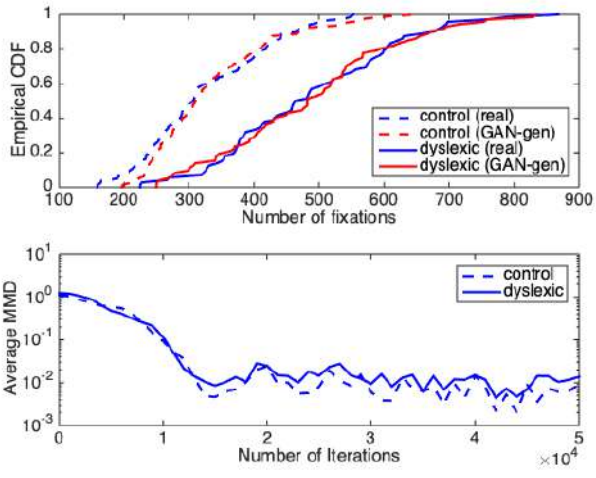

Fig. 9. Real and synthetic empirical CDFs for the number of fixations for both classes (top). Average MMD after 20 repetitions for both classes (bottom).

Slightly increased accuracy when the Naïve Bayes classifier is trained on the augmented dataset. We assess the added value of the synthetic data in classification tasks. (Fig. 10). Table III presents the average LOOCV accuracy along with its standard deviation from the classifiers trained on the augmented dataset (original and synthetic data). Results are slightly better for Naïve Bayes classifier yet not significantly different in statistical sense, while they slightly deteriorate for linear SVM. Table IV presents the average accuracy along with its standard deviation when classifiers are trained on the generated data and then evaluated on the original data. Results reveal that the substitution of the original data with synthetic data is feasible without compromising classification performance.

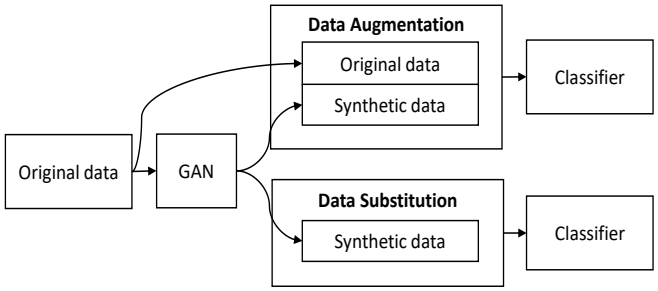

Fig. 10. GAN assessment pipeline. Original data are fed to the GAN, it gets trained and then produces synthetic data. These data are assessed by the classifiers in two ways: combined with the original data (data augmentation) and stand-alone (data substitution).

TABLE III
RESULTS FROM THE CLASSIFIERS TRAINED ON THE AUGMENTED DATASET (GAN-GENERATED AND ORIGINAL DATA) AND TESTED ON THE ORIGINAL DATA.

Classifier	LOOCV accuracy
Linear SVM, LASSO (λ_{1SE})	61.4 ± 3.5
Naïve Bayes, LASSO (λ_{1SE})	72.7 ± 0.9

TABLE IV
RESULTS FROM THE CLASSIFIERS TRAINED ON THE GAN-GENERATED DATA AND TESTED ON THE ORIGINAL DATA.

Classifier	Cross accuracy
Linear SVM, LASSO (λ_{1SE})	67.2 ± 5.1
Naïve Bayes, LASSO (λ_{1SE})	73 ± 0.8

VII. RELATED WORK

Although dyslexia has been extensively studied the last three decades with specialized eye-trackers, there is only a limited number of eye-tracking-based screening systems, partially due to the high cost of eye-trackers up to recently and the debate of the primary cause of dyslexia [8]. The lack of extensive datasets limits significantly the performance of deep-learning architectures. On the other hand, SVM is powerful in case of relatively small datasets. Recent studies have applied ML, and more specifically SVM, for classification of dyslexia on data collected from eye-trackers [15], [20]. For example, Rello and Ballesteros [15] performed a field study that included 97 Spanish language native speakers, aged 11-54 reading 12 different texts. They used a binary polynomial SVM classifier and achieved classification accuracy of 80.18%. Their feature set included the age of the participant, the text number, details about text stylistics, number of visits of an ROI, mean time spent on an ROI, total reading time, mean of fixation duration, number of fixations and sum of all fixations duration. They reported that the reading time, the mean of fixation duration, and the age of the participant have predictive power. Benfatto *et al.* [20] also employed linear SVMs with sequential optimal optimization for dyslexia screening. Their field study in Sweden included 185 children, 97 of them with high risk of dyslexia, speaking Swedish as a first language. All the subjects were reading from paper a short text adapted to their age while their eye movements were recorded. The subjects were equipped with head-mounted goggles with arrays of infrared transmitters and detectors, arranged around each eye. A chin and forehead rest was deployed to minimize head movements and stabilize the viewing distance. Their feature set, produced using a dynamic dispersion threshold algorithm, consists of 168 features. They also distinguished saccades to progressive and regressive ones. A recursive feature elimination algorithm identified the dominant features. They achieved accuracy of 95.6%±4.5% using 48 features of the original feature space. Al-Edaily *et al.* [21] developed Dyslexia Explorer in the Arabic language and performed a study with 14 subjects, 7 of whom with diagnosed dyslexia. Their system is designed to help specialists analyze visual patterns of reading and provide insights into understanding differences between readers with and without dyslexia. Their measurements included fixation duration in each/all ROI, mean fixation duration in each/all ROI, total fixation count for each/ all ROI and backward saccades.

Unlike the above ML-based approaches, Smyrnakis *et al.* [4] developed statistical Bayesian classifiers, using various thresholds and taking into consideration binary correlations. They focused on small age span, critical for dyslexia diagnosis. The size and font of the two texts used was standardized so as to achieve maximal classification accuracy, unlike in [15]. The parameters used for classification involved not only direct eye-tracking parameters, but also relations between eye-tracking parameters and word properties in the texts read. These parameters extend the parameter set used in [20]. Including this set of parameters, it is possible to evaluate word anticipation, which is often problematic in dyslexics [22]. DysLexML has been evaluated by employing the same dataset as in [4], in addition, to the second dataset, collected from a larger field study using a commercial eye-tracker. DysLexML outperforms RADAR in terms of classification accuracy: 97.10% vs. 94.2% for the baseline text. For the easy text, RADAR reports 87.9% accuracy, while DysLexML, with K-means with k equal to 2, exhibits an accuracy of 89.39%. As mentioned, the classifier of DysLexML with the best accuracy on noise-free data is the linear SVM classifier on features automatically selected by the

LASSO regression at λ_{1SE}. Furthermore, it exhibits a robust performance under fixation position noise (added artificially). Its robustness and dimensionality reduction are two innovative aspects of this work. To the best of our knowledge, we are the first to examine the impact of noise on the fixation positions on the accuracy of the classification. Unlike other screening systems for dyslexia, DysLexML *automatically* identifies the features with the largest discriminating power. It achieves high accuracy using only a small set of features (4 and 6, for the 1st and 2nd dataset, respectively). Rello and Ballesteros [15] used only 3 features for their classification by manually selecting the features.

Finally, modern data augmentation techniques which rely on deep neural networks and adversarial learning have been successfully applied in image classification [23], [24] and speech recognition [25] tasks. Nevertheless, the application of such techniques in health monitoring systems where the sample size is limited henceforth the convergence of the training not guaranteed has not been explored. To the best of our knowledge, we are the first to perform dyslexia data generation using GANs. It is also among the first published attempts of training GANs using *small datasets*.

VIII. Conclusions

DysLexML's feature selection, via LASSO with λ of one standard error enabled dimensionality reduction, without compromising the accuracy. In the first field study, the mean and median saccade length, the number of short forward movements, and the number of multiply fixated words are the four features with the most prominent predictive power for the baseline text, while for the easier text only the mean saccade length and the number of short forward movements were selected. In the second larger-scale study, the number of fixations, median fixation duration, number of short forward movements, median length of medium forward movements, number of multiply fixated words, and age are the six dominant features.

From the analysis of the small field scale dataset, the text difficulty does play an important role in the diagnosis: Easy, less challenging, text reduces the power of the word-specific features, as they disappear from the dominant feature set. The text choice has to be relevant to the subjects age and acquired reading skills. The selected features are easily interpreted and capture the prior knowledge about eye movements of dyslexic children. To the best of our knowledge, DysLexML uses the smallest feature set compared to the other related studies.

Given our vision for screening systems that operate in less-controlled, larger-scale environments (e.g., potentially in kindergartens or homes) with commercial eye-trackers, reaching a larger population, the robustness of the system under noise is critical. As a first step towards this objective, we added synthetic noise at the fixation positions and assessed its impact on the accuracy. For noise levels smaller than σ equal to 20 pixels, the performance of the system remains robust. Encouraged by the robustness under noise, the team has been performing follow-up larger-scale field study using inexpensive non-specialized eye-trackers in more diverse settings (e.g., different countries and under silent and out-loud reading). One of the future work plans is the identification of different classes of reading difficulties. Finally, our proposed synthetic data generation method based on GANs is able to create synthetic samples of high relevance with the original dataset. We plan to use such synthetic datasets to train more sophisticated ML and deep learning methods to improve the classification accuracy. This work sets the basis for developing a screening tool that can reach a larger and more diverse population, in less controlled environments, aiming for early intervention and potentially larger social impact.

References

[1] T. Høien and I. Lundberg, *Dyslexia: From Theory to Intervention*. Springer, 2000.

[2] C. Hulme and M. Snowling, "Reading disorders and dyslexia," *Current Opinion in Pediatrics*, vol. 8, no. 6, pp. 731–735, 2016.

[3] A. Casale, "Identifying dyslexic students : The need for computer-based dyslexia screening in higher education," in *Estro: Essex Student Research Online*, vol. 1, 2010.

[4] I. Smyrnakis, V. Andreadakis, V. Selimis, M. Kalaitzakis, T. Bachourou, G. Kaloutsakis, G. D. Kymionis, S. Smirnakis, and I. M. Aslanides, "Radar: A novel fast-screening method for reading difficulties with special focus on dyslexia," *PLOS ONE*, vol. 12, no. 8, pp. 1–26, 2017.

[5] S. Bellocchi, M. Muneaux, M. Bastien-Toniazzo, and S. Ducrot, "I can read it in your eyes: What eye movements tell us about visuo-attentional processes in developmental dyslexia," *Research in Developmental Disabilities*, vol. 34, no. 1, pp. 452–460, 2013.

[6] G. F. Eden, J. F. Stein, H. M. Wood, and F. B. Wood, "Differences in eye movements and reading problems in dyslexic and normal children," *Vision Research*, vol. 34, no. 10, pp. 1345–1358, 1994.

[7] F. J. Martos and J. Vila, "Differences in eye movements control among dyslexic, retarded and normal readers in the spanish population," *Reading and Writing*, vol. 2, no. 2, pp. 175–188, 1990.

[8] K. Rayner, "Eye movements in reading and information processing: 20 years of research.," *Psychological Bulletin*, vol. 124, no. 3, pp. 372–422, 1998.

[9] R. D. Elterman, L. A. Abel, R. B. Daroff, L. F. Dell'Osso, and J. L. Bornstein, "Eye movement patterns in dyslexic children," *Journal of Learning Disabilities*, vol. 13, no. 1, pp. 16–21, 1980.

[10] M. De Luca, E. Di Pace, A. Judica, D. Spinelli, and P. Zoccolotti, "Eye movement patterns in linguistic and non-linguistic tasks in developmental surface dyslexia," *Neuropsychologia*, vol. 37, no. 12, pp. 1407–1420, 1999.

[11] D. Salvucci and J. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, pp. 71–78, ACM, 2000.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[13] V. Fonti and E. N. Belitser, "Feature selection using lasso," 2017.

[14] J. Hyönä and R. K. Olson, "Eye fixation patterns among dyslexic and normal readers: effects of word length and word frequency," *Journal of Experimental Psychology: Learning, Memory, & Cognition*, vol. 2, 1995.

[15] L. Rello and M. Ballesteros, "Detecting readers with dyslexia using machine learning with eye tracking measures," in *Proceedings of the 12th Web for All Conference*, pp. 16:1–16:8, ACM, 2015.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, Curran Associates, Inc., 2014.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013. arXiv:1312.6114 [stat.ML].

[18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014. arXiv:1712.04621 [cs.LG].

[19] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.

[20] M. Nilsson Benfatto, G. Öqvist Seimyr, J. Ygge, . Pansell, A. Rydberg, and C. Jacobson, "Screening for dyslexia using eye tracking during reading," *PLOS ONE*, vol. 11, no. 12, pp. 1–16, 2016.

[21] A. Al-Edaily, A. Al-Wabil, and Y. Al-Ohali, "Dyslexia explorer: A screening system for learning difficulties in the arabic language using eye tracking," in *Human Factors in Computing and Informatics*, pp. 831–834, Springer Berlin Heidelberg, 2013.

[22] F. Huettig and B. Susanne, "Delayed anticipatory spoken language processing in adults with dyslexia–evidence from eye-tracking," *Dyslexia*, vol. 21, no. 2, pp. 97–122, 2015.

[23] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017. arXiv:1712.04621 [cs.CV].

[24] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," 2019. arXiv:1904.08779 [eess.AS].