UNIVERSITAS
GADJAH MADA
Analisis Konsistensi Feature Importance Berbasis SHAP melalui Cross-Model Validation pada Model Tree-Based untuk Deteksi Phishing URL
Rakyan Pangrukti Wibana, Dr. Mardhani Riasetiawan, SE Ak, M.T.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

ANALISIS KONSISTENSI FEATURE IMPORTANCE BERBASIS SHAP MELALUI CROSS-MODEL VALIDATION PADA MODEL TREE-BASED UNTUK DETEKSI PHISHING URL

Oleh

RAKYAN PANGRUKTI WIBANA
22/497076/PA/21394

Perkembangan pesat teknologi internet telah meningkatkan ancaman siber, khususnya serangan *phishing* melalui URL berbahaya. Pendekatan berbasis *machine learning*, terutama model *tree-based* seperti Random Forest, XGBoost, dan LightGBM, terbukti efektif dalam mendeteksi *phishing* URL, namun meningkatnya kompleksitas model menuntut transparansi melalui metode *Explainable Artificial Intelligence* (XAI), khususnya *SHapley Additive exPlanations* (SHAP). Meskipun SHAP telah banyak digunakan untuk interpretasi model individual, konsistensi *feature importance* lintas model masih jarang dikaji. Oleh karena itu, penelitian ini menganalisis konsistensi *feature importance* berbasis SHAP melalui pendekatan *cross-model validation* pada ketiga model tersebut menggunakan dataset dengan 87 fitur dan 11.256 sampel. Ketiga model mencapai akurasi di atas 95%, dengan XGBoost tertinggi sebesar 96,14%. *Cross-model validation* menghasilkan rata-rata *Spearman rank correlation* sebesar 0,8541 dan rata-rata *Jaccard similarity* sebesar 0,4693 pada ambang batas *top-20*, mencerminkan tingkat keselarasan yang tinggi antar model. Melalui operasi irisan himpunan, teridentifikasi 10 fitur konsisten yang diakui penting oleh ketiga model, dengan *page_rank* dan *google_index* sebagai fitur terkuat. Validasi menunjukkan bahwa 10 fitur tersebut mampu mempertahankan akurasi di atas 94% dengan percepatan inferensi hingga 2,4×, membuktikan efisiensi dan stabilitas interpretasi lintas model.

Kata kunci: deteksi *phishing* URL, *explainable AI*, *feature importance*, SHAP, *cross-model validation*, model *tree-based*

UNIVERSITAS
GADJAH MADA
Analisis Konsistensi Feature Importance Berbasis SHAP melalui Cross-Model Validation pada Model Tree-Based untuk Deteksi Phishing URL
Rakyan Pangrukti Wibana, Dr. Mardhani Riasetiawan, SE Ak, M.T.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

CONSISTENCY ANALYSIS OF SHAP-BASED FEATURE IMPORTANCE THROUGH CROSS-MODEL VALIDATION ON TREE-BASED MODELS FOR PHISHING URL DETECTION

By

RAKYAN PANGRUKTI WIBANA
22/497076/PA/21394

The rapid development of internet technology has increased cybersecurity threats, particularly phishing attacks through malicious URLs. Machine learning approaches, especially tree-based models such as Random Forest, XGBoost, and LightGBM, have proven effective in detecting phishing URLs, yet the growing model complexity demands transparency through Explainable Artificial Intelligence (XAI), particularly SHapley Additive exPlanations (SHAP). Although SHAP has been widely used for individual model interpretation, the consistency of feature importance across models remains underexplored. Therefore, this study analyzes the consistency of SHAP-based feature importance through a cross-model validation approach on these three models using a dataset of 87 features and 11,256 samples. All three models achieved accuracy above 95%, with XGBoost reaching the highest at 96.14%. Cross-model validation yielded an average Spearman rank correlation of 0.8541 and an average Jaccard similarity of 0.4693 at the top-20 threshold, reflecting a high level of alignment across models. Through set intersection, 10 features were consistently identified as important across all three models, with *page_rank* and *google_index* as the strongest indicators. Validation showed that these 10 features were able to maintain accuracy above 94% with up to 2.4× faster inference, proving the efficiency and stability of interpretation across models.

Keywords: phishing URL detection, explainable AI, feature importance, SHAP, cross-model validation, tree-based models