BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan internet dan media sosial di indonesia dalam satu dekade terakhir telah menunjukkan pertumbuhan yang sangat pesat. Berdasarkan laporan DataReportal tahun 2024, tercatat lebih dari 212 juta masyarakat indonesia telah terhubung ke internet, dengan sebagian besar merupakan pengguna aktif media sosial yang berasal dari generasi milenial dan Gen Z [1]. kelompok usia ini sangat aktif dalam berinteraksi secara digital, baik untuk kebutuhan informasi, hiburan, maupun transaksi ekonomi. Namun, di balik tingginya penetrasi digital tersebut, kesadaran masyarakat terhadap keamanan siber masih tergolong rendah. Banyak pengguna belum memahami prinsip dasar keamanan digital, seperti mengenali tautan URL web mencurigakan, membedakan antara situs asli dan palsu, serta melindungi data pribadi di dunia maya. Ketidaktahuan ini menjadikan masyarakat sebagai target yang rentan terhadap berbagai jenis serangan siber, khususnya serangan *website phishing* yang marak terjadi belakangan ini [2].

Perkembangan internet dan media sosial di indonesia dalam satu dekade terakhir telah menunjukkan pertumbuhan yang sangat pesat. Berdasarkan laporan DataReportal tahun 2024, tercatat lebih dari 212 juta masyarakat indonesia telah terhubung ke internet, dengan sebagian besar merupakan pengguna aktif media sosial yang berasal dari generasi milenial dan Gen Z. kelompok usia ini sangat aktif dalam berinteraksi secara digital, baik untuk kebutuhan informasi, hiburan, maupun

transaksi ekonomi. Namun, di balik tingginya penetrasi digital tersebut, kesadaran masyarakat terhadap keamanan siber masih tergolong rendah. Banyak pengguna belum memahami prinsip dasar keamanan digital, seperti mengenali tautan URL web mencurigakan, membedakan antara situs asli dan palsu, serta melindungi data pribadi di dunia maya. Ketidaktahuan ini menjadikan masyarakat sebagai target yang rentan terhadap berbagai jenis serangan siber, khususnya serangan website phishing yang marak terjadi belakangan ini [3].

Phishing termasuk salah satu bentuk serangan siber berbasis rekayasa sosial yang menunjukkan peningkatan signifikan dalam beberapa tahun terakhir. Pada masa pandemi COVID 19, intensitas serangan ini melonjak hingga 220% lebih tinggi dibandingkan periode sebelum pandemi akibat meningkatnya aktivitas digital masyarakat global. Memasuki tahun 2022, ancaman phishing kian terlihat nyata dengan tercatat lebih dari 1,27 juta serangan pada kuartal ketiga, menjadikannya periode dengan insiden tertinggi sepanjang sejarah pemantauan [4]. Kondisi tersebut berlanjut pada tahun 2023, di mana laporan Anti Phishing Working Group mengungkapkan terdapat lebih dari 1,3 juta insiden phishing hanya dalam satu kuartal, memperlihatkan bahwa pelaku semakin agresif menargetkan pengguna [5]. Kejadian ini tidak menurun pada 2024, jumlah serangan phising hanya dalam satu kuartal kembali melonjak sekitar 27% dibanding tahun sebelumnya, dengan rata rata lebih dari 1,2 miliar url phishing dikirim setiap harinya dan biaya kerugian per pelanggaran data mencapai 4,8 Juta USD, karakteristik serangan pun semakin berbahaya sekitar 76,4% serangan phishing di tahun 2024 bersifat polimorfik yaitu menggunakan variasi kecil pada email untuk

menghindari deteksi dan sekitar 73,8% di antaranya memanfaatkan kecerdasan buatan (AI) untuk menghasilkan pesan yang sangat meyakinkan [6]. Fakta ini menegaskan bahwa metode deteksi tradisional, seperti daftar hitam (blacklist) URL atau pengecekan manual domain, tidak lagi cukup efektif, sehingga dibutuhkan pendekatan adaptif berbasis machine learning dan teknik analisis perilaku guna menghadapi evolusi serangan phishing modern.

Seiring dengan kemajuan teknologi, pendekatan berbasis kecerdasan buatan (Artificial Intelligence atau AI), khususnya machine learning (ML), mulai digunakan untuk mendeteksi URL web Phishing secara otomatis dan lebih akurat. ML memungkinkan sistem untuk mempelajari pola-pola dari data historis dan mengenali ciri-ciri URL web mencurigakan, seperti panjang URL yang tidak wajar, penggunaan simbol tertentu, hingga struktur domain yang tidak biasa. Salah satu algoritma machine learning yang menunjukkan performa unggul dalam tugas klasifikasi phishing adalah CatBoost (Categorial Boosting). Algoritma ini dikembangkan oleh yandex dan memiliki keunggulan dalam menangani data kategorial, mengurangi risiko overfiting, serta menghasilkan model prediktif yang akurat tanpa banyak CatBoost dapat mencapai akurasi deteksi hingga 97,3%, mengungguli algoritma populer lainnya seperti Random Forest, XGBoost, dan LightGBM [7].

Meskipun demikian, sebagian besar implementasi sistem deteksi web phishing saat ini masih terbatas pada platform desktop atau berbasis web. Padahal, menurut DataReportal, lebih dari 68% pengguna internet global mengakses web melalui perangkat mobile, terutama android. Fakta ini menunjukkan adanya

kebutuhan akan solusi keamanan siber yang mobile friendly, mampu memberikan perlindungan secara real time, dan sesuai dengan pola penggunaan internet masyarakat indonesia. Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk merancang dan membangun sebuah sistem deteksi url web phishing berbasis aplikasi android dengan memanfaatkan algoritma CatBoost Classifier. Aplikasi ini memungkinkan pengguna untuk memasukkan tautan secara langsung, lalu memperoleh hasil klasifikasi instan berupa tingkat kemungkinan apakah URLweb tersebut aman atau berbahaya. Dataset yang digunakan akan berasal dari sumber terbuka seperti kaggle serta basis data lokal yang relevam. Evaluasi performa sistem akan dilakukan menggunakan metrik akurasi, percision, recall, dan f1-score [8].

Dengan pengembangan sistem ini, diharapkan tidak hanya memberikan solusi teknis untuk mendeteksi phishing secara efisien, tetapi juga meningkatkan kesadaran masyarakat indonesia terhadap pentingnya keamanan siber, terutama dalam menghadapi ancaman URL web phishing yang semakin kompleks dan sulit dikenali secara manual.Seiring dengan kemajuan teknologi, pendekatan berbasis kecerdasan buatan (*Artificial Intelligence* atau *AI*), khususnya *machine learning* (ML), mulai digunakan untuk mendeteksi URL web *Phishing* secara otomatis dan lebih akurat. ML memungkinkan sistem untuk mempelajari pola-pola dari data historis dan mengenali ciri-ciri URL web mencurigakan, seperti panjang URL yang tidak wajar, penggunaan simbol tertentu, hingga struktur domain yang tidak biasa. Salah satu algoritma *machine learning* yang menunjukkan performa unggul dalam tugas klasifikasi *phishing* adalah *CatBoost* (*Categorial Boosting*). Algoritma ini dikembangkan oleh *yandex* dan memiliki keunggulan dalam menangani data

kategorial, mengurangi risiko overfiting, serta menghasilkan model prediktif yang akurat tanpa banyak *CatBoost* dapat mencapai akurasi deteksi hingga 97,3%, mengungguli algoritma populer lainnya seperti *Random Forest*, *XGBoost*, dan *LightGBM* [5].

Meskipun demikian, sebagian besar implementasi sistem deteksi web *phishing* saat ini masih terbatas pada *platform* desktop atau berbasis *web*. Padahal, menurut DataReportal, lebih dari 68% pengguna internet global mengakses web melalui perangkat mobile, terutama android. Fakta ini menunjukkan adanya kebutuhan akan solusi keamanan siber yang *mobile friendly*, mampu memberikan perlindungan secara *real time*, dan sesuai dengan pola penggunaan internet masyarakat indonesia. Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk merancang dan membangun sebuah sistem deteksi url web *phishing* berbasis aplikasi android dengan memanfaatkan *algoritma CatBoost Classifier*. Aplikasi ini memungkinkan pengguna untuk memasukkan tautan secara langsung, lalu memperoleh hasil klasifikasi instan berupa tingkat kemungkinan apakah URLweb tersebut aman atau berbahaya. Dataset yang digunakan akan berasal dari sumber terbuka seperti *kaggle* serta basis data lokal yang relevam. Evaluasi performa sistem akan dilakukan menggunakan metrik akurasi, *percision*, *recall*, dan *f1-score* [6].

Dengan pengembangan sistem ini, diharapkan tidak hanya memberikan solusi teknis untuk mendeteksi *phishing* secara efisien, tetapi juga meningkatkan kesadaran masyarakat indonesia terhadap pentingnya keamanan siber, terutama dalam menghadapi ancaman URL web *phishing* yang semakin kompleks dan sulit dikenali secara manual.

1.2 Tujuan dan Manfaat

1.2.1. Tujuan

Tujuan dari penelitian ini adalah untuk merancang dan mengembangkan sebuah sistem deteksi *URL web phishing* yang mampu mengidentifikasi *URL* berbahaya secara otomatis menggunakan metode pembelajaran mesin, dengan memanfaatkan algoritma *CatBoost Classifier*.

1.2.1. Manfaat

Dalam penelitian yang dilakukan terdapat manfaat penelitian sebagai berikut:

1. Memberikan solusi keamanan yang dapat membantu pengguna internet dalam mengidentifikasi situs *phishing*, sehingga dapat meningkatkan keamanan data pribadi dan informasi sensitif lainnya.
2. Menambah referensi dan wawasan mengenai penerapan algoritma *CatBoost Classifier* untuk deteksi situs *phishing*, khususnya di bidang machine learning dan keamanan siber.
3. Memberikan referensi dan dasar penelitian yang dapat digunakan untuk pengembangan lebih lanjut dalam sistem deteksi *phishing* atau penelitian lain yang terkait dengan keamanan situs web dan analisis *machine learning*.

1.3 Tinjauan Pustaka

Penelitian ini bertujuan untuk mengembangkan aplikasi berbasis *mobile* yang berfungsi untuk mendeteksi situs web *phishing* secara otomatis menggunakan *algoritma CatBoost Classifier*. Aplikasi ini diharapkan mampu memberikan informasi secara cepat dan akurat terkait tingkat probabilitas sebuah *URL web* termasuk dalam kategori phishing atau bukan. Selain itu, aplikasi juga menyediakan fitur riwayat deteksi serta rekomendasi tindakan berdasarkan hasil klasifikasi. Berikut adalah beberapa penelitian terdahulu yang dijadikan sebagai acuan dan pembanding dalam pengembangan sistem deteksi *URL web phishing*. Berikut merupakan beberapa penelitian terdahulu yang relavan dengan konteks penelitian ini :

Penelitian sebelumnya yang merancang sistem deteksi phishing berbasis machine learning yang dapat diakses secara *real-time* melalui web dan Telegram. Sistem ini dikembangkan menggunakan *framework Flask* sebagai *backend* serta mengintegrasikan model *XGBoost* yang telah dioptimasi melalui *hyperparameter* tuning untuk memperoleh performa klasifikasi terbaik. Dataset yang digunakan terdiri dari lebih dari 58.000 entri URL *phishing* dan *non-phishing* dengan lebih dari 100 fitur yang mencakup struktur URL, parameter, hingga konten halaman. Sistem ini juga menerapkan pendekatan *adversarial learning* yang memungkinkan pengguna memberikan pelabelan ulang terhadap URL, sehingga dapat digunakan untuk pelatihan ulang model agar lebih adaptif terhadap pola *phishing* terbaru. Evaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* menunjukkan bahwa *XGBoost* menghasilkan akurasi tertinggi sebesar 96,14%.

Selain melalui antarmuka web, sistem ini terintegrasi dengan *bot* Telegram yang memungkinkan pengguna memverifikasi URL secara praktis melalui aplikasi pesan instan tanpa bergantung pada perangkat tertentu. Meskipun sistem ini telah memberikan hasil yang cepat dan akurat, belum tersedia implementasi dalam bentuk aplikasi mobile berbasis Android, serta belum mencakup fitur-fitur lanjutan seperti riwayat deteksi per pengguna, notifikasi otomatis, atau penyimpanan hasil deteksi secara personal [7].

Penelitian lain juga mengembangkan sistem identifikasi situs *phishing* dengan membandingkan beberapa algoritma klasifikasi seperti *Decision Tree*, *Random Forest*, dan *Support Vector Machine* untuk mendeteksi potensi serangan *phishing* berbasis URL. Sistem ini dibangun dengan tujuan untuk mengurangi risiko serangan siber akibat akses ke situs palsu, dengan memanfaatkan dataset dari UCI *Machine Learning Repository* yang terdiri atas fitur-fitur seperti panjang URL, keberadaan simbol "@", jumlah *subdomain*, serta keabsahan sertifikat SSL. Pengujian dilakukan menggunakan *confusion matrix* dan pengukuran akurasi sebagai indikator utama performa model. Hasil evaluasi menunjukkan bahwa *algoritma Random Forest* memberikan performa terbaik dengan akurasi sebesar 96,07%, diikuti oleh *Decision Tree* dan *SVM* yang juga menunjukkan hasil cukup tinggi. Meskipun belum menggunakan *algoritma boosting* seperti *CatBoost*, penelitian ini memberikan kontribusi penting dalam memahami efektivitas model klasik dalam mendeteksi phishing, dan dapat dijadikan sebagai dasar pembanding untuk pengembangan sistem deteksi phishing berbasis algoritma yang lebih canggih di platform web maupun mobile [8].

Penelitian lain juga bertujuan untuk meningkatkan efektivitas deteksi phishing dengan memanfaatkan *algoritma machine learning* dan teknik *Explainable AI* (XAI). Studi ini membandingkan tiga *algoritma klasifikasi*, yaitu *CatBoost*, *XGBoost*, dan *Explainable Boosting Machine* (EBM), dalam mendeteksi *URL phishing* berdasarkan fitur-fitur penting seperti *length_url*, *time_domain_activation*, dan *Page_rank*. Data yang digunakan diperoleh dari berbagai sumber seperti UCI Machine Learning Repository, Kaggle, dan Mendeley, dengan teknik *Recursive Feature Elimination* (RFE) diterapkan untuk menyeleksi fitur yang paling relevan. Hasil evaluasi menunjukkan bahwa *CatBoost* memiliki akurasi tertinggi dan robust meskipun jumlah fitur dikurangi, sedangkan XGBoost unggul dalam efisiensi waktu eksekusi. Untuk meningkatkan interpretabilitas model, digunakan metode SHAP (*SHapley Additive exPlanations*) yang menjelaskan kontribusi masing-masing fitur terhadap prediksi model. Temuan ini menegaskan bahwa pemilihan fitur yang efektif dan penerapan model interpretatif dapat meningkatkan kinerja sistem deteksi phishing. Meskipun EBM menunjukkan akurasi tinggi, model ini memerlukan waktu komputasi lebih lama dibanding dua algoritma lainnya. Penelitian ini memberikan kontribusi penting dalam pengembangan sistem deteksi phishing berbasis web yang tidak hanya akurat, tetapi juga dapat dijelaskan secara transparan kepada pengguna dan pengembang sistem keamanan siber [9].

Penelitian yang membahas pengembangan sistem deteksi *phishing* berbasis URL dengan menggunakan *algoritma Gradient Boosting* dan *CatBoost*. Sistem ini dirancang untuk menggantikan pendekatan sebelumnya yang hanya menggunakan

Random Forest, dengan tujuan meningkatkan efektivitas dalam mengidentifikasi situs phishing. Dataset yang digunakan bersumber dari *UCI Machine Learning Repository*, dan fitur yang diekstraksi mencakup berbagai indikator dari struktur URL, seperti panjang karakter, simbol mencurigakan, dan informasi domain. Penelitian ini memanfaatkan teknik pembelajaran mesin untuk membandingkan pendekatan *boosting* terhadap model *ensemble konvensional*. Temuan mereka menunjukkan bahwa algoritma boosting dapat memberikan hasil yang lebih stabil dan responsif terhadap pola phishing yang kompleks. Pendekatan ini dinilai relevan untuk diimplementasikan pada sistem otomatis deteksi phishing, baik untuk lingkungan desktop maupun perangkat bergerak, karena mampu memanfaatkan fitur URL tanpa memerlukan konten halaman secara menyeluruh[10].

Penelitian lain dibuat dengan mengkaji deteksi situs phishing menggunakan pendekatan machine learning berbasis URL. Penelitian ini menekankan pentingnya fitur-fitur URL seperti panjang URL, jumlah subdomain, kehadiran karakter tertentu, serta komponen teknis lainnya yang kerap dimanipulasi pada situs phishing. Dengan menggunakan dataset publik dan sejumlah algoritma klasifikasi, tim peneliti mengevaluasi efektivitas pendekatan berbasis URL dalam mengidentifikasi situs palsu. Salah satu keunggulan dari penelitian ini adalah pendekatannya yang mengandalkan *client-side detection*, sehingga tidak memerlukan analisis isi halaman atau struktur HTML secara menyeluruh, yang dapat mempercepat proses klasifikasi dan mempermudah integrasi ke dalam sistem keamanan web atau browser. Penelitian ini memberikan kontribusi dengan menegaskan bahwa deteksi phishing dapat dilakukan secara efisien hanya melalui

analisis karakteristik URL, dan model yang dibangun dapat digunakan untuk memperkuat perlindungan terhadap serangan siber berbasis web tanpa ketergantungan pada daftar hitam tradisional [11].

Tabel 1.1 Gap Penelitian

NO	Penelitian Sebelumnya	GAP PENELITIAN	
		Penelitian Terdahulu	Penelitian Saat Ini
1.	Tahun : 2025 Judul : Deteksi Website Phising Menggunakan Teknik Machine Learning	Algoritma : *XGBoost* Akurasi : 96,14% Teknologi : Flask, dataset URL phising dan non-phising dengan 100 fitur, integrasi bot telegram, dilengkapi dengan adversial learning, belum adanya pengembangan mobile dan fitur riwayat deteksi	Algoritma : Machine Learning Catboost Classifier Teknologi :Python Akurasi : 97,3% Dikembangkan menjadi aplikasi berbentuk mobile
2.	Tahun : 2025 Judul : Penerapan Algoritma Klasifikasi pada Machine Learning untuk Deteksi Phishing	Algoritma : Decision Tree, Random Forest, SVM Akurasi : 96,07% Teknologi : Dataset dari UCI Machine Learning Repository, fitur panjang url simbol @, jumlah subdomain dan sertifikat SSL, evaluasi dengan confusion matrix, belum menggunakan boosting,	Algoritma : Machine Learning Catboost Classifier Teknologi :Python Akurasi : 97,3% Dikembangkan menjadi aplikasi berbentuk mobile

		digunakan sebagai pembanding model, belum adanya pengembangan mobile dan fitur riwayat deteksi	
3.	Tahun : 2024 Judul : Enhancing Phishing Detection through Feature Importance Analysis and Explainable AI: A Comparative Study of CatBoost, XGBoost, and EBM Models	Algoritma : Catboost, XGBoost, Explainable Boosting Machine Akurasi : 95% Teknologi : Dataset dari UCI, kaggle, fitur : length_url, time_domain_activation, page_rank, seleksi fitur dengan RFE, interpretabilitas dengan SHAP, menekankan transparansi predeksi model, belum adanya pengembangan mobile dan fitur riwayat deteksi	Algoritma : Machine Learning Catboost Classifier Teknologi :Python Akurasi : 97,3% Dikembangkan menjadi aplikasi berbentuk mobile
4.	Tahun : 2022 Judul : URL Based Phishing Website Detection by Using Gradient and Catboost Algorithms	Algoritma : Gradient Boosting, CatBoost Akurasi : 91% Teknologi : Dataset dari UCI, fitur url : panjang karakter, simbol mencurigakan, info domain, fokus pada boosting untuk	Algoritma : Machine Learning Catboost Classifier Teknologi :Python Akurasi : 97,3% Dikembangkan menjadi aplikasi berbentuk mobile

		menggantikan random forest, tidak bergantung pada analisis konten halaman, hanya mengembangkan di website desktop	
5.	Tahun : 2024 Judul : Comparative Analysis of Nature-Inspired Metaheuristic Techniques for Optimizing Phishing Website Detection	Algoritma : Genetic Algorithm, Particle Swarm Optimization, Ant Colony Optimization Akurasi : 95% Teknologi : client-side detection. Fitur : panjang, subdomain, simbol. Tidak tergantung daftar hitam, hanya untuk di browser/website	Algoritma : Machine Learning Catboost Classifier Teknologi :Python Akurasi : 97,3% Dikembangkan menjadi aplikasi berbentuk mobile

1.4 Data Penelitian

1. Dataset Kaggle

   Dataset yang diperoleh merupakan sumber dari salah satu website https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector, yang berisi lebih dari 11.000 data URL situs web. Setiap baris data mencakup 30 parameter atau fitur hasil ekstraksi dari URL yang merepresentasikan karakteristik masing-masing situs, serta label kelas yang menunjukkan apakah situs tersebut tergolong phishing atau tidak. Nilai pada

setiap fitur dikodekan dalam bentuk -1, 0, atau 1, di mana angka 1 merepresentasikan kondisi legitimate atau aman sesuai perilaku normal sebuah website, angka 0 menandakan kondisi suspicious atau mencurigakan karena memiliki indikasi tertentu namun belum dapat dipastikan phishing, sedangkan angka -1 mengindikasikan situs dengan karakteristik phishing. Adapun label kelas sebagai target klasifikasi bersifat biner, dengan nilai 1 untuk menandakan situs phishing dan -1 untuk menandakan situs yang legitimate atau aman. Secara keseluruhan, dataset ini terdiri atas 11.054 sampel dengan 32 atribut, yang meliputi 30 fitur hasil ekstraksi URL, 1 atribut indeks, serta 1 label kelas sebagai target. Struktur tersebut memberikan representasi yang cukup komprehensif dalam membedakan antara situs phishing dan situs aman. Dataset tersebut dapat dilihat pada Gambar 1.1.

	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedirec
0	0	1	1	1	1	1	-1	0	1	-1	...	1	
1	1	1	0	1	1	1	-1	-1	-1	-1	...	1	
2	2	1	0	1	1	1	-1	-1	-1	1	...	1	
3	3	1	0	-1	1	1	-1	1	1	-1	...	-1	
4	4	-1	0	-1	1	-1	-1	1	1	-1	...	1	

Gambar 1.1 Dataset website detector

Berikut ini merupakan penjelasan semua fitur yang terdapat didalam dataset gambar diatas.

Tabel 1.2 Penjelasan Fitur

NO	Nama	Keterangan
1.	Index	Nomor urut atau identitas unik untuk setiap baris data dalam dataset. Fitur ini tidak digunakan untuk analisis model, melainkan hanya sebagai penanda.
2.	UsingIP	Menunjukkan apakah alamat situs web menggunakan alamat IP (contoh: http://192.168.0.1) alih-alih nama domain. Phishing sering menggunakan IP untuk menghindari deteksi nama domain resmi.
3.	ShortURL	Menunjukkan apakah URL menggunakan layanan pemendek tautan seperti bit.ly, goo.gl, atau tinyurl.com. Phisher sering menyamarkan URL asli agar korban tidak curiga.
4.	Symbol@	Memeriksa apakah terdapat simbol @ dalam URL. Simbol ini sering digunakan untuk mengarahkan browser mengabaikan bagian awal URL dan mengarahkan ke domain palsu.
5.	Redirecting//	Menghitung jumlah tanda // dalam URL setelah protokol (http:// atau https://). Banyak tanda // di tengah URL dapat menjadi indikasi pengalihan mencurigakan.
6.	PrefixSuffix-	Menunjukkan adanya tanda hubung - di domain utama (contoh: secure-bank-login.com). Banyak situs phishing menggunakan tanda hubung untuk meniru domain resmi.

7.	SubDomains	Menghitung jumlah subdomain dalam URL. URL dengan subdomain berlebihan sering dipakai untuk menipu pengguna (contoh: secure.login.bank.com.fake-site.org).
8.	HTTPS	Menunjukkan apakah situs menggunakan protokol HTTPS. Meskipun bukan jaminan aman, ketiadaan HTTPS pada situs login sangat mencurigakan.
9.	DomainRegLen	Lama masa registrasi domain (dalam bulan atau tahun). Domain phishing biasanya berumur pendek.
10.	Favicon	Memeriksa apakah favicon (ikon kecil di tab browser) berasal dari domain yang sama. Favicon yang diambil dari domain lain bisa menjadi tanda situs palsu.
11.	NonStdPort	Menunjukkan apakah situs menggunakan port non-standar selain 80 (HTTP) atau 443 (HTTPS). Port yang tidak umum bisa mengindikasikan aktivitas mencurigakan.
12.	HTTPSDomainURL	Memeriksa apakah kata "https" muncul di bagian domain URL (contoh: https-secure-login.com). Teknik ini sering dipakai untuk menipu.
13.	RequestURL	Mengukur proporsi elemen eksternal (gambar, skrip, dll.) yang diambil dari domain lain. Situs phishing cenderung mengambil banyak elemen dari luar.
14.	AnchorURL	Mengukur persentase tautan (anchor tags <a>) yang mengarah ke domain lain.

15.	LinksInScriptTags	Menghitung jumlah link eksternal yang ada di tag <script> atau <meta>. Banyaknya link eksternal bisa jadi tanda phishing.
16.	ServerFormHandler	Menunjukkan apakah formulir (<form>) mengirim data ke server lain atau ke email langsung. Situs phishing sering menggunakan form handler eksternal.
17.	InfoEmail	Memeriksa apakah ada alamat email dalam URL atau kode halaman. Ini mencurigakan karena situs resmi jarang mengirim data login lewat email.
18.	AbnormalURL	Menunjukkan apakah URL sesuai dengan domain asli yang terdaftar. URL abnormal sering menandakan situs tiruan.
19.	WebsiteForwarding	Mengukur jumlah pengalihan otomatis sebelum halaman termuat. Situs phishing sering menggunakan banyak forwarding untuk menghindari deteksi.
20.	StatusBarCust	Memeriksa apakah skrip memanipulasi teks di status bar browser. Teknik ini digunakan untuk menipu pengguna tentang tujuan link.
21.	DisableRightClick	Menunjukkan apakah klik kanan dinonaktifkan untuk mencegah pengguna memeriksa elemen atau menyalin teks.
22.	UsingPopupWindow	Memeriksa apakah situs menggunakan banyak jendela pop-up. Pop-up sering digunakan phishing untuk menampilkan form login palsu.
23.	IframeRedirection	Menunjukkan apakah situs menggunakan iframe tersembunyi untuk menampilkan konten dari domain lain.

24.	AgeofDomain	Usia domain dalam bulan atau tahun. Domain phishing umumnya baru terdaftar.
25.	DNSRecording	Menunjukkan apakah domain memiliki catatan DNS yang valid. Domain phishing sering tidak memiliki rekam DNS lengkap.
26.	WebsiteTraffic	Mengukur jumlah kunjungan situs berdasarkan data seperti Alexa Rank. Situs phishing cenderung memiliki trafik rendah.
27.	PageRank	Nilai peringkat halaman berdasarkan jumlah dan kualitas backlink. Situs phishing umumnya memiliki PageRank rendah.
28.	GoogleIndex	Memeriksa apakah situs terindeks oleh Google. Situs phishing biasanya tidak terindeks.
29.	LinksPointingToPage	Menghitung jumlah tautan eksternal yang mengarah ke halaman tersebut. Nilai rendah bisa menunjukkan situs baru atau mencurigakan.
30.	StatsReport	Mengacu pada laporan statistik atau reputasi domain dari layanan pihak ketiga seperti WHOIS, PhishTank, atau VirusTotal.
31.	Class	Label target klasifikasi. • 1 = Phishing • -1 = Legit

2. Dataset *Whitelist*

Data ini dikumpulkan secara manual berasal dari sumber terpercaya seperti daftar domain pemerintah (go.id) dari Kemkominfo, domain perbankan resmi dari OJK, e-commerce maupun institusi populer. Dataset whitelist yang digunakan dalam penelitian ini berisi daftar domain resmi

yang telah diverifikasi sebagai *Default Trusted Domain*, terutama dari sektor perbankan di Indonesia. Dataset ini terdiri dari dua kolom, yaitu domain yang berisi alamat situs, dan description yang menjelaskan status keamanan domain tersebut. Jumlah dataset yang digunakan sebanyak 1003 domain. Tujuan utamanya adalah memastikan bahwa situs-situs terpercaya ini dikenali oleh sistem, sehingga dapat mengurangi *false positive*, mempercepat proses validasi URL, dan menjaga akurasi deteksi secara keseluruhan.

1	Column1	Column2
2	domain	description
3	bi.go.id	Default trusted domain
4	bni.co.id	Default trusted domain
5	bri.co.id	Default trusted domain
6	btn.co.id	Default trusted domain
7	bankmandiri.co.id	Default trusted domain
8	bankraya.co.id	Default trusted domain
9	arthagraha.com	Default trusted domain
10	ir.bankbsi.co.id	Default trusted domain
11	kbbank.co.id	Default trusted domain

Gambar 1. 2 Dataset Whitelist