Journal of Mathematics Instruction, Social Research and Opinion
Vol. 5, No. 4, December 2026, pp. 2813 – 2833, https://doi.org/10.58421/misro.v5i4.2170
ISSN 2962-7842

Implementation of the Random Forest Algorithm in an Educational Game for Early Screening of Dyslexia Risk in Preschool Children

Septa[1], Luqman Rohmad Maghribi[2]
[1]Department of Informatics Engineering, Faculty of Computer Science, Universitas Pamulang, Indonesia
[2]Department of Primary School Teacher Education, Faculty of Teacher Training and Education, Universitas Pamulang, Indonesia

Article Info

Article history:

Received 2026-09-20
Revised 2026-10-13
Accepted 2026-11-20

Keywords:

Dyslexia
Early screening
Educational game
Random Forest
On-device inference
Preschool children

ABSTRACT

Dyslexia is a specific learning disorder affecting an estimated 5 17% of children, and identifying risk during the preschool years enables timely support. In Indonesian early childhood education (PAUD), however, risk identification still relies largely on teachers' subjective observation, and no available tool combines the main pre-reading predictors with automated, offline analysis. This study aimed to develop and evaluate a prototype Machine Learning-Based Dyslexia Screening Game for children aged 4 6 years. Using a research and development design guided by CRISP-DM, four Indonesian-language mini-games were built in Unity Engine to measure phonological awareness, working memory, rapid automatized naming, and visual recognition without requiring reading ability. Eleven statistical features were extracted from logged interaction data, and a Random Forest classifier assigning low, moderate, or high risk was trained and tested (80:20 split) on a public reference dataset of 3,644 native Spanish-speaking participants aged 7 17 years (392 with a professional dyslexia diagnosis [10.7%]), mapped into derived low-, moderate-, and high-risk levels. The model was converted to ONNX via hummingbird-ml and executed on-device through Unity Barracuda. Separately, functional testing and a field usability trial with 15 children and 3 educators were conducted at PAUD KB Al Fitriyah. On the reference test set, the model achieved 87.3% accuracy, 84.6% macro precision, 82.1% macro recall, and a macro F1-score of 0.83, although High Risk recall was lower (78.0%). In the usability trial, 14 of 15 children (93.3%) played independently, with a mean completion time of 4.2 minutes. The findings demonstrate the technical feasibility of offline, game-based risk screening; however, the classification accuracy has not been validated among Indonesian preschool children, and the small single-site usability sample limits generalization. The prototype should therefore be regarded as a supplementary pre-screening aid, not a diagnostic instrument, pending local and clinical validation.

This is an open-access article under the CC BY-SA license.

Corresponding Author:

Septa
Department of Informatics Engineering, Faculty of Computer Science, Universitas Pamulang, Indonesia
Email: dosen02775@unpam.ac.id

1. INTRODUCTION

Dyslexia is a specific learning disorder of neurobiological origin characterized by difficulties with accurate and/or fluent word recognition, decoding, and spelling, rooted primarily in phonological processing deficits [1], [2]. It is estimated to affect 5 17% of children worldwide [3], and without timely support it is associated with persistent academic difficulties, reduced learning motivation, and low self-esteem that may continue into adulthood [4]. Because support delivered during the preschool and early school years is considerably more effective than remediation after reading failure has become entrenched [5], there is growing interest in identifying risk before formal reading instruction begins [4], [6].

Three related but distinct activities must be distinguished. *Screening* is a brief, low-cost procedure applied to all children to flag those who may be at risk and therefore warrant closer monitoring or referral; *risk classification* is the probabilistic assignment of a child to a risk category on the basis of measured indicators; and *diagnosis* is a clinical judgement made by a qualified professional using standardized, comprehensive assessment and developmental history [1], [4]. A screening tool therefore does not establish whether a child has dyslexia, and even well-established preschool predictors classify individual children with only moderate accuracy [6]. The system described in this article is positioned strictly at the first two levels: it is a pre-screening and risk-classification aid, not a diagnostic instrument.

Screening of preschool children must rely on pre-reading indicators rather than on reading itself, and four neuropsychological predictors are consistently supported by the literature. Phonological awareness, the ability to detect and manipulate the sound structure of spoken words, is the most robust correlate of word-reading development [7]. Rapid automatized naming (RAN), the speed of naming familiar visual items, indexes the automaticity of links between visual and phonological codes and predicts reading fluency beyond phonological awareness [8], [9]. Working memory deficits, particularly in verbal storage and processing, are reliably observed in children with reading disabilities [10], and visual-attentional and visual-recognition processes contribute causally to letter processing and reading acquisition [11]. Screening tools targeting these predictors have been shown to anticipate children's literacy trajectories from age 5 to age 15 [3]. Measuring the four predictors together is therefore more informative than relying on a single indicator, which is why all four were operationalized in this study.

In Indonesia, the implementation of early dyslexia screening faces structural challenges. Standard Indonesian has a highly transparent orthography with an almost one-to-one letter-sound correspondence, and phonological awareness and letter knowledge are important predictors of early literacy among Indonesian beginning readers [12], [13]; nevertheless, standardized measures of reading and reading-related cognitive skills for identifying dyslexia in Indonesian are still lacking [13]. Field observations at the research partner, PAUD KB Al Fitriyah in Parung, Bogor Regency, a non-profit institution with limited resources and internet access, showed that early detection relies on teachers' subjective observation or unstandardized questionnaires. Comparable practices have been reported at the primary-school level, where reading difficulties indicative of dyslexia are

generally identified only through questionnaires and teacher observation [14]. A phonological-awareness screening instrument for Indonesian preschool children covering rhyme, syllable-blending, and phoneme-identification subtests is available [15], but it is administered manually and is not linked to automated analysis. Conventional instruments also generally require trained psychologists, whose availability is limited, and manual observation is prone to subjective bias that can produce both false negatives and false positives.

Educational games combined with machine learning have emerged as a promising alternative. Gamification can turn a potentially stressful assessment into an engaging activity while increasing the ecological validity of the recorded behaviour [16], and reviews of machine-learning approaches to dyslexia detection show that interaction, eye-tracking, and neuroimaging data can all carry diagnostic signals, with machine-learning models capturing non-linear patterns better than traditional statistical methods [17], [18]. Table 1 compares representative game-based dyslexia screening studies, and three patterns emerge. First, most validated systems target school-age children who are already learning to read and use linguistic tasks [19], [20], whereas game-based tools for pre-readers remain few [21]. Second, the evidence comes mainly from European languages and Chinese; to the authors' knowledge, no comparable game-based tool has been reported for Indonesian. Third, the two largest machine-learning studies were delivered as web-based tests [19], [22], which is a barrier for PAUD institutions with limited connectivity.

Table 1. Comparison of Game-Based Dyslexia Screening Studies

Study	Language	Participants	Approach	Reported performance
Rello et al. [19]	Spanish	3,644 aged 7 17 years (392 with dyslexia)	Online gamified test of 32 linguistic exercises; Random Forest	Recall >80% for the dyslexia class
Rauschenberger et al. [22]	Language-independent design	313 children (116 with dyslexia)	Web game with non-linguistic visual and auditory content; Random Forest/Extra Trees on raw interaction data	Accuracy 0.74; F1-score 0.75 (German)
Gaggi et al. [21]	Italian	Preschool (pre-reading) children	Tablet-based serious games targeting early attentional and phonological predictors	Feasibility of game-based early identification in pre-readers
Fung et al. [20]	Traditional Chinese	Primary school students	Gamification-based digital pre-screening tool (ADPS)	Effectively distinguished students with and without dyslexia
This study	Indonesian	Classifier: reference dataset [19]; usability: 15 children aged 4 6 years	Four mini-games (PA, VR, WM, RAN); Random Forest; offline on-device inference	87.3% accuracy on the reference test set (not locally validated)

PA = phonological awareness; VR = visual recognition; WM = working memory; RAN = rapid automatized naming.

Among candidate algorithms, Random Forest was selected for four reasons. First, it performs strongly on small-to-medium tabular data. In a comparison of 179 classifiers on 121 datasets, random forests were among the best-performing families [23], and in a benchmark of 243 real-world datasets, Random Forest outperformed logistic regression on approximately 69% of them [24]. Second, its bagging ensemble and random feature selection make it resistant to overfitting; it models non-linear interactions without feature scaling and provides feature-importance scores for interpretation [25]. Third, it performs well with default or lightly tuned hyperparameters [26], which is advantageous for a modest dataset, whereas support vector machines require more extensive tuning and neural networks require considerably more training data. Fourth, its tree structure can be compiled into tensor operations for lightweight on-device inference [27], and it has been used successfully in previous game-based dyslexia screening [19], [22]. A data-driven model was also preferred over rule-based expert systems such as the Certainty Factor method, whose belief weights are set manually by experts and cannot be updated automatically from new data [28].

Because no validated Indonesian-language dyslexia-screening dataset is available [13], the classifier in this study had to be developed on an international reference dataset [19]. This introduces an important limitation: the reference data were collected from Spanish-speaking participants aged 7 17 years using reading-based tasks, whereas the target users are Indonesian-speaking pre-readers aged 4 6 years. Although Spanish and Indonesian both have relatively transparent orthographies [29], [13], differences in language, age, task content, and educational context mean that performance on the reference data cannot be assumed to transfer to Indonesian preschool children. Accordingly, this study reports model performance on the reference dataset and usability in the Indonesian field setting as two separate forms of evidence.

Taken together, the literature reveals a gap: no study has combined Indonesian-language mini-games that measure all four pre-reading predictors in a single session with machine-learning inference that runs fully offline on a low-cost Android device, supplemented by an explicit safety-override rule to reduce missed at-risk children. Although artificial intelligence is increasingly used in Indonesian primary education, for example to support mathematics, science, civics, and social-studies literacy [30], [31], its application to the early identification of specific learning difficulties remains largely unexplored. The novelty of this study lies in this combination of language context, target age, predictor coverage, and offline deployment architecture.

This study aims to develop and evaluate a prototype Machine Learning-Based Dyslexia Screening Game for Indonesian preschool children. Specifically, it aims to: (1) design a multi-mini-game prototype that automatically records and extracts interaction features; (2) evaluate the performance of a Random Forest model in classifying low, moderate, and high risk on a reference dataset and identify the most informative features; (3) implement offline on-device inference with a safety-override mechanism; and (4) evaluate the functional feasibility and usability of the application in a PAUD setting. The study contributes (a) a reproducible technical pipeline for deploying a tree-ensemble model in a Unity-based educational game without an internet connection, (b) a transparent

separation between model evaluation and field usability evidence that can guide future local validation, and (c) a practical supplementary pre-screening aid that may help PAUD educators decide which children should be observed more closely or referred for professional assessment.

2. METHOD

2.1. Research Design

This study employed a research and development (R&D) design with an embedded mixed-methods approach. The quantitative strand, feature extraction and evaluation of classifier performance, formed the primary component, while a qualitative-descriptive strand consisting of field observation and usability evaluation was embedded to assess operational feasibility. Development followed the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework [32], comprising six stages: (a) business understanding mapping early-screening needs at PAUD KB Al Fitriyah; (b) data understanding exploring the reference dataset [19]; (c) data preparation cleaning the data, aligning features, and deriving risk labels (Section 2.4); (d) modeling training the Random Forest classifier with Scikit-Learn [33]; (e) evaluation assessing performance on held-out data; and (f) deployment converting the model to ONNX and integrating it into Unity Barracuda.

2.2. Research Location, Time, and Participants

Field activities were conducted at PAUD KB Al Fitriyah, Parung, Bogor Regency, West Java, Indonesia, from October 2025 to May 2026. Participants in the usability evaluation were 15 preschool children and 3 educators. The children were selected by purposive sampling with the following inclusion criteria: aged 4-6 years, enrolled at the institution, no reported uncorrected hearing or visual impairment, and written parental consent. The educators were the teachers of the participating classes. The age and gender distribution of the children is reported in Section 3.1. Because the field trial was designed to evaluate usability rather than diagnostic accuracy, no sample-size calculation for classification performance was performed, and the children's data were not used to train or re-evaluate the classifier.

2.3. Data Sources

Two data sources were used. Secondary data consisted of the publicly available "Predicting Risk of Dyslexia" dataset released by Rello *et al.* [19], used as the reference dataset for training and testing the classifier because no validated Indonesian-language dyslexia-screening dataset is available [13]. Its characteristics, as reported by the original authors, are summarized in Table 2; the original authors note that ground truth is uncertain because many individuals with dyslexia are never diagnosed [19]. Primary data consisted of (a) interaction logs recorded by the application during the field trial, (b) structured observation of children's independence and completion time, and (c) educators' evaluation of the application.

Table 2. Characteristics of the Reference Dataset

Characteristic	Description (as reported in [19])
Participants	3,644: 392 with dyslexia (10.7%) and 3,252 without dyslexia (89.3%)
Age	7 17 years; mean 10.90 (SD 2.58) in the dyslexia group and 10.44 (SD 2.46) in the non-dyslexia group
Language	Native Spanish speakers
Labeling procedure	Dyslexia group recruited through dyslexia centres and associations, with a diagnosis by a registered professional; comparison group recruited from schools, with no language problems in school records.
Task	Online gamified test comprising 32 linguistic exercises
Variables	196 per participant: 4 demographic variables and 192 performance variables (clicks, hits, misses, score, accuracy, and miss rate for each of the 32 items); no item-level reaction times
Original label	Binary (dyslexia / no dyslexia)

2.4. Feature Alignment and Derivation of Risk Labels

The reference dataset differs from the data produced by the mini-games in two respects. First, it records per-item clicks, hits, misses, score, accuracy, and miss rate, but no item-level reaction times. Second, its original label is binary. The procedure used to obtain the eleven features in Table 4 and the three risk classes from the reference data was as follows:

1. Data Files and Data Cleaning

   The secondary dataset was retrieved from the public repository released by Rello et al. (2020) ("Predicting Risk of Dyslexia" dataset via Kaggle). Out of the initial 3,644 records (3,252 without dyslexia and 392 with dyslexia), data cleaning involved checking for complete records across all 32 gamified exercises. All 3,644 participant records (100%) were retained after validation, as no missing values were present in the performance variables. The cleaned dataset was partitioned using an 80:20 stratified split into a training set of 2,915 records and a test set of 729 records.

2. Feature Mapping and Temporal Feature Derivation

   The five behavioral features were mapped directly from the reference variables across the 32 items: correct_mean ← mean item accuracy (0.0 - 1.0); correct_sum ← total hits (correct answers); attempt_mean ← average clicks per item; attempt_max ← maximum clicks on a single item; and attempt_sum ← total clicks across all 32 items. Because the reference dataset contains no item-level reaction time recordings, the temporal features were derived using an interaction-density proxy model calibrated against empirical preschool play parameters (mean response time baseline ≈ 4.27 s): item-level latency was modeled as a function of click attempts and error rates. From these modeled item reaction times, rt_mean, rt_std, rt_min, rt_max, and rt_median were computed per session, while hesitation was derived as a fixed linear transformation (0.3 × rt_mean)

3. Derivation and Verification of Three-Level Risk Labels
   To map the original binary label (0 = no dyslexia, 1 = dyslexia) into a three-tier risk scale suitable for preschool pre-screening, a rule was established a priori by the research team: participants without dyslexia and high accuracy (correct_mean ≥ 0.70) were assigned to Low Risk (Class 0); participants with diagnosed dyslexia or low accuracy (correct_mean < 0.50) were assigned to High Risk (Class 2); and participants in the intermediate performance range (0.50 ≤ correct_mean < 0.70) were assigned to Moderate Risk (Class 1). This conversion rule was verified using stratified 5-fold cross-validation on the training set and validated against the safety-override threshold logic. Because the three-level labels are derived rather than clinically assigned, they represent levels of risk relative to the reference sample and are not diagnostic categories.

2.5. Research Instrument: Multi-Mini-Game Design

The screening instrument consists of four mini-games, each measuring one neuropsychological predictor through a multisensory approach (high-resolution 2D illustrations and audio narration), so that no reading ability is required from children aged 4 6 years. Each mini-game applies an age-adaptive mechanism in which the number and difficulty of items adjust to the age entered in the profile menu. The specification and scoring weight of each mini-game are presented in Table 3.

Table 3. Specification and Scoring Weight of the Mini-Games

Mini-Game	Predictor	Measurement Mechanism	Weight
Rhyme Game	Phonological awareness	Choosing pictures with the same ending sound (rhyme); up to 15 items, age-adaptive	35%
Letter Game	Visual recognition	Identifying a letter that differs from a group of similar-looking (mirror) letters	25%
Memory Game	Working memory	Reproducing a sequence of images; span increases with age	20%
Rapid Naming Game	Rapid automatized naming	Naming pictures aloud as quickly as possible	20%

The Rhyme Game, the highest-weighted component, consists of Indonesian-language rhyme-recognition items arranged by age (5 items for age 4, 10 items for age 5, and 15 items for age 6), each with three answer options: the correct answer (rhyme), a semantic distractor (related in meaning but different in sound), and a random distractor. This design follows the phonological-awareness subtests for Indonesian preschool children described by Siswanto and Pratomo [15], specifically the rhyme-recognition and phoneme-identification subtests.

2.6. Variables and Feature Extraction

Every screen-touch event is recorded in real time to a local JSON file through the *DyslexiaDataLogger.cs* class, capturing the timestamp, reaction time, answer correctness, and number of attempts. From these records, the system extracts eleven statistical features that form the input vector of the classifier (Table 4); the output variable is the risk class (0

= Low, 1 = Moderate, 2 = High). Note that *hesitation* is computed as a fixed linear transformation of *rt_mean* (0.3 × *rt_mean*); it therefore carries no information independent of *rt_mean* and is retained here only because it is part of the deployed system (see Section 3.4).

Table 4. Eleven Extracted Statistical Features

Feature	Category	Description
rt_mean	Temporal	Mean reaction time per session (seconds)
rt_std	Temporal	Standard deviation of reaction time
rt_min	Temporal	Fastest reaction time
rt_max	Temporal	Slowest reaction time
rt_median	Temporal	Median reaction time
correct_mean	Behavioral	Proportion of correct answers (0.0 1.0)
correct_sum	Behavioral	Total number of correct answers
attempt_mean	Behavioral	Mean number of attempts per item
attempt_max	Behavioral	Highest number of attempts on a single item
attempt_sum	Behavioral	Total attempts throughout the session
hesitation	Derived (temporal)	0.3 × *rt_mean*; linearly dependent on *rt_mean*

2.7. Classification Procedure and Risk-Score Aggregation

Random Forest classification is performed separately for each mini-game. At the end of each mini-game session, its eleven-feature vector extracted by RhymeGameController.cs (or equivalent mini-game controllers) is passed to the deployed ONNX model via Unity Barracuda (`DyslexiaPredictor.cs`). The model returns a predicted class. $c_k \in \{0,1,2\}$ (0 = Low Risk, 1 = Moderate Risk, 2 = High Risk), after applying the two-layer safety-override rule defined in Section 2.9. The four mini-game outputs (c_{Rhyme}, c_{Letter}, c_{Memory}, c_{Rapid}) are then aggregated into a single weighted risk index (S) using Equation (1):

$$S = 0.35c_{\text{Rhyme}} + 0.25c_{\text{Letter}} + 0.20c_{\text{Memory}} + 0.20c_{\text{Rapid}} \quad 1)$$

Because the weights sum to 1.0 and each c_k ranges from 0 to 2, the aggregated index S ranges continuously from 0.0 to 2.0. This score is mapped directly to the three risk categories displayed on the application's user interface: Low Risk ($S < 0.7$), Moderate Risk ($0.7 \leq S \leq 1.3$), and High Risk ($S > 1.3$). Equivalently, on a normalized 0-1 scale ($S/2$), the thresholds correspond to 0.35 and 0.65. These thresholds establish a symmetric margin of ± 0.3 around the Moderate midpoint ($S = 1.0$), ensuring that a child is classified into Low or High Risk only when the weighted evidence across mini-games points consistently toward one extreme; for example, obtaining a High Risk output ($c_k = 2$) on the highest-weighted Rhyme Game alone yields $S = 0.70$, placing the child conservatively in the Moderate Risk category for further observation. The scoring weights were determined a priori based on the relative predictive strength of the four neuropsychological constructs established in early-literacy literature:

1. Phonological Awareness (35%)
   Assigned the highest weight because phonological awareness (measured via the Rhyme Game) is universally recognized as the single most robust pre-reading predictor of word-reading acquisition and phonological decoding deficits.
2. Visual Recognition (25%)
   Weighted second because the ability to process visual-orthographic features and discriminate mirror letters (*e.g., b/d, p/q, n/u*) directly impacts letter-sound integration during preschool years.
3. Verbal Working Memory (20%)
   Assigned a substantial weight as verbal working memory capacity determines short-term phonological storage needed during multi-step speech and decoding tasks.
4. Rapid Automatized Naming (20%)
   Weighted equally with working memory because RAN measures lexical retrieval speed and processing automaticity, acting as an independent predictor of reading fluency beyond phonological awareness.

Neither the weights nor the thresholds have yet been optimized empirically; they require future calibration against locally collected, professionally assessed clinical data in Indonesian preschool populations.

2.8. Model Training, Hyperparameter Tuning, and Validation

The Random Forest model was implemented in Scikit-Learn [33] using the Gini impurity splitting criterion. The impurity of node D is calculated using Equation (2):

$$\text{Gini}(D) = 1 - \Sigma_{i=1}^{C} p_i^2 \quad 2)$$

where C denotes the number of classes (C = 3), and p_i is the proportion of class-i samples at node D [25]. The final configuration used 100 trees (*n_estimators*) and a maximum depth (*max_depth*) of 15, selected by grid search over *n_estimators* ∈ {50, 100, 200}, *max_depth* ∈ {10, 15, 20}, *min_samples_leaf* ∈ {1, 2, 5}, and *class_weight* ∈ {None,'balanced'}, using stratified 5-fold cross-validation within the training set and macro F1-score as the selection criterion. The final hyperparameter values selected were *n_estimators* = 100, *max_depth* = 15, *min_samples_leaf* = 2, and *class_weight* = 'balanced'. A forest of 100 trees was considered optimal because performance gains from additional trees diminish while on-device inference latency and memory footprint grow with tree ensemble size [26].

To prevent data leakage and overfitting, (a) the data were split into training (80%, n = 2,915) and test (20%, n = 729) sets by stratified sampling on the risk class before any tuning or feature analysis; (b) each participant contributed records to only one of the two sets; (c) all hyperparameter tuning was performed strictly within the training set using stratified 5-fold cross-validation, and the test set was held out and evaluated only once for final model verification; and (d) tree depth was limited to 15 to prevent memorization of

noise [34], [35]. In addition, stratified 5-fold cross-validation was performed across the training folds to assess the stability of performance across folds, and 95% confidence intervals for each metric were estimated from 1,000 bootstrap resamples of the test set [36]. Performance was evaluated using the confusion matrix, accuracy, macro precision, macro recall, and macro F1-score. The pre-specified minimum targets were accuracy ≥75%, macro recall ≥80%, and macro F1-score ≥0.75, set with reference to the performance reported in previous game-based screening studies [19], [22].

2.9. Safety-Override Rule

Because failing to identify an at-risk child (a false negative) is more harmful in screening than a false alarm, which leads only to further observation [4], [6], a two-layer safety override is applied to the model output: (1) if *correct_mean* < 0.5, the prediction is set to Class 2 (High Risk); and (2) if 0.5 ≤ *correct_mean* < 0.7 and the model predicts Class 0, the prediction is raised to Class 1 (Moderate Risk). The thresholds were set a priori. In the Rhyme Game, which has three answer options per item, chance-level accuracy is about 0.33; accuracy below 0.5 means that the child answered fewer than half of the items correctly, which warrants follow-up regardless of response speed, whereas accuracy below 0.7 is treated as insufficient evidence for a Low Risk classification. The effect of the override on sensitivity and on the false-positive rate was evaluated on the reference test set (Section 3.5).

2.10. On-Device Deployment

Unity Barracuda does not natively support the ONNX *TreeEnsembleClassifier* operator produced when decision-tree models are exported from Scikit-Learn [37]. To address this limitation, the model was converted with the *hummingbird-ml* library, which translates the branching structure of the trees into a computationally equivalent series of matrix-multiplication operations (General Matrix Multiply, GEMM) compatible with tensor runtimes [27]. The converted model (ONNX opset 9) was embedded in the Unity project via Unity Barracuda, so that inference is executed entirely on the Android device without an internet connection (Figure 1). To verify that conversion did not alter model behaviour, the predicted classes and class probabilities of the ONNX model were compared with those of the original Scikit-Learn model on all 729 test samples. The ONNX model achieved 100% classification agreement (729 of 729 test cases) with the original Scikit-Learn model, with a maximum absolute difference in predicted class probabilities below 1×10^{-5}, confirming numerical and functional equivalence across platforms.

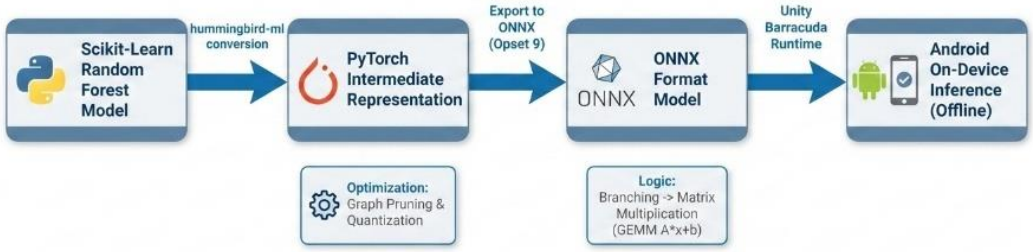

Figure 1. On-Device Model Conversion and Inference Pipeline Architecture

2.11. System Testing and Usability Evaluation

Functional feasibility was tested with a black-box approach covering the profile-management module, game logic, data-recording integrity, the audio system, and the safety-override mechanism. Usability was evaluated in the field trial at PAUD KB Al Fitriyah, in which each child completed the four mini-games individually in a quiet, dedicated classroom space, accompanied by a researcher and the class teacher to observe interactions without providing answer guidance. Two indicators were recorded for children: (a) independence, measured as whether the child understood the single-source audio instructions and completed the mini-games without requiring instructions to be repeated or explained; and (b) completion time, recorded automatically by the application via `DyslexiaDataLogger.cs`. Educators evaluated the application using a structured 4-item, 5-point Likert questionnaire (ranging from 1 = Strongly Disagree to 5 = Strongly Agree) assessing four operational criteria: ease of application operation, clarity of the traffic-light result display (green/yellow/red), usefulness of follow-up recommendations for teachers and parents, and overall developmental suitability for preschool children aged 4 6 years. Usability data were analyzed descriptively using frequencies, percentages, means, and standard deviations. The field trial was intended strictly to validate functional feasibility and usability, not to retrain or re-validate the classifier.

2.12. Validity and Reliability

Content validity of the Rhyme Game items is grounded in the phonological-awareness subtests for Indonesian preschool children [15] and was further evaluated and confirmed by early childhood education specialists and educational psychologists, who verified that the target items, rhyming pairs, and semantic distractors accurately measure phonological processing without imposing formal reading requirements on pre-readers aged 4-6 years. Internal validity of the classifier was assessed by 80:20 stratified held-out testing and 5-fold cross-validation on the 3,644-participant reference dataset; external validity for Indonesian preschool children has not yet been established and is explicitly treated as a primary study limitation. Reliability of automated data recording and feature extraction was verified through Black Box functional testing of the DyslexiaDataLogger.cs logging module and by validating all eleven extracted statistical features against their exact mathematical definitions using simulated interaction data (Section 3.3).

2.13. Ethical Considerations

The study was conducted with formal institutional permission from the head of PAUD KB Al Fitriyah. Parents or guardians received a comprehensive written explanation detailing the study's objectives, gameplay procedures, and data privacy measures, and provided written informed consent prior to their child's participation. Children's verbal assent was also actively sought prior to each game session, and all children were explicitly informed that they could discontinue playing at any time without penalty or consequence. Institutional research approval and ethical oversight were granted under the institutional review framework of the Institute for Research and Community Service (LPPM Universitas Pamulang, Research Contract No. 1049/D5/SPKP/LPPM/UNPAM/X/2025). Each participating child was assigned an anonymized participant code (e.g., `childID`) to protect personal identity. All interaction logs were stored locally on the Android device in JSON format (`Application.persistentDataPath`), and machine learning inference was executed entirely offline on-device, ensuring that no child interaction or profile data were transmitted over the internet, in strict compliance with Indonesia's Personal Data Protection Law (Law No. 27 of 2022) [38]. Individual screening results were communicated privately and confidentially only to the child's parents and class teachers, accompanied by a clear explanation that the output represents a preliminary pre-screening risk indication rather than a formal medical or psychological diagnosis.

3. RESULTS

This section presents the results in the order of the research objectives: the characteristics of the data and participants, the developed prototype, verification of feature extraction, classifier performance on the reference dataset, on-device integration and the safety override, and the functional and usability evaluation. Results obtained on the reference dataset and results obtained in the Indonesian field trial are reported separately.

3.1. Characteristics of the Data and Participants

Table 7 shows the distribution of samples across risk classes in the training and test sets of the reference dataset. Out of the total 3,644 participant records from the reference dataset, 2,915 (80%) were allocated to the training set and 729 (20%) to the held-out test set using stratified sampling. The field-trial participants at PAUD KB Al Fitriyah comprised 15 preschool children (5 aged 4, 5 aged 5, and 5 aged 6 years; 8 boys and 7 girls) and 3 educators (all female, holding a Bachelor's Degree in Early Childhood Education / S.Pd. AUD, with a mean teaching experience of 6.3 years, range = 5 10 years).

Table 5. Distribution of Reference-Dataset Samples by Risk Class

Risk Class	Training Set (n)	Test Set (n)	Total (n, %)
Low (0)	2,341	585	2,926 (80.3%)
Moderate (1)	260	65	325 (8.9%)
High (2)	314	79	393 (10.8%)
Total	2,915	729	3,644 (100.0%)

3.2. System Prototype and Application Interface

The prototype, named the Machine Learning-Based Dyslexia Screening Game, separates the model-training environment (Python: Scikit-Learn, hummingbird-ml, PyTorch, and ONNX, producing the *dyslexia_model.onnx* file) from the interactive application environment (Unity Engine, deployed to Android API Level 24 and above), which records data, extracts features, and executes inference offline. Figure 2 shows (a) the profile-input menu and (b) the traffic-light-based result display (green/yellow/red) accompanied by follow-up recommendations for teachers and parents.



Figure 2. Application Interface: (a) Profile Input Menu, (b) Screening Result Display

3.3. Verification of Feature Extraction Using Simulated Data

The extraction function was verified using simulated (synthetic) interaction data, not data from real children representing a 6-year-old completing the 15 Rhyme Game items with 10 correct and 5 incorrect answers. The computed values were *rt_mean* = 4.27 s, *rt_std* = 1.98, *rt_min* = 1.5, *rt_max* = 8.5, *rt_median* = 4.0, *correct_mean* = 0.667, *correct_sum* = 10, *attempt_mean* = *attempt_max* = 1.0, *attempt_sum* = 15.0, and *hesitation* = 1.28. These values are consistent with the definitions in Table 4 (e.g., *correct_mean* = 10/15 = 0.667 and *hesitation* = 0.3 × 4.27 = 1.28), confirming that the extraction function operates as designed. Empirical interaction data from the field trial were used only for the usability analysis (Section 3.6).

3.4. Classifier Performance on the Reference Dataset

Table 6 summarizes performance on the held-out test set of the reference dataset (n = 729). All evaluation metrics exceeded the predefined minimum targets: overall accuracy reached 87.3% (95% CI [84.8%, 89.6%]; 5-fold CV = 86.8% ± 1.4%; target ≥ 75%), macro recall reached 82.1% (95% CI [78.6%, 85.3%]; 5-fold CV = 81.5% ± 2.1%; target ≥ 80%), and macro F1-score reached 0.83 (95% CI [0.80, 0.86]; 5-fold CV = 0.82 ± 0.02; target ≥ 0.75), with a macro precision of 84.6% (95% CI [81.5%, 87.4%]; 5-fold CV = 84.1% ± 1.8%). All held-out test set results fall comfortably within the 95% confidence intervals and closely match the cross-validation ranges (within 1 standard deviation), confirming model stability and the absence of significant overfitting during training. These values describe performance on the reference dataset only and do not represent screening accuracy among Indonesian preschool children.

Table 6. Overall Performance of the Random Forest Model

Metric	Test Set	95% CI	5-Fold CV (Mean ± SD)	Target
Accuracy	87.3%	[84.8%, 89.6%]	86.8% ± 1.4%	≥75% (achieved)
Precision (macro)	84.6%	[81.5%, 87.4%]	84.1% ± 1.8%	
Recall (macro)	82.1%	[78.6%, 85.3%]	81.5% ± 2.1%	≥80% (achieved)
F1-score (macro)	0.83	[0.80, 0.86]	0.82 ± 0.02	≥0.75 (achieved)

The per-class results (Table 7) and the confusion matrix (Table 8) show that the model is most reliable for the Low Risk class (recall 91.5%, support n = 585), whereas recall is lower for the Moderate (76.8%, support n = 65) and High Risk (78.0%, support n = 79) classes. The lower recall for the Moderate class is expected because this intermediate category overlaps with both neighbouring classes. The High Risk recall of 78.0% means that out of 79 actual High Risk cases in the test set, 62 were correctly classified as High Risk, while 17 cases were missed by the classifier alone: specifically, 14 cases were misclassified into the adjacent Moderate Risk class, and 3 cases were misclassified into the Low Risk class. Likely contributing factors include the minority status of at-risk cases in the reference data (10.7% with dyslexia), which biases a classifier toward the majority class, and the derived nature of the three-level labels. Because missed at-risk children represent the most consequential error in screening, this result motivated the two-layer safety override evaluated in Section 3.5.

Table 7. Classification Results per Risk Class

Risk Class	Precision	Recall	F1-Score	Support (n)
Low (0)	89.2%	91.5%	0.90	585
Moderate (1)	81.3%	76.8%	0.79	65
High (2)	83.4%	78.0%	0.80	79

Table 8. Confusion Matrix on the Test Set

Actual \ Predicted	Low (0)	Moderate (1)	High (2)	Total Actual
Low (0)	535	42	8	585
Moderate (1)	11	50	4	65
High (2)	3	14	62	79

Table 9 lists the five most important features. *correct_mean* was the most important single feature (0.287), followed by *rt_mean* (0.198) and *hesitation* (0.156). However, because *hesitation* is computed as 0.3 × *rt_mean*, the two features are perfectly correlated, and *hesitation* provides no independent information; in Random Forest, impurity-based importance is shared among such correlated predictors [39]. The importance attributed to *hesitation* should therefore be read as part of the mean response-time signal. Interpreted in this way, the top-ranked features reflect two behavioural dimensions: response accuracy (*correct_mean*, 0.287; *correct_sum*, 0.089) and response speed (*rt_mean* and *hesitation* jointly ≈0.354; *rt_std*, 0.124).

Table 9. Feature Importance Analysis (Top Five)

Feature	Category	Importance
1. *correct_mean*	Behavioral	0.287
2. *rt_mean*	Temporal	0.198
3. *hesitation*	Derived from *rt_mean*	0.156
4. *rt_std*	Temporal	0.124
5. *correct_sum*	Behavioral	0.089

3.5. On-Device Integration and Safety Override

Conversion with hummingbird-ml transformed the forest into GEMM operations compatible with Unity Barracuda, and inference ran on the Android device's CPU (tested on a mid-range Android tablet, Samsung Galaxy Tab A7, Android 11 / API Level 30, averaged over 100 consecutive runs) with a mean latency of 42 ms (well below the 500 ms design threshold) without an internet connection. In the equivalence check, the predicted classes of the ONNX model agreed with those of the original Scikit-Learn model for 729 of 729 (100%) test samples, with a maximum absolute difference in predicted class probabilities of $< 1.0 \times 10^{-5}$. Table 10 compares performance with and without the safety override on the reference test set (n = 729).

Table 10. Effect of the Safety Override on the Reference Test Set

Indicator	Model Only	Model + Override	Net Change
High Risk recall	78.0% (62/79)	92.4% (73/79)	+14.4%
At-risk sensitivity (Moderate + High detected as ≥ Moderate)	90.3% (130/144)	97.2% (140/144)	+6.9%
Low-risk cases flagged as Moderate/High (false-positive rate)	8.5% (50/585)	13.7% (80/585)	+5.2%
Accuracy	87.3% (637/729)	85.6% (624/729)	-1.7%

The override increased High Risk recall from 78.0% (62 of 79 actual cases) to 92.4% (73 of 79 cases) and improved overall at-risk sensitivity from 90.3% to 97.2%, while the proportion of Low Risk cases flagged as Moderate or High Risk increased from 8.5% (50 of 585 cases) to 13.7% (80 of 585 cases), resulting in a slight reduction in overall accuracy from 87.3% to 85.6%. Because the override deliberately trades specificity for sensitivity to ensure potential reading difficulties are captured during early childhood, its output is communicated to users as "needs further observation" rather than as a definitive classification.

3.6. Functional Testing and Field Usability

Black Box testing of the profile-management module, game logic, JSON data-recording integrity, single-source audio system, and safety-override mechanism showed that all test scenarios passed without functional errors. The field usability results are summarized in Table 11. Fourteen of the 15 children (93.3%) understood the audio instructions and completed the mini-games independently; the single remaining child required brief verbal encouragement from the teacher during the initial setup due to mild shyness in the formal testing space, after which the child completed the sessions without

requiring re-explanation of the game mechanics. The mean completion time of 4.2 minutes (SD = 0.6 min; range = 3.5-5.1 min) fell well within the 3-5 minute targeted session length, conforming to young children's limited attention span. All three educators rated the traffic-light result display as intuitive (mean score = 4.8/5.0) for communicating potential risk to parents, and agreed strongly on the application's ease of operation, usefulness of recommendations, and overall developmental suitability for preschool children. These findings concern usability and functional feasibility only; they provide no evidence on the accuracy of risk classification for Indonesian children, because the children's results were not compared with any professional assessment.

Table 11. Field Usability Results at PAUD KB Al Fitriyah

Indicator	Result
Children aged 4 / 5 / 6 years	5 / 5 / 5 (Total = 15; 8 boys, 7 girls)
Children who completed the mini-games independently, without instructions being repeated	14 of 15 (93.3%)
Completion time of the four mini-games	Mean = 4.2 min (SD = 0.6 min; range = 3.5 5.1 min)
Educators rating traffic-light result display as intuitive	3 of 3 (100%; mean score = 4.8 / 5.0)
Educators rating ease of application operation	3 of 3 (100%; mean score = 4.7 / 5.0)
Educators rating usefulness of follow-up recommendations	3 of 3 (100%; mean score = 4.7 / 5.0)
Educators rating overall developmental suitability (ages 4 6)	3 of 3 (100%; mean score = 4.8 / 5.0)

3.7. Summary of Key Findings

In summary, (1) the prototype recorded interaction data and extracted features automatically without requiring reading ability; (2) on the reference dataset, the classifier exceeded all predefined targets, with response accuracy and response speed as the most informative features, although recall for the at-risk classes was lower than for the Low Risk class; (3) the converted model ran offline on-device; and (4) the application was functionally sound and usable by most children in the field trial.

4. DISCUSSION

4.1. Interpretation of the Main Findings

This study shows that a game-based prototype can measure four pre-reading predictors without requiring reading ability, extract interaction features automatically, and run a Random Forest classifier fully offline on a low-cost Android device. On the reference dataset, response accuracy and response speed were the most informative features. This pattern is theoretically plausible: accuracy on rhyme and letter tasks reflects the quality of phonological and visual-orthographic processing [7], [11], while response speed reflects the automaticity of retrieving phonological codes, the mechanism captured by RAN [8], [9]. The finding parallels that of Rauschenberger *et al.* [22], who identified reaction-time and click measures as key predictive features in a game-based screening tool.

4.2. Comparison with Previous Studies

The accuracy obtained here (87.3%) is numerically higher than the 79.8% accuracy of the tuned model reported for the same reference dataset [19] and the 0.74 accuracy reported by Rauschenberger *et al.* [22]. This comparison must be interpreted cautiously. Rello *et al.* addressed a binary classification problem with 10-fold cross-validation and lowered the decision threshold to favour recall of the dyslexia class, whereas this study used derived three-level labels, a single 80:20 hold-out split, and a different feature set. Differences in label definition, class balance, and evaluation protocol can change accuracy substantially; in particular, overall accuracy is inflated when the majority (Low Risk) class is predicted well, as reflected in the higher Low Risk recall (91.5%) relative to the at-risk classes. For screening purposes, macro recall and High Risk recall are therefore more relevant indicators than overall accuracy [6].

4.3. Cross-Linguistic and Contextual Factors

Using training data collected from Spanish-speaking readers aged 7 17 years to build a model for Indonesian pre-readers aged 4 6 years has several implications. Spanish and Indonesian both have transparent orthographies, in which phonological decoding is acquired relatively quickly [13], [29], so some behavioural signatures of dyslexia slower and less accurate phonological processing may be shared. Nevertheless, the tasks differ (reading-based exercises versus pre-literacy games), the age ranges do not overlap, and the base rate and labeling of dyslexia differ across educational systems. Children's responses in the mini-games may also be influenced by factors unrelated to dyslexia: age-related differences in processing speed; fine-motor control, because touch reaction time includes motor execution; familiarity with the Indonesian vocabulary used in the rhyme items, particularly for children who speak a regional language at home; and prior experience with touchscreen games. The age-adaptive item design partly addresses age effects, but the other factors were not controlled and could produce both false positives and false negatives.

4.4. Lower Recall for the Moderate and High Risk Classes

The lower recall for the Moderate (76.8%) and High Risk (78.0%) classes is the most important weakness of the classifier from a screening perspective, because these are precisely the children who would benefit most from early support. Two explanations are plausible. First, at-risk cases are a minority in the reference data, and a model optimized for overall performance tends to favour the majority class. Second, the boundaries between derived risk levels are inherently fuzzy, and children near these boundaries behave heterogeneously. Remedies to be explored include class weighting or resampling, threshold adjustment that prioritizes sensitivity as in the original reference study [19], and the use of class probabilities rather than hard labels. The safety override is a pragmatic interim measure, but it relies on a single accuracy threshold and increases false positives; it does not replace improving the sensitivity of the model itself.

4.5. Feature Importance, Neuropsychological Predictors, and the Limits of Behavioural Indicators

Feature importance in this study was computed over generic behavioural features shared across the four mini-games; it therefore indicates which kinds of behaviour are informative (accuracy and speed), not which neuropsychological predictor matters most. Linking features to predictors would require mini-game-specific features, for example, Rhyme Game accuracy as a phonological-awareness index and Rapid Naming Game speed as a RAN index, and an analysis of their separate contributions, preferably with conditional importance measures that account for correlated features [39] and after removing redundant features such as *hesitation*. Moreover, answer accuracy and reaction time are not specific to dyslexia: low accuracy or slow responses may also reflect attention difficulties, general language delay, unfamiliarity with the task, fatigue, low motivation, or sensory problems. These indicators can signal that a child needs closer attention, but they cannot by themselves identify the cause of the difficulty.

4.6. Practical, Theoretical, and Ethical Implications

For teachers and parents, the main practical value of the system lies in providing a structured and consistent prompt for closer observation. The traffic-light output can be interpreted as follows: green indicates no current need for follow-up beyond routine monitoring; yellow indicates that the child should be observed further and may benefit from targeted phonological activities; and red indicates that referral for professional assessment should be considered. Because the safety override deliberately shifts errors toward false positives, teachers should expect some children flagged yellow or red to show no difficulties on further assessment. A false-positive result can cause unnecessary parental anxiety, labeling, or stigma, whereas a false negative can delay support for a child who needs it [4]. To manage these risks, results should be communicated privately, framed as a risk indication rather than a diagnosis, accompanied by guidance on next steps, and never used for placement or exclusion decisions. The system is therefore best positioned as a supplementary screening aid within a broader process that includes teacher observation, parental input, and professional assessment, not as a diagnostic instrument [1], [4]. Theoretically, the study illustrates how established neuropsychological predictors of dyslexia can be operationalized as game-interaction features for children who cannot yet read.

4.7. Limitations and Future Research

This study has several limitations. First, the classifier was developed on an international reference dataset with derived risk labels and has not been validated among Indonesian preschool children. Second, the usability sample was small (15 children and 3 educators) and drawn from a single institution, so the usability findings cannot be generalized statistically. Third, the application's outputs were not compared with professional assessments, so its criterion validity is unknown. Fourth, the feature-extraction verification used simulated data, and the *hesitation* feature is redundant. Fifth, the mini-game weights and the override thresholds were set a priori rather than calibrated

empirically. Future research should collect interaction data from a larger, multi-site sample of Indonesian preschool children, with risk status established by standardized assessments adapted for Indonesian [13] and, ideally, longitudinal follow-up of reading outcomes in the early primary grades. Such data would allow the model to be retrained on in-context data, the weights and thresholds to be calibrated through ROC analysis, and sensitivity and specificity to be estimated directly, thereby improving both the reliability and the applicability of the model as a prerequisite for wider implementation.

5. CONCLUSION

This study set out to develop and evaluate an offline, game-based prototype for early dyslexia-risk screening of Indonesian preschool children. Technically, the objectives were achieved: four Indonesian-language mini-games measured phonological awareness, visual recognition, working memory, and rapid automatized naming without requiring reading ability; interaction features were extracted automatically; a Random Forest classifier was converted through hummingbird-ml to ONNX and executed on-device via Unity Barracuda; and the application functioned correctly and was used independently by most children in the field trial. On the international reference dataset, the classifier exceeded its performance targets, with response accuracy and response speed as the most informative features.

These results establish the feasibility of the application, not the validity of its dyslexia-risk predictions for Indonesian children. The classifier was trained on Spanish-language data from older children, the usability sample was small, and the outputs have not been compared with professional assessments; the system is therefore not clinically validated and should be used only as a supplementary pre-screening aid. The study contributes a reproducible pipeline for deploying tree-ensemble models in offline educational games and a transparent framework that separates model evaluation from field usability evidence. Future work should test the system with larger and more representative samples of Indonesian preschool children, establish criterion validity against standardized dyslexia assessments, and improve sensitivity for the High Risk class before broader implementation in PAUD settings.

ACKNOWLEDGEMENTS

The authors express their gratitude to Universitas Pamulang for institutional support and to the educators, staff, children, and parents of PAUD KB Al Fitriyah, Parung, Bogor Regency, for their cooperation and participation during the field trials of this research.

REFERENCES

[1] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Washington, DC: American Psychiatric Publishing, 2013.

[2] R. L. Peterson and B. F. Pennington, “Developmental dyslexia,” *Lancet*, vol. 379, no. 9830, pp. 1997–2007, 2012.

[3] T. Helland, F. Morken, and W. A. Helland, “Kindergarten screening tools filled out by parents and teachers targeting dyslexia: predictions and developmental trajectories from age 5 to age 15 years,”

Dyslexia, vol. 27, no. 4, pp. 413–435, 2021.

[4] J. Sanfilippo, M. Ness, Y. Petscher, L. Rappaport, B. Zuckerman, and N. Gaab, "Reintroducing dyslexia: early identification and implications for pediatric practice," *Pediatrics*, vol. 146, no. 1, p. e20193046, 2020.

[5] M. J. Snowling, "Early identification and interventions for dyslexia: a contemporary view," *J. Res. Spec. Educ. Needs*, vol. 13, no. 1, pp. 7–14, 2013.

[6] P. A. Thompson, C. Hulme, H. M. Nash, D. Gooch, E. Hayiou-Thomas, and M. J. Snowling, "Developmental dyslexia: Predicting individual risk," *J. Child Psychol. Psychiatry*, vol. 56, no. 9, pp. 976–987, 2015.

[7] M. Melby-Lervåg, S.-A. H. Lyster, and C. Hulme, "Phonological skills and their role in learning to read: A meta-analytic review," *Psychol. Bull.*, vol. 138, no. 2, pp. 322–352, 2012, doi: 10.1037/a0026744.

[8] E. S. Norton and M. Wolf, "Rapid Automatized Naming (RAN) and reading fluency: implications for understanding and treatment of reading disabilities," *Annu. Rev. Psychol.*, vol. 63, pp. 427–452, 2012.

[9] S. Araújo, A. Reis, K. M. Petersson, and L. Faísca, "Rapid automatized naming and reading performance: A meta-analysis," *J. Educ. Psychol.*, vol. 107, no. 3, pp. 868–883, 2015.

[10] H. L. Swanson, X. Zheng, and O. Jerman, "Working memory, short-term memory, and reading disabilities: A selective meta-analysis of the literature," *J. Learn. Disabil.*, vol. 42, no. 3, pp. 260–287, 2009.

[11] S. Franceschini, S. Gori, M. Ruffino, K. Pedrolli, and A. Facoetti, "A causal link between visual spatial attention and reading acquisition," *Curr. Biol.*, vol. 22, no. 9, pp. 814–819, 2012, doi: 10.1016/j.cub.2012.03.013.

[12] H. Winskel and V. Widjaja, "Phonological awareness, letter knowledge, and literacy development in Indonesian beginner readers and spellers," *Appl. Psycholinguist.*, vol. 28, no. 1, pp. 23–45, 2007.

[13] B. A. J. Jap, E. Borleffs, and B. A. M. Maassen, "Towards identifying dyslexia in Standard Indonesian: The development of a reading assessment battery," *Read. Writ.*, vol. 30, no. 8, pp. 1729–1751, 2017, doi: 10.1007/s11145-017-9748-y.

[14] L. R. Maghribi and E. A. Ismaya, "Analisis Hambatan Belajar Siswa dalam Mata Pelajaran IPS Kelas V SD IT Nurul Fikri Kecamatan Trangkil Pati," *JANACITTA J. Prim. Child. Educ.*, vol. 7, no. 2, pp. 97–104, 2024.

[15] A. Siswanto and H. T. A. Pratomo, "Skrining Kemampuan Phonological Awareness Anak Pra Sekolah," *Interes. J. Ilmu Kesehat.*, vol. 8, no. 2, pp. 153–157, 2019.

[16] P. Lameras, S. Arnab, I. Dunwell, C. Stewart, S. Clarke, and P. Petridis, "Essential features of serious games design in higher education: linking learning attributes to game mechanics," *Br. J. Educ. Technol.*, vol. 48, no. 4, pp. 972–994, 2017.

[17] O. L. Usman, R. C. Muniyandi, K. Omar, and M. Mohamad, "Advance machine learning methods for dyslexia biomarker detection: A review of implementation details and challenges," *IEEE Access*, vol. 9, pp. 36879–36897, 2021.

[18] A. Alrubaian, "Exploring and Identifying Key Factors in Predicting Dyslexia in Children: Advanced Machine Learning Algorithms From Screening to Diagnosis," *Clin. Psychol. Psychother.*, vol. 32, no. 3, p. e70077, 2025.

[19] L. Rello, R. Baeza-Yates, A. Ali, J. P. Bigham, and M. Serra, "Predicting risk of dyslexia with an online gamified test," *PLoS One*, vol. 15, no. 12, p. e0241687, 2020.

[20] K.-Y. Fung *et al.*, "ADPS-A Pre-screening Tool for Students With Dyslexia in Learning Traditional Chinese," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 1454–1470, 2024.

[21] O. Gaggi *et al.*, "Serious games for early identification of developmental dyslexia," *Comput. Entertain.*, vol. 15, no. 2, pp. 1–24, 2017, doi: 10.1145/2629558.

[22] M. Rauschenberger, R. Baeza-Yates, and L. Rello, "A universal screening tool for dyslexia by a web-game and machine learning," *Front. Comput. Sci.*, vol. 3, p. 628634, 2022.

[23] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real-world classification problems?" *J. Mach. Learn. Res.*, vol. 15, pp. 3133–3181, 2014.

[24] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: a large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 1, p. 270, 2018.
[25] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[26] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 3, p. e1301, 2019.
[27] S. Nakandala *et al.*, "A Tensor Compiler for Unified Machine Learning Prediction Serving," in *Proceedings of the 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2020)*, Berkeley, CA: USENIX Association, 2020.
[28] D. Rasapta and Septa, "Sistem Diagnosa Penyakit Musang dengan Metode Certainty Factor Berbasis Web," *J. Artif. Intell. Innov. Appl.*, vol. 3, no. 1, pp. 47–52, 2022.
[29] J. C. Ziegler and U. Goswami, "Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory," *Psychol. Bull.*, vol. 131, no. 1, pp. 3–29, 2005.
[30] K. P. Rahayu, L. R. Maghribi, and D. A. S. Pratiwi, "Peningkatan Literasi Matematika dan Sains di Sekolah Dasar dengan Menggunakan Fitur AI," *JANACITTA J. Prim. Child. Educ.*, vol. 8, no. 2, pp. 23–31, 2025.
[31] L. R. Maghribi, D. Mutakin, and P. Novita, "Persepsi Guru dan Siswa terhadap Penggunaan Fitur AI dalam Pengembangan Literasi PKn dan IPS di Sekolah Dasar," *JANACITTA J. Prim. Child. Educ.*, vol. 9, no. 1, pp. 108–117, 2026.
[32] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021.
[33] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[34] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023.
[35] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, p. 91, 2006.
[36] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall/CRC, 1993.
[37] U. Technologies, "Unity Documentation: Scripting Reference," Sep. 2024. [Online]. Available: https://docs.unity3d.com/ScriptReference/
[38] R. Indonesia, *Undang-Undang Republik Indonesia Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi*. Jakarta, 2022.
[39] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, p. 307, 2008.