Skema Pendanaan: UTAMA

**LAPORAN PENELITIAN**

# MODEL KLASIFIKASI DETEKSI SITUS PHISHING MENGGUNAKAN METODE MACHINE LEARNING XGBOOST, RANDOM FOREST DAN FEATURE SELECTION


**TIM PENELITI**

Ketua : Dr. Ir. Mardi Hardjianto, M.Kom. (920024)
Anggota : Agnes Aryasanti, M.Kom. (120059)
Putri Hayati, S.ST., M.Kom. (170067)

**FAKULTAS TEKNOLOGI INFORMASI**
**UNIVERSITAS BUDI LUHUR**

**Agustus 2025**

**HALAMAN PENGESAHAN**
**LAPORAN PENELITIAN**

**Judul Penelitian** : Model Klasifikasi Deteksi Situs Phishing Menggunakan Metode Machine Learning XGBOOST, Random Forest dan Feature Selection

**Bidang Penelitian** : Cyber Security

**Ketua Peneliti**

a. Nama Lengkap : Dr. Ir. Mardi Hardjianto, M.Kom.
b. NIP/NIDN/ID-SINTA : 920024/0422036901/5977626
c. Jabatan Fungsional : Lektor
d. Program Studi : S2-Pasca Sarjana Magister Ilmu Komputer
e. Nomor HP : 081586603011
f. Alamat e-mail : mardi.hardjianto@budiluhur.ac.id

**Anggota Peneliti (1)** Model Klasifikasi Deteksi Situs Phishing Menggunakan

a. Nama Lengkap : Agnes Aryasanti, M.Kom.
b. NIP/NIDN/ID-SINTA : 120059/0322018502/6043631

**Anggota Peneliti (2)**

a. Nama Lengkap : Putri Hayati, S.ST., M.Kom.
b. NIP/NIDN/ID-SINTA : 170067/0308029102/6145520

**Mahasiswa**

a. Nama Lengkap : Samuel Prasetyo
b. NIM : 2111501157

**Lama Penelitian** : 6 bulan

**Biaya Penelitian**

a. Sumber Universitas Budi Luhur : Rp 14.000.000,-
b. Sumber lain (sebutkan jika ada) : Rp -

Jakarta, 31 Agustus 2025

Mengetahui,
Dekan Fakultas Teknologi Informasi

(Dr. Ir. Achmad Solichin, M.T.I.)
NIP 050023

Ketua Pelaksana

(Dr. Ir. Mardi Hardjianto, M.Kom.)
NIP 920024

Menyetujui,
Direktur Riset dan Pengabdian Kepada Masyarakat

(Dr. Ir. Prudensius Maring, M.A.)
NIP 190043

# RINGKASAN

Kemajuan teknologi informasi sangat mempengaruhi berbagai aspek kehidupan. Kemudahan dan kenyamanan banyak didapatkan dalam kemajuan teknologi ini, namun juga menghadirkan tantangan seperti ancaman kejahatan siber. Salah satu ancaman adalah phising. Phishing adalah salah satu bentuk kejahatan siber yang dapat merugikan orang lain dan termasuk tindakan yang melanggar hukum. Teknik penipuan yang dilakukan adalah memanfaatkan sebuah situs palsu untuk memperoleh informasi sensitif, seperti data pribadi, username, password, nomor kartu kredit, NIK, dll. Data yang diperoleh dapat dimanfaatkan untuk memeras, menipu atau bahkan dijual ke pihak lain untuk disalahgunakan. Penipuan menggunakan phishing sudah ada sejak tahun 1996. Di Indonesia, pada kuartal I tahun 2023 dilaporkan ada sebanyak 26.675 kasus phishing, lebih banyak dibandingkan pada kuartal empat 2022 yang hanya 6.106 kasus. Serangan phishing ini terus berkembang dengan teknik yang semakin canggih, seperti penggunakan nama domain yang mirip dan social engineering, sehingga deteksi ancaman phishing yang akurat menjadi kebutuhan yang mendesak. Berbagai penelitian telah dilakukan untuk mencari solusi dalam menangani serangan phishing. Penelitian ini bertujuan mengembangkan sistem deteksi phishing melalui fitur-fitur URL menggunakan metode machine learning XGBoost, Random Forest dan Feature Selection. Penggunaan feature selection untuk mengidentifikasi fitur-fitur yang paling berpengaruh dalam klasifikasi, serta meningkatkan efisiensi dan akurasi model. Evaluasi model XGBoots dan Random Forest, dengan optimasi seleksi fitur menggunakan Feature Importance dalam memprediksi situs web phishing atau non-phishing. Hasil dari model Random Forest menggunakan 30 fitur menghasilkan kinerja terbaik, dengan nilai akurasi 97,4%, sensitivitas 96,6% dan spesifisitas 98,4%.

# PRAKATA

Puji dan syukur kami haturkan ke hadirat Tuhan Yang Maha Esa yang telah mencurahkan semua karunia sehingga diberikan keyakinan dan kemudahan dalam menyusun dan menyiapkan Laporan Penelitian ini dengan judul ”Model Klasifikasi Deteksi Situs Phishing Menggunakan Metode Machine Learning XGBOOST, RANDOM FOREST dan FEATURE SELECTION”.

Pada kesempatan ini, peneliti ingin menyampaikan rasa terima kasih yang tak terhingga kepada:
1. Bapak Prof. Dr. Agus Setyo Budi, M.Sc. selaku Rektor Universitas Budi Luhur.
2. Bapak Dr. Ir. Achmad Solichin, S.Kom., M.T.I. selaku Dekan Fakultas Teknologi Informasi Universitas Budi Luhur.
3. Prof. Dr. Ir. Prudensius Maring, M.A. selaku Direktur Penelitian dan Pengabdian Kepada Masyarakat Universitas Budi Luhur.
4. Untuk semua pihak yang telah membantu penelitian ini, mohon maaf apabila ada kesalahan yang terucap dan terbersit, mohon dibukakan pintu maaf yang selapang-lapangnya.

Akhirnya dengan segala kerendahan hati, kami berharap agar penelitian ini dapat memberikan manfaat bagi peneliti, lingkungan kampus dan sekitarnya. Saran dan kritik yang membangun sangat diharapkan guna perbaikan penelitian mendatang.

Jakarta, Agustus 2025
Penulis

# DAFTAR ISI

# DAFTAR TABEL

# DAFTAR GAMBAR

# DAFTAR LAMPIRAN

# BAB I
# PENDAHULUAN

## 1. Latar Belakang

Penggunaan internet di Indonesia makin meluas pasca pandemi COVID-19 melanda, yang akhirnya membatasi pergerakan masyarakat secara langsung. Perubahan kebiasaan ini menjadikan aktivitas fisik menjadi terbatas sehingga harus beralih ke aktivitas daring melalui platform-platform digital. Kemajuan teknologi dan layanan yang memanfaatkan internet, telah membuat perubahan dalam cara manusia, berbisnis, bekerja, belajar, berbelanja dan berinteraksi sosial dalam kesehariannya. Berbagai pekerjaan dapat diselesaikan dengan mudah dan efisien akibat dampak dari perkembangan teknologi(Makbull Rizki, 2022). Banyaknya aktivitas ini menjadikan internet berfungsi sebagai media yang memudahkan dan menyediakan fasilitas untuk mendukung kegiatan tersebut seperti website. Banyaknya website baru yang bermunculan tumbuh sebagai media kolaborasi antar pengguna internet. Awalnya website hanya berfungsi sebagai alat penyampaian informasi satu arah. Namun, dengan kemajuan teknologi informasi dan komunikasi, website kini telah berkembang menjadi platform interaktif, media komunikasi, dan sarana untuk melakukan berbagai transaksi online.

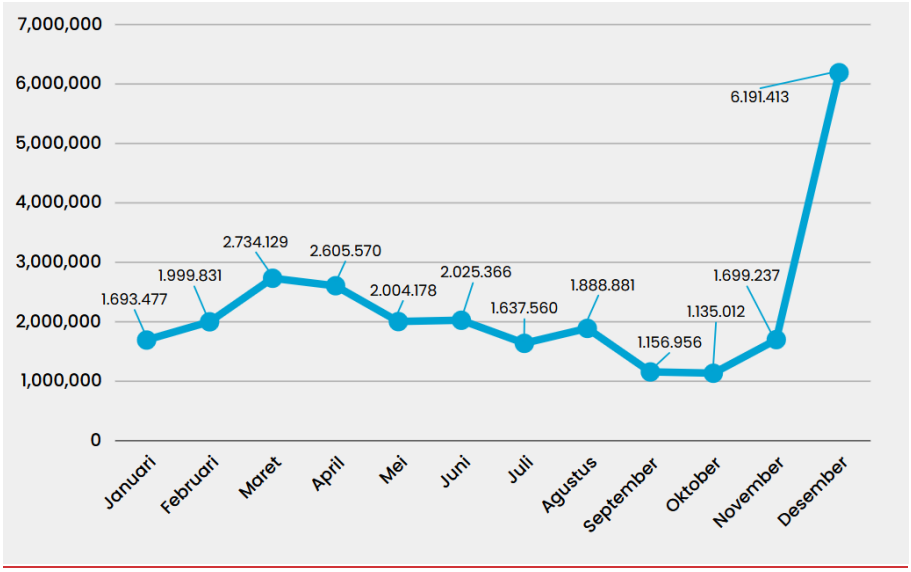
Kemajuan teknologi dan internet selain memberikan manfaat positif juga berdampak negatif seperti kejahatan siber yang memanfaatkan kelemahan sistem dan kurangnya kesadaran dari pengguna(Saputro, Sugiarto and Nugraha, 2024). Kondisi ini sering dimanfaatkan oleh pelaku kejahatan dunia maya untuk mencuri informasi sensitif, seperti data pribadi, kata sandi email, hingga informasi finansial termasuk data kartu kredit dan akun perbankan online, tanpa diketahui oleh korbannya. Untuk melancarkan aksinya, pelaku kejahatan siber seringkali menggunakan situs phishing sebagai alat utama.

Situs phishing dirancang khusus agar menyerupai situs asli, baik dari tampilan, konten, maupun URL domain, untuk menipu pengguna internet. Tujuannya adalah membuat korban percaya bahwa mereka sedang mengakses situs resmi yang sah. Desain situs phishing ini dibuat semirip mungkin dengan situs asli, sehingga sulit dibedakan oleh mata awam. Selain itu, beberapa situs phishing juga dirancang untuk menyajikan informasi atau petunjuk palsu yang menyesatkan. Ketika korban tertipu dan memasukkan informasi yang diminta, pelaku kejahatan dapat memanfaatkan data tersebut pada situs asli untuk melakukan berbagai tindakan yang tidak diinginkan, yang berujung pada kerugian besar bagi korban, mulai dari kerugian finansial hingga kehilangan data. Metode ini biasanya dilakukan melalui e-mail, pesan teks, atau situs web palsu yang berisi lampiran atau tautan berbahaya(Muftiadi, Agustina and Evi, 2022). Tujuan utama dari pelaku phishing adalah untuk melakukan penipuan, guna mendapatkan data sensitif seperti username dan password, nomor PIN, data diri pribadi dari korban, atau sebagai alat penyebaran malware seperti ransomware(Putri and Wijayanto, 2022; BSSN, 2024).

PhishLabs dalam laporan tren ancaman dan inteligen pada kuartal kedua 2023, serangan phishing yang menargetkan platform media sosial meningkat lebih dari 23,4%. Hal ini merupakan jumlah serangan terbesar di media sosial dalam dua tahun dan menempatkan diurutan kedua setelah sektor keuangan(Ryan, 2023). Penggunakan domain palsu dalam penipuan phishing juga dilaporkan terbanyak menggunakan Country Code Top Level Domain (ccTLD) 45,91% diikuti dengan Generic Top Level Domain (gTLD) 41,73%. ccTLD yang banyak digunakan adalah negara Polandia, Kolombia dan Indonesia(Ryan, 2023).

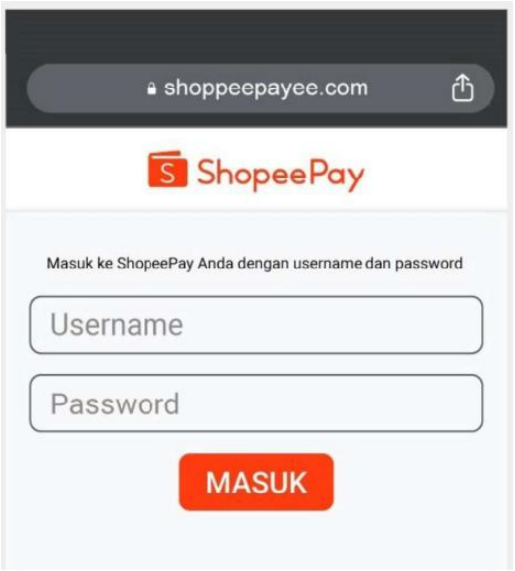
Berdasarkan laporan tahun 2024 ID-SIRTII/CC, mengenai statistik monitoring anomali trafik, serangan phishing menempati urutan petama dari anomali trafik, atau sekitar 26.771.610 aktivitas. Menurut Kaspersky Security Awareness Platform, dari seluruh serangan siber yang terjadi, 91 persen dimulai dengan email phishing dan dari keseluruhan kebocoran data 32 persen disebabkan oleh penggunaan teknik phishing yang tepat(Kusumawardhani, 2022).

Phishing dapat menjadi pintu masuk bagi malware atau akses tidak sah lainnya ke sistem yang berpotensi mengakibatkan serangan lanjutan. **Error! Reference source not found.** merupakan grafik aktivitas phishing tahun 2024.




Gambar 1 Aktivitas Phishing Tahun 2024

Situs jual beli online dan layanan perbankan menjadi target utama serangan phishing, karena potensi keuntungan yang tinggi jika pelaku berhasil. Situs populer seperti Shopee dan PayPal sering menjadi sasaran. Selain itu, tidak sedikit juga situs media sosial seperti Facebook, X, Tiktok dan Instagram yang menjadi target para pelaku kejahatan dunia maya. Selain digunakan untuk mencuri data, situs phishing juga dimanfaatkan untuk melakukan berbagai bentuk penipuan yang mengatasnamakan situs resmi serta sebagai sarana penyebaran malware atau virus komputer oleh pelaku kejahatan siber. Berikut ini disajikan perbandingan antara halaman situs phishing dan situs asli yang menyerupai layanan keuangan daring ShopeePay. Dalam contoh ini, situs phising memiliki alamat URL: http://shoppeepayee.com/ seperti yang ditunjukan pada Gambar 2.




Gambar 2 Contoh Situs Phishing

Maraknya serangan phishing akhir-akhir ini membuat banyak orang tertipu. Situs phishing sulit dibedakan dari situs asli karena kemiripan konten, domain, atau struktur HTML. Deteksi manual tidak efisien dan rentan terhadap kesalahan, sehingga diperlukan suatu model pendeteksian yang akurat apakah sebuah situs merupakan phishing atau tidak. Metode pengklasifikasian dibutuhkan untuk mendeteksi situs dan link phishing agar memperoleh hasil yang efektif.

## 2. Pendekatan Pemecahan Masalah

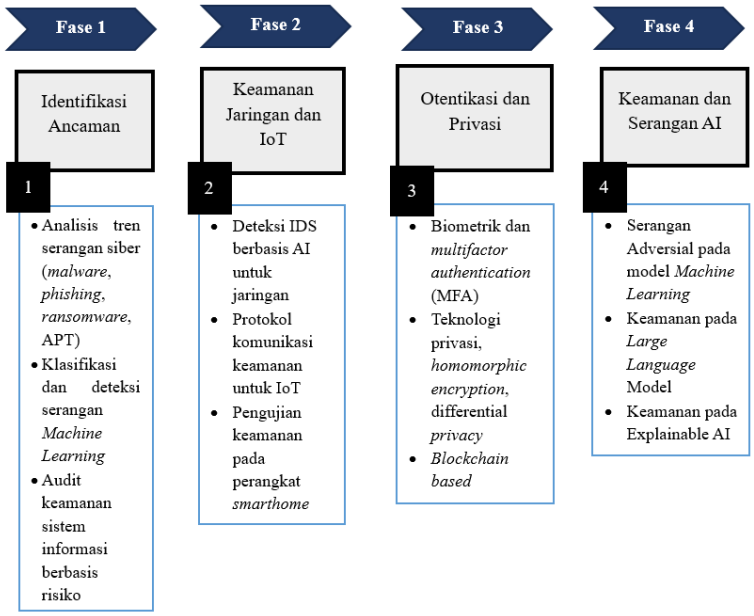
Penyelesaian dari masalah adalah dibuatnya sebuah model untuk mendeteksi apakah suatu situs phishing atau tidak. Pendekatan yang dilakukan adalah berbasis machine learning karena kemampuannya mengenali pola kompleks dalam data fitur situs. Pada penelitian ini, kami mencoba melakukan pengujian dan analisis model XGBoost dengan Random Forest dalam mendeteksi URL phishing dan non phishing. Dalam pengujian dan analisis, model XGBoost dan Random Forest dioptimasi dengan feature selection Recursive Feature Elimination(RFE).

## 3. State Of The Art dan Kebaruan

Deteksi situs phishing atau bukan sudah banyak dilakukan peneliti sebelumnya. Banyak metode yang dapat digunakan untuk mengklasifikasi situs phishing atau bukan. Metode yang digunakan adalah eXtreme Gradient Boost (XGBoost), Gradient Boost, Probabilistic Neural Networks (PNN), Random Forest, Cat Boost, Naïve Bayes, K-Nearest Neighbor, dan Support Vector Machine. Pada penelitian terdahulu, para peneliti melakukan perbandingan antar metode, seperti yang dilakukan Musa, dkk (2019) membandingkan metode XGBoost, PNN, Random Forest, Naïve Bayes dan K-NN menggunakan 30 fitur. Deekshitha (2022) membandingkan metode Gradient Boost, Cat Boost dan Random Forest, sedangkan Mahmud and Wirawan (2024) membandingkan metode Decision Tree, Random Forest dan KNN menggunakan 24 fitur. Ada juga yang membandingkan metode Naïve Bayes, Decision Tree dan Random Forest menggunakan jumlah fitur 30 seperti yang dilakukan (Windarni *dkk.*, 2023). Putri and Wijayanto (2022) membandingkan metode Naïve Bayes, Random Forest, Decision Tree dan Support Vector Machine menggunakan jumlah fitur 9. Pada penelitian terdahulu, sayangnya jumlah fitur yang digunakan bersifat statik. Pada penelitian yang kami lakukan, jumlah fitur yang digunakan akan dicoba dengan jumlah yang bervariasi untuk mendapatkan hasil yang optimal.

## 4. Peta Jalan Penelitian

Peta jalan penelitian lima tahun ke depan ditunjukan pada **Error! Reference source not found.**. Penelitian dibagi menjadi empat fase. Pada fase satu dilakukan analisi tren serangan seperti *malware*, *phishing*, *ransomware*, dan APT. Menggunakan *Machine Learning* untuk klasifikasi dan deteksi serangan serta audit keamanan sistem berbasis risiko. Selanjutnya pada fase dua, akan dilakukan pengamanan jaringan serta perangkat *Internet of Things*. Diantaranya membuat Intrusion Detection System (IDS) berbasis AI, protokol komunikasi aman untuk IoT, pengujian keamanan perangkat seperti *smart home*/*smart grid*. Fase tiga berfokus perlindungan data pribadi dan sistem otentikasi, meliputi Biometrik dan *multi-factor authentication* (MFA), teknologi privasi seperti *homomorphic encryption* dan *differential privacy*, dan identitas digital berbasis *blockchain*. Fase keempat fokus pada keamanan AI dan serangan terhadap AI. ancaman dan perlindungan terhadap sistem berbasis kecerdasan buatan yaitu serangan adversarial terhadap model ML, keamanan sistem berbasis LLM dan rekomendasi, dan Pengembangan AI yang terpercaya dan dapat dijelaskan (XAI).




Gambar 3 Peta Jalan Penelitian

# BAB II
# METODE

Ada beberapa mekanisme yang digunakan untuk mendeteksi atau mengidentifikasi website asli atau phishing. Beberapa teknik yang digunakan untuk mendeteksi situs phishing telah dikelompokkan menjadi empat kategori utama, yaitu List Based Methods, Search Engine Based Methods, Visual Similarity Based Methods dan Machine Learning Based Methods (Jain *et al.*, 2020; Ozker and Sahingoz, 2020; Baduwal *et al.*, 2023).

Pada penelitian ini digunakan metode Machine Learning Based Methods. Deteksi phishing menggunakan Machine Learning Based Methods melibatkan pemanfaatan teknik-teknik pembelajaran mesin untuk mengidentifikasi pola-pola karakteristik dari situs web phishing. Pendekatan ini melibatkan pelatihan model pembelajaran mesin dengan menggunakan kumpulan data yang mencakup fitur dari URL phishing dan non-phishing. Model ini belajar mengenali fitur-fitur yang membedakan antara kedua jenis situs, seperti tata letak halaman, kata kunci, elemen visual, dan pola perilaku. Setelah pelatihan, model dapat diterapkan untuk menganalisis situs web baru dan mengklasifikasikannya sebagai phishing atau non-phishing website berdasarkan pola yang telah dipelajari (Alam *et al.*, 2020; Ozker and Sahingoz, 2020; Zamir *et al.*, 2020)

Dalam mengklasifikasi sebuah situs phishing atau tidak phishing, digunakan Uniform Resource Locator(URL). Dalam konteks analisis URL menggunakan model pembelajaran mesin, URL phishing dan non-phishing diekstrak dalam fitur-fitur untuk membentuk dataset yang digunakan dalam pelatihan dan pengujian. Fitur yang digunakan dalam penelitian dini berjumlah 30. Fitur-fitur yang digunakan dan bagaimana cara memberikan nilai pada fitur ditunjukkan pada Tabel 1.

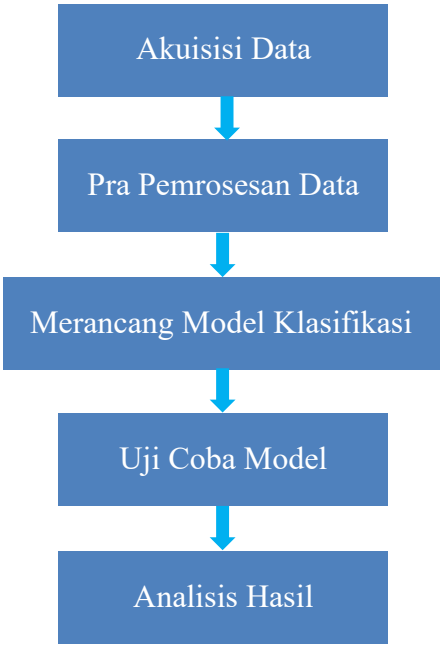
Tabel 1 Fitur Yang Digunkan dan Cara Memberikan Nilai Pada Fitur

| No. | Fitur | Nilai |
|---|---|---|
| 1. | having_IP_Address | -1 = URL menggunakan IP Addr<br>1 = Tidak menggunakan |
| 2. | URL_Length | -1 = Panjang URL > 75<br>0 = 30-75<br>1 = < 30 |
| 3. | Shortining_Service | -1 = Menggunakan ShortURL<br>1 = tidak menggunakan ShortURL |
| 4. | having_At_Symbol | -1 = URL mengandung @<br>1 = URL tidak mengandung @ |
| 5. | Double slash redirecting | -1 = mengandung //<br>1 = tidak mengandung // |
| 6. | Prefix Suffix | -1 = URL terdapat karakter -<br>1 = tidak terdapat karakter - |
| 7. | Having Sub Domain | -1 = banyak subdomain (>= 3)<br>0 = subdomain sedang (1-2)<br>1 = subdomain sedikit (0-1) |
| 8. | SSLfinal State | -1 = Situs tidak menggunakan SSL atau sertifikat tidak valid<br>0 = Situs menggunakan SSL tapi dengan sertifikat yang kurang terpercaya atau ada masalah (misal sertifikat self-signed)<br>1 = Situs menggunakan SSL dengan sertifikat valid dan terpercaya |
| 9. | Domain registeration length | -1 = Masa aktif domain pendek (< 1 tahun)<br>1 = Masa aktif domain panjang (≥ 1 tahun) |
| 10. | Favicon | -1= Favicon berasal dari domain berbeda atau favicon tidak ditemukan<br>1 = Favicon berasal dari domain yang sama |
| 11. | Port | -1 = menggunakan port non-standard<br>1 = menggunakan port standard/ default (80/443) |
| 12. | HTTPS token | -1 = Kata https muncul di bagian domain URL (bukan protokol)<br>1 = Tidak ada https di domain, hanya di protokol |
| 13. | Request URL | Apakah resource (gambar, skrip, iframe, dll) berasal dari domain lain? |

| | | |
|---|---|---|
| | | -1 = Persentase tinggi (> 61%)<br>0 = Persentase sedang (22%–61%)<br>1 = Persentase rendah (≤ 22%) |
| 14. | URL of Anchor | Apakah anchor tag (<a href="..">mengarah ke domain luar?<br><br>-1 = Proporsi tinggi (> 67%)<br>0 = Proporsi sedang (31% – 67%)<br>1 = Proporsi rendah (≤ 31%) |
| 15. | Links in tags | <meta>, <script>, dan <link> yang berasal dari domain di luar domain utama.<br><br>-1 = Proporsi tinggi (> 81%)<br>0 = Proporsi sedang (17% – 81%)<br>1 = Proporsi rendah (≤ 17%) |
| 16. | SFH | -1 = Empty "about:blank" domain berbeda<br>0 = Domain berbeda tapi tidak kosong<br>1 = Action ke domain sendiri |
| 17. | Submitting to email | Apakah form dikirim via email (mailto:)?<br><br>-1 = Formulir dikirim ke alamat email<br>1 = Formulir dikirim ke server (bukan email) |
| 18. | Abnormal URL | -1 = Domain utama tidak ditemukan di dalam konten halaman (form, link, gambar, dll)<br>1 = Domain utama terdapat dalam konten halaman |
| 19. | Redirect | -1 = Redirect lebih dari 4 kali<br>0 = Redirect 1 sampai 4 kali<br>1 = Tidak ada redirect |
| 20. | on_mouseover | -1 = Ada penggunaan onmouseover yang mencurigakan<br>1 = Tidak ditemukan penggunaan mencurigakan |
| 21. | RightClick | -1 = Klik kanan diblokir<br>1 = Klik kanan diizinkan |
| 22. | popUpWidnow | -1 = Ada penggunaan popup<br>1 = Tidak ada popup |
| 23. | Iframe | -1 = Ada penggunaan <iframe> tak terlihat<br>1 = Tidak ada <iframe> |
| 24. | age_of_domain | -1 = umur Domain <= 6 bulan<br>1 = umur Domain > 6 bulan |
| 25. | DNSRecord | -1 = Tidak ada record DNS aktif atau DNS tidak valid<br>1 = DNS record valid dan ada |
| 26. | web_traffic | -1 = Lalu lintas sangat rendah atau tidak tersedia (indikasi phishing)<br>0 = Lalu lintas sedang atau tidak dapat dipastikan<br>1 = Lalu lintas tinggi (indikasi situs legit) |
| 27. | Page_Rank | -1 = PageRank rendah atau tidak tersedia → indikasi phishing<br>1 = PageRank tinggi → indikasi situs legit (bukan phishing) |
| 28. | Google_Index | -1 = Situs tidak terindeks oleh Google → indikasi phishing<br>1 = Situs terindeks oleh Google → indikasi situs legit |
| 29. | Links_pointing_to page | -1 = Tidak ada tautan eksternal yang mengarah ke halaman → indikasi phishing<br>0 = Sedikit tautan eksternal (jumlah minim, mungkin 1-2) → netral/ambigu<br>1 = Banyak tautan eksternal → indikasi situs legit |
| 30. | Statistical_report | -1 = Tidak ada laporan statistik atau laporan tersebut negatif → indikasi phishing<br>1 = Ada laporan statistik atau laporan valid → indikasi situs legit |

## 1. Langkah Penelitian

Langkah-langkah penelitian yang digunakan untuk menggambarkan urutan proses kegiatan secara umum di tunjukkan pada Gambar 4.




Gambar 4 Langkah-Langkah Penelitian

## 2. Akuisisi Data

Dalam studi ini, peneliti memanfaatkan *dataset* publik yang bersumber dari repositori GitHub *Phishing Database*, yang tersedia di tautan https://github.com/Phishing-Database/Phishing.Database. *Repository* ini disponsori oleh PyFunceble, sebuah alat yang digunakan untuk memeriksa apakah suatu situs valid, aktif, atau tidak aktif. Pengambilan data dilakukan dari *repository* publik Phishing. *Database* karena beberapa alasan berikut:

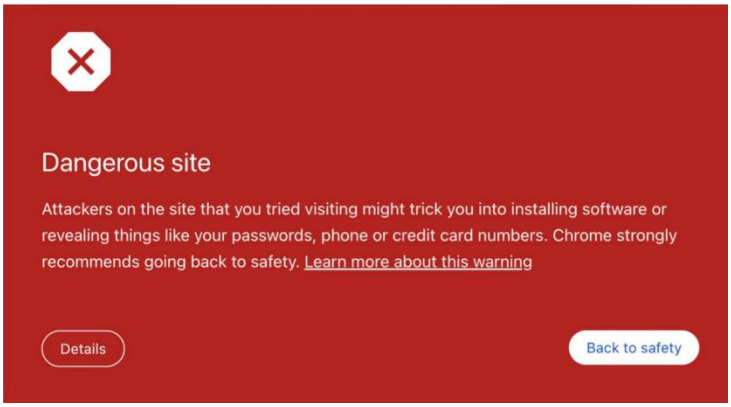
a. Pembaruan rutin
   *Repository* ini diperbarui setiap hari, bahkan beberapa *dataset* diperbarui setiap beberapa jam, sehingga data yang diambil selalu merupakan data terbaru.
b. Validitas situs *phishing*
   Situs-situs yang tercantum dalam *repository* benar-benar merupakan situs *phishing*. Hal ini dibuktikan ketika penulis mencoba membuka salah satu situs dan langsung diblokir oleh browser karena terdeteksi sebagai website berbahaya.
c. Kemudahan akses data
   Karena sifatnya sebagai *repository* publik, penulis dapat dengan mudah mengunduh *dataset* yang tersedia tanpa perlu melakukan registrasi atau membayar.
d. Kelengkapan *dataset*
   *Repository* ini menyediakan berbagai jenis *dataset*, mulai dari link, domain, hingga IP address. Hal ini memudahkan penulis dalam melakukan kurasi dan analisis data sesuai kebutuhan.

Penggunaan *dataset* dari repositori tersebut memberikan sejumlah keunggulan dibandingkan dengan pengumpulan data secara mandiri, khususnya dalam konteks penelitian mengenai URL *phishing* dan non-*phishing*.
Beberapa alasan yang mendasari pemanfaatan *dataset* ini antara lain:

a. Validitas dan reliabilitas data: *Dataset* pada GitHub *Phishing Database* umumnya telah melalui proses verifikasi serta pemutakhiran oleh kontributor dari komunitas keamanan siber, yang meningkatkan akurasi dan integritas data.
b. Keberagaman sumber data: *Dataset* dikompilasi dari berbagai laporan *phishing* aktual, yang mencerminkan variasi dan kompleksitas serangan dunia nyata—sebuah aspek penting dalam membangun model yang mampu melakukan generalisasi.
c. Kolaborasi komunitas terbuka: GitHub sebagai platform terbuka memungkinkan kontribusi dan kolaborasi dari berbagai peneliti serta praktisi keamanan siber, yang memperkaya kualitas dan keberlanjutan data.
d. Kemudahan replikasi dan validasi: Penggunaan *dataset* yang telah dipakai dalam berbagai studi terdahulu mempermudah proses validasi hasil serta perbandingan performa antar pendekatan analitis.
e. Efisiensi waktu dan sumber daya: Menggunakan *dataset* yang telah tersedia memungkinkan peneliti untuk fokus pada pengembangan dan evaluasi model tanpa harus melewati proses pengumpulan dan pembersihan data yang memakan waktu.

*Dataset* yang digunakan dalam penelitian ini terdiri dari dua kategori utama, yaitu phishing (label: -1) netral/ambigu (label: 0) dan non-*phishing* (label: 1). Data ini mencakup berbagai fitur yang berkaitan dengan karakteristik URL, alamat IP, serta atribut teknis lainnya yang relevan dalam pelatihan model pembelajaran mesin untuk klasifikasi situs web *phishing*. Sebuah web *phishing* bila diakses menggunakan *browser*, maka akan muncul tampilan seperti yang ditunjukan pada Gambar 5. Tujuan dari pemanfaatan *dataset* ini adalah untuk mendalami pola dan karakteristik umum dari situs *phishing*, serta mengembangkan pendekatan deteksi yang lebih efektif. Dengan memahami fitur-fitur utama yang menjadi indikator *phishing*, penelitian ini berupaya berkontribusi pada penguatan sistem deteksi dan pencegahan serangan phishing sebagai bagian dari upaya peningkatan keamanan siber secara keseluruhan.

Di *repository Phishing*. *Database*, terdapat beberapa pilihan *dataset* mulai dari IP, domain, dan link. Granularitas data juga beragam, ada yang di-*update* setiap jam, hari, atau gabungan dari awal. Selain itu, *dataset* juga telah diklasifikan menjadi aktif, tidak aktif, dan tidak valid. Hal ini sangat memudahkan penulis dalam proses pengambilan data. Tabel 2 merupakan contoh situs *phishing* yang aktif dan pada Tabel *3* ditampilkan contoh situs bukan *Phishing*.



Gambar 5 Tampilan Ketika Membuka Situs *Phishing*

Tabel 2. Contoh Situs *Phishing*

| No | Situs *Phishing* |
|---|---|
| 1 | http://24e72176-d57a-4cdf-8d12-27b3d3ca13ce-00-35ep2mx3gzoxb.sisko.replit.dev |
| 2 | http://abuniz123.github.io/metamanager |

| | |
|---|---|
| 3 | http://ad-manenger-bill-prsj.vercel.app/authenticate/business.helpfacebook.com-help_159551727005616.html |
| 4 | http://algobuy.shop |
| 5 | http://ann.spacemotors.top |
| 6 | http://ayaanwebdeveloper.github.io/fb-clone |
| 7 | http://bbplc-ee.weeblysite.com |
| 8 | http://bt-103312.weeblysite.com |
| 9 | http://btinternet-106995.weeblysite.com |
| 10 | http://btinterrnett-104587.weeblysite.com |
| 11 | http://communication-ergo-additional-their.vercel.app/facebook.html |
| 12 | http://contactinstant.vercel.app/contact/us.html |
| 13 | http://eanop1.tahory.site |
| 14 | http://event.playaspirant.com |
| 15 | http://facebook-peach-chi.vercel.app |
| 16 | http://facebook-peach-chi.vercel.app/auth |
| 17 | http://facebookv01.vercel.app |
| 18 | http://fagebook-login.vercel.app |
| 19 | http://fkl-32655.bubbleapps.io |
| 20 | http://grounded-absolute-123701.framer.app |
| 21 | http://imasmonline.durablesites.com |
| 22 | http://insta-pro.pk |
| 23 | http://ledger-aktualisieren.pages.dev |
| 24 | http://litv.oilcompanyca.top |
| 25 | http://location-mauve-six.vercel.app |
| 26 | http://nnt-flame.vercel.app/metabusinesscenter.html |
| 27 | http://p95xg.com |
| 28 | http://pasxfuul.com |
| 29 | https://0000006738.vercel.app/sec.html?error=interaction_required&erro |
| 30 | https://2025-local.click |

Tabel 3. Contoh Situs Bukan *Phishing*

| Rank | Root Domain | Linking Root Domains | Domain Authority |
|---|---|---|---|
| 1 | www.google.com | 15,236,114 | 100 |
| 2 | www.blogger.com | 31,311,113 | 100 |
| 3 | youtube.com | 24,336,912 | 100 |
| 4 | linkedin.com | 13,291,390 | 99 |
| 5 | support.google.com | 5,720,703 | 99 |
| 6 | cloudflare.com | 8,211,585 | 99 |
| 7 | microsoft.com | 5,593,547 | 99 |
| 8 | apple.com | 6,849,526 | 99 |
| 9 | en.wikipedia.org | 7,201,596 | 98 |
| 10 | play.google.com | 4,012,038 | 98 |
| 11 | wordpress.org | 12,511,154 | 98 |
| 12 | docs.google.com | 3,642,278 | 98 |
| 13 | mozilla.org | 2,593,193 | 98 |
| 14 | maps.google.com | 6,190,949 | 98 |
| 15 | youtu.be | 5,434,247 | 98 |
| 16 | drive.google.com | 2,681,591 | 97 |
| 17 | bp.blogspot.com | 18,327,022 | 97 |
| 18 | sites.google.com | 2,401,535 | 97 |
| 19 | googleusercontent.com | 3,994,187 | 97 |
| 20 | accounts.google.com | 2,557,208 | 97 |
| 21 | t.me | 1,826,000 | 97 |
| 22 | europa.eu | 2,437,683 | 97 |
| 23 | plus.google.com | 10,955,614 | 97 |
| 24 | whatsapp.com | 4,778,976 | 97 |
| 25 | adobe.com | 2,880,183 | 96 |
| 26 | facebook.com | 61,926,417 | 96 |
| 27 | policies.google.com | 3,521,103 | 96 |
| 28 | uol.com.br | 694,206 | 96 |
| 29 | istockphoto.com | 3,728,189 | 96 |
| 30 | vimeo.com | 3,628,948 | 96 |

Dari *dataset* Tabel 3 dapat dilihat bahwa situs dengan Links Pointing To Page/ jumlah backlink dan domain authority tinggi dapat dipastikan bukan situs phishing. Dalam mengklasifikasi URL apakah phishing atau bukan phishing, salah satu fiturnya adalah links pointing to page.

## 3. Pra Pemrosesan Data

Tahapan pra pemrosesan adalah tahapan pengolahan awal setelah kegiatan akuisisi data. Tahapan ini mencari nilai fitur-fitur dari tiap-tiap URL. Pada Tabel 4, dalam *dataset* terdapat 31 fitur dan penjelasan mengenai berbagai variabel atau atributnya. Dari 30 di antaranya adalah fitur independen, dan 1 merupakan fitur dependen, yang menjadi label untuk klasifikasikan URL *Phishing* (1) dan Non-*Phishing* (-1). Nilai label untuk klasifikasi URL *Phishing* atau Non-*Phishing* tersebut sesuai dengan kondisi URL yang ada di *Dataset*. Setiap URL akan diberi nilai sesuai dengan fitur-fitur.

Tabel 4 Fitur-Fitur dalam *Dataset*

| Index | Fitur | Nilai | | |
|---|---|---|---|---|
| | | Legitimate | Suspicious | Phishing |
| 1 | UsingIP | -1 | | 1 |
| 2 | LongURL | -1 | 0 | 1 |
| 3 | ShortURL | -1 | | 1 |
| 4 | Symbol@ | -1 | | 1 |
| 5 | Redirecting// | -1 | | 1 |
| 6 | PrefixSuffix- | -1 | | 1 |
| 7 | SubDomains | -1 | 0 | 1 |
| 8 | HTTPS | -1 | 0 | 1 |
| 9 | DomainRegLen | -1 | | 1 |
| 10 | Favicon | -1 | | 1 |
| 11 | NonStdPort | -1 | | 1 |
| 12 | HTTPSDomainURL | -1 | | 1 |
| 13 | RequestURL | -1 | | 1 |
| 14 | AnchorURL | -1 | 0 | 1 |
| 15 | LinksInScriptTags | -1 | 0 | 1 |
| 16 | ServerFormHandler | -1 | 0 | 1 |
| 17 | InfoEmail | -1 | | 1 |
| 18 | AbnormalURL | -1 | | 1 |
| 19 | WebsiteForwarding | -1 | 0 | 1 |
| 20 | StatusBarCust | -1 | | 1 |
| 21 | DisableRightClick | -1 | | 1 |
| 22 | UsingPopupWindow | -1 | | 1 |
| 23 | IframeRedirection | -1 | | 1 |
| 24 | AgeofDomain | -1 | | 1 |
| 25 | DNSRecording | -1 | | 1 |
| 26 | WebsiteTraffic | -1 | | 1 |
| 27 | PageRank | -1 | | 1 |
| 28 | GoogleIndex | -1 | | 1 |
| 29 | LinksPointingToPage | -1 | | 1 |
| 30 | StatsReport | -1 | | 1 |
| 31 | Class | -1 | | 1 |

Setelah URL diberi nilai, maka tiap-tiap fitur dapat berisi nilai (1) artinya phishing, (0) artinya suspicious dan (-1) artinya non-phishing, untuk mengklasifikasikan URL masuk yang potensi sebagai phishing/ suspicious atau non-phishing. Contoh dataset yang sudah dinilai ke-30 fitur ditunjukkan pada Tabel 5.

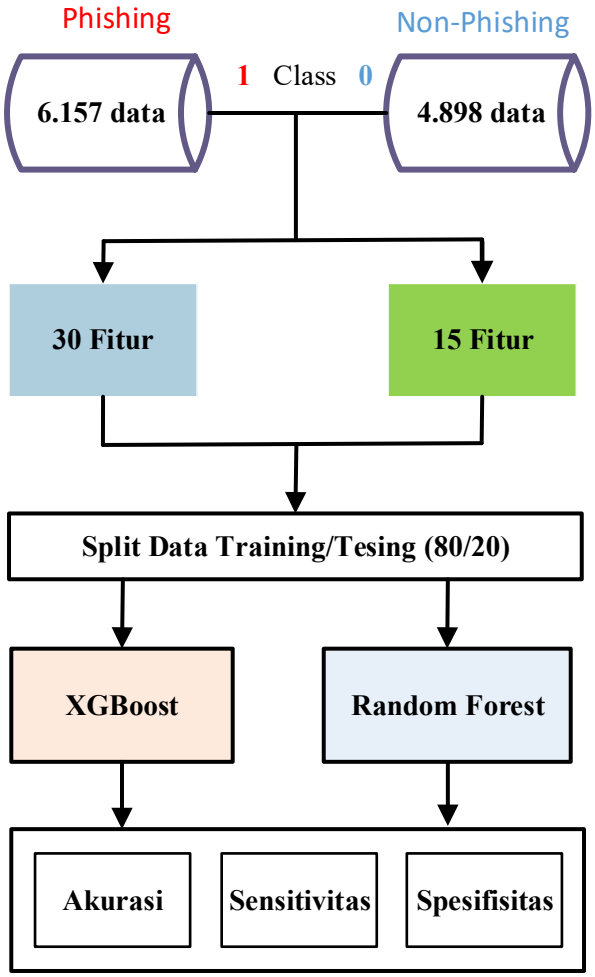
Tabel 5 Contoh *Dataset* Deteksi Phishing

| Using IP | LongURL | ShortURL | Symbol@ | Redirecting// | PrefixSuffix- | Subdomains | HTTPS | DomainRegLen | Favicon | NonStdPort | HTTPSDomainURL | RequestURL | AnchorURL | LinksInScriptTags | ServerFormHandler | InfoEmail | AbnormalURL | WebsiteForwarding | StatusbarCust | DisableRightClick | UsingPopupWindow | IframeRedirection | AgeofDomain | DNSRecording | WebsiteTraffic | PageRank | GoogleIndex | LinkPointingToPage | StatsReport | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | -1 | 0 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | -1 |
| 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | 0 | -1 | -1 |
| 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | 1 | 0 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 |
| 1 | 0 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 0 | -1 | 1 | 1 | 0 | -1 | 1 | -1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | 1 |
| -1 | 0 | -1 | 1 | -1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 0 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 |
| 1 | 0 | -1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 0 | -1 | -1 |
| 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 0 | 1 | -1 |
| 1 | 0 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | -1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 0 | 1 | -1 |
| 1 | 1 | 1 | 1 | 1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | 0 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | -1 | -1 | 1 | 1 |
| 1 | 1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 0 | -1 | -1 |
| -1 | 1 | -1 | 1 | -1 | -1 | 0 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | -1 | 1 | 1 | 0 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 | 1 | 0 | 1 | -1 |
| 1 | 1 | -1 | 1 | 1 | -1 | 0 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | -1 |
| 1 | 1 | -1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 |
| 1 | -1 | -1 | -1 | 1 | -1 | 0 | 0 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 0 | 1 | -1 |
| 1 | -1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 1 | -1 | -1 |
| 1 | -1 | 1 | 1 | 1 | -1 | -1 | 0 | 1 | 1 | -1 | 1 | 1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | -1 |
| 1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 | 1 |
| 1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 0 | -1 | -1 | -1 | 0 | -1 | -1 | -1 | -1 | 1 | -1 | 0 | -1 | 1 | 0 | -1 | 1 |
| 1 | 0 | -1 | 1 | 1 | -1 | 0 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 0 | -1 | -1 | -1 | 0 | -1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 |
| 1 | 0 | 1 | 1 | 1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 | 0 | -1 | 1 |
| 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | 0 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | -1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | -1 | 1 | 0 | -1 | 1 | 1 | 1 | 1 | 0 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | 1 |
| 1 | -1 | -1 | -1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | -1 | 0 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | 0 | 1 | 1 |
| 1 | -1 | 1 | 1 | 1 | -1 | 0 | 1 | -1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 1 | 1 |
| 1 | -1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | -1 | -1 | -1 |
| 1 | -1 | -1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 0 | -1 | 1 |
| 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | 1 |
| 1 | -1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |

## 4. Merancang Model Klasifikasi

Penelitian ini mengusulkan model pendeteksian apakah sebuah URL merupakan phishing atau bukan. Metode yang digunakan untuk mendeteksi adalah XGBoost dan Random Forest. Pendeteksian ini menggunakan 30 fitur dan akan diseleksi fitur-fitur mana yang tidak berpengaruh yang nantinya tidak digunakan. Pada langkah awal kami mencari data situs-situs yang merupakan phishing dan bukan phishing. Dari situs ini, dilakukan penilaian ke 30 fitur yang digunakan. Setelah diperoleh nilai dari fitur-fitur, data dibagi dua, 80% untuk data latih, 20% untuk data uji. Data latih yang telah diperoleh, dilakukan pelatihan dengan metode XGBoost dan Random Forest untuk menghasilkan model. Dari model ini dilakukan pengujian dengan data uji untuk memperoleh nilai akurasi, sensitivitas dan spesifisitas.

Model yang dibuat, dilakukan seleksi fitur untuk mencari fitur-fitur mana yang tidak terlalu berpengaruh dalam proses klasifikasi. Dari kedua model, dicari fitur-fitur mana yang sama dan cukup berpengaruh. Fitur-fitur yang berpangaruh ini dilakukan pelatihan ulang untuk memperoleh model baik dengan metode XGBoost maupun Random Forest. Model dengan fitur terseleksi, dilakukan pengujian untuk memperoleh nilai akurasi, sensitivitas dan spesifisitas. Alur proses pengklasifikasian pendeteksian phishing ditunjukkan pada Gambar 6.




Gambar 6 Alur Proses Pengklasifikasian Phishing

## 5. Pengujian Model Deteksi Situs Phishing

Model deteksi situs phishing yang dibuat perlu diuji apakah telah sesuai yang diharapkan atau tidak. Pengujian dilakukan dengan menggunakan beberapa URL baik yang phishing maupun yang non-phishing. Pengujian yang dilakukan akan menghasilkan klasifikasi dari URL yang dimasukan, yaitu phishing atau non-phishing. Model deteksi situs phishing yang dibuat perlu dievaluasi. Pengevaluasian disajikan melalui nilai akurasi, spesifisitas, dan sensitivitas yang didapat dari *confusion matrix* seperti dinyatakan dalam Persamaan (*1*)(*2*)(*3*). Ada empat kondisi yang terjadi dalam pengukuran ini. Kondisi pertama adalah URL berupa situs phishing dan sistem melaporkan phishing. Kondisi demikian disebut dengan istilah True Positive (TP). Kondisi kedua adalah URL berupa situs phishing dan sistem melaporkan non-phishing. Kondisi demikian disebut dengan istilah False Negative(FN). Kondisi ketiga adalah bila URL berupa situs non-phishing dan sistem melaporkan non-phishing, maka informasinya benar atau disebut True Negative (TN). Kondisi keempat bila URL berupa situs non-phishing dan sistem melaporkan phishing, maka informasi yang diberikan sistem tidak benar atau disebut False Positive(FP). Keempat kondisi ini ditunjukkan pada Tabel 6.

$$Akurasi = \frac{TP + TN}{TP + FN + TN + FP} \; X \; 100\% \qquad (1)$$

$$Spesifisitas = \frac{TN}{TN + FP} \; X \; 100\% \quad (2)$$

$$Sensitivitas = \frac{TP}{TP + FN} \; X \; 100\% \quad (3)$$

Tabel 6 Confusion Matrix

| | | Nilai Sebenarnya | |
|---|---|---|---|
| | | Phishing | Non-Phishing |
| Nilai Prediksi | Phishing | True Positive (TP) | False Positive (FP) |
| | Non-Phishing | False Negative (FN) | True Negative (TN) |

# BAB III
# HASIL PELAKSANAAN PENELITIAN

Berdasarkan model yang telah dirancang sebelumnya, berikut diberikan implementasi dan hasil pengujiannya.

## 1. Akuisisi Data dan Pra Pemrosesan

Data yang digunakan untuk penelitian ini adalah 30 nilai dari fitur-fitur tiap URL dan 1 kelas. Ke 30 nilai fitur ini berisi angka -1 (non-phishing), 0 (dicurigai phishing) dan 1 (phishing), sedangkan nilai kelas berisi angaka 0 (situs non-phishing) dan 1 (situs phishing). Dalam penelitian ini menggunakan data sebanyak 11.055 yang terbagi menjadi 4.898 data situs non-phishing dan 6.157 data situs phishing. Data penelitian ini dibagi dua, 80% untuk data pelatihan dan 20% untuk data pengujian. Contoh data pelatihan ditunjukkan pada Tabel *7*, sedangkan data pengujian ditunjukkan pada Tabel 8.

Tabel 7 Contoh Dataset untuk Pelatihan

| Nilai dari Fitur-Fitur | Kelas |
|---|---|
| -1,1,1,1,-1,-1,-1,-1,-1,-1,1,1,-1,1,-1,1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,1,-1 | 0 |
| 1,1,1,1,1,-1,0,1,-1,1,1,-1,1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,1,1 | 0 |
| 1,0,1,1,1,-1,-1,-1,-1,1,1,-1,1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,1,-1,1,0,-1 | 0 |
| 1,0,1,1,1,-1,-1,-1,1,1,1,-1,-1,0,0,-1,1,1,0,1,1,1,1,-1,-1,1,-1,1,-1,1 | 0 |
| 1,0,-1,1,1,-1,1,1,-1,1,1,1,1,0,0,-1,1,1,0,-1,1,-1,1,-1,-1,0,-1,1,1,1 | 1 |
| -1,0,-1,1,-1,-1,1,1,-1,1,1,-1,1,0,0,-1,-1,-1,0,1,1,1,1,1,1,1,-1,1,-1,-1 | 1 |
| 1,0,-1,1,1,-1,-1,-1,1,1,1,1,-1,-1,0,-1,-1,-1,0,1,1,1,1,1,-1,-1,-1,1,0,-1 | 0 |
| 1,0,1,1,1,-1,-1,-1,1,1,1,-1,-1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,0,1 | 0 |
| 1,0,-1,1,1,-1,1,1,-1,1,1,-1,1,0,1,-1,1,1,0,1,1,1,1,1,-1,1,1,1,0,1 | 1 |
| 1,1,-1,1,1,-1,-1,1,-1,1,1,1,1,0,1,-1,1,1,0,1,1,1,1,1,-1,0,-1,1,0,1 | 0 |
| 1,1,1,1,1,-1,0,1,1,1,1,1,-1,0,0,-1,-1,-1,0,1,1,1,1,-1,1,1,1,1,-1,-1 | 1 |
| 1,1,-1,1,1,-1,1,-1,-1,1,1,1,1,1,-1,-1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,0,-1 | 0 |
| -1,1,-1,1,-1,-1,0,0,1,1,1,-1,-1,-1,1,-1,1,1,0,-1,1,-1,1,1,-1,-1,-1,1,0,1 | 0 |
| 1,1,-1,1,1,-1,0,-1,1,1,1,1,-1,-1,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,1,1 | 0 |
| 1,1,-1,1,1,1,-1,1,-1,1,1,-1,1,0,1,1,1,1,0,1,1,1,1,1,-1,1,-1,1,-1,1 | 1 |
| 1,-1,-1,-1,1,1,-1,0,0,1,1,1,1,-1,-1,0,-1,1,1,0,1,1,1,1,1,-1,-1,-1,1,0,1 | 0 |
| 1,-1,-1,1,1,1,-1,1,1,-1,1,1,-1,1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,0,-1,1,1,-1 | 0 |
| 1,-1,1,1,1,1,-1,-1,0,1,1,-1,1,1,1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,1,1,1,-1,1,1,-1 | 0 |
| 1,1,1,1,1,1,-1,-1,1,1,1,1,1,-1,-1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,-1,1,1,1,-1,-1 | 1 |
| 1,1,1,1,1,-1,-1,-1,1,1,1,1,1,0,0,-1,-1,-1,0,-1,-1,-1,-1,1,-1,0,-1,1,0,-1 | 1 |
| 1,0,-1,1,1,-1,0,1,-1,1,1,1,1,1,0,0,-1,-1,-1,0,-1,1,-1,1,-1,1,1,1,-1,1,-1,-1 | 1 |
| 1,0,1,1,1,-1,0,1,1,1,1,-1,-1,-1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,1,-1,-1,1,0,-1 | 1 |
| 1,1,1,1,1,-1,-1,-1,-1,1,1,-1,1,0,0,-1,1,1,0,1,1,1,1,1,1,0,-1,1,-1,1 | 1 |
| 1,1,1,1,1,-1,1,0,-1,1,1,1,1,1,0,0,-1,1,1,0,1,1,1,1,1,1,1,-1,1,-1,-1 | 1 |
| 1,-1,-1,-1,1,1,-1,1,1,-1,1,1,-1,-1,0,0,-1,1,1,0,1,1,1,1,1,1,-1,-1,1,0,1 | 1 |
| 1,-1,1,1,1,-1,0,1,-1,1,1,1,1,1,1,0,-1,1,1,0,1,1,1,1,-1,1,1,-1,1,0,1 | 1 |
| 1,-1,1,1,1,-1,0,-1,1,1,1,-1,-1,-1,-1,-1,-1,-1,0,1,1,1,1,1,1,0,-1,1,-1,-1 | 1 |
| 1,-1,-1,1,1,1,-1,1,1,1,1,1,-1,1,0,-1,-1,-1,0,1,1,1,1,1,-1,0,-1,1,0,-1 | 1 |
| 1,-1,-1,1,-1,1,-1,1,-1,1,1,1,1,1,1,0,-1,1,1,1,1,1,1,1,-1,-1,-1,1,-1,-1 | 1 |
| 1,-1,1,1,1,1,-1,-1,1,-1,1,1,1,1,1,1,1,0,-1,1,1,0,1,1,1,1,-1,1,1,-1,1,0,1 | 1 |

Tabel 8 Contoh Dataset Pengujian

| Nilai dari Fitur-Fitur | Kelas |
|---|---|
| -1,-1,1,-1,1,-1,1,-1,1,-1,-1,-1,-1,-1,1,-1,-1,1,0,-1,-1,-1,-1,1,1,-1,-1,1,1,-1 | 0 |
| -1,-1,-1,-1,-1,-1,1,1,1,-1,1,1,1,-1,0,0,-1,1,-1,0,1,1,1,1,1,-1,-1,-1,1,0,1 | 1 |
| 1,-1,1,-1,1,-1,1,1,-1,1,1,1,1,0,1,1,1,1,0,1,1,1,1,1,1,-1,-1,1,0,1 | 1 |
| -1,-1,1,-1,1,-1,-1,-1,-1,-1,-1,-1,-1,-1,1,-1,-1,1,0,-1,1,-1,-1,1,1,1,-1,1,0,1 | 0 |
| 1,-1,1,1,1,1,-1,1,1,1,1,1,-1,0,-1,-1,1,1,0,1,1,1,1,1,1,-1,-1,1,1,1 | 1 |
| -1,-1,1,1,1,-1,0,0,-1,1,1,1,1,-1,0,-1,1,1,0,1,1,1,1,1,1,1,-1,1,1,1 | 0 |
| -1,-1,1,1,1,-1,0,-1,-1,1,1,1,1,-1,0,-1,1,1,0,1,1,1,1,1,1,1,-1,1,1,-1 | 0 |
| 1,1,1,1,1,-1,1,-1,-1,1,1,1,1,0,1,-1,1,1,0,1,1,1,1,1,1,1,1,0,-1 | 0 |
| -1,-1,1,-1,1,-1,-1,0,1,1,1,1,-1,-1,1,1,1,1,0,1,1,1,1,1,1,1,-1,1,1,-1 | 0 |
| -1,-1,1,-1,1,-1,0,0,1,1,1,1,1,0,-1,-1,1,1,0,1,1,1,1,1,1,1,-1,-1,1,-1 | 0 |
| 1,-1,1,1,1,-1,0,-1,-1,1,1,1,-1,-1,-1,-1,1,1,0,1,1,1,1,1,1,1,-1,1,0,-1 | 0 |
| 1,1,1,-1,1,1,1,1,-1,-1,-1,-1,1,1,-1,-1,-1,1,0,1,1,-1,-1,-1,1,-1,1,1,0,1 | 1 |

## 2. Hasil Pengujian

### a. Pengujian Dengan 30 Fitur

Pada pengujian pertama dilakukan menggunakan 30 fitur menggunakan metode XGBoost dan Random Forest. Dari hasil pengujian yang dilakukan menggunakan XGBoost maka dapat dihitung nilai True Positive (TP), True Negative (TN), False Negative (FN) dan False Positive (FP), seperti ditunjukkan pada Tabel *9*.

Tabel 9 Nilai TP, TN, FN, FP Metode XGBoost 30 Fitur

| | |
|---|---|
| TP | 1208 |
| TN | 943 |
| FN | 37 |
| FP | 23 |

Penguji sistem deteksi situs phishing menggunakan metode XGBoost dengan 30 fitur memperoleh

$$\text{Nilai Akurasi} = \frac{TP + TN}{TP + FN + TN + FP} X\ 100\%$$

$$= \frac{1208 + 943}{1208 + 37 + 943 + 23} X\ 100\%$$

$$= 97{,}3\%$$

$$\text{Nilai Sensitivitas} = \frac{TP}{TP + FN} X\ 100\%$$

$$= \frac{1208}{1208 + 37} X\ 100\%$$

$$= 97\%$$

$$\text{Nilai Spesifisitas} = \frac{TN}{TN + FP} X\ 100\%$$

$$= \frac{943}{943 + 23} X\ 100\%$$

$$= 97{,}6\%$$

Pengujian menggunakan metode Random Forest dengan 30 fitur maka diperoleh TP, TN, FN, FP seperti ditunjukkan pada Tabel 10.

Tabel 10 Nilai TP, TN, FN, FP Metode Random Forest 30 Fitur

| | |
|---|---|
| TP | 1216 |
| TN | 937 |
| FN | 43 |
| FP | 15 |

Penguji sistem deteksi situs phishing menggunakan metode Random Forest dengan 30 fitur memperoleh

$$\text{Nilai Akurasi} = \frac{TP + TN}{TP + FN + TN + FP} X\ 100\%$$

$$= \frac{1216 + 937}{1216 + 43 + 937 + 15} X\ 100\%$$

$$= 97{,}4\%$$

$$\text{Nilai Sensitivitas} = \frac{TP}{TP + FN} X\ 100\%$$

$$= \frac{1216}{1216 + 43} X\ 100\%$$

$$= 96{,}6\%$$

$$\text{Nilai Spesifisitas} = \frac{TN}{TN + FP} X\ 100\%$$

$$= \frac{937}{937 + 15} X\ 100\%$$

$$= 98{,}4\%$$

## b. Pengujian Dengan Seleksi Fitur

Pada pengujian kedua, kami melakukan seleksi fitur dimana fitur-fitur yang tidak berpengaruh akan dihapus atau tidak digunakan. Seleksi fitur ini menggunakan metode feature importance dan dipilih 15 fitur. Pada pengujian kedua, kami menghitung bobot tiap-tiap fitur dari data model yang dibuat menggunakan XGBoost dan Random Forest dengan 30 fitur. Dari kedua bobot fitur-fitur tersebut, kami memilih 15 fitur teratas yang sama dari kedua metode tersebut. Ke 15 fitur tersebut ditunjukkan pada Tabel 11.

Tabel 11 Fitur-Fitur Setelah Diseleksi

| No | Fitur |
|---|---|
| 1 | UsingIP |
| 2 | PrefixSuffix- |
| 3 | SubDomains |
| 4 | HTTPS |
| 5 | DomainRegLen |
| 6 | RequestURL |

| | |
|---|---|
| 7 | AnchorURL |
| 8 | LinksInScriptTags |
| 9 | ServerFormHandler |
| 10 | AgeofDomain |
| 11 | DNSRecording |
| 12 | WebsiteTraffic |
| 13 | PageRank |
| 14 | GoogleIndex |
| 15 | LinksPointingToPage |

Pada pengujian kedua dengan 15 fitur ini akan digunakan metode XGBoost dan Random Forest kembali. Dari hasil pengujian yang dilakukan menggunakan XGBoost maka dapat dihitung nilai True Positive (TP), True Negative (TN), False Negative (FN) dan False Positive (FP), seperti ditunjukkan pada Tabel 12.

Tabel 12 Nilai TP, TN, FN, FP Metode XGBoost 15 Fitur

| | |
|---|---|
| TP | 1207 |
| TN | 937 |
| FN | 43 |
| FP | 24 |

Penguji sistem deteksi situs phishing menggunakan metode XGBoost dengan 15 fitur memperoleh

$$\text{Nilai Akurasi} = \frac{TP + TN}{TP + FN + TN + FP} X\ 100\%$$

$$= \frac{1207 + 937}{1207 + 43 + 937 + 24} X\ 100\%$$

$$= 97{,}0\%$$

$$\text{Nilai Sensitivitas} = \frac{TP}{TP + FN} X\ 100\%$$

$$= \frac{1207}{1207 + 43} X\ 100\%$$

$$= 96{,}6\%$$

$$\text{Nilai Spesifisitas} = \frac{TN}{TN + FP} X\ 100\%$$

$$= \frac{937}{937 + 24} X\ 100\%$$

$$= 97{,}5\%$$

Pengujian menggunakan metode Random Forest dengan 15 fitur maka diperoleh TP, TN, FN, FP seperti ditunjukkan pada Tabel *13*.

Tabel 13 Nilai TP, TN, FN, FP Metode Random Forest 15 Fitur

| | |
|---|---|
| TP | 1213 |
| TN | 931 |
| FN | 49 |
| FP | 18 |

Penguji sistem deteksi situs phishing menggunakan metode Random Forest dengan 15 fitur memperoleh

$$\text{Nilai Akurasi} = \frac{TP + TN}{TP + FN + TN + FP} X\ 100\%$$

$$= \frac{1213 + 931}{1213 + 49 + 931 + 18} X\ 100\%$$

$$= 97{,}0\%$$

$$\text{Nilai Sensitivitas} = \frac{TP}{TP + FN} X\ 100\%$$

$$= \frac{1213}{1213 + 49} X\ 100\%$$

$$= 96{,}1\%$$

$$\text{Nilai Spesifisitas} = \frac{TN}{TN + FP} X\ 100\%$$

$$= \frac{931}{931 + 18} X\ 100\%$$

$$= 98{,}1\%$$

**3. Analisis Hasil**

Berdasarkan hasil pengujian, model yang diusulkan dapat mendeteksi sebuah URL apakah termasuk phishing atau bukan. Hasil pengujian menggunakan metode XGBoost dan Random Forest dengan jumlah fitur 30 dan 15 ditunjukkan pada Tabel *14*.

Tabel 14 Hasil Pengujian

| | | Akurasi | Sensitivitas | Spesifisitas |
|---|---|---|---|---|
| XGBoost | 30 Fitur | 97,3% | 97,0% | 97,6% |
| | 15 Fitur | 97,0% | 96,6% | 97,5% |
| Random Forest | 30 Fitur | 97,4% | 96,6% | 98,4% |
| | 15 Fitur | 97,0% | 96,1% | 98,1% |

Dari hasil ujicoba dapat diketahui bahwa nilai akurasi tertinggi pada metode XGBoost sebesar 97,4% menggunakan 30 fitur sedangkan 15 fitur hanya 97,0%. Pada metode Random Forest akurasi tertinggi juga menggunakan 30 fitur sebesar 97,4% dibandingkan dengan 15 fitur hanya 97,0%. Keberhasilan memprediksi sebuah URL phishing adalah benar-benar phishing nilai tertinggi sebesar 97% dengan metode XGBoost menggunakan 30 fitur, sedangkan kerberhasilan memprediksi URL bukan phishing adalah benar-benar bukan phishing, nilai tertinggi sebesar 98,4% dengan metode Random Forest menggunakan 30 fitur.

.

# BAB IV
# KESIMPULAN DAN SARAN

## 1. Kesimpulan

Berdasarkan hasil ujicoba pada penelitian ini maka dapat diambil kesimpulan sebagai berikut:

a. Metode XGBoost dan Random Forest menggunakan 30 fitur menghasil nilai akurasi yang tidak beda jauh, sedangkan menggunakan 15 fitur, menghasilkan nilai akurasi yang sama.
b. URL bukan phishing diprediksi bukan phishing, metode Random Forest lebih baik dibandingkan XGBoost.
c. URL phishing diprediksi sebagai phishing, metode XGBoost lebih baik dibandingkan Random Forest.

## 2. Saran

Saran-saran yang diberikan untuk mengembangkan penelitian ini adalah

a. Model deteksi URL phishing atau bukan ini perlu dikembangkan dengan mencoba beberapa seleksi fitur lainnya agar memperoleh fitur yang optimal dalam pendeteksian.
b. Waktu proses perlu diperhitungkan agar memperoleh akurasi dan waktu yang diideal dalam pendeteksian. Semakin banyak fitur yang digunakan, semakin lama waktu yang dibutuhkan untuk pendeteksian.

# DAFTAR PUSTAKA


Alam, M.N. *et al.* (2020) 'Phishing attacks detection using machine learning approach', *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, (Icssit), pp. 1173–1179. Available at: https://doi.org/10.1109/ICSSIT48917.2020.9214225.

Baduwal, M. *et al.* (2023) 'Survey On Machine Learning Paradigms For Phishing Website Detection', *Article in International Journal of Engineering & Technology* [Preprint], (June). Available at: https://www.researchgate.net/publication/370658818.

BSSN (2024) *Lanskap Keamanan Siber Indonesia*. Available at: https://www.bssn.go.id/wp-content/uploads/2025/02/LANSKAP-KEAMANAN-SIBER-2024-1.pdf (Accessed: 15 April 2025).

Deekshitha, B. (2022) 'URL Based Phishing Website Detection by Using Gradient and Catboost Algorithms', *International Journal for Research in Applied Science and Engineering Technology*, 10(6), pp. 3717–3722. Available at: https://doi.org/10.22214/ijraset.2022.43986.

Jain, A.K. *et al.* (2020) 'PhishSKaPe: A Content based Approach to Escape Phishing Attacks', *Procedia Computer Science*, 171(2019), pp. 1102–1109. Available at: https://doi.org/10.1016/j.procs.2020.04.118.

Kusumawardhani, N.Q. (2022) *91 Persen Serangan Siber Berawal dari Kerentanan Ini*. Available at: https://tekno.republika.co.id/berita/rejeqc368/91-persen-serangan-siber-berawal-dari-kerentanan-ini (Accessed: 15 April 2025).

Mahmud, A.F. and Wirawan, S. (2024) 'Deteksi Phishing Website menggunakan Machine Learning Metode Klasifikasi', 13(4), pp. 2540–9719. Available at: http://sistemasi.ftik.unisi.ac.id.

Makbull Rizki (2022) 'Perkembangan Sistem Pertahanan/Keamanan Siber Indonesia dalam Menghadapi Tantangan Perkembangan Teknologi dan Informasi', *Politeia: Jurnal Ilmu Politik*, 14(1), pp. 54–62. Available at: https://doi.org/10.32734/politeia.v14i1.6351.

Muftiadi, A., Agustina, T.P.M. and Evi, M. (2022) 'Studi kasus keamanan jaringan komputer: analisis ancaman phising terhadap layanan online banking', *Hexatech: Jurnal Ilmiah Teknik*, 1(2), pp. 60–65. Available at: https://doi.org/10.55904/hexatech.v1i2.346.

Musa, H. *et al.* (2019) 'A comparative analysis of phishing website detection using XGBOOST algorithm', *Journal of Theoretical and Applied Information Technology*, 97(5), pp. 1434–1443.

Ozker, U. and Sahingoz, O.K. (2020) 'Content Based Phishing Detection with Machine Learning', in *2020 International Conference on Electrical Engineering, ICEE 2020*. Istambul. Available at: https://doi.org/10.1109/ICEE49691.2020.9249892.

Putri, N.B. and Wijayanto, A.W. (2022) 'Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing', *Komputika : Jurnal Sistem Komputer*, 11(1), pp. 59–66. Available at: https://doi.org/10.34010/komputika.v11i1.4350.

Ryan, J. (2023) *Phishing Sites Impersonating Social Media Jump in Q2*. Available at: https://www.phishlabs.com/blog/phishing-sites-impersonating-social-media-jump-in-q2 (Accessed: 17 April 2025).

Saputro, I.A., Sugiarto, L. and Nugraha, F.S. (2024) 'ANALISIS KESADARAN

MASYARAKAT TERHADAP BAHAYA INTERNET PHISHING MENGGUNAKAN K-MEANS CLUSTERING', *Jurnal Satuan Tulisan Riset dan Inovasi Teknologi*, 9(2), pp. 139–146.

Windarni, V.A. *et al.* (2023) 'Deteksi Website Phishing Menggunakan Teknik Filter Pada Model Machine Learning', *Information System Journal*, 6(01), pp. 39–43. Available at: https://doi.org/10.24076/infosjournal.2023v6i01.1268.

Zamir, A. *et al.* (2020) 'Phishing web site detection using diverse machine learning algorithms', *Electronic Library*, 38(1), pp. 65–80. Available at: https://doi.org/10.1108/EL-05-2019-0118.

## LAMPIRAN

**Lampiran 1. Realisasi Penggunaan Anggaran**
Dana Disetujui: Rp 14.000.000,

| Jenis Pembelajaan | Komponen | Item | Kuantitas | Biaya Satuan | Total |
|---|---|---|---|---|---|
| Belanja Bahan | ATK | Kertas, Tinta, Pulpen | 1 | 265.000 | 265.000 |
| Belanja Bahan | Bahan penelitian (habis pakai) | Materai | 4 | 12.000 | 48.000 |
| Pengumpulan Data | Honor pembantu peneliti | Pengambilan data situs *phishing* dan buka *phishing* | 1 | 750.000 | 750.000 |
| Pengumpulan Data | Transport | transport rapat dalam kota (3 kali PP) | 6 | 75.000 | 450.000 |
| Pengumpulan Data | Konsumsi | Biaya konsumsi makan rapat (3 x 4 orang) | 12 | 85.000 | 1.020.000 |
| Analisis Data | Honor pengolah data | Honor pembersihan data | 1 | 850.000 | 850.000 |
| Analisis Data | Honor programmer | Honor pembuatan program web | 1 | 3.000.000 | 3.000.000 |
| Sewa Peralatan | Peralatan penelitian | Sewa server VPS Rumah Web Size L | 1 | 2.747.000 | 2.747.000 |
| Sewa Peralatan | Peralatan penelitian | Sewa domain check*phishing* .com | 1 | 180.000 | 180.000 |
| Sewa Peralatan | Peralatan penelitian | Sewa internet 6 bulan | 6 | 150.000 | 900.000 |
| Pelaporan penelitian | Honor staf administrasi | Administrasi laporan, keuangan, surat-menyurat | 1 | 750.000 | 750.000 |
| Pelaporan Penelitian | Biaya Konsumsi Rapat | Biaya Konsumsi Rapat evaluasi dan Pelaporan (4 orang) | 4 | 85.000 | 340.000 |

| | | | | | |
|---|---|---|---|---|---|
| Pelaporan Penelitian | Biaya Publikasi Sinta ½ | Jurnal Bereputasi Nasional (terakreditasi SINTA 1/2) | 1 | 2.500.000 | 2.500.000 |
| Lainnya | Biaya pendaftaran HKI | HKI Aplikasi | 1 | 200.000 | 200.000 |

**Lampiran 2. Biodata Ketua dan Anggota Tim Peneliti**
**Biodata Ketua Peneliti**

**A. Identitas Diri**

1. Nama Lengkap (dengan gelar) : Dr. Ir. Mardi Hardjianto, M.Kom.
2. Jenis Kelamin : Laki-Laki
3. Jabatan Fungsional : Lektor
4. NIP/NIDN/ID-SINTA : 920024/0422036901/5977626
5. Tempat, Tanggal Lahir : Jakarta, 22 Maret 1969
6. E-mail : mardi.hardjianto@budiluhur.ac.id
7. Nomor Handphone : 081586603011
8. Alamat : Jl. Menteng II/31 Grogol Petamburan, Jakarta

**B. Riwayat Pendidikan**

| | S1 | S2 | S3 |
|---|---|---|---|
| Nama Perguruan Tinggi | STMIK Budi Luhur | Universitas Indonesia | Universitas Gadjah Mada |
| Bidang Ilmu | Teknik Informatika | Ilmu Komputer | Ilmu Komputer |
| Tahun Masuk-Lulus | 1987-1993 | 1995-1999 | 2012-2018 |

**C. Pengalaman Penelitian (5 Tahun Terakhir)**

| No. | Tahun | Judul Penelitian | Pendanaan | |
|---|---|---|---|---|
| | | | Sumber* | Jumlah (Rp) |
| 1. | 2023 | Sistem Presensi Dengan Pendeteksian Wajah Menggunakan Metode K-NN Serta Optimasi Particle Swarm Optimization dan Histogram Of Oriented Gradients di Universitas Budi Luhur | Universitas Budi Luhur | 12.500.000 |
| 2. | 2023 | Pengembangan Metode Threshold Berbasis Akselerometer Smartphone untuk Deteksi Manusia Jatuh | Universitas Budi Luhur | 10.000.000 |

**D. Publikasi Artikel Ilmiah dalam Jurnal (5 Tahun Terakhir)**

| No. | Judul Artikel Ilmiah* | Nama Jurnal | Volume/Nomor/ Tahun |
|---|---|---|---|
| 1. | Perbandingan Penggunaan Bit Steganografi Metode Least Significant Bit (LSB) m-bit pada Citra Digital | Jurnal TiCOM | Vol. 13, No. 3, Mei 2025 |
| 2. | A graph neural network model application in point cloud structure for prolonged sitting detection system based on smartphone sensor data | ETRI Journal | Vol. 47, No. 2, April, 2025 |
| 3. | Meatcheck: Deteksi Kualitas Daging Sapi Berbasis Mobile Deep Learning | Jurnal Ilmiah Informatika Komputer | Vol. 30, No. 1, April 2025 |

| | | | |
|---|---|---|---|
| 4. | Pembuatan GH3RTZ Framework untuk Proses Reconnaissance dan Deteksi Celah Keamanan | Jurnal TiCOM | Vol. 13, No. 2, Januari 2025 |
| 5. | Optimasi Support Vector Machines (SVM) Menggunakan Particle Swarm Optimization (PSO) pada Analisis Sentimen Ulasan Pengguna Layanan Bukalapak di Twitter | Jurnal Pendidikan Teknologi Informasi | Vol. 4, No. 3, November 2024 |
| 6. | Deteksi Celah Keamanan SQL Injection Pada Website Dengah Menggunakan Metode Penetration Testing | Jurnal TELEMATIKA MKOM | Vol. 16, No. 2, September 2024 |
| 7. | Deteksi Website Phishing Dari Analisis Url Menggunakan Algoritma Random Forest | Jurnal Bit | Vol. 21, No. 2, September 2024 |
| 8. | Pengenalan Wajah Secara Realtime Menggunakan Adaboost Viola-Jones dan 2D DWT-PCA dengan Struktur Index KNN-KD Tree | Jurnal Pendidikan Teknologi Informasi | Vol. 4, No. 1, Maret 2024 |

**E. Pemakalah Seminar Ilmiah (5 Tahun Terakhir)**

| No. | Nama Temu Ilmiah/Seminar | Judul Artikel Ilmiah* | Waktu dan Tempat |
|---|---|---|---|
| 1. | | | |

**F. Perolehan HKI (5 Tahun Terakhir)**

| No. | Judul/Tema HKI* | Tahun | Jenis | Nomor P/ID |
|---|---|---|---|---|
| 1. | Program Sistem Presensi dengan Pendeteksian Wajah | 2025 | Program Komputer | 000938105 |
| 2. | Kriptografi Menggunakan Metode Base-32 Dikombinasi Dengan Kata Sandi | 2024 | Program Komputer | 000612493 |
| 3. | Kode Program Untuk Mengenkripsi Data atau Informasi Rahasia Menggunakan Base64 dan Kata Sandi | 2024 | Program Komputer | 000576347 |
| 4. | Kode Program Deteksi Jatuh Pada Manusia Menggunakan Metode Threshold Berbasis Akselerometer Pada Smartphone | 2023 | Program Komputer | 000509657 |

Jakarta, 30 Agustus 2025
Ketua Peneliti,

(Dr. Ir. Mardi Hardjianto, M.Kom.)

**Biodata Anggota 1**

Identitas Diri

1. Nama Lengkap : Agnes Aryasanti, S.Kom., M.Kom
2. Jenis Kelamin : Perempuan
3. Jabatan Fungsional : Lektor
4. NIP/NIDN/ID-SINTA : 120059/0322018502/ 6043631
5. Tempat, Tanggal Lahir : Gumawang, 22 Januari 1985
6. E-mail : agnes.aryasanti@budiluhur.ac.id
7. Nomor Handphone : 085945766802
8. Alamat : Jl. H. Thosin, Ciledug Tangerang Banten

Riwayat Pendidikan

| | **S1** | **S2** | **S3** |
|---|---|---|---|
| **Nama Perguruan Tinggi** | Universitas Budi Luhur | Universitas Budi Luhur | - |
| **Bidang Ilmu** | Sistem Informasi | Ilmu Komputer | - |
| **Tahun Masuk – Lulus** | 2002-2008 | 2012-2014 | - |

Pengalaman Penelitian (5 Tahun Terakhir)

| No | Tahun | Judul Penelitian | Pendanaan | |
|---|---|---|---|---|
| | | | Sumber* | Jumlah (Rp) |
| 1. | 2024 | Sistem Presensi Dengan Pendeteksian Wajah Menggunakan Metode K-NN Serta Optimasi Particle Swarm Optimization dan Histogram Of Oriented Gradients di Universitas Budi Luhur | Universitas Budi Luhur | 12.500.000,- |
| 2. | 2023 | Pengembangan Metode Threshold Berbasis Akselerometer Smartphone untuk Deteksi Manusia Jatuh | Universitas Budi Luhur | 10.000.000,- |
| 3. | 2019 | Analisis Perbandingan Metode Profile Matching Dan Simple Additive Weighting Untuk Menentukan Kelayakan Pemberian Pinjaman Pada Perusahaan Freight Forwarding | Universitas Budi Luhur | 15.000.000,- |

Publikasi Artikel Ilmiah dalam Jurnal (5 Tahun Terakhir)

| No | Tahun | Judul Artikel Ilmiah | Nama Jurnal | Volume/Nomor / Tahun |
|---|---|---|---|---|
| 1. | 2024 | Pengembangan Deteksi Jatuh pada Manusia Menggunakan Metode Threshold Berbasis Data Akselerometer pada Smartphone | Jurnal Edukasi dan Penelitian Informatika | 10/1/2024 |
| 2. | 2023 | Implementasi Sistem Pengelolaan Rekap Jual Beli pada Toko Busana | Journal Shift | 03/02/2023 |
| 3. | 2023 | Implementasi Keamanan File Menggunakan Metode Kriptografi Base-64 dan Steganografi Least Significant Bit (LSB) Random 2-Bit Berbasis Web | Journal TICOM (Technology of Information and Communication) | 11/02/2023 |
| 4. | 2022 | Implementasi Tanda Tangan Digital Menggunakan Algoritme RSA dan SHA-512 dengan Salt Berbasis Web | Jurnal Ticom : Technology of Information and Communication | 10/03/2023 |
| 5. | 2022 | Penerapan Sensor Ultrasonik HC-SR04 dan Hujan untuk Memantau Ketinggian Air dan Pendeteksi Hujan | Jurnal Media Informatika Budidarma | 06/02/2022 |
| 6. | 2022 | Implementasi E-Commerce Menggunakan Wordpress pada Toko Jack'O Shirt | Jurnal Ticom: Technology of Information and Communication | 10/02/2022 |
| 7. | 2022 | Rancangan Sistem Pemantau Kebocoran Gas dan Api Berbasis Internet of Things Menggunakan MQ-2 dan Flame Sensor Module | Jurnal Ticom: Technology of Information and Communication | 10/02/2022 |
| 8. | 2019 | Analisis Perbandingan Metode Profile Matching Dan Simple Additive Weighting Untuk Menentukan Kelayakan Pemberian Pinjaman Pada Perusahaan Freight Forwarding | Budi Luhur Information Technology | 16/01/2019 |

Pemakalah Seminar Ilmiah (5 Tahun Terakhir)

| No | Tahun | Nama Temu Ilmiah/ Seminar | Nama Jurnal | Waktu dan Tempat |
|---|---|---|---|---|
| | | | | |

Perolehan HKI (10 Tahun Terakhir)

| No | Judul HKI | Tahun | Jenis | No Pencatatan |
|---|---|---|---|---|
| 1. | Program Sistem Presensi dengan Pendeteksian Wajah | 2025 | Program Komputer | 000938105 |
| 2. | Kode Program Deteksi Jatuh pada Manusia Menggunakan Metode Threshold Berbasis Akselerometer pada Smartphone | 2023 | Program Komputer | EC00202376704 |
| 3. | Pelatihan Pembuatan Surat Lamaran Kerja Berbasis Digital Branding | 2023 | Jurnal | EC00202382852 |
| 4. | Pembekalan Ilmu Perhitungan Digital Berbasis Stand Alone Dan Cloud Melalui Excel Dan Spreadsheet | 2023 | Jurnal | EC00202380688 |
| 5. | Peningkatan Keahlian Dalam Kalkulasi Statistik Microsoft Excel Pada PKBM Dharma Putra Mandiri Tangerang | 2020 | Karya Tulis | 000178543 |

Jakarta, 30 Agustus 2025

Anggota Peneliti,

(Agnes Aryasanti, S.Kom., M.Kom.)

**Biodata Anggota 2**

Identitas Diri

1. Nama Lengkap : Putri Hayati, S.ST., M.Kom
2. Jenis Kelamin : Perempuan
3. Jabatan Fungsional : Asisten Ahli
4. NIP/NIDN/ID-SINTA : 170067/0308029101/ 6145520
5. Tempat, Tanggal Lahir : Lhokseumawe, 08 Februari 1991
6. E-mail : putri.hayati@budiluhur.ac.id
7. Nomor Handphone : 082362361515
8. Alamat : Jl. Shangrila III No. 82 B, Jakarta Selatan

Riwayat Pendidikan

| | **S1** | **S2** | **S3** |
|---|---|---|---|
| **Nama Perguruan Tinggi** | Politeknik Negeri Lhokseumawe | Universitas Budi Luhur | - |
| **Bidang Ilmu** | Teknik Informatika | Ilmu Komputer | - |
| **Tahun Masuk – Lulus** | 2009 - 2013 | 2015-2017 | - |

Pengalaman Penelitian (5 Tahun Terakhir)

| No | Tahun | Judul Penelitian | Pendanaan | |
|---|---|---|---|---|
| | | | Sumber* | Jumlah (Rp) |
| 1. | 2023 | Aplikasi Jaringan Augmented Reality (Ajar) Untuk Pengenalan Perangkat Dan Jaringan Komputer | Universitas Budi Luhur | 10.000.0000,- |
| 2. | 2018 | Intelligent Mobile Application untuk Kesehatan Mental Mahasiswa Universitas Budi Luhur Menggununakan Fuzzy Inference System | Universitas Budi Luhur | 15.000.000,- |
| 3. | 2017 | Penerapan Sistem IT Asset Management Dengan Penggabungan Metode Kamo dan Metode Model View Controller Dalam Upaya Meningkatkan Kualitas Layanan Dukungan IT Infrastruktur: Studi Kasus Direktorat Teknologi Informasi Kekayaan Intelektual | Universitas Budi Luhur | 15.000.000,- |

Publikasi Artikel Ilmiah dalam Jurnal (5 Tahun Terakhir)

| No | Tahun | Judul Artikel Ilmiah | Nama Jurnal | Volume/Nomor/ Tahun |
|---|---|---|---|---|
| 1. | 2025 | Penerapan E-Commerce Berbasis Content Management System Dengan Metode Waterfall Untuk Promosi Produk Digital | Jiki(Jurnal Ilmu Komputer Dan Informatika) | 06/06/2025 |
| 2. | 2025 | Sistem Informasi Kehadiran Sopir Pada Pt Blue Bird Tbk Berbasis Web | Idealis: Indonesia Journal Information Sistem | 01/01/2025 |
| 3. | 2025 | Rancang Bangun Web Service E-Budgeting Gereja Toraja Jemaat Bintaro Menggunakan Metode RESTful API Berbasis Web dan Android | SKANIKA: Sistem Komputer dan Teknik Informatika | 01/01/2025 |
| 4. | 2024 | Pemanfaatan Aplikasi Android untuk Penjadwalan Mata Pelajaran SMK Media Informatika | PengabdianMu: Jurnal Ilmiah Pengabdian kepada Masyarakat | 09/06/2024 |
| 5. | 2023 | Desain Aplikasi Pembelajaran Perangkat Jaringan Berbasis Augmented Reality Dengan Gamification Model Canvas | Fountain of Informatics Journal | 08/11/2023 |
| 6. | 2023 | Pelatihan Microsoft Word Bagi Siswa Pusat Kegiatan Belajar Masyarakat Negeri 11 Manggarai Jakarta Selatan | Literasi | 03/02/2023 |
| 7. | 2023 | Peningkatan Keterampilan Microsoft Word Dan Powerpoint Untuk Staf Administrasi Pada Gereja Orthodox Indonesia | KRESNA: Jurnal Riset dan Pengabdian Masyarakat | 03/01/2023 |

Pemakalah Seminar Ilmiah (5 Tahun Terakhir)

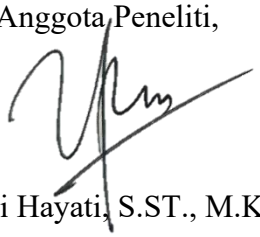
| No | Tahun | Nama Temu Ilmiah/ Seminar | Nama Jurnal | Waktu dan Tempat |
|---|---|---|---|---|
| | | | | |

Perolehan HKI (10 Tahun Terakhir)

| No | Judul HKI | Tahun | Jenis | No Pencatatan |
|---|---|---|---|---|
| 1. | Laporan Akhir Penelitian, Aplikasi Jaringan Augmented Reality (AJAR) Untuk Pengenalan Perangkat Dan Jaringan Komputer | 2024 | Laporan Penelitian | EC00202405959 |
| 2. | Buku Panduan Petunjuk Penggunaan Aplikasi Penjadwalan Mata Pelajaran Berbasis Android SMK Media Informatika Jakarta | 2023 | Buku Panduan | EC00202413492 |

Jakarta, 30 Agustus 2025

Anggota Peneliti,



(Putri Hayati, S.ST., M.Kom.)

## Lampiran 3. Surat Perjanjian Kontrak Penelitian

UNIVERSITAS BUDI LUHUR
Kampus Pusat : Jl. Raya Ciledug - Petukangan Utara - Jakarta Selatan 12260
Telp : 021-5853753 (hunting), Fax : 021-5853489, http://www.budiluhur.ac.id
FAKULTAS TEKNOLOGI INFORMASI
FAKULTAS EKONOMI DAN BISNIS
FAKULTAS ILMU SOSIAL DAN STUDI GLOBAL
FAKULTAS TEKNIK
FAKULTAS KOMUNIKASI DAN DESAIN KREATIF

**SURAT PERJANJIAN KONTRAK PENELITIAN**
Nomor A/UBL/DRPM/000/005/05/25

Pada hari ini, Jumat 16 Mei 2025 Semester Genap Tahun Ajaran 2024/2025, kami yang bertandatangan di bawah ini:

1. **Dr. Ir. Prudensius Maring, M.A.**, selaku Direktur Riset dan Pengabdian Kepada Masyarakat Universitas Budi Luhur, selanjutnya disebut PIHAK PERTAMA.
2. **Dr. Ir. Mardi Hardjianto, S.Kom., M.Kom.**, selaku Peneliti selanjutnya disebut PIHAK KEDUA.

Kedua belah pihak menyatakan bersepakat untuk membuat perjanjian kontrak penelitian sebagai berikut:

**Pasal 1**
**Judul Penelitian**

PIHAK PERTAMA dalam jabatannya tersebut di atas, memberikan tugas kepada PIHAK KEDUA untuk melaksanakan penelitian yang berjudul: Model Klasifikas Deteksi Situs Phishing Menggunakan Metode Machine Learning XGBOOST, Random Forest dan Feature Selection

**Pasal 2**
**Personalia Penelitian**

| | | |
|---|---|---|
| Peneliti Utama | : | Dr. Ir. Mardi Hardjianto, S.Kom., M.Kom. |
| Anggota Peneliti | : | Agnes Aryasanti, S.Kom., M.Kom. |
| | | Putri Hayati, S.ST., M.Kom |

**Pasal 3**
**Waktu dan Biaya Penelitian**

1. Waktu penelitian adalah 6 bulan, terhitung sejak tanggal 10 Maret 2025 sampai dengan 10 September 2025.
2. Biaya pelaksanaan penelitian ini dibebankan pada Yayasan Pendidikan Budi Luhur Cakti Tahun 2025 dengan nilai kontrak sebesar Rp 14,000,000.00 (empat belas juta rupiah)

**Pasal 4**
**Cara Pembayaran**

Pembayaran biaya penelitian diberikan secara bertahap, sebagai berikut:

1. Tahap pertama sebesar 50% dari nilai kontrak, setelah surat perjanjian kontrak penelitian ini ditandatangani oleh kedua belah pihak.
2. Tahap kedua sebesar 50% dari nilai kontrak, setelah PIHAK KEDUA menyerahkan Laporan Hasil Penelitian kepada PIHAK PERTAMA.

**Pasal 5**
**Keaslian Penelitian dan Ketidakterikatan dengan Pihak Lain**

1. PIHAK KEDUA bertanggungjawab atas keaslian judul penelitian sebagaimana disebutkan dalam Pasal 1 Surat Perjanjian Kontrak Penelitian ini (bukan duplikat/jiplakan/plagiat) dari penelitian orang lain.
2. PIHAK KEDUA menjamin bahwa judul penelitian tersebut bebas dari ikatan dengan pihak lain atau tidak sedang didanai oleh pihak lain.

3. PIHAK KEDUA menjamin bahwa judul penelitian tersebut bukan merupakan penelitian yang SEDANG ATAU SUDAH selesai dikerjakan, baik didanai oleh pihak lain maupun oleh sendiri.
4. PIHAK PERTAMA tidak bertanggungjawab terhadap tindakan plagiat yang dilakukan oleh PIHAK KEDUA.
5. Apabila dikemudian hari diketahui ketidakbenaran pernyataan ini, maka kontrak penelitian DINYATAKAN BATAL, dan PIHAK KEDUA wajib mengembalikan dana yang telah diterima kepada Yayasan Pendidikan Budi Luhur Cakti sebagai pemberi dana.

**Pasal 6**
**Monitoring Penelitian**

1. PIHAK PERTAMA berhak untuk:
   a. Melakukan pengawasan administrasi, monitoring, dan evaluasi terhadap pelaksanaan penelitian.
   b. Memberikan sanksi jika dalam pelaksanaan penelitian terjadi pelanggaran terhadap isi perjanjian oleh peneliti.
   c. Bentuk sanksi disesuaikan dengan tingkat pelanggaran yang dilakukan.
2. Pemantauan kemajuan penelitian dikoordinasikan oleh PIHAK PERTAMA.
3. Pelaksanaan kemajuan penelitian dilaksanakan pada tanggal 04 Juli 2025.
4. Format Laporan Kemajuan dan teknis pelaksanaannya diatur oleh PIHAK PERTAMA.

**Pasal 7**
**Laporan Akhir Penelitian**

PIHAK KEDUA wajib menyerahkan laporan akhir dalam bentuk softcopy, paling lambat tanggal 22 Agustus 2025.

**Pasal 8**
**Sanksi**

Segala kelalaian baik disengaja maupun tidak, sehingga menyebabkan keterlambatan menyerahkan laporan hasil penelitian dengan batas waktu yang telah ditentukan akan mendapatkan sanksi sebagai berikut:

1. Tidak diperbolehkan mengajukan usulan penelitian pada semester berikutnya bagi ketua dan anggota peneliti.
2. PIHAK KEDUA diberikan kesempatan perpanjangan waktu penelitian selama 2 (dua) minggu sampai dengan tanggal 05 September 2025.
3. Jika setelah masa perpanjangan tersebut PIHAK KEDUA tidak dapat menyelesaikan penelitiannya, PIHAK KEDUA diwajibkan mengembalikan dana yang sudah diterima kepada Yayasan Pendidikan Budi Luhur Cakti dengan cara mengembalikan tunai kepada PIHAK PERTAMA.

**Pasal 9**
**Penutup**

Perjanjian ini berlaku sejak ditandatangani dan disetujui oleh PIHAK PERTAMA dan PIHAK KEDUA.

Jakarta, 16 Mei 2025

| PIHAK PERTAMA | PIHAK KEDUA |
|---|---|
| METERAI TEMPEL DC8A5AMX318046225 | |
| Dr. Ir. Prudensius Maring, M.A. | Dr. Ir. Mardi Hardjianto, S.Kom., M.Kom. |
| NIP. 190043 | NIP. 920024 |

**Lampiran 4. Catatan Harian**

| No | Tanggal | Kegiatan |
|---|---|---|
| 1. | 10 -20 Maret 2025 | Mencari jurnal dan mempelajari tentang Deteksi Situs Phishing |
| 2. | 10 April 2025 | Pengumpulan data dan analisis |
| 3. | 15 Mei-28 Juni 2023 | Merancang Model Klasifikasi Deteksi Situs Phishing |
| 4. | 11 Juli -Agustus 2025 | Implementasi |
| 5. | 12 Agustus 2025 | Pembuatan Laporan Kemajuan |
| 6. | 15 Agustus 2025 | Pengujian |
| 7. | 21 Agustus 2025 | Pembuatan Laporan Penelitian |

**Lampiran 5. Artikel Ilmiah (draft/submitted/accepted/published)**

# Classification Model for Phishing Website Detection Using Machine Learning Methods: XGBoost, Random Forest, and Feature Selection

**Mardi Hardjianto*[1], Agnes Aryasanti[2], Putri Hayati[3]**

[1,3]Information Technology, Faculty of Information Technology, Budi Luhur University, Indonesia
[2]Information System, Faculty of Information Technology, Budi Luhur University, Indonesia

Email: [1]mardi.hardjianto@budiluhur.ac.id

Received : Jun 9, 2025; Revised : Nov 20, 2025; Accepted : Nov 22, 2025; Published : Dec 11, 2025
Phone Number : 081586603011

**Abstract**

The advancement of information technology has significantly influenced various aspects of life, offering convenience and comfort while also posing challenges such as the growing threat of cybercrime. One of the most prevalent threats is phishing, a cybercrime that exploits fake websites to steal sensitive information such as personal data, usernames, passwords, credit card numbers, and national identification numbers. Such data can later be used for extortion, fraud, or even sold to third parties for misuse. Phishing attacks have been recorded since 1996, and in Indonesia alone, 26,675 phishing cases were reported in the first quarter of 2023, a sharp increase compared to 6,106 cases in the fourth quarter of 2022. These attacks continue to evolve with increasingly sophisticated techniques, including domain name spoofing and social engineering, making accurate phishing detection an urgent necessity. This study aims to develop a phishing detection system using URL-based features with machine learning methods, namely XGBoost, Random Forest, and Feature Selection. Feature selection is applied to identify the most influential attributes in classification, thereby improving both efficiency and accuracy. The threat landscape has also shifted, with cybercriminals gaining easier access to malicious tools, ranging from malware rentals to advanced persistent threats (APT). APT attacks typically employ phishing and social engineering as initial entry points, followed by malware deployment and data theft, often leading to disclosure or trade on the dark web. In this research, the evaluation of XGBoost and Random Forest with Recursive Feature Elimination (RFE) optimization demonstrates that the XGBoost model using 30 features achieved the best performance, with an accuracy of 97.4%, precision of 98.5%, and recall of 95.5%.

***Keywords :*** *Phishing Website Detection; Machine Learning; XGBoost; Random Forest; Feature Selection; Cybersecurity; URL Features*



## 1. INTRODUCTION

The widespread adoption of the internet in Indonesia increased significantly after the COVID-19 pandemic, which limited physical interactions and accelerated the shift toward online activities through digital platforms. Advances in information and communication technology have transformed websites from simple one-way information channels into interactive platforms that support communication, collaboration, and online transactions (Makbull Rizki, 2022). These changes have made human activities in business, education, shopping, and social interaction more efficient and convenient.

Despite its benefits, technological progress has also created vulnerabilities that are exploited by cybercriminals. One of the most common and damaging methods is phishing, which deceives users by imitating legitimate websites in design, content, and domain URLs. Victims are tricked into disclosing sensitive data such as usernames, passwords, credit card numbers, or personal identification information, which attackers then use for fraud, extortion, or distribution of malware (Muftiadi, Agustina, and Evi, 2022; Putri and Wijayanto, 2022; BSSN, 2024).

The scale of phishing attacks continues to escalate globally. PhishLabs reported that phishing targeting social media platforms rose by more than 23.4% in the second quarter of 2023, ranking second only to the financial sector. Fraudulent domains are frequently registered under Country Code Top Level Domains (ccTLD) and Generic Top Level Domains (gTLD), with Poland, Colombia, and Indonesia among the most exploited ccTLDs (Ryan, 2023). In Indonesia, the ID-SIRTII/CC report recorded 26.7 million phishing-related anomalies in 2024, while Kaspersky reported that 91% of cyberattacks begin with phishing emails, with 32% of data breaches directly linked to phishing (Kusumawardhani, 2022).

These findings highlight the urgent need for accurate phishing detection systems to address the evolving threat landscape. To this end, machine learning algorithms have been widely applied, with XGBoost and Random Forest emerging as effective classification models. Feature selection techniques, such as Recursive Feature Elimination (RFE), are further employed to identify the most influential attributes, thereby improving efficiency and enhancing detection accuracy.

E-commerce platforms and online banking services have become primary targets of phishing attacks due to the high potential gains for perpetrators if successful. Popular sites such as Shopee and PayPal are often exploited, along with social media platforms including Facebook, X, TikTok, and Instagram, which are also frequent targets of cybercriminals. In addition to stealing data, phishing websites are also used to carry out various types of fraud impersonating official sites, as well as to serve as a medium for spreading malware or computer viruses. The following section presents a comparison between a phishing site and the legitimate website of the ShopeePay online financial service.

The increasing prevalence of phishing attacks has deceived many users, as phishing websites are often difficult to distinguish from legitimate ones due to similarities in content, domain names, and even HTML structures. Manual detection is not only inefficient but also highly prone to error, highlighting the need for an accurate detection model capable of determining whether a website is phishing or legitimate. To address this challenge, classification-based machine learning methods can be applied to detect phishing sites and links more effectively, thereby improving reliability and reducing user risk.

## 2. METHOD

Several approaches have been developed to detect or identify whether a website is legitimate or phishing. These techniques are generally grouped into four main categories: List-Based Methods, Search Engine-Based Methods, Visual Similarity-Based Methods, and Machine Learning-Based Methods (Jain et al., 2020; Ozker and Sahingoz, 2020; Baduwal et al., 2023).

This study adopts the Machine Learning-Based approach, which leverages machine learning techniques to identify characteristic patterns of phishing websites. The process involves training a model using a dataset composed of features extracted from both phishing and legitimate URLs. The model learns to distinguish attributes that differentiate the two classes, such as page layout, keywords, visual elements, and behavioral patterns. Once trained, the model can analyze new websites and classify them as either phishing or legitimate based on the patterns it has learned (Alam et al., 2020; Ozker and Sahingoz, 2020; Zamir et al., 2020).

In this research, Uniform Resource Locator (URL)-based features are used as the primary input for classification. Both phishing and non-phishing URLs are processed to extract their features, which are then used to construct the dataset for training and testing the classification models.

### 2.1. Research Roadmap

The five-year research roadmap is presented in Figure 1. The study is divided into four phases. In the first phase, trend analysis of cyberattacks such as malware, phishing, ransomware, and Advanced Persistent Threats (APT) will be conducted. Machine learning will be applied for attack classification and detection, as well as risk-based system security audits.

The second phase will focus on network and Internet of Things (IoT) device security. This includes the development of Artificial Intelligence (AI)-based Intrusion Detection Systems (IDS), secure communication protocols for IoT, and security testing of devices such as smart homes and smart grids.

The third phase emphasizes personal data protection and authentication systems, covering biometric and multi-factor authentication (MFA), privacy-preserving technologies such as homomorphic encryption and differential privacy, and blockchain-based digital identity. The fourth phase will concentrate on AI security and threats targeting AI systems. This includes adversarial attacks on machine learning models, security for Large Language Model (LLM)-based and recommendation systems, as well as the development of trustworthy and explainable AI (XAI).

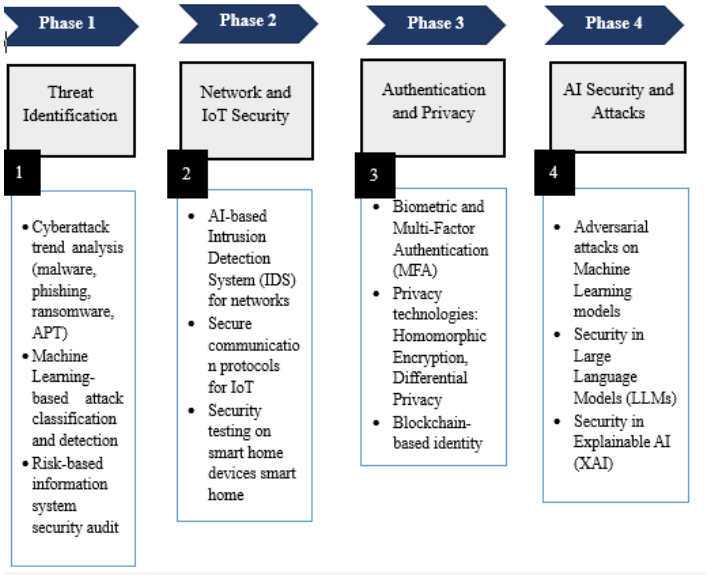



Figure 1. Research Roadmap

## 2.2. Reseacrh Process

The research steps used to illustrate the general sequence of activities are presented in Figure 2.

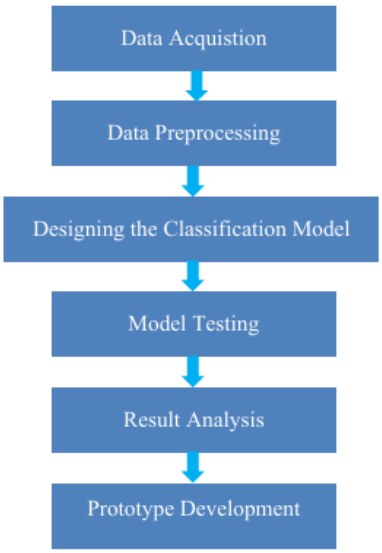



Figure 2. Research Process

In this study, the researchers utilized a public dataset sourced from the GitHub *Phishing Database* repository available https://github.com/Phishing-Database/Phishing.Database. *Reposit.* This repository is sponsored by PyFunceble, a tool used to check whether a website is valid, active, or inactive.

The dataset used in this study consists of three main categories: phishing (label: -1), neutral/ambiguous (label: 0), and non-phishing (label: 1). The data includes various features related to URL characteristics, IP addresses, and other technical attributes relevant for training machine learning models in phishing website classification. When a phishing website is accessed through a browser, its display appears similar to the example shown in Figure 3.

The purpose of utilizing this dataset is to examine the patterns and common characteristics of phishing websites, as well as to develop more effective detection approaches. By identifying the key features that serve as phishing indicators, this study aims to contribute to strengthening detection systems and preventing phishing attacks as part of broader cybersecurity enhancement efforts.

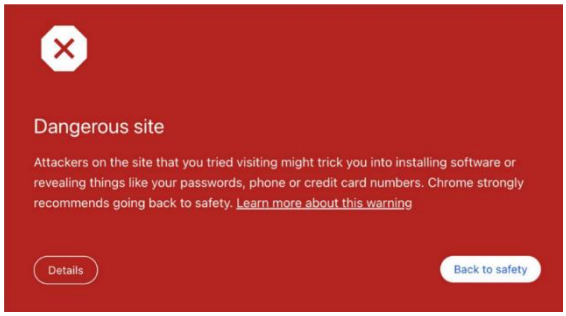



Figure 3. Display Accessing a *Phising Website*

Table 1. *Phising Website*

| No | Situs *Phishing* |
|---|---|
| 1 | http://24e72176-d57a-4cdf-8d12-27b3d3ca13ce-00-35ep2mx3gzoxb.sisko.replit.dev |
| 2 | http://abuniz123.github.io/metamanager |
| 3 | http://ad-manenger-bill-prsj.vercel.app/authenticate/business.helpfacebook.com-help_159551727005616.html |
| 4 | http://algobuy.shop |
| 5 | http://ann.spacemotors.top |
| 6 | http://ayaanwebdeveloper.github.io/fb-clone |
| 7 | http://bbplc-ee.weeblysite.com |
| 8 | http://bt-103312.weeblysite.com |
| 9 | http://btinternet-106995.weeblysite.com |
| 10 | http://btinterrnett-104587.weeblysite.com |
| 11 | http://communication-ergo-additional-their.vercel.app/facebook.html |
| 12 | http://contactinstant.vercel.app/contact/us.html |
| 13 | http://eanop1.tahory.site |
| 14 | http://event.playaspirant.com |
| 15 | http://facebook-peach-chi.vercel.app |
| 16 | http://facebook-peach-chi.vercel.app/auth |
| 17 | http://facebookv01.vercel.app |
| 18 | http://fagebook-login.vercel.app |
| 19 | http://fkl-32655.bubbleapps.io |
| 20 | http://grounded-absolute-123701.framer.app |
| 21 | http://imasmonline.durablesites.com |

| | |
|---|---|
| 22 | http://insta-pro.pk |
| 23 | http://ledger-aktualisieren.pages.dev |
| 24 | http://litv.oilcompanyca.top |
| 25 | http://location-mauve-six.vercel.app |
| 26 | http://nnt-flame.vercel.app/metabusinesscenter.html |
| 27 | http://p95xg.com |
| 28 | http://pasxfuul.com |
| 29 | https://0000006738.vercel.app/sec.html?error=interaction_required&erro |
| 30 | https://2025-local.click |

## 3. RESULT

The preprocessing stage is the initial step following data acquisition. This stage involves extracting the feature values from each URL. As shown in Table 2, the dataset contains 31 features along with descriptions of their respective variables or attributes. Of these, 30 are independent features, while 1 serves as the dependent feature, which acts as the label to classify URLs as Phishing (1) or Non-*Phishing* (-1). The label values for classifying Phishing or Non-Phishing URLs correspond to the actual conditions of the URLs in the dataset. Each URL is assigned values according to its features.

Table 2. Dataset Features

| Index | | Nilai | | |
|---|---|---|---|---|
| | | Legitimate | Suspicious | Phishing |
| 1 | UsingIP | -1 | | 1 |
| 2 | LongURL | -1 | 0 | 1 |
| 3 | ShortURL | -1 | | 1 |
| 4 | Symbol@ | -1 | | 1 |
| 5 | Redirecting// | -1 | | 1 |
| 6 | PrefixSuffix- | -1 | | 1 |
| 7 | SubDomains | -1 | 0 | 1 |
| 8 | HTTPS | -1 | 0 | 1 |
| 9 | DomainRegLen | -1 | | 1 |
| 10 | Favicon | -1 | | 1 |
| 11 | NonStdPort | -1 | | 1 |
| 12 | HTTPSDomainURL | -1 | | 1 |
| 13 | RequestURL | -1 | | 1 |
| 14 | AnchorURL | -1 | 0 | 1 |
| 15 | LinksInScriptTags | -1 | 0 | 1 |
| 16 | ServerFormHandler | -1 | 0 | 1 |
| 17 | InfoEmail | -1 | | 1 |
| 18 | AbnormalURL | -1 | | 1 |
| 19 | WebsiteForwarding | -1 | 0 | 1 |
| 20 | StatusBarCust | -1 | | 1 |
| 21 | DisableRightClick | -1 | | 1 |
| 22 | UsingPopupWindow | -1 | | 1 |
| 23 | IframeRedirection | -1 | | 1 |
| 24 | AgeofDomain | -1 | | 1 |
| 25 | DNSRecording | -1 | | 1 |
| 26 | WebsiteTraffic | -1 | | 1 |
| 27 | PageRank | -1 | | 1 |
| 28 | GoogleIndex | -1 | | 1 |
| 29 | LinksPointingToPage | -1 | | 1 |
| 30 | StatsReport | -1 | | 1 |
| 31 | Class | -1 | | 1 |

After the URLs are assigned values, each feature may contain a value of (1) indicating phishing, (0) indicating suspicious, and (-1) indicating non-phishing, in order to classify incoming URLs as potential phishing, suspicious, or non-phishing. An example of a dataset with values assigned to the 30 features is presented in Table 3.

Table 4. Phishing Detection Datasets

| Using IP | LongURL | ShortURL | Symbol@ | Redirecting// | PrefixSuffix- | Subdomains | HTTPS | DomainRegLen | Favicon | NonStdPort | HTTPSDomainURL | RequestURL | AnchorURL | LinksInScriptTags | ServerFormHandler | InfoEmail | AbnormalURL | WebsiteForwarding | StatusbarCust | DisableRightClick | UsingPopupWindow | IframeRedirection | AgeofDomain | DNSRecording | WebsiteTraffic | PageRank | GoogleIndex | LinkPointingToPage | StatsReport | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | -1 | 0 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | -1 |
| 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | 0 | -1 | -1 |
| 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | 1 | 0 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 |
| 1 | 0 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 0 | -1 | 1 | 1 | 0 | -1 | 1 | -1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | 1 |
| -1 | 0 | -1 | 1 | -1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 0 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 |
| 1 | 0 | -1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 0 | -1 | -1 |
| 1 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 0 | 1 | -1 |
| 1 | 0 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | -1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 0 | 1 | -1 |
| 1 | 1 | 1 | 1 | 1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | 0 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 |
| 1 | 1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 0 | -1 | -1 |
| -1 | 1 | -1 | 1 | -1 | -1 | 0 | 0 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | -1 | 1 | 1 | 0 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 | 1 | 0 | 1 | -1 |
| 1 | 1 | -1 | 1 | 1 | -1 | 0 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | -1 |
| 1 | 1 | -1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 |
| 1 | -1 | -1 | -1 | 1 | -1 | 0 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 0 | 1 | -1 |
| 1 | -1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 1 | -1 | -1 |
| 1 | -1 | 1 | 1 | 1 | -1 | -1 | 0 | 1 | 1 | -1 | 1 | 1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | -1 |
| 1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | -1 | -1 | 1 |
| 1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 0 | -1 | -1 | -1 | 0 | -1 | -1 | -1 | -1 | 1 | -1 | 0 | -1 | 1 | 0 | -1 | 1 |
| 1 | 0 | -1 | 1 | 1 | -1 | 0 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 0 | -1 | -1 | -1 | 0 | -1 | 1 | -1 | 1 | -1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 |
| 1 | 0 | 1 | 1 | 1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 | 0 | -1 | 1 |
| 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 1 | -1 | 1 | 0 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | -1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | -1 | 1 | 0 | -1 | 1 | 1 | 1 | 1 | 0 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | 1 |
| 1 | -1 | -1 | -1 | 1 | -1 | 1 | 1 | -1 | 1 | 1 | -1 | -1 | 0 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | 0 | 1 | 1 |
| 1 | -1 | 1 | 1 | 1 | -1 | 0 | 1 | -1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | -1 | 1 | 0 | 1 | 1 |
| 1 | -1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | -1 | -1 | -1 |
| 1 | -1 | -1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | 0 | -1 | -1 | -1 | 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 0 | -1 | 1 |
| 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | 1 |
| 1 | -1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |

Based on the previously designed model, the following section presents its implementation and evaluation results. *Data Acquisition*, the data used in this study consist of URLs along with the values of their respective features. A total of 11,056 datasets were utilized. An example of the data employed in this research is presented in Table 4. From the overall dataset, 80% was allocated for training data and 20% for testing data.

Table 4. Training Dataset

| Feature Values | Class |
|---|---|
| -1,1,1,1,-1,-1,-1,-1,-1,1,1,-1,1,-1,1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,-1,1,1,-1 | Non-Phishing |
| 1,1,1,1,1,-1,0,1,-1,1,1,-1,1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,1,1 | Non-Phishing |
| 1,0,1,1,1,-1,-1,-1,-1,1,1,-1,1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,1,-1,1,0,-1 | Non-Phishing |
| 1,0,1,1,1,-1,-1,-1,1,1,1,-1,-1,0,0,-1,1,1,0,1,1,1,1,-1,-1,1,-1,1,-1,1 | Non-Phishing |
| 1,0,-1,1,1,-1,1,1,-1,1,1,1,1,0,0,-1,1,1,0,-1,1,-1,1,-1,-1,0,-1,1,1,1 | Phishing |
| -1,0,-1,1,-1,-1,1,1,-1,1,1,-1,1,0,0,-1,-1,-1,0,1,1,1,1,1,1,1,-1,1,-1,-1 | Phishing |
| 1,0,-1,1,1,-1,-1,-1,1,1,1,1,1,-1,-1,0,-1,-1,-1,0,1,1,1,1,1,-1,-1,-1,1,0,-1 | Non-Phishing |
| 1,0,1,1,1,-1,-1,-1,1,1,1,1,-1,-1,0,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,0,1 | Non-Phishing |
| 1,0,-1,1,1,-1,1,1,-1,1,1,-1,1,0,1,-1,1,1,0,1,1,1,1,1,-1,1,1,1,0,1 | Phishing |
| 1,1,-1,1,1,-1,-1,1,-1,1,1,1,1,1,0,1,-1,1,1,0,1,1,1,1,1,-1,0,-1,1,0,1 | Non-Phishing |

| | |
|---|---|
| 1,1,1,1,1,-1,0,1,1,1,1,1,-1,0,0,-1,-1,-1,0,1,1,1,1,-1,1,1,1,1,-1,-1 | Phishing |
| 1,1,-1,1,1,-1,1,-1,-1,1,1,1,1,-1,-1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,0,-1 | Non-Phishing |
| -1,1,-1,1,-1,-1,0,0,1,1,1,-1,-1,-1,1,-1,1,1,0,-1,1,-1,1,1,-1,-1,-1,1,0,1 | Non-Phishing |
| 1,1,-1,1,1,-1,0,-1,1,1,1,1,-1,-1,-1,-1,1,1,0,1,1,1,1,-1,-1,0,-1,1,1,1 | Non-Phishing |
| 1,1,-1,1,1,1,-1,1,-1,1,1,-1,1,0,1,1,1,1,0,1,1,1,1,1,-1,1,-1,1,-1,1 | Phishing |
| 1,-1,-1,-1,-1,1,-1,0,0,1,1,1,1,-1,-1,0,-1,1,1,0,1,1,1,1,1,-1,-1,-1,1,0,1 | Non-Phishing |
| 1,-1,-1,1,1,-1,1,1,-1,1,1,-1,1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,0,-1,1,1,-1 | Non-Phishing |
| 1,-1,1,1,1,-1,-1,0,1,1,-1,1,1,0,-1,-1,-1,-1,0,1,1,1,1,-1,1,1,-1,1,1,-1 | Non-Phishing |
| 1,1,1,1,1,-1,-1,1,1,1,1,-1,-1,0,-1,-1,-1,-1,0,1,1,1,1,1,-1,-1,1,1,-1,-1 | Phishing |
| 1,1,1,1,1,-1,-1,1,-1,1,1,1,1,0,0,-1,-1,-1,0,-1,-1,-1,-1,1,-1,0,-1,1,0,-1 | Phishing |
| 1,0,-1,1,1,-1,0,1,-1,1,1,1,1,0,0,-1,-1,-1,0,-1,1,-1,1,-1,1,1,-1,1,-1,-1 | Phishing |
| 1,0,1,1,1,-1,0,1,1,1,1,-1,-1,0,-1,-1,-1,-1,0,1,1,1,1,-1,1,-1,-1,1,0,-1 | Phishing |
| 1,1,1,1,1,-1,-1,-1,-1,1,1,-1,1,0,0,-1,1,1,0,1,1,1,1,1,1,0,-1,1,-1,1 | Phishing |
| 1,1,1,1,1,-1,1,0,-1,1,1,1,1,0,0,-1,1,1,0,1,1,1,1,1,1,1,-1,1,-1,1 | Phishing |
| 1,-1,-1,-1,1,-1,1,1,-1,1,1,-1,-1,0,0,-1,1,1,0,1,1,1,1,1,1,-1,-1,1,0,1 | Phishing |
| 1,-1,1,1,1,-1,0,1,-1,1,1,1,1,1,0,-1,1,1,0,1,1,1,1,-1,1,1,-1,1,0,1 | Phishing |
| 1,-1,1,1,1,-1,0,-1,1,1,1,-1,-1,-1,-1,-1,-1,-1,0,1,1,1,1,1,1,0,-1,1,-1,-1 | Non-Phishing |
| 1,-1,-1,1,1,1,-1,1,1,1,1,1,-1,1,0,-1,-1,-1,0,1,1,1,1,1,-1,0,-1,1,0,-1 | Phishing |
| 1,-1,-1,1,-1,1,-1,1,-1,1,1,1,1,1,1,0,-1,1,1,1,1,1,1,1,-1,-1,1,-1,1,-1,1 | Phishing |
| 1,-1,1,1,1,-1,-1,1,-1,1,1,1,1,1,1,0,-1,1,1,0,1,1,1,1,-1,1,1,1,1,0,1 | Phishing |

## 4. DISCUSSIONS

The developed phishing website detection model needs to be tested to determine whether it performs as expected. The testing process is carried out using several URLs, both phishing and non-phishing. The test produces classifications of the input URLs as either phishing or non-phishing. The phishing detection model also requires evaluation. This evaluation is presented through the accuracy value obtained from the confusion matrix, as expressed in Equation (1).

There are four possible conditions in this measurement. The first condition occurs when the URL is a phishing site and the system correctly reports it as phishing. This case is referred to as True Positive (TP). The second condition occurs when the URL is a phishing site but the system incorrectly reports it as non-phishing. This is referred to as False Negative (FN). The third condition occurs when the URL is a non-phishing site and the system correctly reports it as non-phishing, which is referred to as True Negative (TN). The fourth condition occurs when the URL is a non-phishing site but the system incorrectly reports it as phishing, which is referred to as False Positive (FP). These four conditions are presented in Table 5.

$$Akurasi = \frac{TP + TN}{TP + FN + TN + FP} \; X \; 100\% \qquad (4)$$

Tabel 15 Confusion Matrix

| | | Actual Value | |
|---|---|---|---|
| | | Phishing | Non-Phishing |
| Nilai Prediksi | Phishing | True Positive (TP) | False Positive (FP) |
| | Non-Phishing | False Negative (FN) | True Negative (TN) |

## 5. CONCLUSION

Based on the results of the study on phishing detection through URLs using the XGBoost method and the testing produced an accuracy of xx%. The analysis of the conducted experiments indicates that this system is capable of detecting whether a URL is phishing or non-phishing.The conclusion is the essence of the entire paper. Made in paragraph form, and not in list form.

## REFERENCES


[1] Alam, M.N. *et al.* (2020) 'Phishing attacks detection using machine learning approach', *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, (Icssit), pp. 1173–1179. Available at: https://doi.org/10.1109/ICSSIT48917.2020.9214225.

[2] Baduwal, M. *et al.* (2023) 'Survey On Machine Learning Paradigms For Phishing Website Detection', *Article in International Journal of Engineering & Technology* [Preprint], (June). Available at: https://www.researchgate.net/publication/370658818.

[3] BSSN (2024) *Lanskap Keamanan Siber Indonesia*. Available at: https://www.bssn.go.id/wp-content/uploads/2025/02/LANSKAP-KEAMANAN-SIBER-2024-1.pdf (Accessed: 15 April 2025).

[4] Deekshitha, B. (2022) 'URL Based Phishing Website Detection by Using Gradient and Catboost Algorithms', *International Journal for Research in Applied Science and Engineering Technology*, 10(6), pp. 3717–3722. Available at: https://doi.org/10.22214/ijraset.2022.43986.

[5] Jain, A.K. *et al.* (2020) 'PhishSKaPe: A Content based Approach to Escape Phishing Attacks', *Procedia Computer Science*, 171(2019), pp. 1102–1109. Available at: https://doi.org/10.1016/j.procs.2020.04.118.

[6] Kusumawardhani, N.Q. (2022) *91 Persen Serangan Siber Berawal dari Kerentanan Ini*. Available at: https://tekno.republika.co.id/berita/rejeqc368/91-persen-serangan-siber-berawal-dari-kerentanan-ini (Accessed: 15 April 2025).

[7] Mahmud, A.F. and Wirawan, S. (2024) 'Deteksi Phishing Website menggunakan Machine Learning Metode Klasifikasi', 13(4), pp. 2540–9719. Available at: http://sistemasi.ftik.unisi.ac.id.

[8] Makbull Rizki (2022) 'Perkembangan Sistem Pertahanan/Keamanan Siber Indonesia dalam Menghadapi Tantangan Perkembangan Teknologi dan Informasi', *Politeia: Jurnal Ilmu Politik*, 14(1), pp. 54–62. Available at: https://doi.org/10.32734/politeia.v14i1.6351.

[9] Muftiadi, A., Agustina, T.P.M. and Evi, M. (2022) 'Studi kasus keamanan jaringan komputer: analisis ancaman phising terhadap layanan online banking', *Hexatech: Jurnal*

*Ilmiah Teknik*, 1(2), pp. 60–65. Available at: https://doi.org/10.55904/hexatech.v1i2.346.

[10] Musa, H. *et al.* (2019) 'A comparative analysis of phishing website detection using XGBOOST algorithm', *Journal of Theoretical and Applied Information Technology*, 97(5), pp. 1434–1443.

[11] Ozker, U. and Sahingoz, O.K. (2020) 'Content Based Phishing Detection with Machine Learning', in *2020 International Conference on Electrical Engineering, ICEE 2020*. Istambul. Available at: https://doi.org/10.1109/ICEE49691.2020.9249892.

[12] Putri, N.B. and Wijayanto, A.W. (2022) 'Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing', *Komputika : Jurnal Sistem Komputer*, 11(1), pp. 59–66. Available at: https://doi.org/10.34010/komputika.v11i1.4350.

[13] Ryan, J. (2023) *Phishing Sites Impersonating Social Media Jump in Q2*. Available at: https://www.phishlabs.com/blog/phishing-sites-impersonating-social-media-jump-in-q2 (Accessed: 17 April 2025).

[14] Saputro, I.A., Sugiarto, L. and Nugraha, F.S. (2024) 'ANALISIS KESADARAN MASYARAKAT TERHADAP BAHAYA INTERNET PHISHING MENGGUNAKAN K-MEANS CLUSTERING', *Jurnal Satuan Tulisan Riset dan Inovasi Teknologi*, 9(2), pp. 139–146.

[15] Windarni, V.A. *et al.* (2023) 'Deteksi Website Phishing Menggunakan Teknik Filter Pada Model Machine Learning', *Information System Journal*, 6(01), pp. 39–43. Available at: https://doi.org/10.24076/infosjournal.2023v6i01.1268.

[16] Zamir, A. *et al.* (2020) 'Phishing web site detection using diverse machine learning algorithms', *Electronic Library*, 38(1), pp. 65–80. Available at: https://doi.org/10.1108/EL-05-2019-0118.

## Lampiran 6. HKI

**PERMOHONAN PENDAFTARAN CIPTAAN**

I. Pencipta :

| | | |
|---|---|---|
| 1. Nama | : | Agnes Aryasanti |
| 2. Kewarganegaraan | : | Indonesia |
| 3. Alamat lengkap | : | Kp. Dukuh RT. 001 RW.007 Sudimara Selatan, Ciledug 15151 dan kode pos |
| 4. No. HP & E-mail | : | 085945766802/agnes.aryasanti@budiluhur.ac.id |
| 1. Nama | : | Mardi Hardjianto |
| 2. Kewarganegaraan | : | Indonesia |
| 3. Alamat lengkap dan kode pos | : | Jl. Mekarsari Raya Blok H/7 Cimanggis, Depok 16452 |
| 4. No. HP & E-mail | : | 081586603011/ mardi.mardjianto@budiluhur.ac.id |
| 1. Nama | : | Putri Hayati |
| 2. Kewarganegaraan | : | Indonesia |
| 3. Alamat lengkap dan kode pos | : | Jl. Lilawangsa Sumur Bor Desa Simpang Empat Kecamatan Banda Sakti, Lhokseumawe 24351 |
| 4. No. HP & E-mail | : | 082362361515/putri.hayati@budiluhur.ac.id |

II. Jenis dari judul ciptaan yang dimohonkan : Model Klasifikasi Deteksi Situs Phishing

III. Tanggal dan tempat di-umumkan untuk pertama kali di wilayah Indonesia atau di luar wilayah Indo-nesia :

IV Deskripsi singkat ciptaan : Model Klasifikasi Deteksi Situs Phishing

Jakarta, 30 Agustus 2025

Tanda Tangan :
Nama Lengkap : Mardi Hardjianto

Tanda Tangan :
Nama Lengkap : Agnes Aryasanti

Tanda Tangan :
Nama Lengkap : Putri Hayati