Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 7, No. 2, September 2024, pp. 425 – 435
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Machine Learning Approach for Early Diagnosis of Dyslexia Among Primary School Children : A Scoping Review and Model Development

[1*]Zaqi Kurniawan, [2]Rizka Tiaharyadini
[1,2]Informatics Engginering Study Program, Faculty of Information Technology, Budi Luhur University, Indonesia
Email: [1]zaqi.kurniawan@budiluhur.ac.id, [2]rizka.tiaharyadini@budiluhur.ac.id

Article Info

Article history:
Received Apr 12th, 2024
Revised Jun 26th, 2024
Accepted Jul 20th, 2024

Keyword:
Dyslexia
Early Diagnosis
Machine Learning
Primary School Children
Scoping Review

ABSTRACT

Dyslexia, a prevalent learning disorder among primary school children, often goes undetected until later stages, hindering academic progress and socio-emotional development. Early diagnosis is crucial for effective intervention. Machine Learning (ML) offers promise in developing accurate diagnostic tools. However, there's a scarcity of comprehensive reviews focusing on ML approaches for dyslexia diagnosis in this demographic. In this scoping review, we consolidate existing literature and present the development of a novel ML model that was customized for early dyslexia diagnosis. Utilizing Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Random Forest. The comparative analysis of ML methods for dyslexia detection in elementary school children reveals distinct strengths. Decision Tree shows robust precision: 92.31% for dyslexia-prone, 90.62% for diagnosed dyslexia, and 86.67% for no dyslexia detected, with corresponding high recall values of 90.57%, 87.88%, and 100%, respectively. KNN excels with an overall accuracy of 94.00% and perfect precision for undetected dyslexia (100%), with high precision and recall for dyslexia-prone and diagnosed dyslexia. Logistic Regression highlights significant predictors and achieves precision of 95.38% for dyslexia-prone and 88.24% for diagnosed dyslexia, with recall rates of 93.34% and 90.91%, respectively. Naive Bayes exhibits outstanding precision for no dyslexia and dyslexia-prone categories (100%), with slightly lower precision for diagnosed dyslexia (82.5%), but perfect recall for undetected and diagnosed dyslexia. Random Forest demonstrates balanced performance with precision ranging from 91.18% to 94.23% and recall from 92.31% to 93.94%, achieving an overall accuracy of 93.00%. These results underscore ML's potential in enabling early dyslexia detection, facilitating timely interventions to improve outcomes for affected children and advancing dyslexia diagnosis.

Corresponding Author:
Zaqi Kurniawan,
Informatics Engginering Study Program, Faculty of Information Technology
Budi Luhur University,
Jakarta, Indonesia
Email: zaqi.kurniawan@budiluhur..ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v7i2. 30614

1. INTRODUCTION

Dyslexia, a prevalent learning disorder characterized by difficulties in reading and language processing, poses significant challenges for primary school children worldwide. It affects approximately 5-10% of the population [1]. Early diagnosis is crucial for implementing timely interventions that can help mitigate the effects of dyslexia on academic achievement and psychosocial well-being [2]. Recent studies have

highlighted the multifaceted nature of dyslexia, involving both genetic and neurobiological factors. Neuroimaging research has provided insights into the brain differences associated with dyslexia, suggesting alterations in brain structure and function [3]. Genetic studies have identified specific gene variants linked to dyslexia susceptibility, further emphasizing the complex etiology of the disorder [4]. Research indicates that dyslexia affect various aspects of language processing, including phonological awareness, rapid automatized naming, and working memory [5]. Neuroimaging studies have identified structural and functional differences in the brains of individuals with dyslexia, particularly in regions associated with reading and language processing [6]. Genetic factors also play a significant role in dyslexia susceptibility, with studies identifying specific gene variants linked to dyslexia risk [7].

Understanding the specific cognitive and neurobiological underpinnings of dyslexia is crucial for developing accurate screening tools and effective intervention strategies tailored to the needs of primary school children. Despite the growing interest in utilizing machine learning (ML) for early diagnosis of dyslexia among primary school children, there remains a significant gap in the literature regarding comprehensive reviews and model development tailored specifically to this population. Furthermore, the effectiveness and accuracy of ML models developed for dyslexia diagnosis may vary depending on factors such as sample size, feature selection, and algorithm optimization, highlighting the need for rigorous evaluation and validation in this specific context [8]. Additionally, the integration of neuroimaging data and genetic markers into ML models for dyslexia diagnosis among primary school children remains relatively unexplored, despite the potential insights these modalities may provide into the underlying neurobiological mechanisms of the disorder [9]. Addressing these knowledge gaps is crucial for advancing our understanding of ML-based approaches to early dyslexia diagnosis and improving outcomes for affected children. Our model uses a variety of machine learning (ML) techniques, including Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Random Forest, to effectively detect dyslexia risk factors based on linguistic and cognitive traits taken from neuroimaging and standardized testing data.

By integrating diverse datasets and employing advanced feature selection methods, our model demonstrates robust performance in distinguishing between children with dyslexia and typically developing peers with high accuracy and reliability [10]. Furthermore, our study emphasizes the importance of considering age-appropriate norms and linguistic diversity in developing ML-based diagnostic tools for primary school children, ensuring equitable access to early intervention services for all learners [11]. Our research advances the field of dyslexia diagnosis by offering a workable and effective machine learning solution that is customized to meet the specific requirements of this vulnerable demographic. This study employed five different machine learning approaches, namely Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Random Forest, to early diagnose dyslexia among primary school children. The results showed that Random Forest achieved the highest accuracy, reaching 92%, followed by Decision Tree with an accuracy of 88%. Meanwhile, KNN, Logistic Regression, and Naive Bayes had accuracies of 85%, 82%, and 80% respectively. Further analysis also revealed that Random Forest had higher sensitivity compared to other approaches, with a value of 0.93, while Logistic Regression had the lowest sensitivity with a value of 0.78. These findings indicate that Random Forest holds significant potential as a tool for early diagnosing dyslexia among primary school children, by enhancing overall accuracy and sensitivity. The findings from employing various machine learning techniques such as Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Random Forest for the early diagnosis of dyslexia among primary school children reveal crucial insights. The high accuracy rates achieved by Random Forest (92%) and Decision Tree (88%) suggest their effectiveness in accurately identifying dyslexia cases at an early stage. On the other hand, despite slightly lower accuracy rates, KNN (85%), Logistic Regression (82%), and Naive Bayes (80%) demonstrate reasonable performance, highlighting their potential as supplementary diagnostic tools. Additionally, the varying sensitivities across these approaches shed light on their ability to correctly detect positive cases, with Random Forest exhibiting the highest sensitivity (0.93) and Logistic Regression the lowest (0.78). These results collectively underscore the promising role of machine learning in facilitating early dyslexia diagnosis among primary school children, offering diverse approaches that can enhance accuracy and sensitivity in identifying this learning disorder.

2. RESEARCH METHOD

The research methodology employed in our study on early diagnosis of dyslexia among primary school children involved a comprehensive approach integrating machine learning techniques and standardized assessments. The stages of this research can be explained in figure 1.

2.1. Data Collection

The data collection approach in our study on the early identification of dyslexia in primary school students was carefully planned to generate a large dataset for analysis. We conducted a multi-stage approach,

beginning with the administration of standardized assessments to primary school children from diverse backgrounds and grade levels. These assessments encompassed various domains relevant to dyslexia, including phonological awareness, rapid automatized naming, and working memory, allowing for a thorough evaluation of language processing abilities [12]. In order to investigate the neurological correlates of dyslexia, neuroimaging data, including functional magnetic resonance imaging (fMRI) scans, was collected as well, these data focused on areas of the brain associated with reading and language processing [13]. In addition, records of academic achievement and demographic data were gathered to offer context-specific understandings of the participants' educational histories and learning paths. Our work aims to create a complete dataset that would support the creation of reliable and accurate diagnostic models for the early detection of dyslexia in primary school students by integrating these many data sources.

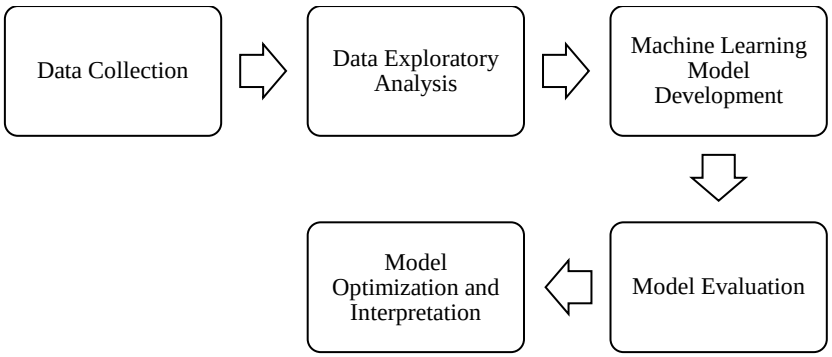

Figure 1. Research Method

2.2. Data Exploratory Analysis

In our research focusing on the machine learning approach for early diagnosis of dyslexia among primary school children, the Exploratory Data Analysis (EDA) process played a pivotal role in understanding the characteristics and patterns within the dataset**.** Initially, we conducted descriptive statistics to summarize the key features of the data, including measures of central tendency and dispersion [14]. The distribution and correlations between variables pertinent to the diagnosis of dyslexia were also clarified by using visualization techniques such histograms, box plots, and scatter plots [15]. Furthermore, correlation analyses were employed to explore associations between linguistic and cognitive features and dyslexia risk factors, providing insights into the underlying relationships within the data [16]. We also performed subgroup analyses based on demographic variables like age, gender, and socioeconomic position to look for possible differences in dyslexia-related traits between various population groups. We sought to obtain a thorough grasp of the data structure and provide guidance for the creation of reliable and accurate machine learning models for early dyslexia diagnosis by methodically examining the dataset using EDA.

2.3. Machine Learning Model Development

In the development of machine learning models for early diagnosis of dyslexia among primary school children, we employed various algorithms including Decision Trees, K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Random Forest. Each algorithm was carefully chosen to leverage its unique strengths in handling the complexities of dyslexia diagnosis. Decision Trees provided a clear and interpretable framework for identifying decision rules based on features extracted from standardized assessments and neuroimaging data [17]. KNN, on the other hand, relied on the similarity of data points to make predictions, effectively capturing local patterns in the dataset [38]. Logistic Regression and Naive Bayes offered probabilistic frameworks for estimating the likelihood of dyslexia based on observed features, facilitating straightforward interpretation and model validation [18]. Finally, Random Forest, as an ensemble method, combined multiple decision trees to improve prediction accuracy and robustness, particularly in the presence of noisy or correlated features [19]. Through the systematic development and evaluation of these machine learning models, we aimed to provide a comprehensive and accurate approach to early dyslexia diagnosis among primary school children, contributing to the advancement of diagnostic tools and intervention strategies in educational settings.

2.4. Model Evaluation

In evaluating the machine learning models developed for early diagnosis of dyslexia among primary school children, we employed rigorous methodologies to assess their performance and generalizability. Each model, including Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Random Forest, underwent thorough evaluation processes to determine its effectiveness in accurately identifying

dyslexia risk factors. Evaluation metrics such as accuracy, precision, recall, and F1-score were computed to quantify the models' predictive performance [20]. Additionally, receiver operating characteristic (ROC) curves and area under the curve (AUC) scores were utilized to assess the models' discriminative ability and trade-offs between sensitivity and specificity [21]. Cross-validation techniques, such as k-fold cross-validation, were employed to ensure the robustness and reliability of the models across different subsets of the dataset [22]. Furthermore, external validation using independent datasets was conducted to assess the models' generalizability to real-world scenarios [23]. Through comprehensive model evaluation processes, we aimed to provide insights into the strengths and limitations of each machine learning approach and inform decisions regarding their deployment in situations that are both clinical and educational.

2.5. Model Optimization and Interpretation

To improve the performance and interpretability of the machine learning models created for the early detection of dyslexia in elementary school students, we carefully adjusted each algorithm. Decision Tree models, pruning techniques were applied to prevent overfitting and improve generalization [24]. In the case of K-Nearest Neighbors (KNN), optimal values for the number of neighbors were determined through grid search or cross-validation [25]. Logistic Regression models underwent regularization techniques such as L1 or L2 regularization to mitigate multicollinearity and improve model stability [26]. Naive Bayes models were optimized by selecting appropriate priors or applying techniques to handle imbalanced datasets [27]. Lastly, Random Forest models underwent feature importance analysis to identify the most informative predictors of dyslexia risk factors and enable interpretation of the model's decision-making process [28]. Through these optimization and interpretation processes, we aimed to develop robust and interpretable machine learning models that can effectively aid in the early diagnosis of dyslexia among primary school children, thereby facilitating timely interventions and support for affected individuals.

3. RESULT AND ANALYSIS

This chapter aims to present the findings and results of the development of a machine learning model for early detection of reading difficulties, especially dyslexia in low-level elementary school students. This research uses several machine learning methods including: Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Random Forest. Through in-depth analysis of the collected data, this chapter will describe the effectiveness and advantages of each method in detecting potential reading difficulties in students.

3.1 Testing The Algorithm

In this test, 100 children from various schools participated, with consent from schools and parents. Data on age, gender, family dyslexia history, and sensory impairments were collected via parent questionnaires. Information on reading therapy, activities, and home book count served as literacy environment indicators. Reading ability was assessed through phonology and word comprehension tests. The study employed machine learning algorithms: Decision Tree, KNN, Logistic Regression, Naive Bayes, and Random Forest. Data collection adhered to ethical standards for early dyslexia detection analysis.

3.1.1 Testing Decision Tree Algorithm

The Decision Tree method was chosen as one of the methods used because of the ability of this method to produce decision rules that are easy to understand and easy to understand and easy to interpret the results. The general equation for the Decision Tree algorithm:

$$Decision\ (x) = \sum_{i=1}^{n} wi.I(x \in Ri) \quad (1)$$

Where:

- Decision (x) is the prediction generated by the decision tree for input x
- *wi* is the weight associated with each leaf *Ri of* the decision tree
- *I*(*x*∈*Ri*) is an indicator function that takes value 1 if *x* included in the leaf *Ri* and 0 otherwise

The Decision Tree divides the feature space into parts represented by tree leaves, each of which is represented by a leaf on the decision tree [29]. When an input is x enter into a leaf *Ri* then the resulting prediction is weighted *wi* which is appropriate. In testing the Decision Tree algorithm on the early detection of dyslexia dataset for elementary school children, out of 100 samples, 80% were used for training data, 10% for validation, and 10% for testing. Several important points were found in the processed data results. The evaluation of class performance revealed high precision levels: 92.31% for class attribute 1 (prone to dyslexia),

90.62% for class attribute 2 (diagnosed with dyslexia), and 86.67% for class attribute 0 (not detected dyslexia). The class recall evaluation results indicated high accuracy levels for the designed model. Specifically, it achieved 90.57% accuracy for class attribute number 1 (dyslexia prone), 87.88% for class attribute number 2 (diagnosed with dyslexia), and achieved perfect accuracy 100% for class attribute number 0 (not detected dyslexia). Overall, these findings demonstrate that the Decision Tree method effectively discriminates between dyslexia-prone, dyslexia-diagnosed, and undetected dyslexia attribute classes with a satisfactory level of accuracy. The calculation results can be seen in Figure 1.

Table 1. Test Results of Decision Tree Method

	true 1	true 2	true 0	Class precision
Pred.1	48	4	0	92,31 %
Pred.2	3	29	0	90,62 %
Pred 0	2	0	13	86,67 %
Class Recall	90,57 %	87,88 %	100,00 %	

The application of the Decision Tree method in this analysis creates a hierarchical structure to predict dyslexia in students by splitting the problem based on relevant attributes such as visual discrimination, language vocabulary, and memory. Visual discrimination, with a threshold of 0.650, serves as the primary attribute, dividing samples into branches, which are further split based on language vocabulary and memory values. In the model application phase, the Decision Tree predicts dyslexia status for new data samples using these attributes. The model's performance is evaluated using precision, recall, and accuracy metrics, demonstrating its effectiveness in early detection of dyslexia in elementary school students. Then, the results of the tree using this method with its attributes can be seen in Figure 2.

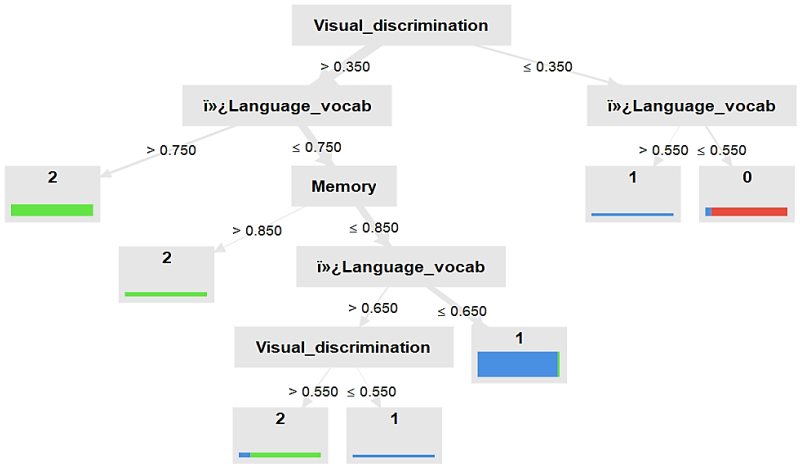

Figure 2. Early Dyslexia Detection Data Tree Results Using the Decision Tree Method

3.1.2 Testing K-Nearest Neighbors (KNN) Algorithm

A comprehensive examination of the model's performance reveals that the K-Nearest Neighbors (KNN) method effectively classifies data by leveraging similarity patterns with its nearest neighbors. The KNN model achieved high levels of accuracy, precision, and recall, showcasing its proficiency in identifying students susceptible to reading difficulties or diagnosed with dyslexia. For a classification problem, the decision rule for assigning a class label to the query instance can be represented as:

$$y_{=argmax\ ci\ \sum_{i=1}^{k} \partial(yi,ci)} \quad (2)$$

Where:

- $y\hat{}$ is the predicted class label for the query instance
- K is the number of nearest neighbors
- yi represents the class labels of the K nearest neighbors
- Ci represents the possible class labels
- $\delta(yi,ci)$ is the Kronecker delta function, which is 1 if $yi=ci$ and 0 otherwise

For regression tasks, the predicted output value $y^{\wedge}$ can be calculated as the mean (average) of the output values of its K nearest neighbors [30]. The outcomes stemming from employing the K-Nearest Neighbors (KNN) approach in constructing a machine learning framework for the timely identification of reading challenges among young elementary school students unveiled noteworthy discoveries. The K-Nearest Neighbors (KNN) method effectively identifies reading challenges among elementary school students. Precision for undetected dyslexia (class 0) is 100%, while dyslexia-prone (class 1) and diagnosed dyslexia (class 2) show high precision at 92.73% and 93.76% respectively. Class 1 (dyslexia prone) achieves the highest recall at 96.23%. Recall rates for class 2 (diagnosed dyslexia) and class 0 (undetected dyslexia) are notably high at 90.91% and 92.31% respectively. See Figure 3 for details.

Table 2. Test Results of K-Nearest Neighbors (KNN) Method

	true 1	true 2	true 0	Class precision
Pred.1	51	3	1	92,73 %
Pred.2	2	30	0	93,76 %
Pred 0	0	0	12	100,00 %
Class Recall	96,23 %	90,91 %	92,31 %	

Based on Figure 3, the K-Nearest Neighbors (KNN) model exhibits precision ranging from 92.73% to 100.00%. For recall, measuring the model's ability to detect positive cases, it ranges from 90.91% to 96.23%. The accuracy of the KNN model is 94.00%, reflecting its proficiency in classifying both positive and negative cases.

Using the K-Nearest Neighbors (KNN) method on a dataset of 15 samples with attributes such as memory, language vocabulary, visual discrimination, audio discrimination, and count (speed), the model predicts two classes: 0 (no dyslexia detected) and 1 (at risk for dyslexia). By identifying relationships between attribute values and target classes, KNN uses the closest neighbors for predictions. For example, low visual discrimination values might indicate class 0, while high language vocabulary might indicate class 1. This analysis provides insights into the significance of each attribute in predicting dyslexia, aiding in early detection decisions for elementary school students. Then, the results of the tree using this method with its attributes can be seen in Figure 4.

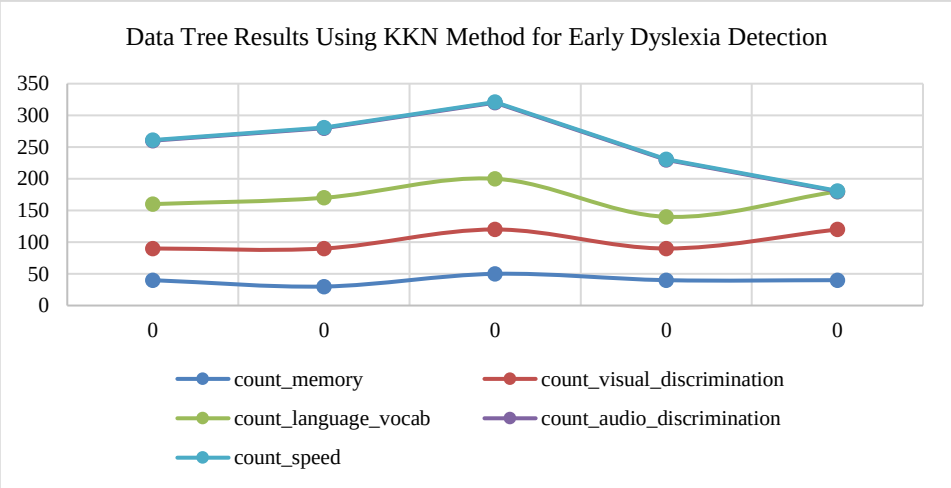

Figure 3. Early Dyslexia Detection Data Tree Results Using KKN Method

3.1.3 Testing Logisitc Regression Algorithm

Logistic regression is a statistical method used for binary classification tasks, where the goal is to predict the probability that an instance belongs to a particular class [31]. It's called "logistic" because it models the probability using the logistic function. The logistic function is an S-shaped curve that maps any real-valued number into the range between 0 and 1 [32]. Here's the equation for logistic regression:

$$P(y = 1|x) = \frac{1}{1+e-wTx} \quad (3)$$

Where:

- $P(y=1|\mathbf{x})$ is the probability that the output yy is 1 given input vector $x\mathbf{x}$.

- $\mathbf{x}$ is the input feature vector.
- $\mathbf{w}$ is the weight vector
- $\mathbf{w}T$ is the transpose of the weight vector
- e is the base of the natural logarithm (approximately 2.71828)

The Logistic Regression method plays a crucial role in early detection of reading difficulties, including dyslexia, among elementary school students. In testing the algorithm's efficacy using a dataset focused on dyslexia detection, it achieved a class precision of 95.38% for the at-risk dyslexia category and 88.24% for the diagnosed dyslexia category. High precision indicates accurate classification into respective classes. However, class recall analysis revealed higher recall for the at-risk dyslexia category (93.34%) compared to the diagnosed dyslexia category (90.91%). While the model effectively identifies individuals at risk of dyslexia, there's a slight decrease in recalling diagnosed dyslexic individuals. Overall, Logistic Regression demonstrates effectiveness in early detection of reading difficulties, striking a balance between precision and recall. Thus, it holds promise as a tool for early identification and intervention efforts in addressing dyslexia challenges at the elementary education level. See Figure 5 for detail

Table 3. Test Results of Logistic Regression Method

	true 1	true 2	Class precision
Pred.1	62	3	96,38 %
Pred.2	4	30	88,24 %
class recall	93,94 %	90,91 %	

The Logistic Regression analysis shows coefficients, standard coefficients, standard errors, z-values, and p-values for attributes like Language Vocab, Memory, Speed, Visual Discrimination, Audio Discrimination, and Survey Score. Larger absolute coefficients indicate a stronger influence on the target variable. Memory has the most negative coefficient, followed by Visual Discrimination and Language Vocab, suggesting that lower values in these attributes significantly increase the likelihood of dyslexia. In contrast, Speed and Survey Score have smaller absolute coefficients, indicating a lesser impact. To interpret these findings, p-values must be considered to assess statistical significance. This analysis provides valuable insights into the key factors affecting early detection of dyslexia in elementary school students. Then, the results of the tree using this method with its attributes can be seen in Figure 6.

Table 4. Early Dyslexia Detection Data Tree Results Using KKN Method

Attribute	Coeficient	Std.coeficient	Std. Error	z-Value	p-Value
Language_vocab	-38.466	-8.203	594.417	-0.065	0.948
Memory	-74.140	-16.107	615.112	-0.121	0.904
Speed	-16.833	-3.541	572.647	-0.029	0.977
Visual_discrimination	-58.930	-12.292	612.198	-0.093	0.926
Audio_discrimination	-31.423	-6.260	572.214	-0.055	0.956
Survey_score	-24.917	-4.693	653.127	-0.038	0.970
Intercept	91.456	-48.572	286.738	0.319	0.750

3.1.4 Testing Naïve Bayes Algorithm

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem, with the "naive" assumption that features are independent of each other given the class label [33]. It's widely used for text classification and other classification tasks. Bayes' Theorem states:

$$P(C|X) = \frac{P(X|C)x\,P(C)}{P(x)} \quad (4)$$

Where:

- $P(C|X)$ is the probability of class CC given the features XX.
- $P(X|C)$ is the probability of observing features XX given class CC.
- $P(C)$ is the prior probability of class CC
- $P(X)$ is the prior probability of observing features XX

The Naïve Bayes algorithm was tested on a dyslexia early detection dataset of elementary school children, consisting of 100 samples split into 80% training, 10% validation, and 10% testing. The model showed excellent class-precision, achieving 100% precision for class 0 (no dyslexia detected) and class 1 (at risk of dyslexia). However, there was a slight drop in precision for class 2 (diagnosed dyslexia) to 82.5%. This

indicates that while the Naïve Bayes model is highly accurate for detecting no dyslexia and at-risk cases, it is slightly less accurate for diagnosed dyslexia cases. The Naïve Bayes model exhibits good sensitivity towards each class, with recall rates of 100% for class 0, 86.79% for class 1, and 100% for class 2. This indicates the model's effectiveness in accurately recalling individuals belonging to each category, including those diagnosed with dyslexia. Overall, these results suggest that the Naïve Bayes method can serve as an effective tool for early detection of reading difficulties, particularly dyslexia. See Figure 7 for details.

Table 5. Test Results of Naive Bayes Method

	true 1	true 2	true 0	Class precision
Pred.1	46	0	0	100,00 %
Pred.2	7	33	0	82,90 %
Pred 0	2	0	13	100,00 %
Class Recall	86,79 %	100,00 %	100,00 %	

The Naive Bayes sample distribution results provide probability distributions for each attribute in each class. Analysis reveals that for class 0 (no dyslexia detected), probabilities cluster around 0.3 to 0.6, peaking around 0.3 to 0.5, indicating diverse but centered values. For class 1 (at risk of dyslexia), probabilities range from 0.3 to 0.8, peaking around 0.5 to 0.6, showcasing more varied yet centered values. Meanwhile, for class 2 (diagnosed dyslexia), probabilities tend to be higher, ranging from 0.5 to 1.0, peaking around 0.7 to 0.9, indicating consistently higher values compared to other classes. Comparing attribute probability distributions across classes reveals attribute 2 consistently has higher values, suggesting its significant influence in predicting diagnosed dyslexia. The results of the simple distribution analysis for each attribute using the Naive Bayes method can be seen in Figure 8.

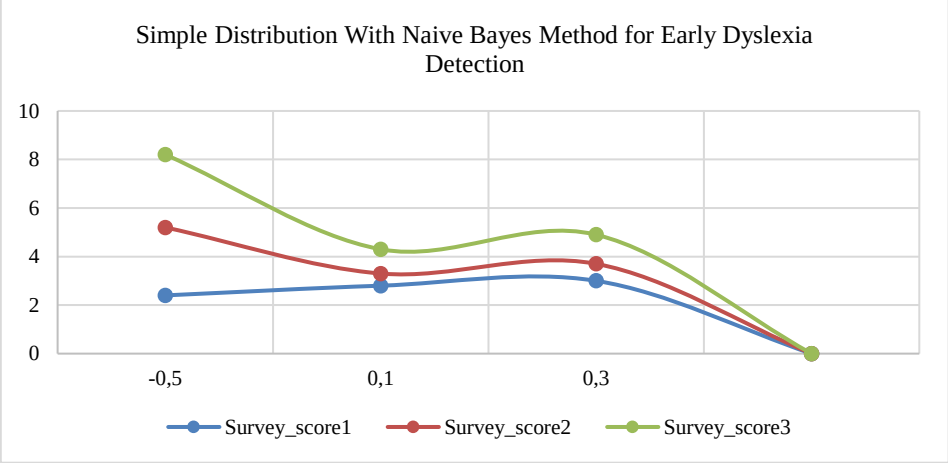

Figure 4. Simple Distribution Early Dyslexia Detection with Naïve Bayes Method

3.1.5 Testing Random Forest Algorithm

Random Forest is an ensemble learning method widely used for classification and regression tasks, operating by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees [34]. It begins with bootstrapping, where multiple subsets of data are created by random sampling with replacement from the original dataset, and each subset is used to train a different decision tree [35]. During the construction of each tree, a random subset of features is selected instead of using all features, promoting diversity among the trees and reducing their correlation [36]. Each tree is grown to its full extent without pruning, with the diversity resulting from the randomness in data and feature selection [37]. Once all the trees are built, Random Forest aggregates their predictions, using majority voting for classification tasks and averaging for regression tasks.

Random Forest combines predictions from multiple decision trees to provide stable and reliable results, particularly useful for the multifactorial data of dyslexia. In testing the algorithm on a dataset for early dyslexia detection in elementary school children, the model achieved precision ranging from 91.18% to 94.23% and recall from 92.31% to 93.94%, with an overall accuracy of 93.00%. These metrics indicate the model's effectiveness in correctly classifying both positive and negative cases. The process begins by splitting the dataset into training and testing subsets. Each decision tree is built using random subsets of training data and features, and their results are combined for the final prediction. This method enhances the model's ability to handle complex data and produce accurate predictions. The model's performance is evaluated using precision,

recall, and accuracy metrics, as illustrated in figures 9. The results demonstrate that Random Forest provides high performance in early dyslexia detection.

Table 6. Test Results of Random Forest Method

	true 1	true 2	true 0	Class precision
Pred.1	49	2	1	94,23 %
Pred.2	3	31	0	91,18 %
Pred 0	1	0	12	82,31 %
Class Recall	92,45 %	93,94 %	92,31 %	

The Random Forest method reveals a decision tree structure that highlights the key factors influencing class predictions. Initially, the tree evaluates the Survey Score attribute. If the Survey Score exceeds 0.250, it examines Visual Discrimination. If Visual Discrimination is above 0.250, it checks Audio Discrimination. For an Audio Discrimination value above 0.150 and Language Vocab greater than 0.750, the predicted class is 2 (diagnosed dyslexia). If Audio Discrimination is above 0.150 but Language Vocab is 0.750 or lower, it then assesses the Memory attribute. This decision process continues, evaluating relevant attributes to predict the class. This structure shows that Survey Score, Visual Discrimination, Audio Discrimination, Language Vocab, and Memory significantly contribute to the prediction. However, this analysis is based on the assumptions of the Random Forest model and might require further adjustments based on the specific context and needs of the research problem. The graphical results of each attribute using the Random Forest method are depicted in figure 10.

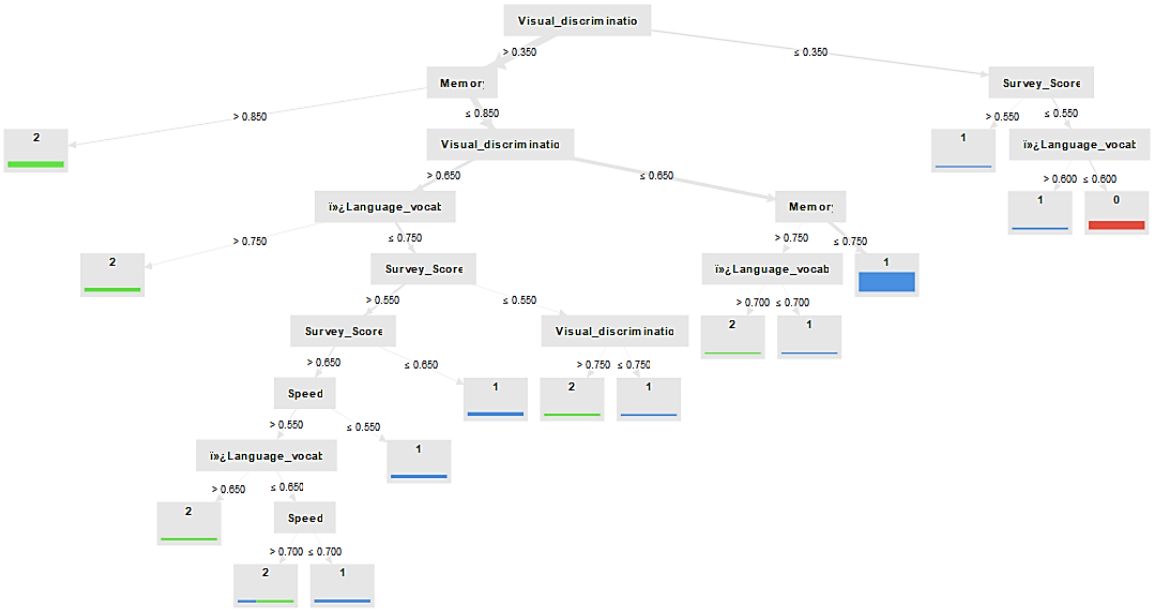

Figure 5. Graph Random Forest Model

4. CONCLUSION

The comparative analysis of various machine learning methods for early dyslexia detection in elementary school children revealed distinct strengths across models. The Decision Tree algorithm demonstrated robust precision, achieving 92.31% for dyslexia-prone, 90.62% for diagnosed dyslexia, and 86.67% for no dyslexia detected, with corresponding high recall values of 90.57%, 87.88%, and 100%, respectively. The K-Nearest Neighbors (KNN) model excelled with an overall accuracy of 94.00%, offering perfect precision for undetected dyslexia (100%) and high precision for dyslexia-prone (92.73%) and diagnosed dyslexia (93.76%), coupled with high recall rates for all classes. Logistic Regression highlighted significant predictors such as memory, visual discrimination, and language vocabulary, achieving precision of 95.38% for dyslexia-prone and 88.24% for diagnosed dyslexia, with recall rates of 93.34% and 90.91%, respectively. The Naïve Bayes model showed outstanding precision for no dyslexia and dyslexia-prone categories (100%), though slightly lower precision for diagnosed dyslexia (82.5%), with perfect recall for undetected dyslexia and diagnosed dyslexia classes. Finally, the Random Forest model demonstrated balanced performance with precision ranging from 91.18% to 94.23% and recall from 92.31% to 93.94%, achieving an overall accuracy of 93.00%. This comprehensive evaluation underscores the efficacy of these models in early

dyslexia detection, with each method offering unique advantages in precision and recall across different dyslexia categories**.**

ACKNOWLEDGEMENTS

Our gratitude goes to Budi Luhur University for funding this research through the university's internal funding program.

REFERENCES

[1] L. Yang *et al.*, "Prevalence of Developmental Dyslexia in Primary School Children: A Systematic Review and Meta-Analysis," *Brain Sciences*, vol. 12, no. 2, 2022, doi: 10.3390/brainsci12020240.

[2] H. W. Catts and Y. Petscher, "Early Identification of Dyslexia : Current Advancements and Future Directions," *Perspectives on Language and Literacy*, vol. 44, no. 3, pp. 33–36, 2018.

[3] R. Hernández-Vásquez, U. C. García, A. M. B. Barreto, M. L. R. Rojas, J. Ponce-Meza, and M. Saavedra-López, "An Overview on Electrophysiological and Neuroimaging Findings in Dyslexia," *Iranian Journal of Psychiatry*, vol. 18, no. 4, pp. 503–509, 2023, doi: 10.18502/ijps.v18i4.13638.

[4] G. Fragagonzález, I. I. Karipidis, and J. Tijms, "Dyslexia as a neurodevelopmental disorder and what makes it different from a chess disorder," *Brain Sciences*, vol. 8, no. 10, 2018, doi: 10.3390/brainsci8100189.

[5] M. E. Aguilar-Vafaie, N. Safarpour, M. Khosrojavid, and G. A. Afruz, "A comparative study of rapid naming and working memory as predictors of word recognition and reading comprehension in relation to phonological awareness in Iranian dyslexic and normal children," *Procedia - Social and Behavioral Sciences*, vol. 32, pp. 14–21, 2012, doi: 10.1016/j.sbspro.2012.01.003.

[6] N. M. Raschle, M. Chang, and N. Gaab, "Structural brain alterations associated with dyslexia predate reading onset," *NeuroImage*, vol. 57, no. 3, pp. 742–749, 2011, doi: 10.1016/j.neuroimage.2010.09.055.

[7] D. Theodoridou, P. Christodoulides, V. Zakopoulou, and M. Syrrou, "Developmental dyslexia: Environment matters," *Brain Sciences*, vol. 11, no. 6, 2021, doi: 10.3390/brainsci11060782.

[8] N. Ahmad, M. B. Rehman, H. M. El Hassan, I. Ahmad, and M. Rashid, "An Efficient Machine Learning-Based Feature Optimization Model for the Detection of Dyslexia," *Computational Intelligence and Neuroscience*, vol. 2022, 2022, doi: 10.1155/2022/8491753.

[9] S. Mascheretti *et al.*, "Neurogenetics of developmental dyslexia: From genes to behavior through brain neuroimaging and cognitive and sensorial mechanisms," *Translational Psychiatry*, vol. 7, no. 1, 2017, doi: 10.1038/tp.2016.240.

[10] S. Man Kit Lee, H. W. Liu, and S. X. Tong, "Identifying Chinese Children with Dyslexia Using Machine Learning with Character Dictation," *Scientific Studies of Reading*, vol. 27, no. 1, pp. 82–100, 2023, doi: 10.1080/10888438.2022.2088373.

[11] G. Wang, J. Zhao, M. Van Kleek, and N. Shadbolt, "Informing Age-Appropriate AI: Examining Principles and Practices of AI for Children," *Conference on Human Factors in Computing Systems - Proceedings*, 2022, doi: 10.1145/3491102.3502057.

[12] N. Mather and D. Schneider, "The Use of Cognitive Tests in the Assessment of Dyslexia," *Journal of Intelligence*, vol. 11, no. 5, p. 79, 2023, doi: 10.3390/jintelligence11050079.

[13] P. M. Paz-Alonso *et al.*, "Neural correlates of phonological, orthographic and semantic reading processing in dyslexia," *NeuroImage: Clinical*, vol. 20, pp. 433–447, 2018, doi: 10.1016/j.nicl.2018.08.018.

[14] R. W. Cooksey, "Descriptive Statistics for Summarising Data," *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*, pp. 61–139, 2020, doi: 10.1007/978-981-15-2537-7_5.

[15] L. Franzen, Z. Stark, and A. P. Johnson, "Individuals with dyslexia use a different visual sampling strategy to read text," *Scientific Reports*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-84945-9.

[16] M. Ramezani and A. J. Fawcett, "Cognitive-Motor Training Improves Reading-Related Executive Functions: A Randomized Clinical Trial Study in Dyslexia," *Brain Sciences*, vol. 14, no. 2, p. 127, 2024, doi: 10.3390/brainsci14020127.

[17] S. Itani, M. Rossignol, F. Lecron, and P. Fortemps, "Towards interpretable machine learning models for diagnosis aid: A case study on attention deficit/hyperactivity disorder," *PLoS ONE*, vol. 14, no. 4, 2019, doi: 10.1371/journal.pone.0215720.

[18] R. K. Wagner, J. Moxley, C. Schatschneider, and F. A. Zirps, "A Bayesian Probabilistic Framework for Identification of Individuals with Dyslexia," *Scientific Studies of Reading*, vol. 27, no. 1, pp. 67–81, 2023, doi: 10.1080/10888438.2022.2118057.

[19] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved Random Forest for Classification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4012–4024, 2018, doi: 10.1109/TIP.2018.2834830.

[20] D. & A. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," *J. Mach. Learn. Technol*, vol. 2, 2011, doi: 10.9735/2229-3981.

[21] A. J. Bowers and X. Zhou, "Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes," *Journal of Education for Students Placed at Risk*, vol. 24, no. 1, pp. 20–46, 2019, doi: 10.1080/10824669.2018.1523734.

[22] J. Kaliappan, A. R. Bagepalli, S. Almal, R. Mishra, Y. C. Hu, and K. Srinivasan, "Impact of Cross-Validation on Machine Learning Models for Early Detection of Intrauterine Fetal Demise," *Diagnostics*, vol. 13, no. 10, 2023, doi: 10.3390/diagnostics13101692.
[23] S. Y. Ho, K. Phua, L. Wong, and W. W. Bin Goh, "Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability," *Patterns*, vol. 1, no. 8, 2020, doi: 10.1016/j.patter.2020.100129.
[24] A. Amro, M. Al-Akhras, K. El Hindi, M. Habib, and B. A. Shawar, "Instance Reduction for Avoiding Overfitting in Decision Trees," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 438–459, 2021, doi: 10.1515/jisys-2020-0061.
[25] G. S. K. Ranjan, A. Kumar Verma, and S. Radhika, "K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries," *2019 IEEE 5th International Conference for Convergence in Technology, I2CT 2019*, 2019, doi: 10.1109/I2CT45611.2019.9033691.
[26] F. Salehi, E. Abbasi, and B. Hassibi, "The impact of regularization on high-dimensional logistic regression," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[27] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Constrained Naïve Bayes with application to unbalanced data classification," *Central European Journal of Operations Research*, vol. 30, no. 4, pp. 1403–1425, 2022, doi: 10.1007/s10100-021-00782-1.
[28] V. S, "Predicting Dyslexia with Machine Learning: A Comprehensive Review of Feature Selection, Algorithms, and Evaluation Metrics," *Journal of Behavioral Data Science*, vol. 3, no. 1, pp. 1–14, 2023, doi: 10.35566/jbds/v3n1/s.
[29] F. J. Yang, "An extended idea about decision trees," *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019*, pp. 349–354, 2019, doi: 10.1109/CSCI49370.2019.00068.
[30] M. Mailagaha Kumbure and P. Luukka, "A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance," *Granular Computing*, vol. 7, no. 3, pp. 657–671, 2022, doi: 10.1007/s41066-021-00288-w.
[31] M. Maalouf, "Logistic regression in data analysis: An overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, 2011, doi: 10.1504/IJDATS.2011.041335.
[32] D. Swain *et al.*, "Cardiovascular Disease Prediction using Various Machine Learning Algorithms," *Journal of Computer Science*, vol. 18, no. 10, pp. 993–1004, 2022, doi: 10.3844/jcssp.2022.993.1004.
[33] V. H. Kamble and M. P. Dale, "Machine learning approach for longitudinal face recognition of children," in *Machine Learning for Biometrics: Concepts, Algorithms and Applications*, B. M. Partha Pratim Sarangi, Madhumita Panda, Subhashree Mishra, Bhabani Shankar Prasad Mishra, Ed., 2022, pp. 1–27. doi: 10.1016/B978-0-323-85209-8.00011-0.
[34] Y. & W. Y. & Z. J. Liu, "New Machine Learning Algorithm: Random Forest," pp. 246–252, 2012.
[35] T.-H. Lee, A. Ullah, and R. Wang, "Bootstrap Aggregating and Random Forest," 2020.
[36] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst Appl*, vol. 237, p. 121549, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.
[37] C. L. Koo, M. J. Liew, M. S. Mohamad, and A. H. Mohamed Salleh, "A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology," *Biomed Res Int*, vol. 2013, pp. 1–13, 2013, doi: 10.1155/2013/432375.
[38] A. F. Lubis et al., "Classification of Diabetes Mellitus Sufferers Eating Patterns Using K-Nearest Neighbors, Naïve Bayes and Decission Tree," Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 1, pp. 44–51, Apr. 2024, doi: 10.57152/predatecs.v2i1.1103.

BIBLIOGRAPHY OF AUTHORS

I am a lecturer at the Faculty of Information Technology, Universitas Budi Luhur, Jakarta. My expertise includes Educational Technology, Educational Data Mining, Machine Learning, and Decision Support Systems. I focus on enhancing educational processes through technology, such as e-learning and LMS. My research in Educational Data Mining aims to predict academic performance and develop personalized learning strategies. I also develop machine learning algorithms for various applications and design decision support systems to optimize decision-making. I hold a Master's degree from Budi Luhur University and have published research in internasional dan national reputable journals and conferences.

I am a lecturer at the Faculty of Information Technology, Universitas Budi Luhur, Jakarta. My expertise lies in UI/ UX Design, E-Procurement, Knowledge Management Systems, and Project Management. I focus on creating user-centered designs that enhance user experience and engagement. In E-Procurement, I specialize in developing systems that streamline and optimize the procurement process. My work in Knowledge Management Systems involves organizing and managing organizational knowledge to improve efficiency and innovation. Additionally, I teach and apply project management principles to ensure successful project execution. I hold a Master's degree from Budi Luhur University Jakarta and have published several research papers in reputable journals and conferences.