Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests

Article *in* ICT Express · March 2022

DOI: 10.1016/j.icte.2022.02.011

2 authors:

Shahriar Kaisar
RMIT University
36 PUBLICATIONS 1,003 CITATIONS
SEE PROFILE

Abdullahi Chowdhury
University of South Australia
37 PUBLICATIONS 516 CITATIONS
SEE PROFILE

The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com
ScienceDirect

ICT Express 8 (2022) 563–568

www.elsevier.com/locate/icte

Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests

Shahriar Kaisar[a,*], Abdullahi Chowdhury[b]

[a] Department of Information Systems and Business Analytics, RMIT University, Australia
[b] Faculty of Engineering, Computer and Mathematical Sciences, University of Adelaide, Australia

Received 11 August 2021; received in revised form 29 December 2021; accepted 28 February 2022
Available online 15 March 2022

Abstract

Developmental Dyslexia is a learning disorder often discovered in school-aged children who face difficulties while reading or spelling words even though they may have average or above-average levels of intelligence. This ultimately results in anger, frustration, low self-esteem, and other negative feelings. Early detection of Dyslexia can be highly beneficial for dyslexic children as their learning needs can be properly addressed. Researchers have used several testing techniques for early discovery where the data is collected from reading and writing tests, online games, Magnetic reasoning imaging (MRI) and Electroencephalography (EEG) scans, picture and video recording. Several Machine learning techniques have also been used in this regard recently. However, existing works did not focus on the problem of the imbalanced dataset where the percentage of dyslexic participants is much higher compared to non-dyslexic participants, which is expected to be the case for pre-screening among a random population. This paper addresses the imbalanced dataset obtained from dyslexia pre-screening tests and proposes an oversampling and ensemble-based machine learning technique for the detection of Dyslexia. Simulation results show that the proposed approach improves the detection accuracy of the minority class, i.e., dyslexic patients from 80.61% to 83.52%.

Keywords: Dyslexia; Imbalanced data; Ensemble technique; Machine learning; Oversampling

1. Introduction

The word ‘Dyslexia’ is originated from the Greek language, which refers to difficulty with words. This is not a disease rather a type of specific learning difficulty (SLD) where a person finds it difficult to read, spell or write even though they may possess average or above-average levels of intelligence. Among native English speakers, 10% of Australians suffer from dyslexia while the rate is much higher (20%) for other nationals including the citizens of Canada and the UK. A similar trend is present among other language users [1,2]. Developmental Dyslexia (i.e., dyslexia which is genetic, present from birth, and develops over time) is generally observed among the younger school-going population who performs poorly in schools and often faces difficulties during their studies. This ultimately results in negative emotions, such as anger, frustration, depression, anxiety, and low self-esteem [3]. Therefore, early detection of Dyslexia is critical to provide them with the necessary support and foster their development. This paper addresses the detection of dyslexia through pre-screening tests to help identify potential dyslexic children at the early stages of their academic careers and provide them with additional support for an equitable learning experience.

Although early detection of dyslexia is crucial for effective remediation, conventional techniques are expensive and require professional oversight [4]. Recently, machine learning techniques have become popular in predicting Dyslexia from the data collected through online tests, such as reading and writing exercises, online games, tracking eye movement, or collecting EEG scans and MRI data while the participants engage in reading or writing tasks [5]. Among these techniques, online gamified tests have recently gained popularity among researchers as they are cost-effective, easier to conduct, and can cover a wide participant base [2,4,6]. In this case, participants engage in online games while information about their performance is collected and later analyzed for detecting dyslexia. However, a large-scale user study of such tests produces an imbalanced dataset as the percentage of non-dyslexic participants is likely to be much higher compared

* Corresponding author.
E-mail addresses: shahriar.kaisar@rmit.edu.au (S. Kaisar), Abdul.Chowdhury@adelaide.edu.au (A. Chowdhury).

Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2022.02.011

to the percentage of dyslexic participants [6], which is not explicitly addressed in the existing literature.

In real-life scenarios, the number of observations will not be the same for all classes in a random collection sample. For example, in malware or dyslexia detection, most of the observations are likely to be from the "Not-malware" or "Not-dyslexia" class, and a small portion of the data would represent "malware" or "dyslexia" class, respectively. Machine learning or Deep learning classifiers and learner models demonstrate biasness to the larger portion of observed data (known as majority class) for this type of imbalanced dataset to increase the detection accuracy. However, these learner models generally show poor performance in detecting the data belonging to the smaller class (known as minority class). In this case, the oversampling methods can be useful to increase minority class observations by generating more minority class samples. Different oversampling methods use different techniques to generate additional synthetic minority class samples closer to the original minority class distribution and have been effectively used in bankruptcy detection [7], cervical cancer identification [8], and intrusion detection [9]. In this work, we have used different oversampling techniques to investigate which method provides a better result in detecting the minority class data without impacting the overall detection accuracy.

Machine learning and Deep learning methods are getting popular in the medical domain for medical diagnosis and decision-making purposes [10–13]. Sarivougioukas and Vagelatos [10] used denotational mathematics as a framework for modeling a deep neural network to enhance the quality of medical decision making. Authors in [11,12], and [13] have used Naive Bayes, K-nearest Neighbor, and Random Forest classifier to detect breast cancer, lung cancer, and liver injuries, respectively. All these works used a single classifier or single ensemble technique. A combination of ensemble methods rather than a single classifier is expected to provide better results as they can incorporate learning from multiple weak learners to make a decision [14]. Therefore, we have also incorporated an ensemble classifier to improve the performance of our proposed approach. To the best of our knowledge, no prior work on dyslexia detection and pre-screening tests incorporated oversampling and ensemble-based machine learning techniques to improve the detection accuracy of minority class as well as the overall performance, which is the novelty of our current work.

2. Related work

The traditional dyslexia detection techniques include standardized tests conducted by psychologists where participants are provided with reading or writing tasks, phonological awareness, and working memory tests. A poor performance in these tests is attributed to Dyslexia. However, these techniques require the presence of a specialist, are time-consuming and costly, and hence machine learning-based detection techniques are becoming increasingly popular.

Researchers have used different types of tests including reading [15–18], writing [19,20] and online games [2,4,6] to collect different types of data, such as text [4,6,17], image [18, 20], video [18], Eye movement tracking [15], MRI scans [16] and EEG scans [19]. Afterward, the collected data is processed and analyzed using different machine learning techniques to predict and classify dyslexic and non-dyslexic participants. However, the above-mentioned works mostly used a single classifier whose performance can be further improved using an ensemble classifier, which is explored in this article. Among these methods, the online game-based methods are most cost-effective and can cover a wide participant base within a short time, and do not require any specialized tools, such as EEG headset, camera, eye tracker, or MRI scanner. However, a large-scale user study of online game-based data collection method generally results in an imbalanced dataset as reported in [4,6]. In those cases, although the overall performance of machine learning techniques remains fairly stable (80%–85% [4,6], the detection of the minority class, i.e., the dyslexic participant requires further improvement, which is investigated in this work using oversampling techniques.

Oversampling methods are used in different domains to increase the detection accuracy of the less occurred cases (i.e., minority class). For example, fault diagnosis in [21], real-time accident detection in [22], and breast cancer detection in [23] employed different oversampling techniques. Therefore, oversampling methods are likely to produce high detection accuracy for minority class detection in dyslexia pre-screening tests, which is not addressed in the literature.

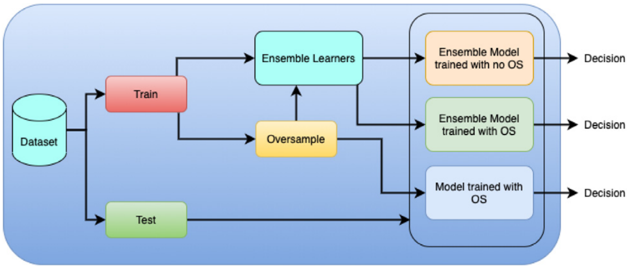


Fig. 1. Proposed Approach.

3. Proposed approach

Fig. 1 shows a schematic representation of the proposed approach. In this case, the data can be collected from online gamified tests for pre-screening of Dyslexia as discussed in [4, 6]. The collected data can be further divided into training and test dataset. We have used 70%–30% splits for training and test data. The proposed approach uses three different models. For the first model, the training dataset is directly used as an input to ensemble learners, while in the second model, the training dataset passes through an oversampling module before being used by ensemble learners and finally, the third model only uses an oversampling technique with a single classifier (i.e., baseline classifier). The test dataset is then applied to these three different models for performance evaluation. The details of the dataset, oversampling techniques, and ensemble learners are discussed below.

3.1. Dataset

We have used the dataset generated by authors in [6] who conducted an online gamified test with 3644 participants and collected data about 196 different features. The first four features are demographic features (e.g., Age, Gender, Native Language, and Language subjects), while the remaining 192 features are based on 32 different questions, which were asked to each participant. Each question was measured with six points (clicks, hits, misses, score, accuracy, and miss rate). The final column of the dataset is the output column. This column has two classes, Yes and No. 'Yes' represents the participant may have Dyslexia and should go for further testing while 'No' represents the participant does not have Dyslexia. The Dyslexia negative count for this dataset is 3252 (89.218%), and Dyslexia positive count is 392 (10.754%).

3.2. Oversampling methods

Since the above-mentioned dataset consists of only 10.754% minority class data points, the learner model has very little data to learn from the minority class. Therefore, we have used three widely used oversampling methods, namely Synthetic Minority Over-sampling Technique (SMOTE), Borderline-SMOTE [11], and Adaptive Synthetic (ADASYN) [24] to address this issue.

SMOTE [11] generates new synthetic samples from the minority class to generate a class-balanced or nearly class-balanced training set. This approach takes a number (K) of Nearest Neighbors (KNN) for each minority class sample ($\mathbf{x}$). After that, it interpolates the selected sample and one of the randomly selected neighbors ($\mathbf{x}_{ch}$) of the selected sample and creates a new synthetic sample x_{new} using interpolation:

$$\mathbf{x}_{new} = \mathbf{x} + (\mathbf{x}_{ch} - \mathbf{x}) \times random(0, 1) \tag{1}$$

where $random(0, 1)$ is a value between 0 and 1.

One of the major drawbacks of SMOTE is it does not work well if the observation appears in the minority class, but they are outlying.

To resolve the issue, Borderline-SMOTE [12] ignores all minority class samples as noise points if all of the neighbors of the sample are in the majority class. It selects the minority samples whose neighbors have both minority and majority class samples. The ADASYN [24] finds the impurity of the minority class sample by considering the number of majority class neighbor samples. If the impurity is higher, it generates more synthetic samples for those minority class samples. For each minority class sample x_i, ADASYN randomly chooses one minority data sample x_{zi} from K nearest neighbors and creates a synthetic sample (s_i) using:

$$s_i = x_i + (x_{zi} x_i) \times \beta \tag{2}$$

where $(x_{zi}\ x_i)$ is the difference vector in n dimensional spaces, and β is a random number: $\beta \in [0, 1]$.

3.3. Baseline learners or single classifiers

We have used different baseline learners, which are discussed below.

Logistic regression (LR) is a supervised machine learning technique that employs a linear binary classification procedure to categorize data into a set of discrete classes [25]. Using the following equation, an input value (x) is directed to anticipate an output value (y), which is a binary number (0 or 1).

$$y = \frac{e^{(b_0+b_1x)}}{1 + e^{(b_0+b_1x)}} \tag{3}$$

here, y is the prediction, b_0 is the bias or intercept, and b_1 is the coefficient value of x.

Support Vector Machine (SVM) is a useful technique for data classification and regression applications. It looks for the optimal separating hyperplane (OSH). It also uses a kernel function to translate data into a higher-dimensional space for the development of OSH. To perform better, the rule of thumb is to restrict a higher constraint on the normalization rather than limiting the rate of error [26]

In the hyperplane $\mathbf{w}.\mathbf{x} + b = 0$, the output y_i for response x_i is:

$$y_i(w.x_i + b) \geq 1 - \varepsilon_i, \triangledown i \tag{4}$$

where ε_i is slack variable used to achieve the optimum solution.

Decision Tree (DT) has nodes and leaves, which are formed throughout the learning process, i.e. tree construction. Each node of the tree works on an attribute, and the leaves that branch out from this attribute are used to examine the class label based on the attribute's value. The sequence continues until the final class label is calculated by traversing all leaves [26]. The tree construction is based on the information gain calculated by the entropy of sample S and attributes A_j as

$$Infogain(A_j) = Entropy(S) - Entropy(A_j) \tag{5}$$

3.4. Ensemble learners

Ensemble learners combine the prediction from multiple weak learners. The prediction of weak learners can be combined using different methods. In our approach, we have used the voting model and analyzed the performance of both hard and soft voting methods. In the hard voting method, the prediction output for each classifier is given as 1 or 0. The majority votes received (number of 1 or 0) are considered as the final output. In soft voting, every classifier in the ensemble model calculates the probability of the target sample to be in a target class. The final output is determined by the weighted prediction probability of all classifiers. We have used adaptive boosting (AdaBoost), gradient boosting (GB) and extreme gradient boosting (XGBoost) as our ensemble learning models.

Adaboost is a method for combining numerous weak classifiers in an iterative process. This method works by training multiple weak classifiers for the same training data, modifying the sample weight based on the results of each training and the accuracy of the previous overall classification, and then training the next weak classifier with the new data after the weight has been changed. AdaBoost learns from the weighted error rate of the sample that was classified incorrectly. Gradient Boosting (GB) also learns from the sample classified incorrectly, but unlike AdaBoost, it learns from the residual error. GB is an approach for combining weak learners to build a strong learner. The categorization in this approach is based on the residuals of the previous iteration, with the influence of each attribute being examined one by one until the desired accuracy is reached. The residuals are calculated using a loss function that is optimized via gradient descent. XGBoost, on the other hand, uses gradient descent architecture for the ensemble learning method.

Different types of datasets may have different types of distributions, feature values, and outliers. To address this, our proposed approach generates synthetic minority samples using different oversampling methods rather than using the original dataset (with no oversampling) or a single oversampling method to provide the learning models better opportunities to learn from the imbalanced data. The proposed approach also uses an ensemble classifier rather than a single classifier to produce the best detection result with minimum false detection. Although we have used the proposed approach for dyslexia detection, it can also be used to handle imbalanced classification problems in cybersecurity, fraud detection, medical decision making, and other similar domains.

4. Performance evaluation

We have implemented our proposed approach using Python, and the results obtained from the experiment are presented in Table 1. We have used commonly used metrics, such as Accuracy, Precision, Recall, and ROC, to measure the performance of different models. The table shows the performance of single classifiers, such as support vector machine (SVM), logistic regression (LR), and decision tree (DT), as well as the performance of ensemble classifiers, such as GB and XGB, on the testing dataset. In the first two cases, no oversampling methods were used. In our next step of the experiments, we added three oversampling techniques, namely SMOTE, ADASYN, and Borderline, and measured their performance with single classifiers, such as LR, DT, and SVM. This shows the improvements we obtained due to the incorporation of the oversampling method. Finally, we employed oversampling and ensemble methods together to see their combined impact.

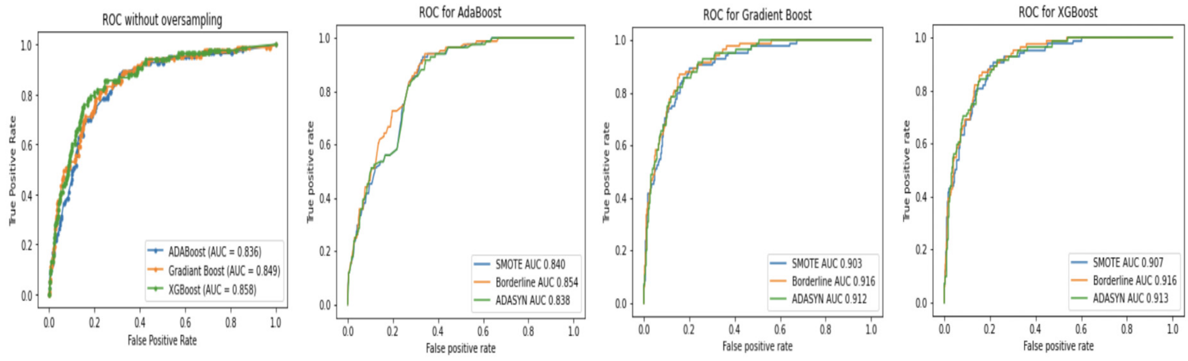
Table 1 suggests that among the single classifiers, SVM achieved the highest accuracy (89.1%) when no oversampling was applied. On the other hand, ensemble learner GradientBoost achieved an accuracy of 90.2% when no oversampling method was applied. However, if we closely look at the recall column, we can see that both of them achieved poor results (SVM 23.8% and GB 13.4%). For an imbalanced dataset, the recall value carries more importance than the accuracy as even a naive learner (predicting every sample as non-dyslexic in this dataset) will still achieve 89.2% accuracy. On the other hand, a lower recall actually suggests that a high number of participants who are Dyslexic were classified as non-dyslexic during the prediction phase, which can be very harmful to diagnosis purposes [6]. On the contrary, when we applied oversampling technique, the recall value increased up to 56% for the SVM technique suggesting more dyslexic participants were correctly classified. Table 1 also indicates that incorporation of oversampling technique with ensemble learners significantly improved the performance in terms of recall value and accuracy. Borderline oversampling with XGB achieved the highest accuracy of 89.8%, while the combination of ADASYN and XGB achieved the highest recall rate (89.6%). Although the precision value has decreased due to oversampling and ensemble technique suggesting that more non-dyslexic participants were classified as dyslexic, however, for this kind of test, it is more important to identify participants who are likely to develop dyslexia rather than sending someone without dyslexia to a specialist [6].

Table 1
Performance Matrices.

		Accu	Prec	Rec	ROC
No oversampling single classifier	LR	0.827	0.681	0.023	0.781
	DT	0.845	0.841	0.154	0.807
	SVM	0.891	0.813	0.238	0.781
No oversampling ensemble	ADABoost	0.886	0.685	0.019	0.834
	GB	0.902	0.75	0.134	0.866
	XGB	0.899	0.67	0.127	0.863
Single classifier with oversampling	LR	0.834	0.716	0.291	0.809
	DT	0.855	0.825	0.569	0.827
	SVM	0.723	0.794	0.563	0.809
Ensemble (GB) with oversampling	SMOTE	0.875	0.48	0.83	0.898
	Borderline	0.888	0.54	0.831	0.906
	ADASYN	0.883	0.483	0.861	0.901
Ensemble (XGB) with oversampling	SMOTE	0.887	0.486	0.835	0.896
	Borderline	0.898	0.523	0.835	0.899
	ADASYN	0.883	0.475	0.896	0.90

Fig. 2 shows Receiver Operating Characteristics (ROC) curves for different cases. The AUC for AdaBoost, GradientBoost, and XGBoost is 0.836, 0.849, and 0.858, respectively, without oversampling (ref Fig. 2a). After applying different oversampling methods, from Fig. 2(b), 2(c), and 2(d), we can see that the AUC values increased for all three ensemble methods. Borderline-SMOTE increased AUC from 0.836, 0.849, and 0.858 to 0.854, 0.916, and 0.916 for AdaBoost, GradientBoost, and XGBoost, respectively. Such improvements are achieved due to the use of oversampling and ensemble methods.

Table 2 shows a comparison of the result reported in [6] and our approach. We have used the results achieved by ADASYN with XGB for this comparison. The table shows that the Accuracy, Recall, and Roc reported in [6] are 0.798, 0.806, and 0.837, respectively, while the proposed approach improved those metrics and achieved 0.883, 0.896, and 0.90.



(a) (b) (c) (d)

Fig. 2. ROC curves: (a) ROC Curve with no oversampling, (b)ROC Curve for AdaBoosting with oversampling, (c) ROC Curve for Gradient Boosting with oversampling, and (d) ROC Curve for XGBoost with oversampling.

Table 2
Comparison with other Model.

Method	Accuracy	Recall	ROC
Rello et al. [6]	0.798	0.806	0.837
XGB with ADASYN in our Approach	0.898	0.835	0.899

5. Conclusion

Online gamified tests and machine learning techniques have recently gained popularity as pre-screening tests for dyslexia detection. However, the dataset obtained from such tests is naturally imbalanced due to a high number of non-dyslexic participants. Existing literature for dyslexia detection did not explicitly address this class imbalance problem and achieved a lower detection rate for dyslexic participants. The imbalanced classification problem can be addressed using oversampling techniques and ensemble learners, which are investigated in this paper. Simulation results suggest that incorporation of such techniques improves the detection accuracy of the minority class, i.e., dyslexic participants.

CRediT authorship contribution statement

Shahriar Kaisar: Conceptualization, Methodology, Formal analysis, Editing, Writing – original draft, Writing - review & editing. **Abdullahi Chowdhury:** Conceptualization, Methodology, Formal analysis, Programming, Editing, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Asvestopoulou, V. Manousaki, A. Psistakis, I. Smyrnakis, V. Andreadakis, I.M. Aslanides, M. Papadopouli, DysLexML: Screening tool for dyslexia using machine learning, 2019, arXiv preprint arXiv:1903.06274.

[2] M. Rauschenberger, L. Rello, R. Baeza-Yates, J.P. Bigham, Towards language independent detection of dyslexia with a web-based game, in: Proceedings of the 15th International Web for All Conference, Lyon, France, 2018, pp. 1–10.

[3] S.S.A. Hamid, N. Admodisastro, A. Kamaruddin, A study of computer-based learning model for students with dyslexia, in: 9th Malaysian Software Engineering Conference, MySEC, KL, Malaysia, IEEE, 2015, pp. 284–289.

[4] L. Rello, E. Romero, M. Rauschenberger, A. Ali, K. Williams, J.P. Bigham, N.C. White, Screening dyslexia for english using HCI measures and machine learning, in: Proceedings of the International Conference on Digital Health, Lyon, France, 2018, pp. 80–84.

[5] S. Kaisar, Developmental dyslexia detection using machine learning techniques : A survey, Elsevier ICT Express 6 (3) (2020) 181–184.

[6] L. Rello, R. Baeza-Yates, A. Ali, J.P. Bigham, M. Serra, Predicting risk of dyslexia with an online gamified test, PLoS One 15 (12) (2020) 1–15.

[7] T. Le, M.T. Vo, B. Vo, M.Y. Lee, S.W. Baik, A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction, Complexity, Hindawi (2019) 1–13.

[8] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, S. Annamalai, Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier, J. Med. Syst. 43 (9) (2019) 1–19.

[9] D. Gonzalez-Cuautle, A. Hernandez-Suarez, G. Sanchez-Perez, L.K. Toscano-Medina, J. Portillo-Portillo, J. Olivares-Mercado, H.M. Perez-Meana, A.L. Sandoval-Orozco, Synthetic minority oversampling technique for optimizing classification tasks in botnet and intrusion-detection-system datasets, Appl. Sci. 10 (3) (2020) 794.

[10] J. Sarivougioukas, A. Vagelatos, Modeling deep learning neural networks with denotational mathematics in UbiHealth environment, Int. J. Softw. Sci. Comput. Intell. (IJSSCI) 12 (3) (2020) 14–27.

[11] K.-J. Wang, B. Makond, K.-H. Chen, K.-M. Wang, A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients, Appl. Soft Comput. 20 (2014) 15–24.

[12] K.-J. Wang, A.M. Adrian, K.-H. Chen, K.-M. Wang, A hybrid classifier combining borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: A case study in Taiwan, Comput. Methods Programs Biomed. 119 (2) (2015) 63–76.

[13] Y.-Q. Liu, C. Wang, L. Zhang, Decision tree based predictive models for breast cancer survivability on imbalanced data, in: 3rd International Conference on Bioinformatics and Biomedical Engineering, Beijing, China, IEEE, 2009, pp. 1–4.

[14] I. Cvitić, D. Peraković, M. Periša, B. Gupta, Ensemble machine learning approach for classification of IoT devices in smart home, Int. J. Mach. Learn. Cybern. (2021) 1–24.

[15] M.N. Benfatto, G.O. Seimyr, J. Ygge, T. Pansell, A. Rydberg, C. Jacobson, Screening for dyslexia using eye tracking during reading, PLoS One 11 (12) (2016).

[16] P. Płoński, W. Gradkowski, I. Altarelli, K. Monzalvo, M. van Ermingen-Marbach, M. Grande, S. Heim, A. Marchewka, P. Bogorodzki, F. Ramus, et al., Multi-parameter machine learning approach to the neuroanatomical basis of developmental dyslexia, Hum. Brain Map. 38 (2) (2017) 900–908.

[17] R.U. Khan, J.L.A. Cheng, O.Y. Bee, Machine learning and dyslexia: Diagnostic and classification system (DCS) for kids with learning disabilities, Int. J. Eng. Technol. 7 (3.18) (2018) 97–100.

[18] S.S.A. Hamid, N. Admodisastro, N. Manshor, A. Kamaruddin, A.A.A. Ghani, Dyslexia adaptive learning model: student engagement prediction using machine learning approach, in: 3rd International Conference on Soft Computing and Data Mining, Johor Bahru, Malaysia, Springer, 2018, pp. 372–384.

[19] H. Perera, M.F. Shiratuddin, K.W. Wong, K. Fullarton, EEG signal analysis of writing and typing between adults with dyslexia and normal controls, Int. J. Interact. Multimed. Artif. Intel. 5 (1) (2018) 62.

[20] K. Spoon, D. Crandall, K. Siek, Towards detecting Dyslexia in children's handwriting using neural networks, in: Proceedings of the 36th International Conference on Machine Learning, California, USA, 2019, pp. 1–5.

[21] W. Zhang, X. Li, X.-D. Jia, H. Ma, Z. Luo, X. Li, Machinery fault diagnosis with imbalanced data using deep generative adversarial networks, Measurement 152 (2020) 107377.

[22] A.B. Parsa, H. Taghipour, S. Derrible, A.K. Mohammadian, Real-time accident detection: coping with imbalanced data, Accid. Anal. Prev. 129 (2019) 202–210.

[23] J. Nahar, T. Imam, K.S. Tickle, A.S. Ali, Y.-P.P. Chen, Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer, Expert Syst. Appl. 39 (16) (2012) 12371–12377.

[24] H. He, Y. Bai, E.A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: IEEE International Joint Conference on Neural Networks, Hongkong, China, IEEE, 2008, pp. 1322–1328.

[25] S. Sperandei, Understanding logistic regression analysis, Biochem. Med. 24 (1) (2014) 12–18.

[26] S.S. Shafin, S.A. Prottoy, S. Abbas, S.B. Hakim, A. Chowdhury, M. Rashid, et al., Distributed denial of service attack detection using machine learning and class oversampling, in: International Conference on Applied Intelligence and Informatics, Springer, 2021, pp. 247–259.