MULTIMODAL LEARNING MENGGUNAKAN EFFICIENTNETV2 DAN DISTILBERT UNTUK DETEKSI *WEBSITE* ILEGAL DAN IMPLEMENTASINYA PADA *BROWSER EXTENSION*

Sahal Maghfud[1], Ulfa Khaira[2] , Akhiyar Waladi[3]

[1,2,3]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Jambi, Jambi, Indonesia
e-mail: [1]sahalcocgood@gmail.com , [2]ulfakhaira@unja.ac.id,
[3]akhiyar.waladi@unja.ac.id

ABSTRAK

Tingginya penetrasi internet Indonesia memicu penyebaran konten ilegal yakni judi *online*, pornografi, serta pembajakan digital. Pemblokiran domain dinilai kurang efektif karena sangat mudah dihindari menggunakan domain baru maupun jaringan VPN. Penelitian ini menerapkan pendekatan *multimodal learning* yang menggabungkan informasi visual dari EfficientNetV2 M, data teks dari DistilBERT multilingual, dan struktur HTML berdasar frekuensi elemen. Seluruhnya digabungkan lewat teknik *early fusion* menjadi vektor berdimensi 2.187 lalu diklasifikasikan oleh Multi *Layer* Perceptron ke empat kategori normal, judi, pornografi, pembajakan. Kontribusi utama riset ini adalah menyusun dataset mandiri sejumlah 16.224 sampel melalui *web scraping* berlabel manual. Evaluasi membuktikan bahwa gabungan teks dan gambar memberikan performa terbaik dengan akurasi 88,74 persen, Macro F1 Score 0,8275, beserta *Macro Recall* 0,812. Demi menjaga ketahanan menghadapi variasi elemen web nyata, ekstensi browser Chromium tetap memakai fusi teks, gambar, dan HTML. Ekstensi sukses mengklasifikasikan 36 dari 40 *website* baru. Kesalahan pada pembajakan terjadi karena kurangnya data latih serta kemiripan visual.

Kata kunci: *multimodal learning*; deteksi *website* ilegal; EfficientNetV2-M; DistilBERT; *browser extension*

ABSTRACT

To address the ineffectiveness of domain blocking for illegal content such as online gambling, pornography, and piracy in Indonesia due to VPNs or new domains, this study proposes a Chromium extension based on multimodal learning. It fuses visual features using EfficientNetV2 M, text using multilingual DistilBERT, and HTML structure via early fusion into a vector of 2,187 dimensions, which is then classified by an MLP into four categories namely normal, online gambling, pornography, and piracy. Using a completely novel dataset consisting of 16,224 samples, the text and image combination achieved the best performance with an accuracy of 88.74%, a Macro F1 Score of 0.8275, and a Macro Recall of 0.812. For real world robustness, the extension utilizes all three modalities comprising text, image, and HTML, successfully classifying 36 of 40 unseen websites. Misclassifications occurred only in the digital piracy category due to limited training data and high visual similarity to legitimate websites..

Keywords*: multimodal learning; illegal website detection; EfficientNetV2-M; DistilBERT; browser extension*

1. PENDAHULUAN

Tingginya penggunaan internet di Indonesia (72,78% penduduk pada 2024) turut meningkatkan penyebaran konten ilegal seperti perjudian *online*, pornografi, penipuan, dan pelanggaran hak cipta [1]. Fenomena ini berpotensi merusak moral masyarakat, khususnya generasi muda [2] Kemkomdigi mencatat lebih dari 3,1 juta konten negatif ditindak selama Oktober 2024–Oktober 2025, dengan dominasi perjudian *online* dan pornografi, sementara kerugian akibat judi *online* dan pembajakan digital mencapai triliunan rupiah.

Konten ilegal menimbulkan berbagai dampak negatif, seperti kerugian hak cipta dan risiko *malware* akibat pembajakan, kecanduan serta gangguan mental akibat judi *online*, dan pengaruh terhadap fungsi kognitif maupun perilaku seksual akibat pornografi [3], [4], [5]. Oleh karena itu, diperlukan mekanisme deteksi yang efektif.

Pemblokiran situs oleh pemerintah masih dapat dihindari melalui VPN, *encrypted DNS*, maupun domain baru dan *mirror site* [6], [7]. Kondisi ini menunjukkan perlunya deteksi berbasis konten yang tetap efektif meskipun alamat domain berubah. Berbagai penelitian telah menggunakan *machine learning* untuk mendeteksi situs ilegal melalui teks, gambar, maupun struktur HTML dengan akurasi tinggi [8], [9], [10]. Namun, sebagian besar masih mengandalkan satu modalitas sehingga berisiko kehilangan informasi saat situs melakukan penyamaran.

Tabel 1 Penelitian terdahulu

Studi	Tahun	Dataset	Modalitas	Klasifier	Akurasi	Jenis Klasifikasi
Simanjuntak & Muhammad [8]	2025	Konten teks *website* judi *online*	Teks	CNN dan Word2Vec	100%	Biner
Çolhak et al.[9]	2024	*Screenshot website phishing*	Visual (*Screenshot*)	DenseNet	95,28%	Biner
Asdaghi et al.[10]	2020	*Source code* halaman *web spam*	Struktur HTML	*Machine Learning*	95,8%	Biner
Chen et al.[11]	2020	Daftar situs populer pornografi & perjudian	Visual (Spa-BoVW) dan Tekstual (HTML)	SVM, Doc2Vec, & *Random Forest*	> 99%	Biner
Wang et al.[12]	2022	*Website* perjudian	Visual dan Teks (OCR)	*Fine-tuned* ResNet34 &	> 99%	Biner

Studi	Tahun	Dataset	Modalitas	Klasifier	Akurasi	Jenis Klasifikasi
				BiLSTM-*Self-Attention*		
Penelitian Ini	2026	16.224 sampel dari *web scraping* (Situs Indonesia)	Visual (EfficientNetV2-M), Teks (DistilBERT), & Struktur HTML	*Multi-Layer Perceptron* (MLP) dengan *Early Fusion*	88,74%	Multi-kelas

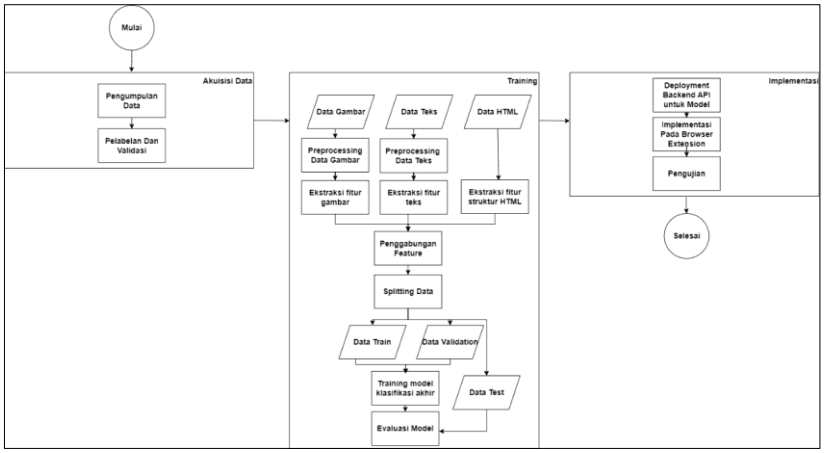
Penelitian [11] mengembangkan sistem deteksi situs pornografi dan perjudian menggunakan kombinasi fitur visual dan tekstual melalui mekanisme fusi keputusan. Dengan SVM, Doc2Vec, Random Forest, dan Spa-BoVW, model mencapai akurasi, presisi, dan *recall* di atas 99%, menunjukkan efektivitas pendekatan multimodal. Penelitian [12] menggunakan fusi multimodal berbasis fitur visual ResNet34 dan fitur teks dari OCR yang diproses dengan BiLSTM-*Self-Attention*. Pendekatan ini mencapai akurasi di atas 99% dan berhasil memanfaatkan informasi semantik yang sering diabaikan sistem konvensional. Penelitian [13] mengimplementasikan ekstensi Chrome berbasis *blacklist domain* dan penyensoran kata pada DOM untuk membatasi akses ke situs berbahaya. Meskipun berhasil melakukan pemblokiran dan sinkronisasi data, pendekatan ini kurang adaptif terhadap perubahan domain dan penyamaran konten.

Meskipun studi [11] dan [12] menunjukkan performa tinggi, keduanya masih berfokus pada klasifikasi biner dan dataset berbahasa asing. Kompleksitas konten ilegal berbahasa Indonesia memerlukan pendekatan multikelas yang mampu menangani variasi visual dan linguistik secara lebih komprehensif.

Berdasarkan kondisi tersebut, penelitian ini menerapkan pendekatan *multimodal learning* yang menggabungkan fitur visual, tekstual, dan struktur HTML untuk mendeteksi situs ilegal secara multikelas. Fitur visual diekstraksi menggunakan EfficientNetV2-M [14], sedangkan fitur teks menggunakan DistilBERT [15]. Ketiga modalitas diintegrasikan melalui *early fusion* untuk meningkatkan kemampuan deteksi terhadap berbagai bentuk penyamaran konten pada situs web berbahasa Indonesia. Model kemudian diimplementasikan pada *browser extension* guna membantu pengguna mengidentifikasi dan membatasi akses ke situs yang terdeteksi ilegal. Kontribusi utama penelitian ini adalah pembangunan dataset multimodal berisi 16.224 sampel situs web Indonesia, pengembangan arsitektur fusi tiga modalitas, evaluasi melalui tujuh skenario ablasi, serta implementasi sistem klasifikasi multikelas secara *real-time* pada *browser extension*.

2. METODE PENELITIAN

Penelitian ini dilaksanakan melalui beberapa tahapan sistematis untuk memastikan kualitas dan validitas hasil analisis. Tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Pengumpulan Data

Dataset terdiri atas situs normal dan ilegal. Situs ilegal dikumpulkan menggunakan kata kunci representatif yang diperluas melalui Semrush *Keyword Magic Tool*, kemudian URL diekstraksi dari SERP Bing dan diproses melalui *web scraping*. Pengumpulan data juga diperluas dengan memanfaatkan outbound link antar situs ilegal. Situs normal diperoleh dari Semrush Rank berdasarkan metrik SE Traffic Indonesia. Sekitar 5.000 domain teratas kategori *mobile* dan *desktop* digunakan sebagai representasi situs populer, serupa dengan pendekatan penelitian [11].

Data yang dikumpulkan mencakup teks berupa judul dan 1.500 karakter awal konten, *screenshot* halaman utama tanpa *scroll*, serta file HTML. Ketiga data tersebut dipilih untuk merepresentasikan informasi tekstual, visual, dan struktural situs Pelabelan dan validasi dilakukan secara manual untuk memastikan ketepatan kategori, membedakan situs ilegal dari situs informatif terkait, serta mencegah kesalahan klasifikasi akibat tingginya trafik situs ilegal.

Pre-processing Data Teks

Pre-processing teks meliputi perbaikan *encoding* menggunakan ftfy, penghapusan URL dengan *regular expression*, normalisasi spasi, dan normalisasi kata tidak baku menggunakan leksikon bahasa Indonesia informal [16]. Tahapan ini bertujuan menjaga konsistensi dan kualitas representasi linguistik. Contoh *Pre-processing* data teks dapat dilihat pada Tabel 2

Tabel 2 *Pre-processing* data teks

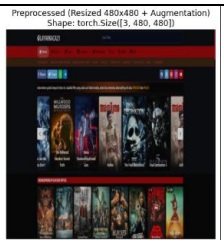
Teks Sebelum *Pre-processing*	Teks Sesudah *Pre-processing*
situs nonton film *online* bioskop sub indo gratis slide 10 to 16 of 8	situs menonton film *online* bioskop sub indonesia gratis slide 10 to16 of 8

Akses cepat melalui Domain alternatif kami https://jav.si Ingin mendapatkan ravished in rough	Akses cepat melalui Domain alternatif kami Ingin mendapatkan ravished ini rough

Pre-processing Data Gambar

Pre-processing gambar dilakukan melalui resize ke 480×480 piksel, konversi menjadi *tensor* dengan rentang nilai [0,1], serta normalisasi menggunakan parameter ImageNet agar sesuai dengan kebutuhan EfficientNetV2-M.. Contoh hasil *pre-processing* gambar dapat dilihat pada Tabel 3.

Tabel 3 Contoh hasil *pre-processing* data gambar

Gambar sebelum *Pre-processing*	**Gambar sesudah *Pre-processing***
	
	

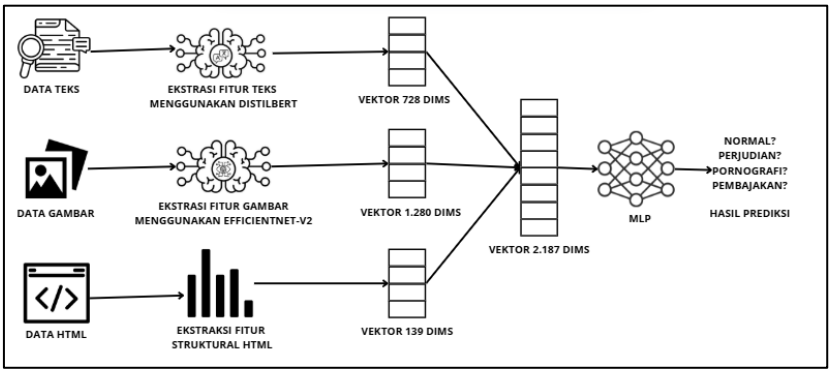
Ekstraksi fitur

Fitur visual diekstraksi menggunakan EfficientNetV2-M yang memanfaatkan *compound scaling* untuk menyeimbangkan akurasi dan efisiensi. Hasil akhirnya berupa vektor fitur berukuran 1.280 dimensi. Fitur teks diekstraksi menggunakan DistilBERT *multilingual cased* yang menghasilkan representasi kontekstual berdimensi 768 melalui enam lapisan Transformer berbasis *self-attention* Fitur struktur HTML diperoleh dengan menghitung frekuensi kemunculan 139 elemen HTML berdasarkan dokumentasi MDN dan pendekatan penelitian [17], kemudian dikonversi menjadi vektor numerik 139 dimensi.

Penggabungan Fitur dan Pembagian Data

Penggabungan fitur dilakukan menggunakan teknik *early fusion*, yaitu dengan menggabungkan langsung ketiga vektor fitur dari modalitas visual (1.280 dimensi), teks (768 dimensi), dan struktural HTML (139 dimensi) menjadi satu vektor fitur tunggal berukuran 2.187 dimensi. Pendekatan ini memungkinkan model klasifikasi mempelajari hubungan lintas-modalitas sehingga keputusan klasifikasi tidak hanya bergantung pada

satu aspek saja, melainkan pada keseluruhan representasi digital dari situs tersebut. Proses penggabungan fitur divisualisasikan pada Gambar 2.



Gambar 2. Proses penggabungan fitur

Setelah fitur digabungkan, dataset dibagi menjadi tiga bagian dengan proporsi 70% data *training*, 15% data validasi, dan 15% data *testing*.

Pelatihan Model Klasifikasi

Tahap penentuan kelas akhir dilakukan dengan memproses vektor fitur gabungan 2.187 dimensi hasil pendekatan *early fusion* menggunakan arsitektur *Multi-Layer Perceptron* (MLP) yang diimplementasikan melalui *framework* PyTorch. Pada dasarnya, MLP merupakan arsitektur jaringan saraf tiruan yang terdiri dari lapisan masukan, lapisan tersembunyi , dan lapisan keluaran yang saling terhubung penuh , serta memanfaatkan fungsi aktivasi non-linear untuk memodelkan pola data yang kompleks [18].

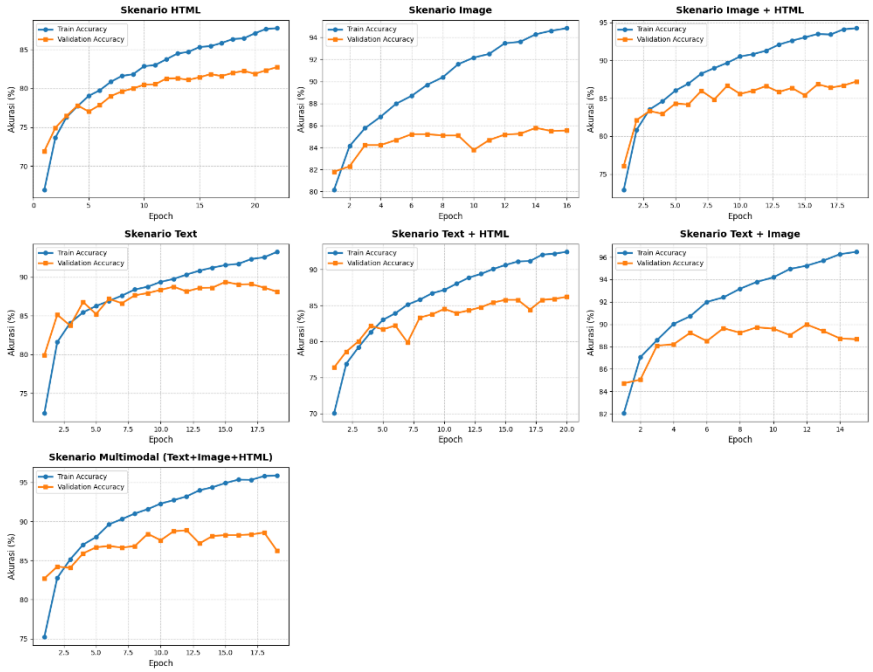
Model memetakan fitur ke empat kategori, yaitu normal, judi *online*, pornografi, dan pembajakan. Untuk mengukur kontribusi tiap modalitas, dilakukan pengujian pada modalitas tunggal, dua modalitas, dan tiga modalitas sekaligus. Arsitektur model yang digunakan ditunjukkan pada Tabel 4.

Tabel 4 Arsitektur MLP yang digunakan

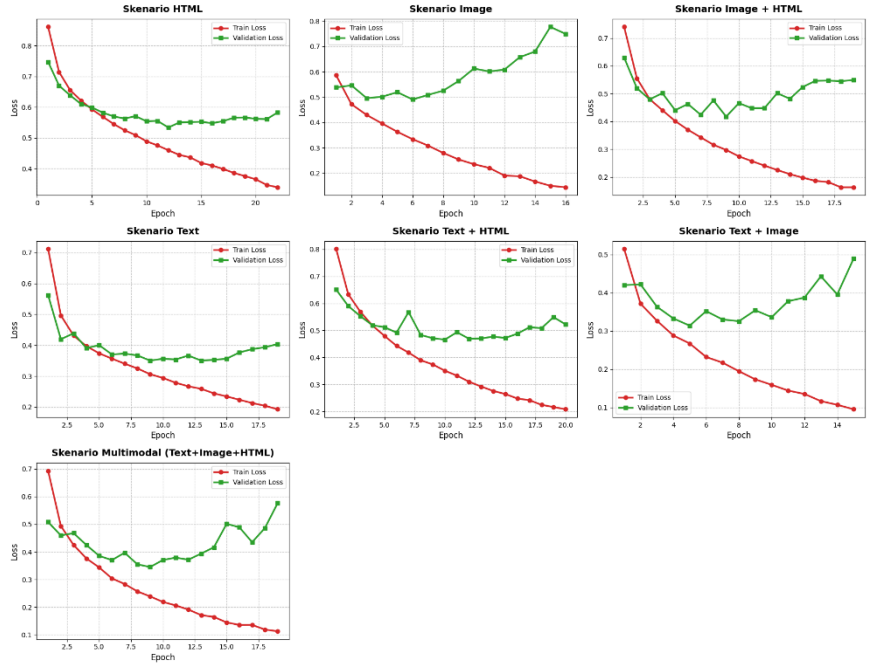
Nama *Layer*	Input Dimensi	Output Dimensi	Parameter
Layer Normalization	2187	2187	4.374
Linear *Layer*	2187	512	1.120.256
ReLU *Activation*	512	512	0
Dropout (0.2)	512	512	0
Linear *Layer*	512	128	65.664
ReLU *Activation*	128	128	0
Output *Layer*	128	4	516
Softmax	4	4	0
Total Parameter:			1,190,810

Proses pelatihan menggunakan *optimizer* Adam dengan *learning rate* 0,0005 dan *batch size* 16. Untuk mencegah *overfitting*, diterapkan mekanisme *early stopping* yang

memantau nilai *validation loss* pada setiap *epoch* dengan parameter *patience* 5 dan *min_delta* 0,001. Model dengan nilai *validation loss* terbaik disimpan sebagai *checkpoint* dan digunakan sebagai model akhir.



Gambar 3 Kurva Akurasi Pelatihan dan Validasi pada Berbagai Skenario



Gambar 4 Kurva nilai *loss* Pelatihan dan Validasi pada Berbagai Skenario

Berdasarkan Gambar 3 dan Gambar 4 , model mampu mempelajari pola fitur dengan baik, ditandai oleh peningkatan akurasi dan penurunan *loss* yang konsisten pada

fase pelatihan. Namun, performa pada fase validasi cenderung mencapai titik jenuh (*plateau*) lebih cepat di kisaran akurasi 82%–89%. Terlihat jelas adanya indikasi *overfitting* pada beberapa skenario kompleks, di mana kurva *validation loss* mulai bergerak naik meskipun metrik pelatihan terus membaik. Kondisi inilah yang membuat penerapan *early stopping* dan checkpointing menjadi sangat penting; proses pelatihan berhasil dihentikan tepat di titik *validation loss* terendah sebelum *overfitting* memburuk, sehingga menghasilkan model yang tangguh dan optimal dalam menggeneralisasi data baru.

Evaluasi

Metrik evaluasi yang digunakan meliputi *Overall Accuracy*, *Precision*, *Recall*, dan F1-Score. Rumus evaluasi model dijabarkan pada Persamaan (1) hingga (4),

$$Overall\ Accuracy = \frac{\sum TP}{\text{Total Seluruh Data}} \quad (1)$$

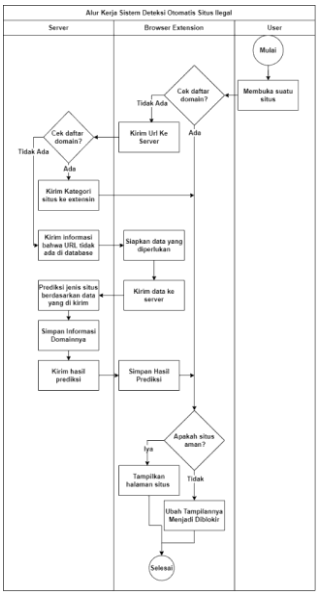
$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Deployment* dan Implementasi *Browser Extension

Model di-deploy menggunakan FastAPI yang menyediakan *endpoint* untuk prediksi multimodal, pengecekan domain yang telah diproses, dan pelaporan kesalahan klasifikasi oleh pengguna. Lalu *browser extension* Chrome secara otomatis mengambil *screenshot*, mengekstraksi teks, dan menganalisis struktur HTML sebelum mengirim data ke server. Situs yang terdeteksi ilegal akan diblokir, sementara pengguna dapat melakukan koreksi melalui autentikasi PIN.. Alur Kerja Sistem Deteksi Otomatis Situs Ilegal disajikan dalam Gambar 5.



Gambar 5. Alur Kerja Sistem Deteksi Otomatis Situs Ilegal

Pengujian Sistem

Pengujian terbagi menjadi dua tahap, diawali dengan *black box testing* guna memvalidasi fungsionalitas *browser extension* seperti klasifikasi data dari lokal maupun server, prediksi di server, tampilan halaman *blocked*, perubahan status oleh pengguna, serta *setup* PIN awal. Tahap selanjutnya mengevaluasi performa model pada kondisi nyata dengan mengakses langsung URL baru melalui browser, menggunakan data yang sepenuhnya di luar dataset pelatihan, validasi, maupun *testing* sebelumnya..

3. HASIL DAN PEMBAHASAN

Pengumpulan data

Hasil pengumpulan data dari dua sumber utama, yaitu top domain Semrush berupa 11.037 tautan unik *website* normal dan pencarian berbasis kata kunci berupa 5.897 tautan unik *website* ilegal, menghasilkan 23.860 data awal. Setelah melalui proses validasi dan deduplikasi, jumlah tersebut disaring menjadi 20.447 data. Melalui tahap akhir berupa *scraping*, pelabelan, dan pembersihan eror, diperoleh total 16.224 sampel data final bermodalitas tekstual, visual, dan struktur HTML. Distribusi akhir dari keseluruhan data tersebut mencakup 8.780 *website* normal, 4.184 judi *online*, 2.160 pornografi, dan 1.100 *website* pembajakan.

Penggabungan fitur dan pembagian data

Dataset multimodal yang digunakan untuk tahapan *training* model merupakan hasil penyatuan vektor dari ketiga jenis ekstraksi fitur. Detail susunan kolom dan sampel representasi numerik dari hasil penggabungan fitur ini dapat dilihat pada Tabel 5.

Tabel 5 Contoh data hasil penggabungan fitur

Fitur Teks			**Fitur Gambar**			**Fitur Struktur HTML**		
text_0	**...**	**text_767**	**image_0**	**...**	**image_1**	**html_0**	**...**	**html_138**
-0.0551	...	-0.28167	-0.21335	...	-0.19400	278	...	0
-0.1725	...	0.020611	-0.18617	...	-0.05156	17	...	0
0.16593	...	-0.03487	0.241445	...	0.10971	100	...	0

Dataset multimodal yang telah digabungkan selanjutnya dibagi untuk kebutuhan pelatihan dan evaluasi model klasifikasi. Hasil pembagian data ke dalam set *training*, *validation*, dan *testing* dapat dilihat pada Tabel 6. Meskipun distribusi kelas tidak seimbang, data asli tetap digunakan karena teknik penyeimbangan yang diuji sebelumnya tidak memberikan peningkatan performa yang signifikan.

Tabel 6 Hasil pembagian data

Label	**Data *Training***	**Data *Validation***	**Data *Testing***	**Total**
0 (Normal)	6145	1318	1317	8780
1 (Judi *Online*)	2929	627	628	4184
2 (Pornografi)	1512	324	324	2160
3 (Pembajakan)	770	165	165	1100

Evaluasi

Kinerja model dievaluasi secara mendalam menggunakan metrik klasifikasi utama, yaitu *Recall*, *Precision*, dan F1-Score untuk masing-masing kelas, beserta nilai rata-rata makronya. Evaluasi dilakukan pada berbagai skenario modalitas untuk mengukur kontribusi masing-masing serta kombinasinya. Adapun pengujian ini mencakup empat kategori kelas, yaitu: Kelas 0 (Normal), Kelas 1 (Judi *Online*), Kelas 2 (Pornografi), dan Kelas 3 (Pembajakan).

Tabel 7. Nilai *recall* per kelas untuk masing masing modalitas

Modalitas	***Recall***			
	Normal	**Judi *online***	**Pornografi**	**Pembajakan**
Teks	0.9263	0.9045	0.7593	0.5939
Gambar	0.9506	0.8105	0.7932	0.3758
Struktur HTML	0.9355	0.7834	0.5802	0.3758

Teks dan HTML	0.937	0.8662	0.6358	0.4848
Gambar dan HTML	0.9506	0.871	0.7531	0.4121
Teks dan Gambar	0.9226	0.9315	0.8179	0.5758
Teks, Gambar, dan HTML	0.9522	0.9076	0.7315	0.5515

Berdasarkan pengujian *recall* pada Tabel 7 menunjukkan seluruh modalitas mampu mengenali situs normal dengan baik (>0,92). Kombinasi teks dan gambar memberikan *recall* tertinggi pada kategori judi *online* dan pornografi, sedangkan kategori pembajakan menjadi yang paling sulit dideteksi..

Tabel 8 Nilai *precision* per kelas untuk masing masing modalitas

Modalitas	***Precision***			
	Normal	**Judi *online***	**Pornografi**	**Pembajakan**
Teks	0.8777	0.9016	0.9318	0.6533
Gambar	0.8336	0.8946	0.9113	0.7654
Struktur HTML	0.7948	0.8897	0.7932	0.6596
Teks dan HTML	0.8417	0.8962	0.8477	0.678
Gambar dan HTML	0.8575	0.9224	0.8531	0.7158
Teks dan Gambar	0.8987	0.8931	0.9397	0.6552
Teks, Gambar, dan HTML	0.8739	0.9283	0.908	0.7339

Berdasarkan hasil pengujian pada Tabel 8, Kombinasi teks dan gambar menghasilkan *precision* terbaik pada kategori normal dan pornografi. Penambahan fitur HTML meningkatkan *precision* kategori judi *online*, sementara kategori pembajakan menunjukkan *precision* lebih tinggi daripada *recall* sehingga model cenderung konservatif dalam memberikan prediksi.

Tabel 9 Nilai F1-score per kelas untuk masing masing modalitas

Modalitas	**F1-Score**			
	Normal	**Judi *online***	**Pornografi**	**Pembajakan**
Teks	0.9014	0.903	0.8367	0.6222
Gambar	0.8883	0.8505	0.8482	0.5041

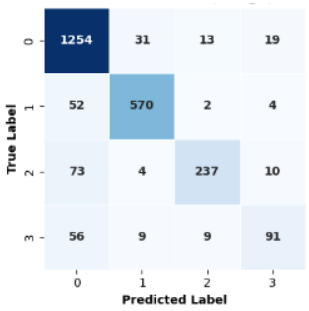
Struktur HTML	0.8594	0.8332	0.6702	0.4788
Teks dan HTML	0.8868	0.881	0.7266	0.5654
Gambar dan HTML	0.9017	0.896	0.8	0.5231
Teks dan Gambar	0.9105	0.9119	0.8746	0.6129
Teks, Gambar, dan HTML	0.9113	0.9179	0.8103	0.6298

Melalui Tabel 9 dapat dilihat bahwa kombinasi tiga modalitas menghasilkan F1-Score terbaik untuk kategori normal, sedangkan kombinasi teks dan gambar unggul pada kategori judi *online* dan pornografi. Deteksi pembajakan masih menjadi tantangan utama pada seluruh skenario..

Tabel 10 Nilai macro average masing masing modalitas

Modalitas	**Macro *Precision***	***Macro Recall***	**Macro F1-Score**	***Overall Accuracy***
Teks	0.8411	0.7960	0.8158	0.8759
Gambar	0.8512	0.7325	0.7728	0.8546
Struktur HTML	0.7843	0.6687	0.7104	0.8110
Teks dan HTML	0.8159	0.7310	0.7650	0.8480
Gambar dan HTML	0.8372	0.7467	0.7802	0.8673
Teks dan Gambar	0.8467	0.8120	0.8275	0.8874
Teks, Gambar, dan HTML	0.8610	0.7857	0.8173	0.8841

Pada Tabel 10 menunjukkan bahwa kombinasi modalitas Teks dan Gambar menghasilkan akurasi tertinggi sebesar 0.8874 serta *Macro F1-Score* terbaik di angka 0.8275. Kombinasi ini memiliki agresivitas deteksi (*Macro Recall*) yang lebih baik, yaitu sebesar 0.8120. Sebaliknya, penggunaan fitur Struktur HTML secara independen justru menghasilkan nilai kinerja terendah dengan *Macro F1-Score* 0.7104, menunjukkan bahwa hanya mengandalkan kode struktural halaman web tidaklah cukup untuk membedakan kategori konten berbahaya.



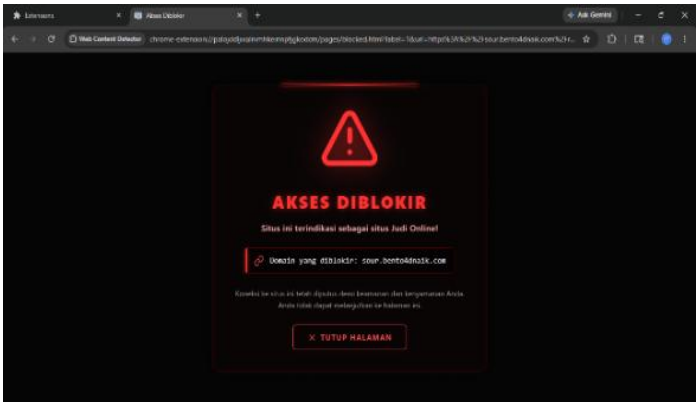
Gambar 6 *Confusion Matrix Model Skenario Teks dan Gambar*

Untuk melengkapi metrik evaluasi sebelumnya, pola kesalahan prediksi divisualisasikan melalui *Confusion Matrix* pada Gambar 6. *Confusion Matrix* menunjukkan model sangat baik mengenali situs normal, namun masih sering mengklasifikasikan beberapa situs judi, pornografi, dan pembajakan sebagai situs normal. Hal ini mengindikasikan adanya kemiripan fitur antara sebagian situs ilegal dan situs legal.

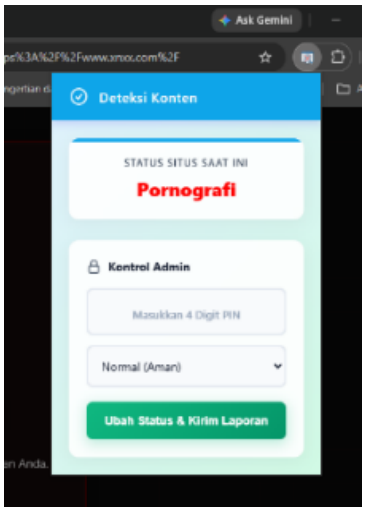
Meskipun kombinasi Teks dan Gambar memberikan performa terbaik, *browser extension* tetap mengimplementasikan tiga modalitas, yaitu Teks, Gambar, dan Struktur HTML. Keputusan ini bertujuan membuktikan bahwa arsitektur fusi yang lebih kompleks dapat diterapkan dengan baik sekaligus menjadi dasar pengembangan teknik fusi data yang lebih optimal pada penelitian selanjutnya, khususnya terkait pembobotan fitur HTML.

Implementasi

Sistem yang telah dirancang kemudian direalisasikan ke dalam bentuk *browser extension*. Tampilan halaman pemblokiran saat situs terdeteksi ilegal dapat dilihat pada Gambar 7, sedangkan antarmuka *pop-up* ekstensi ditunjukkan pada Gambar 8.



Gambar 7 Tampilan halaman pemblokiran



Gambar 8. Antarmuka Popup

Pengujian

Pengujian fungsionalitas dilakukan pada fitur-fitur utama *browser extension* dan *Backend* API untuk memastikan sistem berjalan sesuai kebutuhan. Hasil pengujian dari berbagai skenario tersebut ditunjukkan pada Tabel 11

Tabel 11 Pengujian fitur sistem

Fitur yang Diuji	**Skenario dan Hasil yang Diharapkan**	**Hasil Pengujian**
Setup PIN Awal	Pengguna membuat PIN saat instalasi dan PIN berhasil disimpan	Berhasil
Halaman *Blocked*	*Website* ilegal diakses dan halaman blokir ditampilkan	Berhasil
Klasifikasi Data Lokal	*Website* tersedia di penyimpanan lokal dan hasil klasifikasi ditampilkan tanpa request ke server	Berhasil
Klasifikasi Data Server	*Website* tersedia di server dan sistem menampilkan hasil prediksi	Berhasil
Klasifikasi Data Baru	*Website* belum ada di lokal maupun server dan sistem menampilkan hasil klasifikasi dari server	Berhasil
Perubahan Status *Website*	Pengguna mengubah label *website* dan perubahan berhasil disimpan serta dilaporkan ke server	Berhasil
Autentikasi PIN	Pengguna memasukkan PIN yang salah dan sistem menolak perubahan data	Berhasil

Berdasarkan Tabel 11, seluruh fitur sistem telah berhasil berjalan sesuai skenario yang ditentukan. Hal ini membuktikan bahwa sistem telah berfungsi dengan baik dalam mendukung proses deteksi dan pemblokiran *website* ilegal secara otomatis. Pengujian pada 40 *website* baru menunjukkan 36 *website* berhasil diklasifikasikan dengan benar. Kesalahan hanya terjadi pada kategori pembajakan akibat keterbatasan data latih dan kemiripan dengan situs legal. Namun, evaluasi lebih luas masih diperlukan untuk memperoleh gambaran performa yang lebih representatif.

4. KESIMPULAN

Sistem deteksi *website* ilegal multimodal yang menggabungkan analisis visual berbasis EfficientNetV2-M, teks melalui DistilBERT, dan struktur HTML berhasil dibangun. Meskipun pengujian membuktikan bahwa kombinasi teks dan gambar memberikan performa terbaik dengan akurasi 88,74 persen, *Macro F1-Score* 0,8275, dan *Macro Recall* 0,812, sistem ini tetap diimplementasikan secara utuh menggunakan ketiga modalitas yakni teks, gambar, dan HTML ke dalam ekstensi peramban Chromium. Model ini sangat unggul mengenali situs normal, judi *online*, dan pornografi, namun masih kesulitan pada kategori pembajakan digital akibat terbatasnya data latih. Ekstensi peramban berfungsi penuh dengan serangkaian fitur seperti klasifikasi otomatis, pemblokiran, verifikasi PIN, serta pelaporan koreksi. Pengujian pada data baru menunjukkan sistem andal mendeteksi judi dan pornografi dengan skor kepercayaan tinggi, walau misklasifikasi masih terjadi pada situs pembajakan atau situs dengan tampilan yang kurang representatif. Oleh karena itu, penelitian selanjutnya disarankan untuk memperbanyak sampel data pembajakan, mengintegrasikan metode interpretabilitas seperti SHAP, serta mengeksplorasi implementasi pada perangkat seluler atau *edge device* guna memperluas jangkauan sistem.

DAFTAR PUSTAKA

[1] S. Mariyam, "Regulasi Konten Ilegal Pada Media Berbasis Teknologi Informasi," *cita huk. indonesa.*, vol. 1, no. 2, Nov 2022, doi: 10.57100/jchi.v1i2.19.

[2] N. Afiqa, S. Sarah, N. K. Aghna, A. F. J. M. Aziz, dan S. Supriyono, "Degradasi Moralitas Generasi Muda di Era Globalisasi: Relevansi Pendidikan Kewarganegaraan sebagai Benteng Karakter," *J. Pendidik. Tambusai*, vol. 9, no. 3, hlm. 37084–37088, 2025.

[3] A. Iqbal, M. N. Aman, R. Rejendran, dan B. Sikdar, "Unveiling the Connection Between *Malware* and Pirated Software in Southeast Asian Countries: A Case Study," *IEEE Open J. Comput. Soc.*, vol. 5, hlm. 62–72, 2024, doi: 10.1109/OJCS.2024.3364576.

[4] L. Rafiqah dan H. Rasyid, "The Dampak Judi *Online* terhadap Kehidupan Sosial Ekonomi Masyarakat," *Al-Mutharahah J. Penelit. Dan Kaji. Sos. Keagamaan*, vol. 20, no. 2, hlm. 282–290, Des 2023, doi: 10.46781/al-mutharahah.v20i2.763.

[5] C. Afriliani, N. A. Azzura, dan J. R. B. Sembiring, "Faktor Penyebab dan Dampak dari Kecanduan Pornografi di Kalangan Anak Remaja Terhadap Kehidupan Sosialnya," *Harmony J. Pembelajaran IPS Dan PKN*, vol. 8, no. 1, hlm. 7–14, Agu 2023, doi: 10.15294/harmony.v8i1.61470.

[6] F. P. Handoko dan I. M. H. Wijaya, "Efektivitas Permenkominfo No. 19 Tahun 2014 Tentang Penanganan Situs Internet Bermuatan Negatif Terhadap Penyalahgunaan Aplikasi Virtual Private Network," *J. Mhs. Huk. Saraswati*, vol. 03, no. 1, hlm. 866–877, 2023.

[7] A. R. Julian, S. Suwarno, dan P. Syah, "Upaya Dan Tantangan Diskomdigi Dalam Pemblokiran Situs Judi *Online* Di Provinsi Lampung," vol. 4, no. 1, hlm. 111–123, 2025.

[8] N. Simanjuntak dan A. H. Muhammad, "Analisis Perbandingan Algoritma SVM dan CNN dalam Mendeteksi *Website* Judi *Online* Berdasarkan Konten Teks," *Bull.*

Comput. Sci. Res., vol. 5, no. 4, hlm. 361–371, Jun 2025, doi: 10.47065/bulletincsr.v5i4.586.

[9] F. Çolhak, M. İ. Ecevit, dan H. Dağ, "Transfer Learning for *Phishing* Detection: *Screenshot*-Based *Website* Classification," dalam *2024 9th International Conference on Computer Science and Engineering (UBMK)*, Antalya, Turkiye: IEEE, Okt 2024, hlm. 1–6. doi: 10.1109/UBMK63289.2024.10773490.

[10] F. Asdaghi, A. Soleimani, dan M. Zahedi, "A Novel Set of Contextual Features for Web *Spam* Detection," *Int. J. Nonlinear Anal. Appl.*, vol. 11, no. 1, Jan 2020, doi: 10.22075/ijnaa.2020.4297.

[11] Y. Chen, R. Zheng, A. Zhou, S. Liao, dan L. Liu, "Automatic Detection of Pornographic and Gambling Websites Based on Visual and Textual Content Using a Decision Mechanism," *Sensors*, vol. 20, no. 14, hlm. 3989, Jul 2020, doi: 10.3390/s20143989.

[12] C. Wang, M. Zhang, F. Shi, P. Xue, dan Y. Li, "A *Hybrid* Multimodal Data Fusion-Based Method for Identifying Gambling Websites," *Electronics*, vol. 11, no. 16, hlm. 2489, Agu 2022, doi: 10.3390/electronics11162489.

[13] R. F. Ramadhan dan A. Fauzan, "Pembatasan Internet Berbasis Ekstensi Web pada Chrome Browser," *Proc. Ser. Phys. Form. Sci.*, vol. 6, hlm. 192–199, Okt 2023, doi: 10.30595/pspfs.v6i.869.

[14] M. Tan dan Q. Le, "EfficientNetV2: Smaller Models and Faster *Training*," dalam *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021, hlm. 10096–10106.

[15] V. Sanh, L. Debut, J. Chaumond, dan T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," dalam *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS*, 2019.

[16] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, dan A. Jamal, "Colloquial Indonesian Lexicon," dalam *2018 International Conference on Asian Language Processing (IALP)*, Bandung, Indonesia: IEEE, Nov 2018, hlm. 226–229. doi: 10.1109/IALP.2018.8629151.

[17] A. Almomani *dkk.*, "*Phishing Website* Detection With Semantic Features Based on *Machine Learning* Classifiers: A Comparative Study," *Int. J. Semantic Web Inf. Syst.*, vol. 18, no. 1, hlm. 1–24, Feb 2022, doi: 10.4018/IJSWIS.297032.

[18] A. Zhang, Z. Lipton, M. Li, dan A. J. Smola, *Dive into deep learning*. Cambridge New York Port Melbourne New Delhi Singapore: Cambridge University Press, 2024.