International Journal of Engineering & Technology, 7 (3.18) (2018) 97-100

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Technical paper

Machine Learning and Dyslexia: Diagnostic and Classification System (DCS) for Kids with Learning Disabilities

Rehman Ullah Khan[1]*, Julia Lee Ai Cheng[1], Oon Yin Bee[1]

[1]Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
**Corresponding author E-mail:krullah@unimas.my*

Abstract

New generation is the future of every nation, but dyslexia which is a learning disability is spoiling the new generation. Most experts are using manual techniques to diagnose dyslexia. Machine learning algorithms are capable enough to learn the knowledge of experts and intelligently diagnose and classify dyslexics. This research proposes such a machine learning based diagnostic and classification system. The system is trained by human expert classified data of 857 school children scores in various tests. The data was collected in another fundamental research of designing special tests for dyslexics. Twenty-fifth percentile was used as threshold. The scores equal to the threshold and below were marked as indicators of children who were likely to have dyslexia while the scores above the threshold were considered to be indicators of children who were non-dyslexic. The system has three components: the diagnostic module is a pre-screening application that can be used by experts, trained users and parents for detecting the symptoms of dyslexia. The second module is classification, which classifies the kids into two groups, non-dyslexics and suspicious for dyslexia. A third module is an analysis tool for researchers. The results show that 20.7% of students seem to be dyslexic out of 257 in the testing data set which has confirmed by human expert.

Keywords: *Classification of dyslexics; Diagnosis of dyslexics; Dyslexia; Learning disabilities; Machine learning systems.*

1. Introduction

Dyslexia is a specific learning disability that is neurobiological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities [1]. These difficulties typically result from a deficit in the phonological component of language that is often unexpected in relation to other cognitive abilities and the provision of effective classroom instruction. Secondary consequences may include problems in reading comprehension and reduced reading experience that can impede the growth of vocabulary and background knowledge.

Malaysia is a relatively young nation, having a population of 31.66 million [2]. Five percent of the Malaysian population is primary school going children [2]. Education is one of the vehicles for achieving the country's vision. In Malaysia, 4-10 % of students have been found to have the signs and symptoms of dyslexia [3].

Despite the heterogeneity in reading disabilities, dyslexia has been the predominant focus in Malaysia[4]. Additionally, norm-referenced assessment for the Malaysian population is non-existent. Given that Malay is the national language and English is the lingua franca, bilingualism among Malaysians is expected [5]. Reading disabilities have detrimental consequences in children lives. All such children need help, but due to limited resources, logistics, and social stigma, they are unable to get the training that they require.

Therefore, there is a need for developing automated interventions to provide diagnosis to support children with reading disabilities including dyslexia. These automated interventions should be standardized, objective, repeatable, low-cost, and can be deployed outside of the clinic or at fingertips of the children. Therefore, we propose Diagnosis and Classification System to diagnose category and intensity of disability and then classify kids with learning disabilities based on the kids score from different tests. This system will provide the necessary literacy foundation in reading and writing starting at the earliest age possible (i.e., preschool). Given the uniqueness of the Malaysian education system that focuses on both English and Malay, the research and design of this Diagnostic and Classification System will contribute to the extant literature on technologies in learning difficulties, bilingualism, literacy diagnosis, and reading disabilities.

With this glaring need in research and practice, this research was conducted to provide an early diagnosis to schoolchildren in Malaysia, more specifically in a sample of children in Kuching, Sarawak. Given the lack of expertise in the area of diagnosis, the ultimate objectives of this research are

1. To develop and evaluate diagnostic and classification system for kids with learning disabilities.
2. To estimate dyslexia prevalence using predictions from the diagnostic and automated classification system.

2. Literature Review

Becoming literate is an important milestone in a child's scholarly life and a pathway to academic success into adulthood [6,7]. However, around the world and in Malaysia, there are children who continue to fall between the "cracks" and are left behind [8, 9]. Drawing from studies overseas, the researchers know that approximately 12% of the United States school population exhibit characteristics of reading disabilities that are heterogeneous [10]. Dyslexia affects about 10-15% of the school-age population [9]. Studies in Malaysia on dyslexia have reported similar findings [9].

The other two reading disabilities namely, hyperlexia (adequate decoding but inadequate comprehension skills) and language learning disabilities (difficulties in both decoding and comprehension), affect 15% and 36% of poor readers, respectively. Shaywitz [11] reported that the prevalence of dyslexia hovered around 5% to 17% among children. In her earlier study in 1992, Shaywitz and colleagues reported that 28% and 17% of children diagnosed in Primary 1 would still be classified as having dyslexia in Grade 3 and Grade 6, respectively [12].

Students with learning disabilities lose their academic goals [13]. Therefore, the learning gap between the children with learning disabilities and those without increases with time [14]. Ultimately, these children with learning disabilities face problems in reading possess reduced motivation, experience obstacles in continuing education, and struggle with limited employment opportunities [15]. Therefore, it is important that parents and teachers are given the proper tools to diagnose the symptoms of dyslexia in young children [16].

In line with these concerns and the nation's concern about its competitive edge in the international arena, the Malaysian government recently introduced the preliminary report of the Malaysian Education Blueprint [17] describing the government's goals for improving the nation's education system. One area that requires much focus to meet such goals is the availability and accessibility of technological tools that provide the independence to both parents and teachers in providing the necessary diagnostic assistance to their children/students. This study is a step towards the achievement of the blueprint's goals. The ultimate goal of this study is to provide an easy way to teachers, researchers and developers who want to provide early diagnosis and intervention to those students who lag behind their peers developmentally and instructionally [18].

Mass identification of students with learning disabilities could be a challenge for the government to help the nation's education system and helping those children with dyslexia to improve their abilities in leaning. Data processing for the results collected manually from different schools involved huge effort of analysis from the human expert. Besides, with the limited number of human experts in Malaysia, it could be challenged to dedicate their knowledge for the diagnosis process. Machine learning could be an effective way to imitate and duplicate the knowledge of human expert. Thus, providing a reliable platform for mass diagnosis and classification of the dyslexia disorder.

Machine Learning algorithms are best suited for such diagnostic and classification system and can solve predictive problems [19]. Machine learning approach is very suitable for a situation where the amount of data is large and having complicated structure of interrelated attributes [20]. Prediction and analysis of such data is challenging because of analyzing individual effect of interrelated attributes. Such problems can be solved by machine learning algorithms [21, 22]. The goal of classification is to predict the class from the input variables. The classification algorithms analyze the input data and generate an output. As a supervised task, classification needs predefined target for each sample. The goal is that the classification algorithm will learn enough to match the prediction of a new sample with the targets as much as possible. Therefore, a successful learner should be able to progress from individual examples to broader generalization [23].

A classification algorithm adjusts its parameters by learning from the training data. This process is called training of the system. There are a bunch of classification algorithms like kNN or k Nearest Neighbors, decision tree, and naive Bayes etc. The researchers have tested several classification algorithms. Although kNN is an old algorithm but it was selected based on its batter accuracy for this particular data. After proper training, the system builds a knowledge-base that can accurately label unseen data [22]. This process is also known as supervised learning [24] [25]. Classification systems are widely used in other domains. For example, doctors are keen to conduct better diagnosis; businessmen need decision-making system to decide; and banks and other credit institutions need to predict finance information of companies and individuals [26-28].

There are three categories of machine learning techniques for classification: statistical pattern recognition [29], machine learning techniques for induction of decision trees or production rules [30, 31] and connectionist [32]. All these categories can be applied to the same problem. Only the generalization procedure is different to best perform in noisy data. Fisher & McKusick used back propagation neural nets for classification [22]. They found that back propagation neural nets perform well in noisy data but they are slow.

3. System Design and Development

During the system analysis and design phase, a typical function-oriented design was adopted. In function-oriented design the system is comprised of many smaller sub-systems known as function. These functions are capable of performing significant tasks in the system. Python programming language was selected as a tool for development. Because Python has built in machine learning algorithms. Python is also open source and cross platform which can run on different operating systems. We used scikit-learn, which is a powerful open source machine learning libraries for Python. Scikit-learn provide algorithms for machine learning tasks including classification, regression, dimensionality reduction, and clustering. It also provides modules for extracting features, processing data, and evaluating models. As our data need to be classified into two distinguished classes, "no dyslexia" and "seems to be dyslexic in spelling and reading". Therefore, we selected K Nearest Neighbors Classifier algorithm from scikit-learn. KNN classifier is best suitable for binary classification.

Similarly, the researchers used pandas and numpy libraries of Python programming language for fast data manipulation. We have used TKinter graphical user interface of python to design the user interface of the system.

The system provides the following three (3) main functionalities and also shown in Fig. 1 below.

3.1. Check for Dyslexia Category

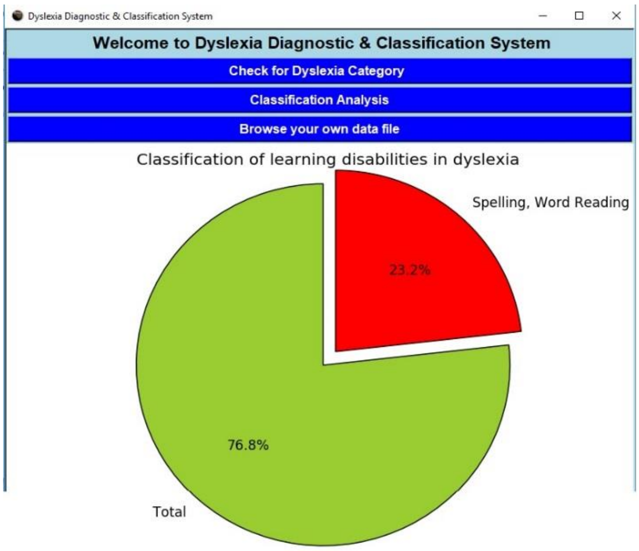
After training, the system builds a knowledge-base from data. This module can now predict new cases. So, professionals can test the accuracy of the system. Parents and teachers can predict learning disabilities in initial stages of the children. A test case is shown below in Fig. 3 below.

3.2. Classification Analysis

This module shows that 23.2% of the total data seems to be dyslexic in spelling and reading while 76.8% of data is non-dyslexic.

3.3. Module for Non-IT Background Researchers

This module is for non-IT background researchers. If they have data about dyslexia and can design and develop a diagnostic and classification system. So, they can easily browse their file to analyze their data.


Dyslexia Diagnostic & Classification System

Welcome to Dyslexia Diagnostic & Classification System

Check for Dyslexia Category
Classification Analysis
Browse your own data file
Classification of learning disabilities in dyslexia
Spelling, Word Reading
23.2%
76.8%
Total

Fig. 1: Interface of DCS with the three modules.

4. Methodology

5.1 Sample Data and Data Preprocessing

During this study, the system was developed based on data collected from 857 primary-1, school students. This sample data had been collected in a prior study using specially designed tests [33, 34]. Informed consent was obtained with the help of the school principal and the class teachers from all the participants' parents prior to the study. A list of 10 words based on words taught in Primary-1 was chosen for the Bahasa Malaysia test. These words were extracted from the Malaysian Primary-1 textbooks [35-37]. Machine learning algorithms need the data to be preprocessed so that the machine could learn from it. There is currently no benchmark in the Malay language which we could compare with. However, we used the widely used cut point of 25th percentile [15, 38]. Therefore, an expert in dyslexia (the second author) manually labeled the data into "no dyslexia" and "seems to be dyslexic in spelling and reading" groups.

5.2 Machine Learning Procedure

How machine can learn? Let's understand this phenomenon by an analogy of animals learning. When rats get a new food, they will first eat very small amounts, and subsequent feeding will depend on the flavor of the food and its physiological effect. If the food is poisonous they will associate this food with illness and will not eat in future. Clearly, learning takes place here. So, rats learn to avoid poisonous food [23].

As rats or human beings learn, similarly computer/machine can also learn how to recognize a kid with symptoms of dyslexia and then classify. The machine will memorize all such cases that had been labeled by human expert. When a new case comes in then the machine will search similar case in the already stored cases. If a match found then the machine will label it as a dyslexic otherwise non-dyslexic. The ability of machine to categorize an unseen case is learning. A successful learning system should be able to progress from individual examples to broader generalization [23].

To train a system for learning needs that the data should be preprocessed

4.2.1 Null data is not acceptable for machine learning algorithms. So, we preprocessed our data and removed null values.

4.2.2 The data must be in a separate table where each column represents an attribute. So, we compiled data from 857 primary school students into a table having attributes of spelling and reading.

4.2.3 There must be a separate one column table containing predefined labels assigned by human expert. So, an expert in dyslexia (the second author) manually labeled our data using twenty-fifth percentile. All the kids having scores of twenty-fifth percentile and below where labeled as suspicious for dyslexia in spelling and reading. The kids having scores above twenty-fifth percentile where as non-dyslexic.

5. Results

5.1 Experimental Evaluation

To validate the system, we used the standard way of splitting the data into training data set (70%) comprising 599 respondents and test data set (30%) comprising 257 respondents. For this we used using train_test_split function of scikit-learn. The training set contains a predefined known output and the system was trained by this trading data set. As mentioned in the section 2.1 in detail that after training, the system learns and builds knowledge of these known examples. Then successfully trained system can broadly generalize the result. The testing set is used to validate the system prediction on this subset. We used the test dataset to validate our system prediction. We calculated the accuracy of the system using scikit-learn matrics.accuracy_score function. The results show that 23% of children were at risk for dyslexia in the training data and 20.7% in the testing data with 99% of accuracy.

To further validate the system, we calculated the R^2 score of the system. R^2 is a "number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s)"[23]. Basically, this score shows the prediction accuracy of the system. This score shows that our system is 99% accurate in prediction and classification.

We also benchmark our results with a human expert in dyslexia. We selected 30 participants and evaluated manually by an expert in dyslexia (second author). Then we evaluated the same 30 respondents scores by the system, whose evaluation was 100% same to the human expert.

6. Discussion

Artificial intelligence and machine learning is matured enough to help human being in different fields of life. Soon in future these technologies will become more humanistic and will improve human being's capabilities.

Given that the machine generated a prevalence of 23.2% of the large sample to be at risk for dyslexia, this seemingly inflated prevalence of 23.2% by the machine diagnostic is not surprising given that the children who participated in this study were on average about 7 years old (i.e., Grade 1 equivalent). This age group and its corresponding prevalence is similar to the age group and prevalence reported in earlier studies. Shaywitz and colleagues [12] reported that 28% of children identified as having dyslexia in 1st grade would be classified as having dyslexia in 3rd grade. In addition, Shaywitz and colleagues [12,11] concluded that the diagnosis of dyslexia seems to be unstable of time and varies across the developmental spans of the children. There is one caveat with use of these cut points, however. The classification is only meant to identify children who are at risk of dyslexia, of which the prevalence data should then be used to provide the necessary early intervention/remediation, which hopefully will reduce the prevalence of dyslexia among school children. Thus, a future area of study that deserves focus is the longitudinal and trajectory of classification of children who are at risk of dyslexia across their developmental lifespan. Whether the difference in orthography (i.e., Malay orthography versus English orthography) would contribute to a higher or lower prevalence also deserves further investigation.

7. Conclusion

In this research, the researchers have successfully design and developed machine learning based diagnostic and classification system for kids with learning disabilities. The system's diagnosis and classification is validated and confirmed by a human expert. It is user friendly and easy to use system for researchers, trained users and parents to timely diagnose the symptoms of dyslexia.

Acknowledgement

This research was supported by Research and Innovation Management Center (RIMC) Universiti Malaysia Sarawak (UNIMAS) under Short Grant Scheme, Grant No. F04/SGS/1453/2016(05).

References

[1] International Dyslexia Association. Definition of Dyslexia. 2018 16-08-2018]; Available from: https://dyslexiaida.org/definition-of-dyslexia/.

[2] Department of statistics. Department of statistics Malaysia official portal. 2016 [cited 2017 10-7-2017]; Available from: https://www.dosm.gov.my/v1/index.php?r=column/cone&menu_id=dDM2enNvM09oTGtQemZPVzRTWENmZz09.

[3] Dyslexia Association of Sarawak. Dyslexia Association of Sarawak. 2017 [cited 2017 11-7-2017]; Available from: http://www.dyslexia-swk.com/default.asp?pageid=173.

[4] Ong, P.H., et al., Dyslexia among undergraduates in Malaysian universities: a mixed-methods study of prevalence, academic performances, academic difficulties and coping strategies. International Journal of Diversity in Organization, 2009. 9(1): p. 43-55.

[5] Aziz, U.A. Bilingual approach to learning. 2008 [cited 2015 May 25]; Available from: http://www.malaysianbar.org.my/general_opinions/comments/ungku_a._aziz_bilingual_approach_to_learning.html.

[6] Snow, C.E., et al., Is literacy enough? Pathways to academic success for adolescents. 2007: Paul H Brookes Publishing.

[7] Grajo, E.M., Lenin Mobile Apps Make Reading Fun for Children with Dyslexia. 2012 [cited 2015 20/5/2015]; Available from: http://www.newswise.com/articles/mobile-apps-make-reading-fun-for-children-with-dyslexia-slu-occupational-therapist-says.

[8] Gomez, C., Dyslexia in Malaysia. International book of dyslexia: A guide to practice and resources, 2004: p. 158-163.

[9] Vellutino, F.R., et al., Specific reading disability (dyslexia): What have we learned in the past four decades? Journal of child psychology and psychiatry, 2004. 45(1): p. 2-40.

[10] Catts, H.W., T.P. Hogan, and M.E. Fey, Subgrouping poor readers on the basis of individual differences in reading-related abilities. Journal of Learning Disabilities, 2003. 36(2): p. 151-164.

[11] Shaywitz, S.E., Dyslexia. New England Journal of Medicine, 1998. 338(5): p. 307-312.

[12] Shaywitz, S.E., et al., Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability. New England Journal of Medicine, 1992. 326(3): p. 145-150.

[13] Francis, D.J., et al., Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. Journal of Educational psychology, 1996. 88(1): p. 3.

[14] Stanovich, K.E., Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. Reading research quarterly, 1986: p. 360-407.

[15] Fletcher, J.M., et al., Classification of Learning Disabilities: An Evidence-Based Evaluation. Executive Summary. 2001.

[16] Griffin, P., M.S. Burns, and C.E. Snow, Preventing reading difficulties in young children. 1998: National Academies Press.

[17] Malaysian Ministry of Education, Malaysia Education Blueprint 2013-2025 2013-2025

[18] Foorman, B.R., et al., The role of instruction in learning to read: Preventing reading failure in at-risk children. Journal of Educational psychology, 1998. 90(1): p. 37.

[19] Rello, L. and M. Ballesteros. Detecting readers with dyslexia using machine learning with eye tracking measures. in Proceedings of the 12th Web for All Conference. 2015. ACM.

[20] Cui, Z., et al., Disrupted white matter connectivity underlying developmental dyslexia: A machine learning approach. Human brain mapping, 2016. 37(4): p. 1443-1458.

[21] Michalski, R. and R. Chilansky, Learning by being told and learning from examples. Int. J. of Policy Analysis and Information System, 1980.

[22] Fisher, D.H. and K.B. McKusick. An empirical comparison of ID3 and back-propagation. in IJCAI. 1989.

[23] Shalev-Shwartz, S. and S. Ben-David, Understanding machine learning: From theory to algorithms. 2014: Cambridge university press.

[24] Weiss, S.M. and I. Kapouleas, An empirical comparison of pattern recognition, neural nets and machine learning classification methods. Readings in machine learning, 1990: p. 177-183.

[25] Weiss, S.M. and C.A. Kulikowski, Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. 1991: Morgan Kaufmann Publishers Inc.

[26] Galindo, J. and P. Tamayo, Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Computational Economics, 2000. 15(1): p. 107-143.

[27] Kononenko, I., Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 2001. 23(1): p. 89-109.

[28] Aronson, J.E., T.-P. Liang, and E. Turban, Decision support systems and intelligent systems. 2005: Pearson Prentice-Hall.

[29] Duda, R. and P. Hart, Pattern classification and scene analysis. 1973.

[30] Quinlan, J.R., Induction of decision trees. Machine learning, 1986. 1(1): p. 81-106.

[31] Quinlan, J.R., C4. 5: programs for machine learning. 2014: Elsevier.

[32] McClelland, J.L. and D.E. Rumelhart, Explorations in parallel distributed processing: A handbook of models, programs, and exercises. 1989: MIT press.

[33] Otaiba, S.A., et al., Predicting kindergarteners' end-of-year spelling ability based on their reading, alphabetic, vocabulary, and phonological awareness skills, as well as prior literacy experiences. Learning Disability Quarterly, 2010. 33(3): p. 171-183.

[34] Lee, J.A.C. and S. Al Otaiba, End-of-Kindergarten Spelling Outcomes: How Can Spelling Error Analysis Data Inform Beginning Reading Instruction? Reading & Writing Quarterly: Overcoming Learning Difficulties, 2016.

[35] Abdul Malek, S.S., K. Yusuf, and A.H. Lin, Bahasa Malaysia tahun 1 sekolah kebangsaan: buku teks. 2010: Dewan Bahasa Pustaka.

[36] Abdul Malek, S.S., Bahasa Malaysia tahun 1 sekolah kebangsaan buku aktiviti jilid 1. Kurikulum Standard Sekolah Rendah, ed. Y. Khadijah and L. Hazlin. 2012, Kuala Lumpur: Dewan Bahasa Pustaka.

[37] Abdul Malek, S.S., Bahasa Malaysia tahun 1 sekolah kebangsaan buku aktiviti jilid 2. Kurikulum Standard Sekolah Rendah, ed. Y. Khadijah and L. Amir Hazlin. 2012, Kuala Lumpur: Dewan Bahasa Pustaka.

[38] Hasbrouck, J. and G.A. Tindal, Oral reading fluency norms: A valuable assessment tool for reading teachers. The Reading Teacher, 2006. 59(7): p. 636-644.