INDEXIA: Informatic and Computational Intelligent Journal
Vol.07, No.02, Nov 2025, hlm. 88-96
P-ISSN: 2657-0424
E-ISSN: 2657-0432

IMPLEMENTASI *OPTICAL CHARACTER RECOGNITION* DAN ALGORITMA *AHO-CORASICK* UNTUK DETEKSI BAHAN BERBAHAYA PADA PRODUK *SKINCARE*

Alfin Syahrina[1,*)], Bambang Krismono[2], Muhammad Zulfikri[3]
[1, 2,3)] Program Studi Ilmu Komputer, Universitas Bumigora
Jl. Ismail Marzuki No.22, Cilinaya, Kec. Cakranegara, Kota Mataram, Nusa Tenggara Barat 83127
e-mail: alfinsyahrinafina@gmail.com[1)]

(Naskah masuk : 8 Juli 2025 Diterima untuk diterbitkan : 23 Juli 2025)

ABSTRAK

Perkembangan industri skincare di Indonesia memunculkan tantangan terhadap keberadaan bahan kimia berbahaya dalam produk. Pengguna awam kesulitan mengidentifikasi bahan berbahaya dalam daftar komposisi pada kemasan. Penelitian ini bertujuan mengembangkan sistem deteksi otomatis berbasis Tesseract OCR untuk ekstraksi teks dari gambar kemasan, serta algoritma Aho-Corasick untuk mendeteksi bahan berbahaya. Dataset terdiri atas 5.328 bahan skincare dari Kaggle dan 1.004 bahan berbahaya dari CDPH yang diklasifikasikan ke dalam empat kategori risiko. Uji coba pada 30 gambar produk menunjukkan akurasi ekstraksi Tesseract OCR mencapai 93,43% (Word Accuracy) dan 97,06% (Character Accuracy). Deteksi bahan berbahaya menggunakan Aho-Corasick mencapai akurasi 100%. Hasil ini menunjukkan sistem efektif membantu konsumen dalam mengenali bahan berbahaya pada produk skincare.

***Kata Kunci**: OCR, Tesseract, Aho-Corasick, skincare, bahan berbahaya, ekstraksi teks, deteksi otomatis.*

ABSTRACT

The growth of the skincare industry in Indonesia presents challenges regarding the presence of harmful chemical substances in products. General consumers often struggle to identify hazardous ingredients listed on product packaging. This study aims to develop an automated detection system using Tesseract OCR for text extraction from packaging images and the Aho-Corasick algorithm for detecting harmful ingredients. The dataset consists of 5,328 skincare ingredients from Kaggle and 1,004 hazardous substances from CDPH, classified into four risk categories. Experiments on 30 product images showed that Tesseract OCR achieved a text extraction accuracy of 93.43% (Word Accuracy) and 97.06% (Character Accuracy). The detection of hazardous substances using the Aho-Corasick algorithm reached 100% accuracy. These results indicate that the system is effective in assisting consumers in identifying harmful ingredients in skincare products.

***Keywords**: OCR, Tesseract, Aho-Corasick, skincare, harmful ingredients, text extraction, automatic detection.*

I. PENDAHULUAN

Industri kosmetik dan perawatan kulit di Indonesia mengalami pertumbuhan pesat dalam beberapa tahun terakhir. Namun, masih banyak produk skincare yang mengandung bahan kimia berbahaya seperti *mercury, hydroquinone, formaldehyde, butylated hydroxyanisole (BHA),* dan *parabens,* yang berisiko menyebabkan gangguan ginjal, sistem saraf, kanker, dan kelainan janin [1][2]. Paparan rutin terhadap zat ini bersifat kumulatif dan dapat menimbulkan dampak kesehatan serius dalam jangka panjang [3]. Produk dengan kandungan tersebut masih banyak beredar secara bebas, baik di toko maupun secara daring [4]. Identifikasi bahan berbahaya biasanya dilakukan secara manual dengan membaca label komposisi, namun metode ini membutuhkan waktu, ketelitian, dan pemahaman istilah teknis atau asing, yang sulit dilakukan oleh konsumen awam. Oleh karena itu, diperlukan sistem otomatis yang mampu mengekstraksi teks dari gambar kemasan dan mendeteksi bahan berbahaya dengan cepat dan akurat. Teknologi *Optical Character Recognition (OCR)*, khususnya *Tesseract*, banyak digunakan untuk ekstraksi teks dari citra. Penelitian sebelumnya menunjukkan bahwa *Tesseract* memiliki akurasi dan kecepatan yang baik dibandingkan *PyOCR* dan *tesserOCR* [5][6][7]. Untuk tahap

deteksi bahan, digunakan algoritma *Aho-Corasick* yang efisien dalam pencarian banyak pola sekaligus dan terbukti unggul dibandingkan algoritma seperti KMP [8][9].

Berbeda dengan penelitian yang menggabungkan OCR dan *Named Entity Recognition (NER)* untuk identifikasi bahan halal atau haram [10], penelitian ini menggunakan kombinasi *Tesseract OCR* dan *Aho-Corasick* secara langsung untuk mendeteksi bahan berbahaya dalam *skincare. Dataset* yang digunakan mencakup 5.328 bahan *skincare* dari *Kaggle* dan 1.004 bahan berbahaya dari *California Department of Public Health (CDPH)* yang dikategorikan dalam empat jenis risiko: *cancer*, *developmental*, *female reproductive*, dan *male reproductive*. Penelitian ini akan menggunakan *dataset* bahan *skincare* sebanyak 5.328 data yang diperoleh dari *Kaggle* dan 1.004 data bahan berbahaya dari *Dataset Search*, dengan fokus pada produk *skincare* wajah. Bahan-bahan tersebut dikategorikan dalam empat kelompok risiko, yaitu: *cancer, development, female reproductive*, dan *male reproductive*. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem yang mampu mendeteksi bahan berbahaya dalam komposisi produk *skincare* secara otomatis dengan judul **"**Implementasi *Optical character recognition* dan Algoritma *Aho-Corasick* untuk Deteksi Bahan Berbahaya pada Produk *Skincare*.

II. METODE PENELITIAN

Penelitian ini menggunakan pendekatan metode *prototype* yang memungkinkan pengembangan sistem dilakukan secara bertahap dan iteratif. Metode ini dinilai cocok karena memberikan fleksibilitas dalam merancang, menguji, dan menyempurnakan sistem berbasis *Optical Character Recognition* (OCR) dan algoritma *Aho-Corasick* sesuai kebutuhan fungsional.

2.1 Pengumpulan Kebutuhan

Penelitian ini menggunakan dua jenis dataset utama. Dataset pertama terdiri dari 5.328 data bahan *skincare* yang bersumber dari *Kaggle*, mencakup nama-nama bahan dalam bahasa Indonesia, Inggris, dan Latin, serta disimpan dalam format *JSON* untuk memudahkan proses pencocokan. Dataset kedua berisi 1.004 data bahan berbahaya yang diperoleh dari situs resmi *California Department of Public Health* (CDPH). Data ini diklasifikasikan ke dalam empat kategori risiko, yaitu *cancer* (625 data), *developmental* (237 data), *female reproductive* (60 data), dan *male reproductive* (82 data). Contoh gambar dataset sebagimana tersaji pada Gambar 1.

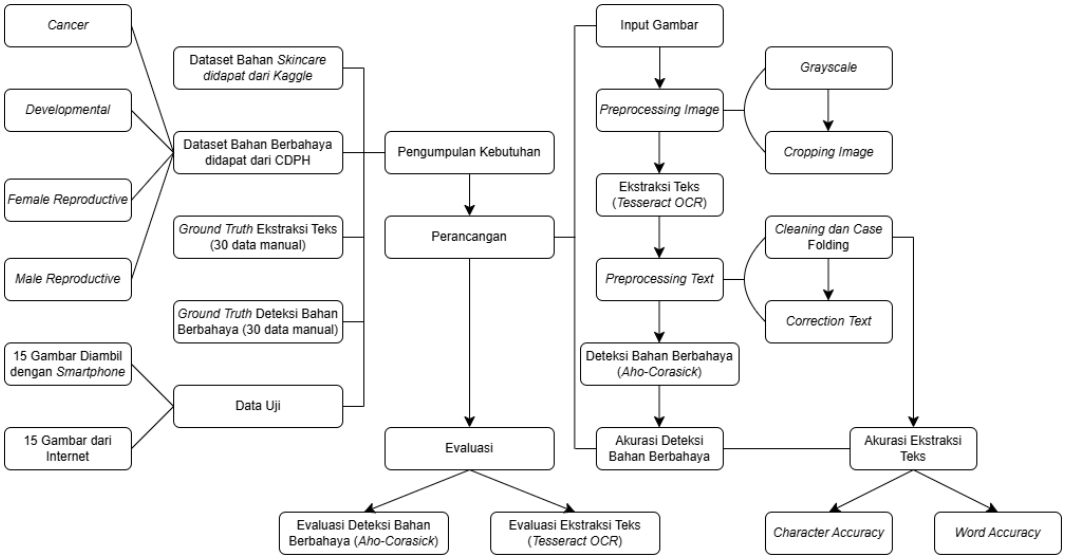
Gambar 1. Contoh Gambar Kemasan Produk *Skincare*

Selain dua dataset utama tersebut, disusun pula dua jenis *ground truth* secara manual. *Ground truth* pertama digunakan untuk evaluasi ekstraksi teks, yaitu hasil pencatatan manual terhadap seluruh bahan komposisi (*ingredients*) yang tercantum dalam gambar kemasan *skincare. Ground truth* kedua digunakan untuk evaluasi deteksi bahan berbahaya, berupa hasil pencatatan bahan berbahaya yang muncul dalam hasil ekstraksi OCR dari gambar kemasan. Masing-masing *ground truth* terdiri dari 30 data.

INDEXIA: Informatic and Computational Intelligent Journal
Vol.07, No.02, Nov 2025, hlm. 88-96
P-ISSN: 2657-0424
E-ISSN: 2657-0432

Data uji yang digunakan berupa 30 gambar kemasan produk *skincare*. Sebanyak 15 gambar diambil menggunakan kamera smartphone *Xiaomi Redmi Note 8* dengan resolusi 48 MP dan ukuran 1825 × 1825 piksel, sedangkan 15 gambar lainnya diperoleh dari internet dan mencakup berbagai merek populer seperti Wardah, Viva, Kahf, dan *Clean & Clear*.

2.2 Alur Penelitian



Gambar 2. Alur Penelitian

Flowchart pada Gambar 2. menyajikan alur keseluruhan sistem yang dikembangkan untuk melakukan ekstraksi teks dan deteksi bahan berbahaya pada produk *skincare*. Alur penelitian ini dimulai dari tahap pengumpulan kebutuhan data, dilanjutkan dengan perancangan dan implementasi sistem, hingga tahap evaluasi akurasi. Tahap pengumpulan kebutuhan mencakup pengumpulan beberapa jenis data penting, antara lain dataset bahan *skincare* yang diperoleh dari situs *Kaggle* dan dataset bahan berbahaya yang diambil dari *California Department of Public Health* (CDPH). Dataset bahan berbahaya ini terdiri dari beberapa kategori seperti *cancer, developmental, female reproductive,* dan *male reproductive*. Selain itu, disiapkan juga *ground truth* untuk ekstraksi teks sebanyak 30 data yang dibuat secara manual sebagai acuan pembanding hasil ekstraksi OCR. *Ground truth* untuk deteksi bahan berbahaya juga disiapkan sebanyak 30 data manual untuk evaluasi sistem deteksi.

Data uji yang digunakan dalam penelitian terdiri dari 30 gambar kemasan produk *skincare,* yang terbagi menjadi dua kelompok, yaitu 15 gambar hasil pengambilan menggunakan *smartphone* dan 15 gambar yang diunduh dari internet. Gambar-gambar tersebut dijadikan input untuk proses selanjutnya, yaitu *preprocessing image*. Pada tahap ini, gambar terlebih dahulu dikonversi ke format *grayscale* untuk menyederhanakan elemen visual yang akan diproses, lalu dilakukan *cropping* untuk memotong bagian yang tidak relevan, sehingga hanya bagian komposisi bahan yang dianalisis lebih lanjut.

Gambar hasil *preprocessing* kemudian diekstrak menggunakan *Tesseract OCR* untuk memperoleh teks bahan-bahan yang tercantum pada kemasan produk. Teks yang dihasilkan dari proses OCR selanjutnya diproses kembali melalui tahap pembersihan teks (*cleaning*), diikuti dengan *case folding* untuk menyamakan format huruf menjadi huruf kecil. Setelah itu, dilakukan koreksi terhadap hasil OCR agar mendekati teks asli sesuai ejaan bahan sebenarnya. Teks yang telah dibersihkan dan dikoreksi kemudian digunakan dalam proses deteksi bahan berbahaya. Proses ini dilakukan dengan mencocokkan setiap kata dalam teks dengan dataset bahan berbahaya menggunakan algoritma *Aho-Corasick*, yang berfungsi untuk mengidentifikasi keberadaan bahan berbahaya dalam daftar komposisi yang ditemukan.

Tahap akhir dari alur penelitian ini adalah evaluasi sistem, yang dibagi menjadi dua bagian utama. Evaluasi pertama adalah akurasi ekstraksi teks yang dilakukan terhadap hasil dari *Tesseract OCR*, dengan dua metrik pengukuran yaitu *Character Accuracy* untuk mengukur kesesuaian antar karakter, dan *Word Accuracy* untuk mengukur kesesuaian antar kata terhadap *ground truth*. Evaluasi kedua adalah akurasi deteksi bahan berbahaya, yang dilakukan dengan membandingkan hasil deteksi

sistem menggunakan algoritma *Aho-Corasick* dengan *ground truth* deteksi, guna menilai ketepatan sistem dalam mengidentifikasi bahan berbahaya pada produk *skincare*.

2.3 Perancangan Sistem

Perancangan sistem dimulai dari tahap input gambar, yang kemudian diproses melalui tahapan *preprocessing* citra, yaitu konversi ke *grayscale* dan pemotongan area teks yang relevan agar hanya bagian komposisi bahan yang dianalisis. Gambar hasil *preprocessing* kemudian diekstraksi menggunakan *Tesseract OCR*. *Tesseract OCR* merupakan mesin pengenalan karakter optik (*Optical Character Recognition*) *open-source* yang dikembangkan oleh *Google* dan mendukung berbagai bahasa serta struktur teks. Pada penelitian ini digunakan *Tesseract* versi 5.5.0, yang telah mengintegrasikan pendekatan *deep learning* berbasis *Long Short-Term Memory* (LSTM) untuk meningkatkan akurasi pengenalan karakter [11]. Proses kerja *Tesseract* dilakukan melalui beberapa tahap:

1. *Preprocessing* gambar, yaitu konversi gambar ke bentuk biner dan segmentasi area teks.
2. Pengenalan karakter, menggunakan model LSTM untuk mengenali huruf, angka, dan simbol dari blok teks.
3. *Postprocessing*, yaitu normalisasi hasil pengenalan dan koreksi ejaan menggunakan kamus internal atau eksternal [12].

Dalam implementasi sistem ini, *Tesseract* dijalankan dengan konfigurasi --oem 3 (*OCR Engine Mode: LSTM only)* dan --psm 6 (*Page Segmentation Mode: Assume a single uniform block of text*) agar sesuai dengan struktur daftar komposisi bahan pada kemasan *skincare*. Hasil ekstraksi berupa teks mentah kemudian diproses melalui tahap pembersihan teks (*cleaning*) dan *case folding* untuk menyamakan format huruf. Setelah itu, dilakukan koreksi ejaan sederhana terhadap teks hasil OCR agar mendekati bentuk aslinya.

Teks hasil koreksi selanjutnya dianalisis menggunakan algoritma *Aho-Corasick* untuk mendeteksi keberadaan bahan berbahaya. Algoritma *Aho-Corasick* merupakan algoritma pencarian pola (*multi-pattern matching*) yang sangat efisien. Algoritma ini membangun struktur pohon *trie* dari seluruh kata kunci (bahan berbahaya), kemudian menambahkan *fail links* agar proses pencarian dapat dilakukan dalam satu kali pemindaian teks [13][14]. Kompleksitas waktu algoritma ini adalah: $O(n + m + z)$, di mana n adalah panjang teks, m adalah panjang total pola yang dicari, dan z adalah jumlah kecocokan yang ditemukan [15]. Keunggulan *Aho-Corasick* adalah kemampuannya mencocokkan ratusan hingga ribuan entri bahan kimia secara simultan dan efisien, tanpa perlu memindai ulang teks untuk setiap kata kunci. Hasil pencocokan langsung mengembalikan nama bahan serta kategori risikonya (misalnya: *cancer, developmental, female reproductive* dan *male reproductive*). Hasil deteksi bahan berbahaya ini kemudian digunakan dalam tahap evaluasi untuk menilai kinerja sistem secara keseluruhan.

2.4 Evaluasi

Evaluasi sistem dilakukan pada dua aspek, yaitu akurasi ekstraksi teks dan akurasi deteksi bahan berbahaya. Pengukuran akurasi ekstraksi teks terdiri dari dua metrik utama, yaitu *Word Accuracy* (WA) dan *Character Accuracy* (CA). WA mengukur persentase kata yang dikenali dengan benar dibandingkan dengan *ground truth*, sedangkan CA mengukur *persentase* karakter yang dikenali dengan benar. Rata-rata WA dan CA dihitung menggunakan rumus 1 dan 2 sebagai berikut:

$$Rata - rata\ WA = \frac{1}{n}\sum_{i=1}^{n} WA_i \quad (1)$$

$$Rata - rata\ CA = \frac{1}{n}\sum_{i=1}^{n} CA_i \quad (2)$$

Dengan:

- $WA_i = Word\ Accuracy\ pada\ gambar\ ke - i$
- $CA_i = Character\ Accuracy\ pada\ gambar\ ke - i$
- $n = jumlah\ total\ gambar\ yang\ dievaluasi$
- $\sum = Notasi\ penjumlahan$

Sementara itu, evaluasi terhadap deteksi bahan berbahaya dilakukan menggunakan pendekatan *confusion matrix* yang terdiri dari empat elemen, yaitu *True Positive* (TP), *False Positive* (FP), *False Negative* (FN), dan *True Negative* (TN). Nilai-nilai tersebut digunakan untuk menghitung *Accuracy*, *Precision*, *Recall*, dan *F1-Score* dengan rumus 3 sampai 6 sebagai berikut:

INDEXIA: Informatic and Computational Intelligent Journal
Vol.07, No.02, Nov 2025, hlm. 88-96
P-ISSN: 2657-0424
E-ISSN: 2657-0432

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$precision = \frac{TP}{TP+FP} \quad (4)$$

$$recall = \frac{TP}{TP+FN} \quad (5)$$

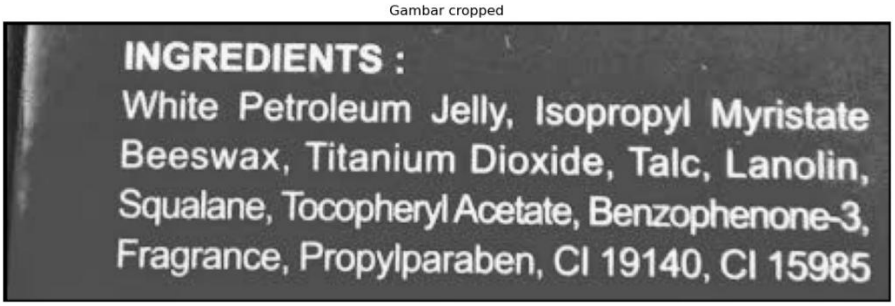
$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

Dengan metode dan pendekatan yang digunakan, sistem ini diharapkan mampu melakukan proses ekstraksi teks dan deteksi bahan berbahaya secara otomatis dan akurat pada produk *skincare*.

III. HASIL DAN PEMBAHASAN

Tahapan awal dalam penelitian ini adalah pengumpulan kebutuhan yang mencakup dua jenis dataset utama. Dataset pertama terdiri dari 5.328 bahan *skincare* yang diperoleh dari *Kaggle* dan disimpan dalam format *JSON* agar mudah digunakan dalam proses pencocokan dengan hasil ekstraksi teks. Dataset kedua berisi 1.004 bahan kimia berbahaya yang diambil dari *California Department of Public Health* (CDPH), yang diklasifikasikan ke dalam empat kategori risiko, yaitu *cancer, developmental, female reproductive*, dan *male reproductive*. Selain dataset utama, disusun pula dua *ground truth* untuk evaluasi sistem. *Ground truth* pertama berisi hasil pencatatan manual dari daftar komposisi bahan yang tertera pada 30 gambar kemasan produk *skincare*, sementara *ground truth* kedua berupa hasil pencocokan bahan berbahaya dari hasil ekstraksi teks dengan dataset bahan berbahaya. Kedua *ground truth* ini digunakan sebagai pembanding untuk mengukur kinerja sistem pada tahap evaluasi.

Data uji dalam penelitian ini terdiri dari 30 gambar kemasan produk *skincare*. Sebanyak 15 gambar diambil langsung menggunakan kamera *smartphone Xiaomi Redmi Note 8* dengan resolusi 1825x1825 piksel, sementara 15 gambar lainnya diperoleh dari internet dengan beragam merek seperti Wardah, Viva, Kahf, dan *Clean & Clear*. Tahapan perancangan sistem dimulai dari proses *preprocessing image* yang meliputi konversi ke format *grayscale* untuk meningkatkan kontras dan mengurangi *noise*, serta pemotongan (*cropping*) bagian gambar yang memuat teks "*Ingredients*" atau "Komposisi". Contoh hasil sistem ditunjukkan pada Gambar 3, yang memperlihatkan hasil *cropping* dari area komposisi bahan pada kemasan produk *skincare*.



Gambar 3. Hasil *Cropping Image*

Hasil ekstraksi teks dari gambar tersebut menggunakan *Tesseract OCR* adalah sebagai berikut:
Hasil Ekstraksi Teks:
INGREDIENTS : White Petroleum Jelly, Isopropyl Myristate Beeswax, Titanium Dioxide, Talc, Lanolin,

'Squalane, Tocopheryl Acetate, Benzophenone-3, Fragrance, Propylparabens, Cl 19140, Cl 15985

Proses ekstraksi teks dilakukan menggunakan *Tesseract OCR* dengan konfigurasi --oem 3 dan --psm 6 yang disesuaikan untuk mengenali blok teks horizontal. Setelah teks berhasil diekstraksi, dilakukan *preprocessing text* berupa *cleaning* untuk menghapus karakter tidak relevan dan *case folding* untuk menyamakan huruf menjadi huruf kecil. Proses ini diikuti dengan koreksi teks menggunakan metode *string similarity* dengan nilai ambang (*cutoff*) ≥ 0.6 untuk memperbaiki kesalahan pengenalan karakter akibat OCR. Gambar 3 menunjukkan hasil koreksi teks.

Tabel 1. Hasil Koreksi Teks

No	*Ingredients (Extracted)*	*Ingredients (Corrected)*
0	*aqua*	*aqua*
1	*mineral oil*	*mineral oil*
2	*stearic acid*	*stearic acid*
3	*cetyl alcohol*	*cetyl alcohol*
4	*iethanolamine*	*triethanolamine*
5	*methy\paraben*	*methylparaben*
6	*perfume*	*perfume*
7	*propylparaben*	*propylparaben*
8	*camellia sinensis extract*	*camellia sinensis extract*
9	*tea tree (melaleuca altern ia) oil*	*tea tree (melaleuca alternifolia) oil*
10	*bht*	*bht*
11	*ci 19140*	*ci 19140*
12	*ci 42090*	*ci 42090*

Tahap selanjutnya adalah deteksi bahan berbahaya menggunakan algoritma *Aho-Corasick*. Algoritma ini digunakan untuk mencocokkan bahan hasil *preprocessing* dengan daftar bahan berbahaya berdasarkan empat kategori risiko. Hasil deteksi ini divisualisasikan melalui pewarnaan dan keterangan kategori pada masing-masing bahan sebagimana ditunjukkan pada Tabel 2.

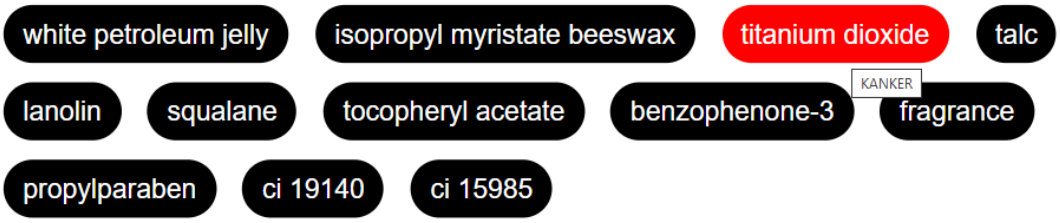
Tabel 2. Entitas atau Kategori untuk Deteksi Bahan Berbahaya

Entitas	***Color***	**Contoh *Visual***
Cancer	*Red*	aqua
Perkembangan	*Orange*	acrylates copolymer
Male Reproductive	*Blue*	propylene glycol
Female Reproductive	*Pink*	cocamidopropyl betaine
Cancer & Perkembangan	*Lime*	glycerin
Cancer & *Male Reproductive*	*Brown*	decyl glucoside
Cancer & *Female Reproductive*	*Gray*	sodium lauroyl sarcosinate
Perkembangan & *Male Reproductive*	*Dark slate gray*	potassium cocoyl glycinate
Perkembangan & *Female Reproductive*	*Yellow*	sodium pca
Male Reproductive & *Female Reproductive*	*Green*	sodium laureth sulfate
Cancer, Perkembangan & *Male Reproductive*	*Coral*	sodium chloride
Cancer, Perkembangan & *Female Reproductive*	*Khaki*	potassium cocoate

INDEXIA: Informatic and Computational Intelligent Journal
Vol.07, No.02, Nov 2025, hlm. 88-96
P-ISSN: 2657-0424
E-ISSN: 2657-0432

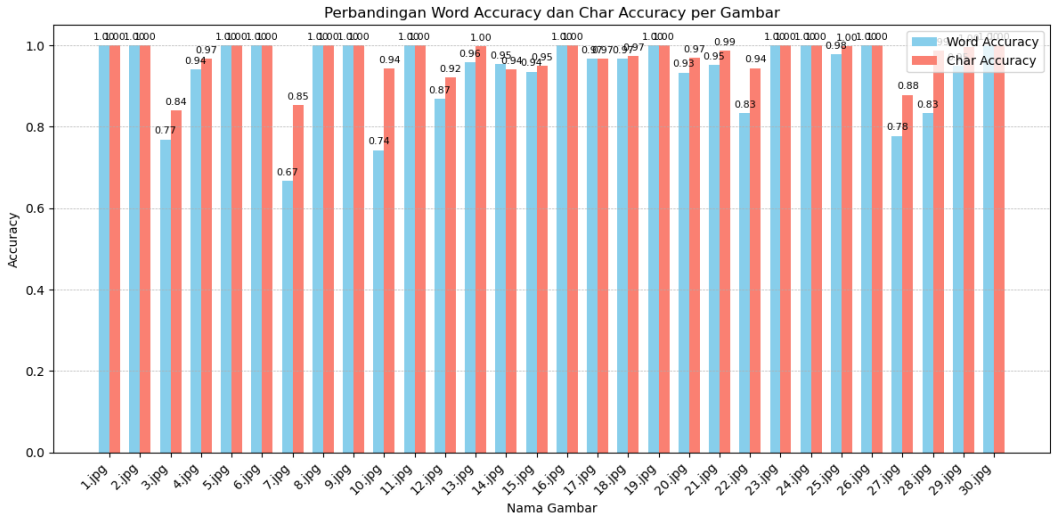
Perkembangan, *Male Reproductive & Female Reproductive*	*Teal*	phenoxyethanol
Semua Kategori	*Turquoise*	sodium hydroxide
Tidak Berbahaya	*Black*	disodium edta

Hasil visualisasi dari deteksi bahan berbahaya ditunjukkan pada Gambar 4, di mana setiap bahan yang berhasil dikenali diberi warna sesuai tingkat risiko. Warna merah menunjukkan bahan berbahaya (misalnya kategori *cancer*), sementara warna hitam menunjukkan bahan tidak berbahaya.



Gambar 4. Visualisasi hasil deteksi bahan berbahaya berdasarkan kategori risiko

Untuk mengevaluasi performa sistem, dilakukan pengukuran terhadap akurasi ekstraksi teks dan deteksi bahan berbahaya. Evaluasi menyeluruh dilakukan terhadap 30 gambar. Total *Word Accuracy* adalah 28.0287 dan *Character Accuracy* adalah 29.1171. Dengan total 30 gambar, rata-rata *Word Accuracy* sistem adalah 93,43% dan rata-rata *Character Accuracy* mencapai 97,06%. Satu gambar, yaitu 7.jpg, tercatat memiliki nilai *Word Accuracy* di bawah 70% (66,67%), sementara semua gambar lainnya menunjukkan *Character Accuracy* di atas 70%. Gambar 3 menunjukkan grafik perbandingan akurasi kata dan karakter hasil testing 30 gambar. Dalam evaluasi deteksi bahan berbahaya, Sistem menunjukkan kinerja yang sangat baik dalam mendeteksi bahan berbahaya, dengan capaian *accuracy, precision, recall*, dan *F1-score* sebesar 100%. Rincian evaluasi disajikan pada Tabel 3. Evaluasi lebih lanjut dilakukan dengan mengelompokkan bahan berdasarkan jumlah kategori risikonya. Untuk bahan dengan satu kategori (seperti *cancer, developmental, male reproductive*, atau *female reproductive*), sistem menunjukkan hasil deteksi dengan semua metrik evaluasi mencapai 100%. Tidak ditemukan bahan dengan dua, tiga, atau empat kategori dalam data uji.



Gambar 3. Grafik Perbandingan Akurasi Kata dan karakter dari 30 Gambar

Tabel 3. Hasil *Confusion Matrix* Deteksi Bahan Berbahaya

Jenis Evaluasi	TP	TN	FP	FN
Keseluruhan	4	627	0	0
Satu Kategori	4	627	0	0
Tidak Berbahaya	627	4	0	0

Untuk bahan yang tidak berbahaya, seluruh prediksi yang dilakukan sistem sesuai, sehingga akurasi deteksi bahan tidak berbahaya juga mencapai 100%. Hasil evaluasi ini menunjukkan bahwa sistem yang dibangun mampu mengekstraksi teks dari gambar kemasan dengan akurat dan mendeteksi bahan berbahaya dengan sangat baik, sehingga dapat digunakan sebagai alat bantu konsumen dalam mengidentifikasi keamanan produk *skincare* secara otomatis.

IV. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan sistem deteksi bahan berbahaya pada produk *skincare* berbasis gambar kemasan dengan memanfaatkan *Tesseract OCR* dan algoritma *Aho-Corasick*. Hasil evaluasi terhadap 30 gambar uji menunjukkan bahwa sistem memiliki kinerja tinggi dalam proses ekstraksi teks, dengan rata-rata *Word Accuracy (WA)* sebesar 93,43% dan *Character Accuracy (CA)* sebesar 97,06%. Meskipun terdapat satu gambar dengan WA di bawah 70%, seluruh gambar tetap menunjukkan CA di atas 70%, menandakan sistem cukup andal dalam mengenali karakter teks.

Pada tahap deteksi bahan berbahaya, sistem mencapai performa sempurna dengan nilai *accuracy*, *presicion*, *recall*, dan *F1-score* sebesar 100%. Deteksi berjalan optimal terutama pada bahan yang termasuk dalam satu kategori risiko (seperti *Cancer*, *Developmental*, *Male Reproductive dan Female Reproductive*), sedangkan bahan berkategori ganda belum ditemukan dalam data uji. Sistem juga terbukti mampu mengidentifikasi bahan yang tidak berbahaya dengan akurasi sempurna. Sistem ini hanya mendukung teks beraksara Latin, sehingga belum mampu mengenali bahan yang ditulis dalam aksara *non-Latin* seperti Jepang, Korea, atau Mandarin.

Ke depan, penelitian ini memiliki potensi besar untuk dikembangkan lebih lanjut. Salah satu arah pengembangan yang dapat dilakukan adalah dengan menambah variasi dan jumlah data uji agar sistem dapat diuji secara lebih luas dan representatif. Selain itu, peningkatan pada tahapan *preprocessing*, baik gambar maupun teks, akan memberikan dampak signifikan terhadap akurasi ekstraksi dan deteksi. Penelitian selanjutnya juga dapat mempertimbangkan penggunaan algoritma OCR yang lebih canggih untuk melakukan perbandingan performa dengan *Tesseract OCR*. Di sisi lain, integrasi pendekatan berbasis *deep learning* dalam proses deteksi bahan berbahaya dapat menjadi alternatif yang menjanjikan untuk meningkatkan ketepatan klasifikasi.

DAFTAR PUSTAKA

[1] A. Asroni, G. Indrawan, dan L. J. Erawati Dewi, "Implementasi Hirarki Dataset Dalam Membangun Model Language Aksara Bali Menggunakan Framework Tesseract OCR," *J. Resist. (Rekayasa Sist. Komputer)*, vol. 6, no. 1, hal. 20–28, 2023, doi: 10.31598/jurnalresistor.v6i1.1345.

[2] Syafriwan Nasution, "Rancang Bangun Aplikasi Batak Angkola Dictionary," vol. 3, no. 2, hal. 91–102, 2018.

[3] O. Lazhar dan B. Djamel, "Simd implementation of the Aho-Corasick algorithm using intel Avx2," *Scalable Comput.*, vol. 20, no. 3, hal. 563–576, 2019, doi: 10.12694/scpe.v20i3.1572.

[4] R. Haryanti, S. Auliya, dan M. Abdassah, "Artikel Ulasan: Tinjauan Bahan Berbahaya dalam Krim Pencerah Kulit," *Farmaka*, vol. 16, no. 2, hal. 214–224, 2018.

[5] A. N. Rahmawati, S. A. Wibowo, dan U. Sunarya, "Analisis Sistem Optical Character Recognition (Ocr) Pada Dokumen Digital Menggunakan Metode Tesseract Perfomance Analysis of Optical Character Recognition (Ocr) System on Digital Documents Using Tesseract Method," *e-Proceeding Eng.*, vol. 8, no. 5, hal. 4777–4785, 2021.

[6] Y. H. Tiara Susilo Putri, Anggalana, "Analisis Yuridis Perlindungan Konsumen Terhadap Kosmetik Kecantikan yang Tidak Layak Edar (Studi pada Badan Pengawas Obat Makanan BPOM Bandar Lampung)," vol. 3, no. 1, hal. 335–347, 2024.

[7] N. Yusuf, A. Wahyu, dan H. Habo, "Pengaruh Penggunaan Kosmetik (Whitening Cream) Terhadap Kadar Merkuri (Hg) Pada Perawat Magang Program Studi Profesi Ners Universitas Muslim Indonesia," *Wind. Heal. J. Kesehat.*, vol. 2, no. 3, hal. 206–217, 2019, doi: 10.33368/woh.v0i0.170.

[8] T. W. Ramdhani, I. Budi, dan B. Purwandari, "Optical Character Recognition Engines Performance Comparison in Information Extraction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, hal. 120–127, 2021, doi: 10.14569/IJACSA.2021.0120814.

[9] S. H. Naibaho, Nailufar Farha Afifah, Yuyun Umaidah, dan Nono Heryana, "Analysis of Student Reading Interest in UNSIKA Library with K-Means Algorithm," *Antivirus J. Ilm. Tek. Inform.*, vol. 18, no. 1, hal. 82–94, 2024, doi: 10.35457/antivirus.v18i1.2926.

[10] Daniel Jurafsky and James H. Martin, *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 2024. doi: 10.9783/9780812200027.toc.

[11] R. Devi, B. Kumar, dan P. Student, "Special issues on Computer Applications Image Processing Principles and Applications," hal. 56–59, 2020, [Daring]. Tersedia pada: www.internationaljournalssrg.org

[12] Reza Eka Alfarisi, "Rancang Bangun Aplikasi Terjemahan BahasaJepang - Indonesia Berbasis AndroidMenggunakan Tesseract OCR.," 2020.

[13] S. G. Ashish Gangurde, Gauri Dhumal, Shivam Gavandi, "Algoritma Pencocokan Pola dan Aplikasinya," 2023. https://medium.com/@gavadesnehal2/pattern-matching-algorithms-and-its-applications-30c95eaddaff

[14] N. F. Sulaeman dan M. Murnawan, "Implementasi Algoritma Aho-Corasick pada Pencarian di Aplikasi Lost and Found," *J. Edukasi dan Penelit. Inform.*, vol. 9, no. 3, hal. 509, 2023, doi: 10.26418/jp.v9i3.68389.

[15] B. Komalasari, Rita. Angelina, Joan., Meilani, *Pengantar ilmu komputer: teori komprehensif perkembangan ilmu komputer terkini*, no. January. 2023.