Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230722

A study on dyslexia detection using machine learning techniques for checklist, questionnaire and online game based datasets

S Santhiya and C S KanimozhiSelvi

Department of Artificial Intelligence, Kongu Engineering College, Perundurai, Tamil Nadu, India

santhiya123cse@gmail.com

Abstract. Learning disabilities are one of the most common developmental disorders in children. Learning is fundamental to a child's overall development. Children struggle with daily activities such as reading, speaking, organizing things, and so on. The specific learning disorders are classified into dyslexia, dysgraphia, and dyscalculia. Children who find difficulty in reading and are unable to differentiate speech sounds are said to have dyslexia. Dysgraphia and dyscalculia deal with written and mathematical calculations. Early diagnosis and detection are essential for early recovery from diseases. The proposed article presents methodologies and techniques used for detecting dyslexia. The primary contribution of this paper is a comparative analysis of various machine learning algorithms for diagnosing dyslexia, including SVM, KNN, Logistic Regression, K-mean Clustering, Oversampling, and Ensemble methods. Deep learning methods such as CNN and LeNet architecture have been used to identify dyslexia. The proposed study examines recent advances in detecting dyslexia using machine learning and deep learning approaches and identifies prospective research areas for the future.

Keywords: Learning Disability, Dyslexia, Machine Learning, Deep Learning.

1. Introduction

Education gives knowledge and good behavior. It improves one's ability to think, work, and act smartly. Children with learning disabilities might not learn as fast as normal children of the same age. They have difficulty with speaking, reading, writing, doing calculations, organizing things, and so on [1]. The specific learning disorders are classified into dyslexia, dysgraphia, and dyscalculia. Individuals with this problem may be unable to participate fully and competently in academic an activity, which leads to poor academic achievement. Even though the understanding of learning disabilities has evolved in recent years, diagnosing and assessing the severity of the disorder remains difficult. Teachers are important in evaluating children and advising parents to take their children to the doctor. A medical practitioner may have difficulty in diagnosing learning disability problems because they differ from one child to another.

Machine learning and deep learning methods are used to detect dyslexia. Machine learning learns from examples and progressively improves prediction accuracy and decision-making with experience over time. Existing machine learning techniques used for predicting learning disabilities are SVM

DOI: 10.54254/2755-2721/5/20230722

[2][3], Logistic Regression[4][5], Naive Bayes, K-NN[6], Random Forest, and decision tree[7]. The goal of the review paper is to examine current advances in dyslexia identification and future research opportunities using machine learning and deep learning methodologies.

2. Literature Review

Psychologists perform standardized tests like reading and writing, memory tasks, and phonological awareness in the conventional identification method of dyslexia. Test scores are evaluated, and the scores help to identify whether the person is affected by dyslexia or not. Performances are evaluated based on scores. Poor-scoring children are identified as being dyslexic. The conventional method necessitates the presence of psychologists. These methods require more time and are expensive.

Machine learning is becoming increasingly popular for medical diagnosis and decision-making in the medical field. Data collection is the first and most important step in machine learning. The machine learning steps are depicted in Fig. 1. The quality and quantity of data collection determine the output efficiency. The authors used different tests to collect data. specially designed tests, checklists, questionnaires, online games, reading tests for eye tracking, MRI scans, EEG signals, images, and video, and web-based independent game tests for auditory and visual checking. Some tests require materials like customized tools and cameras, eye trackers, MRI scanners, EEG headsets, and corneal reflection systems.

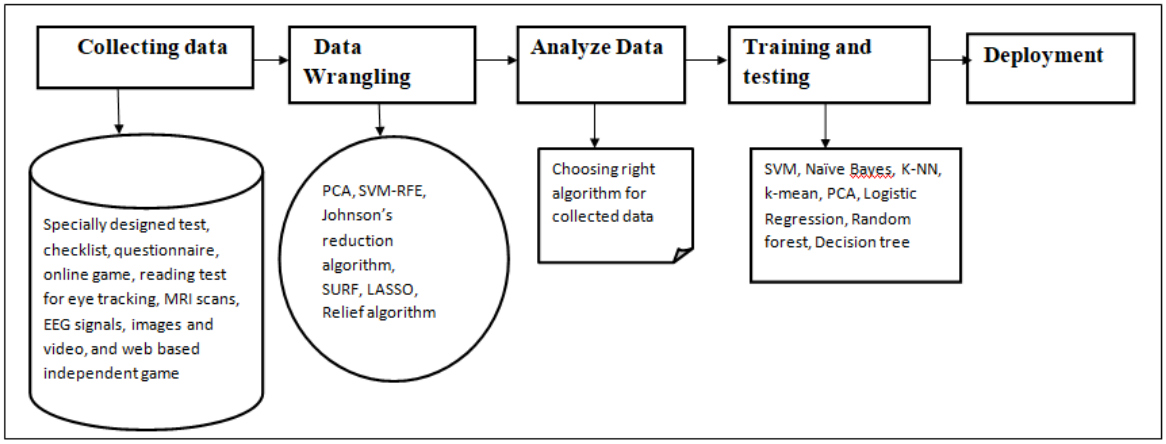

Figure 1. DNA Schematic.

Machine learning algorithms are used to identify dyslexia. The researchers used various methods such as supervised, unsupervised, and ensemble approaches, and the convolutional neural network to detect dyslexia is described below. Table 1. to Table 3. shows the comparison of different algorithms used by researchers for detecting and improving the accuracy of dyslexia.

DOI: 10.54254/2755-2721/5/20230722

Table 1. Comparison of algorithms and accuracy score for the dataset collected through checklist.

Year, Author Name	Number of attributes	Test type	Data set	Techniques	Machine Learning Algorithm	Accuracy score
[7], 2010, Julie M.David & Kannan Balakrishnan	16	Checklist	125 real data sets	Decision Trees and Clustering	J48 Algorithm with K-mean clustering	77.60%
[8], 2010, Julie M.David & Kannan Balakrishnan	16	Checklist	513 real data sets	Rough Sets	Naïve Bayes algorithm with Johnson's reduction algorithm	93.37%
[9], 2010, Julie M.David & Kannan Balakrishnan	16	Checklist	1100 real data sets	SVM	Sequential minimal optimization algorithm(SMO) in SVM	97.86%
[10], 2011, Julie M.David & Kannan Balakrishnan	16	Checklist	513 real data sets	Comparison of Rough set with SVM	Naïve Bayes batch classifier with Johnson's reduction algorithm	
[11], 2013, Julie M. David, Kannan Balakrishnan	16	Checklist	1020 real data set	Comparison of existing classifier and modified classifier with data preprocessing	Existing classifiers- ANN, J48, SVM, Naive Bayes. Modified Classifier- New ANFIS, New ANN, New Fuzzy	New ANN-99.03%, New Fuzzy -99.42%, New ANFIS-100%

Learning difficulties are predicted using different rules drawn from the decision tree of the J48 algorithm. K-mean clustering identifies the various indications and symptoms found in a child with LD [7]. Attributes in rough sets are reduced and classified using Johnson's reduction algorithm and the Naive Bayes algorithm [8]. The author compares decision trees with rough set theory for detecting LD. It is found that the rough set performs better in terms of accuracy and categorization. SVM [9] is performed using the sequential minimal optimization technique, and decision trees are constructed using the J48 algorithm. The Naive Bayes Batch classifier [10] is used for rough set classification, and the results are compared to those from the SMO algorithm in the SVM study. The resulting SVM is very complex compared to the rough set method. The new ANFIS method [11] achieves the highest accuracy and a comparison is depicted in Fig.2

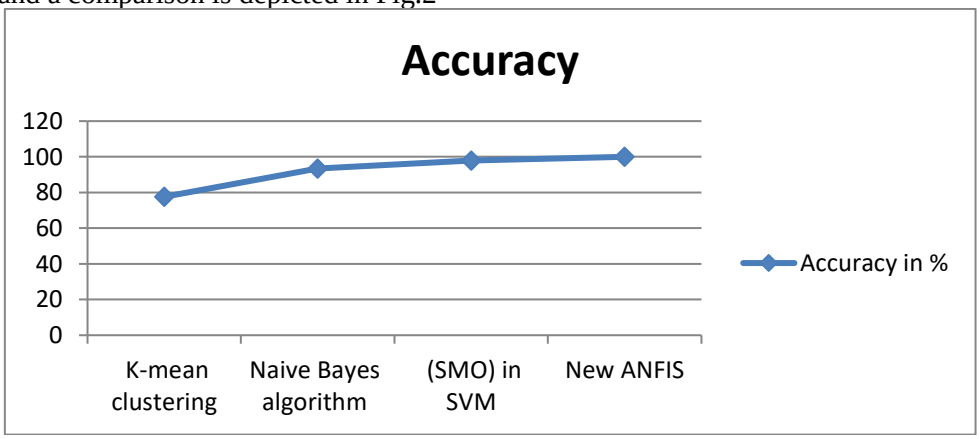

Figure 2. Performance of ML classifier for checklist dataset.

DOI: 10.54254/2755-2721/5/20230722

Table 2. Comparison of algorithms and accuracy score for the dataset collected through the Questionnaire gamet.

Year, Author Name	Number of attributes	Test type	Dataset	Techniques	Machine Learning Algorithm	Accuracy score
[12], 2016, K.Ambili & P.Afsar	16	Questionnaire	1124 instances	Artificial Neural Network		Accuracy, Learning Time, Error Rate
[13], 2016, K.Ambili & P.Afsar	16	Questionnaire	1124 instances	Naive Bayes Algorithm and ANN	fusion of Naive Bayes and Neural network classifier	Accuracy, Learning Time, Error Rate

The Naive Bayes algorithm outperforms the back propagation neural network in terms of accuracy. Both algorithm's outputs are combined and assigned random weights. It identified the combination that produces the highest level of accuracy. The Naive Bayes-Neural Network Fusion Technique performs better than each algorithm used alone [13].

Table 3. Comparison of algorithms and accuracy score for the dataset collected through the online-based game**.**

Year, Author Name	Number of attributes	Test type	Languages	Techniques	Machine Learning Algorithm	Accuracy score
[14], 2016, Rello L and Williams K, et al	6	Online web-based Game	English and Spanish	Support Vector Machine		85.85%
[15], 2020, Rello L, Baeza-Yates R et al	196	online gamified test	Spanish	Random Forests	Standard information the gain in decision trees, 10-fold cross-validation.	76.80%
[16], 2022, Shahriar Kaisar & Abdullahi Chowdhury	196	online gamified test	Spanish	Oversampling technique and Ensemble classifier	AdaBoost, Gradient Boost and XGBoost	88.3%, 89.6%, and 90%.
[17], 2020, Rauschenberger M & Baeza-Yates et al	ALL-33 Features,ES-41 Features, DE-38Features	web based game	German-DE, Spain-ES, ALL language	Random Forests and Extra Trees	Random Forest (RF), Extra Trees (ETC), Gradient Boosting (GB)	RF-74%German, ETC-69%spanish,GB-61%

A Wilcoxon Signed-Rank test evaluates the variation between groups for nonparametric data. The SVM model detects dyslexia based on the data collected through web-based games [14]. Random forest [15] was applied for the online gamified test, which produced the highest accuracy compared to the web-based game test [17] using random forest. Imbalanced data is obtained during pre-screening tests. The author suggests using oversampling and ensemble methods for identifying dyslexia. Adaptive boosting, gradient boosting, and extreme gradient boosting were used as ensemble models. The ensemble technique ADASYN with XGB achieved an accuracy of 90% [16] compared to other algorithms in [14], [15], and [17] for the data collected through an online game [18][19][20]. The comparison is depicted in Fig. 3.

DOI: 10.54254/2755-2721/5/20230722

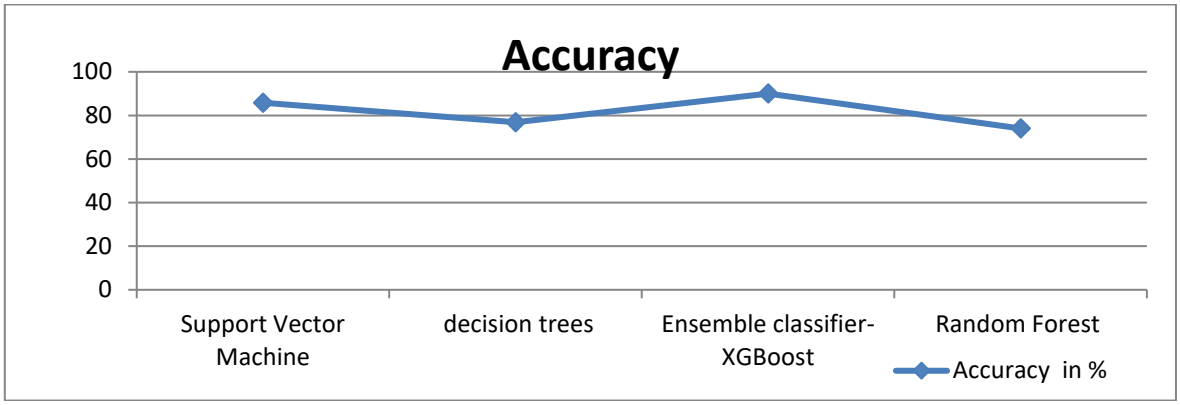

Figure 3. Performance of ML classifier for Online Gamified dataset.

3. Conclusion

The authors used various techniques for detecting dyslexia. The performances of various methods in ML are compared for further improvement in the future. The ensemble method achieved 90% accuracy for data collected through an online game. Accuracy needs to be improved for the data collected through online games. Online gamified data collections are language-independent. Online gamified data collection does not require any customized tool, is less expensive, and can reach all users. The investigation aids us in providing a complete examination of each model, including methods, algorithms, datasets used, and various performance measures such as accuracy, specificity, sensitivity, etc.

A novel proposed methodology is to be adopted to increase the performance of the ML approaches. The design of interactive multimedia and machine learning-based mobile and computer-aided intelligent diagnostic and therapeutic applications shall be developed to help special educators in diagnosing, training, assessing, and monitoring those children. Detecting dyslexia alone will not help children recover from dyslexia. The therapy game application will be developed to help the children recover from the disorder.

References

[1] Shanmugavadivel, K., Sathishkumar, V. E., Kumar, M. S., Maheshwari, V., Prabhu, J., & Allayear, S. M. (2022). Investigation of Applying Machine Learning and Hyperparameter Tuned Deep Learning Approaches for Arrhythmia Detection in ECG Images. Computational

[2] Pavithra, E., Janakiramaiah, B., Narasimha Prasad, L. V., Deepa, D., Jayapandian, N., & Sathishkumar, V. E. (2022). Visiting Indian Hospitals Before, During and After COVID. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.

[3] Kogilavani, S. V., Sathishkumar, V. E., & Subramanian, M. (2022, May). AI Powered COVID-19 Detection System using Non-Contact Sensing Technology and Deep Learning Techniques. In 2022 18th International Conference on Distributed Computing in Sensor Systems (DCOSS) (pp. 400-403). IEEE.

[4] Subramanian, M., Sathishkumar, V. E., Ramya, C., Kogilavani, S. V., & Deepti, R. (2022, May). A Lightweight Depthwise Separable Convolution Neural Network for Screening Covid-19 Infection from Chest CT and X-ray Images. In 2022 18th International Conference on Distributed Computing in Sensor Systems (DCOSS) (pp. 410-413). IEEE.

[5] Subramanian, M., Lv, N. P., & VE, S. (2022). Hyperparameter optimization for transfer learning of VGG16 for disease identification in corn leaves using Bayesian optimization. Big Data, 10(3), 215-229.

[6] Sathishkumar, V. E., & Cho, Y. (2019, December). Cardiovascular disease analysis and risk assessment using correlation based intelligent system. In Basic & clinical pharmacology & toxicology (Vol. 125, pp. 61-61). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA:

DOI: 10.54254/2755-2721/5/20230722

WILEY.

[7] Balakrishnan J M D 2010 *Significance of classification techniques in prediction of learning disabilities* arXiv preprint arXiv:1011.0628

[8] David J M & Balakrishnan K 2010 *Machine learning approach for prediction of learning disabilities in school-age children* International Journal of Computer Applications, 9(11), pp 7-14

[9] Julie M D & Kannan B 2010 *Prediction of learning disabilities in school age children using decision tree.* In Recent Trends in Networks and Communications Springer, Berlin, Heidelberg, pp 533-542

[10] David J M & Balakrishnan K 2011 *Prediction of key symptoms of Learning Disabilities in school-age ChildrenUsing rough sets*. International Journal of Computer and Electrical Engineering, 3(1), p 163

[11] David J M & Balakrishnan K 2013 *Performance improvement of fuzzy and neuro fuzzy systems: prediction of learning disabilities in school-age children*. International Journal of Intelligent Systems and Applications, 5(12), p 34

[12] Ambili K & Afsar P 2016 *A framework for learning disability prediction in school children using artificial neural network* International Journal of Advanced Research in Science, Engineering and Technology, 3(6).

[13] Ambili K & Afsar P 2016 *A framework for learning disability prediction in school children using naïve Bayes-neural network fusion technique* J Inf Knowl Res Comput Eng, 4(01).

[14] Rello L Williams K Ali A White N C & Bigham J P 2016 *Dytective: towards detecting dyslexia across languages using an online game*. In Proceedings of the 13th International Web for All Conference pp. 1-4

[15] Rello L Baeza-Yates R Ali A Bigham J P & Serra M 2020. *Predicting risk of dyslexia with an online gamified test*. Plos one, 15(12), e0241687.

[16] Kaisar S & Chowdhury A 2022. *Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests*. ICT Express.

[17] Rauschenberger M Baeza-Yates R & Rello L 2020 *Screening risk of dyslexia through a web-game using language-independent content and machine learning*. In Proceedings of the 17th International Web for All Conference pp. 1-12.

[18] Sathishkumar, V. E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications, 153, 353-366.

[19] VE, S., & Cho, Y. (2020). A rule-based model for Seoul Bike sharing demand prediction using weather data. European Journal of Remote Sensing, 53(sup1), 166-183.

[20] Krishnamoorthy, N., Prasad, L. N., Kumar, C. P., Subedi, B., Abraha, H. B., & Sathishkumar, V. E. (2021). Rice leaf diseases prediction using deep neural networks with transfer learning. Environmental Research, 198, 111275.