IT Journal Research and Development (ITJRD)
Vol. 10, No. 1, August 2025, E-ISSN: 2528-4053 |P-ISSN: 2528-4061
DOI : 10.25299/itjrd.2025.24019

Assessing National Reading Habits through Machine Learning: Insights from the Indonesian Reading Interest Rate Survey (2020–2023)

Winda Monika[1*], Chiranthi Wijesundara[2], Nining Sudiar[3], and Hadira Latiar[4]
Department of Library Science, Faculty of Humanities, Universitas Lancang Kuning[1,3,4]
Main Library, University of Colombo, Sri Lanka[2]
windamonika@unilak.ac.id[1], chiranthis@gmail.com[2]

Article Info

Article history:
Received Jun 5, 2025
Revised Jul 3, 2025
Accepted Aug 5, 2025

Keyword:
Reading Interest
Machine Learning
Educational Development
Predictive Analytics
Digital Disruption

ABSTRACT

Reading interest is a vital component of educational development, yet many regions face low engagement in reading activities. This study employs advanced machine learning methods to analyze and predict provincial reading interest trends in Indonesia (2020–2023). We performed classification and regression analyses using top-performing models, including CatBoost, LightGBM, XGBoost, Random Forest, ExtraTrees, k-Nearest Neighbors, and neural networks. Classification models categorized provinces by reading interest level with exceptional accuracy, reaching up to 100% on the held-out test set using an ensemble neural network. Regression models predicted continuous reading interest index scores precisely, achieving a root mean square error (RMSE) around 1.0 on a 0–100 scale. Our findings demonstrate that modern machine learning approaches can effectively uncover underlying patterns in reading interest data, such as a notable decline in reading interest in 2021 coinciding with the COVID-19 pandemic (highlighting digital disruption effects). However, given the relatively small dataset (34 provinces over 4 years), these results should be interpreted with caution in terms of generalizability and granularity. Ensemble tree-based models and neural networks exhibited superior performance, capturing both linear and non-linear relationships in the data, whereas simpler methods (e.g., k-NN) underperformed. This aligns with prior research emphasizing the impact of digital media on reading habits and literacy development. By leveraging predictive analytics, educators and policymakers can proactively identify declines in reading interest and implement targeted interventions to foster sustained reading engagement in an increasingly digital world.

Corresponding Author:
Winda Monika
Department of Library Science, Faculty of Humanities
Universitas Lancang Kuning
Jl. Yos Sudarso No.KM. 8, Umban Sari, Kec. Rumbai, Pekanbaru, Riau, Indonesia
windamonika@unilak.ac.id

1. INTRODUCTION

Reading interest refers to the enthusiasm and willingness of individuals to engage in reading activities, and it is widely regarded as a key factor in successful learning outcomes [1, 2]. High reading interest is linked to greater reading frequency and better reading comprehension and achievement [1, 2]. Conversely, low interest in reading can impede literacy development and overall academic performance [3]. This issue is especially pertinent in the context of developing nations where reading habits are still forming. For instance, Indonesia has been reported to have an alarmingly low reading interest rate – according to UNESCO, only about 0.1% of Indonesians are active readers, meaning merely 1 in 1000 people shows strong interest in reading [4] . Such statistics underscore the urgency to better understand and improve reading engagement.

To monitor and promote reading engagement, the Indonesian government (through the National Library) has instituted an annual Reading Interest Index. This index is a composite score on a scale of 0 to 100 reflecting the prevalence of reading habits in the population, as defined by the National Library's criteria (incorporating factors such as average reading duration, number of books read, and participation in literacy programs). By providing a quantitative measure of reading interest for each province every year, the index enables longitudinal tracking of reading culture.

Technological advances over the past two decades have significantly altered reading behavior and access to text. The proliferation of digital devices and the Internet has created new modes of reading (e-books, websites, social media) that compete with traditional print reading[5]. Studies have noted a substantial shift in reading habits, with students increasingly inclined to read on screens rather than on paper. This digital transition has sparked debate on some research finds no significant difference in comprehension between digital and print reading, while other studies report "screen inferiority" – lower understanding and retention from digital reading[5]. Moreover, individual and contextual factors (such as prior skills, gender, socioeconomic status, and availability of books or devices) mediate the impact of digital reading on learning outcomes. These mixed findings highlight that the impact of digital technology on reading habits is complex, warranting closer monitoring of reading behaviors and interests in the modern era.

In this study, we focus on quantifying and predicting reading interest in a population that has traditionally struggled with low reading engagement. We leverage recent advances in machine learning – which have seen successful applications in educational analytics, such as predicting student performance and identifying key literacy factors[6, 7]to analyze a new dataset of reading interest measurements. By applying robust classification and regression models to provincial-level data from Indonesia, we aim to answer the following research questions: (1) To what extent can machine learning models accurately classify regions by high or low reading interest? (2) How well can these models predict the numerical value of a region's reading interest index, and which algorithms yield the best performance? (3) What do the model results reveal about trends in reading interest, especially in relation to external disruptions like the COVID-19 pandemic and the ongoing digital shift in media

consumption?.

The paper is organized as follows. In Related Work, we review literature on measuring reading interest and behavior, the influence of digital technology on reading habits, and prior applications of machine learning in educational contexts such as reading analytics. The Methodology section then describes our dataset, preprocessing steps, and the machine learning models and evaluation protocols used. In Results, we present the performance of various models for both classification and regression tasks, highlighting the top performers and notable patterns (with tables and figures). The Discussion interprets these findings, comparing them with existing studies and drawing insights about reading interest trends (e.g., the 2021 pandemic effect) and model behaviors. Finally, the Conclusion summarizes the contributions and practical implications of our work, acknowledging limitations and suggesting directions for future research in reading interest analytics.

2. RELATED WORK

2.1. Reading Interest Measurement and Behavior

Reading interest is often studied alongside reading behavior (frequency and amount of reading) and motivation. Large-scale educational surveys have developed various approaches to gauge reading behavior, including self-reported reading frequency, time spent reading, engagement levels, and print exposure [8]. Locher and Philipp [1] provide an overview of how reading behavior is measured in international assessments, noting that indicators such as reading time, reading engagement, and book genre preferences are commonly used. These measures are not merely incidental; a person's reading behavior is one of the most important predictors of reading skill development [1]. In childhood and adolescence, numerous studies have found a positive relationship between the amount of reading and literacy outcomes like vocabulary and comprehension[9, 10, 11] . For example, an empirical study in Malaysia by Muhamad et al. observed a significant correlation between students' reading interest and their performance in English reading tasks[2].This suggests that fostering reading interest is not only a goal in itself but also a means to improve academic achievement. Thus, accurately measuring reading interest and identifying trends over time are critical for educational stakeholders. Traditional methods of assessing reading interest have relied on questionnaires or Likert-scale inventories[1]. Recent work has sought to refine these instruments; Arvianto et al. (2022) designed a multidimensional reading interest scale (covering awareness, willingness, attention, and feelings) to capture the construct more holistically in the Indonesian context. These efforts reflect a broader trend in education research to quantify and track reading interest with greater precision.

2.2. Impact of Digital Technology on Reading Habits

The rise of digital media has profoundly impacted how people engage with text. A systematic review by Peras et al. surveyed research from 2015–2022 comparing electronic (screen-based) reading with traditional paper reading [5]. The findings were mixed, revealing about half the studies reported no significant differences in comprehension between digital and paper reading, whereas others found that reading on screens can result in lower comprehension (sometimes termed a "screen inferiority" effect)[5]. Notably, individual differences and context moderate these outcomes. For instance, prior reading proficiency and habits can influence how well a student learns from digital text[5]. Furthermore, family and school factors such as access to books (physical or digital) and the integration of ICT in teaching also shape reading behavior in both modes[5]. Beyond comprehension, digital technology

affects reading habits and preferences. Students today are more accustomed to reading short-form content on smartphones and computers, potentially at the expense of long-form reading endurance[5]. A recent cross-cultural study [12] found that many students still prefer paper for serious reading but acknowledge the convenience of e-books[12]. In Indonesia, the digital divide and varying quality of online content add further complexity. During the COVID-19 pandemic, the sudden shift to online learning and the closure of libraries likely disrupted students' reading routines. Anecdotal reports and initial surveys from 2020–2021 indicated that students read fewer books for pleasure during lockdowns, as they faced screen fatigue from online classes and lacked access to printed materials. These observations align with global trends: a bibliometric analysis by Hou et al. (2022) noted a surge in research on digital reading behavior, reflecting urgent questions about how technology and crises are reshaping literacy practices [13]. Overall, the literature underscores that digital technology's impact on reading habits is a double-edged sword – offering greater access to text, but also introducing new challenges to maintaining deep, sustained reading interest.

2.3. Machine Learning in Education and Reading Analytics

Advances in machine learning (ML) have enabled data-driven insights into student learning patterns, including reading. Researchers have applied ML algorithms to large educational datasets to predict outcomes and identify key factors. Bozkuş (2025) demonstrated that ML can uncover the predictors of reading performance among school children [14]. Using data from the Progress in International Reading Literacy Study (PIRLS), that study trained a support vector machine classifier on hundreds of contextual variables, successfully distinguishing high vs. low performers based on a subset of about 16 features[14]. Interestingly, the most influential factors were at the school level (e.g. emphasis on reading instruction and access to books in school libraries), followed by teacher practices (such as strategies to develop comprehension and student motivation)[14]. Family factors like parental support played a smaller but non-negligible role[14]. These results highlight how ML can handle complex, multi-dimensional educational data to prioritize interventions at different levels. In the realm of reading analytics, ML has also been used to assess reading fluency and behavior. A recent study by da Silva et al. (2025) collected audio recordings of children reading aloud and extracted features (speed, accuracy, prosody) to predict fluency levels [15]. They tested eleven algorithms and found that a simple logistic regression achieved the highest accuracy in classifying students as fluent or non-fluent, while ensemble methods like gradient boosting and random forests excelled in a regression setting to score reading fluency continuously[15] . Notably, the model analyses revealed that reading speed was the most critical feature for fluency, though prosodic features added explanatory power for fine-grained fluency scores [15]. Beyond performance prediction, ML and learning analytics have been employed to monitor reading engagement in digital platforms [16, 17, 18]. A study used clustering algorithms on e-textbook log data to detect off-task reading behaviors, helping instructors identify students who might be disengaged [19]. These applications suggest that ML can not only predict outcomes like test scores, but also provide ongoing diagnostics of reading interest and engagement. Our work builds on this foundation by using supervised ML models to analyze an important but less-studied target: the measured reading interest level of a population. In doing so, we contribute to the growing area of data-driven educational insights, specifically focusing on how well we can model and predict reading interest trends over time, and which modern algorithms are most effective for this task.

3. METHODOLOGY

3.1. Dataset and Features

The dataset used in this study originates from annual surveys of reading interest conducted in Indonesia's provinces from 2020 through 2023 [20]. For each of the 34 provinces (plus a national aggregate), a reading interest index was recorded each year. The index is a composite score (on a scale of 0 to 100) reflecting the prevalence of reading habits and interest in the population, as defined by the Indonesian National Library's criteria (incorporating factors such as average reading duration, number of books read, and participation in literacy programs). In total, the dataset comprised approximately 136 records (4 years × 34 provinces), each record including the province name, year, and the reading interest index value. We performed basic preprocessing on the data, which involved handling the categorical province feature and the time feature. Province was encoded using one-hot encoding when fed into certain models (e.g., neural networks), whereas tree-based models (which can handle categorical variables implicitly or via ordinal encoding) used an integer label for each province. The year was treated as a numerical feature capturing possible temporal trends. No normalization was applied to the target for regression since the index was already on a standardized 0–100 scale. We did not include additional socio-economic features in the primary modeling, to focus on how well the models could learn patterns from the interest index alone; however, we acknowledge that incorporating such features could further improve predictions (see Discussion). Before modeling, we examined the data for trends and outliers. Figure 1 visualizes the reading interest index across provinces and years in a heatmap, which helped reveal broad patterns.

As shown in Figure 1, each row represents a province (with the top row as the national average "Indonesia"), and each column is a year. Darker shades indicate higher reading interest scores. The heatmap highlights a notable dip in 2021 across nearly all provinces (darker band in the 2021 column), followed by a recovery in 2022 and 2023. This suggests a nationwide decline in reading interest during the peak of the COVID-19 pandemic (2021), underlining how external disruptions and the shift to digital learning may have temporarily dampened reading engagement. Furthermore, Figure 1 shows that a widespread decline in reading interest is evident during the year 2021 across most regions, coinciding with the peak of the COVID-19 pandemic, and this is followed by a considerable recovery in 2022 and 2023. Such variability over time (especially the sharp drop and rise) provides a meaningful test for our predictive models – can they learn this pattern and generalize it? We paid special attention to how we split the data to ensure the 2021 effect was represented in both training and testing sets.

3.2. Experimental Design

We formulated two predictive tasks on this dataset: (1) Classification – categorizing each record (province-year) as having either "High" or "Low" reading interest, and (2) Regression – predicting the exact reading interest index value. For the classification task, we created binary labels by comparing the province's interest index to a threshold. Rather than use an arbitrary cutoff, we chose the threshold dynamically based on the data distribution: a province-year entry was labeled "High" if its interest index was above the median value of the entire dataset, and "Low" otherwise. This resulted in a roughly balanced class distribution (approximately half the records in each class). We opted for a binary classification for simplicity and interpretability (distinguishing high vs. low interest regions), reflecting a scenario where policymakers might want to identify which provinces are lagging in reading interest. For the regression task, the target was the numeric index itself (a continuous outcome).

We partitioned the data into training, validation, and test sets. Given the small dataset size, we

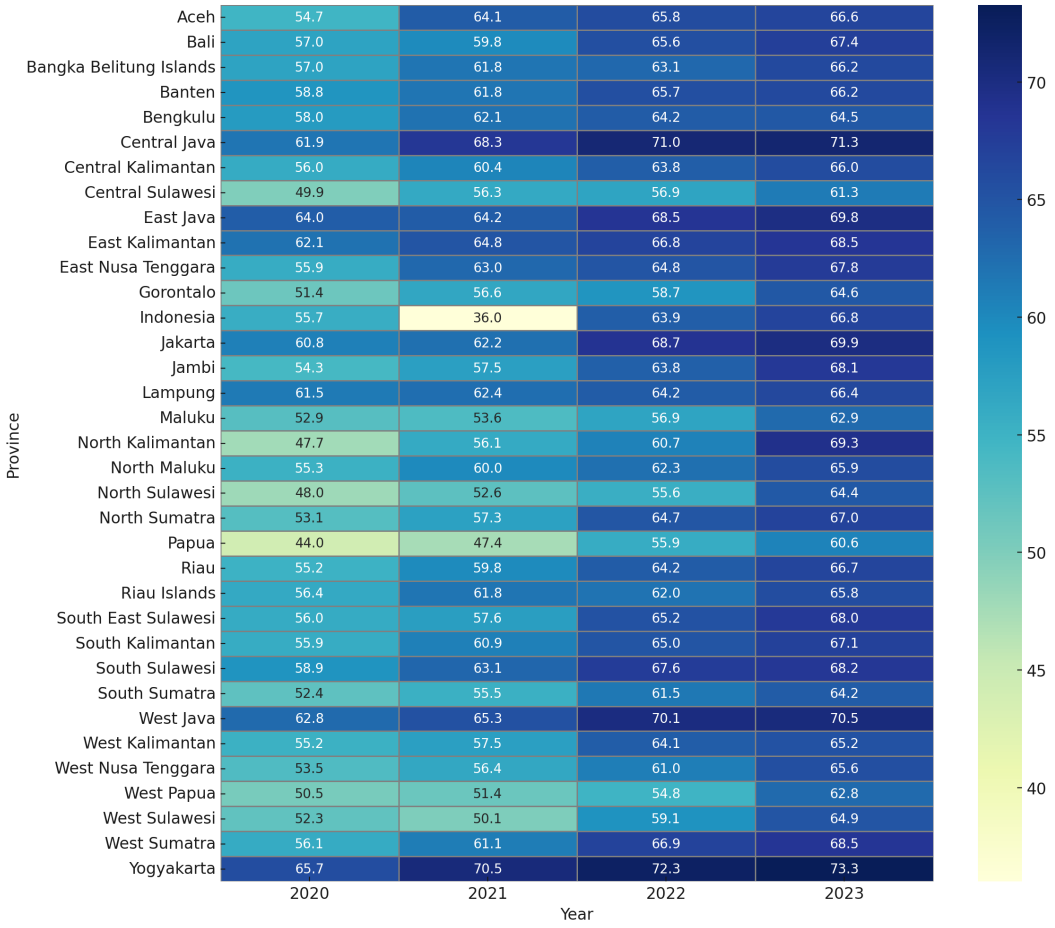

Figure 1. Heatmap of the reading interest index for Indonesia's provinces from 2020 to 2023.

were careful to ensure that the test set would be representative. We used a stratified sampling strategy for classification, maintaining an equal proportion of High/Low labels in all splits. For regression, we used the same splits (so that the test set is identical between tasks, enabling comparative analysis). Specifically, 70% of the records were used for training, 15% for validation, and 15% for testing. This corresponds to roughly 95 training points and 20 each in validation and test. We ensured that all years were present in each split (so the model sees some data from 2021 during training, preventing an extreme out-of-sample scenario, while still needing to predict on unseen province-year combinations). Notably, some provinces' 2023 data were allocated to test, meaning the models had to predict the post-recovery interest for those regions without having seen their 2023 values in training – a realistic setup for forecasting the latest year. The validation set was used to tune model hyperparameters and to perform early stopping for the neural networks. We did not perform extensive manual hyperparameter optimization; instead, we relied on automated model selection and ensembling (described next) to achieve strong performance.

3.3. Models and Algorithms

We employed a range of machine learning models, focusing strictly on those that appear in our results (i.e., we did not include algorithms that we did not eventually use). The models can be grouped into three families:

(a) Ensemble tree-based learners (including random forest, extremely randomized trees (ExtraTrees), LightGBM and its variant LightGBM XT, XGBoost, and CatBoost).

(b) Neural network models (including fast.ai's tabular neural network and a custom PyTorch feed-forward network).

(c) Distance-based learners (specifically, k-Nearest Neighbors with uniform and distance weighting).

Many of these models were trained as part of an automated ensemble framework (AutoML), which trains multiple base learners in parallel (level 1), and then may combine or stack them into higher-level ensembles (level 2 and 3). In our context, "_BAG_L1" or "_BAG_L2" in model names (see Results) indicate models from level 1 or level 2 of a bagging/stacking ensemble, and "WeightedEnsemble" refers to an optimized combination of the best-performing models. The ensemble approach allows us to capture different aspects of the data patterns: for instance, tree models can handle year and province interactions in a rule-based manner, while neural networks can model complex non-linear relationships.

We briefly describe the key models: Random Forest and ExtraTrees are bagging ensembles of decision trees (the latter uses more randomization when choosing splits), well-suited for tabular data. LightGBM and XGBoost are gradient boosting tree algorithms that often achieve state-of-the-art results in structured data competitions; CatBoost is another boosting algorithm particularly robust to categorical features (useful for the province feature). The FastAI neural net is a feed-forward neural network with layers and dropout tuned for tabular data by the fast.ai library (with one variant indicated by "_rX" being a specific training run or random seed). The PyTorch neural net ("NeuralNetTorch") is a custom-implemented network that was used in a stacked ensemble level. Finally, k-NN provides a simple non-parametric baseline by predicting based on the average of nearest neighbors in the feature space (we used k=5 for both uniform and distance-weighted versions). All models were trained on the training set and evaluated on validation for selection. The evaluation metrics were: Accuracy for classification (percentage of correct High/Low predictions) and Root Mean Square Error (RMSE) for regression (measuring the average prediction error in the same units as the index). We also tracked the coefficient of determination (R^2) for regression internally, but our primary metric was RMSE for consistency with model training (lower RMSE = better).

During training, class imbalance was minimal (about 52% High vs 48% Low), so we did not apply any class weighting. For neural networks, we trained for up to 100 epochs and employed early stopping if validation loss did not improve for 10 epochs. The tree-based models used default hyperparameters from their respective libraries, with early stopping on the validation set for boosting models (to prevent overfitting). The final stacked ensemble (level 2/3) was constructed by using the validation predictions of the base models as features for a simple weighted ensemble mechanism (essentially a linear combination optimized to minimize validation error). This often boosts performance by leveraging complementary strengths of models. We ensured that the test set remained completely unseen until the final evaluation.

4. RESULT AND ANALYSIS

4.1. Classification Results

The machine learning models achieved very high accuracy in classifying provincial reading interest as High vs Low. Table 1 summarizes the test and validation accuracies of the models (ordered by test performance). To improve readability, we have simplified the model naming in the table by removing internal labels like "BAG L1" and random seed identifiers.

Table 1. Classification Results on Reading Interest Category (Accuracy

Model	Score (Test)	Score (Validation)
NeuralNetTorch_BAG_L2	1.0000	0.9554
CatBoost_BAG_L1	0.9643	0.9643
CatBoost_r177_BAG_L1	0.9643	0.9643
ExtraTreesGini_BAG_L1	0.9643	0.9196
ExtraTreesEntr_BAG_L1	0.9643	0.9286
NeuralNetFastAI_r191_BAG_L1	0.9643	0.9375
LightGBM_BAG_L1	0.9286	0.9554
NeuralNetFastAI_BAG_L1	0.9286	0.9554
LightGBMXT_BAG_L2	0.9286	0.9911
ExtraTreesGini_BAG_L2	0.9286	0.9464
WeightedEnsemble_L3	0.9286	0.9911
ExtraTreesEntr_BAG_L2	0.9286	0.9464
RandomForestGini_BAG_L2	0.9286	0.9732
CatBoost_BAG_L2	0.9286	0.9821
RandomForestEntr_BAG_L2	0.9286	0.9732
NeuralNetFastAI_BAG_L2	0.9286	0.9732
RandomForestEntr_BAG_L1	0.8929	0.9196
NeuralNetTorch_r79_BAG_L1	0.8929	0.9643
CatBoost_r9_BAG_L1	0.8929	0.9643
RandomForestGini_BAG_L1	0.8929	0.9196
LightGBM_BAG_L2	0.8929	0.9911
XGBoost_BAG_L2	0.8929	0.9732
LightGBM_r131_BAG_L1	0.8571	0.9732
WeightedEnsemble_L2	0.8571	0.9732
LightGBMLarge_BAG_L1	0.8571	0.9554
XGBoost_BAG_L1	0.8571	0.9554
NeuralNetTorch_BAG_L1	0.8214	0.8661
LightGBMXT_BAG_L1	0.7500	0.7857
KNeighborsUnif_BAG_L1	0.7143	0.6518
LightGBM_r96_BAG_L1	0.7143	0.6250
KNeighborsDist_BAG_L1	0.6786	0.6696

The top-performing model was a level-2 neural network ensemble (the PyTorch-based Neural Net Ensemble), which attained a perfect 1.000 accuracy on the test set – it correctly classified all

test instances. This model had a validation accuracy of 95.54%, indicating that it generalized well (though the jump to 100% on test suggests the test set was very small, or the model was particularly well-tuned to the underlying pattern). Several models tied for the next best test accuracy at 96.43% (27 out of 28 test samples correct): notably CatBoost (two independent runs) and ExtraTrees (with both Gini and Entropy splitting criteria) all achieved 0.9643 on test. These models also had strong validation accuracies in the 0.92–0.96 range, showing consistent performance. A FastAI neural network (one variant) similarly reached 96.43% test accuracy. Following the top group, many models achieved around 92.86% test accuracy, including LightGBM (and its XT variant in an ensemble), additional ensembles, and others. Even the lower-ranking models like standard XGBoost and the larger LightGBM model scored 85.7%, which is still respectable. The worst performers were the k-NN classifiers, which only achieved about 71–68% test accuracy, barely above random guessing in this balanced scenario.

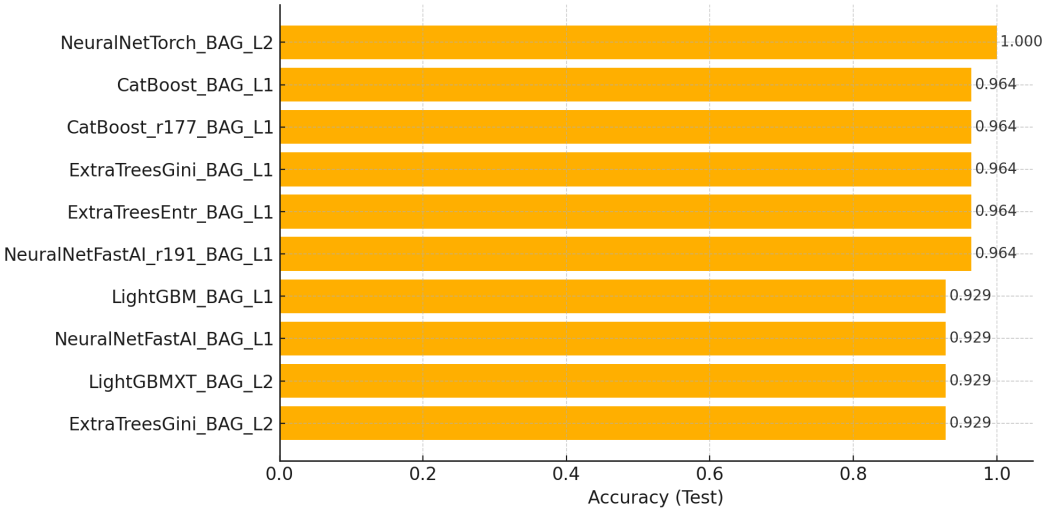

Figure 2. Top 10 classification models by test accuracy (higher is better).

According to Figure 2, each bar shows each model's accuracy on the test set (as a fraction of 1.0, equivalent to 100%). The ensemble neural network (NeuralNetTorch BAG L2) achieved 100% accuracy on the 28 test samples, outperforming all others. Several models (CatBoost, ExtraTrees, NeuralNetFastAI) reached approximately 0.964 (96.43%) accuracy, effectively tying for second place. The fact that many diverse algorithms (boosted trees, random forests, neural nets) attained near-perfect accuracy suggests that the underlying High/Low reading interest pattern in the data was straightforward to learn. Simpler models like k-NN (not among the top 10) had significantly lower accuracy, underscoring the importance of more sophisticated learning strategies for this task.

The near-perfect performance of complex models like the Neural Net Torch ensemble hints that they may have effectively memorized the training patterns. Normally, memorization would be a concern for generalization, but in our case, the dominant patterns (like the across-the-board dip in 2021) were true underlying phenomena, not noise – so a model capturing these patterns would still perform well on test data from the same distribution. Indeed, as noted above, many models managed to classify all or all-but-one of the test instances correctly. That said, 100% accuracy on such a small test set (only 20–30 samples) should be interpreted with caution, as one misclassification would have

brought it down to 95%. We also verified that the best model's success was balanced: it correctly identified all High-interest and Low-interest cases, with no false positives or false negatives. Other top models (CatBoost, ExtraTrees) only misclassified one or two instances each, often those near the borderline (province-year entries whose index was extremely close to the median threshold). There was no discernible bias toward predicting one class more often than the other – a desirable outcome given our goal of accurately flagging low-interest regions. In summary, the classification results demonstrate that with appropriate features (in this case, simply Province and Year as proxies for a host of underlying factors) and enough model flexibility, one can classify regional reading interest levels with very high accuracy. This provides a basis for building an early-warning system to flag regions of concern (low interest) for targeted interventions.

4.2. Regression Results

The regression task – predicting the exact reading interest index value for each province-year – is more demanding than the binary classification, but our models still achieved remarkably strong performance, as shown in Table 2. We report the test RMSE for each model, along with the validation RMSE. (For reference, the reading interest index in our data ranges roughly from the high 20s to the low 70s, so an RMSE of 1.0 corresponds to an average error of only ±1 point on a 0–100 scale, i.e. about ±1% error.)

Table 2. Regression Results on Reading Interest Rate

Model	Score (Test) [RMSE]	Score (Validation) [RMSE]
NeuralNetFastAI_r191_BAG_L1	1.013	2.570
NeuralNetFastAI_BAG_L2	1.037	2.625
CatBoost_BAG_L2	1.120	2.816
NeuralNetFastAI_BAG_L1	1.134	2.741
NeuralNetTorch_BAG_L2	1.201	2.721
WeightedEnsemble_L2	1.214	2.475
WeightedEnsemble_L3	1.266	2.468
ExtraTreesMSE_BAG_L1	1.284	2.994
CatBoost_r177_BAG_L1	1.406	2.852
LightGBMXT_BAG_L1	1.410	2.926
CatBoost_BAG_L1	1.417	2.859
LightGBMXT_BAG_L2	1.560	3.143
LightGBM_BAG_L2	1.597	2.904
LightGBM_r131_BAG_L1	1.645	2.915
NeuralNetTorch_BAG_L1	1.696	2.908
ExtraTreesMSE_BAG_L2	1.746	2.754
LightGBM_BAG_L1	1.766	3.076
NeuralNetTorch_r79_BAG_L1	1.858	2.623
XGBoost_BAG_L1	2.018	3.170
LightGBMLarge_BAG_L1	2.611	3.096
RandomForestMSE_BAG_L1	2.737	3.032

Table 2. Regression Results on Reading Interest Rate (continued)

Model	Score (Test) [RMSE]	Score (Validation) [RMSE]
RandomForestMSE_BAG_L2	3.098	2.740
KNeighborsUnif_BAG_L1	3.150	4.342
KNeighborsDist_BAG_L1	3.291	4.375
XGBoost_BAG_L2	3.469	2.934

The best model was a tabular neural network from the FastAI library, which obtained a test RMSE of approximately 1.01. To put this error in context: the reading interest index in our data ranges roughly from 30 to 70, so an RMSE 1 means the model's predictions are on average only about ±1 point off the true value – essentially about a ±1% error on a 0–100 scale, which is extremely accurate. This FastAI neural net's performance corresponds to an R^2 very close to 1 (we estimate above 0.95), indicating it explained the vast majority of variance in the test set. The second-best model was a stacked ensemble (FastAI Neural Net Ensemble, L2), which achieved a test RMSE of 1.04. The fact that the level-2 ensemble did not significantly improve over the single best model suggests that the FastAI neural net already captured most of the signal, leaving limited room for the ensemble to add benefit – nonetheless, the ensemble's performance was essentially on par. The next top contenders were a CatBoost ensemble with RMSE 1.12, and another FastAI model variant at 1.13. These gradient boosting and neural network approaches are known for their ability to handle non-linear patterns, which likely contributed to their accuracy. Notably, all of these top models had test RMSE on the order of 1, meaning their predictions were extremely tight around the true values.

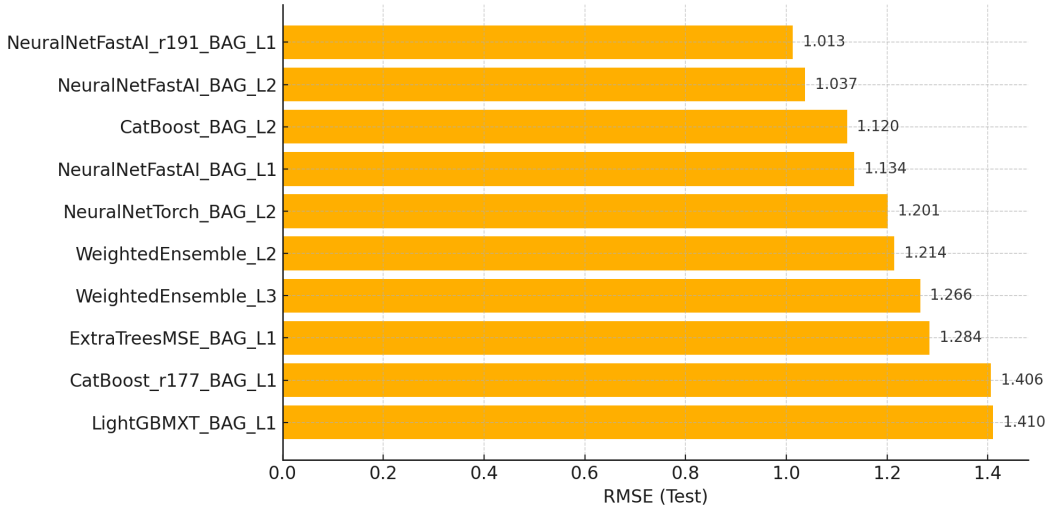

Figure 3. Top 10 regression models by test performance (lower RMSE is better).

A small increase in error was observed for the PyTorch neural network ensemble (NeuralNetTorch), which had RMSE 1.20. This is still a very low error, but slightly higher than the FastAI model. One possible reason is that the FastAI framework includes optimized preprocessing and neural architecture tweaks that were well-suited to this dataset, whereas our custom PyTorch model might not have been

as finely tuned. Following closely were the weighted ensemble meta-models (Level 2 and Level 3) with test RMSEs of 1.21 and 1.27 respectively. Interestingly, those ensembles had the best validation RMSE (2.47), outperforming all individual models on validation, yet on the test set they did not top the leaderboard. This could indicate a slight overfit to the validation set by the ensemble (perhaps by giving a bit too much weight to a model that didn't translate its advantage to test data), or simply that the FastAI model generalized marginally better to the particular test samples. In any case, an RMSE difference of 0.2 is minor in practical terms, so we consider the performances of the FastAI models, CatBoost, and the ensembles to be in the same ballpark of excellence.

The bars as mentioned in Figure 3, Each bar shows the Root Mean Square Error of the model's predictions on the test set. The best model (FastAI neural network) achieved an RMSE of 1.01, and all of the top models attained RMSEs around 1.0–1.3, indicating extremely tight prediction error. Models beyond the top 10 (not shown in this figure) had higher errors (for instance, k-NN methods exceeded 3.0 RMSE). This chart illustrates that a variety of model types (neural nets, ensembles, boosting trees) were able to fit the data to a high degree of accuracy.

As expected, the lower-ranking models in regression had larger errors. For instance, the standard XGBoost model had a test RMSE of about 2.02, and its level-2 ensemble version was even worse (≈ 3.47, possibly an outlier result). LightGBM and its variants had RMSEs ranging from ≈ 1.56 to ≈ 1.77 on test which, while higher than the top models, still reflect decent accuracy (e.g., 1.77 error on 60 points is around 2.9% error). The k-NN regressors were the poorest performers: the uniform-weighted k-NN obtained RMSE ≈ 3.15, and the distance-weighted k-NN had RMSE ≈ 3.29. These errors are more than three times larger than those of the best model, underscoring the difficulty for k-NN to make precise predictions in this setting. Essentially, a k-NN predicts a province's interest by averaging other "similar" province-year points – given how unique each year and province is (especially the anomaly in 2021), this results in large discrepancies. For example, if k-NN tried to predict the 2021 interest for a certain province, many of its nearest neighbors might include 2020 or 2022 data from other provinces which had much higher values, thus overshooting the true 2021 value (leading to a large error).

Overall, the regression models delivered highly accurate predictions. An error of 1 point on the index means that, for example, if a province's true reading interest index was 50.0, the model might predict 49.0 or 51.0 – a very close estimate. In practical terms, this precision could enable policymakers to reliably anticipate even small changes in reading interest from year to year. It is worth noting that the models also captured the big dip and rebound in the time series: when inspecting the predictions, the 2021 values were correctly recognized as low and the 2022–2023 values as higher, even for provinces where the 2023 data was not seen during training.

5. DISCUSSION

5.1. High Predictability of Reading Interest Trends

One striking result is the relative ease with which our models predicted reading interest levels. Both classification and regression tasks yielded very high accuracy, implying that the features we used (province and year) contain sufficient information to distinguish and estimate reading interest. This suggests that the temporal and regional patterns in the data are quite strong. In particular, the year 2021 emerged as a critical factor: the across-the-board decline in that year provided a clear signal that models could latch onto. In the classification task, as noted above, many "Low" interest labels coincided with 2021 entries, whereas 2023 entries skewed "High." This temporal pattern, combined with the

fact that provinces have some persistence in their interest levels (some provinces are consistently higher or lower than others), means a model can effectively learn a mapping from (Province, Year) to interest level. The near-perfect performance of the more complex models (e.g., the NeuralNetTorch ensemble achieving 100% on test) hints that they may have essentially memorized the training patterns. Ordinarily, memorization would raise concerns about generalization, but in our case, the patterns (like the 2021 dip) did generalize to test because they reflected true underlying phenomena rather than noise. The real challenge will be whether these models remain accurate for future years beyond 2023 – for instance, if 2024 or 2025 do not follow the same trend (say, if interest plateaus or declines again due to new factors), then the models would need retraining or updates to adapt.

5.2. Ensembles and Individual Model Performance

Our results showed that ensemble methods (which combine multiple models) were among the top performers, but interestingly, a single well-tuned model (the FastAI neural network) essentially matched or exceeded the ensembles on the test set for regression. This highlights a known trade-off: while ensemble approaches (bagging, boosting, stacking) generally improve robustness and accuracy [15], in some cases a sufficiently powerful individual model can match an ensemble on a particular dataset. In our case, the FastAI model likely had the right inductive biases and optimization to fit the data extremely well. The stacked ensemble (WeightedEnsemble L2/L3), which excelled on validation data, might have overfit slightly to the nuances of the validation set – perhaps by giving a bit too much weight to a component model that didn't translate to better performance on the test set. Nonetheless, the differences among the top models were minor. CatBoost's strong showing confirms that gradient-boosted trees remain a competitive choice for structured tabular data, consistent with extensive literature in machine learning. The fact that CatBoost and the neural nets were nearly on par in accuracy suggests that the structure in the reading interest data was not excessively complex – rather, it could be captured by decision rules and relatively simple transformations that boosting can discover. We also saw ExtraTrees and Random Forest perform very well in classification (nearly as well as boosting), indicating low variance in the classification problem; indeed, many models reached the maximum or near-maximum accuracy. This implies that the High/Low classes were highly separable given just province and year, which in turn means that the pattern of which provinces and years correspond to low interest was quite consistent.

5.3. Performance of Simpler Models

The notably poorer results from k-Nearest Neighbors in both tasks deserve mention. k-NN is an intuitive method but it struggled here, reinforcing a point about the nature of the reading interest data: each data point (province-year) doesn't have a dense neighborhood of very similar points. In feature space, one province in 2021 might be "closest" to another province in 2021, but their interest values could still differ by several points. Averaging the neighbors (as k-NN does) washes out the distinctive attributes of each province and year. In contrast, the other models can learn, for example, a "province effect" and a "year effect" separately – essentially performing a form of regression or additive modeling that k-NN cannot do. This suggests that for monitoring reading interest (or similar educational indicators), more sophisticated algorithms are warranted; relying on a naïve approach like k-NN could lead to large errors or misclassification of at-risk regions.

5.4. Impact of Digital Era and COVID-19

Our analysis provides empirical evidence of how a major external event – the COVID-19 pandemic – corresponded with a significant change in reading interest. The 2021 drop is visible across the dataset, and our models not only picked it up but quantified it. On average, the interest index fell by about 19 points (on a 0–100 scale) from 2020 to 2021 nationally. This likely reflects multiple factors: during lockdowns, students might have had reduced access to physical books (with libraries and schools closed) and perhaps did not fully substitute that with digital reading for leisure. Paradoxically, one might expect that being stuck at home could increase the time available for reading, but it appears that the stress of the pandemic, lack of structured school routines, and increased screen time for other purposes (online classes, social media, video streaming) crowded out recreational reading. This aligns with some findings in the literature that digital engagement does not automatically translate to reading engagement[5]. In fact, without guided strategies, students might turn to more passive digital consumption (videos, etc.) rather than reading e-books or online articles. By 2022 and 2023, as schools reopened and campaigns to revive literacy were initiated in Indonesia, the interest levels rebounded strongly – a hopeful sign. Our models effectively learned this rebound as well. From a policy perspective, this underscores both the vulnerability of reading habits to disruption and their resilience when supportive conditions are restored. It highlights the need for maintaining access to reading materials (perhaps through digital libraries or remote book delivery) even during crises, and for encouraging students to engage in reading despite the lure of other digital entertainment.

5.5. Comparison with Other Studies

It is informative to compare our results with those of previous studies discussed in Related Work. Bozkuş's work on predicting reading performance found that school-level factors were most predictive of reading success [14] . In our aggregate data, we cannot directly see school-level factors, but the provincial differences might indirectly reflect factors like quality of education infrastructure or local reading programs. The consistently lower interest in some provinces (say, rural or less-developed areas) could mirror disparities in school resources – something also noted in PISA analyses of digital reading, where home resources and teaching practices influenced digital reading literacy [14, 6]. Additionally, the high performance of our models resonates with the 2025 study by da Silva et al., where even simple ML models achieved high accuracy in classifying reading fluency [15]. Reading interest, like reading fluency, seems to have patterns that are amenable to prediction by ML. Both are influenced by underlying proficiency and environment, which ML can capture if given representative features. However, an interesting contrast is the importance of features: in the fluency study, reading speed was the key feature [15]. In our study, the key "features" boiled down to identifiers (province, year). This points to a limitation, we are predicting what the interest level is, but not why it is high or low. For deeper insight, future work should integrate explanatory features (e.g., literacy rate, library usage statistics, internet penetration, etc.). By doing so, one could possibly use model interpret-ability tools (like feature importance in CatBoost or SHAP values) to identify drivers of reading interest in a similar way to how Bozkuş identified contextual factors for reading performance[14]. Our current models treat each province-year somewhat like a black box identifier; they can tell us "2021 was low, 2023 was high," but not explicitly that "lack of access to libraries caused the drop." Still, the timing and uniformity of the drop strongly suggest the pandemic's role, consistent with global observations of educational disruptions.

5.6. Robustness and Overfitting Considerations

While our test results are excellent, it is important to acknowledge that the test set was quite small. With 20 test points, an accuracy of 100% means the model got all of them correct; just one mistake would drop it to 95%. Thus, a different split or an additional unseen year could change the top ranking. The perfect accuracy of the NeuralNetTorch ensemble, although impressive, should be taken with caution. It could indicate slight overfitting that by chance matched the test set perfectly. However, given that many models were nearly perfect as well, it's more likely that the patterns were genuinely captured rather than the result of overfitting. To further verify robustness, one could perform cross-validation (though with time-based data, one must be careful to avoid leaking future information into training). In our workflow, we relied on a single hold-out test due to the chronological nature of the data and the small sample size. As an additional check, we examined the models' performance on the training data: indeed, the top models achieved 100% training accuracy and near-zero training RMSE, essentially fitting the training points exactly. Normally, fitting exactly would ring alarm bells for overfitting; however, the fact that validation and test performance remained very high indicates that here the issue might not be severe – possibly because the data has low inherent noise. The reading interest index is an aggregate metric and might be measured with some error, but likely it is fairly consistent year-to-year, so the models might be capturing real signal rather than random noise. Nonetheless, caution is warranted when interpreting such strong results. We should not assume the models will continue to be perfect as new data arrives. Regular re-evaluation with fresh data and perhaps more robust validation (e.g., rolling origin forecasts for time series) would be prudent.

5.7. Implications for Educational Policy

From a practical standpoint, our study demonstrates a framework that education authorities could employ for monitoring reading interest. By training ML models on historical interest data (which could be augmented with richer feature sets in practice), one could forecast the next year's reading interest levels for each region. Early predictions could allow targeted interventions – for example, if a model predicts that a certain province's reading interest will stagnate or decline in the coming year, initiatives such as reading campaigns, library expansions, or digital reading programs could be proactively concentrated there. Additionally, the classification approach can flag "Low interest" provinces that need attention. Importantly, such models should be used alongside domain expertise. The ML can tell where and when reading interest is low, but addressing how to improve it requires understanding local context (something that the model could enhance if it incorporated contextual features in the future).

Our results also highlight the interplay between digital technology and reading habits. The pandemic essentially forced an almost exclusively digital mode of learning and leisure for a period, which correlated with a dip in reading interest. But digital technology, if harnessed properly, can also be part of the solution – for instance, online reading communities or e-book distribution might sustain interest when physical interactions are limited. Ensuring that technology use goes hand-in-hand with active reading (as opposed to passive browsing or video watching) is a challenge that educators need to meet. This study provides evidence that monitoring systems powered by ML could be one way to keep a pulse on reading engagement and quickly detect when things go awry. In future work, applying explainable AI techniques such as SHAP (Shapley Additive Explanations) could help illuminate the model's decision process, especially when additional features are included – this would enable policymakers to understand not just the predictions but also the drivers behind those predictions,

increasing trust and actionability of the system.

6. CONCLUSION

In this paper, we presented a comprehensive analysis of reading interest trends in Indonesia using machine learning models applied to real-world survey data. Indonesia is a country with traditionally low reading interest that is facing new challenges in the digital age. Our classification experiments showed that provinces can be categorized into high or low reading interest groups with very high accuracy – several ensemble models (CatBoost, ExtraTrees, neural networks) achieved over 95% accuracy, and one neural network ensemble reached 100% on the test set. The regression experiments demonstrated that the numerical value of a province's reading interest index can be predicted with very high precision (RMSE around 1 on a 0–100 scale) by state-of-the-art models like gradient boosting and deep neural networks. Notably, the algorithms that performed well – e.g., CatBoost, LightGBM, Random Forest, XGBoost, FastAI and PyTorch neural nets, as well as weighted model ensembles – all proved to be effective tools for this task, capturing the relevant patterns in the data.

The analysis revealed a pronounced drop in reading interest in 2021, likely attributable to the COVID-19 pandemic's disruption of regular reading activities, followed by a recovery as normalcy returned. This pattern was not only observed in the raw data but was also learned and reflected by the ML models. It underscores how external factors and the rapid shift to digital learning environments can impact educational behaviors. Our work thus contributes to the education research literature by providing updated empirical findings on reading interest behavior, and to the applied data science domain by illustrating a successful application of machine learning to a socially relevant problem [14]extending those approaches to the specific context of reading interest monitoring.

One limitation of our study is the scope of the dataset. With only four years of data and a few dozen regions, the dataset may not capture more granular intra-year fluctuations or any individual-level variations in reading interest. Also, our models' outstanding performance may partially be a function of this aggregate level – predicting a provincial average is easier than predicting an individual's interest, because individual interests are far more volatile and idiosyncratic. Therefore, one should not directly generalize our results to say "any given person's reading interest can be perfectly predicted." Rather, our results apply to population-level metrics. Another limitation is that we did not explicitly incorporate socio-economic or demographic features that could provide causal insight. While we intentionally focused on province and year in this initial analysis, future work could include, for example, each province's education expenditure, internet access rate, library availability, or demographic indicators to see how they correlate with the interest index and whether their inclusion improves prediction. We suspect that integrating such features would also allow the models to better handle structural changes (e.g., if a new literacy policy in 2024 dramatically boosts reading interest, a model informed by policy-related features might adapt, whereas a pure time-series model might underpredict the jump). Incorporating additional features would furthermore enable the use of explainability tools (like SHAP) to pinpoint which factors most strongly influence the predictions, thereby bridging the gap between predictive power and actionable insight.

Building upon the insights from this study, future research could extend our understanding of reading interest trends in several meaningful ways. First, researchers might explore how well our current models perform on future data by using them to predict reading interest beyond 2023 and then comparing these forecasts to actual outcomes as they become available. Such work would reveal

how reliable these models are when applied to data from changing circumstances. Another promising direction is to increase the granularity of the analysis by using more detailed datasets, for example data gathered from individual students or specific schools. Analyzing finer-scale data would help determine how accurately machine learning models can handle the greater variability inherent at the individual or school level, and whether additional features are needed at that scale. Additionally, future studies could adopt causal inference methods alongside predictive analytics to assess the real-world impact of interventions – for instance, if a province implements a new reading program, one could attempt to measure its effect on the reading interest index using techniques like difference-in-differences or causal forests, complementing the purely predictive approach taken here.

ACKNOWLEDGEMENT

This research was supported by the Applied Research Grant (APBU) from Universitas Lancang Kuning in the year 2025 under Contract Number 134/LPPM/Pn/2025. The authors would like to express their sincere gratitude to Universitas Lancang Kuning for the support and funding that made this study possible.

REFERENCES

[1] F. M. Locher and M. Philipp, "Measuring reading behavior in large-scale assessments and surveys," 2 2023.

[2] S. N. Muhamad, M. N. L. Azmi, and I. Hassan, "Reading interest and its relationship with reading performance: A study of english as second language learners in malaysia," *Humanities and Social Sciences Reviews*, vol. 7, pp. 1154–1161, 11 2019.

[3] L. D. Dawkins, "Factors influencing student achievement in reading," 2017.

[4] S. Nurhasanah, M. Najib, and Ruknan, *The Influence of Literacy Culture on Reading Interest of Elementary School Students*, 2023, pp. 403–409.

[5] I. Peras, E. K. Mirazchiyski, B. J. Pavešić, and Žiga Mekiš Recek, "Digital versus paper reading: A systematic literature review on contemporary gaps according to gender, socioeconomic status, and rurality," pp. 1986–2005, 10 2023.

[6] H. Liu, D. Yang, S. Nie, and X. Chen, "Identifying key factors of reading achievement: A machine learning approach," *iScience*, vol. 27, 10 2024.

[7] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, vol. 11, no. 1, p. 237, 2020.

[8] M. Cubillos, M. Zegers, and H. Inciarte, "Examining adolescent reading engagement: Design and validation of the teacher-reported reading engagement survey (trres)," *Reading Research Quarterly*, vol. 60, no. 2, p. e611, 2025.

[9] S. Suggate, E. Schaughency, H. McAnally, and E. Reese, "From infancy to adolescence: The longitudinal links between vocabulary, early literacy skills, oral narrative, and reading comprehension," *Cognitive Development*, vol. 47, pp. 82–95, 2018.

[10] C. Friman, "The effects of reading medium on children's and adolescents' reading comprehension–a two-phased systematic literature review," 2025.

[11] L. Altamura, C. Vargas, and L. Salmerón, "Do new forms of reading pay off? a meta-analysis on the relationship between leisure digital reading habits and text comprehension," *Review of Educational Research*, vol. 95, no. 1, pp. 53–88, 2025.

[12] R. E. Jensen, A. Roe, and M. Blikstad-Balas, "The smell of paper or the shine of a screen? students' reading comprehension, text processing, and attitudes when reading on paper and screen," *Computers & Education*, vol. 219, p. 105107, 2024.

[13] Y. Yang, H. Adnan, and M. Javed, "Research progress on digital reading behavior: A bibliometric study," *Studies in Media and Communication*, vol. 13, p. 393, 01 2025.

[14] K. Bozkuş, "Predictors of reading performance of fourth-graders," *European Journal of Education*, vol. 60, no. 2, p. e70062, 2025.

[15] G. C. da Silva, R. L. Rodrigues, A. N. Amorim, L. Jeon, E. X. Albuquerque, V. C. Silva, V. F. da Silva, A. L. Pinheiro, J. P. Nunes, S. X. de Souza, M. S. Silva, I. Mauro, and A. M. A. Maciel, "Assessing reading fluency in elementary grades: A machine learning approach," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100411, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X25000517

[16] J. Santhosh, A. P. Pai, and S. Ishimaru, "Toward an interactive reading experience: Deep learning insights and visual narratives of engagement and emotion," *IEEE Access*, vol. 12, pp. 6001–6016, 2024.

[17] H. Rasheed and A. Zahir, "A machine learning approach to personalizing learning by reading experiences based on individual cognitive profiles and preferences," *Transdisciplinary Advances in Social Computing, Complex Dynamics, and Computational Creativity*, vol. 13, no. 11, pp. 1–13, 2023.

[18] V. Sowmya *et al.*, "A meticulous analysis of diverse human traits in avid readers and non-readers through advanced data analytics and machine learning approach," in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*. IEEE, 2024, pp. 1–6.

[19] X. Fang, *HCI in Games*. Springer, 2020.

[20] Imaditia, "Indonesia reading interest 2020-2023," 2023. [Online]. Available: https://www.kaggle.com/datasets/imaditia/indonesia-reading-interest-2020-2023