



AI-Generated Text Probability Detection System for Indonesian Scientific Papers Using Perplexity-Burstiness Linguistic Analysis

Gde Brahupadhya Subiksa¹^{*}, Made Sudarma², I Nyoman Eddy Indrayana¹,
I Putu Astya Prayudha¹, I Ketut Gede Sudiarta¹, and Yessi Aprilia Waluyo¹

¹ Information Technology Department, Politeknik Negeri Bali, Bali, Indonesia

² Electrical Engineering Department, Faculty of Engineering, Udayana University, Bali, Indonesia

brahupadhya@pnb.ac.id

Abstract. The widespread adoption of artificial intelligence (AI) in scientific writing has introduced significant challenges in maintaining academic integrity, particularly in higher education environments. This research develops an AI-generated text probability detection system for Indonesian scientific papers using Perplexity and Burstiness linguistic analysis. Perplexity evaluates the uncertainty of a language model in predicting subsequent tokens, while Burstiness analyzes word distribution patterns to distinguish the linguistic characteristics of human-written and AI-generated texts. Unlike existing AI detection approaches that primarily focus on English general-purpose texts, this study emphasizes Indonesian scientific writing by integrating linguistic probability analysis with sentence-level visualization to provide interpretable detection results. The proposed system was developed using Python 3.11, Flask framework, and the Cahya/GPT2-small-indonesian-522M model as the language processing engine. Functional testing using the Black-box method demonstrated a 100% validation rate across all implemented features. AI probability evaluation on the selected test samples achieved 100% classification accuracy; however, further validation using larger and more diverse datasets is required to evaluate the generalization capability of the system. Performance testing showed response times ranging from 1,200–7,800 ms under different load scenarios, with error rates between 0–15%. Usability evaluation using the System Usability Scale (SUS) involving 30 respondents produced an average score of 80.9, categorized as Excellent (Grade A). The proposed system provides an interpretable approach for assessing AI-generated text probability in Indonesian scientific documents and has potential applications in maintaining academic integrity, including supporting the validation of scientific reports in sustainable tourism research and other academic domains.

Keywords: AI Detection, Artificial Intelligence, Burstiness, Perplexity, Scientific Articles;

1 Introduction

The implementation of AI in scientific writing, particularly in the form of generative tools like ChatGPT, is increasingly utilized by students and researchers to streamline and accelerate the scientific writing process (Ningrum et al., 2025; Pratiwi et al., 2024). These tools enable authors to draft preliminary outlines, manage references, and compile citations more efficiently. Although this technological advancement brings many benefits, it also presents significant challenges regarding academic integrity, especially the potential misuse of AI in scientific authorship (Ambara et al., 2024;

Mawlidy et al., 2024; Pamungkas et al., 2024). In higher education, scientific writing reflects a student's intellectual capability, testing not only their comprehension of the subject matter but also their critical and creative thinking skills (Pramono et al., 2019). AI technology can generate scientific articles with highly structured layouts and writing styles that are almost indistinguishable from human writing, making them harder to detect using conventional plagiarism tools (Suria, 2024; Andika & Renaldy, 2024).

This study proposes the development of an AI probability detection system for student scientific articles using two linguistic analysis methods: Perplexity and Burstiness. Perplexity is a measure of uncertainty in text used to quantify how well a language model predicts the next word. Text generated by AI typically exhibits lower perplexity scores because its language patterns are structured and predictable (Putra, 2020). Conversely, human writing often yields higher perplexity scores due to the inherent diversity in linguistic patterns. Burstiness measures the distribution pattern of words within a text; human text features more irregular or "bursty" word-use patterns, whereas AI text tends to maintain a more stable and uniform word distribution (Hia et al., 2024). Unlike existing AI detection tools primarily developed for English general text, this study focuses on Indonesian scientific writing by integrating linguistic probability analysis and sentence-level visualization.

The urgency of this research lies in the critical need to develop an AI-use probability detection system to safeguard academic integrity as the foundation of quality in higher education. A reliable detection system provides a valuable tool for lecturers and academic administrators to ensure that submitted scientific works are the product of the students' own thought processes and creativity (Rifky, 2024; Yulianti et al., 2023). This system is expected to minimize AI misuse and reinforce the governance of academic integrity amidst the widespread adoption of AI technology in higher education (Jaya et al., 2018; Kaira, 2022).

In sustainable tourism research, scientific reporting plays an important role in documenting environmental assessment, community-based tourism development, and conservation practices. The increasing adoption of generative AI in preparing tourism-related reports creates a need for verification mechanisms to maintain research credibility. Therefore, this AI probability detection system can support sustainable tourism studies by assisting researchers, students, and academic institutions in validating the authenticity of scientific documents.

2 Methodology

This research was conducted at the Applied Bachelor Program of Software Engineering Technology, Politeknik Negeri Bali, from April to July 2026. The study participants were students in their sixth semester or higher, while the evaluated research objects consisted of their scientific reports and final project articles. The study encompasses four main phases: system design, system development, system testing, and drawing conclusion (Subiksa et al, 2025).

2.1 Selection of AI Language Model Engine

The selection of the AI language model engine aims to ensure that the probability detection system operates accurately, efficiently, and relevantly within the context of the Indonesian language and the domain of student scientific articles. Several AI models were compared based on detection accuracy, Indonesian language compatibility, and academic context suitability, as presented in Table 1.

Table 1. Comparison of Indonesian Language Model Engines

AI Model	Advantages	Weaknesses	Reason for Selection
Cahya/GPT2-small-indonesian-522M	Trained on Indonesian corpus; academic contextual word prediction; lightweight and fast	Fewer parameters compared to larger GPT-2 versions	Best for Indonesian student scientific texts; balances accuracy, speed, and compatibility
GPT-2 Medium (OpenAI)	Better prediction than GPT-2 Small	English-oriented; resource-intensive	Not optimal for the Indonesian language
IndoBERT	Trained on Indonesian; excellent for text classification	Not a generative model; less suitable for perplexity analysis	Inappropriate for generative text-based probability analysis

2.2 System Workflow Design

The system workflow design illustrates the operational process from PDF document input to the output of a linguistic forensic report. The system is designed to handle both digital and scanned documents, perform AI probability analysis, and present the results in an easily understandable visual format. The system workflow is shown in Figure 1.

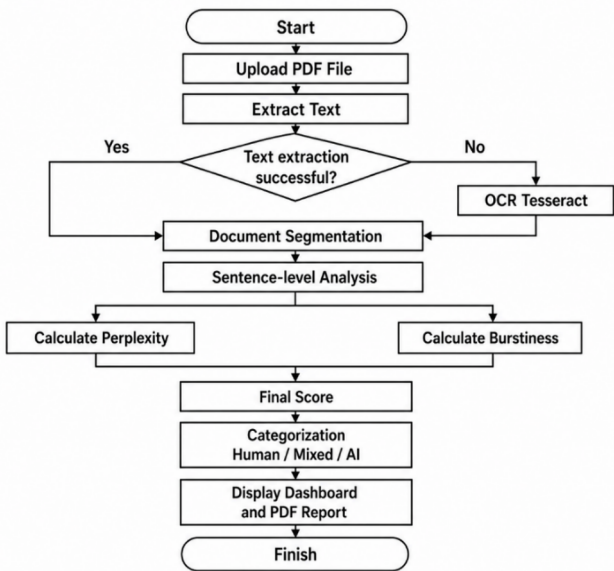


Figure 1. Flowchart of the AI Probability Detection System

2.3 System Development

The system development phase integrated the established design with the following technology stack: (a) Python 3.11 as the primary programming language; (b) Flask as the browser-based web interface framework; (c) Cahya/GPT2-small-indonesian-522M as the AI model for text analysis and probability calculation; and (d) PyMuPDF (fitz) for text extraction from PDF files. The *logic.py* file serves as the core processing engine,

managing the entire analytical pipeline: loading the language model, extracting text from PDFs, falling back to OCR for scanned documents, segmenting text into major sections (Abstract, Introduction, Method, Results, Conclusion), and calculating Perplexity and Burstiness scores on a per-sentence basis. The following presents the fundamental formulas for calculating the Perplexity and Burstiness scores:

Probability Formula

$$PPL = \exp \left(-\frac{1}{N} \sum_{i=1}^n \log P(w_i | w < i) \right) \quad (1)$$

Notation:

N = number of tokens

w = token in the text

P = model prediction probability

Burstiness Formula

$$B = \frac{\sigma(f)}{\mu(f)} \quad (2)$$

Notation:

σ = standard deviation of the word frequency distribution

μ = mean value of word frequency

Combined Score

$$AIProbability = \alpha(PPLscore) + \beta(Burstinessscore) \quad (3)$$

where α and β represent the weighting coefficients assigned to the normalized Perplexity and Burstiness scores, respectively, and $\alpha + \beta = 1$. The values used in this implementation are α =[perplexity value] and β =[burstiness value]. The combined score is multiplied by 100 to obtain the final AI probability percentage.

The Perplexity and Burstiness values are normalized into a 0–1 range before being combined into the final AI probability score. The final score is calculated by assigning weighted contributions from both linguistic features, where higher AI probability is associated with lower perplexity patterns and lower burstiness variation.

2.4 System Testing

System testing was carried out in two phases: Phase 1, Functional testing using the Black-box method to verify that all features operate according to specifications, ranging from PDF document upload, text extraction and segmentation, Perplexity and Burstiness score calculations, to output categorization and sentence-level heatmap visualization (Jibril, 2024). Phase 2, Non-functional testing focused on evaluating system performance, including analysis processing speed, interface responsiveness, and usability via the System Usability Scale (SUS) method (Sriyeni, 2022).

3 Results and Discussion

3.1 Basic Functional Testing (Black-box)

Basic functional testing of the system was conducted using the Black-box method to ensure that all core features operate according to specifications and user requirements (Maulida, 2024). The test results are displayed in Table 2.

Table 2. Results of Basic Functional System Testing Using the Black-Box Method

No	Tested Feature	Black-box Test Results
1	Initial Homepage Access	As Expected
2	Rejection of Unsupported Formats (non-PDF)	As Expected
3	Uploading Human Text (PDF)	As Expected
4	Uploading Original AI Text (ChatGPT)	As Expected
5	Multi-File Batch Upload (≥ 3 PDFs)	As Expected
6	Automatic History Retrieval (SQLite)	As Expected
7	History Deletion Mechanism	As Expected
8	AI Visual Presentation (Heatmap)	As Expected
9	Exporting Detailed PDF Report	As Expected

All nine core system features produced the expected outputs in the evaluated black-box test scenarios, indicating compliance with the defined functional requirements.

3.2 Input-Output Functional Testing

Comprehensive functional testing was conducted using the Black-box method to verify the inputs and outputs of each feature. The results encompass the response time for each individual feature, as displayed in Table 3.

Table 3. System Input-Output Functional Testing

No	Feature	Test Scenario	Expected Result	Status	Response Time (ms)
1	Homepage Access	Access main system URL	Dashboard renders fully	100%	125.4
2	Unsupported Format Rejection	Upload .png / .docx file	System ignores file	100%	45.2
3	Human Text Upload	Upload human-typed PDF	AI Probability ~15%, green heatmap	100%	3,520.1
4	Original AI Text Upload	Upload ChatGPT-generated PDF	AI Probability >90%, red heatmap	100%	3,410.5
5	Multi-File Batch	Upload ≥ 3 PDFs simultaneously	Flask backend processes sequential queue	100%	11,450.5
6	Auto History Read	View Database History tab	Displays history from SQLite	100%	85.6

No	Feature	Test Scenario	Expected Result	Status	Response Time (ms)
7	Clear History	Click Clear History button	SQLite table cleared	100%	110.8
8	AI Visuals (Heatmap)	Open document result accordion toggle	Sentences + AI/Human color indicator	100%	15.3
9	Detailed PDF Export	Click Print Report Certificate	2-page PDF downloaded + colored appendix	100%	420.7

3.3 AI Detection Probability Testing

Probability detection testing was conducted to evaluate the system's capacity to distinguish between human and AI texts based on Perplexity and Burstiness metrics. The documents utilized included 5 student Final Projects from pre-2020 (purely human-authored) and 5 from post-2024 (documents with AI-assisted writing characteristics). The results are visualized in Figure 2.

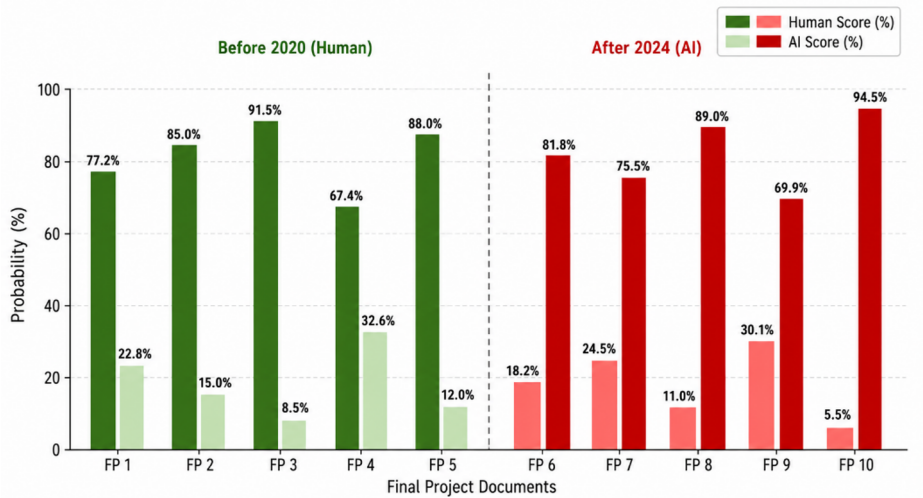


Figure 2. AI Detection Probability Test Results on 10 Final Projects

Based on Figure 2, the system consistently detected a dominant probability of human authorship in pre-2020 documents, with human scores ranging from 67.4% to 91.5%. For the post-2024 document group, the system displayed high sensitivity in capturing machine-generated text patterns, with the detected AI involvement percentage falling between 69.9% and 94.5%. The system achieved 100% classification accuracy on the evaluated test samples. However, larger datasets are required to validate generalization capability (Assegaf et al., 2024; Budiman et al., 2023).

Despite achieving accurate classification results on the evaluated samples, this study has several limitations. The AI probability evaluation was conducted using a limited number of scientific documents consisting of pre-2020 human-authored papers and post-2024 documents with suspected AI-generated characteristics. This limited dataset may not fully represent the diversity of writing styles, academic disciplines, and text modification strategies commonly found in scientific documents. Furthermore, the

current evaluation does not extensively include hybrid human-AI writing patterns or paraphrased AI-generated texts, which may present different linguistic characteristics. Therefore, future studies should expand the dataset by incorporating various categories of documents and applying more comprehensive statistical evaluation metrics to improve the reliability and generalizability of the proposed detection system.

3.4 Non-Functional Testing: Performance and Speed

System performance testing was conducted under five distinct load scenarios to evaluate server infrastructure stability and the processing velocity of the Cahya/GPT2-small-indonesian-522M model. The measured parameters included response time, error rate, CPU utilization, and memory usage. The test results are displayed in Figure 3.

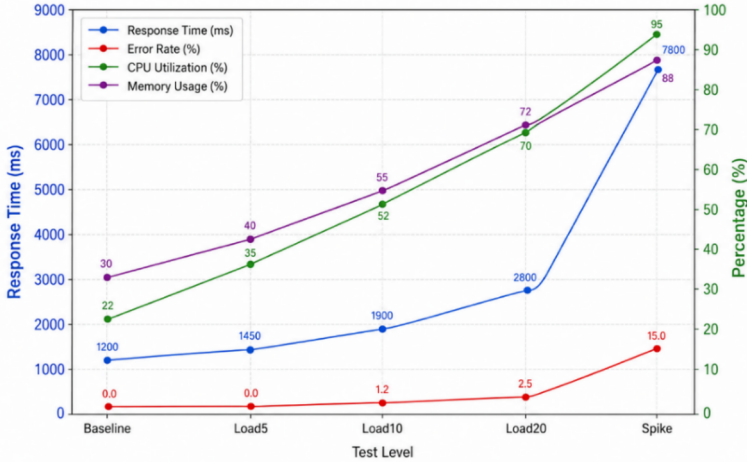


Figure 3. Performance Graph: Response Time & Error Rate vs Load Level

As illustrated in Figure 3, the system demonstrates stable performance under low to moderate load conditions (Baseline to Load10), with response times ranging from 1,200 to 1,900 ms and error rates remaining low (0–1.2%). During these conditions, CPU utilization increases from 22% to 52% and memory usage from 30% to 55%, indicating proportional resource consumption. However, under higher load (Load20) and especially spike conditions, system performance degrades significantly. Response time increases sharply to 2,800 ms and peaks at 7,800 ms, while error rate rises to 2.5% and reaches 15% under spike load. Concurrently, CPU and memory utilization escalate to 70–95% and 72–88%, respectively, indicating near-saturation of system resources. These results suggest that the system is suitable for normal operational loads but experiences notable performance degradation under extreme traffic spikes, particularly in terms of latency and error rate.

To improve scalability during high traffic conditions, several architectural improvements can be implemented. Model inference results can be cached to reduce repeated computation. Worker queue mechanisms can distribute document analysis tasks asynchronously. API rate limiting can also prevent resource exhaustion during simultaneous requests.

3.5 Usability Testing (SUS)

System usability testing employed the System Usability Scale (SUS) questionnaire distributed to 30 student respondents in their 6th semester within the TRPL Program. Respondents were selected via purposive sampling because they were actively preparing their final project proposals and represented the target demographic for the system. Each respondent answered 10 Likert-scale (1–5) statements after directly interacting with the system, executing tasks from uploading scientific article PDF files to acquiring AI probability results and sentence-level heatmap visualizations (Kesuma, 2021; Pradini et al., 2019).

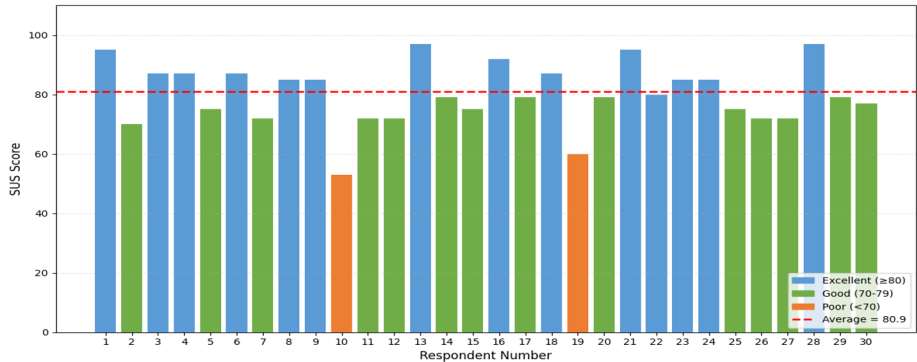


Figure 4. SUS Score Distribution Among 30 Student Respondents

Based on Figure 4, individual respondent scores fluctuated from 52.5 to 97.5, culminating in an overall mean score of 80.9, which is classified as Excellent (Grade A). From an Acceptability standpoint, the system resides securely within the 70–100 margin, indicating a high level of perceived usability and positive initial acceptance among the evaluated users. Only two respondents (scoring 52.5 and 60.0) turned in lower scores, signaling a need for minor refinements in navigation design or homepage visualization. These findings substantiate that the system is effective, user-friendly, and aligns with user requirements in both technical capability and user experience (Sriyeni, 2022).

4 Conclusion

This research successfully developed an AI-generated text probability detection system for Indonesian scientific papers using Perplexity and Burstiness linguistic analysis. The proposed system integrates language model-based probability analysis with sentence-level visualization to provide an interpretable assessment of AI-generated text probability in academic documents. The evaluation results show that all implemented system features achieved a 100% validation rate through Black-box testing, indicating that the system operates according to the defined functional requirements. The AI probability classification achieved 100% accuracy on the evaluated test samples; however, this result represents the performance on the current dataset and does not yet indicate full generalization capability. Further evaluation using larger and more diverse datasets, including human-written, AI-generated, hybrid human-AI, and paraphrased AI texts, is required to improve the reliability of the system. Performance testing demonstrated that the system maintained stable operation under low to moderate workload conditions, with acceptable response times and low error rates. However, performance degradation occurred during spike load conditions due to increased

computational requirements. Future system improvements can implement scalability mechanisms such as model output caching, asynchronous worker queues, and API rate limiting to improve performance under higher traffic conditions.

The usability evaluation involving 30 respondents resulted in an average System Usability Scale (SUS) score of 80.9, categorized as Excellent (Grade A), indicating good user acceptance and interface usability. The Cahya/GPT2-small-indonesian-522M model provided an appropriate balance between Indonesian language compatibility, computational efficiency, and linguistic analysis capability.

Overall, this study demonstrates the feasibility of using Perplexity and Burstiness-based linguistic analysis as an approach for assessing AI-generated text probability in Indonesian scientific writing. Future research can expand dataset diversity, improve evaluation methodology, and extend the application of this system to scientific reporting validation in sustainable tourism studies, environmental assessment reports, and other academic publications related to green tourism.

Acknowledgment

The authors express their gratitude to the Institute for Research and Community Service (P3M) of Politeknik Negeri Bali for funding support through the PNB DIPA Regular Research Scheme for the Fiscal Year 2026, as well as to all students of the TRPL Program who participated as respondents in this study.

References

- Ambara, M. P., Sudiarta, M., & Suryaniadi, S. M. (2024). Digitalization test materials and information system for scheduling competency test sites at Politeknik Negeri Bali's professional certification institute. *Dalam International Conference on Sustainable Green Tourism Applied Science-Social Applied Science 2024 (ICoSTAS-SAS 2024)* (pg. 856–865). Atlantis Press.
- Andika, E. A., & Renaldy, R. (2024). Perkembangan sistem berbasis artificial intelligence di bidang kesehatan. *Prosiding Seminar Nasional Informatika*, 593–598.
- Assegaf, I. A., Syastra, M. T., Kurniawan, T. R., & Soleh, M. (2024). Pengembangan chatbot konsultasi kesehatan mental berbasis Open AI model GPT-3.5 Turbo menggunakan media WhatsApp. *Jurnal Informatika Teknologi dan Sains (JINTEKS)*, 6(4), 785–793.
- Budiman, A. W., Setiawan, A., & Nugroho, S. (2023). Pengembangan sistem layanan informasi berbasis web dengan memanfaatkan AI pada ChatGPT. *Jurnal Teknologi dan Sistem Informasi Bisnis*, 5(4), 592–602.
- Hia, A. S., Andika, N., & Supratikta, H. (2024). Perancangan sistem informasi penjadwalan kuliah berbasis AI (Artificial Intelligence). *Multidisciplinary Indonesian Center Journal (MICJO)*, 1(3), 1198–1206.
- Jaya, H., Sabran, S., Idris, M., Djawad, Y. A., Ilham, A., & Ahmar, A. S. (2018). Kecerdasan buatan. *Fakultas MIPA Universitas Negeri Makassar*.
- Jibril, M. (2024). Pengujian sistem informasi e-modul pada SMPN 1 Tempuling menggunakan Black Box testing. *Jurnal Perangkat Lunak*, 6(2), 327–332.
- Kaira, M. F. (2022). Implementasi model bahasa OpenAI GPT-3 untuk aplikasi text content generator berbasis web dan aplikasi mobile [Skripsi, Universitas].
- Kesuma, D. P. (2021). Penggunaan metode System Usability Scale untuk mengukur aspek usability pada media pembelajaran daring di Universitas XYZ. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, 8(3), 1615–1626.

- Maulida, R. (2024). Komparasi respons ChatGPT dan Gemini terhadap command pattern identik dengan metode Black Box. *Jurnal Teknik Informatika*, 10(2), 68–71.
- Mawliidy, E. R., Dio, R., & Lorensa, L. (2024). Kemampuan artificial intelligence terhadap pendeteksian fraud: Studi literatur. *Akurasi: Jurnal Studi Akuntansi dan Keuangan*, 7(1), 89–104.
- Ningrum, T. S., Anspratiwi, V. D., & Wahyudi, M. (2025). Etika dan dampak sosial dari penerapan artificial intelligence dalam sistem informasi manajemen. *Jurnal Riset Sistem Informasi*, 2(1), 6–15.
- Pamungkas, B. A., Dharmawan, A., & Yusup, Y. (2024). Development of Android education application for Arabic alphabet with artificial intelligence for children. *Sistemasi: Jurnal Sistem Informasi*, 13(4), 1449–1463.
- Pradini, R. S., Kriswibowo, R., & Ramdani, F. (2019). Usability evaluation on the SIPR website uses the System Usability Scale and Net Promoter Score. Dalam 2019 International Conference on Sustainable Information Engineering and Technology (SIET) (pg. 280–284). IEEE.
- Pramono, W. A., Az-Zahra, H. M., & Rokhmawati, R. I. (2019). Evaluasi usability pada aplikasi MyTelkomsel dengan menggunakan metode usability testing. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 3(3), 2951–2959.
- Pratiwi, N. P. D., Sanjaya, I. B., & Adiningrat, G. P. (2024). Analisis kecepatan dan kehandalan pelayanan jasa perawatan dan perbaikan terhadap tingkat kepuasan pelayanan Unit Penunjang Akademik Perawatan dan Perbaikan Politeknik Negeri Bali. *Jurnal Syntax Admiration*, 5(9), 3512–3524.
- Putra, O. E. (2020). Implementasi artificial intelligence pada sistem pengawasan pasien rumah sakit. *J Teknol*, 10(2), 28–41.
- Rahmadan, M., & Gunawan, C. E. (2024). Perancangan data flow diagram aplikasi tabungan sampah PT Pusri Palembang. *Prosiding Seminar Nasional Mini Riset Mahasiswa*.
- Rifky, S. (2024). Dampak penggunaan artificial intelligence bagi pendidikan tinggi. *Indonesian Journal of Multidisciplinary on Social and Technology*, 2(1), 37–42.
- Sari, V. B. M., Solichin, M. R., Sholikha, S. M., Priastuti, D. N., Cahyani, F., & Noviana, N. E. (2023). Peningkatan kualitas penulisan artikel ilmiah terakreditasi Sinta melalui pelatihan interaktif. *Community Development Journal: Jurnal Pengabdian Masyarakat*, 4(5), 10269–10275.
- Selvina, O., Setyaningrum, V., & Atmaja, D. S. (2023). Pengembangan bahan ajar chatbot berbasis artificial intelligence pada materi sistem pernapasan makhluk hidup di sekolah dasar. *Muallimuna: Jurnal Madrasah Ibtidaiyah*, 8(2), 88–99.
- Sriyeni, Y. (2022). Analisis usability aplikasi investasi digital menggunakan metode Heuristic Evaluation dan System Usability Scale. *Jurnal Ilmiah Informatika Global*, 13(2).
- Subiksa, G. B., Indrayana, I. N. E., Prayudha, I. P. A., Sudiarta, I. K. G., & Waluyo, Y. A. (2025, October). ICOSTAS-EAS 2025, 101–111.
- Suria, O. (2024). A statistical analysis of System Usability Scale (SUS) evaluations in online learning platform. *Journal of Information Systems and Informatics*, 6(2), 992–1007.
- Udil, P. A. (2021). Pelatihan penulisan artikel ilmiah penelitian tindakan kelas untuk publikasi pada jurnal ilmiah. *Jurnal Nasional Pengabdian Masyarakat*, 2(1), 21–27.
- Widiyastuti, N. E., etc. (2023). Inovasi & pengembangan karya tulis ilmiah: Panduan lengkap untuk penelitian dan mahasiswa. PT. Sonpedia Publishing Indonesia.
- Yulianti, G., Bernardi, B., Permana, N., & Wijayanti, F. A. K. W. (2023). Transformasi pendidikan Indonesia: Menerapkan potensi kecerdasan buatan (AI). *Journal of Information Systems and Management (JISMA)*, 2(6), 102–106.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

