ANALISIS PERBANDINGAN KINERJA MODEL *MACHINE LEARNING* SERTA IDENTIFIKASI FITUR DOMINAN DALAM KLASIFIKASI *URL PHISHING*

HANA PERTIWI

PROGRAM MAGISTER ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul "Analisis Perbandingan Kinerja Model *Machine Learning* serta Identifikasi Fitur Dominan dalam Klasifikasi URL *Phishing*" adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Agustus 2026

Hana Pertiwi
G6501222028

RINGKASAN

HANA PERTIWI. Analisis Perbandingan Kinerja Model *Machine Learning* serta Identifikasi Fitur Dominan dalam Klasifikasi URL *Phishing*. Dibimbing oleh MEDRIA KUSUMA DEWI HARDHIENATA dan SHELVIE NIDYA NEYMAN.

Perkembangan era digital yang semakin meningkat membuat ancaman kejahatan siber semakin besar. Berdasarkan Laporan Lanskap Keamanan Siber Indonesia Tahun 2024, Badan Siber dan Sandi Negara menyebutkan bahwa *phishing* menjadi salah satu anomali terbanyak dengan aktivitas sebesar 26.771.610. Peningkatan jumlah distribusi URL *Phishing* dalam beberapa tahun ini menuntut adanya metode untuk mendeteksi URL *Phishing* yang dapat mengurangi risiko bagi pengguna. Salah satu solusi untuk menangani aksi kejahatan siber tersebut adalah dengan mengadopsi sistem *machine learning*. Penelitian ini menganalisis model *machine learning* menggunakan data komprehensif yang telah dikumpulkan oleh Badan Siber dan Sandi Negara pada tahun 2021 s.d. 2024. Rekomendasi hasil analisis yang didapat memberikan model terbaik untuk penanganan *phishing* dan mengidentifikasi karakteristik URL *phishing* khususnya yang marak digunakan di Indonesia. Penelitian ini mengidentifikasi fitur dominan untuk mengurangi *missing feature* berdasarkan fitur keamanan. Analisis dilakukan untuk menentukan algoritma yang memberikan performa paling optimal dalam mendeteksi URL *Phishing* dengan membandingkan kinerja tiga model algoritma *decision tree* yaitu algoritma *Random Forest, XGBoost*, dan *LightGBM*. Hasil menunjukkan bahwa *LightGBM* memiliki performa lebih unggul dibandingkan *Random Forest* dan *XGBoost* dengan tingkat akurasi sebesar 97,85% dengan MAE sebesar 0,0215. Secara keseluruhan dari setiap hasil *Recall, Precision*, dan *F1-Score LightGBM* memiliki nilai yang lebih tinggi dibandingkan dengan model lain. *LightGBM* menunjukkan performa terbaik dalam memodelkan hubungan *non-linier* secara efisien melalui pendekatan *leaf-wise tree growth* sehingga lebih adaptif untuk data URL *phishing* yang memiliki karakter lebih kompleks. Dari model terbaik dievaluasi dengan analisis *feature importance* dan analisis *Shapley Additive exPlanations (SHAP)* telah teridentifikasi bahwa fitur yang memiliki kontribusi paling besar dalam penentuan URL *Phishing* adalah *Count_Slash_Website*. Analisis data BSSN dalam penelitian ini menghasilkan sejumlah fitur dominan yang membentuk karakteristik URL, sehingga menjadi ciri khas serangan *phishing* yang umum digunakan di Indonesia dalam hal ini salah satunya adalah jumlah penggunaan garis miring (*Count_Slash_Website) pada* URL.

Kata kunci: *LightGBM, Phishing, Random Forest, SHAP,* URL *Phishing, XGBoost*.

.

SUMMARY

HANA PERTIWI. Comparative Analysis of Machine Learning Model Performance and Identification of Dominant Features in *Phishing* URL Classification. Supervised by MEDRIA KUSUMA DEWI HARDHIENATA and SHELVIE NIDYA NEYMAN.

The rapid advancement of the digital era has significantly increased the risk of cybercrime, Based on the 2024 Indonesian Cybersecurity Landscape Report, the National Cyber and Crypto Agency (BSSN) noted that *phishing* has become one of the most frequent anomalies, with activity reaching 26,771,610 incidents. The increasing distribution of *phishing* URLs in recent years necessitates effective detection methods to reduce risks for users. One of the solutions to address this issue is the adoption of machine learning-based systems. This study analyzes machine learning models using a comprehensive dataset collected by BSSN from 2021 to 2024. The results of this analysis provide recommendations for the most effective model in handling *phishing* detection and identifying the characteristics of *phishing* URLs, particularly those commonly found in Indonesia. Furthermore, this study identifies dominant features to reduce missing feature issues based on security-related attributes for forensic purposes. A comparative analysis is conducted to determine the most optimal algorithm for *phishing* URL detection by evaluating three decision tree-based models, namely *Random Forest*, *XGBoost*, and *LightGBM*. The results show that *LightGBM* achieves superior performance, with an accuracy of 97.85% and a Mean Absolute Error (MAE) of 0,0215, outperforming *Random Forest* and *XGBoost*. Overall, *LightGBM* also demonstrates higher recall, precision, and F1-score values compared to the other models. This performance is attributed to its ability to efficiently model non-linear relationships through a leaf-wise tree growth approach, making it more adaptive to the complex characteristics of *phishing* URL data. Additionally, feature importance analysis and Shapley Additive exPlanations analysis reveals that Count_Slash_Website is the most influential feature in determining *phishing* URLs. The analysis of BSSN data in this study identified several dominant features that define URL characteristics, which serve as distinctive markers of *phishing* attacks prevalent in Indonesia, one of which is the number of slashes (Count_Slash_Website) present in the URL.

Keywords: *LightGBM*, *Phishing*, *Random Forest*, SHAP, *XGBoost*.

ANALISIS PERBANDINGAN KINERJA MODEL *MACHINE LEARNING* SERTA IDENTIFIKASI FITUR DOMINAN DALAM KLASIFIKASI *URL PHISHING*

HANA PERTIWI

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Ilmu Komputer

PROGRAM MAGISTER ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026

Tim Penguji pada Ujian Tesis:
1. Dr. Mushthofa, S.Kom., M.Sc

Judul Tesis : Analisis Perbandingan Kinerja Model *Machine Learning* serta Identifikasi Fitur Dominan dalam Klasifikasi URL *Phishing*
Nama : Hana Pertiwi
NIM : G6501222028

Disetujui oleh

Pembimbing 1:
Dr. Medria Kusuma Dewi Hardhienata, S.Komp.

Pembimbing 2:
Dr. Shelvie Nidya Neyman, S.Kom., M.Si.

Diketahui oleh

Ketua Program Studi:
Irman Hermadi, S.Kom., M.S., Ph.D.
NIP 197503112006041009

Dekan Sekolah Sains Data, Matematika dan Informatika:
Prof. Dr. Ir. Agus Buono, M.Si., M.Kom
NIP. 196607021993021001

Tanggal Ujian:
3 Agustus 2026

Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan kepada Tuhan atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Judul penelitian ini adalah "Analisis Perbandingan Kinerja Model *Machine Learning* serta Identifikasi Fitur Dominan dalam Klasifikasi URL *Phishing*".

Terima kasih penulis ucapkan kepada para pembimbing, Dr. Medria Kusuma Dewi Hardhienata, S.Komp. dan Dr. Shelvie Nidya Neyman, S.Kom., M.Si yang telah membimbing dan banyak memberi saran. Terimakasih kepada seluruh dosen yang telah memberikan ilmu selama menjalani studi di IPB. Ungkapan terima kasih juga disampaikan kepada ayah, ibu dan mama serta seluruh keluarga terutama suami tercinta Oktavianus Abrianta Sembiring, S.H., M.H. dan kedua putri tercinta yang telah memberikan dukungan, doa, dan kasih sayangnya sehingga karya ilmiah ini terselesaikan dengan baik. Penulis mengakui penggunaan teknologi kecerdasan buatan dalam membantu memperbaiki struktur penulisan, tata bahasa, serta memberikan saran tambahan dalam penyusunan naskah penelitian ini. Penggunaan teknologi ini hanya sebagai alat bantu dalam proses penulisan. Seluruh isi, analisis, interpretasi data, dan kesimpulan yang disajikan dalam penelitian ini sepenuhnya merupakan tanggung jawab penulis.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Agustus 2026

Hana Pertiwi

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN