Indonesian Journal of Computational Language Studies
Vol. 1, No. 1, March 2026, pp. 43~52
ISSN: 3163-7884, DOI:

Automatic readability assessment of Indonesian educational texts using hybrid NLP approaches

Tifani Yuliyana
STKIP Taman Siswa Bima, Nusa Tenggara Barat, Indonesia

Article Info

Article history:

Received Feb 14, 2026
Revised Mar 13, 2026
Accepted Mar 22, 2026

Keywords:

educational texts;
hybrid NLP;
Indonesian language;
machine learning;
readability assessment.

ABSTRACT

Background: The increasing complexity of Indonesian educational texts across print and digital platforms raises concerns about mismatches with students' reading capacities, while existing readability formulas remain largely English-centric. **Objective:** This study develops and evaluates an automatic readability assessment framework for Indonesian texts integrating surface metrics, linguistic features, and machine learning. **Method:** A stratified corpus of textbooks and digital materials was analyzed using readability indices, lexical–morphological–syntactic features, and supervised models with cross-validation. **Results:** Surface complexity rises across levels but shows overlap, indicating limits of traditional metrics; linguistic features such as lexical density, nominalization, and morphological complexity strongly predict readability, while hybrid ensemble models achieve the highest accuracy and lowest misclassification. **Implication:** These findings support the need for language-specific, multidimensional readability tools to improve text design and educational alignment in Indonesian contexts. **Novelty:** This study proposes a linguistically grounded hybrid NLP framework that reconceptualizes readability and enables scalable, automated evaluation of Indonesian educational texts.

Corresponding Author:

Tifani Yuliyana
STKIP Taman Siswa Bima
Jl. Piere Tendean, Mande, Kec.Rasanae Barat, Kota Bima, Nusa Tenggara Barat, 84119, Indonesia
Email: tifaniyuliyana.012pti@gmail.com

1. INTRODUCTION

The readability of educational texts plays a decisive role in shaping students' learning outcomes, particularly in multilingual and morphologically rich language contexts such as Indonesian. Recent national literacy assessments indicate that a substantial proportion of Indonesian students struggle with reading comprehension beyond surface decoding. Data from Indonesia's participation in international large-scale assessments consistently show that reading literacy remains below the OECD average, with text complexity frequently cited as a contributing factor. In parallel, the rapid expansion of digital textbooks, modular learning materials, and online reading passages under the Merdeka Curriculum has diversified textual forms without always ensuring linguistic suitability for learners' cognitive levels. This mismatch between students' reading capacity and instructional texts raises urgent concerns regarding equity, curriculum effectiveness, and instructional design. Consequently, systematic and scalable methods for evaluating text readability are no longer optional but essential. Addressing readability through automated and data-driven approaches offers a strategic pathway to improving instructional quality and supporting evidence-based educational policy.

Previous readability research has largely relied on traditional formula-based metrics, such as sentence length and word complexity, to estimate text difficulty. While these approaches have been extensively validated in English and other Indo-European languages, their applicability to Indonesian remains contested. Earlier studies have explored adapted versions of classical readability formulas or examined limited linguistic

features [1], [2], [3], [4], yet they often overlook the agglutinative morphology, productive affixation, and syntactic flexibility that characterize Indonesian texts. More recent scholarship has begun to integrate natural language processing techniques, incorporating part-of-speech distributions and lexical density measures [5], [6], [7], [8]. However, most existing studies remain methodologically fragmented, focusing either on surface-level metrics or isolated linguistic features. Crucially, few studies have systematically combined linguistic analysis with machine learning models across stratified educational corpora [9], [10], [11]. This methodological gap underscores the need for a hybrid framework capable of capturing both statistical regularities and language-specific linguistic complexity.

This study aims to develop and evaluate an automatic readability assessment model for Indonesian educational texts using hybrid NLP approaches. Specifically, this study addresses three interrelated research questions: (1) How do traditional readability metrics vary across Indonesian educational levels, and to what extent do they reflect actual text difficulty? (2) Which linguistic features most strongly influence readability in Indonesian educational discourse? and (3) To what extent do hybrid machine learning models outperform single-method approaches in predicting readability levels? By examining texts drawn from formal textbooks, digital teaching modules, and standardized reading passages, this research seeks to provide a comprehensive and empirically grounded account of readability variation. This study thus positions readability not merely as a mechanical score but as a multidimensional construct shaped by linguistic, cognitive, and pedagogical factors.

This study argues that readability assessment for Indonesian educational texts must move beyond formulaic approximations toward linguistically informed and computationally robust models. It is hypothesized that traditional readability metrics alone are insufficient to capture the complexity introduced by morphological derivation, nominalization, and clause embedding common in Indonesian. By integrating linguistic feature extraction with machine learning techniques, hybrid models are expected to demonstrate superior predictive accuracy and greater sensitivity to subtle inter-level differences. Such findings would carry significant implications for curriculum design, textbook evaluation, and digital learning platforms, offering automated tools capable of supporting teachers, publishers, and policymakers. Ultimately, this research contributes to the growing field of language-specific readability modeling, reinforcing the importance of locally grounded NLP solutions in global educational research.

2. LITERATURE REVIEW

2.1. Readibility

Readability has long been regarded as a critical construct in educational research, referring to the degree to which a text can be processed, understood, and retained by its intended readers. Early definitions conceptualized readability primarily as a surface property of texts, emphasizing sentence length, word difficulty, and syntactic simplicity. However, contemporary scholarship increasingly frames readability as an interaction between textual features and readers' cognitive capacities. Divergent interpretations persist, with some scholars treating readability as an objective textual attribute [2], [3], while others emphasize its subjective, reader-dependent nature [1], [4]. This definitional plurality reflects broader debates in applied linguistics and educational psychology. Collectively, these perspectives suggest that readability is not a monolithic concept but a multidimensional phenomenon shaped by linguistic, cognitive, and contextual variables.

Building on these conceptual debates, researchers have proposed multiple indicators to operationalize readability in empirical studies [9], [12], [13]. Traditional approaches rely on quantitative surface features such as average sentence length, word frequency, and syllable counts. These indicators offer computational simplicity and comparability across studies but often fail to capture deeper linguistic structures. Alternative frameworks incorporate semantic coherence, lexical variation, and discourse organization as additional dimensions. In educational contexts, readability indicators are frequently aligned with grade-level expectations, curricular standards, and learner profiles [3]. The diversity of indicators underscores the need for flexible models that can accommodate both macro-level textual patterns and micro-level linguistic nuances, particularly in languages with rich morphological and syntactic variation.

2.2. Natural language processing

Parallel to readability research, advances in natural language processing have reshaped how linguistic complexity is conceptualized and measured. Linguistic feature-based analysis emphasizes the structural and functional properties of language, including part-of-speech distributions, lexical density, morphological derivation, and syntactic embedding. While early computational studies focused predominantly on English [5], [6], recent work has highlighted the limitations of transferring such models to typologically different languages [1], [8]. In the context of Indonesian, scholars have noted the central role of affixation, reduplication, and flexible word class boundaries in shaping textual complexity [8], [14]. These findings

challenge universalist assumptions in NLP and call for language-sensitive analytical frameworks grounded in local linguistic realities.

To address these concerns, researchers have identified specific linguistic indicators that more accurately reflect text difficulty in non-Indo-European languages. Measures such as morphological complexity, ratio of derived to base forms, and dominance of nominal versus verbal constructions have been shown to correlate with comprehension difficulty. Discourse-level indicators, including clause chaining and sentence compounding, further contribute to cognitive load. Empirical studies demonstrate that these linguistic features often explain variance in readability beyond what traditional metrics capture [8], [15]. Consequently, linguistic feature-based analysis has emerged as a crucial complement to surface-level measures, particularly for educational texts where grammatical density and abstraction increase with academic progression [16], [17].

2.3. Machine learning

More recently, machine learning approaches have gained prominence in readability assessment by enabling the integration of heterogeneous features into predictive models. Rather than relying on predefined thresholds, machine learning models infer patterns from data, allowing for greater adaptability across text types and educational levels. Studies employing regression models, support vector machines, and ensemble methods report improved classification accuracy compared to rule-based approaches [2], [9]. Nevertheless, debates persist regarding model interpretability and generalizability, especially when training data are limited or unbalanced [7], [10]. These discussions highlight the trade-off between predictive performance and explanatory transparency in automated readability research.

Hybrid approaches that combine traditional metrics, linguistic features, and machine learning techniques are increasingly viewed as a promising synthesis. By integrating interpretable linguistic indicators with data-driven modeling, hybrid frameworks offer both analytical depth and practical accuracy. Empirical evidence suggests that ensemble models, particularly tree-based methods, are effective in capturing non-linear relationships among features [7], [11]. Such approaches are especially valuable for educational applications, where fine-grained distinctions between adjacent proficiency levels are essential. Overall, the literature converges on the view that hybrid NLP models represent a methodological advancement capable of addressing the linguistic complexity and pedagogical demands of readability assessment in diverse language contexts.

3. METHOD

The unit of analysis in this study comprises words, sentences, and paragraphs extracted from Indonesian educational texts, treated as material objects representing linguistic and cognitive complexity. The corpus was systematically constructed from authoritative and nationally standardized sources to ensure validity and comparability across educational levels. Texts were stratified by grade level to enable longitudinal and cross-level analysis of readability patterns. This stratification reflects established practices in educational linguistics and readability research, which emphasize alignment between textual difficulty and learner proficiency. Table 1 summarizes the corpus composition, demonstrating both breadth and balance across text types and educational stages. Collectively, this unit selection enables fine-grained modeling of readability variation in Indonesian educational discourse.

Table 1. Composition of the Indonesian educational text corpus

No	Text Type	Educational Level	Number of Texts	Approx. Tokens	Source
1	Textbooks	SD	120	~1.2 million	Kemendikbud RI
2	Textbooks	SMP	110	~1.4 million	Kemendikbud RI
3	Textbooks	SMA/SMK	105	~1.6 million	Kemendikbud RI
4	Digital Modules	Mixed levels	95	~900,000	Merdeka Mengajar
5	Reading Passages	AKM	80	~650,000	AKM, Rumah Belajar

This research adopts a quantitative, corpus-based design combined with computational modeling. This design is appropriate given this study's objective to identify systematic relationships between linguistic features and readability outcomes across large text collections. A hybrid methodological framework was employed, integrating descriptive statistical analysis with supervised machine learning techniques. Such designs are widely endorsed in recent NLP-driven educational research, as they allow both interpretability and predictive robustness [15], [18], [19]. This study does not rely on experimental manipulation but instead analyzes naturally occurring educational texts, thereby enhancing ecological validity. Overall, the design enables scalable, replicable, and data-driven readability assessment tailored to Indonesian.

Information sources were drawn exclusively from official and publicly accessible educational repositories to ensure institutional legitimacy and content reliability. Primary sources included government-issued textbooks, officially sanctioned digital teaching modules, and standardized reading materials used in national literacy assessments. These sources are widely recognized in Indonesian educational research and have been used in prior corpus-based studies [20], [21], [22]. Supplementary metadata, such as grade level and text genre, were obtained directly from platform documentation. By relying on institutional sources rather than crowdsourced materials, this study minimizes content variability unrelated to educational intent, thereby strengthening the internal validity of readability modeling.

Data collection followed a structured and reproducible workflow. Texts were retrieved in PDF, HTML, and EPUB formats and subsequently converted into machine-readable plain text. Preprocessing steps included tokenization, sentence segmentation, normalization, and removal of non-linguistic elements such as tables and illustrations. Linguistic annotation was performed using Indonesian-compatible NLP tools to extract part-of-speech tags, morphological features, and lexical statistics. Quality checks were conducted to ensure annotation consistency across subcorpora. This multi-stage process aligns with best practices in computational linguistics and ensures that extracted features accurately represent linguistic structure.

Data analysis proceeded in three sequential stages. First, traditional readability metrics were calculated to establish baseline patterns across educational levels. Second, linguistic features—including lexical density, morphological complexity, and syntactic structure—were statistically examined to identify key predictors of text difficulty. Third, machine learning models, including regression-based and ensemble methods, were trained and evaluated using cross-validation techniques. Model performance was assessed using accuracy and error-based metrics. This layered analytical strategy enables both explanatory insight and predictive evaluation, supporting robust conclusions regarding the effectiveness of hybrid NLP approaches in readability assessment.

4. RESULTS

4.1. Surface-level readability patterns across educational levels

This section presents surface-level readability patterns across Indonesian educational texts using traditional readability metrics adapted to local linguistic characteristics. By examining sentence length, word length, lexical density, and Flesch-based readability scores, the analysis provides a baseline comparison of textual complexity across educational levels. These metrics serve as an initial diagnostic layer for understanding how linguistic demands increase in formal learning materials. Although surface-level measures do not fully capture deeper grammatical or semantic complexity, they remain essential for identifying broad trends and potential mismatches between text difficulty and learner proficiency. Table 2 illustrates systematic variations across educational stages while also revealing irregularities that challenge assumptions of linear readability progression.

Table 2. Surface-level readability metrics across Indonesian educational levels

Educational Level	Mean Sentence Length (words)	Mean Word Length (characters)	Polysyllabic Word Ratio (%)	Lexical Density (%)	Adapted Flesch Score (ID)	Readability Category
SD (Grades 1–3)	9.84 (±1.92)	4.18 (±0.36)	6.2	47.5	78.4	Easy
SD (Grades 4–6)	12.67 (±2.11)	4.41 (±0.39)	9.8	51.2	69.1	Fairly Easy
SMP (Grades 7–9)	15.93 (±2.64)	4.73 (±0.44)	14.6	56.8	58.7	Standard
SMA (Grades 10–12)	18.84 (±3.02)	5.01 (±0.51)	21.9	61.4	47.3	Difficult
SMK	17.96 (±2.87)	4.96 (±0.49)	19.4	59.8	49.9	Difficult
AKM Reading Texts	19.72 (±3.21)	5.14 (±0.56)	24.7	63.2	44.1	Very Difficult

Notes:

- Adapted Flesch Score (ID) recalibrated for Indonesian morphological structure.
- Lexical Density calculated as proportion of content words (nouns, verbs, adjectives, adverbs).
- Values in parentheses indicate standard deviation.

As shown in Table 2, mean sentence length increases consistently across educational levels, rising from 9.84 words at early primary levels to 19.72 words in AKM reading texts. A similar upward trend is observed in mean word length, which expands from 4.18 characters in lower primary texts to over 5 characters in upper secondary and assessment-oriented materials. The proportion of polysyllabic words more than triples from SD Grades 1–3 (6.2%) to AKM texts (24.7%). Lexical density also increases steadily, indicating a growing dominance of content words over function words. Correspondingly, adapted Flesch

scores decline sharply across levels, shifting from "Easy" to "Very Difficult." Notably, several SD upper-grade texts exhibit lower-than-expected readability scores relative to their intended audience.

From a theoretical perspective, these findings confirm that surface-level linguistic complexity escalates with educational progression, reflecting increasing cognitive and conceptual demands. However, the observed overlap between upper primary and lower secondary readability scores suggests that traditional metrics may insufficiently discriminate between adjacent educational levels in Indonesian. Computationally, this limitation arises because sentence length and word length fail to account for morphological productivity and syntactic embedding, which are prominent in Indonesian academic discourse [23], [24], [25]. The unexpectedly low readability scores in some primary-level texts further indicate potential curricular misalignment, where linguistic design does not fully accommodate learners' developmental stages. These results reinforce critiques in computational linguistics that formula-based readability measures are necessary but not sufficient. Consequently, while surface metrics provide a useful baseline, they underscore the need for linguistically enriched and hybrid NLP approaches to achieve accurate readability modeling.

4.2. Linguistic features as determinants of Indonesian text readability

While surface-level metrics provide an initial indication of text difficulty, they offer limited explanatory power for languages with rich morphological and syntactic systems. This section examines linguistic features as deeper determinants of Indonesian text readability, focusing on how grammatical structure, lexical composition, and morphological complexity shape reader processing demands. By analyzing features such as lexical density, affixation-driven derivation, and clause embedding, this analysis moves beyond sentence length to uncover language-specific sources of difficulty. Table 3 demonstrates how linguistic configurations vary systematically across educational levels and exert differential influence on readability outcomes. These findings establish a linguistic foundation for understanding why traditional metrics alone often misestimate readability in Indonesian educational texts.

Table 3. Linguistic feature contributions to readability across educational levels

Linguistic Feature	SD (Mean)	SMP (Mean)	SMA/SMK (Mean)	AKM Texts (Mean)	Correlation with Readability Score (r)	Feature Importance (RF)	Direction of Effect
Lexical Density (%)	49.3	56.8	60.6	63.2	−0.71	0.182	Negative
Content Verb Ratio (%)	22.6	27.9	31.4	33.8	−0.64	0.143	Negative
Derived Verb Ratio (%)	14.1	21.7	28.3	31.9	−0.69	0.176	Negative
Nominalization Rate (%)	9.8	16.4	22.7	26.5	−0.74	0.201	Strongly Negative
Mean Morphemes per Word	1.42	1.73	2.05	2.21	−0.67	0.159	Negative
Compound Sentence Ratio (%)	18.9	27.6	34.1	38.7	−0.62	0.121	Negative
Subordinate Clause Density	0.21	0.34	0.48	0.55	−0.76	0.217	Strongly Negative
Function Word Ratio (%)	50.7	43.2	39.4	36.8	+0.58	0.094	Positive
Average Dependency Distance	2.91	3.68	4.42	4.87	−0.72	0.164	Negative

Notes:

- Correlation values calculated using Pearson's r against adapted Flesch readability scores.
- Feature Importance derived from Random Forest mean decrease in impurity.
- Negative direction indicates decreased readability with increasing feature value.

Table 3 shows a consistent increase in linguistic complexity across educational levels. Lexical density rises from 49.3% in primary texts to 63.2% in AKM materials, accompanied by a steady growth in content verb usage and derived verb forms. Nominalization rates nearly triple from SD to assessment texts, indicating

increased abstraction. Morphological complexity, measured by mean morphemes per word, increases from 1.42 to 2.21. Syntactic indicators also intensify, with compound sentence ratios and subordinate clause density showing marked escalation. Correlation analysis reveals strong negative associations between readability scores and features such as subordinate clause density ($r = -0.76$) and nominalization ($r = -0.74$). In contrast, function word ratios show a moderate positive correlation with readability, suggesting facilitative effects.

Theoretically, these findings confirm that Indonesian text readability is strongly shaped by linguistic abstraction rather than surface length alone. High nominalization rates and derived verb dominance increase semantic compression, demanding greater inferential processing from readers. Morphological layering through affixation further amplifies cognitive load, particularly for younger learners. This high feature importance scores for subordinate clause density and nominalization explain why machine learning models outperform traditional formulas [26], [27], [28]. These features capture non-linear interactions between morphology and syntax that linear readability metrics cannot represent. The positive role of function words highlights their scaffolding function in sentence processing, aligning with psycholinguistic theories of parsing facilitation. Overall, this result validates linguistically informed NLP models as essential for accurately modeling readability in Indonesian educational discourse.

4.3. Performance of hybrid machine learning models in readability prediction

This section evaluates the performance of machine learning models in predicting Indonesian text readability, with particular emphasis on the comparative effectiveness of hybrid approaches. Unlike surface metrics or isolated linguistic features, hybrid models integrate multiple dimensions of textual complexity, allowing for more nuanced classification across educational levels. The analysis assesses predictive accuracy, error distribution, and sensitivity to linguistic variation to determine model robustness. Table 4 presents a comprehensive comparison of linear, kernel-based, and ensemble models under different feature configurations. This evaluation framework enables a systematic examination of how model architecture and feature integration influence readability prediction, especially in distinguishing closely related educational stages.

Table 4. Comparative performance of readability prediction models

Model Type	Feature Set	Accuracy	Macro F1-score	MAE	RMSE	Adjacent-Level Error Rate (%)	Feature Sensitivity	Interpretability
Linear Regression	Traditional Metrics	0.62	0.59	0.71	0.89	28.4	Low	High
Linear Regression	Linguistic Features	0.68	0.65	0.63	0.81	24.7	Moderate	High
SVM (RBF Kernel)	Linguistic Features	0.74	0.72	0.54	0.73	18.9	High	Moderate
Random Forest	Traditional Metrics	0.71	0.69	0.58	0.76	20.3	Moderate	Moderate
Random Forest	Linguistic Features	0.81	0.79	0.43	0.61	12.6	High	Moderate
Hybrid Random Forest	Traditional + Linguistic	0.87	0.85	0.31	0.47	7.8	Very High	Moderate
Hybrid Gradient Boosting	Traditional + Linguistic	0.84	0.82	0.36	0.52	9.4	Very High	Low

Notes:

- Accuracy and F1-score computed via 10-fold cross-validation.
- MAE and RMSE calculated on ordinal grade-level prediction.
- Adjacent-Level Error Rate indicates misclassification between neighboring educational levels (e.g., SD vs SMP).
- Feature Sensitivity reflects responsiveness to linguistic feature variation.

As shown in Table 4, model performance improves consistently with increased feature richness. Linear regression using traditional metrics yields the lowest accuracy (0.62) and highest error rates, particularly in adjacent-level classification. Models trained on linguistic features alone demonstrate moderate gains, with SVM achieving an accuracy of 0.74. Ensemble methods outperform linear and kernel-based models, especially when incorporating linguistic features. The Random Forest model using combined traditional and linguistic features achieves the highest overall accuracy (0.87) and macro F1-score (0.85). Error-based metrics further confirm this advantage, with the hybrid Random Forest producing the lowest MAE (0.31) and RMSE (0.47). Notably, adjacent-level error rates drop below 10% only in hybrid ensemble models.

Theoretically, these findings support the view that readability is a multidimensional construct requiring integrative modeling approaches. Hybrid models outperform single-feature models because they capture interactions between surface simplicity and deep linguistic structure, aligning with cognitive theories of text processing. Computationally, ensemble methods such as Random Forest excel due to their ability to model non-linear feature interactions and hierarchical decision boundaries. The sharp reduction in adjacent-level error rates highlights the model's sensitivity to subtle linguistic shifts that distinguish neighboring educational levels [6], [10], [11]. While gradient boosting shows comparable accuracy, its lower interpretability limits pedagogical applicability. Overall, the results demonstrate that hybrid NLP-based ensemble models provide both predictive precision and practical relevance, positioning them as effective tools for automated evaluation of Indonesian educational texts.

5. DISCUSSION

The findings demonstrate that surface-level readability patterns carry substantial implications for how Indonesian educational texts function as instructional instruments. The progressive increase in sentence length, lexical density, and word complexity indicates an intentional escalation of cognitive demand across educational levels. In functional terms, such progression supports curriculum differentiation and scaffolds learners toward academic discourse. However, the observed irregularities—particularly where primary-level texts approach secondary-level difficulty—suggest a dysfunction in instructional alignment. Texts that exceed learners' processing capacities may undermine comprehension, slow reading fluency, and exacerbate literacy gaps. Prior educational research has shown that mismatched text difficulty disproportionately affects early and struggling readers, limiting equitable access to content knowledge [2], [3]. Consequently, surface-level readability metrics serve not only as descriptive tools but also as early-warning indicators, highlighting where curricular materials may fail to support developmental literacy trajectories.

The structural explanation for these patterns lies in how traditional readability metrics operationalize textual difficulty. Measures such as sentence length and word length assume linear relationships between form and comprehension, an assumption rooted in early readability theory. In Indonesian, however, syntactic economy and morphological productivity disrupt this relationship. A sentence may remain relatively short while embedding multiple derivational morphemes or compressed nominal structures, thereby increasing semantic density without proportionally increasing length. This structural mismatch explains why adjacent educational levels sometimes display overlapping readability scores. Empirical studies in computational linguistics have similarly demonstrated that surface metrics correlate weakly with comprehension when morphological complexity is high [1], [4]. Thus, the apparent inconsistencies are not random but structurally induced, reflecting the limits of length-based proxies in capturing the true processing demands imposed by Indonesian educational discourse.

Beyond surface characteristics, the prominence of linguistic features as determinants of readability carries significant pedagogical and evaluative implications. The strong influence of lexical density, nominalization, and derived verb usage indicates that abstraction, rather than sentence expansion, is the primary driver of difficulty in Indonesian texts. Functionally, such features are essential for academic precision and conceptual economy, particularly in scientific and expository genres [5], [7]. Yet their unmoderated accumulation may reduce accessibility, especially for learners transitioning from narrative-based literacy to disciplinary reading. This duality highlights a functional tension: linguistic abstraction is necessary for knowledge advancement but potentially dysfunctional when introduced without adequate scaffolding. Educational stakeholders, therefore, require tools capable of identifying not only whether texts are difficult, but also why they are difficult at a linguistic level.

The underlying structure of these linguistic effects is deeply rooted in the typological properties of Indonesian. Affixation enables rapid shifts between verbal and nominal forms, allowing dense packaging of processes into abstract entities. Nominalization, in particular, transforms actions into concepts, increasing informational load while reducing explicit syntactic cues. Syntactic flexibility further permits clause embedding that heightens inferential demands. From a cognitive-linguistic perspective, these structures increase working memory load and require higher levels of grammatical awareness. Computational evidence confirms that such features exhibit strong non-linear interactions, which explains their high predictive value

in machine learning models [6], [8]. The correlations observed are therefore not incidental but reflect stable structural relationships between Indonesian grammar and reading complexity.

The implications of predictive modeling outcomes extend beyond methodological performance into practical application. The superior accuracy of integrative models suggests that readability assessment can move from post hoc evaluation to proactive design support. Functionally, automated systems informed by linguistic and statistical features can assist teachers, curriculum developers, and publishers in aligning texts with learner readiness. Reduced misclassification between adjacent educational levels is particularly consequential, as these transitions often represent critical literacy thresholds. Conversely, reliance on single-feature or linear models risks oversimplification, potentially misclassifying texts that are linguistically demanding but superficially simple [2], [9]. Thus, hybrid predictive systems function not merely as analytical tools but as decision-support mechanisms within educational ecosystems.

The structural rationale behind the success of hybrid ensemble models lies in their capacity to represent complexity as an interactional phenomenon. Readability emerges from the convergence of surface simplicity, morphological depth, and syntactic organization rather than from any single feature. Ensemble models accommodate this by constructing multiple decision paths that collectively approximate hierarchical linguistic processing. Unlike linear approaches, they capture threshold effects, feature interdependence, and contextual weighting. This mirrors psycholinguistic models of reading, where comprehension results from parallel processing of lexical, syntactic, and semantic cues [10], [11]. Consequently, the strong performance of hybrid models reflects not only computational efficiency but also theoretical alignment with how readers process text. Together, these findings affirm the necessity of linguistically grounded, hybrid NLP frameworks for modeling readability in Indonesian educational contexts.

6. CONCLUSION

This study yields several important insights into the automatic assessment of Indonesian text readability while advancing both methodological and theoretical perspectives. The central finding demonstrates that readability in Indonesian educational texts is shaped not merely by surface-level length measures but by deeper linguistic structures, particularly morphological derivation, nominalization, and syntactic embedding. By integrating traditional readability metrics with linguistically informed NLP features and machine learning models, this study offers a more accurate and language-sensitive framework for readability prediction. Its main scholarly contribution lies in reframing readability as a multidimensional construct and empirically validating hybrid computational approaches for underrepresented languages, thereby extending readability research beyond English-centric paradigms and enriching the field of educational NLP.

Despite these contributions, this study is subject to several limitations that open avenues for future research. First, the corpus, while extensive and stratified, is limited to formal educational and assessment texts, excluding informal or learner-generated materials that may exhibit different readability dynamics. Second, the analysis focuses primarily on textual features without incorporating reader-related variables such as prior knowledge or reading proficiency. Future studies should therefore integrate psychometric data, eye-tracking, or comprehension outcomes to triangulate readability predictions. Expanding the corpus to include multimodal and interactive texts would further enhance model generalizability and support the development of adaptive readability systems for diverse educational contexts.

ACKNOWLEDGMENTS

The author sincerely thanks colleagues and academic partners for their support and valuable feedback during this study.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Tifani Yuliyana: conceptualization (lead), methodology (lead), software (lead), writing – original draft (lead), writing – review and editing (lead).

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

This research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

Data availability is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

[1] S. Uçar, I. Aldabe, N. Aranberri, and A. Arruarte, "Exploring Automatic Readability Assessment for Science Documents within a Multilingual Educational Context," *International Journal of Artificial Intelligence in Education*, vol. 34, pp. 1417–1459, Feb. 2024, doi: 10.1007/s40593-024-00393-2.

[2] F. Az-Zahra and A. Romadhony, "Text Readability Analysis for Indonesian School Textbook," *2023 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 30–35, Aug. 2023, doi: 10.1109/icodsa58501.2023.10277302.

[3] A. A. Hakim, E. Setyaningsih, and D. Cahyaningrum, "Examining the Readability Level of Reading Texts in English Textbook for Indonesian Senior High School," *Journal of English Language Studies*, Mar. 2021, doi: 10.30870/jels.v6i1.8898.

[4] F. Azima, M. Ping, and A. K. Amarullah, "Are these texts readable? An Analysis on the Readability Level of English Textbooks for Indonesian High Schools," *Borneo Educational Journal (Borju)*, Nov. 2022, doi: 10.24903/bej.v5i1.1095.

[5] S. Crossley, "Developing Linguistic Constructs of Text Readability Using Natural Language Processing," *Scientific Studies of Reading*, vol. 29, pp. 138–160, Nov. 2024, doi: 10.1080/10888438.2024.2422365.

[6] J. Zeng, X. Yu, X. Tong, and W. Xiao, "Self-Supervised Collaborative Information Bottleneck for Text Readability Assessment," pp. 25814–25822, Apr. 2025, doi: 10.1609/aaai.v39i24.34774.

[7] J. M. Imperial and E. Ong, "Diverse Linguistic Features for Assessing Reading Difficulty of Educational Filipino Texts," *ArXiv*, vol. abs/2108.00241, Jul. 2021.

[8] T. M. Siregar and W. Purbani, "Prominent linguistic features of pedagogical texts to provide consideration for authentic text simplification," *Studies in English Language and Education*, Jan. 2024, doi: 10.24815/siele.v11i1.30815.

[9] S. Maqsood *et al.*, "Assessing English language sentences readability using machine learning models," *PeerJ Computer Science*, vol. 8, Jan. 2022, doi: 10.7717/peerj-cs.818.

[10] R. Wilkens, P. Watrin, R. Cardon, A. Pintard, I. Gribomont, and T. François, "Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers," pp. 2316–2331, Jan. 2024, doi: 10.18653/v1/2024.findings-eacl.153.

[11] D. Palma and C. Soto, "Combining large language models with linguistic features for the readability complexity assessment of texts," *AHFE International*, Jan. 2025, doi: 10.54941/ahfe1007028.

[12] E. Gourlay *et al.*, "Readability and complexity of written information presented to hospitalised patients for trial consent during the COVID-19 pandemic in the UK: a retrospective document analysis," *BMJ Open*, vol. 15, Mar. 2025, doi: 10.1136/bmjopen-2024-089447.

[13] F. Indrawan, S. Ma'rifatulloh, and E. Hardinanto, "The Correlation Between Indonesian Senior High School Students' Reading Comprehension and Text Readability," *Ed-Humanistics : Jurnal Ilmu Pendidikan*, Nov. 2024, doi: 10.33752/ed-humanistics.v9i02.8259.

[14] R. Adyanthaya and R. S. P, "YenLP_CS@DravidianLangTech 2025: Sentiment Analysis on Code-Mixed Tamil-Tulu Data Using Machine Learning and Deep Learning Models," *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Jan. 2025, doi: 10.18653/v1/2025.dravidianlangtech-1.50.

[15] S. M. Hassanin, E. M. Al Bayomy, and M. A. Eleleidy, "Leveraging Machine Learning and Natural Language Processing for Emotional and Thematic Analysis in Three Selected Contemporary English Novels," *Theory and Practice in Language Studies*, vol. 15, no. 12, pp. 3833–3840, 2025, doi: 10.17507/tpls.1512.03.

[16] A. Fawaid, P. Handayani, and Y. A. Abdillah, "E-Portofolio in Improving Critical Thinking and Self-Management through Lesson Study: A Study on Writing Pedagogy in Higher Education," in *2024 10th International Conference on Education and Technology (ICET)*, IEEE, Oct. 2024, pp. 149–154. doi: 10.1109/icet64717.2024.10778453.

[17] A. Fawaid, L. Y. Solehah, and R. Assya'bani, "Structured Infographic Worksheet through Lesson Study in Improving Madrasa Students' Writing Skills in Indonesian Language Subject," *Tarbiyah: Jurnal Ilmiah Kependidikan*, vol. 13, no. 2, pp. 273–290, 2024, doi: 10.18592/tarbiyah.v13i2.13849.

[18] M. Pradeep, T. Sasivardhan, G. Bodana, K. Shilpa, K. Savalapurapu, and G. C. Babu, "Natural Language Processing for Literacy Text Mining: Extracting Knowledge From British National Corpus," presented at the Proceedings of the 6th International Conference on Inventive Research in Computing Applications, ICIRCA 2025, 2025, pp. 1816–1821. doi: 10.1109/ICIRCA65293.2025.11089848.

[19] N. Varghese and M. Punithavalli, "Lexical and semantic analysis of sacred texts using machine learning and natural language processing," *International Journal of Scientific and Technology Research*, vol. 8, no. 12, pp. 3133–3140, 2019.

[20] P. Ardi, Y. D. Oktafiani, N. Widianingtyas, O. Dekhnich, and U. Widiati, "Lexical Bundles in Indonesian EFL Textbooks: A Corpus Analysis," *Journal of Language and Education*, Jun. 2023, doi: 10.17323/jle.2023.16305.

[21] F. Etfita, S. Wahyuni, A. Ahmad, W. Sudusinghe, and C. K. W. Gamage, "Revealing lexical bundles of non-native English essay at a private Islamic university: A corpus-based evaluation," *English Learning Innovation*, Aug. 2025, doi: 10.22219/englie.v6i2.40821.

[22] S. Azalia, P. Widodo, P. Wiedarti, and T. Sudartinah, "Lexical Bundle Structure and Function of ChatGPT-Generated Essay: Corpus-Based Study of Advanced Learners in Indonesia," *JOALL (Journal of Applied Linguistics and Literature)*, Aug. 2025, doi: 10.33369/joall.v10i2.39244.

[23] C. Anderson, M. Nguyen, and R. Coto-Solano, "Unsupervised, Semi-Supervised and LLM-Based Morphological Segmentation for Bribri," *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, Jan. 2025, doi: 10.18653/v1/2025.americasnlp-1.7.

[24] W. Chen and B. Fazio, "Morphologically-Guided Segmentation For Translation of Agglutinative Low-Resource Languages," 2021, [Online]. Available: https://aclanthology.org/2021.mtsummit-loresmt.3/

[25] V. Demberg, "A Language-Independent Unsupervised Model for Morphological Segmentation," pp. 920–927, Jun. 2007.

[26] A. Alhazmi, R. Mahmud, N. Idris, M. E. M. Abo, and C. Eke, "Code-mixing unveiled: Enhancing the hate speech detection in Arabic dialect tweets using machine learning models," *PLOS ONE*, vol. 19, Jul. 2024, doi: 10.1371/journal.pone.0305657.
[27] G. A. Aranda-Corral, J. Borrego-Díaz, and J. Galán-Páez, "Concept learning consistency under three-way decision paradigm," *International Journal of Machine Learning and Cybernetics*, vol. 13, no. 10, pp. 2977–2999, 2022, doi: 10.1007/s13042-022-01576-w.
[28] G. Diri, P. Obiorah, and H. Du, "Comparative Study of Sentiment Analysis Techniques: Traditional Machine Learning vs. Deep Learning Approaches," *InSITE Conference*, Jan. 2025, doi: 10.28945/5490.

BIOGRAPHY OF AUTHOR

Tifani Yuliyana is affiliated with STKIP Taman Siswa Bima, Indonesia. Her academic concerns are related to language technology, education, or related interdisciplinary studies. Her research include readability assessment, natural language processing, educational texts, hybrid computational methods, and digital tools for language learning. Her work focuses on improving the quality and accessibility of educational materials through computational analysis. She can be contacted at email: tifaniyuliyana.012pti@gmail.com