Journal of Computer Science and Computational Intelligence
May 2026 | Vol. 2 | No. 1 | Page 47-56
E-ISSN: 3109-7383 P-ISSN: 3110-8393
DOI: https://doi.org/10.30598/algorithm.v2i1.47-56

Klasifikasi Topik Riset Ilmu Komputer di Kawasan ASEAN Menggunakan Algoritma *Random Forest*

(Classification of Computer Science Research Topics in the ASEAN Region Using the Random Forest Algorithm)

Victoria Daniela Melatawun[1], Citra Fathia Palembang[2*], Noval Febrian Pattiasina[3]

[1,2,3] Program Studi Ilmu Komputer, Fakultas Sains dan Teknologi, Universitas Pattimura
[*]Corresponding Author e-mail: *fpchiet@gmail.com

Manuscript submitted: ***15th April 2026*** | ***Manuscript revision:*** ***20th April 2026*** | ***Accepted for publication:*** ***5th May 2026***

Abstract

The growth of scientific publications in computer science across Southeast Asia (ASEAN) over the past decade reflects the increasing research capacity of its member states. However, few studies have systematically mapped the distribution of computer science research topics in this region using a machine learning approach, particularly in identifying under-researched sub-fields (research gaps). This study aims to classify computer science research topics from scientific publications of ASEAN countries for the period 2015–2025, while simultaneously identifying open research gaps. Data were obtained from OpenAlex, an open-access bibliographic database covering more than 245 million global scientific publications, with a total of 43,263 papers collected from 10 ASEAN countries. Each paper was represented using the Term Frequency-Inverse Document Frequency (TF-IDF) method with 10,000 features and classified into 22 computer science sub-fields using a Random Forest algorithm with 200 estimators and a 5-fold cross-validation scheme. Model evaluation yielded an accuracy of 93.87%, a weighted F1-Score of 0.9320, and a Cross-Validation Accuracy of 93.73% ± 0.63%. Artificial Intelligence dominated computer science research across ASEAN, peaking at 1,168 papers in 2020. Computer Vision (F1=0.00), Bioinformatics (F1=0.12), and Robotics (F1=0.18) were identified as the sub-fields with the largest research gaps. It should be noted that sub-field labeling was performed automatically, and therefore manual validation by domain experts remains necessary.

Keywords: *ASEAN; Computer Science; Random Forest; Text Classification; TF-IDF*

How to cite this article:
V. D. Melatawun, C. F. Palembang, and N. F. Pattiasina, "Klasifikasi Topik Riset Ilmu Komputer di Kawasan ASEAN Menggunakan Algoritma Random Forest", *algorithm*, vol. 2, no. 1, pp. 47-56

Journal homepage: https://ojs3.unpatti.ac.id/index.php/algorithm
Research Article Open Access

1. PENDAHULUAN

Ilmu komputer merupakan salah satu bidang yang mengalami pertumbuhan publikasi paling pesat di tingkat global. Menurut data OpenAlex, jumlah publikasi ilmu komputer di seluruh dunia meningkat lebih dari dua kali lipat dalam satu dekade terakhir, didorong oleh perkembangan Kecerdasan Buatan (*Artificial Intelligence*/AI), *Cloud Computing*, *Internet of Things* (IoT), dan *Cybersecurity*[1]. Di kawasan Asia Tenggara (ASEAN), tren serupa mulai terlihat dengan Indonesia, Malaysia, dan Singapura menunjukkan peningkatan signifikan dalam output publikasi ilmiah. Indonesia sendiri tercatat memiliki lebih dari 800.000 *paper* ilmu komputer yang terindeks di OpenAlex. Suranto et al. menemukan bahwa jumlah artikel dari kawasan ini meningkat konsisten dari 306 artikel pada 2018 menjadi 877 artikel pada puncaknya di 2022[2].

Namun, kuantitas publikasi yang tinggi tidak selalu mencerminkan pemerataan penelitian di seluruh sub-bidang ilmu komputer. Sahria menunjukkan bahwa distribusi topik riset di Indonesia sangat terkonsentrasi pada beberapa sub-bidang tertentu [3]. Zhu dan Lei mengidentifikasi bahwa distribusi riset cenderung tidak merata dalam bibliometrik klasifikasi teks global [4]. Kondisi ini menunjukkan bahwa perkembangan riset ilmu komputer tidak hanya perlu dilihat dari jumlah publikasi, tetapi juga dari sebaran tema, cakupan sub-bidang, dan konsistensi kontribusi ilmiah pada setiap area kajian. Ketimpangan distribusi topik dapat menyebabkan beberapa sub-bidang berkembang pesat, sementara sub-bidang lain kurang terpantau dan kurang memperoleh perhatian akademik. Tantangan tambahan muncul dari sistem pelabelan otomatis yang digunakan basis data bibliografi seperti OpenAlex, dimana kata kunci tertentu dapat menyebabkan misklasifikasi lintas domain. Identifikasi *research gap* yang akurat karenanya memerlukan pendekatan yang mempertimbangkan potensi *noise* pelabelan otomatis ini.

Breiman memperkenalkan *Random Forest* sebagai metode *ensemble* yang terbukti efektif untuk klasifikasi teks multikelas [5]. Satria et al. menerapkan *Random Forest* untuk klasifikasi teks berbahasa Indonesia dengan hasil memuaskan [6]. Wati et al. membuktikan efektivitas TF-IDF untuk klasifikasi sentimen masyarakat Indonesia [7]. Goh et al. menunjukkan bahwa *machine learning* untuk klasifikasi abstrak penelitian mampu mencapai performa kompetitif dengan anotasi manusia[8]. Berdasarkan karakteristik tersebut, kombinasi TF-IDF dan *Random Forest* menjadi pendekatan yang relevan untuk mengolah data tekstual bibliografis, khususnya dalam mengubah informasi judul, abstrak, dan kata kunci menjadi representasi numerik yang dapat diklasifikasikan secara otomatis. Pendekatan ini juga memberikan keuntungan karena relatif mudah diterapkan, memiliki interpretasi yang lebih jelas dibandingkan beberapa model kompleks, serta mampu menangani data dengan jumlah kelas yang beragam.

Penelitian ini memiliki tiga kontribusi utama: (1) membangun model klasifikasi otomatis sub-bidang ilmu komputer dari publikasi ASEAN menggunakan *Random Forest* dan TF-IDF, (2) memetakan distribusi topik riset per negara ASEAN dan tren per tahun 2015–2025, dan (3) mengidentifikasi *research gap* secara kuantitatif sekaligus mendokumentasikan tantangan *noise* pelabelan otomatis lintas domain sebagai temuan metodologis yang penting. Dengan demikian, penelitian ini tidak hanya berkontribusi pada pengembangan model klasifikasi teks berbasis *machine learning*, tetapi juga menyediakan gambaran empiris mengenai arah perkembangan riset ilmu komputer di kawasan ASEAN. Hasil penelitian diharapkan dapat menjadi dasar awal bagi peneliti, institusi pendidikan tinggi, dan pembuat kebijakan dalam memahami sub-bidang yang dominan maupun sub-bidang yang masih kurang mendapat perhatian.

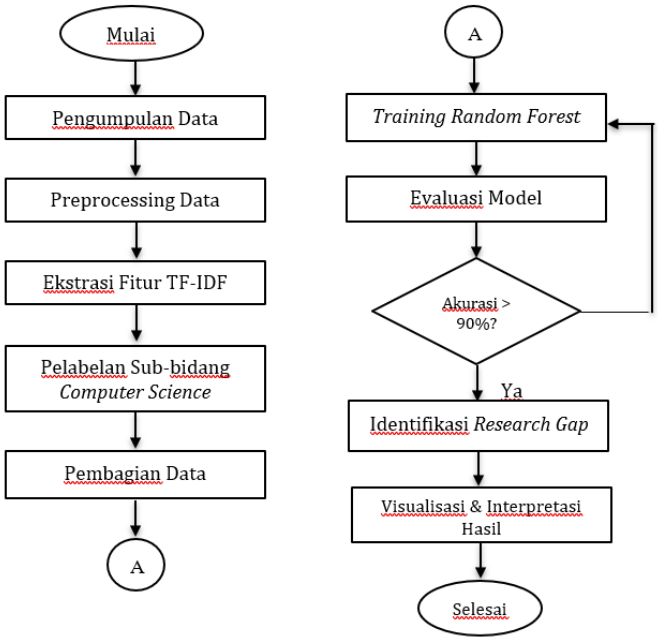
2. METODE PENELITIAN

Metodologi penelitian ini melibatkan serangkaian tahapan yang meliputi pengumpulan data,

pre-processing, ekstraksi fitur, pelabelan sub-bidang, pembagian data, klasifikasi, evaluasi model, serta visualisasi dan interpretasi hasil. Setiap tahapan dilakukan secara sistematis untuk memastikan bahwa data publikasi yang digunakan dapat diolah menjadi informasi yang representatif dalam memetakan topik riset ilmu komputer di kawasan ASEAN. Proses pengolahan data diawali dengan pengumpulan data bibliografi, kemudian dilanjutkan dengan pembersihan data dan transformasi teks agar dapat digunakan dalam proses klasifikasi berbasis *machine learning*.

Tahapan ekstraksi fitur dilakukan menggunakan metode TF-IDF untuk mengubah data teks menjadi representasi numerik. Selanjutnya, model *Random Forest* digunakan untuk mengklasifikasikan publikasi ke dalam sub-bidang ilmu komputer yang telah ditentukan. Hasil klasifikasi kemudian dievaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* untuk menilai kemampuan model dalam mengenali masing-masing kategori. Setelah model memperoleh performa yang memadai, hasil klasifikasi digunakan untuk mengidentifikasi distribusi topik riset, tren publikasi, serta potensi *research gap* pada berbagai sub-bidang ilmu komputer di kawasan ASEAN.

Secara umum, alur penelitian ini dirancang untuk menggambarkan hubungan antara proses pengolahan data, pembentukan model klasifikasi, evaluasi performa, hingga interpretasi hasil penelitian. Visualisasi alur kerja diperlukan agar setiap tahap penelitian dapat dipahami secara runtut dan menunjukkan keterkaitan antara proses teknis dan tujuan analisis. Tahapan-tahapan tersebut ditampilkan pada Gambar 1.



Gambar 1. *Flowchart* Tahapan Penelitian

2.1 Pengumpulan Data

Data dikumpulkan dari OpenAlex API menggunakan filter bidang ilmu komputer (Concept ID: C41008148) untuk 10 negara ASEAN pada periode 2015–2025. *Filter Concept* ID OpenAlex berbasis deteksi kata kunci otomatis sehingga terdapat kemungkinan paper dari disiplin lain ikut terkumpul apabila mengandung terminologi ilmu komputer dalam abstraknya [1].

Tabel 1. Jumlah paper yang dikumpulkan per negara ASEAN

Negara	*Paper* Tersedia	*Paper* Diambil
Indonesia	805.968	5.000
Malaysia	182.201	5.000
Singapura	108.321	5.000
Thailand	68.398	5.000
Vietnam	55.010	5.000
Filipina	39.677	5.000
Myanmar	4.079	4.079
Kamboja	4.768	4.768
Brunei	3.542	3.542
Laos	874	874
Total	**1.272.838**	**43.263**

2.2 *Preprocessing* dan Ekstrasi Fitur TF-IDF

Setiap paper direpresentasikan sebagai gabungan teks dari judul, abstrak, dan konsep OpenAlex. Label klasifikasi ditentukan berdasarkan sub-bidang ilmu komputer utama dari kolom concepts OpenAlex. Teks diubah menjadi vektor numerik menggunakan TF-IDF dengan konfigurasi: maksimal 10.000 fitur, *stop words* bahasa Inggris, n-gram range (1,2), dan *minimum document frequency* 2. Label dengan kurang dari 30 sampel dibuang.

2.3 Pelatihan dan Evaluasi Model

Data dibagi 80% data latih dan 20% data uji dengan *stratified sampling*. Model *Random Forest* dilatih dengan 200 estimator dan *class_weight* = *balanced*[5]. Apabila akurasi belum mencapai 90%, dilakukan *tuning hyperparameter* dan proses diulangi. Evaluasi menggunakan metrik *Accuracy*, *Precision*, *Recall*, F1-*Score*, *Confusion Matrix*, dan *Cross-Validation* 5-*fold*.

2.4 Identifikasi Research Gap

Research gap diidentifikasi melalui dua indikator: (1) sub-bidang dengan F1-*score* rendah mengindikasikan minimnya data representatif, dan (2) analisis distribusi *paper* per sub-bidang secara keseluruhan ASEAN, bukan per negara secara spesifik mengingat potensi noise pelabelan otomatis lintas domain [4].

3. HASIL DAN PEMBAHASAN

3.1 Performa Model

Performa Model *Random Forest* ditunjukkan pada Tabel 2. Akurasi 93,87% dan *Cross-Validation Accuracy* 93,73% ± 0,63% menunjukkan model tidak mengalami *overfitting*. Nilai standar deviasi kecil (0,63%) mengindikasikan kestabilan model yang tinggi. Mengacu pada rumus *Accuracy* (1) dan *F1-Score Weighted* (5), nilai-nilai ini mencerminkan performa model yang konsisten di semua kelas. Hasil ini sebanding dengan temuan Goh et al. [8] dan melampaui hasil Satria et al. [6] dalam klasifikasi teks berbahasa Indonesia.

Tabel 2. Hasil evaluasi model *Random Forest*

Metrik Evaluasi	Nilai
Accuracy	93,87%
F1-Score (Weighted)	0,9320
CV Accuracy (5-fold)	93,73% ± 0,63%

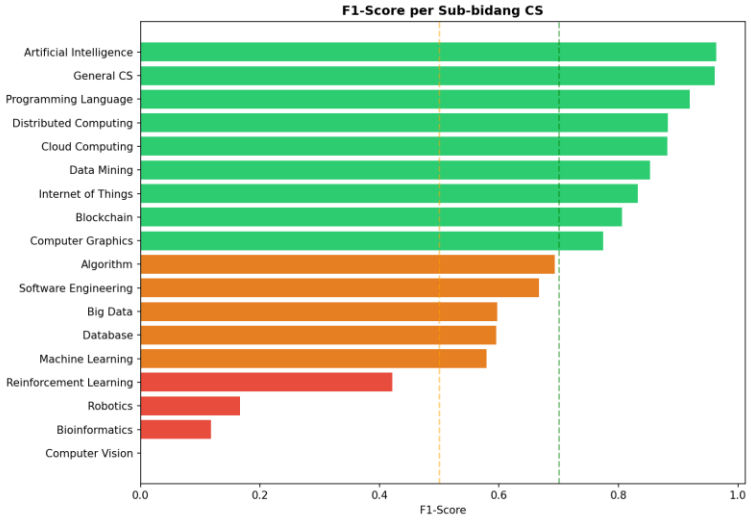
Metrik Evaluasi	Nilai
Jumlah Estimator	200 pohon
Jumlah Fitur TF-IDF	10.000
Split Data	80% Train / 20% Test

3.2 Performa per Sub-bidang

Tabel 3 merangkum nilai *Precision, Recall*, dan F1-*score* untuk setiap sub-bidang, dan Gambar 2 menampilkan visualisasi F1-Score per sub-bidang dengan kode warna: hijau (F1 ≥ 0,7), oranye (0,5 ≤ F1 < 0,7), dan merah (F1 < 0,5).

Tabel 3. Rangkuman Nilai nilai *Precision*, *Recall*, dan F1-*score*

Sub-bidang	*Precision*	*Recall*	*F1-Score*	*Support*
Artificial Intelligence	0,94	0,99	0,96	1.753
General Computer Science (CS)	0,94	0,98	0,96	5.791
Programming Language	1,00	0,85	0,92	84
Distributed Computing	0,93	0,84	0,88	76
Cloud Computing	0,92	0,85	0,88	84
Data Mining	0,86	0,85	0,85	106
Internet of Things	0,81	0,85	0,83	108
Blockchain	0,91	0,72	0,81	40
Computer Graphics	0,92	0,67	0,77	18
Algorithm	0,88	0,57	0,69	256
Software Engineering	0,00	0,85	0,68	-
Big Data	0,95	0,43	0,60	46
Database	0,96	0,43	0,60	123
Machine Learning	1,00	0,41	0,58	81
Reinforcement Learning	-	-	0,43	-
Robotics	-	-	0,18	-
Bioinformatics	1,00	0,06	0,12	16
Computer Vision	0,00	0,00	0,00	7

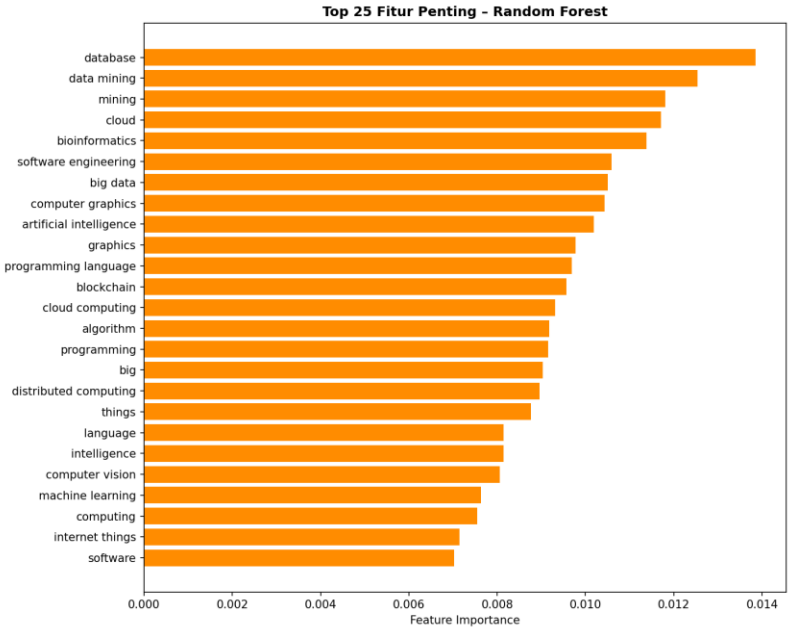


Gambar 2. F1-Score per Sub-bidang

Dari Gambar 3. *Artificial Intelligence* (F1=0,96) dan *General CS* (F1=0,96) memiliki performa terbaik. Rendahnya F1-Score pada *Database* (0,60) dan *Machine Learning* (0,58) sebagian dapat disebabkan *noise* pelabelan otomatis, di mana kata kunci "*database*" atau "learning" muncul dalam paper dari disiplin non-ilmu komputer. Konsisten dengan temuan Xiao et al.[9], kata dengan diskriminabilitas kategori tinggi berkontribusi paling besar dalam klasifikasi *multiclass*.

3.3 *Feature Importance*

Tabel 3. merangkum nilai Precision, Recall, dan F1-Score untuk setiap sub-bidang. **Gambar 2** menampilkan visualisasi F1-Score per sub-bidang dengan kode warna: hijau (F1 ≥ 0,7), oranye (0,5 ≤ F1 < 0,7), dan merah (F1 < 0,5). Gambar 4**.** menampilkan 25 fitur paling berpengaruh dalam model *Random Forest*.

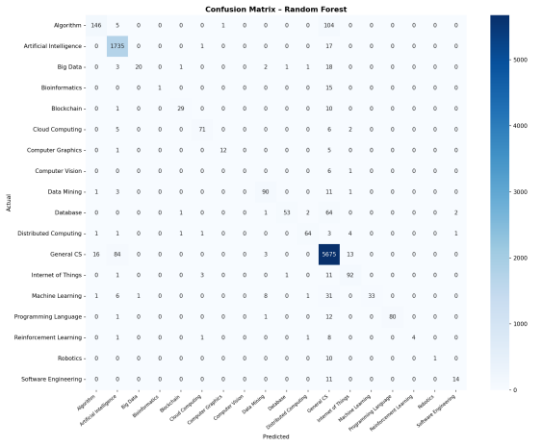


Gambar 3. Top 25 fitur paling berpengaruh dalam model *Random Forest*

Fitur paling berpengaruh adalah "database" (0,014), "data mining" (0,013), "mining" (0,010), dan "cloud" (0,010). Tingginya feature importance kata "database" mengkonfirmasi fenomena noise pelabelan. Kata ini sangat diskriminatif secara statistik karena muncul di berbagai konteks lintas domain. Konsisten dengan Xiao et al. [9], terminologi domain spesifik menjadi pembeda utama antar kelas.

3.4 *Confusion Matrix* dan Analisis Kesalahan Klasifikasi

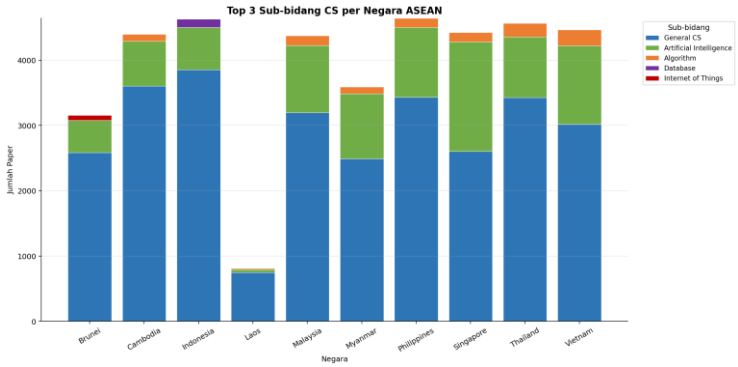
General CS memiliki prediksi benar tertinggi (5.675 paper). *Artificial Intelligence* sangat akurat (1.735 benar). Kesalahan terbesar terjadi pada *Algorithm* yang sering salah ke General CS (104 kasus), sebagian dapat dijelaskan oleh *noise* pelabelan pada paper matematika dan fisika. Sub-bidang *Computer Vision* (7 sampel) dan *Bioinformatics* (16 sampel) hampir seluruhnya salah diklasifikasikan akibat ketidakseimbangan data yang ekstrem, mengkonfirmasi nilai *Recall* rendah. Gambar 4 menampilkan *confusion matrix* hasil prediksi model pada data uji.



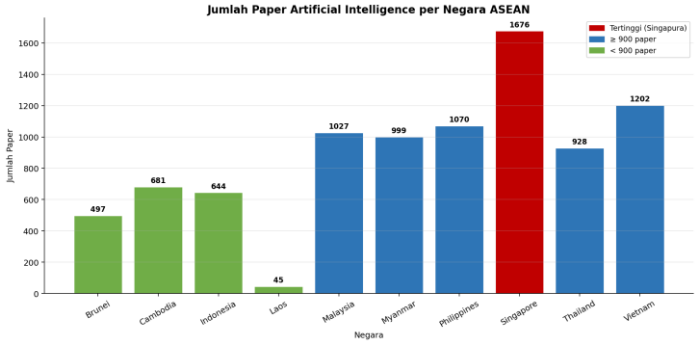
Gambar 4. *Confusion matrix*

3.5 Distribusi Topik Riset per Negara ASEAN

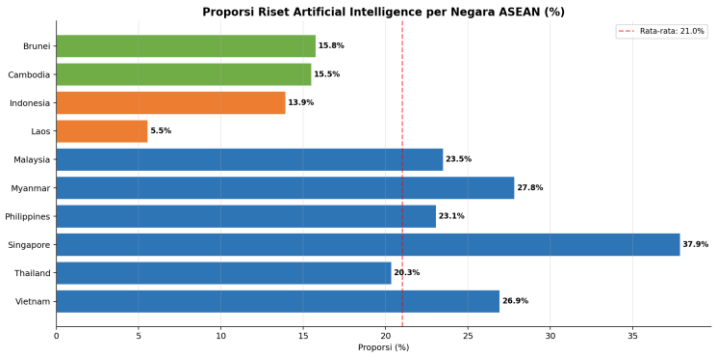
Gambar 5, 6, dan 7 menampilkan distribusi sub-bidang ilmu komputer per negara ASEAN. Interpretasi perlu mempertimbangkan potensi noise pelabelan otomatis lintas domain [4] .



Gambar 5. Distribusi top 3 sub-bidang Ilmu Komputer per negara ASEAN



Gambar 6. Proporsi riset *artificial intelligenc*e per negara ASEAN (%)

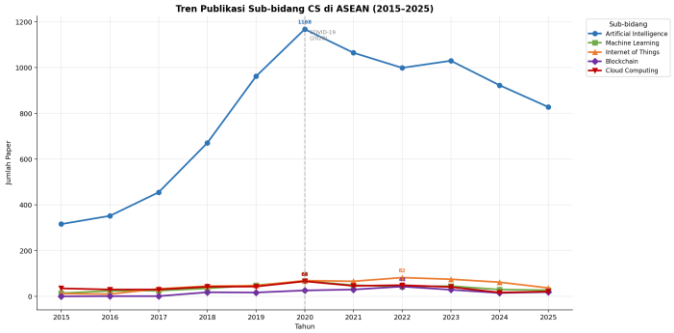


Gambar 7. Proporsi riset *artificial intelligenc*e per negara ASEAN (%)

Dari Gambar 5, 6, dan 7, General CS mendominasi di semua negara ASEAN. *Artificial Intelligence* secara konsisten menjadi sub-bidang terbesar kedua dengan Singapura memimpin (1.676 paper), diikuti Vietnam (1.202), Filipina (1.070), dan Malaysia (1.027). Temuan ini relatif lebih dapat dipercaya karena terminologi AI cukup spesifik. Sub-bidang ketiga di sebagian besar negara adalah *Algorithm*, namun validasi manual diperlukan mengingat terminologi generiknya [3]. Singapura memiliki proporsi AI tertinggi.

3.6 Tren Topik Riset Ilmu Komputer per Tahun (2015–2025)

Gambar 8 menampilkan tren jumlah paper untuk 5 sub-bidang Ilmu Komputer terpilih di kawasan ASEAN.



Gambar 8. Tren publikasi sub-bidang Ilmu Komputer ASEAN (2015–2025)

Tabel 4. Tren Jumlah Paper Sub-Bidang Ilmu Komputer Terpilih di ASEAN

Sub-bidang	2015	2018	2020	2022	2024	2025
Artificial Intelligence	316	671	1.168	999	923	828
Internet of Things	12	45	69	82	62	37
Blockchain	0	18	26	43	15	19
Cloud Computing	35	42	66	49	17	21
Machine Learning	14	35	67	43	30	27

Artificial Intelligence mengalami pertumbuhan paling signifikan dari 316 paper (2015) menjadi puncaknya 1.168 paper pada 2020. IoT tumbuh stabil dari 12 paper (2015) menjadi 82 paper (2022). Blockchain tumbuh dari nol paper (2015) menjadi 43 paper (2022). *Cloud Computing* mengalami penurunan signifikan pada 2024–2025. Data tren ini menggunakan sub-bidang dengan terminologi spesifik sehingga lebih tidak rentan terhadap *noise* pelabelan dibandingkan sub-bidang dengan

terminologi generik seperti Database atau Algorithm [4].

3.7 *Research Gap*

Berdasarkan gabungan analisis F1-*Score*, *feature importance*, *confusion matrix*, dan distribusi per negara dengan mempertimbangkan potensi *noise* pelabelan otomatis *research gap* teridentifikasi.

Tabel 5. Pemetaan *Research Gap*

Tingkat Gap	*Sub-bidang*	*F1-Score*	*Support*	*Indikasi*
Sangat Tinggi	Computer Vision	0,00	7	Hampir tidak ada paper
Sangat Tinggi	Bioinformatics	0,12	16	Sangat sedikit paper
Sangat Tinggi	Robotics	0,18	-	Kurang representasi
Tinggi	Reinforcement Learning	0,43	-	Data tidak seimbang
Sedang	*Machine Learning*	0,58	81	Sebagian noise pelabelan
Sedang	*Database & Big Data*	0,60	46–123	Sebagian noise pelabelan

Computer Vision, *Bioinformatics*, dan *Robotics* merupakan *research gap* paling valid karena terminologinya sangat spesifik sehingga tidak rentan terhadap *noise* pelabelan lintas domain. *Machine Learning* dan *Database* ditandai asterisk (*) karena sebagian rendahnya *F1-Score* kemungkinan disebabkan *noise* pelabelan otomatis bukan semata-mata karena sedikitnya riset [3], [4].

Penelitian ini memiliki beberapa keterbatasan. Pertama dan paling signifikan, pelabelan sub-bidang dilakukan secara otomatis berdasarkan metadata OpenAlex tanpa validasi manual oleh pakar domain [10]. Investigasi terhadap sampel data menunjukkan adanya *noise* pelabelan lintas domain yang cukup signifikan. Paper dari biologi, ekologi, pendidikan, dan fisika ikut terkumpul karena mengandung kata kunci Ilmu Komputer dalam abstraknya. Hal ini memengaruhi akurasi analisis distribusi per sub-bidang, terutama untuk sub-bidang dengan terminologi generik[1]. Kedua, data tahun 2025 belum sepenuhnya terindeks. Ketiga, negara-negara ASEAN dengan populasi akademik kecil seperti Laos (874 paper), Brunei (3.542 paper), dan Myanmar (4.079 paper) memiliki representasi tidak proporsional dibandingkan Indonesia (805.968 paper). Keempat, penelitian ini hanya mencakup *paper* berbahasa Inggris. Keterbatasan pertama membuka peluang penelitian lanjutan yang mengintegrasikan validasi manual atau teknik filtering berbasis konten yang lebih canggih.

4. KESIMPULAN

Penelitian ini berhasil mengklasifikasikan topik riset Ilmu Komputer dari 43.263 paper ASEAN (2015–2025) menggunakan *Random Forest* berbasis *Term Frequency-Inverse Document Frequency* (TF-IDF) dengan akurasi 93,87%, *F1-Score* 0,9320, dan *Cross-Validation Accuracy* 93,73% ± 0,63%. *Artificial Intelligence* mendominasi riset Ilmu Komputer di seluruh ASEAN dengan puncak 1.168 paper pada 2020, dan Singapura memiliki proporsi riset *Artificial Intelligence* tertinggi. *Computer Vision* (F1=0,00), *Bioinformatics* (F1=0,12), dan *Robotics* (F1=0,18) teridentifikasi sebagai sub-bidang dengan *research gap* terbesar dan paling valid, sementara *Machine Learning* dan *Database* memerlukan validasi manual lebih lanjut karena potensi *noise* pelabelan otomatis lintas domain. Temuan metodologis penting penelitian ini adalah teridentifikasinya fenomena *noise* pelabelan otomatis yang signifikan pada data OpenAlex. Penelitian selanjutnya disarankan menggunakan model *deep learning* berbasis *Bidirectional Encoder Representations from Transformers* (BERT), menerapkan teknik *oversampling*, mengintegrasikan validasi manual untuk mengurangi *noise* pelabelan, serta memperluas cakupan ke publikasi berbahasa lokal ASEAN.

REFERENSI

[1] J. Priem, H. Piwowar, and R. Orr, "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," vol. 27330, pp. 2018–2022, 2022, [Online]. Available: http://arxiv.org/abs/2205.01833.

[2] B. Suranto *et al.*, "State of digitalization in the Southeast Asia region – bibliometric analysis," *Qual. Quant.*, pp. 1053–1080, 2025, doi: 10.1007/s11135-025-02296-3.

[3] Y. Sahria, "Implementasi Teknik Web Scraping pada Jurnal SINTA Untuk Analisis Topik Penelitian Kesehatan Indonesia," *URECOL (Unversity Res. Colloqium)*, pp. 297–306, 2020, [Online]. Available: http://repository.urecol.org/index.php/proceeding/article/view/1079.

[4] H. Zhu and L. Lei, "The Research Trends of Text Classification Studies (2000–2020): A Bibliometric Analysis," *SAGE Open*, vol. 12, no. 2, 2022, doi: 10.1177/21582440221089963.

[5] L. BREIMAN, "RFRSF: Employee Turnover Prediction Based on *Random Forest*s and Survival Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[6] W. Satria and M. Riasetiawan, "Essay Answer Classification With Smote *Random Forest* and Adaboost in Automated Essay Scoring," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 17, no. 4, p. 359, 2023, doi: 10.22146/ijccs.82548.

[7] R. Wati, S. Ernawati, and H. Rachmi, "Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH," *J. Manaj. Inform.*, vol. 13, no. 1, pp. 84–93, 2023, doi: 10.34010/jamika.v13i1.9424.

[8] Y. C. Goh, X. Q. Cai, W. Theseira, G. Ko, and K. A. Khor, "Evaluating human versus machine learning performance in classifying research abstracts," *Scientometrics*, vol. 125, no. 2, pp. 1197–1212, 2020, doi: 10.1007/s11192-020-03614-2.

[9] L. Xiao, Q. Li, Q. Ma, J. Shen, Y. Yang, and D. Li, *Text classification algorithm of tourist attractions subcategories with modified TF-IDF and Word2Vec*, vol. 19, no. 10 October. 2024.

[10] A. Aghaei Chadegani *et al.*, "A comparison between two main academic literature collections: Web of science and scopus databases," *Asian Soc. Sci.*, vol. 9, no. 5, pp. 18–26, 2013, doi: 10.5539/ass.v9n5p18.