ABSTRAK

Keamanan siber telah menjadi tantangan kritikal seiring meningkatnya ancaman *phishing* yang memanipulasi URL dan konten halaman web untuk memperoleh informasi sensitif pengguna. Pendekatan *lexical-based features*, yang hanya mengandalkan struktur *URL*, semakin mudah dielakkan oleh teknik *phishing* modern yang memanfaatkan *subdomain* sah untuk menyamarkan konten berbahaya. Penelitian ini mengusulkan pendekatan *hybrid phishing detection* menggunakan algoritma XGBoost yang mengombinasikan 10 fitur leksikal dan 19 fitur konten halaman web, dievaluasi melalui enam skenario eksperimen (S1–S6). Analisis interpretabilitas dilakukan secara komparatif menggunakan tiga metode *feature importance*, yaitu *Built-in Gain*, *Permutation Importance*, dan *SHAP*.

Model dilatih pada *dataset* PhishOFE (101.063 sampel), sedangkan validasi generalisasi dilakukan pada *dataset* Mendeley dan kombinasi OpenPhish+Tranco. Hasil evaluasi *intra-dataset* menunjukkan skenario *hybrid* (S2) mencapai *F1-score* tertinggi sebesar 0,9902, melampaui pendekatan *lexical-based* (S1: 0,9308) dan *content-based* (S3: 0,9529). Ketiga metode *feature importance* menghasilkan peringkat yang konsisten, dengan enam fitur dominan yang muncul pada *top-10* ketiga metode: *URLLength*, *NoOfSubDomain*, *NoOfHyperlink*, *NoOfDigits*, *NoOfCSS*, dan *NoOfURLRedirect*. Namun, evaluasi *cross-dataset* menunjukkan penurunan performa yang signifikan akibat *domain shift*: subset fitur hasil *SHAP* (S6) menunjukkan ketahanan terbaik pada *dataset* Mendeley (*F1-score* 0,8147), sementara tidak ada satu metode *feature importance* pun yang secara konsisten menghasilkan subset fitur paling representatif pada kedua *dataset* eksternal. Temuan ini mengindikasikan bahwa keunggulan *hybrid features* atas fitur tunggal bersifat kondisional terhadap distribusi data pengujian, bukan bersifat universal.

Kata kunci: deteksi *phishing,* XGBoost*, hybrid features, feature importance,* SHAP*, domain shift*

ABSTRACT

Cybersecurity has become a critical challenge as phishing threats increasingly manipulate URLs and web page content to obtain users' sensitive information. Lexical-based feature approaches, which rely solely on URL structure, are increasingly easy to evade by modern phishing techniques that exploit legitimate subdomains to disguise malicious content. This research proposes a hybrid phishing detection approach using the XGBoost algorithm, combining 10 lexical features and 19 web page content features, evaluated through six experimental scenarios (S1–S6). Interpretability analysis was conducted comparatively using three feature importance methods: Built-in Gain, Permutation Importance, and SHAP. The model was trained on the PhishOFE dataset (101,063 samples), while generalization validation was performed on the Mendeley dataset and a combined OpenPhish+Tranco dataset. Intra-dataset evaluation results show that the hybrid scenario (S2) achieved the highest F1-score of 0.9902, surpassing the lexical-based approach (S1: 0.9308) and the content-based approach (S3: 0.9529). All three feature importance methods produced consistent rankings, with six dominant features appearing in the top-10 of all three methods: URLLength, NoOfSubDomain, NoOfHyperlink, NoOfDigits, NoOfCSS, and NoOfURLRedirect. However, cross-dataset evaluation revealed a significant performance decline due to domain shift: the SHAP-derived feature subset (S6) showed the best robustness on the Mendeley dataset (F1-score 0.8147), while no single feature importance method consistently produced the most representative feature subset across both external datasets. These findings indicate that the superiority of hybrid features over single-type features is conditional on the test data distribution, rather than universal.

Keywords: *phishing detection, XGBoost, hybrid features, feature importance, SHAP, domain shift*