Optimasi Hybrid SVM-Naïve Bayes, Weighted Voting Deteksi Phishing URL, Seleksi Fitur-Explainable AI

Arizal Sabila Nurhikam[1], Ifani Haryanti[2], Fitri Khoirunnisa[3]
[1]Program Studi Teknik Informatika, Universitas Adhirajasa Reswara Sanjaya
[2,3]Program Studi Teknik Informatika, Universitas Adhirajasa Reswara Sanjaya
e-mail: [1]arizalsabilanurhikam@gmail.com, [2]ifani@ars.ac.id, [3]fitri.khoirunnisa@ars.ac.id

Abstrak

Serangan *phishing* merupakan ancaman besar dalam keamanan siber yang memanfaatkan rekayasa sosial untuk memperoleh informasi sensitif seperti kredensial login atau data perbankan. Salah satu upaya penanggulangannya yang efektif adalah deteksi URL phishing berbasis *machine learning*, yang dapat mengidentifikasi pola berbahaya secara otomatis. Penelitian ini mengusulkan model *hybrid* dengan menggabungkan algoritma *Support Vector Machine* (SVM) dan Naïve Bayes melalui pendekatan *Weighted Voting* untuk meningkatkan akurasi deteksi dan kestabilan model. Untuk mengurangi *overfitting* dan meningkatkan efisiensi klasifikasi, dilakukan seleksi fitur menggunakan SelectKBest, *Recursive Feature Elimination* (RFE), dan *Principal Component Analysis* (PCA). Ketiga metode dibandingkan untuk menilai kontribusinya terhadap performa model. Dalam rangka meningkatkan transparansi dan interpretabilitas, pendekatan *Explainable* AI (XAI) seperti SHAP dan LIME digunakan untuk menganalisis kontribusi fitur terhadap hasil prediksi. Performa model dievaluasi berdasarkan metrik akurasi, presisi, *recall*, dan F1-*score*. Hasil eksperimen menunjukkan bahwa model *hybrid* dengan fitur optimal mampu mencapai akurasi tinggi dan menghasilkan interpretasi prediktif yang informatif. Kombinasi pendekatan algoritmik dan interpretatif ini menawarkan solusi yang akurat sekaligus dapat dijelaskan, sehingga cocok diterapkan dalam sistem deteksi *phishing* yang efektif. Temuan ini diharapkan berkontribusi pada pengembangan sistem keamanan berbasis machine learning yang transparan dan efisien untuk menghadapi ancaman phishing di dunia nyata.

Kata kunci— *Phishing* URL, *Support Vector Machine* (SVM), Naïve Bayes, *Feature Selection, Explainable* AI.

Abstract

Phishing attacks are a major cybersecurity threat, exploiting social engineering to obtain sensitive information such as login credentials or banking data. An effective countermeasure is phishing URL detection using machine learning, which identifies malicious URL patterns automatically. This study proposes a hybrid model combining Support Vector Machine (SVM) and Naïve Bayes via a Weighted Voting approach to improve detection accuracy and model stability. To reduce overfitting and boost classification efficiency, feature selection is applied using SelectKBest, Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA). These methods are compared to evaluate their impact on model performance. For transparency and interpretability, Explainable AI (XAI) techniques—SHAP and LIME—are used to analyze feature contributions to prediction results. Model performance is assessed using accuracy, precision, recall, and F1-score metrics. Experimental results show the hybrid model with optimal features achieves high accuracy and produces informative interpretations. This combination of algorithmic and interpretive approaches provides a solution that is both accurate and explainable, making it suitable for implementation in effective phishing detection systems. The findings are expected to contribute to the development of transparent and efficient machine learning-based cybersecurity systems to address real-world phishing threats.

Keywords—*Phishing URL, Support Vector Machine (SVM), Naïve Bayes, Feature Selection, Explainable AI.*

Corresponding Author:
Ifani Haryanti,
Email: ifani@ars.ac.i

1. PENDAHULUAN

Internet telah menjadi infrastruktur utama dalam menunjang aktivitas manusia modern, mulai dari komunikasi, transaksi finansial, hingga pertukaran informasi. Pemanfaatannya telah meluas ke berbagai sektor seperti perdagangan elektronik, transportasi, layanan kesehatan, administrasi pemerintahan digital, hingga industri keuangan. Di Indonesia, tingkat penetrasi internet terus meningkat signifikan, dengan 73,7% populasi menjadi pengguna aktif pada tahun 2019, naik 8,9% dibandingkan tahun sebelumnya [1]. Namun, di balik manfaatnya, penggunaan internet yang bersifat terbuka juga memunculkan risiko serius, salah satunya adalah serangan *phishing*.

Phishing merupakan bentuk penipuan siber yang dirancang untuk mencuri informasi sensitif seperti kredensial akun dan data keuangan [2]. Teknik ini umumnya dilakukan dengan meniru tampilan situs resmi agar meyakinkan pengguna [3]. Menurut laporan *Anti-Phishing Working Group* (2021), halaman phishing yang menargetkan sektor keuangan, *e-commerce*, dan media sosial hanya dalam kuartal pertama, dan jumlah ini diprediksi terus meningkat. Ancaman ini menunjukkan perlunya sistem deteksi *phishing* yang adaptif, akurat, dan mampu memberikan peringatan dini.

Pendekatan berbasis *machine learning* telah menjadi salah satu solusi untuk mendeteksi *URL phishing* secara otomatis [4]. Namun, tantangan masih muncul, antara lain tingkat akurasi yang belum optimal pada data kompleks, risiko *overfitting* akibat penggunaan fitur berlebihan, serta keterbatasan interpretabilitas model. Oleh karena itu, diperlukan strategi yang mampu mengombinasikan kekuatan beberapa algoritma melalui pendekatan *ensemble*, didukung oleh seleksi fitur yang efektif dan pendekatan *Explainable AI* (XAI) agar hasil prediksi dapat dipahami oleh pengguna.

Penelitian ini mengembangkan model deteksi *phishing* URL menggunakan kombinasi algoritma *Support Vector Machine* (SVM) dan *Naïve Bayes* dalam pendekatan *hybrid* berbasis *Weighted Voting* [5]. Seleksi fitur diterapkan menggunakan tiga metode, yakni *SelectKBest*, *Recursive Feature Elimination* (RFE), dan *Principal Component Analysis* (PCA), guna mengurangi *overfitting* dan memperkuat generalisasi model. Sebagai upaya peningkatan transparansi, pendekatan *Explainable AI* diintegrasikan dengan menggunakan SHAP dan LIME untuk memberikan penjelasan terhadap hasil prediksi . Penelitian ini bertujuan tidak hanya untuk meningkatkan akurasi deteksi *phishing*, tetapi juga memastikan bahwa proses klasifikasi dapat dijelaskan secara logis dan informatif kepada pengguna akhir.

Kebaruan dari penelitian ini terletak pada integrasi tiga komponen utama—model *hybrid*, seleksi fitur berlapis, dan interpretasi berbasis XAI—dalam satu kerangka sistem deteksi *phishing URL*. Pendekatan ini diharapkan mampu menjawab tantangan praktis dalam keamanan siber dan memberikan kontribusi nyata dalam pengembangan sistem deteksi *phishing* yang andal, transparan, dan mudah diterapkan di lingkungan nyata..

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif yang bertujuan untuk menguji performa model klasifikasi *URL phishing* secara numerik melalui evaluasi statistik [6]. Model dikembangkan menggunakan pendekatan *hybrid machine learning*, yaitu kombinasi dari algoritma *Support Vector Machine* (SVM) dan *Naïve Bayes* yang digabungkan melalui teknik *Weighted Voting* [7]. Rancangan penelitian disusun secara eksperimental untuk mengamati pengaruh kombinasi algoritma, seleksi fitur, dan pendekatan *Explainable AI* terhadap akurasi dan interpretabilitas model [8].

Data yang digunakan dalam penelitian ini bersifat sekunder dan diperoleh dari sumber terbuka, yaitu UCI *Machine Learning Repository* [9]. Dataset terdiri dari sejumlah URL yang telah diklasifikasikan sebagai *phishing* atau *legitimate*. Masing-masing URL direpresentasikan oleh fitur numerik yang menggambarkan karakteristik teknis URL, seperti panjang karakter, jumlah simbol khusus, struktur domain, dan penggunaan protokol HTTPS. Data telah mengalami transformasi awal dari bentuk mentah ke bentuk terstruktur, sehingga siap digunakan dalam proses analisis dan klasifikasi.

Prosedur penelitian diawali dengan proses pra-pemrosesan data, yang mencakup normalisasi, penghapusan nilai kosong, serta transformasi fitur agar sesuai dengan kebutuhan algoritma. Tahap berikutnya adalah seleksi fitur untuk mereduksi dimensi dan menghindari *overfitting*. Tiga metode seleksi yang digunakan adalah *SelectKBest*, *Recursive Feature Elimination* (RFE), dan *Principal Component Analysis* (PCA). Pemilihan metode ini bertujuan untuk mempertahankan fitur-fitur paling relevan yang berkontribusi terhadap deteksi *phishing URL*.

Model klasifikasi dibangun dengan menggunakan dua algoritma utama. *Naïve Bayes*, yang berbasis probabilistik dengan asumsi independensi fitur [10]. *Support Vector Machine*, yang bekerja dengan mencari *hyperplane* optimal untuk memisahkan dua kelas [11]. Keduanya digabungkan dalam kerangka *Weighted Voting*, di mana setiap model memberikan kontribusi berdasarkan bobot akurasi yang dimilikinya [12]. Strategi ini digunakan untuk memperoleh hasil klasifikasi yang lebih stabil dan akurat [13].

Evaluasi kinerja dilakukan menggunakan metrik akurasi, presisi, *recall*, *F1-score*, dan ROC-AUC untuk memberikan gambaran menyeluruh terhadap performa model. Tahapan eksperimen dilakukan secara berulang untuk membandingkan hasil dari masing-masing metode seleksi fitur dan kombinasi model yang digunakan.

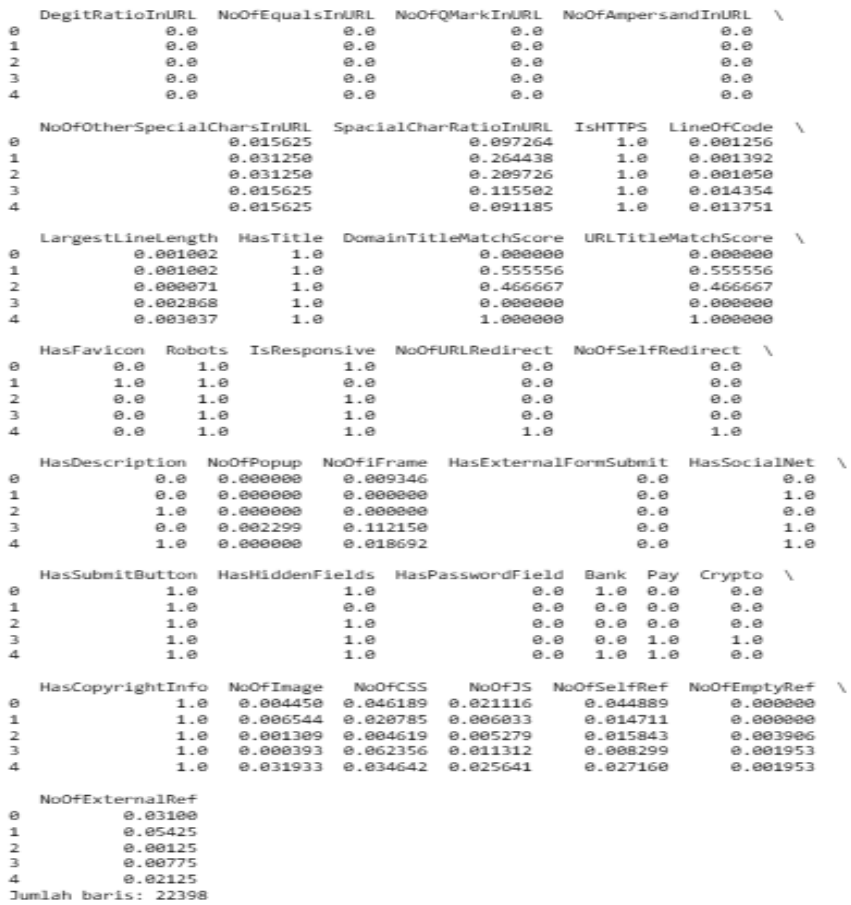
Untuk meningkatkan interpretabilitas, penelitian ini menerapkan pendekatan *Explainable Artificial Intelligence* (XAI) menggunakan SHAP (*SHapley Additive Explanations*) dan LIME (*Local Interpretable Model-Agnostic Explanations*). SHAP digunakan untuk memberikan pemahaman terhadap kontribusi fitur secara global dan local [14], sementara LIME digunakan untuk menjelaskan prediksi individual secara intuitif [15]. Dengan penerapan XAI, hasil prediksi model tidak hanya dapat diandalkan secara akurasi, tetapi juga dapat dijelaskan kepada pengguna, sehingga meningkatkan transparansi dan kepercayaan terhadap sistem deteksi *phishing* yang dikembangkan.

2.1. Tahapan Review

Harap Penelitian ini menyajikan hasil implementasi dan evaluasi dari metode yang telah dijelaskan pada bab sebelumnya setiap tahapan dalam *flowchart* penelitian mulai dari pengumpulan data, praproses, seleksi fitur, pemodelan *hybrid Support Vector Machine* (SVM) dan *Naïve Bayes* dengan skema, hingga tahap evaluasi dan interpretasi fitur penting melalui *Explainable AI* (XAI) akan dijelaskan secara sistematis beserta hasil pengujian dan analisisnya *weighted voting*.

2.2. Gambar dan Tabel

Semua tabel dan gambar yang anda masukkan dalam dokumen harus disesuaikan dengan urutan 1 kolom atau ukuran penuh satu kertas, agar memudahkan bagi reviewer untuk mencermati makna gambar.



	DegitRatioInURL	NoOfEqualsInURL	NoOfQMarkInURL	NoOfAmpersandInURL
0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0

	NoOfOtherSpecialCharsInURL	SpacialCharRatioInURL	IsHTTPS	LineOfCode
0	0.015625	0.097264	1.0	0.001256
1	0.031250	0.264438	1.0	0.001392
2	0.031250	0.209726	1.0	0.001050
3	0.015625	0.115502	1.0	0.014354
4	0.015625	0.091185	1.0	0.013751

	LargestLineLength	HasTitle	DomainTitleMatchScore	URLTitleMatchScore
0	0.001002	1.0	0.000000	0.000000
1	0.001002	1.0	0.555556	0.555556
2	0.000071	1.0	0.466667	0.466667
3	0.002868	1.0	0.000000	0.000000
4	0.003037	1.0	1.000000	1.000000

	HasFavicon	Robots	IsResponsive	NoOfURLRedirect	NoOfSelfRedirect
0	0.0	1.0	1.0	0.0	0.0
1	1.0	1.0	0.0	0.0	0.0
2	0.0	1.0	1.0	0.0	0.0
3	0.0	1.0	1.0	0.0	0.0
4	0.0	1.0	1.0	1.0	1.0

	HasDescription	NoOfPopup	NoOfiFrame	HasExternalFormSubmit	HasSocialNet
0	0.0	0.000000	0.009346	0.0	0.0
1	0.0	0.000000	0.000000	0.0	1.0
2	1.0	0.000000	0.000000	0.0	0.0
3	0.0	0.002299	0.112150	0.0	1.0
4	1.0	0.000000	0.018692	0.0	1.0

	HasSubmitButton	HasHiddenFields	HasPasswordField	Bank	Pay	Crypto
0	1.0	1.0	0.0	1.0	0.0	0.0
1	1.0	0.0	0.0	0.0	0.0	0.0
2	1.0	1.0	0.0	0.0	0.0	0.0
3	1.0	1.0	0.0	0.0	1.0	1.0
4	1.0	1.0	0.0	1.0	1.0	0.0

	HasCopyrightInfo	NoOfImage	NoOfCSS	NoOfJS	NoOfSelfRef	NoOfEmptyRef
0	1.0	0.004450	0.046189	0.021116	0.044889	0.000000
1	1.0	0.006544	0.020785	0.006033	0.014711	0.000000
2	1.0	0.001309	0.004619	0.005279	0.015843	0.003906
3	1.0	0.000393	0.062356	0.011312	0.008299	0.001953
4	1.0	0.031933	0.034642	0.025641	0.027160	0.001953

	NoOfExternalRef
0	0.03100
1	0.05425
2	0.00125
3	0.00775
4	0.02125

Jumlah baris: 22398

Gambar 1. Tahap *pre-processing data*

Proses awal berupa praproses data dilakukan terhadap 22.598 entri URL yang telah diekstraksi menjadi puluhan fitur numerik. Seluruh entri telah dipastikan bebas dari missing *values*, dan fitur-fitur tidak relevan seperti URL, domain, dan *title* dihapus untuk menghindari data leakage. Fitur numerik kemudian dinormalisasi agar seimbang dalam proses pelatihan, mengingat model seperti SVM *sensitive* terhadap skala data. Label target disusun dalam format biner, yakni 0 untuk URL legitimate dan 1 untuk *URL phishing*.

```
--- Informasi Feature Selection ---
Total fitur awal: 50

SelectKBest (Top 10 Fitur dengan F-statistic):
1. URLSimilarityIndex
2. IsHTTPS
3. DomainTitleMatchScore
4. HasFavicon
5. IsResponsive
6. HasDescription
7. HasSocialNet
8. HasSubmitButton
9. HasHiddenFields
10. HasCopyrightInfo
```

Gambar 2. Tahap *SelectKBest*

Metode ini sederhana namun efektif, dengan hasil seleksi *SelectKBest* yang menyoroti kontribusi signifikan dari atribut struktur URL dan metadata halaman dalam klasifikasi *phishing*.

```
RFE (Top 20 Fitur menggunakan Logistic Regression):
1. URLSimilarityIndex
2. CharContinuationRate
3. TLDLegitimateProb
4. TLDLength
5. NoOfSubDomain
6. LetterRatioInURL
7. DegitRatioInURL
8. NoOfOtherSpecialCharsInURL
9. SpacialCharRatioInURL
10. IsHTTPS
11. DomainTitleMatchScore
12. HasFavicon
13. Robots
14. HasDescription
15. HasSocialNet
16. HasSubmitButton
17. HasHiddenFields
18. Crypto
19. HasCopyrightInfo
20. NoOfSelfRef
```

Gambar 3. Tahap *Recursive Feature Elimination* (RFE)

RFE unggul dalam interpretabilitas karena mempertahankan fitur asli yang relevan secara kontekstual, serta menekankan pentingnya kombinasi karakteristik URL dan konten halaman dalam membedakan *URL phishing* dari yang asli.

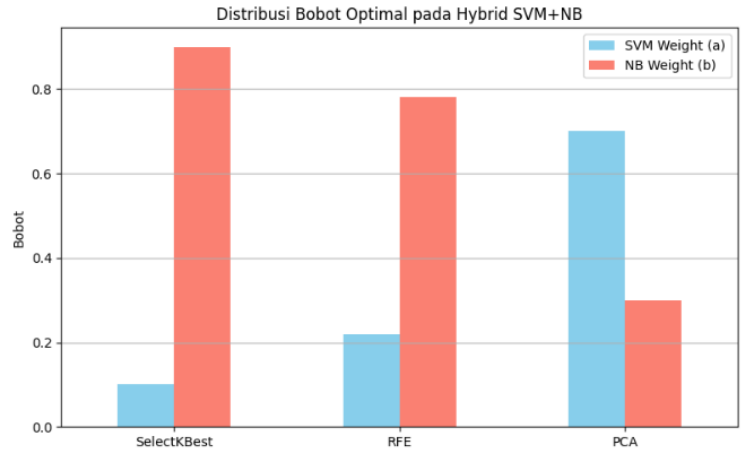
```
PCA (Reduksi ke 20 komponen utama):
Berikut 5 baris pertama hasil transformasi PCA:
          PCA_1     PCA_2     PCA_3     PCA_4     PCA_5     PCA_6     PCA_7  \
12459  0.756779  0.280851 -0.972397 -0.073038  0.586209 -0.797287 -0.411028
8188   0.869933  0.466210 -0.165416  0.428398  0.054904  0.616429  0.339338
9853   1.153140  0.283611  0.582911 -0.624555  0.897352  0.084800  0.320912
21438  0.876899  0.465331  0.385691 -0.166810  0.285243 -0.089960  0.319401
19401  1.399924  0.288329 -0.303217  0.545940  0.728706  0.270251 -0.159270

          PCA_8     PCA_9    PCA_10    PCA_11    PCA_12    PCA_13    PCA_14  \
12459 -0.054221 -0.181855 -0.047263 -0.118817 -0.018210 -0.113049  0.194445
8188  -0.185333  0.299030 -0.892823 -0.042631  0.367280 -0.121352  0.008370
9853  -0.038602  0.127987 -0.360938 -0.150266  0.067594 -0.176858  0.295938
21438 -0.662948  0.529046  0.360298  0.421734 -0.315114 -0.153810 -0.394800
19401  0.247769 -0.116563 -0.268565  0.414271  0.293820  0.020867 -0.073286

         PCA_15    PCA_16    PCA_17    PCA_18    PCA_19    PCA_20
12459  0.091207 -0.146564 -0.035256 -0.078090  0.017379  0.013021
8188  -0.041852 -0.014546  0.028268 -0.032725  0.026888 -0.030021
9853   0.748628 -0.223531  0.123743 -0.183207  0.046227  0.004291
21438 -0.327564 -0.074664  0.115965 -0.021372  0.027377 -0.037805
19401  0.020124  0.025399 -0.253828  0.835472  0.004052 -0.029831
```

Gambar 4. Tahap *Principal Component Analysis* (PCA)

PCA mereduksi dimensi dan mengatasi *overfitting* dengan mengubah fitur asli menjadi kombinasi linier yang mempertahankan variansi data, namun mengorbankan interpretabilitas karena fitur baru tidak merepresentasikan atribut asli secara langsung.

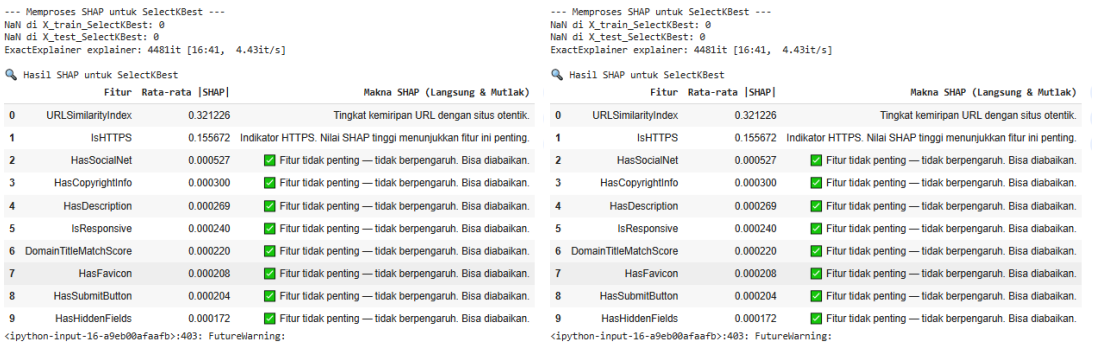
📊 Hasil Evaluasi Terbaik untuk Tiap Metode Seleksi Fitur (Hybrid SVM+NB):

	Accuracy	Precision	Recall	F1-Score	AUC-ROC	SVM Weight (a)	NB Weight (b)
SelectKBest	0.9998	0.9996	1.0000	0.9998	1.0000	0.10	0.90
RFE	0.9998	0.9996	1.0000	0.9998	1.0000	0.22	0.78
PCA	0.9989	0.9985	0.9996	0.9991	0.9999	0.70	0.30



Gambar 5. Evaluasi Kinerja *Model Hybrid*

Evaluasi model menunjukkan bahwa performa *model hybrid* SVM dan *Naïve Bayes* dengan pendekatan *Weighted Voting* menunjukkan bahwa metode seleksi fitur *SelectKBest* dan RFE menghasilkan kinerja yang sangat tinggi dan hampir identik. Keduanya mencatatkan nilai akurasi sebesar 0.9998, *precision* 0.9996, *recall* 1.0000, dan *F1-score* 0.9998, serta AUC-ROC sebesar 1.0000, yang menandakan bahwa model memiliki kemampuan deteksi *phishing* URL secara sempurna tanpa menghasilkan kesalahan klasifikasi negatif palsu *(false negative)*. Bobot optimal yang digunakan untuk *SelectKBest* adalah $\alpha = 0.10$ (SVM) dan $1 - \alpha = 0.90$ *(Naïve Bayes)*, sedangkan untuk RFE adalah $\alpha = 0.22$ (SVM) dan 0.78 *(Naïve Bayes)*. Sementara itu, model dengan fitur yang diekstraksi menggunakan PCA menunjukkan performa yang sedikit lebih rendah, yaitu dengan akurasi sebesar 0.9989, *precision* 0.9985, *recall* 0.9996, *F1-score* 0.9991, dan AUC-ROC sebesar 0.9999. Meskipun tetap menunjukkan kinerja yang sangat baik, hasil ini mengindikasikan bahwa fitur-fitur dari PCA tidak seoptimal *SelectKBest* dan RFE dalam menangkap pola karakteristik *phishing* URL untuk *model hybrid* ini. Bobot optimal untuk metode PCA adalah $\alpha = 0.70$ (SVM) dan 0.30 *(Naïve Bayes)*. Secara keseluruhan, hasil ini menegaskan bahwa pemilihan metode seleksi fitur dan penyesuaian bobot *voting* α secara cermat berpengaruh signifikan terhadap kinerja akhir *model hybrid* dalam tugas klasifikasi *URL phishing.*



```
--- Memproses SHAP untuk SelectKBest ---
NaN di X_train_SelectKBest: 0
NaN di X_test_SelectKBest: 0
ExactExplainer explainer: 4481it [16:41,  4.43it/s]
```

🔍 Hasil SHAP untuk SelectKBest

	Fitur	Rata-rata |SHAP|	Makna SHAP (Langsung & Mutlak)
0	URLSimilarityIndex	0.321226	Tingkat kemiripan URL dengan situs otentik.
1	IsHTTPS	0.155672	Indikator HTTPS. Nilai SHAP tinggi menunjukkan fitur ini penting.
2	HasSocialNet	0.000527	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
3	HasCopyrightInfo	0.000300	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
4	HasDescription	0.000269	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
5	IsResponsive	0.000240	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
6	DomainTitleMatchScore	0.000220	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
7	HasFavicon	0.000208	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
8	HasSubmitButton	0.000204	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
9	HasHiddenFields	0.000172	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.

```
<ipython-input-16-a9eb00afaafb>:403: FutureWarning:
```


```
--- Memproses SHAP untuk SelectKBest ---
NaN di X_train_SelectKBest: 0
NaN di X_test_SelectKBest: 0
ExactExplainer explainer: 4481it [16:41,  4.43it/s]
```

🔍 Hasil SHAP untuk SelectKBest

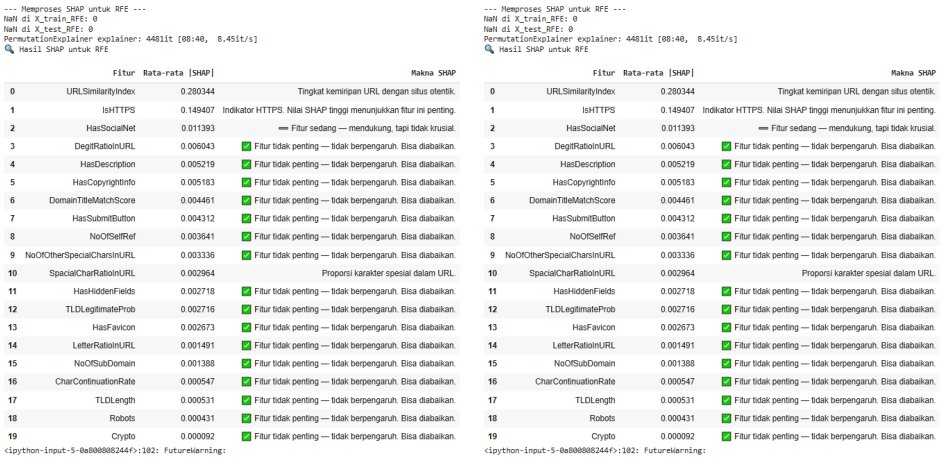
	Fitur	Rata-rata |SHAP|	Makna SHAP (Langsung & Mutlak)
0	URLSimilarityIndex	0.321226	Tingkat kemiripan URL dengan situs otentik.
1	IsHTTPS	0.155672	Indikator HTTPS. Nilai SHAP tinggi menunjukkan fitur ini penting.
2	HasSocialNet	0.000527	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
3	HasCopyrightInfo	0.000300	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
4	HasDescription	0.000269	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
5	IsResponsive	0.000240	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
6	DomainTitleMatchScore	0.000220	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
7	HasFavicon	0.000208	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
8	HasSubmitButton	0.000204	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
9	HasHiddenFields	0.000172	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.

```
<ipython-input-16-a9eb00afaafb>:403: FutureWarning:
```


Gambar 6. Hasil interpretasi XAI dengan SHAP pada *feature selection SelectKBest summary plot & bar chart*

Hasil SHAP pada fitur *SelectKBest* menunjukkan bahwa *URLSimilarityIndex (0.321226)* dan *IsHTTPS (0.155672)* merupakan dua fitur utama yang paling memengaruhi prediksi *phishing*, sementara delapan fitur lainnya memiliki kontribusi sangat rendah dan kurang relevan.

Kedua fitur dominan tersebut menunjukkan korelasi kuat terhadap *output* prediksi dan berkontribusi besar dalam meningkatkan performa model, sedangkan fitur lain seperti *HasSubmitButton* dan *HasHiddenFields* tetap berperan kecil namun penting dalam membentuk batas keputusan klasifikasi.



```
--- Memproses SHAP untuk RFE ---
NaN di X_train_RFE: 0
NaN di X_test_RFE: 0
PermutationExplainer explainer: 4481it [08:40,  8.45it/s]
```

🔍 Hasil SHAP untuk RFE

	Fitur	Rata-rata |SHAP|	Makna SHAP
0	URLSimilarityIndex	0.280344	Tingkat kemiripan URL dengan situs otentik.
1	IsHTTPS	0.149407	Indikator HTTPS. Nilai SHAP tinggi menunjukkan fitur ini penting.
2	HasSocialNet	0.011393	🟰 Fitur sedang — mendukung, tapi tidak krusial.
3	DegitRatioInURL	0.006043	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
4	HasDescription	0.005219	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
5	HasCopyrightInfo	0.005183	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
6	DomainTitleMatchScore	0.004461	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
7	HasSubmitButton	0.004312	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
8	NoOfSelfRef	0.003641	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
9	NoOfOtherSpecialCharsInURL	0.003336	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
10	SpacialCharRatioInURL	0.002964	Proporsi karakter spesial dalam URL.
11	HasHiddenFields	0.002718	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
12	TLDLegitimateProb	0.002716	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
13	HasFavicon	0.002673	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
14	LetterRatioInURL	0.001491	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
15	NoOfSubDomain	0.001388	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
16	CharContinuationRate	0.000547	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
17	TLDLength	0.000531	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
18	Robots	0.000431	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
19	Crypto	0.000092	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.

```
<ipython-input-5-0a800808244f>:102: FutureWarning:
```


```
--- Memproses SHAP untuk RFE ---
NaN di X_train_RFE: 0
NaN di X_test_RFE: 0
PermutationExplainer explainer: 4481it [08:40,  8.45it/s]
```

🔍 Hasil SHAP untuk RFE

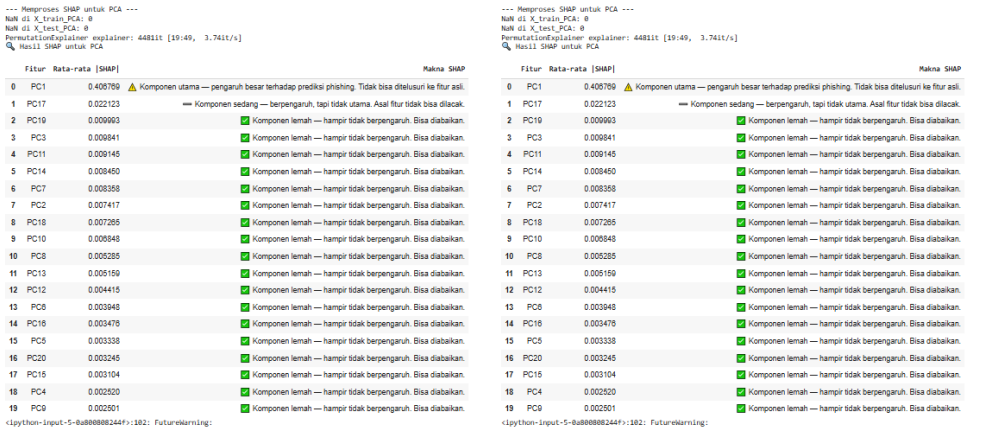
	Fitur	Rata-rata |SHAP|	Makna SHAP
0	URLSimilarityIndex	0.280344	Tingkat kemiripan URL dengan situs otentik.
1	IsHTTPS	0.149407	Indikator HTTPS. Nilai SHAP tinggi menunjukkan fitur ini penting.
2	HasSocialNet	0.011393	🟰 Fitur sedang — mendukung, tapi tidak krusial.
3	DegitRatioInURL	0.006043	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
4	HasDescription	0.005219	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
5	HasCopyrightInfo	0.005183	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
6	DomainTitleMatchScore	0.004461	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
7	HasSubmitButton	0.004312	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
8	NoOfSelfRef	0.003641	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
9	NoOfOtherSpecialCharsInURL	0.003336	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
10	SpacialCharRatioInURL	0.002964	Proporsi karakter spesial dalam URL.
11	HasHiddenFields	0.002718	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
12	TLDLegitimateProb	0.002716	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
13	HasFavicon	0.002673	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
14	LetterRatioInURL	0.001491	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
15	NoOfSubDomain	0.001388	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
16	CharContinuationRate	0.000547	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
17	TLDLength	0.000531	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
18	Robots	0.000431	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.
19	Crypto	0.000092	✅ Fitur tidak penting — tidak berpengaruh. Bisa diabaikan.

```
<ipython-input-5-0a800808244f>:102: FutureWarning:
```


Gambar 7. Hasil interpretasi XAI dengan SHAP pada *feature selection Recursive Feature Elimination* (RFE) *summary plot & bar chart*.

Hasil SHAP pada fitur hasil seleksi RFE menunjukkan bahwa dua fitur utama yang paling berkontribusi terhadap prediksi *phishing* adalah *URLSimilarityIndex (0.280344)* dan *IsHTTPS (0.149407)*, dengan kontribusi sedang dari *HasSocialNet (0.011393)*, sementara 17 fitur lainnya memiliki nilai SHAP sangat rendah dan dianggap tidak signifikan.

Distribusi kontribusi yang timpang ini menegaskan bahwa model hanya mengandalkan sebagian kecil fitur utama, mencerminkan pentingnya interpretabilitas dalam menjelaskan keputusan model, serta efektivitas RFE dalam mempertahankan fitur yang relevan.



--- Memproses SHAP untuk PCA ---
NaN di X_train_PCA: 0
NaN di X_test_PCA: 0
PermutationExplainer explainer: 4481it [19:49, 3.74it/s]
🔍 Hasil SHAP untuk PCA

	Fitur	Rata-rata |SHAP|	Makna SHAP
0	PC1	0.406769	⚠️ Komponen utama — pengaruh besar terhadap prediksi phishing. Tidak bisa ditelusuri ke fitur asli.
1	PC17	0.022123	➖ Komponen sedang — berpengaruh, tapi tidak utama. Asal fitur tidak bisa dilacak.
2	PC19	0.009993	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
3	PC3	0.009841	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
4	PC11	0.009145	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
5	PC14	0.008450	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
6	PC7	0.008358	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
7	PC2	0.007417	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
8	PC18	0.007265	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
9	PC10	0.006848	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
10	PC8	0.005285	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
11	PC13	0.005159	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
12	PC12	0.004415	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
13	PC6	0.003948	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
14	PC16	0.003476	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
15	PC5	0.003338	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
16	PC20	0.003245	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
17	PC15	0.003104	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
18	PC4	0.002520	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
19	PC9	0.002501	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.

<ipython-input-5-0a800802244f>:102: FutureWarning:

--- Memproses SHAP untuk PCA ---
NaN di X_train_PCA: 0
NaN di X_test_PCA: 0
PermutationExplainer explainer: 4481it [19:49, 3.74it/s]
🔍 Hasil SHAP untuk PCA

	Fitur	Rata-rata |SHAP|	Makna SHAP
0	PC1	0.406769	⚠️ Komponen utama — pengaruh besar terhadap prediksi phishing. Tidak bisa ditelusuri ke fitur asli.
1	PC17	0.022123	➖ Komponen sedang — berpengaruh, tapi tidak utama. Asal fitur tidak bisa dilacak.
2	PC19	0.009993	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
3	PC3	0.009841	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
4	PC11	0.009145	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
5	PC14	0.008450	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
6	PC7	0.008358	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
7	PC2	0.007417	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
8	PC18	0.007265	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
9	PC10	0.006848	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
10	PC8	0.005285	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
11	PC13	0.005159	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
12	PC12	0.004415	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
13	PC6	0.003948	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
14	PC16	0.003476	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
15	PC5	0.003338	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
16	PC20	0.003245	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
17	PC15	0.003104	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
18	PC4	0.002520	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.
19	PC9	0.002501	✅ Komponen lemah — hampir tidak berpengaruh. Bisa diabaikan.

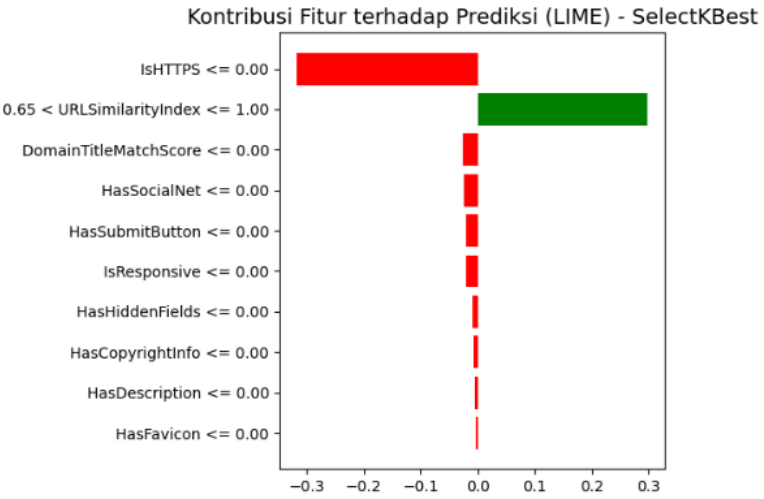
<ipython-input-5-0a800802244f>:102: FutureWarning:

Gambar 8. Hasil interpretasi XAI dengan SHAP pada *feature selection Principal Component Analysis* (PCA) *summary plot & bar chart*

Hasil SHAP untuk fitur hasil transformasi PCA menunjukkan bahwa hanya komponen utama PC1 yang memberikan kontribusi besar (0.406769) terhadap prediksi model, disusul oleh PC17 (0.022123) dengan kontribusi sedang, sementara 18 komponen lainnya memiliki kontribusi sangat kecil dan dianggap tidak signifikan.

Meskipun PCA efektif dalam mereduksi dimensi dan mempertahankan akurasi, keterbatasannya terletak pada interpretabilitas karena setiap komponen merupakan kombinasi linier dari fitur asli, sehingga sulit untuk menelusuri pengaruh fitur awal secara langsung terhadap keputusan model.

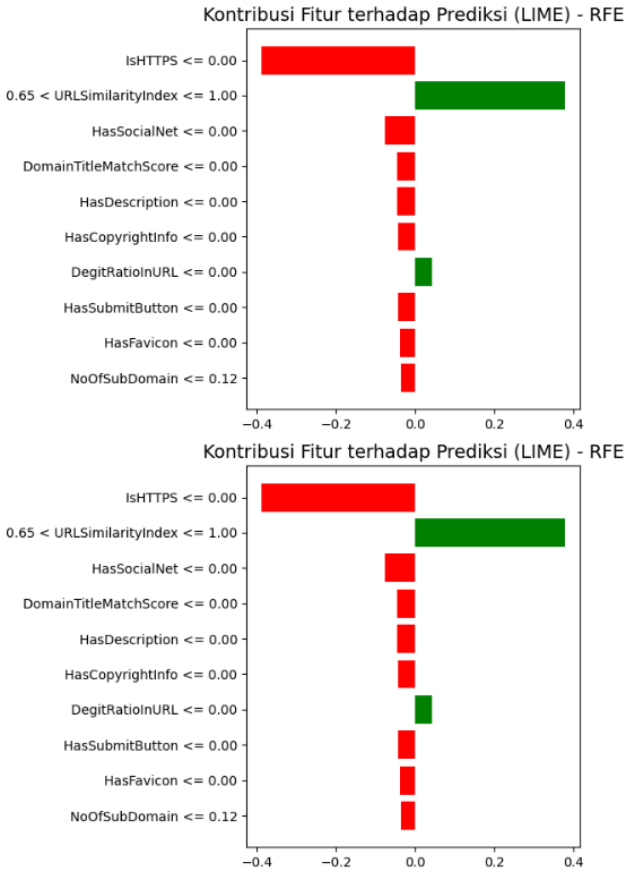
	No	Fitur	Kontribusi	Artinya
0	1	IsHTTPS <= 0.00	-0.3085	Situs tidak pakai HTTPS, tapi model justru menurunkan dugaan phishing karena phisher biasanya justru berusaha tampil aman dengan HTTPS.
1	2	0.65 < URLSimilarityIndex <= 1.00	+0.2928	URL mirip dengan situs asli → mendukung phishing karena meniru tampilan otentik.
2	3	HasSocialNet <= 0.00	-0.0326	Tidak ada tautan media sosial. Biasanya mencurigakan, tapi model menganggap ini mengurangi risiko phishing.
3	4	HasSubmitButton <= 0.00	-0.0244	Tidak ada tombol submit → indikasi aman, model menurunkan kemungkinan phishing.
4	5	HasDescription <= 0.00	-0.0164	Tidak ada deskripsi halaman → tidak informatif, tapi menurunkan risiko phishing.
5	6	DomainTitleMatchScore <= 0.00	-0.0164	Domain tidak cocok dengan judul halaman. Biasanya indikasi phishing, tapi model justru menurunkan skornya.
6	7	IsResponsive <= 0.00	-0.0141	Interpretasi tidak tersedia untuk fitur ini.
7	8	HasHiddenFields <= 0.00	-0.0114	Tidak ada field tersembunyi di form, ini positif dan menurunkan peluang phishing.
8	9	HasCopyrightInfo <= 0.00	-0.0085	Tidak ada copyright info di halaman. Tidak profesional, tapi menurut model tidak terlalu mencurigakan.
9	10	HasFavicon <= 0.00	-0.0053	Interpretasi tidak tersedia untuk fitur ini.



Gambar 9. Hasil interpretasi XAI dengan LIME pada *feature selection SelectKBest summary plot & bar chart*

Hasil interpretasi LIME pada model dengan fitur *SelectKBest* menunjukkan bahwa *URLSimilarityIndex* memberikan kontribusi positif terbesar terhadap prediksi *phishing*, sementara *IsHTTPS* justru memberikan kontribusi negatif, mengindikasikan bahwa model lebih mengandalkan kemiripan URL dengan situs asli dibandingkan aspek keamanan protokol.

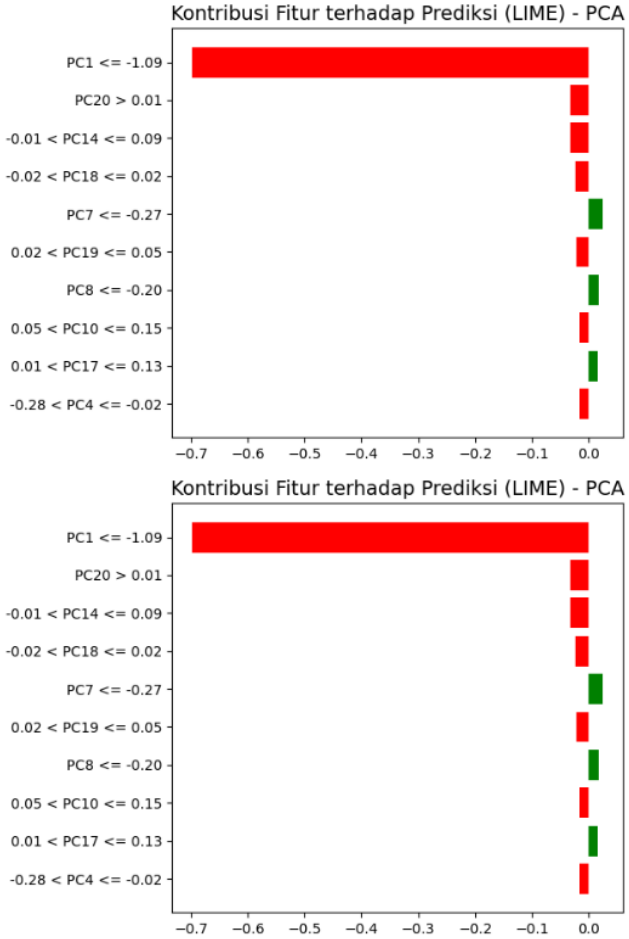
Fitur lain seperti *HasSocialNet*, *HasSubmitButton*, dan *HasDescription* memberikan kontribusi negatif kecil, yang menandakan bahwa ketidakhadiran elemen-elemen tersebut justru menurunkan risiko *phishing* menurut pola data pelatihan, dan secara keseluruhan, fitur *SelectKBest* memberikan informasi yang relevan dan dapat dijelaskan secara lokal pada level *instance*.



Gambar 10. Hasil interpretasi XAI dengan LIME pada *feature selection Recursive Feature Elimination* (RFE) *summary plot & bar chart*

Hasil interpretasi LIME pada model dengan fitur RFE menunjukkan bahwa *URLSimilarityIndex* memberikan kontribusi positif terbesar terhadap prediksi *phishing*, sedangkan *IsHTTPS* memberikan kontribusi negatif signifikan, mengindikasikan bahwa model memandang kemiripan URL sebagai sinyal *phishing* yang lebih kuat dibanding keberadaan *HTTPS*.

Fitur-fitur lain seperti *HasSocialNet*, *DomainTitleMatchScore*, dan *NoOfSubDomain* memberikan kontribusi negatif kecil, sementara *DigitRatioInURL* justru berkontribusi positif meski URL minim angka. Pola ini menegaskan bahwa model mengikuti pola data pelatihan, dan bahwa kombinasi RFE dan LIME mampu menyoroti fitur-fitur paling relevan secara lokal dan kontekstual.



Gambar 11. Hasil interpretasi XAI dengan LIME pada *feature selection Principal Component Analysis* (PCA) *summary plot* & *bar chart*

Hasil interpretasi model menggunakan LIME pada fitur hasil seleksi PCA menunjukkan bahwa PC1 memberikan kontribusi negatif terbesar terhadap prediksi *phishing*, sementara komponen lainnya seperti PC20, PC14, dan PC18 hanya berkontribusi kecil, baik negatif maupun positif.

Dominasi kontribusi PC1 mengindikasikan bahwa sebagian besar informasi prediktif terkonsentrasi pada satu komponen utama, namun karena PCA merupakan kombinasi linier dari fitur asli, interpretasi makna tiap komponen menjadi terbatas dan abstrak, sehingga mengurangi transparansi model.

3. HASIL DAN PEMBAHASAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa tujuan utama penelitian, yaitu membangun model deteksi *phishing URL* berbasis *hybrid Machine Learning* dengan pendekatan *weighted voting* antara algoritma *Support Vector Machine* (SVM) dan *Naïve Bayes*, telah tercapai dengan baik. Model yang dikembangkan menunjukkan performa klasifikasi yang tinggi, terutama setelah penerapan teknik seleksi fitur seperti *SelectKBest* dan *Recursive Feature Elimination* (RFE), yang mampu meningkatkan akurasi sekaligus mengurangi kompleksitas model. Selain itu, penerapan metode *Explainable AI* (XAI) seperti SHAP dan LIME terbukti efektif dalam memberikan interpretasi terhadap kontribusi fitur atau komponen terhadap hasil prediksi, sehingga meningkatkan transparansi dan kepercayaan pengguna terhadap sistem. Penelitian ini memberikan kontribusi penting dalam

konteks keamanan siber, khususnya dalam pengembangan sistem deteksi *phishing* yang andal dan dapat dijelaskan. Ke depan, penelitian lanjutan dapat difokuskan pada integrasi pendekatan ini dalam sistem deteksi *real-time* serta eksplorasi algoritma *ensemble* lainnya untuk peningkatan performa dan skalabilitas model.

4. KESIMPULAN

Penelitian ini menerapkan model deteksi *phishing URL* berbasis *hybrid* antara *Support Vector Machine* (SVM) dan *Naïve Bayes* dengan skema *Weighted Voting*. Proses meliputi praproses data, seleksi fitur, pelatihan model, serta interpretasi menggunakan *Explainable AI* (XAI). Setelah pembersihan data dan penghapusan atribut deskriptif untuk mencegah *data leakage*, diperoleh 22.598 entri numerik-biner.Tiga metode seleksi fitur diuji: *SelectKBest*, *Recursive Feature Elimination* (RFE), dan *Principal Component Analysis* (PCA). *SelectKBest* dan RFE menghasilkan performa terbaik dengan *precision* 0.9996, *recall* 1.0000, *F1-score* 0.9998, dan AUC-ROC 1.0000. Bobot optimal *voting* adalah $\alpha = 0.10$ (SVM) dan 0.90 (*Naïve Bayes*) untuk *SelectKBest*, serta $\alpha = 0.22$ dan 0.78 untuk RFE. PCA menunjukkan kinerja sedikit lebih rendah (*F1-score* 0.9991, AUC 0.9999). Analisis XAI melalui SHAP dan LIME menegaskan bahwa fitur *URLSimilarityIndex* dan *IsHTTPS* paling berpengaruh terhadap hasil prediksi. Secara keseluruhan, kombinasi seleksi fitur tepat dan bobot *Weighted Voting* terbukti meningkatkan akurasi, reliabilitas, serta transparansi model deteksi *phishing URL*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan dalam pelaksanaan penelitian ini. Ucapan terima kasih khusus disampaikan kepada dosen pembimbing yang telah memberikan arahan dan masukan yang sangat berarti, serta kepada keluarga dan rekan-rekan yang senantiasa memberikan semangat dan dukungan selama proses penyusunan jurnal ini. Penulis juga mengapresiasi kontribusi sumber data dari media sosial yang menjadi objek kajian dalam penelitian ini. Semoga hasil penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya dalam bidang analisis sentimen dan pengolahan data teks.

DAFTAR PUSTAKA

[1] A. F. Mahmud and S. Wirawan, “Deteksi Phishing Website menggunakan Machine Learning Metode Klasifikasi,” *Sistemasi: Jurnal Sistem Informasi*, vol. 13, no. 4, pp. 1368–1380, 2024.

[2] A. Mutemi and F. Bacao, “E-commerce fraud detection based on machine learning techniques: Systematic literature review,” *Big Data Mining and Analytics*, vol. 7, no. 2, pp. 419–444, 2024.

[3] A. Safi and S. Singh, “A systematic literature review on phishing website detection techniques,” *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 590–611, 2023.

[4] A. F. Nugraha, R. F. A. Aziza, and Y. Pristyanto, “Penerapan metode Stacking dan Random Forest untuk Meningkatkan Kinerja Klasifikasi pada Proses Deteksi Web Phishing,” *Jurnal Infomedia: Teknik Informatika, Multimedia, dan Jaringan*, vol. 7, no. 1, pp. 39–44, 2022.

[5] D. C. Asogwa, S. O. Anigbogu, I. E. Onyenwe, and F. A. Sani, “Text classification using hybrid machine learning algorithms on big data,” *arXiv preprint arXiv:2103.16624*, 2021.

[6] P. P. Kuantitatif, “Metode Penelitian Kunatitatif Kualitatif dan R&D,” *Alfabeta, Bandung*, 2016.

[7] N. Gharaibeh, O. M. Al-hazaimeh, A. Abu-Ein, and K. M. Nahar, "A hybrid svm naïve-bayes classifier for bright lesions recognition in eye fundus images," *Int J Electr Eng Inform*, vol. 13, no. 3, pp. 530–545, 2021.

[8] L. C. Nnadi, Y. Watanobe, M. M. Rahman, and A. M. John-Otumu, "Prediction of students' adaptability using explainable AI in educational machine learning models," *Applied Sciences*, vol. 14, no. 12, p. 5141, 2024.

[9] S. R. Abdul Samad *et al.*, "Analysis of the performance impact of fine-tuned machine learning model for phishing URL detection," *Electronics (Basel)*, vol. 12, no. 7, p. 1642, 2023.

[10] S. Wang, J. Ren, and R. Bai, "A semi-supervised adaptive discriminative discretization method improving discrimination power of regularized naive Bayes," *Expert Syst Appl*, vol. 225, p. 120094, 2023.

[11] Z. Alhaq, A. Mustopa, S. Mulyatun, and J. D. Santoso, "Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter," *Journal of Information System Management (JOISM)*, vol. 3, no. 1, pp. 16–21, 2021.

[12] A. Rojarath and W. Songpan, "Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems," *Applied Intelligence*, vol. 51, pp. 4908–4932, 2021.

[13] V. C. Osamor and A. F. Okezie, "Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis," *Sci Rep*, vol. 11, no. 1, p. 14806, 2021.

[14] W. Yang *et al.*, "Survey on explainable AI: From approaches, limitations and applications aspects," *Human-Centric Intelligent Systems*, vol. 3, no. 3, pp. 161–188, 2023.

[15] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Inform*, vol. 11, no. 1, p. 10, 2024.