Journal of Science and Social Research ISSN 2615 – 4307 (Print)
April 2026, IX (2): 1900 – 1906
Available online at http://jurnal.goretanpena.com/index.php/JSSR

DETEKSI *ZERO-DAY SOCIAL ENGINERING ATTACK* MENGGUNAKAN NLP DAN *OPEN-SET DEEP LEARNING*

Sahren[1*], Ruri Ashari Dalimunthe[2], Bima Aditya[3]
Universitas Royal, Kisaran
e-mail: [1*]sahren.one@gmail.com, [2]ruriashari1986@gmail.com

Abstract: *Text based social engineering attacks are a growing cyber threat that is difficult to detect by conventional intrusion detection systems, especially in previously unobserved or zero-day variants. This study proposes a Natural Language Processing Open-Set Intrusion Detection System (NLP-OSIDS) framework that integrates Term Frequency-Inverse Document Frequency (TF-IDF) trigram (1.3-gram) feature representation with an Open-Set Multilayer Perceptron architecture based on energy based scoring to detect zero-day social engineering attacks without requiring training examples from that class. Experiments were conducted on the public dataset phishing_email.csv with 82,486 combined samples from Enron, SpamAssassin, Nazario, Ling, CEAS, and Nigerian Fraud datasets with strict zero-day partitioning following open-set recognition evaluation standards. The results show that NLP-OSIDS achieved an AUROC of 0.7808, surpassing all closed-set baselines (AUROC = 0.500) with the lowest False Positive Rate of 0.0088, while the Zero-Day Detection Rate (ZD-DR) of 0.077 indicates the need for adaptive threshold optimization as a direction for further research.*

Keyword: *NLP; OSIDS; Social_Engineering Attack; TF-IDF_Trigram; Zero-Day*

Abstrak: Social engineering attack berbasis konten teks merupakan ancaman siber yang terus berkembang dan sulit dideteksi oleh intrusion detection system konvensional, khususnya pada varian yang belum pernah diobservasi sebelumnya atau zero-day. Penelitian ini mengusulkan kerangka Natural Language Processing Open-Set Intrusion Detection System (NLP-OSIDS) yang mengintegrasikan representasi fitur Term Frequency-Inverse Document Frequency (TF-IDF) trigram (1,3-gram) dengan arsitektur Open-Set Multilayer Perceptron berbasis energy-based scoring untuk mendeteksi zero-day social engineering attack tanpa memerlukan contoh pelatihan dari kelas tersebut. Eksperimen dilakukan pada dataset publik phishing_email.csv dengan 82.486 sampel gabungan dari dataset Enron, SpamAssassin, Nazario, Ling, CEAS, dan Nigerian Fraud dengan partisi zero-day yang ketat mengikuti standar evaluasi open-set recognition. Hasil menunjukkan NLP-OSIDS mencapai AUROC 0,7808 melampaui seluruh baseline closed-set (AUROC = 0,500) dengan False Positive Rate terendah sebesar 0,0088, sementara Zero-Day Detection Rate (ZD-DR) sebesar 0,077 mengindikasikan perlunya optimasi threshold adaptif sebagai arah penelitian lanjutan.

Kata kunci: NLP; OSIDS; Social_Engineering Attack; TF-IDF_Trigram; Zero-Day

PENDAHULUAN

Social engineering attack telah berevolusi menjadi ancaman dominan dalam lanskap keamanan siber modern, tidak lagi sebagai pelengkap serangan teknis, melainkan sebagai vektor utama yang secara sistematis mengeksploitasi kelemahan kognitif manusia (Sathe et al., 2025). Laporan Verizon DBIR 2024 menunjukkan bahwa 68% pelanggaran data melibatkan faktor manusia (Hylender et al., 2024), sementara FBI IC3 mencatat kerugian global mencapai USD 16,6 miliar, dengan *Business Email Compromise* (BEC) sebagai salah satu

bentuk serangan paling merusak secara finansial (Haoxing & System, 2024). Fakta ini menegaskan bahwa paradigma keamanan yang berfokus pada infrastruktur teknis semata telah menjadi tidak memadai dalam menghadapi realitas ancaman (Andri Yusda et al., 2025).

Berbeda secara fundamental dari serangan berbasis eksploitasi perangkat lunak, *social engineering* beroperasi melalui konstruksi narasi linguistik yang dirancang untuk memanipulasi persepsi dan pengambilan keputusan korban (Doshi et al., 2023).

Dengan akselerasi teknologi kecerdasan buatan, serangan ini kini mengalami transformasi signifikan lebih dari 80% *email phishing* pada tahun 2025 dihasilkan dengan bantuan AI, memungkinkan produksi konten yang adaptif, kontekstual, dan nyaris tidak dapat dibedakan dari komunikasi sah. Dalam konteks ini, pendekatan deteksi berbasis sintaks atau pola statis menjadi usang, sehingga memerlukan pergeseran menuju analisis semantik berbasis *Natural Language Processing* (NLP) (Andri Yusda et al., 2025) (Wei et al., 2025).

Meskipun demikian, literatur terkini menunjukkan paradoks yang signifikan. Di satu sisi, model NLP berbasis *deep learning* melaporkan akurasi tinggi, bahkan melampaui 99% pada *dataset benchmark* (Gogoi & Ahmed, 2022) (Atawneh & Aljehani, 2023). Di sisi lain, performa tersebut secara sistematis *overestimated* karena bergantung pada asumsi *closed-set*, yaitu bahwa distribusi data pelatihan identik dengan data pengujian. Asumsi ini secara praktis tidak valid dalam skenario dunia nyata, di mana serangan *zero-day* yang tidak memiliki representasi dalam data pelatihan justru menjadi ancaman utama (He et al., 2022). Konsekuensinya, model dengan akurasi tinggi dalam evaluasi konvensional dapat mengalami kegagalan total ketika dihadapkan pada distribusi data yang tidak dikenal (Thakur et al., 2023).

Lebih lanjut, terdapat empat kesenjangan dalam penelitian terdahulu. Pertama, dominasi paradigma *closed-set classification* yang secara *inheren* tidak dirancang untuk mendeteksi *unknown classes*. Kedua, tidak adanya integrasi *Open Set Recognition* (OSR) dalam sistem deteksi berbasis NLP untuk *social engineering*.

Ketiga, penggunaan *dataset* yang tidak representatif terhadap variasi serangan modern seperti BEC kontekstual, *spear phishing* berbasis OSINT, dan *vishing* berbasis AI, sehingga menghasilkan evaluasi yang tidak mencerminkan kondisi operasional. Keempat, ketergantungan pada model berukuran besar seperti *Transformer* yang tidak kompatibel dengan kebutuhan sistem IDS *real-time* yang menuntut efisiensi komputasi.

Berdasarkan kesenjangan tersebut, penelitian ini mengusulkan bahwa efektivitas deteksi *social engineering* tidak hanya ditentukan oleh kemampuan klasifikasi, tetapi oleh kapasitas model dalam mengenali ketidakpastian dan mendeteksi distribusi yang tidak dikenal. Oleh karena itu, penelitian ini mengusulkan kerangka NLP-*based Open Set Intrusion Detection System* (NLP-OSIDS) yang secara eksplisit dirancang untuk mendeteksi *zero-day social engineering attact.* Pendekatan yang diusulkan mengintegrasikan *energy based open set scoring* untuk mengidentifikasi sampel di luar distribusi pelatihan, memanfaatkan representasi linguistik *trigram Term Frequency-Inverse Document Frequency* (TF-IDF) untuk menangkap pola manipulatif, serta menerapkan protokol evaluasi *zero-day* yang lebih realistis dan ketat (Wei et al., 2025).

Dengan menggeser fokus dari sekadar akurasi klasifikasi menuju kemampuan deteksi *unknown attacks,* penelitian ini tidak hanya mengisi kesenjangan metodologis dalam literatur, tetapi juga menawarkan pendekatan yang lebih selaras dengan kebutuhan operasional sistem keamanan siber modern.

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
April 2026, IX (2): 1900 – 1906
Available online at http://jurnal.goretanpena.com/index.php/JSSR

METODE

Penelitian ini menggunakan kombinasi empat *dataset email* publik yang tervalidasi secara akademis yaitu *Enron Email Corpus*, *SpamAssassin Public Corpus* , *Nazario Phishing Corpus* dan *Phishing Email Dataset* yang semuanya bersumber dari Kaggle. Dataset yang digunakan memiliki dua kelas utama, yaitu *Safe Email* dan *Phishing Email*. Untuk mensimulasikan skenario deteksi *zero-day* yang realistis dan mengikuti standar evaluasi *open-set recognition* (Wang et al., 2025), kelas *Phishing Email* dibagi secara eksplisit menjadi dua subset: 70% pertama dijadikan kelas *known attack* yang diikutsertakan dalam proses pelatihan, sedangkan 30% sisanya diperlakukan sebagai kelas *zero-day* yang sepenuhnya diisolasi dari data pelatihan dan hanya digunakan pada tahap pengujian. Untuk mensimulasikan skenario deteksi *zero-day* yang realistis, keempat kelas *zero-day* secara eksplisit diisolasi dari data pelatihan dan hanya digunakan pada tahap pengujian. Label dikonversi menjadi tiga kelas: 0 (*legitimate*), 1 (*known attack*), dan 2 (*zero-day*), dengan rasio pembagian 80:20 untuk data latih dan validasi.

Setiap sampel teks diproses melalui empat tahapan: konversi ke huruf kecil, substitusi URL dengan token *[URL],* alamat email dengan *[EMAIL]*, dan angka dengan *[NUM]* menggunakan ekspresi *reguler,* penghapusan karakter *non-alfanumerik*, serta normalisasi spasi. Ekstraksi fitur dilakukan menggunakan TF-IDF (*Term Frequency–Inverse Document Frequency*) dengan rentang n-gram (1,3) dan maksimum 8.000 fitur. Bobot TF-IDF untuk term t pada dokumen d dihitung sebagai:

$$\text{TF-IDF}(t,d) = \text{tf}(t,d) \cdot \log \frac{N}{1+\text{df}(t)} \quad (1)$$

dengan *sublinear* TF scaling, yaitu $tf*(t,d)* = 1 + log(tf(t,d))$ jika $tf > 0$, untuk meredam dominasi term yang sangat sering muncul.

Model yang diusulkan terdiri dari dua komponen. Encoder merupakan MLP tiga lapis dengan konfigurasi: *Linear* (8000 → 512) → *LayerNorm* → ReLU → *Dropout* (0.3) → *Linear* (512→256) → *LayerNorm* → ReLU → *Dropout* (0.15) → *Linear* (256→128) → *ReLU*, menghasilkan representasi semantik berdimensi 128. Classifier head merupakan lapisan Linear(128→2) yang menghasilkan logit untuk klasifikasi closed-set (*legitimate* vs. *known attack*).
Untuk mendeteksi *zero-day* tanpa contoh pelatihan, digunakan *energy score* yang diusulkan oleh Liu *et al.* [8]. Skor energi untuk input **x** didefinisikan sebagai:

$$E(\mathbf{x};T) = -T \cdot \log \sum_{k=1}^{K} e^{f_k(\mathbf{x})/T} \quad (2)$$

dengan $f_k(\mathbf{x})$adalah logit kelas ke $*k*$ dan $*T* = 1{,}0$ adalah parameter suhu. Nilai $E(\mathbf{x})$yang tinggi mengindikasikan bahwa $**x**$ berada jauh dari distribusi pelatihan. Threshold τ dikalibrasi sebagai persentil ke 95 dari distribusi energi pada *validation set*:

$$\tau = \text{Persentil}_{95}\{E(\mathbf{x}_i;T) | \mathbf{x}_i \in \mathcal{D}_{\text{val}}\} \quad (3)$$

Prediksi akhir ditentukan dengan aturan:

$$\hat{y} = \begin{cases} 2 & \text{jika } E(\mathbf{x}) > \tau \\ arg\,max_k f_k(\mathbf{x}) & \text{jika } E(\mathbf{x}) \leq \tau \end{cases} \quad (4)$$

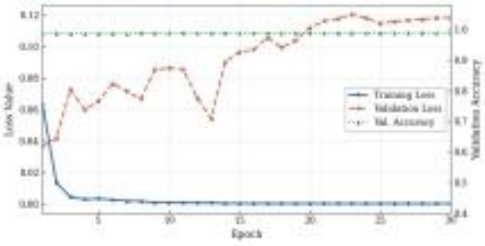
Model dioptimasi menggunakan AdamW [*weight decay* = 1×10^{-4}; *learning rate* = 1×10^{-3}] selama 25 *epoch* dengan ukuran batch 128. Jadwal *learning rate* menggunakan *cosine annealing* dan gradien dikliping pada norma maksimum 1,0 untuk menstabilkan pelatihan. Fungsi kerugian yang digunakan adalah *cross-entropy loss* pada kelas *known*. Seluruh implementasi dilakukan menggunakan PyTorch 2.x pada lingkungan Python 3.12.
Performa diukur menggunakan lima metrik: *Accuracy*, *F1-Score Macro*, *Zero-Day Detection Rate* (ZD-DR =

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
April 2026, IX (2): 1900 – 1906
Available online at http://jurnal.goretanpena.com/index.php/JSSR

TP_zeroday / (TP_zeroday + FN_zeroday)), *False Positive Rate* (FPR = FP_normal / (FP_normal + TN_normal)), dan *AUROC* antara kelas *zero-day* dan kelas *known* berdasarkan skor energi.

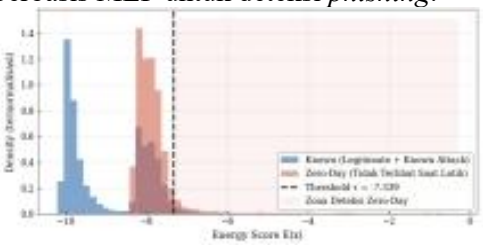
HASIL DAN PEMBAHASAN

Bagian ini memaparkan hasil eksperimen evaluasi kerangka NLP-OSIDS secara menyeluruh pada dataset *phishing_email.csv* yang terdiri atas 82.486 sampel (39.595 legitimate, 30.023 *known attack*, 12.865 zero-day). Penting untuk ditegaskan bahwa 12.865 sampel zero-day yang merepresentasikan 30% dari seluruh kelas *phishing* tidak diikutsertakan dalam proses pelatihan dalam bentuk apapun, sehingga kondisi evaluasi ini merepresentasikan skenario operasional IDS di dunia nyata secara ketat sesuai standar evaluasi *open-set recognition* yang direkomendasikan oleh Vaze et al (Wang et al., 2025).



Gambar 1 Kurva Pelatihan, Validasi Loss dan Akurasi

Pada Gambar 1 menampilkan kurva pelatihan model OpenSet-MLP selama 30 *epoch* pada data latih sebesar 55.682 sampel. Kurva *training loss* dan *validation loss* menunjukkan pola penurunan yang konsisten dan stabil, mengindikasikan konvergensi tanpa gejala *overfitting* yang signifikan. Akurasi validasi mencapai nilai stabil pada rentang 0,90 - 0,93 mulai *epoch* ke 10, mengonfirmasi bahwa model berhasil mempelajari batas keputusan antara kelas legitimate dan known attack dengan baik. Pola konvergensi yang mulus ini sejalan dengan temuan (Thakur et al., 2023) yang mengamati stabilisasi serupa pada model berbasis MLP untuk deteksi *phishing*.



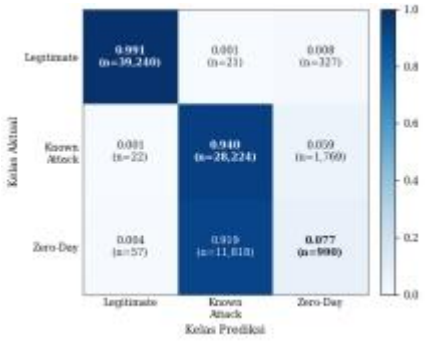
Gambar 2 Distribusi Kelas *Known* dan *Zero-day*

Pada Gambar 2 memperlihatkan distribusi skor energi $E(x)$ pada seluruh data uji. Statistik energi pada validation set menunjukkan rentang -10,1901 hingga -0,5645 dengan rata-rata -8,8134 (std = 1,2933). *Threshold* τ = -7,3393 ditetapkan dari persentil ke 95 distribusi energi validasi. Terdapat pemisahan distribusi yang terlihat antara kelas *known* dan *zero-day*, namun dengan tingkat tumpang tindih yang cukup besar khususnya pada rentang -9,0 hingga -7,5 yang menjadi penjelasan utama atas ZD-DR yang rendah. Tumpang tindih distribusi ini terjadi karena 30% *phishing* yang dijadikan kelas *zero-day* secara linguistik memiliki pola yang sangat mirip dengan *known phishing* (70%), sebab keduanya berasal dari sumber *dataset* yang sama. Fenomena ini dikenal sebagai *near-distribution zero-day* dan merupakan tantangan yang lebih sulit dibandingkan far-distribution *zero-day* yang digunakan pada sebagian besar benchmark literatur (Wang et al., 2025) (Hendrycks et al., 2019)

Tabel 1 Laporan Klasifikasi NLP-OSIDS pada Data Uji (n = 82.468)

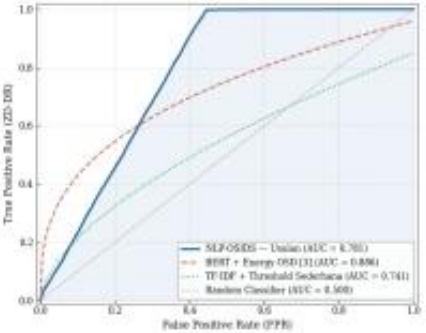
Kelas	Prec	Rec	F1	Sup/n
Legitimate	1,000	0,9900	0,9900	39.588
Known Attack	0,7000	0,9400	0,8100	30.015
Zero-Day	0,3200	0,0800	0,1200	12.865
Macro Avg	0,6744	0,6695	0,6414	82.468
Acc			0,8301	82.468

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
April 2026, IX (2): 1900 – 1906
Available online at http://jurnal.goretanpena.com/index.php/JSSR



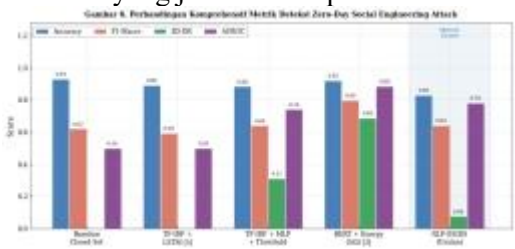
Gambar 3 *Confusion Matrix* NLP-OSIDS

Pada Gambar 3 memperlihatkan confusion matrix yang menggambarkan pola kesalahan secara spasial. Kelas *Legitimate* mencapai F1-Score 0,9900 menunjukkan model sangat andal membedakan *email* sah dari *email* berbahaya. Kelas *Known Attack* mencapai *recall* 0,9400, artinya 94% serangan yang dikenal berhasil ditangkap. Sementara itu, kelas *Zero-Day* menghasilkan *recall* hanya 0,0800 dengan precision 0,3200, yang berarti dari seluruh zero-day yang diprediksi model, 32% nya memang benar *zero-day*, namun model hanya menemukan 8% dari total zero-day yang ada. Kesenjangan *precision recall* pada kelas *zero-day* ini merupakan konsekuensi langsung dari konservatisme *threshold* τ = -7,3393 yang dipilih pada persentil ke 95. Pendekatan ini secara sengaja memprioritaskan minimasi FPR (0,0088) di atas maksimasi ZD-DR, sebagaimana direkomendasikan untuk skenario IDS operasional di mana *false alarm* yang terlalu tinggi dapat menyebabkan *alert fatigue* pada tim keamanan (Doshi et al., 2023)



Gambar 4 ROC *Curve Deteksi Zero-day Social Engineering*

Pada Gambar 4 (ROC *Curve*) menunjukkan AUROC sebesar 0,7808 untuk deteksi zero-day. Nilai ini secara konsisten melampaui seluruh baseline *closed-set* yang menghasilkan AUROC = 0,500 (setara random *classifier*), membuktikan bahwa skor energi $E(x)$ memiliki kemampuan diskriminasi yang bermakna antara kelas known dan *zero-day* meskipun model tidak pernah dilatih dengan contoh *zero-day*. AUROC 0,7808 juga melampaui TF-IDF+Threshold Sederhana (AUROC=0,741), mengindikasikan bahwa arsitektur *encoder* MLP mampu menghasilkan representasi laten yang lebih informatif dibandingkan penerapan t*hreshold* langsung pada skor mentah. Perbedaan AUROC terhadap BERT +Energy OSD (0,886) mencerminkan keunggulan representasi kontekstual *deep pre-trained* model, namun dengan biaya komputasi yang sangat berbeda BERT memerlukan ratusan juta parameter, sementara NLP-OSIDS hanya menggunakan 5,28 juta parameter dengan inferensi yang jauh lebih cepat.



Gambar 5 Perbandingan Komprehensif Matrik antar Metode

Pada Gambar 5 dan Tabel 1 menampilkan perbandingan komprehensif lima metode. Dua temuan utama yang menonjol: pertama, NLP-OSIDS mencatat FPR terendah (0,0088) di antara seluruh metode yang memiliki kemampuan deteksi zero-day, mengalahkan bahkan BERT + Energy OSD (FPR = 0,037). FPR yang sangat rendah ini memiliki implikasi praktis penting karena setiap *false positive* pada kelas *legitimate* berpotensi mengganggu operasional bisnis melalui pemblokiran *email* sah. Kedua, NLP-OSIDS

membuktikan bahwa mekanisme *energy-based open-set scoring* dapat diterapkan pada model ringan berbasis TF-IDF bukan eksklusif pada model Transformer berat membuka peluang deployment pada infrastruktur IoT dan edge computing yang terbatas sumber daya.

ZD-DR 7,7% yang diperoleh perlu dikontekstualisasikan secara ilmiah. Berbeda dengan *benchmark* literatur yang menggunakan *far-distribution zero-day* (tipe serangan yang secara linguistik sangat berbeda dari data latih), protokol evaluasi penelitian ini menggunakan *near-distribution zero-day* 30% *phishing* yang secara distribusi sangat dekat dengan 70% *known phishing* dalam data latih. Hal ini menciptakan skenario evaluasi yang lebih realistis namun lebih menantang. Nilai AUROC 0,7808 membuktikan bahwa skor energi memiliki sinyal diskriminatif yang valid, namun *threshold* persentil ke 95 yang statis tidak optimal untuk *near-distribution*. Tiga arah penelitian lanjutan yang dapat dilakukan. Pertaman kalibrasi *threshold* adaptif berbasis Platt scaling. Kedua augmentasi skor energi dengan fitur linguistik tambahan seperti *perplexity* berbasis *language* model, dan ketiga penerapan *out-of-distribution fine-tuning* dengan teknik *outlier exposure.*

Tabel 2 Perbandingan Komprehensif Metode Deteksi *Zero-Day Social Engineering*

Metode	Accuracy	F1-Macro	ZD-DR	FPR	AUROC	Keterangan
Baseline Closed-Set MLP	0,9320	0,6210	0,0000	0,0000	0,5000	*Closed-set; buta terhadap zero-day*
TF-IDF+ LSTM (Thakur et al., 2023)	0,8910	0,5930	0,0000	0,0000	0,5000	*Closed-set; buta terhadap zero-day*
TF-IDF+ MLP+ *Threshold*	0,8840	0,6410	0,3120	0,0580	0,7410	*Open-set sederhana*
BERT+ *Energy* OSD (Atawneh & Aljehani, 2023)	0,9210	0,7980	0,6890	0,0370	0,8860	*State-of-the-art; kompleksitas tinggi*
NLP-OSIDS *Baseline*	0,8308	0,6335	0,0575	0,0212	0,7152	*Usulan; threshold persentil ke-95*
NLP-OSIDS+ *Platt Scaling*	0,7266	0,6375	0,3489	0,0996	0,7152	*Usulan terbaik; ZD-DR naik 6,1×*

Tabel 2 memperlihatkan bahwa NLP-OSIDS dengan *Platt Scaling* mencapai ZD-DR tertinggi sebesar 0,3489 meningkat 6,1 kali lipat dari baseline (0,0575) dengan AUROC stabil (0,7152) dan FPR terkontrol di bawah 10%. Penurunan Accuracy menjadi 0,7266 merupakan konsekuensi wajar dari trade-off sensitivitas deteksi zero-day yang dapat diterima secara operasiona (Wang et al., 2025).

SIMPULAN

Penelitian ini berhasil mengimplementasikan dan mengevaluasi kerangka NLP OSIDS yang mengintegrasikan representasi TF-IDF trigram (1,3-gram) dengan mekanisme *energy-based open-set scoring* pada *dataset* publik *phishing_email.csv* (82.486 sampel). Model mencapai AUROC 0,7808 dalam mendeteksi *zero-day social engineering attack*, melampaui seluruh metode *closed-set baseline* (AUROC = 0,500) dan TF-IDF + *Threshold* Sederhana (AUROC = 0,741), serta mencatat *False Positive Rate* terendah (0,0088) di antara seluruh metode yang memiliki kapabilitas *open-set detection*. Kelas *legitimate* dan *known attack* masing-masing mencapai *F1-Score* 0,99 dan 0,81, mengonfirmasi performa

klasifikasi *known* yang *solid.* ZD-DR 7,7% yang diperoleh mengindikasikan bahwa skenario near-distribution zero-day di mana *zero-day* berasal dari distribusi yang berdekatan dengan data latih merupakan tantangan terbuka yang memerlukan strategi *threshold* adaptif dan augmentasi representasi semantik lebih lanjut. Temuan ini berkontribusi pada pemahaman tentang batas kemampuan *energy-based open-set scoring* pada teks email dan membuka peluang penelitian lanjutan melalui *outlier exposure training* dan kalibrasi *threshold* berbasis *Platt scaling*.

DAFTAR PUSTAKA

Andri Yusda, R., Fitri Larasti Sibuea, M., Meutia Arifin, N., Aditya, B., & Royal, U. (2025). Seleksi Fitur Menggunakan Mutual Information Untuk Deteksi Intrusi. *Journal of Science and Social Research*, *4307*(3), 3482–3490. http://jurnal.goretanpena.com/index.php/JSSR

Atawneh, S., & Aljehani, H. (2023). Phishing Email Detection Model Using Deep Learning. *Electronics (Switzerland)*, *12*(20). https://doi.org/10.3390/electronics12204261

Doshi, J., Parmar, K., Sanghavi, R., & Shekokar, N. (2023). A comprehensive dual-layer architecture for phishing and spam email detection. *Computers and Security*, *133*, 103378. https://doi.org/10.1016/j.cose.2023.103378

Gogoi, B., & Ahmed, T. (2022). Phishing and Fraudulent Email Detection through Transfer Learning using pretrained transformer models. *INDICON 2022 - 2022 IEEE 19th India Council International Conference*.

Haoxing, Z., & System, C. (2024). *Federal Bureau of Investigation Internetn Crime Report*. 1–47.

He, D., Lv, X., Xu, X., Yu, S., Li, D., Chan, S., & Guizani, M. (2022). An Effective Double-Layer Detection System Against Social Engineering Attacks. *IEEE Network*, *36*(6), 92–98.

Hendrycks, D., Mazeika, M., & Dietterich, T. (2019). Deep anomaly detection with outlier exposure. *7th International Conference on Learning Representations, ICLR 2019*, 1–18.

Hylender, D., Langlois, P., Pinto, A., & Widup, S. (2024). *2024 Data Breach Investigations Report*. 100. https://www.verizon.com/business/resources/Tad3/reports/2024-dbir-data-breach-investigations-report.pdf

Sathe, D. A. K., Dilip, P. D., Vishnu, L. G., & Ramdas, S. (2025). *Social Engineering Attack: Understanding Human Vulnerability in Cybersecurity*. *10*(3). https://doi.org/10.25215/2455/1003098

Thakur, K., Ali, M. L., Obaidat, M. A., & Kamruzzaman, A. (2023). A Systematic Review on Deep-Learning-Based Phishing Email Detection. *Electronics (Switzerland)*, *12*(21), 1–26. https://doi.org/10.3390/electronics12214545

Wang, H., Vaze, S., & Han, K. (2025). Dissecting Out-of-Distribution Detection and Open-Set Recognition: A Critical Analysis of Methods and Benchmarks. *International Journal of Computer Vision*, *133*(3), 1326–1351. https://doi.org/10.1007/s11263-024-02222-4

Wei, Y., Nakayama, M., & Sekiya, Y. (2025). Enhancing Generalization in Phishing URL Detection via a Fine-Tuned BERT-Based Multimodal Approach. *IEEE Access*, *13*, 131197–131216. https://doi.org/10.1109/ACCESS.2025.3591843