DETECTION OF DYSLEXIA AND DYSCALCULIA IN CHILDREN

Jawahar Jonathan, Bharathi K, Haripriya Khajuria, Lavanya J C, M Jyoshitha, Meghna Keshri

.

Dept. of Computer Science & Engineering,Acharya Institute of Technology,Bengaluru, India.

ABSTRACT

Learning disabilities represent a prevalent category of developmental challenges seen in children. The process of learning is pivotal for a child's holistic growth. Children encounter hurdles in their daily tasks such as reading, verbal expression, organizational skills, and similar activities. These specific learning impediments are categorized into dyslexia, dysgraphia, and dyscalculia. Dyslexia refers to difficulties in reading and distinguishing speech sounds. Dysgraphia and dyscalculia pertain to challenges in written expression and mathematical computations, respectively. Timely identification and assessment are crucial for early intervention and treatment. The forthcoming article outlines methodologies and approaches employed in identifying dyslexia. Its key contribution lies in a comparative examination of diverse machine learning algorithms utilized for dyslexia diagnosis, encompassing SVM, KNN, and random forest.

KEYWORDS

Dyslexia, Learning Disorder, Machine learning (ML), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest.

1. INTRODUCTION

Dyslexia stands out as a complex neurological condition garnering considerable attention from modern neuroscience researchers. According to the International Dyslexia Association, dyslexia is characterized by challenges in spelling, language processing, and accurate word recognition.

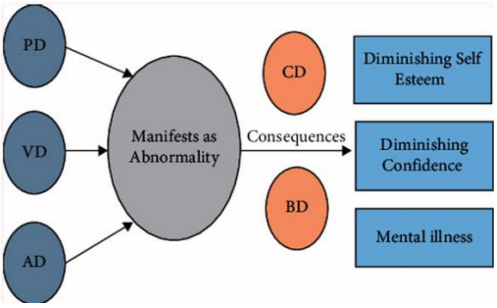

Figure 1: Various actors and consequences in the dyslexia. timeline.

Figure 1 provides an overview of the dyslexia paradigm, highlighting its key components: phonological disorder (PD), visual disorder (VD), and auditory disorder (AD). These issues typically emerge early in life and progress into noticeable abnormalities, significantly influencing

an individual's personality. The repercussions encompass various behavioral deficits (BD) and cognitive deficits (CD). Contrary to common misconceptions, dyslexia does not involve seeing letters or words in reverse. Instead, it stems from difficulties in phonological processing, impeding language manipulation.

For instance, individuals with dyslexia may struggle to identify words when asked to remove specific sounds, leading to delays in reading comprehension. Dyslexia affects a significant portion of the population, ranging from 5% to 17% across different languages. Its onset usually occurs during childhood, progressively impacting academic performance and eroding self-esteem. Meanwhile, advancements in healthcare technologies, particularly Computer-Aided Detection and Diagnostic (CAD) systems, have revolutionized healthcare provisioning.

The proposed framework for dyslexia detection offers several advantages, relying on online tests to quantify behavioral and cognitive deficits without requiring imaging data. Leveraging machine learning techniques like principal component analysis, SVM, KNN, and Random Forest, this framework demonstrates efficacy in dyslexia detection [1].

2. LITERATURE REVIEW

The papers presented offer diverse approaches to detecting and predicting dyslexia through machine learning. They employ various algorithms such as PCA, ANN, KNN, SVM, Random Forest, Naïve Bayes, and LSTM, showcasing a rich landscape of methodologies [2]. Achieving high accuracy rates up to 95%, these models exhibit promise in aiding dyslexia diagnosis and support for affected individuals [4]. However, challenges persist, including the need for more user-friendly interfaces, improvements in dataset accuracy, and addressing inefficiencies in prediction accuracy [1]. Proposed solutions suggest leveraging advanced techniques like CNNs, refining data collection methods, and enhancing algorithmic processing to overcome these hurdles. Each study contributes valuable insights and methodologies, paving the way for more effective and accessible tools in dyslexia detection and support [9]. Through ongoing research and refinement, these machine learning-based approaches hold potential to significantly impact the lives of dyslexic individuals by providing timely interventions and personalized strategies for academic success.

Table 1. Gap Analysis

Title	Findings from Literature	Gaps Identified
An Efficient Machine Learning-Based Feature Optimization Model for the Detection of Dyslexia	PCA and ANN Algorithms on Machine Learning.	The accuracy of the system is 95% which can be further improved by using advanced ML Techniques like CNN.
A Machine Learning-based Predictor to Support University Students with Dyslexia with Personalized Tools and Strategies	Algorithm used here are KNN, SVM and Random Forest.	Though the system is efficient it does not provide a user friendly Interface and is complex to implement.
A study on dyslexia detection using machinelearning techniques for checklist, questionnaire and online game-based datasets	This system uses a combination of Naïve Bayes batch classifier with Johnson's reduction algorithm	The therapy game application used cannot provide accurate data-sets on children with SLD.
Dyslexia Prediction Using Machine Learning algorithms -A Review	The algorithms used are KNN and Random Forest.	Inefficient-Accuracy Rate.

Dyslexia and Dysgraphia prediction: A new machine learning approach	It uses LSTM(Long-Short term memory neural networks).	It often breaks downs the sentences into smaller parts to an extent that it becomes difficult to read and understand even by the machine.

3. PROPOSED METHODOLOGY

The database underwent preprocessing to ensure compatibility with machine learning algorithms for training and testing [7]. Initially, scores were assigned to potential answers across three sets of questions regarding encountered difficulties, support tools, and support strategies. This scoring process followed equivalences detailed in Table 1. For answers denoted as "rarely" within the support tools questions, exclusion from analysis occurred if over 15% of participants selected this option for any question. Otherwise, scores were determined based on the mode of responses, aiming to maintain sample size consistency.

Table 2. Corresponding Binary Scores for Input Options

Answer	Equivalent Binary Score
Rarely	1
Sometimes	2
Frequently	3
Mostly always	4

The subjective nature of participant responses introduces variability, susceptible to personal interpretation. To mitigate this variability, scores for tools and strategies were threshold to yield binary outcomes: useful or useless. The prediction task focused solely on assessing the usefulness of support methodologies, rather than estimating their level of utility. Threshold values were individually selected for each methodology, with ML algorithms trained using various thresholds to maximize prediction accuracy [5]. Tested thresholds included 1, 2, 3, and 4.

Thresholding also extended to difficulty questions but only as one training option. Algorithms were trained using both scores from Table 1 and binary outcomes derived from threshold application. This choice was justified by the lower variability of difficulty scores and the potential for improved algorithm performance with numeric data.

For Random Forest setups, the bootstrap technique was employed, randomly selecting one-third of variables at each decision split across 50 decision trees. Input variables were treated as scores, numeric values, or binary values obtained through thresholding. The same thresholds used for output variable binarization were applied.

In k-Nearest Neighbors setups, input variables were treated as numeric or binary values using the same thresholding method. Euclidean and Hamming distance metrics were utilized depending on variable representation.

Support Vector Machines were trained with three different kernels: linear, polynomial (degree 2), and radial basis function (RBF). Input variables were handled as numeric or binary values, following the thresholding method.

Algorithm performance was assessed by calculating the overall prediction accuracy (A) using a cross-validation approach with 10 folds. The formula for A involved summing correct predictions and dividing by the total number of tests performed in each fold.

$$A = \frac{1}{Nf} \times \sum_{f=1}^{N_f} \frac{N_{cf}}{N_T}, \qquad (1)$$

where N_f is the number of folds used for cross-validation (namely 10), N_{cf} is the number of correct predictions and N_T is the total number of tests performed in each fold.

4. DESIGN

4.1. Input Design

Input design is the process of converting user-originated inputs to a computer understandable format. Input design is one of the most expensive phases of the operation of computerized system and is often the major problem of a system. A large number of problems with a system can usually be tracked backs to fault input design and method. Every moment of input design should be analyzed and designed with utmost care. The system takes input from the users, processes it and produces an output. Input design is link that ties the information system into the world of its users. The system should be user-friendly to gain appropriate information to the user. The decisions made during the input design are:

- To provide cost effective method of input.
- To achieve the highest possible level of accuracy.
- To ensure that the input is understand by the user.

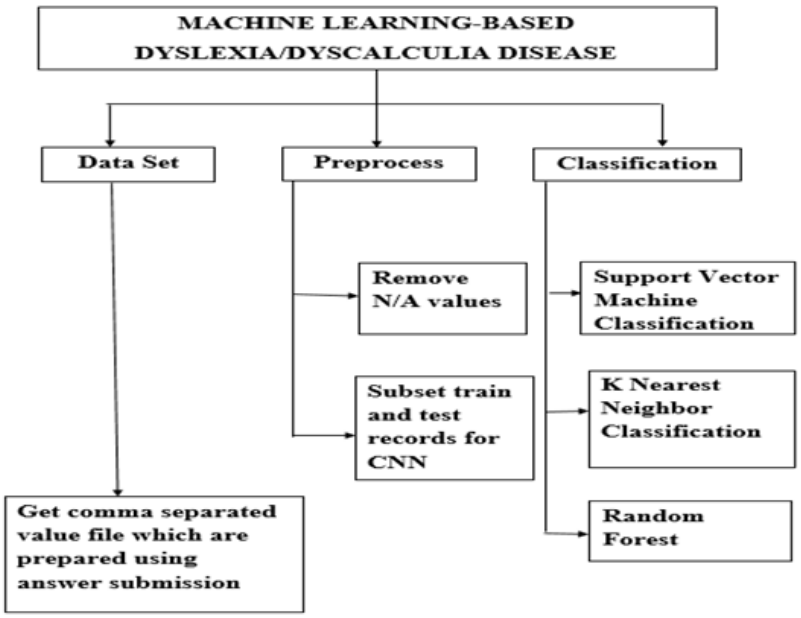

Figure 2: Block representation of system architecture.

System analysis decide the following input design details like, what data to input, what medium to use, how the data should be arranged or coded, data items and transactions needing validations to

detect errors and at last the dialogue to guide user in providing input. Input data of a system may not be necessarily is raw data captured in the system from scratch. These can also be the output of another system or subsystem. The design of input covers all the phases of input from the creation of initial data to actual entering of the data to the system for processing [3]. The design of inputs involves identifying the data needed, specifying the characteristics of each data item, capturing and preparing data from computer processing and ensuring correctness of data. Any Ambiguity in input leads to a total fault in output. The goal of designing the input data is to make data entry as easy and error free as possible.

4.2. Output Design

Output design generally refers to the results and information that are generated by the system for many end-users; output is the main reason for developing the system and the basis on which they evaluate the usefulness of the application.

The output is designed in such a way that it is attractive, convenient and informative. Codings are made with various features, which make the console output more pleasing.

As the outputs are the most important sources of information to the users, better design should improve the system's relationships with user and also will help in decision-making. Form design elaborates the way output is presented and the layout available for capturing information.

5. EXPERIMENTAL RESULTS

Table 3. Comparison of algorithms and their accuracy

ALGORITHMS USED	ACCURACY SCORE
SVM	97.86%
KNN	98.63%
RANDOM FOREST	98.63%

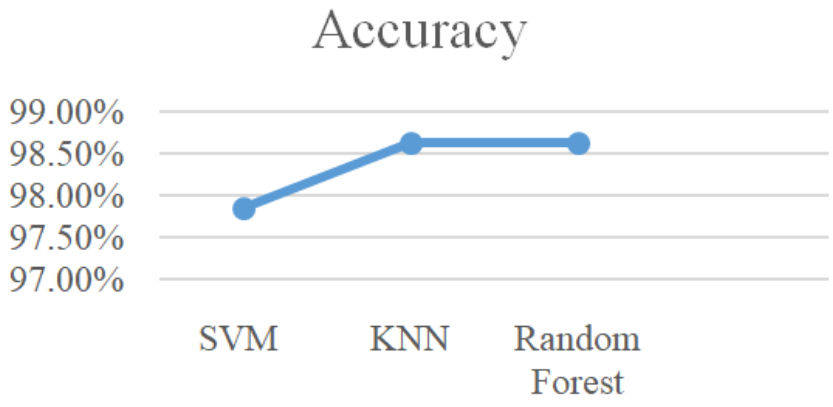

Figure 3: Graphical representation of accuracy of algorithms used.

In the evaluation of different machine learning algorithms, Table 2, for dyslexia detection, SVM achieved an accuracy score of 97.86%, indicating its robust performance in classifying dyslexia and dyscalculia cases. Similarly, both KNN and Random Forest algorithms demonstrated high accuracy scores of 98.63%, (refer Figure 3) showcasing their effectiveness in accurately identifying dyslexia. These results underscore the reliability and efficacy of these machine

learning techniques in aiding dyslexia detection, offering promising avenues for further research and potential clinical applications [6][8].

5. CONCLUSION

Dyslexia affects approximately 10% of the global population, with varying degrees of impact on personality ranging from moderate to severe. In India, the prevalence rate is slightly lower, estimated at around 8%. Detecting this disorder early is crucial for effective intervention in most cases.Collaboration among researchers, clinicians, and experts from diverse fields offers promising prospects for addressing this challenge. Artificial intelligence and machine learning, alongside medical imaging technologies, provide hopeful avenues. This study introduces an online test-based approach for dyslexia detection, suitable considering the age and lifestyle of the subjects. Leveraging acquired data, a machine learning framework incorporating principal component analysis is established. Notably, Random Forest emerges as highly effective in dyslexia detection, alongside SVM and KNN. The study reports an impressive accuracy rate of 98.63% for Random Forest and KNN, indicating significant potential compared to existing methodologies. This research lays groundwork for the development of a dyslexia learning management system, offering promise for future advancements in this domain.

REFERENCES

[1] Shantakumar B. Patil and Y.S. Kumaraswamy, “Intelligent and Effective Heart Prediction System using Data Mining and Artificial Neural Network.

[2] Latha Parthiban and R. Subramanian, “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm”, Disease Prediction System using CANFIS and Genetic Algorithm”.

[3] Charly, K.: “Data Mining for the Enterprise”, 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.

[4] Wu., ER., Peters, W., Morgan, M.W.: “The Next Generation Clinical Decision Support: Linking Evidence to Best Practice”, Journal Healthcare Information Management. 16(4), 50-55, 2002.

[5] Dyslexia Prediction Using Machine Learning Algorithms-A Review- Kasthuri Stephen.

[6] Dyslexia and Dysgraphia prediction: A new machine learning approach- Gilles Richard and Mathieu Serrurier.

[7] Mary K.Obenshain, “Application of Data Mining Techniques to Healthcare Data”, Infection Control and Hospital Epidemiology, vol. 25, no.8, pp. 690–695, Aug. 2004.

[8] Dyslexia and Dysgraphia prediction: A new machine learning approach- Gilles Richard and Mathieu Serrurier.

[9] A study on dyslexia detection using machine learning techniques for checklist, questionnaire and online game based datasets- S Santhiya and CS KanimozhiSelvi.