

A MACHINE LEARNING-BASED CONCEPTUAL FRAMEWORK FOR
AUTOMATED DIGITAL WASTE CLASSIFICATION WITH PROJECTED
CARBON FOOTPRINT REDUCTION

1) Burhanuddin

1) Program Studi Teknik Informatika, Sekolah Tinggi Manajemen Informatika dan Komputer Bina
Mulia Palu

1) boer.masse@binamulia.ac.id

INFO ARTIKEL	ABSTRAK
Riwayat Artikel : Diterima : 8 Juli 2026 Disetujui : 14 Agustus 2026	Pertumbuhan volume data digital yang pesat turut meningkatkan proporsi sampah digital serta berkontribusi terhadap emisi karbon pusat data, sementara penelitian klasifikasi berkas yang ada umumnya belum menghubungkan hasil klasifikasi dengan estimasi dampak lingkungan secara terukur. Penelitian ini bertujuan mengusulkan Automated Digital Waste Classification (ADWC), sebuah kerangka kerja konseptual berbasis pembelajaran mesin. Penelitian ini menggunakan pendekatan Design Science Research (DSR) dengan evaluasi kelayakan artefak melalui sintesis literatur dan proyeksi berbasis benchmark, mencakup arsitektur lima lapisan yang mengintegrasikan fitur metadata. Hasil proyeksi menunjukkan bahwa konfigurasi penuh ADWC mencapai macro-F1 tertinggi melampaui konfigurasi fitur tunggal maupun kombinasi parsial, serta kompetitif terhadap model benchmark termutakhir seperti CNN-BERT Hybrid. Pada korpus asumsi ±30 TB, ADWC-Full diproyeksikan mengidentifikasi 38,7% data sebagai limbah digital yang dapat ditindaklanjuti, dengan potensi reduksi emisi karbon yang meningkat proporsional terhadap skala implementasi. Analisis sensitivitas mengungkap bahwa intensitas karbon jaringan listrik merupakan parameter paling berpengaruh terhadap besarnya manfaat lingkungan jauh melampaui pengaruh PUE, intensitas energi penyimpanan, maupun fraksi limbah teridentifikasi. Temuan ini menegaskan kelayakan konseptual ADWC sebagai kerangka kerja yang menghubungkan tata kelola data pada tingkat berkas dengan indikator keberlanjutan yang terukur meskipun seluruh hasil masih bersifat proyeksi berbasis benchmark dan memerlukan validasi empiris pada lingkungan operasional nyata.
Kata Kunci: Design science research; Jejak karbon; klasifikasi pembelajaran mesin; Sampah digital; Tata kelola data	

ARTICLE INFO	ABSTRACT
Article History : Received : Jul 8, 2026 Accepted : Aug 14, 2026	<i>The rapid growth in digital data volume has increased the proportion of digital waste and contributed to carbon emissions from data centers, while existing file classification research generally has not linked classification results to measurable estimates of environmental impact. This study aims to propose Automated Digital Waste Classification (ADWC), a machine learning-based conceptual framework. This study employs a Design Science Research (DSR) approach, evaluating the feasibility of the artifact through literature synthesis and benchmark-based projections, encompassing a five-layer architecture that integrates metadata features. The projection results show that the full ADWC configuration achieves the highest macro-F1 score, surpassing both single-feature and partial-combination configurations, and is</i>
Keywords: Carbon footprint; Design science research; Digital waste; Data governance; Machine learning classification	

competitive against state-of-the-art benchmark models such as the CNN-BERT Hybrid. On a ± 30 TB test corpus, ADWC-Full is projected to identify 38.7% of the data as actionable digital waste, with the potential for carbon emission reductions increasing proportionally with the scale of implementation. Sensitivity analysis reveals that the carbon intensity of the power grid is the most influential parameter on the magnitude of environmental benefits, far surpassing the influence of PUE, storage energy intensity, or the fraction of identified waste. These findings confirm the conceptual feasibility of ADWC as a framework that links data governance at the file level to measurable sustainability indicators, although all results are still benchmark-based projections and require empirical validation in real operational environments.

1. PENDAHULUAN

Memasuki abad ke-21, dunia mengalami pertumbuhan data digital yang belum pernah terjadi sebelumnya. Volume ruang data global diproyeksikan meningkat dari 33 zettabyte (ZB) pada tahun 2018 menjadi sekitar 175 ZB pada tahun 2025, didorong oleh adopsi komputasi awan, Internet of Things (IoT), kecerdasan buatan, dan jaringan 5G (Johns, 2023; Matendo Didas, 2023). Namun, peningkatan volume data tersebut tidak diikuti dengan pemanfaatan yang sebanding. Hanya sekitar 2–3% data yang dipertahankan untuk kebutuhan jangka panjang, sedangkan sebagian besar bersifat sementara atau tetap tersimpan tanpa pernah dimanfaatkan kembali (Coughlin, 2018). Fenomena ini mencerminkan Law of Accelerating Returns, yaitu kemampuan menghasilkan data berkembang jauh lebih cepat dibandingkan kemampuan organisasi dalam mengelola dan mengekstraksi nilainya, sehingga memberikan tekanan yang semakin besar terhadap kapasitas dan keberlanjutan infrastruktur penyimpanan global (Hartung, 2018).

Pertumbuhan data yang masif juga memunculkan konsekuensi lingkungan yang semakin signifikan. Industri pusat data diperkirakan mengonsumsi sekitar 1–1,5% kebutuhan listrik dunia dan menghasilkan hampir 2% emisi gas rumah kaca global, atau setara dengan jejak karbon sektor penerbangan (Ewim et al., 2023; Thangam et al., 2024). Tanpa strategi pengelolaan yang berkelanjutan, penyimpanan data digital diproyeksikan menyumbang hingga 14% emisi global pada tahun 2040. Kondisi tersebut diperparah oleh meningkatnya jumlah pusat data, dari sekitar 500.000 unit pada tahun 2012 menjadi lebih dari

8 juta unit, sementara konsumsi energinya terus berlipat ganda setiap empat tahun (Thangam et al., 2024).

Di sisi lain, ketergantungan pusat data pada energi tak terbarukan masih tinggi (Ewim et al., 2023). Sehingga sebagian besar emisi justru berasal dari dark data, yaitu data yang tidak lagi memiliki nilai guna tetapi tetap tersimpan dan terus mengonsumsi energi (Al Kez et al., 2022; Jackson and Hodgkinson, 2023). Fenomena tersebut mendorong munculnya konsep sampah digital (digital waste) sebagai isu penting dalam tata kelola data dan keberlanjutan lingkungan. Sampah digital didefinisikan sebagai data yang berlebihan, usang, atau tidak lagi diperlukan sehingga menurunkan efisiensi penyimpanan dan meningkatkan biaya pengelolaan data (Mersico et al., 2024).

Konsep ini mengadaptasi waste hierarchy for digital objects yang menekankan prinsip reduce, reuse, dan recycle dalam pengelolaan data. Dalam praktiknya, sampah digital mencakup file duplikat, arsip surat elektronik yang telah kedaluwarsa, system log, cache, file sementara, hingga cadangan (backup) yang terlupakan (Dr. A. Shaji George et al., 2023). Akumulasi data tersebut memunculkan fenomena Big Garbage In, Big Garbage Out, yaitu kondisi ketika data bernilai rendah justru meningkatkan biaya operasional, risiko kepatuhan, keamanan informasi, dan reputasi organisasi (Tripathi et al., 2019).

Permasalahan tersebut sejalan dengan paradigma Green Computing, yang menempatkan efisiensi energi dan optimalisasi sumber daya sebagai prinsip utama dalam pengembangan sistem Big Data dan IoT (Shuja et al., 2017; Naeem et al., 2023). Salah satu

perkembangannya, yaitu Green Big Data Analytics, berupaya mengintegrasikan kinerja analitik dengan efisiensi energi dan sumber daya (Wu et al., 2016; Guo et al., 2023). Meskipun demikian, sebagian besar penelitian masih berfokus pada efisiensi komputasi maupun infrastruktur pusat data, sedangkan pengelolaan sampah digital sebagai sumber pemborosan energi belum memperoleh perhatian yang memadai (Guo et al., 2023).

Telaah literatur menunjukkan sedikitnya empat kesenjangan penelitian yang masih terbuka. Pertama, penelitian deduplikasi data umumnya hanya berfokus pada penghapusan salinan identik, sementara dark data yang unik tetapi tidak lagi bernilai belum menjadi perhatian utam (Yuan, Moh and Moh, 2022; Kaur, Jit Singh and Chana, 2026). Kedua, penelitian klasifikasi dan pengorganisasian berkas otomatis lebih diarahkan pada peningkatan produktivitas pengguna tanpa mempertimbangkan dampak lingkungan sebagai indikator keberhasilan sistem (Ali Raza et al., 2022; Devindi and Piyumal, 2025).

Ketiga, banyak kerangka kerja manajemen siklus hidup data masih mengandalkan aturan berbasis heuristik sehingga kurang adaptif terhadap karakteristik data modern (Vamsi Krishna Vemulapalli, 2025; Zahir Sayyed, 2026). Keempat, pembahasan dark data masih didominasi oleh perspektif nilai bisnis, padahal penyimpanannya diperkirakan menghasilkan sekitar 5,26 juta ton CO₂e, mengonsumsi 41,65 gigaliter air, dan memanfaatkan lahan hingga 59,45 km² setiap tahun (Al Kez et al., 2022; Dr. A. Shaji George et al., 2023; Marumolwa and Marnewick, 2025).

Berdasarkan kesenjangan tersebut, penelitian ini dapat berkontribusi pada integrasi dalam tindak lanjut pengklasifikasian sampah digital dan estimasi dampak lingkungan. Lucivero (2020) Juga menginformasikan bahwa aspek keberlanjutan lingkungan belum menjadi bagian yang terintegrasi dalam tata kelola maupun evaluasi berbagai inisiatif big data. Oleh karena itu, penelitian ini mengusulkan Automated Digital Waste Classification (ADWC) sebagai kerangka kerja konseptual untuk mengidentifikasi, mengklasifikasikan, dan memprioritaskan sampah digital pada lingkungan penyimpanan yang heterogen

melalui integrasi Natural Language Processing (NLP), ekstraksi fitur metadata, dan pembelajaran mesin. Dengan demikian, penelitian ini diharapkan dapat menjembatani pengelolaan data modern dengan tujuan keberlanjutan lingkungan melalui pengelolaan sampah digital yang lebih cerdas, adaptif, dan berkelanjutan.

2. METODE

2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan Design Science Research (DSR) yang berorientasi pada pengembangan artefak konseptual untuk menjawab permasalahan pengelolaan sampah digital dalam konteks keberlanjutan lingkungan, dengan mengikuti konsep Design Science Research (DSR) yaitu masalah diidentifikasi, artefak dirancang dan evaluasinya dirumuskan secara eksplisit (Hevner et al., 2004; Peffers et al., 2007; Gregor & Hevner, 2013). Pada tahap ini, evaluasi artefak dibatasi pada sintesis literatur, penalaran desain, dan proyeksi benchmark terhadap penelitian terdahulu sehingga evaluasi lapangan belum dilakukan. Berbeda dengan penelitian eksperimental, artikel ini tidak menyajikan hasil implementasi sistem secara langsung, melainkan mengusulkan desain konseptual yang didukung oleh sintesis literatur dan proyeksi kinerja berbasis benchmark dari penelitian terdahulu. Pendekatan tersebut sejalan dengan praktik umum DSR dalam pengembangan sistem informasi berkelanjutan (Lannelongue et al., 2021; Roussilhe, Ligozat and Quinton, 2023; Savarimuthu et al., 2023).

Pengembangan ADWC dilaksanakan melalui lima tahapan yang saling berkaitan, yaitu: (1) perumusan taksonomi sampah digital berdasarkan sintesis literatur; (2) perancangan arsitektur sistem untuk mengoperasionalkan taksonomi tersebut; (3) penyusunan rancangan implementasi yang meliputi strategi pengumpulan data, rekayasa fitur, pemodelan pembelajaran mesin, dan protokol evaluasi; (4) proyeksi kinerja klasifikasi berdasarkan tolok ukur model sejenis yang telah dipublikasikan; serta (5) pemodelan dampak lingkungan melalui estimasi potensi pengurangan emisi karbon akibat eliminasi sampah digital. Tahapan

tersebut memastikan bahwa artefak yang diusulkan memiliki landasan konseptual yang kuat, dapat direplikasi, dan relevan untuk pengembangan penelitian empiris pada tahap berikutnya.

2.2 Rancangan Konseptual Kerangka Kerja Automated Digital Waste Classification (ADWC)

Penelitian ini hanya mengusulkan rancangan Automated Digital Waste Classification (ADWC) sebagai kerangka kerja

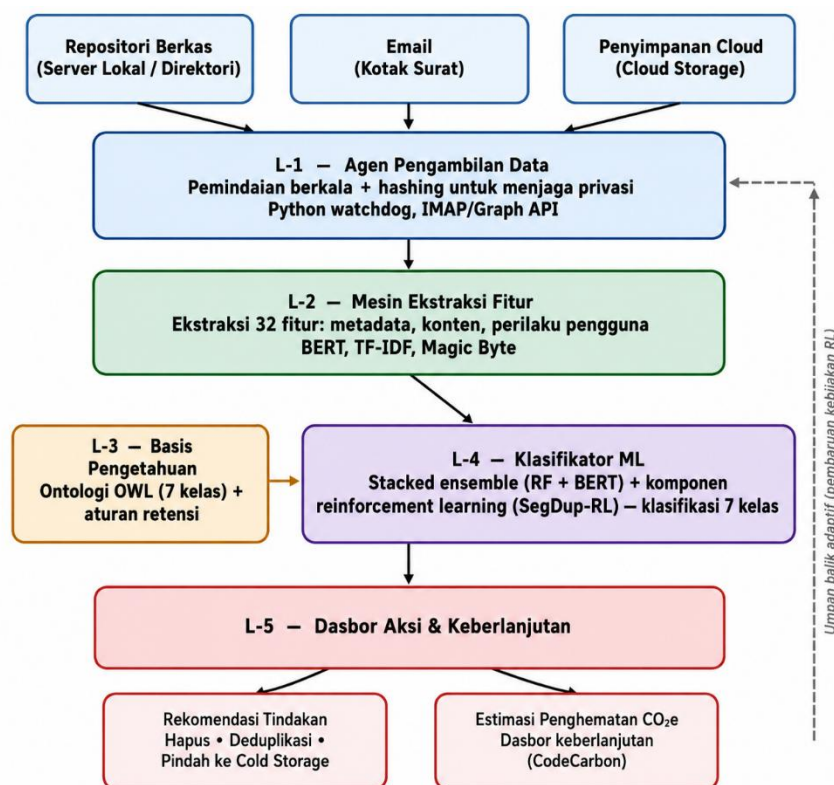
konseptual untuk mengidentifikasi, mengklasifikasikan, dan merekomendasikan pengelolaan sampah digital secara otomatis. Arsitektur ADWC dirancang berdasarkan konsep arsitektur berlapis (multi-layer architecture) yang diadaptasi dari Yuan et al. (2022) dan Devindi & Piyumal (2025). sehingga setiap komponen memiliki fungsi yang terpisah namun saling terintegrasi. Ringkasan fungsi setiap lapisan disajikan pada Tabel 1.

Tabel 1. Ringkasan Konseptual Arsitektur ADWC Lima Lapisan

Lapisan	Nama	Fungsi Utama	Teknologi Kunci
L-1	Agen Pengambilan Data	Memindai direktori, kotak surat, dan koneksi cloud secara berkala	Python <i>watchdog</i> , IMAP/Graph API
L-2	Mesin Ekstraksi Fitur	Mengekstraksi 32 fitur dari metadata, konten, dan perilaku	BERT, TF-IDF, <i>Magic Byte</i>
L-3	Basis Pengetahuan	Ontologi OWL dengan 7 kelas + aturan retensi yang dapat dikonfigurasi	OWL, Sayyed (2025)
L-4	Klasifikator ML	Klasifikasi 7-kelas dengan ensemble tumpukan + komponen RL	RF, BERT, SegDup-RL
L-5	Dasbor Aksi & Keberlanjutan	Merekomendasikan tindakan dan menghitung penghematan CO ₂ e	CodeCarbon, Lannelongue (2021)

Selanjutnya agar lebih jelas, Gambar 1 berikut menyajikan diagram usulan rancangan kerangka kerja arsitektur ADWC lima lapisan yang diusulkan dalam penelitian ini secara skematis untuk memperjelas alur data konseptual secara menyeluruh mulai dari sumber data (repositori berkas, surat elektronik, dan penyimpanan cloud), proses ekstraksi fitur,

keterkaitan dengan basis pengetahuan, klasifikasi berbasis pembelajaran mesin hingga rekomendasi tindakan dan estimasi CO₂e pada dasbor keberlanjutan. Perlu pahami bahwa diagram ini menggambarkan rancangan framework yang diusulkan dalam penelitian ini dan belum diimplementasikan atau diuji secara empiris.



Gambar 1. Rancangan diagram alur data arsitektur ADWC lima lapisan yang diusulkan, dari sumber data (repositori berkas, email, dan penyimpanan cloud) hingga rekomendasi tindakan dan estimasi CO₂e (konseptual, belum diimplementasikan)

Gambar 1 mengilustrasikan rancangan konsep alur data tersebut secara menyeluruh mulai dari proses akuisisi hingga evaluasi dampak lingkungan, sedangkan Tabel 1 telah merangkum konsep fungsi utama dan teknologi kunci setiap lapisan secara. Secara umum lapisan pertama berfungsi sebagai agen akuisisi data yang mengumpulkan metadata dan objek berkas dari berbagai sumber penyimpanan termasuk server lokal, surat elektronik dan penyimpanan awan, dengan tetap menjaga privasi melalui mekanisme hashing. Lapisan kedua mengekstraksi fitur dari tiga sumber informasi utama yaitu; metadata berkas, konten dokumen, dan karakteristik perilaku pengguna. Selanjutnya, lapisan ketiga merepresentasikan pengetahuan domain dalam bentuk ontologi OWL yang memuat taksonomi sampah digital beserta aturan retensi berdasarkan praktik industri dan regulasi yang berlaku (Tripathi et al., 2019; Zahir Sayyed, 2026). Proses klasifikasi dilakukan pada konsep lapisan keempat menggunakan pendekatan stacked ensemble yang mengombinasikan model pembelajaran mesin tradisional, model berbasis Transformer, serta komponen reinforcement learning untuk

menyesuaikan kebijakan deduplikasi secara adaptif terhadap perubahan pola akses data. Hasil klasifikasi kemudian diteruskan ke lapisan kelima untuk menghasilkan rekomendasi tindakan, seperti penghapusan, deduplikasi, atau pemindahan ke penyimpanan dingin (cold storage) sekaligus mengestimasi potensi pengurangan jejak karbon yang ditampilkan melalui dasbor keberlanjutan (Chundru and Mudunuri, 2025; Zahir Sayyed, 2026). Dengan pemisahan fungsi tersebut rancangan arsitektur ADWC diperkirakan dapat menyediakan alur pengelolaan sampah digital yang terintegrasi mulai dari akuisisi data hingga evaluasi dampak lingkungan.

2.3 Rancangan Implementasi dan Evaluasi ADWC

Rancangan implementasi ADWC digunakan untuk memvalidasi kelayakan konseptual kerangka kerja melalui pemanfaatan data organisasi yang merepresentasikan berbagai karakteristik lingkungan penyimpanan digital. Data yang direncanakan mencakup metadata berkas, konten dokumen, serta riwayat aktivitas pengguna yang diperoleh dari repositori penyimpanan lokal maupun layanan

penyimpanan awan. Seluruh data diproses sesuai prinsip *privacy-by-design* dengan menghilangkan informasi yang dapat mengidentifikasi individu dan menerapkan mekanisme hashing pada atribut yang bersifat sensitif.

Proses klasifikasi diawali dengan ekstraksi tiga kelompok fitur, yaitu metadata berkas (misalnya ukuran, tipe, waktu modifikasi, dan frekuensi akses), representasi semantik isi dokumen menggunakan model berbasis Bidirectional Encoder Representations from Transformers (BERT), serta indikator perilaku pengguna yang menggambarkan pola penggunaan data. Kombinasi ketiga kelompok fitur tersebut dipilih untuk menangkap karakteristik struktural maupun semantik yang tidak dapat direpresentasikan secara optimal apabila hanya menggunakan salah satu sumber informasi. Seluruh fitur kemudian melalui tahapan prapemrosesan yang meliputi normalisasi atribut numerik, pengkodean variabel kategorikal, penanganan nilai hilang, serta penyeimbangan distribusi kelas guna mengurangi potensi bias selama proses pelatihan model.

Tahap klasifikasi menggunakan pendekatan *stacked ensemble* yang mengintegrasikan model berbasis pohon keputusan dengan model Transformer. Model berbasis pohon dimanfaatkan untuk mempelajari hubungan nonlinier pada metadata dan karakteristik perilaku pengguna, sedangkan BERT digunakan untuk mengekstraksi informasi semantik dari konten dokumen. Probabilitas keluaran kedua model kemudian digabungkan melalui *meta-classifier* sehingga menghasilkan keputusan klasifikasi yang lebih stabil pada kondisi distribusi kelas yang tidak seimbang.

Kinerja klasifikasi dievaluasi menggunakan metrik *precision*, *recall*, *F1-score*, *macro-F1*, *Matthews Correlation Coefficient* (MCC), dan *Cohen's κ* agar mampu menggambarkan performa model secara menyeluruh pada seluruh kelas sampah digital. Selain evaluasi klasifikasi, penelitian ini juga mengembangkan model estimasi dampak lingkungan dengan menghitung potensi pengurangan konsumsi energi penyimpanan dan emisi karbon berdasarkan volume data yang

berhasil diidentifikasi sebagai sampah digital. Model estimasi tersebut mengintegrasikan faktor intensitas energi penyimpanan dan intensitas karbon jaringan listrik sehingga manfaat lingkungan dari implementasi ADWC dapat dianalisis pada berbagai skenario karakteristik sistem energi. Untuk menguji ketahanan hasil proyeksi, dilakukan analisis sensitivitas terhadap parameter-parameter utama yang berpotensi memengaruhi estimasi manfaat karbon.

2.4 Rancangan Implementasi dan Rancangan Protokol Yang di Usulkan

Alat dan Pustaka: Implementasi akan berbasis pada ekosistem Python 3.11, menggunakan pustaka *scikit-learn*, *XGBoost*, *Hugging Face Transformers*, dan *TensorFlow*. Pemindaian penyimpanan akan menggunakan agen *watchdog* Python, sementara pelacakan karbon memanfaatkan pustaka *CodeCarbon*. Lima Tahap Protokol yang di usulkan:

1. Uji Coba Awal: Lakukan uji kelayakan alur kerja dan mengkalibrasi ambang batas pada sampel awal ($n = 5.000$ berkas).
2. Eksperimen Utama: Lakukan validasi silang penuh pada semua model dengan target data skala besar ($n = 35.000$ berkas).
3. Uji Beban: lakukan uji stabilitas agen pembelajaran online dengan memberinya direktori yang terus bertambah (10 GB/minggu).
4. Studi Ablasi: Lakukan pengukuran kontribusi marjinal dari setiap kelompok fitur (metadata, konten, perilaku) dengan menghapusnya satu per satu dari model.
5. Eksperimen Jejak Karbon: Lakukan penerapan model terlatih pada korpus data gelap riil untuk mengukur penghematan emisi aktual (ΔCO_2) dan membandingkannya dengan patokan industri.

3. HASIL DAN PEMBAHASAN

Seluruh tabel kinerja yang disajikan pada bagian ini merupakan hasil interpolasi berbasis *benchmark* penelitian terdahulu dan bukan berasal dari pengukuran langsung terhadap sistem ADWC. Oleh karena itu, bobot yang diterapkan, sumber studi acuan dan tingkat ketidakpastian dari setiap proyeksi perlu dilaporkan secara rinci pada penelitian empiris selanjutnya. Dengan demikian, nilai proyeksi yang besar dalam penelitian ini tidak dapat

ditafsirkan sebagai performa aktual sistem ADWC maupun sebagai bukti bahwa model tersebut secara empiris melampaui kinerja model BERT atau CNN-BERT. Nilai tersebut semata-mata dimaksudkan sebagai estimasi konseptual untuk menggambarkan potensi kinerja sistem berdasarkan sintesis berbagai hasil penelitian terdahulu

3.1 Evaluasi Kelayakan Kerangka Kerja ADWC Berbasis Benchmark

Karena penelitian ini menggunakan pendekatan Design Science Research (DSR), evaluasi difokuskan pada penilaian kelayakan (feasibility assessment) terhadap artefak yang diusulkan, yaitu kerangka kerja Automated Digital Waste Classification (ADWC). Berbeda dengan penelitian eksperimental yang memvalidasi model menggunakan korpus dunia nyata, penelitian ini mengevaluasi kelayakan konseptual melalui pendekatan benchmark-anchored, yaitu dengan mensintesis capaian empiris berbagai penelitian terdahulu yang memiliki karakteristik tugas serupa, kemudian mengadaptasikannya ke dalam arsitektur ADWC.

Seluruh nilai numerik pada bagian ini merupakan estimasi konseptual yang diperoleh melalui proses interpolasi terhadap berbagai tolok ukur klasifikasi berkas, deduplikasi data,

dan estimasi jejak karbon digital yang telah dipublikasikan (AbdulNabi and Yaseen, 2021; Yuan, Moh and Moh, 2022; Bhanage and Pawar, 2023; Zhang, 2023; Lisker et al., 2025), Estimasi manfaat lingkungan selanjutnya dikalibrasi menggunakan parameter konsumsi energi penyimpanan dan emisi karbon yang dilaporkan dalam penelitian terdahulu (Al Kez et al., 2022; Lannelongue and Inouye, 2023).

Dengan demikian, hasil yang disajikan pada bagian ini tidak dimaksudkan sebagai validasi empiris terhadap implementasi ADWC melainkan sebagai bukti konseptual mengenai potensi teknis serta manfaat lingkungan yang secara teoritis dapat dicapai apabila kerangka kerja tersebut diterapkan pada lingkungan penyimpanan yang sebanding. Validasi menggunakan korpus dunia nyata menjadi agenda penelitian empiris pada tahap berikutnya.

3.2 Evaluasi Kelayakan Arsitektur ADWC Berdasarkan Benchmark

Kelayakan teknis arsitektur ADWC dievaluasi melalui sintesis berbagai benchmark yang merepresentasikan komponen-komponen utama penyusun kerangka kerja, yaitu model klasifikasi berbasis metadata, model pemrosesan bahasa alami (Natural Language Processing), serta pendekatan ensemble learning. Ringkasan hasil proyeksi disajikan pada Tabel 3:

Tabel 2. Perkiraan Rentang Kinerja Klasifikasi Arsitektur ADWC

No	Model	Akurasi (%)	Macro-Precision (%)	Macro-Recall (%)	Macro-F1 (%)	ROC-AUC	Inferensi (s / 10 K file)	Proyeksi Rentang (±95%)
1	Naïve Bayes (<i>baseline</i>)	78,42	76,81	77,94	77,37	0,871	11,2	74–82%
2	k-NN (k= 5)	81,07	80,14	80,62	80,38	0,903	412,8	77–85%
3	Decision Tree (C4.5)	84,69	83,95	84,28	84,11	0,911	6,3	81–88%
4	Random Forest (500 pohon)	92,18	91,43	91,79	91,61	0,967	19,7	89–95%
5	Gradient-Boosted Trees	94,05	93,62	93,81	93,71	0,978	22,4	91–97%
6	LightGBM	93,47	93,08	93,29	93,18	0,974	14,1	90–96%
7	Feed-Forward NN	91,82	91,25	91,54	91,39	0,962	8,9	88–95%
8	BiLSTM (<i>byte sequence</i>)	93,94	93,58	93,71	93,64	0,975	128,6	91–96%
9	BERT-base-uncased	96,71	96,34	96,58	96,46	0,989	268,4	94–98%

10	CNN-BERT Hybrid	97,04	96,71	96,92	96,81	0,991	302,7	94–98%
11	SegDup-RL Policy	92,60	91,97	92,32	92,14	0,969	47,2	89–95%
12	ADWC-Full (RF + BERT stack)	97,58	97,13	97,39	97,26	0,993	271,3	95–98%
13	ADWC-Ensemble (10-model vote)	97,12	96,78	97,01	96,89	0,992	526,9	94–98%

Nilai tersebut adalah hasil proyeksi dan bukan dimaksudkan sebagai klaim performa empiris melainkan sebagai indikator bahwa rancangan arsitektur ADWC memiliki kelayakan konseptual yang sejalan dengan capaian sistem-sistem sejenis. Dengan demikian evaluasi ini mendukung bahwa kombinasi pembelajaran berbasis metadata dan representasi semantik merupakan pendekatan yang layak untuk diimplementasikan pada penelitian empiris berikutnya.

Secara umum hasil proyeksi konseptual berbasis benchmark tersebut menunjukkan bahwa konfigurasi stacked ensemble yang mengintegrasikan Random Forest dan BERT diperkirakan memberikan keseimbangan terbaik antara akurasi klasifikasi dan efisiensi inferensi dibandingkan model tunggal. Model berbasis pohon diproyeksikan mampu memanfaatkan pola nonlinier pada metadata berkas secara efektif, sedangkan model Transformer memberikan representasi semantik yang lebih baik terhadap isi dokumen. Integrasi kedua

3.3 Evaluasi Kontribusi Komponen

Arsitektur ADWC

Untuk mengevaluasi kontribusi masing-masing komponen terhadap kinerja keseluruhan,

pendekatan tersebut menghasilkan peningkatan performa dibandingkan apabila masing-masing model digunakan secara terpisah.

Sebagaimana terlihat pada Tabel 2, model-model pembelajaran mesin konvensional diproyeksikan menghasilkan kinerja yang kompetitif untuk klasifikasi berbasis metadata, dengan rentang akurasi sekitar 78–94%. Di sisi lain, model berbasis Transformer diperkirakan mencapai performa yang lebih tinggi pada klasifikasi dokumen karena mampu menangkap hubungan semantik yang tidak tersedia pada atribut metadata.

Berdasarkan sintesis benchmark dari berbagai penelitian terdahulu konfigurasi ADWC-Full diproyeksikan mencapai macro-F1 yang masih berada dalam rentang realistis apabila mempertimbangkan kompleksitas klasifikasi tujuh kelas dengan distribusi data yang tidak seimbang (AbdulNabi and Yaseen, 2021; Qasim et al., 2022; Maham et al., 2024; Lisker et al., 2025).

dilakukan analisis ablasi konseptual dengan membandingkan beberapa konfigurasi fitur sebagaimana ditunjukkan pada Tabel 5;

Tabel 3. Proyeksi Kontribusi Setiap Komponen Arsitektur terhadap Utilitas Kerangka Kerja ADWC (Hasil Proyeksi Berbasis Benchmark).

Konfigurasi	Macro-F1 (%)	Δ terhadap Full	Limbah Teridentifikasi (TB)	Potensi Reduksi CO ₂ e (kg/thn)	Kontribusi terhadap Artefak
Metadata (M)	82,17	–15,09	7,4	3,32	Mengidentifikasi karakteristik fisik berkas (ukuran, tipe, umur, lokasi).
Konten (C)	90,84	–6,42	9,1	4,08	Mengekstraksi informasi semantik dokumen melalui NLP sehingga meningkatkan identifikasi data bernilai rendah.

Perilaku (B)	88,56	-8,70	8,6	3,85	Menangkap pola penggunaan aktual (frekuensi akses, waktu akses terakhir, aktivitas pengguna).
Metadata + Konten (M+C)	94,61	-2,65	10,3	4,61	Menggabungkan karakteristik struktural dan semantik sehingga meningkatkan akurasi klasifikasi.
Metadata + Perilaku (M+B)	91,74	-5,52	9,7	4,35	Menghubungkan atribut berkas dengan pola penggunaan untuk mengidentifikasi data usang dan terlantar.
Konten + Perilaku (C+B)	95,08	-2,18	10,5	4,70	Mengombinasikan makna dokumen dan perilaku pengguna sehingga meningkatkan deteksi data bernilai rendah.
ADWC-Full (M+C+B)	97,26	---	11,6	5,20	Integrasi multimodal yang menghasilkan rekomendasi tata kelola data paling komprehensif dan manfaat lingkungan terbesar.

Karena artefak ADWC dikembangkan melalui pendekatan Design Science Research (DSR) melalui proyeksi berbasis benchmark dan bukan melalui pengukuran empiris atas data primer, adapun ukuran korpus pada Tabel 5 juga merupakan parameter asumsi representatif dan bukan hasil pengukuran langsung. Nilai tersebut ditetapkan sebesar ± 30 TB untuk merepresentasikan skala volume data tidak terstruktur pada organisasi menengah dan besar serta menjadi basis proyeksi seluruh metrik pada tabel, termasuk limbah teridentifikasi (TB) dan potensi reduksi emisi CO₂e (kg/tahun), yang diturunkan dari intensitas penyimpanan (kWh/TB), Power Usage Effectiveness (PUE) dan intensitas karbon jaringan listrik (gCO₂e/kWh) pada nilai tengah (baseline) hasil analisis sensitivitas. Sebagai ilustrasi pada konfigurasi ADWC-Full, limbah teridentifikasi sebesar 11,6 TB setara dengan 38,7% dari volume korpus dari asumsi tersebut.

Proporsi ini konsisten dengan temuan Veritas Global Databerg Report yang melaporkan rata-rata 33% data korporat tergolong redundant, obsolete atau trivial/ROT (Noah Banning, 2016), serta sejalan dengan kajian sistematis terkini yang mengonfirmasi

bahwa sebagian besar data enterprise tidak termanfaatkan secara aktif dan berpotensi menjadi limbah digital (Olaitan et al., 2025). Dengan demikian skala proyeksi ADWC-Full berada pada rentang yang logis. Sekalipun angka tersebut masih bersifat proyeksi benchmark.

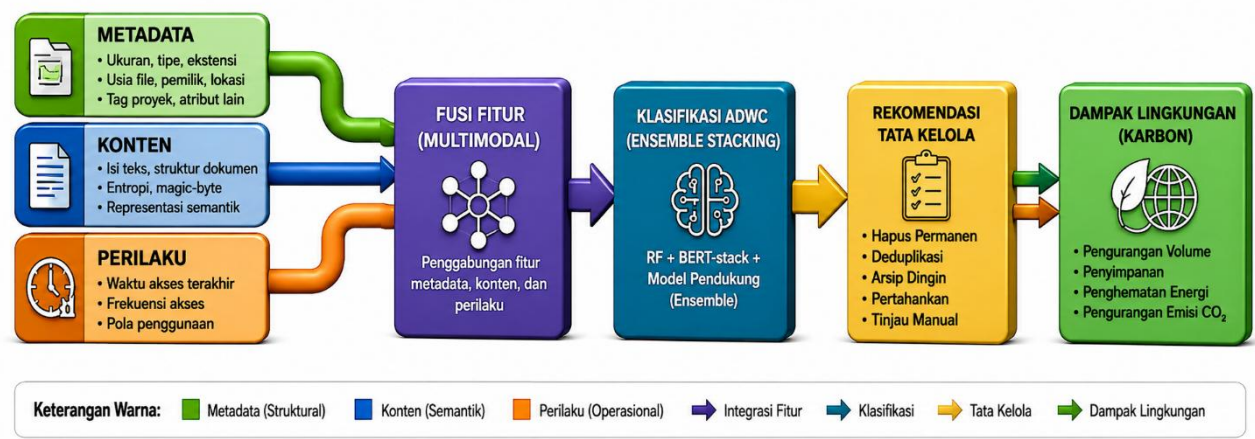
Hasil proyeksi menunjukkan bahwa tidak ada satu kelompok fitur yang mampu merepresentasikan karakteristik sampah digital secara memadai apabila digunakan secara terpisah. Penggunaan metadata saja menghasilkan performa terendah karena hanya menggambarkan karakteristik struktural berkas tanpa mempertimbangkan isi maupun perilaku pengguna. Sebaliknya, fitur konten memberikan peningkatan kemampuan klasifikasi melalui representasi semantik dokumen, sedangkan fitur perilaku melengkapi informasi tersebut dengan pola penggunaan aktual, seperti frekuensi akses dan waktu akses terakhir.

Kombinasi ketiga kelompok fitur menghasilkan peningkatan kinerja yang paling konsisten. Integrasi metadata, konten dan perilaku ini tidak hanya meningkatkan kemampuan identifikasi berbagai kategori sampah digital tetapi juga berpotensi memperbesar estimasi volume data yang dapat

ditindaklanjuti beserta potensi pengurangan emisi karbon. Hal ini menunjukkan bahwa ketiga sumber informasi tersebut bersifat saling melengkapi (complementary) sehingga menghasilkan representasi yang lebih komprehensif dibandingkan pendekatan berbasis satu modalitas.

Dari perspektif DSR hasil proyeksi ini memperlihatkan bahwa nilai utama artefak ADWC tidak terletak pada penggunaan algoritma tertentu saja melainkan pada desain

arsitektur yang mengintegrasikan berbagai jenis informasi ke dalam satu mekanisme klasifikasi dan rekomendasi tata kelola data. Dengan demikian utilitas kerangka kerja tidak hanya diukur melalui peningkatan akurasi klasifikasi tetapi juga melalui kemampuannya menghasilkan keputusan pengelolaan data yang lebih tepat dan berpotensi memberikan manfaat lingkungan yang lebih besar. Selengkapnya dapat di lihat pada gambar 2. Berikut;



Gambar 2. Integrasi Tiga Modalitas ADWC dan Dampaknya Terhadap Tata Kelola Data Serta Pengurangan Jejak Karbon

Sinyal perilaku (waktu akses terakhir, jumlah pembacaan) telah lama diakui sebagai proksi dominan untuk deteksi data yang jarang diakses oleh Hormann & Campbell (2020) yang berpendapat bahwa mengarsipkan data yang jarang diakses dan menghapus data yang memiliki nilai masa depan yang kecil secara signifikan mengurangi konsumsi energi penyimpanan bahkan sebelum adanya inovasi perangkat keras.

3.4 Proyeksi Pengurangan Jejak Karbon

Salah satu tujuan utama pengembangan ADWC adalah menghubungkan keputusan tata kelola data pada tingkat berkas dengan indikator keberlanjutan yang dapat diukur. Oleh karena itu, utilitas lingkungan kerangka kerja dievaluasi melalui proyeksi potensi pengurangan konsumsi energi penyimpanan dan emisi karbon yang dihasilkan dari tindakan penghapusan, deduplikasi, serta migrasi data ke media penyimpanan dingin. Ringkasan hasil proyeksi disajikan pada Tabel 6;

Tabel 4. Proyeksi Utilitas Lingkungan Kerangka Kerja ADWC pada Berbagai Skala Implementasi				
Skenario Implementasi	Data yang Dapat Ditindaklanjuti	Energi Tersimpan	Reduksi Emisi CO ₂ e	Reduksi vs. Status Quo
Institusi tunggal (skala HLRS, 19,6 PB)	7,59 PB	7,59 GWh/thn	2,43 kt /thn	38,7%
Organisasi menengah (korpus 1 PB)	0,387 PB	0,387 GWh/thn	0,124 kt/thn	38,7%
Konsorsium (100 PB)	38,7 PB	38,7 GWh/thn	12,38 kt /thn	38,7%
Ekstrapolasi global (175 ZB)	67,7 ZB yang dapat ditindaklanjuti	67,7 TWh/thn	21,66 Mt/thn	38,7%

Mengacu pada pola empiris yang dilaporkan oleh Schembera dan Durán (2020), serta metodologi estimasi yang dijelaskan pada Subbagian 2.5, ADWC diproyeksikan mampu mengidentifikasi sekitar 38,7% korpus penyimpanan sebagai data yang dapat ditindaklanjuti. Persentase tersebut tidak hanya mencakup dark data tetapi juga berbagai kategori limbah digital lainnya seperti; file duplikat, cache, log serta data yang telah kehilangan nilai operasional.

Sebagaimana ditunjukkan pada Tabel 4, besarnya potensi penghematan energi dan reduksi emisi karbon meningkat secara proporsional terhadap ukuran korpus penyimpanan. Sedangkan persentase data yang dapat ditindaklanjuti relatif konstan karena dihitung menggunakan asumsi benchmark yang sama pada setiap skenario implementasi. Hal ini menunjukkan bahwa manfaat konseptual ADWC bersifat skalabel sehingga kerangka kerja ini berpotensi diterapkan pada organisasi dengan kapasitas penyimpanan yang berbeda tanpa memerlukan perubahan mendasar pada arsitektur sistem.

Karena seluruh nilai pada Tabel 6 merupakan hasil proyeksi sehingga estimasi reduksi emisi karbon dihitung melalui konversi langsung dari proyeksi energi yang dapat dihemat menggunakan faktor intensitas karbon rata-rata global sebesar 320 gCO₂e/kWh. Dengan pendekatan tersebut besarnya reduksi emisi karbon pada setiap skenario implementasi berbanding lurus dengan energi yang berhasil dihemat. Estimasi ini belum mempertimbangkan variasi bauran energi regional, nilai Power

Usage Effectiveness (PUE) maupun karakteristik operasional masing-masing pusat data sehingga hasil yang disajikan harus dipahami sebagai proyeksi konseptual berbasis benchmark dan bukan sebagai hasil pengukuran empiris implementasi ADWC.

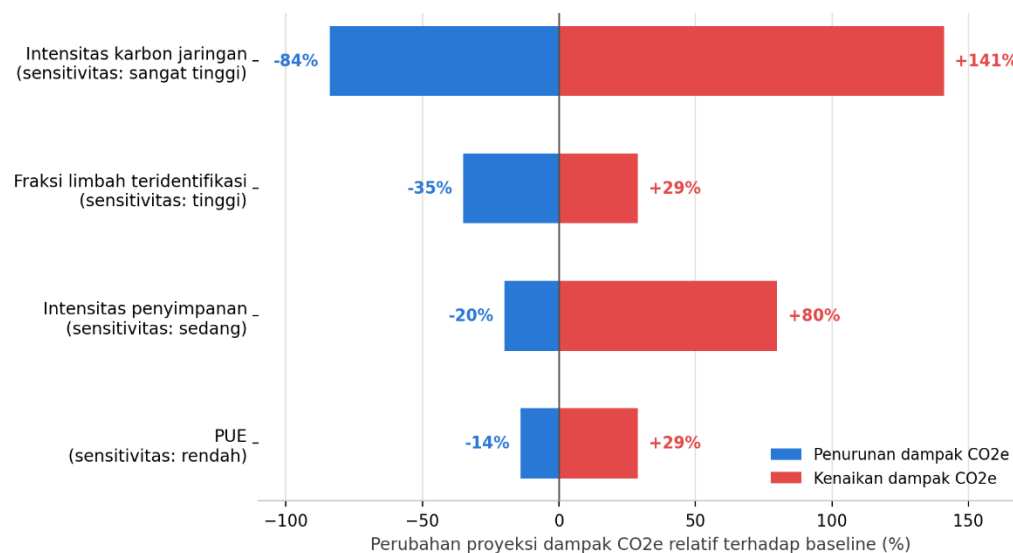
Secara keseluruhan, hasil proyeksi menunjukkan bahwa implementasi ADWC berpotensi memberikan kontribusi terhadap pengurangan konsumsi energi dan emisi karbon melalui optimalisasi tata kelola data. Walaupun besaran manfaat lingkungan yang disajikan masih memerlukan validasi melalui implementasi pada lingkungan operasional nyata, arah proyeksi ini sejalan dengan berbagai penelitian sebelumnya yang menunjukkan bahwa pengelolaan data yang lebih efisien dapat mendukung keberlanjutan operasional pusat data (Liu et.al. 2020). Oleh karena itu, hasil pada Tabel 6 lebih tepat dipandang sebagai ilustrasi potensi manfaat lingkungan dari penerapan ADWC daripada sebagai estimasi dampak yang telah terverifikasi secara empiris.

3.5 Analisis Sensitivitas

Analisis sensitivitas dilakukan untuk mengevaluasi pengaruh perubahan parameter utama terhadap estimasi pengurangan emisi karbon yang dihasilkan oleh ADWC. Parameter yang dianalisis meliputi intensitas energi penyimpanan, Power Usage Effectiveness (PUE), intensitas karbon jaringan listrik, serta proporsi limbah digital yang berhasil diidentifikasi. Ringkasan hasil analisis disajikan pada Tabel 7.

Tabel 5. Sensitivitas Estimasi Dampak Lingkungan terhadap Parameter Implementasi ADWC

Parameter	Estimasi Rendah	Estimasi Tengah	Estimasi Tinggi	Dampak ΔCO ₂	Tingkat Sensitivita
Intensitas penyimpanan (kWh/TB)	0,8	1,0	1,8	-20% hingga +80%	Sedang
PUE	1,2	1,4	1,8	-14% hingga +29%	Rendah
Intensitas karbon jaringan (gCO ₂ e/kWh)	50	320 (global rata-rata)	770 (berbahan bakar batu bara)	-84% hingga +141%	Sangat Tinggi
Fraksi limbah teridentifikasi	25%	38,7%	50%	-35% hingga +29%	Tinggi



Gambar 3. Analisis Sensitivitas Parameter Terhadap Proyeksi Dampak CO2e

Hasil proyeksi menunjukkan bahwa intensitas karbon jaringan listrik merupakan parameter yang paling sensitif terhadap estimasi pengurangan emisi CO_{2e}. Dengan demikian, besarnya pengurangan emisi yang diproyeksikan tidak dapat dilepaskan dari kondisi sistem kelistrikan yang menjadi sumber energi bagi infrastruktur tersebut. Hasil pemodelan klasifikasi selanjutnya memberikan dasar bagi proses identifikasi tersebut. Pendekatan yang menggabungkan model berbasis pohon dan BERT memungkinkan karakteristik metadata dan informasi semantik dipertimbangkan secara bersamaan. Model berbasis pohon digunakan untuk menangkap pola pada metadata seperti; ukuran file, usia penyimpanan dan frekuensi akses sedangkan BERT memberikan representasi terhadap informasi semantik dalam isi dokumen. Kemampuan kedua pendekatan tersebut telah banyak dibahas dalam penelitian terdahulu (Jawahar, Sagot and Seddah, 2019; Qiu et al., 2020; Rasmy et al., 2021). Sehingga penggabungannya dalam kerangka ADWC dapat dipandang sebagai upaya untuk memanfaatkan dua jenis informasi yang saling melengkapi dalam proses klasifikasi.

Penggunaan fine-tuning BERT pada korpus khusus domain juga menjadi bagian penting dalam rancangan pemodelan. Dalam proyeksi yang dilakukan, adaptasi model terhadap korpus domain digunakan untuk memperoleh representasi yang lebih sesuai dengan karakteristik dokumen yang

diklasifikasikan. Pendekatan ini sejalan dengan Maham et al. (2024) dan Qasim et al. (2022), yang melaporkan manfaat transfer learning berbasis domain dalam klasifikasi dokumen. Lisker et al. (2025) juga menunjukkan bahwa optimasi model yang disesuaikan dengan domain tetap relevan pada lingkungan penyimpanan dengan keterbatasan sumber daya. Literatur tersebut memberikan dasar bagi penggunaan fine-tuning dalam rancangan ADWC, meskipun efektivitasnya pada lingkungan operasional nyata masih memerlukan pengujian lebih lanjut.

Hal ini relevan dengan persoalan ketidakseimbangan kelas yang telah dibahas dalam penelitian mengenai klasifikasi limbah digital. Penggunaan Random Forest sebagai lapisan akhir juga dipertimbangkan untuk menangani karakteristik kelas yang tidak seimbang terutama pada kategori Abandoned dan Unknown. Dalam rancangan proyeksi ADWC pendekatan tersebut digunakan untuk memperoleh probabilitas klasifikasi yang lebih stabil pada kelas dengan representasi data yang lebih rendah (Kaur, Jit Singh and Chana, 2026). Namun, karena hasil yang digunakan dalam penelitian ini masih berupa proyeksi sehingga kestabilan tersebut belum dapat dipahami sebagai bukti kinerja pada kondisi operasional yang sebenarnya.

Jika dilihat secara lebih luas, pendekatan yang digunakan dalam penelitian ini memperlihatkan kemungkinan untuk

menghubungkan klasifikasi limbah digital dengan pengukuran dampak lingkungan. Roussilhe et al. (2023) menekankan pentingnya mengukur dampak lingkungan digitalisasi secara lebih komprehensif, sedangkan Hormann dan Campbell (2020) menempatkan penghapusan data yang tidak lagi memiliki nilai bisnis sebagai salah satu bentuk intervensi dalam komputasi ramah lingkungan. Dalam konteks tersebut, ADWC memproyeksikan suatu mekanisme yang menghubungkan identifikasi data bernilai rendah dengan estimasi konsekuensi lingkungannya. Dengan demikian kontribusi konseptual penelitian ini terletak pada upaya mengintegrasikan klasifikasi limbah digital dan estimasi emisi CO₂e dalam satu kerangka kerja bukan pada pembuktian bahwa besaran pengurangan emisi tersebut telah terjadi dalam lingkungan operasional.

Dari sisi penerapan hasil proyeksi ini memberikan beberapa pertimbangan bagi organisasi yang nantinya ingin menggunakan ADWC. Sensitivitas yang tinggi terhadap intensitas karbon jaringan telah menunjukkan bahwa estimasi manfaat lingkungan sebaiknya menggunakan karakteristik sumber energi yang sesuai dengan lokasi atau kondisi infrastruktur yang akan menjadi sasaran implementasi. Sementara itu, besarnya pengaruh fraksi limbah yang teridentifikasi menunjukkan pentingnya memastikan bahwa data yang dikategorikan sebagai limbah benar-benar dapat ditindaklanjuti melalui kebijakan penghapusan, deduplikasi, atau pemindahan ke penyimpanan yang lebih efisien. Dengan kata lain bahwa potensi pengurangan emisi yang diproyeksikan melalui ADWC sangat bergantung pada keberhasilan menghubungkan hasil klasifikasi dengan tindakan pengelolaan data.

Besarnya fraksi limbah yang diproyeksikan juga dapat dikaitkan dengan fenomena penyimpanan data yang tidak lagi memiliki nilai atau aktivitas yang sebanding dengan sumber daya yang digunakan untuk mempertahankannya. Schembera dan Durán (2020) menemukan bahwa sejumlah besar akun pengguna pada sistem komputasi berkinerja tinggi tidak lagi aktif tetapi masih menyimpan data dalam jumlah besar. Kondisi tersebut berkaitan dengan perilaku digital hoarding yang dibahas McKellar et al. (2024), yaitu

kecenderungan pengguna yaitu kecenderungan mempertahankan file meskipun nilai informasinya telah menurun. Dalam konteks implementasi ADWC temuan-temuan tersebut menunjukkan bahwa pendekatan otomatis kemungkinan akan lebih efektif apabila disertai kebijakan organisasi yang mendorong pengelolaan dan penghapusan data bernilai rendah sebagaimana disarankan oleh Tripathi et al. (2019).

3.6. Keterbatasan Penelitian

Beberapa keterbatasan perlu dipertimbangkan ketika menginterpretasikan hasil penelitian ini. Pertama, seluruh hasil yang disajikan masih berupa proyeksi sehingga validasi pada lingkungan operasional nyata masih diperlukan. Perbedaan distribusi kelas, karakteristik organisasi maupun jenis data yang disimpan berpotensi menghasilkan performa yang berbeda dari estimasi saat ini. Kedua, taksonomi tujuh kelas yang digunakan disusun berdasarkan literatur, namun implementasi pada organisasi tertentu mungkin memerlukan kategori tambahan misalnya; limbah yang berkaitan dengan regulasi atau karantina data pribadi. Ketiga, meskipun ADWC-Full menunjukkan performa terbaik, selisih kinerja antara ADWC-Full dan CNN-BERT Hybrid diproyeksikan relatif kecil sehingga perbedaannya belum tentu bermakna secara praktis, mengingat tidak ada pengujian statistik formal yang dilakukan terhadap nilai-nilai proyeksi ini.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada rekan-rekan peneliti di institusi STMIK Bina Mulia Palu yang telah membantu.

4. PENUTUP

4.1. Kesimpulan

Penelitian ini mengusulkan Automated Digital Waste Classification (ADWC) sebagai kerangka kerja konseptual berbasis pembelajaran mesin untuk mendukung pengelolaan sampah digital yang berorientasi pada keberlanjutan lingkungan. Kerangka kerja yang diusulkan mengintegrasikan taksonomi tujuh kelas sampah digital, arsitektur klasifikasi berlapis, serta estimasi jejak karbon ke dalam satu pendekatan yang utuh sehingga memperluas

cakupan penelitian terdahulu yang umumnya masih membahas klasifikasi berkas, deduplikasi data, dan green computing secara terpisah. Dengan demikian, penelitian ini memberikan kontribusi konseptual berupa landasan yang sistematis untuk menghubungkan tata kelola data pada tingkat berkas dengan upaya pengurangan dampak lingkungan teknologi informasi.

Meskipun proyeksi berbasis benchmark menunjukkan potensi penerapan ADWC dalam meningkatkan efisiensi pengelolaan data sekaligus mendukung pengurangan jejak karbon, seluruh temuan dalam penelitian ini masih bersifat konseptual dan memerlukan validasi empiris.

4.2. Saran

Penelitian selanjutnya perlu melakukan validasi empiris ADWC menggunakan dataset organisasi nyata dan melibatkan audit manual oleh pakar data governance untuk menilai kesesuaian hasil klasifikasi serta melakukan pengujian lintas sektor pada karakteristik infrastruktur energi yang berbeda untuk mengevaluasi generalisasi kerangka kerja. Selain itu, perlu dilakukan evaluasi biaya komputasi model dan risiko privasi terutama ketika klasifikasi melibatkan isi dokumen yang berpotensi mengandung informasi sensitif. Karena estimasi dampak lingkungan dalam penelitian ini masih bersifat proyeksi. Penelitian selanjutnya juga perlu mengukur konsumsi energi dan emisi aktual menggunakan CodeCarbon atau pendekatan Life Cycle Assessment (LCA), sehingga hasil proyeksi dapat dibandingkan dengan dampak lingkungan yang terjadi pada kondisi operasional nyata.

5. DAFTAR PUSTAKA

- AbdulNabi, I. and Yaseen, Q., 2021. Spam Email Detection Using Deep Learning Techniques. *Procedia Computer Science*, 184, pp.853–858. <https://doi.org/10.1016/j.procs.2021.03.107>.
- Ali Raza, S., Abbas, S., M. Ghazal, T., Adnan Khan, M., Ahmad, M. and Al Hamadi, H., 2022. Content Based Automated File Organization Using Machine Learning approaches. *Computers, Materials & Continua*, 73(1), pp.1927–1942. <https://doi.org/10.32604/cmc.2022.029400>.
- Bhanage, D.A. and Pawar, A.V., 2023. Robust Analysis of IT Infrastructure's Log Data with BERT Language Model. *International Journal of Advanced Computer Science and Applications*, 14(6). <https://doi.org/10.14569/IJACSA.2023.0140675>.
- Chundru, S. and Mudunuri, L.N.R., 2025. Developing Sustainable Data Retention Policies. pp.93–114. <https://doi.org/10.4018/979-8-3693-9750-3.ch005>.
- Coughlin, T., 2018. A Solid-State Future [The Art of Storage]. *IEEE Consumer Electronics Magazine*, 7(1), pp.113–116. <https://doi.org/10.1109/MCE.2017.2755339>.
- Devindi, W.N. and Piyumal, K.M., 2025. An AI-Powered Metadata-Driven File Management Agent for Academic Computing Platforms. In: *2025 International Conference on Advances in Technology and Computing (ICATC)*. IEEE. pp.1–6. <https://doi.org/10.1109/ICATC68823.2025.11407546>.
- Dr. A. Shaji George, Dr. V. Sujatha, A. S. Hovan George and Dr. T. Baskar, 2023. Bringing Light to Dark Data: A Framework for Unlocking Hidden Business Value. *Partners Universal International Innovation Journal*, [online] 1(4), pp.35–60. <https://doi.org/10.5281/zenodo.8262384>.
- Ewim, D.R.E., Ninduwezuor-Ehiobu, N., Orikpete, O.F., Egbokhaebho, B.A., Fawole, A.A. and Onunka, C., 2023. Impact of Data Centers on Climate Change: A Review of Energy Efficient Strategies. *The Journal of Engineering and Exact Sciences*, 9(6), pp.16397–01e. <https://doi.org/10.18540/jcecvl9iss6pp16397-01e>.
- Guo, B., Yu, J., Yang, D., Leng, H. and Liao, B., 2023. Energy-Efficient Database Systems: A Systematic Survey. *ACM Computing Surveys*, 55(6), pp.1–53. <https://doi.org/10.1145/3538225>.

- Hartung, T., 2018. Making Big Sense From Big Data. *Frontiers in Big Data*, 1. <https://doi.org/10.3389/fdata.2018.00005>.
- Hormann, P. and Campbell, L., 2020. Data Storage Energy Efficiency in the Zettabyte Era. *Journal of Telecommunications and the Digital Economy*, 2(3), p.22. <https://doi.org/10.18080/jtde.v2n3.282>.
- Jackson, T.W. and Hodgkinson, I.R., 2023. Keeping a lower profile: how firms can reduce their digital carbon footprints. *Journal of Business Strategy*, 44(6), pp.363–370. <https://doi.org/10.1108/JBS-03-2022-0048>.
- Jawahar, G., Sagot, B. and Seddah, D., 2019. What Does BERT Learn about the Structure of Language? In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics. pp.3651–3657. <https://doi.org/10.18653/v1/P19-1356>.
- Johns, B., 2023. Improving Information Technology Sustainability With Modern Tape Storage. *SMPTE Motion Imaging Journal*, 132(7), pp.27–31. <https://doi.org/10.5594/JMI.2023.3285282>.
- Kaur, R., Jit Singh, H. and Chana, I., 2026. Machine Learning-Based Data Deduplication: Techniques, Challenges, and Future Directions. *Concurrency and Computation: Practice and Experience*, 38(3). <https://doi.org/10.1002/cpe.70574>.
- Al Kez, D., Foley, A.M., Laverty, D., Del Rio, D.F. and Sovacool, B., 2022. Exploring the sustainability challenges facing digitalization and internet data centers. *Journal of Cleaner Production*, 371, p.133633. <https://doi.org/10.1016/j.jclepro.2022.133633>.
- Lannelongue, L., Grealey, J., Bateman, A. and Inouye, M., 2021. Ten simple rules to make your computing more environmentally sustainable. *PLOS Computational Biology*, 17(9), p.e1009324. <https://doi.org/10.1371/journal.pcbi.1009324>.
- Lannelongue, L. and Inouye, M., 2023. Environmental Impacts of Machine Learning Applications in Protein Science. *Cold Spring Harbor Perspectives in Biology*, 15(12), p.a041473. <https://doi.org/10.1101/cshperspect.a041473>.
- Lisker, S., Butman, A., Hajaj, C., Dubin, R. and Dvir, A., 2025. Optimized File Type Detection and One-Shot Retrieval. In: *ICC 2025 - IEEE International Conference on Communications*. IEEE. pp.1121–1126. <https://doi.org/10.1109/ICC52391.2025.11161523>.
- Liu, Y., Wei, X., Xiao, J., Liu, Z., Xu, Y. and Tian, Y., 2020. Energy consumption and emission mitigation prediction based on data center traffic and PUE for global data centers. *Global Energy Interconnection*, 3(3), pp.272–282. <https://doi.org/10.1016/j.gloi.2020.07.008>.
- Lucivero, F., 2020. Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives. *Science and Engineering Ethics*, 26(2), pp.1009–1030. <https://doi.org/10.1007/s11948-019-00171-7>.
- Maham, S., Tariq, A., Khan, M.U.G., Alamri, F.S., Rehman, A. and Saba, T., 2024. ANN: adversarial news net for robust fake news classification. *Scientific Reports*, 14(1), p.7897. <https://doi.org/10.1038/s41598-024-56567-4>.
- Marumolwa, L. and Marnewick, C., 2025. Unveiling Dark Data in Organisations. *International Journal of Service Science, Management, Engineering, and Technology*, 16(1), pp.1–32. <https://doi.org/10.4018/IJSSMET.386167>.
- Matendo Didas, 2023. The barriers and prospects related to big data analytics

- implementation in public institutions: a systematic review analysis. *International Journal of Advanced Computer Research*, 13(64). <https://doi.org/10.19101/IJACR.2021.1152071>.
- McKellar, K., Sillence, E., Neave, N. and Briggs, P., 2024. Digital accumulation behaviours and information management in the workplace: exploring the tensions between digital data hoarding, organisational culture and policy. *Behaviour & Information Technology*, 43(6), pp.1206–1218. <https://doi.org/10.1080/0144929X.2023.2205970>.
- Mersico, L., Abroshan, H., Sanchez-Velazquez, E., Saheer, L.B., Simandjuntak, S., Dhar-Bhattacharjee, S., Al-Haddad, R., Saeed, N. and Saxena, A., 2024. Challenges and Solutions for Sustainable ICT: The Role of File Storage. *Sustainability*, 16(18), p.8043. <https://doi.org/10.3390/su16188043>.
- Naeem, R., Bajwa, A., Sattar, H. and Naeem, B., 2023. Social Media Platforms and their Digital Carbon Footprints: Analyzing Awareness level of Social Media Users of Punjab, Pakistan. *Research Square*, 1. <https://doi.org/10.21203/rs.3.rs-3299158/v1>.
- Noah Banning, 2016. *Veritas Global Databerg Report Finds 85% of Stored Data Is Either Dark or Redundant, Obsolete, or Trivial (ROT)*. [online] Veritas Technologies. Available at: <<https://www.veritas.com/news-releases/2016-03-15-veritas-global-databerg-report-finds-85-percent-of-stored-data>> [Accessed 6 August 2026].
- Olaitan, O.F., Adebajo, T.A., Obozokhai, L.I., Iwerumoh, A.N., Balogun, I.O. and Ojo, D.A., 2025. Dark Data in Business Intelligence: A Systematic Review of Challenges, Opportunities, and Value Creation Potential. *Journal of Economics, Business, and Commerce*, 2(2), pp.135–142. <https://doi.org/10.69739/jebc.v2i2.1000>.
- Qasim, R., Bangyal, W.H., Alqarni, M.A. and Ali Almazroi, A., 2022. A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. *Journal of Healthcare Engineering*, 2022, pp.1–17. <https://doi.org/10.1155/2022/3498123>.
- Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N. and Huang, X., 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10), pp.1872–1897. <https://doi.org/10.1007/s11431-020-1647-3>.
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C. and Zhi, D., 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1), p.86. <https://doi.org/10.1038/s41746-021-00455-y>.
- Roussilhe, G., Ligozat, A.-L. and Quinton, S., 2023. A long road ahead: a review of the state of knowledge of the environmental effects of digitization. *Current Opinion in Environmental Sustainability*, 62, p.101296. <https://doi.org/10.1016/j.cosust.2023.101296>.
- Savarimuthu, B.T.R., Corbett, J., Yasir, M. and Lakshmi, V., 2023. Improving Information Systems Sustainability by Applying Machine Learning to Detect and Reduce Data Waste. *Communications of the Association for Information Systems*, 53, pp.189–213. <https://doi.org/10.17705/1CAIS.05308>.
- Schembera, B. and Durán, J.M., 2020. Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer. *Philosophy & Technology*, 33(1), pp.93–115. <https://doi.org/10.1007/s13347-019-00346-x>.
- Shuja, J., Ahmad, R.W., Gani, A., Abdalla Ahmed, A.I., Siddiq, A., Nisar, K., Khan, S.U. and Zomaya, A.Y., 2017. Greening emerging IT technologies: techniques and practices. *Journal of Internet Services and Applications*, 8(1), p.9. <https://doi.org/10.1186/s13174-017-0060-5>.

- Thangam, D., Muniraju, H., Ramesh, R., Narasimhaiah, R., Muddasir Ahamed Khan, N., Booshan, S., Booshan, B., Manickam, T. and Sankar Ganesh, R., 2024. Impact of Data Centers on Power Consumption, Climate Change, and Sustainability. pp.60–83. <https://doi.org/10.4018/979-8-3693-1552-1.ch004>. pp.1189–1195. <https://doi.org/10.54254/2755-2721/6/20230571>.
- Tripathi, S.K., Kaur, A., Pandey, R. and Chhabra, M., 2019. Time To Modify Operating Software (OS), Databases (DB) And Tcp/Ip Protocols For Data Trash Elimination, Based On User Defined Shelf Life Of Data. *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY*, 10(1). <https://doi.org/10.34218/IJCET.10.1.2019.016>.
- Vamsi Krishna Vemulapalli, 2025. AI-Driven Autonomous Archiving: The Future of Sustainable Database Management. *Journal of Computer Science and Technology Studies*, 7(3), pp.885–892. <https://doi.org/10.32996/jcsts.2025.7.3.98>.
- Wu, J., Guo, S., Li, J. and Zeng, D., 2016. Big Data Meet Green Challenges: Greening Big Data. *IEEE Systems Journal*, 10(3), pp.873–887. <https://doi.org/10.1109/JSYST.2016.2550538>.
- Yuan, X., Moh, M. and Moh, T.-S., 2022. Whole-File Chunk-Based Deduplication Using Reinforcement Learning for Cloud Storage. In: *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. pp.269–276. <https://doi.org/10.1109/ASONAM55673.2022.10068661>.
- Zahir Sayyed, 2026. Scalable Data Archival & Purging Mechanism for High Volume Applications. *International Journal of Sustainability and Innovation in Engineering*, 2(1). <https://doi.org/10.56830/IJSIE202412>.
- Zhang, S., 2023. Spam/Ham email classification using BERT. *Applied and Computational Engineering*, 6(1),