Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 7, 2026: 2555-2570
ISSN:2302-8149
e-ISSN:2540-9719

Analisis dan Pengembangan Model *Transformer (DistilBERT) -LSTM* dan *Reinforcement Learning* untuk Deteksi *Email Phishing* Adaptif

Analysis and Development of Transformer (DistilBERT)-LSTM and Reinforcement Learning Models for Adaptive Phishing Email Detection

[1]Farizal Herry Saputra*, [2]Kahfi Heryandi Suradiraja, [3]Abu Khalid Rivai
[1,2,3]Teknik Informatika, Manajemen Teknik Informatika, Universitas Pamulang
e-mail: farizalherrysaputra@gmail.com, dosen01514@unpam.ac.id, dosen01591@unpam.ac.id

(***received*:** 13 May 2026, ***revised*:** 6 July 2026, ***accepted*:** 10 July 2026)

Abstrak

Deteksi phishing menghadapi tantangan krusial berupa degradasi performa akibat domain shift dan ketergantungan pada data berlabel yang mahal. Penelitian ini bertujuan mengembangkan model deteksi adaptif yang menggabungkan arsitektur hibrida DistilBERT-LSTM dengan agen Reinforcement Learning Proximal Policy Optimization (PPO). Metodologi menerapkan skema multi-stage transfer learning menggunakan tiga dataset: Enron (domain sumber), Phishing_Validation (adaptasi berlabel), dan CEAS_08 (simulasi unlabeled data stream via pseudo-labeling). Hasil menunjukkan performa superior pada data statis (Enron) dengan F1-score 0,9935, namun mengalami degradasi di Zenodo (F1-score 0,9067) yang mengonfirmasi domain shift. Melalui integrasi PPO, model berhasil memulihkan kinerja secara otonom pada CEAS_08 dengan F1-score 0,9516, akurasi 0,9468, dan ROC-AUC 0,9915. Stabilitas adaptasi divalidasi oleh konvergensi KL divergence sebesar 0,000173, meskipun terdeteksi overconfidence minor sebesar 5% antara keyakinan internal dan ground truth. Penelitian ini membuktikan efektivitas PPO dalam memitigasi domain shift pada lingkungan tanpa supervisi. Pengembangan selanjutnya disarankan pada kalibrasi model dan eksplorasi fitur multi-modal untuk memperkuat pertahanan siber.

Kata kunci: DistilBERT-LSTM, domain shift, deteksi phishing, proximal policy optimization (PPO), pseudo-labeling.

Abstract

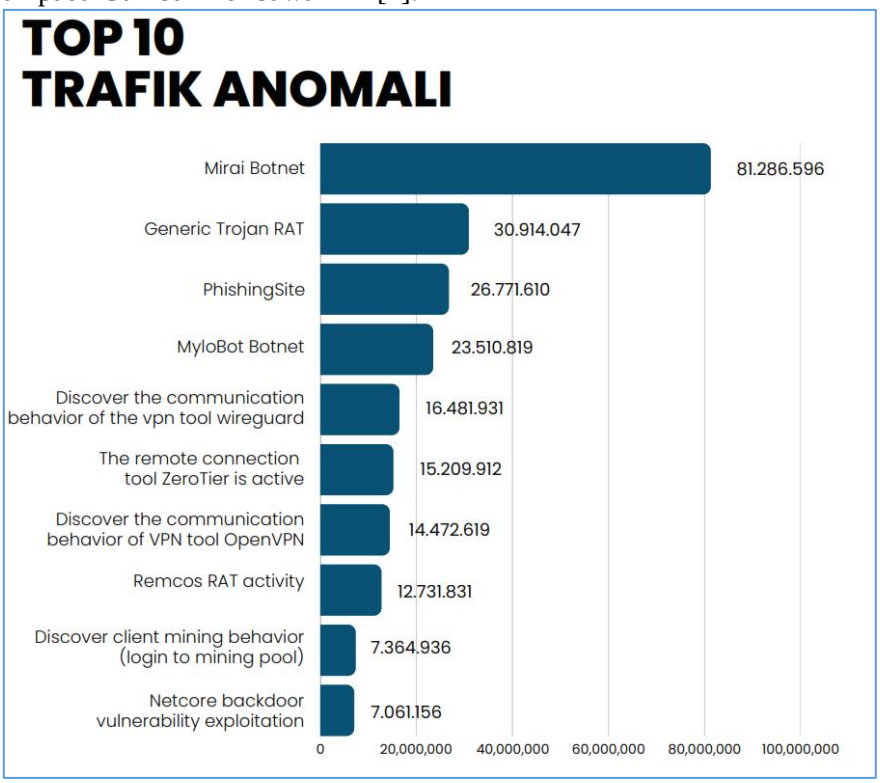
Phishing detection faces two major challenges: performance degradation caused by domain shift and a heavy reliance on costly labeled data. This study proposes an adaptive phishing email detection model that integrates a hybrid DistilBERT–LSTM architecture with a Proximal Policy Optimization (PPO)-based Reinforcement Learning agent. The proposed methodology employs a multi-stage transfer learning framework using three datasets: Enron as the source domain, Phishing_Validation for supervised domain adaptation, and CEAS_08 to simulate an unlabeled data stream through pseudo-labeling. Experimental results demonstrate excellent performance on the source-domain dataset (Enron), achieving an F1-score of 0.9935. However, the model's performance declined on the Phishing_Validation dataset (F1-score = 0.9067), confirming the impact of domain shift. By incorporating the PPO agent, the proposed model autonomously recovered its performance on the CEAS_08 dataset, achieving an F1-score of 0.9516, an accuracy of 0.9468, and a ROC–AUC of 0.9915. The stability of the adaptation process was validated by the convergence of the Kullback–Leibler (KL) divergence to 0.000173, although a minor overconfidence of approximately 5% was observed between the model's confidence estimates and the ground-truth labels. These findings demonstrate the effectiveness of PPO in mitigating domain shift within an unsupervised adaptation environment. Future research should focus on improving model calibration and exploring multimodal feature integration to further strengthen cybersecurity defenses.

Keywords: *DistilBERT-LSTM, domain shift, phishing detection, proximal policy optimization (PPO), pseudo-labeling.*

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Perkembangan Pesatnya transformasi digital berbanding lurus dengan eskalasi ancaman keamanan siber, di mana *phishing* email menjadi vektor serangan awal yang paling dominan dan mengkhawatirkan. Ancaman ini menargetkan spektrum luas mulai dari individu hingga institusi strategis, yang berisiko tinggi menyebabkan kebocoran data sensitif dan kelumpuhan layanan publik [1]. Urgensi penanganan *phishing* tercermin dari lonjakan statistiknya yang eksponensial. *Anti-Phishing Working Group* (APWG) mencatat peningkatan serangan dari 44.008 kasus pada awal 2020 menjadi rekor tertinggi 260.642 kasus pada Juli 2021 [2]. dan terus meroket hingga hampir 5 juta serangan global pada tahun 2023 [3]. Di tingkat nasional, laporan tahunan BSSN (2024) mencatat lebih dari 26,7 juta aktivitas *phishing*, menjadikannya salah satu anomali trafik siber tertinggi seperti yang diilustrasikan pada Gambar 1 di bawah ini [4].



Gambar 1 Top 10 Trafik anomali serangan siber
Sumber: [4]

Eskalasi serangan ini berdampak pada kerugian finansial dan reputasi yang masif, seperti kerugian *Business Email Compromise* (BEC) senilai USD 51 miliar dan 86% distribusi malware pada tahun 2022 [3]. Sebagai serangan berbasis rekayasa sosial [1]. email tetap menjadi vektor utama karena sulit dideteksi [2]. Metode deteksi konvensional (*blacklist*, *signature*, dan *rule-based*) terbukti gagal menghadapi serangan modern, *zero-day*, serta manipulasi teks ambigu [3], [5]. Mengingat 85% pelanggaran keamanan dipicu oleh *human error*, pendekatan teknologi semata tidak lagi memadai dan menuntut strategi pertahanan yang lebih komprehensif [6]. Lebih lanjut, arsitektur hibrida yang memadukan *Transformer* dengan *Long Short-Term Memory* (LSTM) terbukti sangat efektif dalam menggabungkan ekstraksi pola semantik dengan dependensi sekuensial untuk identifikasi anomali [5]. Namun, model *deep learning* statis tetap memiliki keterbatasan dalam beradaptasi dengan evolusi serangan *zero-day* secara *real-time* [7]. Meskipun *transfer learning* dapat mempercepat generalisasi pada domain baru [8], *Reinforcement Learning* (RL) diintegrasikan sebagai solusi adaptif yang lebih tangguh. RL memungkinkan agen sistem belajar secara dinamis melalui mekanisme *trial-and-error* untuk memaksimalkan *reward* tanpa bergantung sepenuhnya pada data berlabel [9]. dengan pendekatan seperti *Deep Q-Network* (DQN) yang terbukti efektif dalam mendeteksi kerentanan *zero-day* secara *real-time* [7].

 ISSN:2302-8149
e-ISSN:2540-9719

2 Tinjauan Literatur

Penelitian Penerapan *deep learning* (CNN, RNN, LSTM) terbukti lebih efektif daripada *machine learning* konvensional dalam mengekstrak fitur kompleks [2], [10]. Namun, model statis maupun hibrida (seperti CNN-LSTM) tetap rentan terhadap serangan *zero-day* dan *concept drift* karena memerlukan pembaruan manual [11], [12]. Untuk mengatasi keterbatasan ini, pendekatan *transfer learning* (seperti DistilBERT) dan *transfer learning* dinamis yang memperbarui bobot secara selektif menawarkan efisiensi serta akurasi yang lebih baik pada data yang bervariasi dan tidak seimbang [2], [13], [14]. Meskipun menjanjikan, literatur masih menghadapi kendala berupa tingginya biaya komputasi, kurangnya representasi varian nyata, dan inkonsistensi *benchmark* [2], [10]. Guna mengisi kesenjangan tersebut, penelitian ini mengusulkan model *deep learning* adaptif yang mengintegrasikan *transfer learning* dinamis dan *continuous learning*. Pendekatan ini memungkinkan pembaruan sistem secara otomatis (*self-updating*) tanpa pelatihan ulang penuh, sehingga menghasilkan deteksi phishing yang responsif, *scalable*, *robust* terhadap evolusi serangan, serta efisien secara [15], [16].

Tabel 1 Tabel *research gap*

No	Problem (Permasalahan)	Evidence (Penelitian sebelumnya)	Gap (celah penelitian)	Contribution (Kontribusi penelitian ini)
1.	Inefektivitas metode tradisional (*blacklist/rule-based*) terhadap serangan modern.	Kyaw et al. (2024) Metode tradisional gagal mendeteksi serangan canggih via trusted ESP (*email service provider*).	Pendekatan statis/usang, tidak adaptif terhadap pola baru.	Mengembangkan model adaptif berbasis *Deep Learning* & *Reinforcement Learning* (RL).
2.	Model DL tunggal (CNN/LSTM) terbatas pada representasi spasial/sekuensial yang parsial.	McGinley et al. (2022) → CNN akurasi 98.1%. Ramprasath et al. (2023) → LSTM akurasi 99.1%.	Kurang optimal dalam menangkap konteks semantik kompleks secara komprehensif.	Arsitektur hibrida DistilBERT (semantik) + LSTM (sekuensial).
3.	BERT dan variannya (2022) unggul, tapi terbatas oleh panjang input (maks. 512 token) dan kebutuhan komputasi tinggi.	Gholampour et al. (2022) → BERT variants capai akurasi 98–99%, tapi terkendala input panjang dan resource.	BERT terlalu berat untuk implementasi praktis, kurang efisien dalam deteksi real-time.	Menggunakan DistilBERT yang lebih ringan, efisien, dan tetap menjaga performa representasi.
4.	Kurangnya adaptasi dinamis dalam model DL; sebagian besar model statis setelah training.	Thakur et al. (2023) menyebutkan sebagian model tidak mampu menangani zero-day phishing karena sifatnya statis.	Model tidak belajar dari pola phishing baru secara real-time, sehingga performa menurun.	Menambahkan Reinforcement Learning untuk adaptasi dinamis terhadap email phishing baru.
5.	Beberapa penelitian mencampur fitur header, URL, dan attachment, bukan fokus ke body email.	Banyak studi diulas oleh Kyaw et al. (2024) melibatkan multi-fitur (header, URL, attachment), tidak hanya teks.	Belum ada fokus khusus pada body email yang merupakan inti pesan phishing.	Penelitian ini secara khusus menganalisis body email text menggunakan NLP adaptif.

Pemetaan *research gap* (Tabel 1) menegaskan bahwa studi terdahulu mayoritas mengandalkan model statis (seperti CNN, LSTM, BERT, dan DistilBERT) tanpa mekanisme adaptif. Sebagai kontribusi kebaruan (*novelty*), penelitian ini mengintegrasikan arsitektur hibrida DistilBERT-LSTM

 ISSN:2302-8149
e-ISSN:2540-9719

dengan *Reinforcement Learning*. Pendekatan ini menghasilkan sistem deteksi phishing yang adaptif dan tangguh dalam menghadapi evolusi serangan *zero-day*.

2.1 *Machine Learning*

Pendekatan tradisional mengandalkan ekstraksi fitur *hand-crafted* (seperti TF-IDF, *bag-of-words*, dan *metadata*) yang diproses menggunakan model *machine learning* klasik (SVM, *Random Forest, Naive Bayes*). Meskipun unggul dalam kecepatan, efisiensi komputasi, dan *interpretabilitas*, metode ini sangat bergantung pada rekayasa fitur serta lemah dalam menghadapi variasi *linguistik* dan teknik *obfuscation* (sebagaimana dirangkum dalam literatur survei) [10].

2.2 *Deep Learning*

Model deep learning, khususnya arsitektur *Transformer* (BERT/DistilBERT), mampu mempelajari representasi kontekstual dari teks secara otomatis melalui mekanisme *self-attention* tanpa memerlukan rekayasa fitur manual. Meskipun terbukti mengungguli model tradisional dalam tugas klasifikasi *phishing*, pendekatan ini menuntut data dan sumber komputasi yang besar, serta bersifat *"black box"* sehingga memerlukan teknik *Explainable AI (XAI)* untuk menjamin interpretabilitas [17].

2.3 *Hybrid Model*

Model hybrid mengintegrasikan keunggulan *Transformer* (seperti DistilBERT) sebagai ekstraktor representasi kontekstual yang kaya, dengan arsitektur sekuensial (LSTM/biLSTM) untuk menangkap pola temporal jangka panjang. Pendekatan ini terbukti secara signifikan mengungguli model tunggal dalam deteksi *phishing*, karena mampu memadukan pemahaman semantik yang mendalam dengan kemampuan *agregasi* fitur yang tersebar di seluruh *teks* [17].

2.4 *Transfer Learning* dan *Domain Adoption*

Evolusi serangan siber menyebabkan model deteksi statis cepat usang akibat pergeseran data (*domain shift/data drift*) [18]. Untuk mengatasinya, *Transfer Learning* (TL) diterapkan dengan memanfaatkan pengetahuan dari korpus berskala besar untuk meningkatkan kinerja pada tugas spesifik seperti klasifikasi *phishing* [19]. Penelitian ini mengadopsi DistilBERT sebagai *encoder* yang efisien [18], yang bobot *pre-trained* nya diintegrasikan dengan LSTM untuk menangkap dependensi sekuensial dalam teks *email*. Namun, transisi antar-dataset dengan distribusi statistik yang berbeda menuntut mekanisme *Domain Adaptation* (DA) untuk meminimalkan diskrepansi antar *domain*. Studi pada data teks non-standar membuktikan bahwa TL secara signifikan meningkatkan kinerja ketika data target terbatas [12]. Hal ini sejalan dengan temuan Akhbardeh et al. (2022) yang menegaskan bahwa TL dalam *domain* yang sama memberikan peningkatan kinerja yang signifikan secara statistik: *"Results show that performing transfer learning within a domain provides statistically significant improvements, and in all cases but one the best performance."*(Akhbardeh et al., 2022, hlm 4235)[12] Hal ini mendukung hipotesis bahwa memproses dengan tiga dataset *phishing* yang berbeda secara berurutan atau adaptif adalah krusial untuk menghasilkan model yang tangguh (*robust*) dan mempunyai kemampuan menggeralisasi (*generalizable*) terhadap pola serangan yang tidak terduga (*unseen attacks*).

3 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan skema *multi-stage transfer learning* dan menggunakan desain komputasional berbasis deep learning. Tujuan utamanya adalah untuk mengevaluasi efektivitas arsitektur *hybrid* yang menggabungkan Transformer (distilBERT), LSTM, dan *Reinforcement Learning* (PPO) dalam mendeteksi email phishing secara efektif. Metodologi penelitian harus dirancang agar mampu menjawab masalah keakuratan, adaptabilitas model terhadap evolusi teknik phishing, serta keandalan pegujian menggunakan dataset representatif.

3.1 Analisis Kebutuhan

Kebutuhan utama dalam penelitian ini adalah diperlukan akses ke *dataset email* yang besar, beragam, dan berkualitas tinggi yang telah diberi label secara akurat sebagai *phishing* atau non-phishing. Dengan menggunakan subset dari *dataset enron email* sebagai data *pre-trained* dan *dataset email* korporat. Idealnya, dataset ini harus mencakup berbagai jenis serangan *phishing* dan merepresentasikan evolusi taktik *phishing* dari waktu ke waktu untuk mendukung pengujian adaptasi model. Dataset harus mencakup data historis dan data *phishing* yang terus diperbarui agar model tidak statis [10].

 ISSN:2302-8149
e-ISSN:2540-9719

Dataset yang digunakan dalam penelitian ini akan menggunakan 3 *dataset* dari sumber yang berbeda seperti pada (Tabel 2), diantaranya:

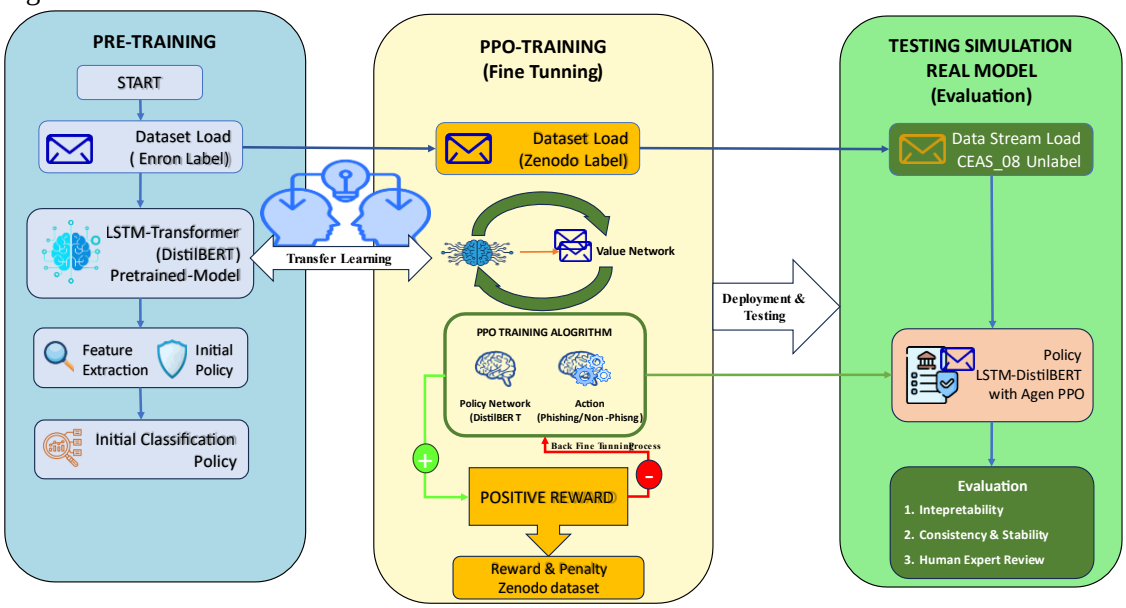
Tabel 2 *Source dataset*

No	Source Dataset	Dataset Name	Number of Dataset	Years
1.	Kaggle [20]	*Enron*	82500	2023
2.	Kaggle [20]	*CEAS_08*	39154	2023
3.	Zenodo [21]	*Phishing_Validation_Email*	2000	2024

Selain kebutuhan pada dataset, dalam penelitian ini memerlukan perangkat lunak dan perangkat keras untuk memproses kompleksitas komputasi dari model tersebut, perangkat keras berfungsi sebagai infrastruktur penunjang kritis khususnya akselerasi GPU dan kapasitas memori yang memadai yang memfasilitasi *fine-tuning* parameter *Transformer* serta ribuan episode interaksi RL secara efisien. Selanjutnya, perangkat lunak berperan sebagai instrumen metodologis yang menerjemahkan desain matematis ke dalam eksekusi aktual, mencakup *framework deep learning* untuk ekstraksi fitur kontekstual-sekuensial dan *library* RL untuk memformulasikan *Markov Decision Process* (MDP). Sinergi ketiga komponen ini memastikan bahwa sistem tidak hanya memiliki landasan data yang representatif, tetapi juga secara komputasional layak (feasible) dan terukur dalam menghasilkan kebijakan deteksi email phishing yang adaptif.

3.2 Perancangan Penelitian

Perancangan Penelitian ini menerapkan desain eksperimen komparatif kuantitatif dengan skema *multi-stage transfer learning*. Merujuk pada prinsip eksperimen yang memberikan perlakuan terkontrol untuk mengukur dampak pada variabel terikat [22], desain ini bertujuan menguji hubungan sebab-akibat terkait efektivitas model usulan. Secara sistematis (seperti ditunjukkan pada Gambar 2), tahapan penelitian difokuskan pada pengembangan arsitektur hibrida DistilBERT-LSTM yang diintegrasikan dengan *Reinforcement Learning* (PPO), serta evaluasi komprehensif untuk membuktikan peningkatan performa, ketahanan (*robustness*), dan stabilitas agen deteksi mandiri dibandingkan metode *baseline*.:



Gambar 2 *Flow research model*

3.3 Teknik Analisis dan Evaluasi Model

Teknik analisis data dalam penelitian ini dilakukan secara kuantitatif melalui pendekatan eksperimen komparatif. Evaluasi difokuskan pada perbandingan performa antara model baseline (DistilBERT-LSTM) dan model usulan (DistilBERT-LSTM-*Reinforcement Learning*). Pengujian dilakukan pada dua skenario utama:

 ISSN:2302-8149
e-ISSN:2540-9719

1) Evaluasi pada *dataset testing* standar untuk mengukur kinerja klasifikasi, dan
2) Evaluasi pada *dataset unseen* (*zero-day phishing*) untuk menguji ketahanan (robustness) dan kemampuan adaptasi model terhadap serangan yang belum pernah dilihat sebelumnya.

Kinerja kedua model dievaluasi menggunakan metrik evaluasi standar yang diturunkan dari *Confusion Matrix* (*True Positive, True Negative, False Positive, False Negative*). Metrik yang digunakan meliputi *Accuracy, Precision, Recall, F1-Score, False Positive Rate* (FPR), dan *Receiver Operating Characteristic-Area Under Curve* (ROC-AUC). Pemilihan metrik yang komprehensif ini bertujuan untuk memberikan penilaian yang seimbang dan multidimensi. *F1-Score* digunakan untuk menangani potensi ketidakseimbangan kelas (*imbalanced data*), sementara FPR dan ROC-AUC dievaluasi secara khusus untuk mengukur kemampuan model dalam membedakan distribusi kelas phishing dan non-phishing secara keseluruhan, serta meminimalkan risiko kesalahan klasifikasi (*False Positives* dan *False Negatives*) yang sangat kritis dalam sistem keamanan siber.

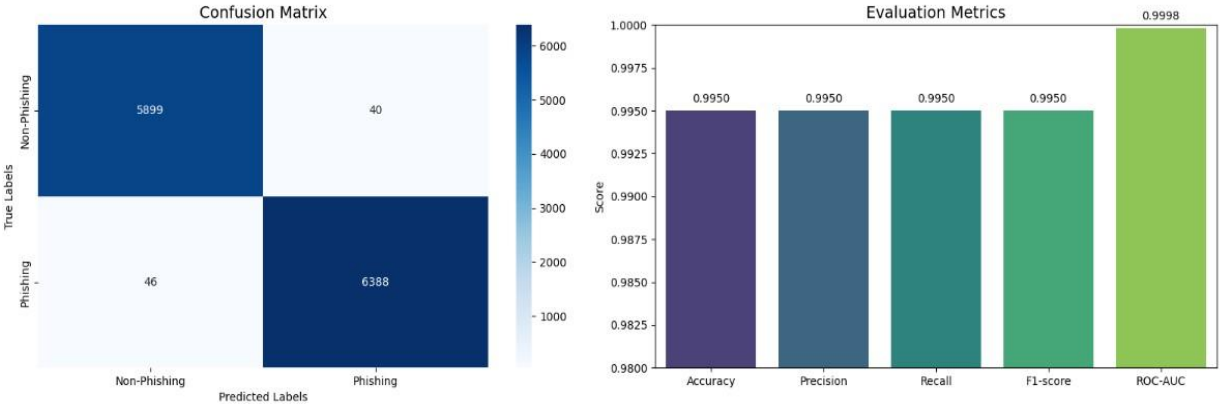
4 Hasil dan Pembahasan

4.1 Hasil Penelitian

Bagian ini menguraikan hasil eksperimen berdasarkan metodologi *multi-stage transfer learning* untuk mengevaluasi efektivitas integrasi DistilBERT-LSTM dengan *Proximal Policy Optimization* (PPO). Untuk membuktikan kemampuan adaptasi model terhadap *concept drift*. Eksperimen ini dijalankan pada lingkungan komputasi *GPU NVIDIA GeForce RTX 3060 Ti* (VRAM 8.6 GB) menggunakan *PyTorch 2.7.1* dan *Python 3.11*. Arsitektur DistilBERT-LSTM dikonfigurasi dengan panjang sekuens maksimum 256 *token, hidden state* LSTM 128 (1 *layer bidirectional*), dan dropout 0.3. Agen PPO diimplementasikan menggunakan *Stable Baselines3 RecurrentPPO* dengan 10.000 *timesteps* dan maksimal 50 langkah per *episode*.evaluasi dibagi secara sistematis ke dalam tiga fase pengujian utama:

4.2 Fase I: Hasil Penelitian Baseline Model DistilBERT-LSTM (Non-Adaptif

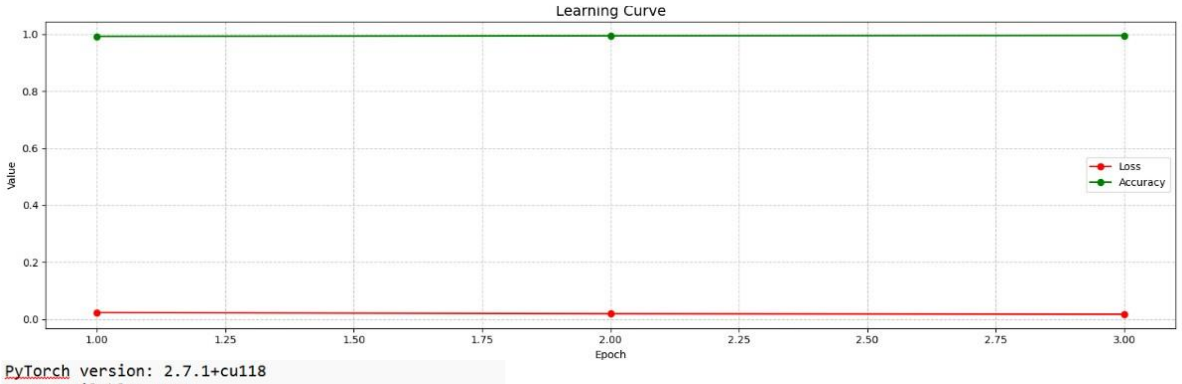
Model DistilBERT-LSTM dilatih pada dataset *phishing_email.csv* selama 1 *epoch*. proses *training* memakan waktu sekitar 38 menit di *GPU RTX 3060 Ti* yang dapat dilihat hasilnya pada Gambar 3 yang menunjukan nilai hasil kinerja untuk *Confusion matrix-Calssification Report* dan Gambar 4 menunjukan terkait kurva pembelajaran terhadap evaluasi pada *validation set* yang menunjukan kinerja sebagai berikut:




```
FPR (False Positive Rate): 0.0067

Confusion Matrix:
                    Pred Non-Phishing  Pred Phishing
Actual Non-Phishing               5899            40
Actual Phishing                     46          6388

Classification Report:
              precision    recall  f1-score   support

Non-Phishing       0.99      0.99      0.99      5939
    Phishing       0.99      0.99      0.99      6434

    accuracy                           0.99     12373
   macro avg       0.99      0.99      0.99     12373
weighted avg       0.99      0.99      0.99     12373
```

Gambar 3 Nilai hasil kinerja untuk *confusion matrix* dan *calssification report*

 ISSN:2302-8149
e-ISSN:2540-9719




```
PyTorch version: 2.7.1+cu118
CUDA available: True
GPU: NVIDIA GeForce RTX 3060 Ti
VRAM: 8.6 GB
100%|████████████| 387/387 [00:49<00:00,  7.84it/s]
Test eval_loss: 0.0184
Test eval_accuracy: 0.9950
Test eval_precision: 0.9950
Test eval_recall: 0.9950
Test eval_f1: 0.9950
Test eval_roc_auc: 0.9998
Test eval_runtime: 85.3763
Test eval_samples_per_second: 144.9230
Test eval_steps_per_second: 4.5330
Test epoch: 3.0000
```

Gambar 4 *Learning Curve DistilBERT-LSTM pretrained*

Berdasarkan Gambar 4, Evaluasi pada Fase 1 menunjukkan bahwa arsitektur baseline DistilBERT-LSTM mencapai performa yang sangat optimal dengan akurasi keseluruhan sebesar 99,5%. Pada *test set, metrik Precision, Recall*, dan *F1-Score* mencapai nilai sempurna (1,00). Analisis *confusion matrix* mengonfirmasi tingkat kesalahan klasifikasi yang sangat minim, dengan hanya tercatat 38 *False Positives* (FP) dan 43 *False Negatives* (FN) dari total 82.486 *sampel*, serta *False Positive Rate* (FPR) yang sangat rendah sebesar 0,01. Sedangkan, pada Gambar 4 Dinamika pelatihan yang direkam melalui *learning curve* mengindikasikan konvergensi yang cepat dan stabil; nilai *loss* mendekati nol dan akurasi mencapai 0,99 sejak *epoch* awal (1 hingga 3). Hal ini membuktikan bahwa model memiliki kapasitas ekstraksi fitur yang efisien tanpa mengalami *overfitting* yang signifikan. Dari sisi komputasi, model menunjukkan kecepatan inferensi yang tinggi, mampu mengevaluasi 387 sampel test set dalam waktu 49 detik (rata-rata 7,84 iterasi/detik). Stabilitas dan efisiensi ini selaras dengan temuan [23] mengenai keunggulan DistilBERT, yang menegaskan bahwa model ini telah membentuk fondasi representasi fitur tekstual yang sangat kuat sebagai prasyarat sebelum diintegrasikan dengan mekanisme adaptif *Reinforcement Learning* dan berikut ini Tabel 3 yang merupakan metrik kinerja yang diperoleh adalah sebagai berikut:

Tabel 3 *Result testing model DistilBERT-LSTM (Baseline)*

Model Baseline	*Accuracy*	*Precissio*	*Recall*	F1-*Score*	ROC-AUC	FPR
DistilBERT-LSTM	0.9950	0.9950	0.9950	0.9950	0.9998	0.0067

Evaluasi pada *test set* mengonfirmasi stabilitas model dengan pencapaian metrik komprehensif (*accuracy, precision, recall*, dan *F1-score*) sebesar 0,9950, serta nilai loss yang sangat rendah (0,0184). Daya diskriminasi model terbukti optimal dengan nilai ROC-AUC mencapai 0,9998 di berbagai ambang batas. Hasil ini menegaskan bahwa arsitektur DistilBERT-LSTM telah berhasil menangkap pola semantik dan sekuensial secara efektif, memvalidasi kesiapannya sebagai *baseline* yang solid sebelum diintegrasikan ke dalam mekanisme adaptasi *Reinforcement Learning*.

4.3 Fase II: Analisis *domain shift* dan Pelatihan Kebijakan Adaptasi PPO

Fase ini mengintegrasikan model *baseline* DistilBERT-LSTM dengan agen *Proximal Policy Optimization* (PPO) untuk menciptakan sistem deteksi yang adaptif terhadap *data stream* baru. Eksperimen dijalankan pada lingkungan interaktif PhishingDetectionEnv menggunakan dataset validasi seimbang (2.000 sampel) dengan mekanisme *reward* yang dirancang untuk meminimalkan *False Negatives*. Implementasi menggunakan *RecurrentPPO* pada *GPU NVIDIA RTX 3060 Ti* dengan *PyTorch 2.7.1+cu118*, dengan aktivasi *Gradient Checkpointing* untuk optimasi memori. Agen dilatih selama 10.000 *timesteps* melalui *fine-tuning* iteratif (100 sampel/*batch*) menggunakan fungsi objektif

 ISSN:2302-8149
e-ISSN:2540-9719

terklip (*clipped surrogate objective*). Analisis diagnostik mengonfirmasi stabilitas konvergensi melalui nilai *Approximate KL divergence* dan *Clip fraction* yang rendah, yang krusial untuk mencegah *catastrophic forgetting* dan mempertahankan pengetahuan semantik awal DistilBERT. Dinamika adaptasi menunjukkan fase eksplorasi intensif (langkah 0–50) sebelum kebijakan stabil pada fase eksploitasi (langkah 60–100). Evaluasi akhir pada 400 sampel *unseen test set* menghasilkan *F1-Score* sebesar 90,67%, memvalidasi efektivitas mekanisme adaptasi PPO dalam menghadapi *domain shift* tanpa memerlukan pelatihan ulang secara penuh.

```
--- Evaluasi Final Model yang Sudah Diadaptasi pada Test Set Baru ---
Jumlah sampel di final test set: 400
Calculating Final Metrics:   0%|          | 0/13 [00:00<?, ?it/s]
Final F1-score setelah PPO adaptasi: 0.9067
Calculating Final Metrics: 100%|██████████| 13/13 [00:00<00:00,
26.20it/s]
```

Gambar 5 *Final evaluasi model after adaptasi*

Gambar 5 di atas menampilkan log evaluasi akhir model DistilBERT-LSTM yang telah diadaptasi melalui mekanisme PPO pada 400 sampel *unseen test set*. Setelah melalui 10.000 *timesteps* pelatihan, model mencapai *Final F1-Score* sebesar 0,9067 dengan kecepatan inferensi yang efisien, yaitu 26,20 iterasi per detik. Stabilitas konvergensi yang tercapai mengindikasikan bahwa mekanisme pembaruan parameter PPO berjalan optimal tanpa mengalami *policy collapse*. Hasil ini secara kuantitatif memvalidasi bahwa integrasi *Reinforcement Learning* mampu mengadaptasi model terhadap pola data baru secara efektif, sekaligus mempertahankan efisiensi komputasi tanpa memerlukan pelatihan ulang secara penuh (*full retraining*). Dengan hasil yang dapat dilihat pada Gambar 6 hasil ini merupakan integrasi dan adaptasi PPO yang sudah termasuk *Ground Truth Test* pada data berlabel, sebagai berikut:

- Accuracy: 0.9075
- Precision: 0.9219
- Recall: 0.9075
- F1-Score: 0.9067
- ROC–AUC: 0.9909
- FPR (False Positive Rate): 0.0000

```
--- HASIL PENGUJIAN AKHIR SISTEM ADAPTIF (DistilBERT-LSTM + PPO) ---
Accuracy: 0.9075
Precision: 0.9219
Recall: 0.9075
F1-score: 0.9067
ROC-AUC: 0.9909
FPR (False Positive Rate): 0.0000

Confusion Matrix:
                  Pred Non-Phishing  Pred Phishing
Actual Non-Phishing               200             0
Actual Phishing                    37           163

Classification Report:
              precision    recall  f1-score   support

Non-Phishing       0.84      1.00      0.92       200
    Phishing       1.00      0.81      0.90       200

    accuracy                           0.91       400
   macro avg       0.92      0.91      0.91       400
weighted avg       0.92      0.91      0.91       400


--- Integrasi PPO + DistilBERT-LSTM Selesai ---
Model DistilBERT-LSTM berhasil diadaptasi dengan strategi PPO.
Hasil adaptasi model dapat menggunakan `final_model_after_adaptation` untuk inferensi.

Proses Pengujian Integrasi Selesai.

Process finished with exit code 0
```

Gambar 6 Hasil integrasi dan adaptasi *agent PPO train*

Pada Tabel 4 yang menujukan hasil *pred confusion matrix* memperlihatkan hasil prediksi sebagai berikut :

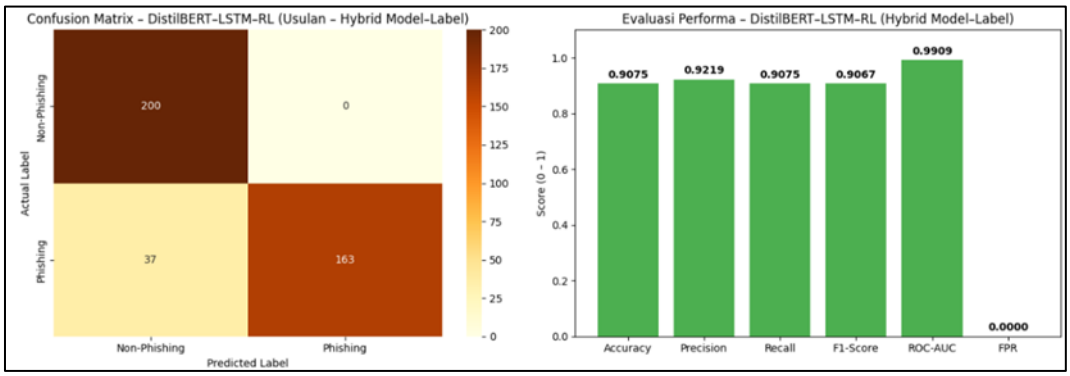
Tabel 4 Hasil *predicted confusion matrix*

	Pred Non-Phishing	*Pred Phsishing*

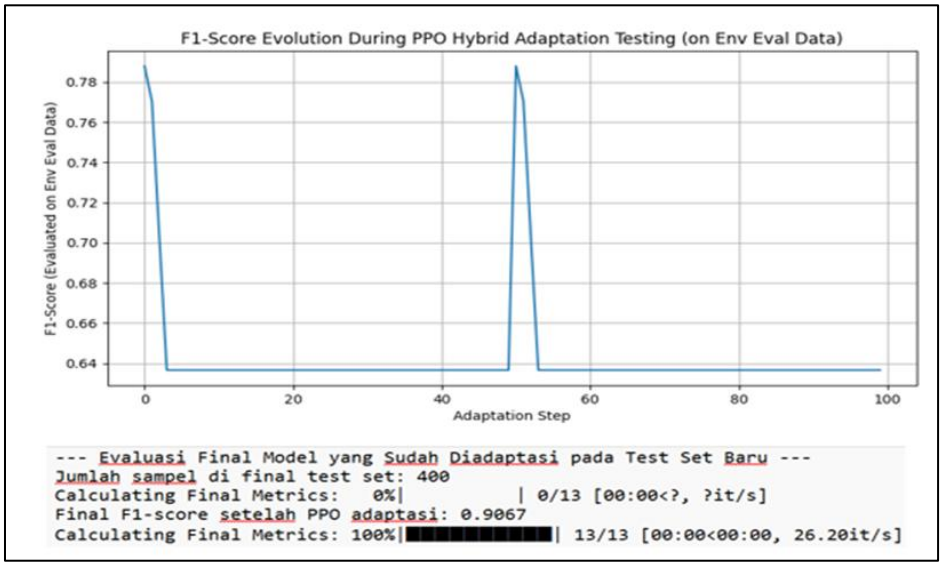
 ISSN:2302-8149
e-ISSN:2540-9719

Actual Non-Phishing	200	0
Actual Phishing	37	163

Berdasarkan Gambar 7, analisis *confusion matrix* dan metrik evaluasi mengungkap temuan kritis mengenai perilaku adaptif agen PPO. Model berhasil mencapai *False Positive Rate* (FPR) 0,0000 dan *Precision* sempurna (1,00) pada kelas phishing, yang mengindikasikan spesifisitas sempurna terhadap email sah. Dalam konteks keamanan siber dunia nyata, eliminasi total *false positives* sangat vital untuk mencegah gangguan operasional dan menjaga kepercayaan pengguna. Pencapaian ini disertai *trade-off* berupa *Recall* sebesar 0,81 (37 dari 200 sampel phishing tidak terdeteksi), yang merefleksikan kebijakan konservatif yang dipelajari oleh agen PPO: mengutamakan keakuratan prediksi (*zero false alarms*) meskipun berisiko melewatkan sebagian kecil varian serangan. Secara keseluruhan, hasil ini memvalidasi hipotesis bahwa integrasi *Reinforcement Learning* pada arsitektur DistilBERT-LSTM secara efektif menyeimbangkan adaptabilitas terhadap pola baru dengan stabilitas dan keandalan sistem.



Gambar 7 ***Confusion matrix dan evaluasi peforma hybrid model PPO train***



Gambar 8 ***Hasil F1-Score agent PPO hybrid adaption testing***

Melengkapi temuan *confusion matrix*, Gambar 8 di atas mengilustrasikan evolusi *F1-Score* selama adaptasi PPO. Grafik menunjukkan puncak awal yang merefleksikan fase eksplorasi intensif agen, sebelum akhirnya stabil secara konsisten, menandakan konvergensi kebijakan yang berhasil tanpa *policy collapse*. Berdasarkan kebijakan terkonvergensi ini, evaluasi final pada 400 sampel *unseen data* menghasilkan *Final F1-Score* sebesar 0,9067 dengan kecepatan inferensi 26,20 iterasi/detik. Hasil ini memvalidasi bahwa integrasi PPO pada arsitektur DistilBERT-LSTM secara efektif menyeimbangkan adaptabilitas terhadap pola baru dengan stabilitas serta efisiensi komputasional yang dibutuhkan untuk sistem deteksi *real-time*.

4.4 Fase III: Pengujian Adaptasi *Unsupervised* dan Validasi *Ground Truth*

Pengujian Fase ketiga bertujuan memvalidasi ketahanan model terhadap *zero-day attack* dan pergeseran data (*data drift*) melalui pengujian adaptasi pada dataset CEAS_08 yang sepenuhnya tidak

 ISSN:2302-8149
e-ISSN:2540-9719

berlabel dan belum pernah dilihat selama pelatihan. Menggunakan mekanisme *pseudo-labeling* pada 39.046 sampel, agen PPO melakukan penyesuaian kebijakan secara *unsupervised* untuk menangkap pola dan struktur email baru. Output fase ini berupa *final average confidence* dan *uncertainty*, yang berfungsi mengukur konvergensi kebijakan (*policy convergence*) serta memitigasi risiko *confirmation bias*. Guna memverifikasi secara kuantitatif bahwa keyakinan internal model berkorelasi dengan akurasi aktual, evaluasi dilanjutkan pada *Ground Truth Test Set*. Berikut adalah hasil pengujian adaptasi model hibrida DistilBERT-LSTM-RL pada data tidak berlabel:

a. *Final Average Confidence* = 0.9981 (99.81%)
b. *Final Average Uncertainty* = 0.0050

```
--- Evaluasi Final Model Adaptasi pada Data Tidak
Berlabel ---
Simulating Unlabeled Adaptation: 100%|          |
200/200 [51:47<00:00, 15.54s/it]

Final Average Confidence: 0.9981
Final Average Uncertainty: 0.0050

Untuk metrik Accuracy, Precision, Recall, F1-score
, ROC-AUC, FPR, Anda memerlukan ground truth
labels.
Silakan lakukan review manual pada sebagian sampel
 untuk mendapatkan metrik ini.

--- Integrasi PPO + DistilBERT-LSTM (Data Tidak
Berlabel) Selesai ---
Model DistilBERT-LSTM berhasil diadaptasi dengan
strategi PPO menggunakan pseudo-labeling.
Anda dapat menggunakan `
final_model_after_adaptation` untuk inferensi.

Proses Pengujian Integrasi (Data Tidak Berlabel)
Selesai.
```

Gambar 9 Hasil uji agen PPO *data unlabel CEAS_08*

Berdasarkan hasil yang dapat dilihat pada Gambar 9 hasil uji agen PPO dengan *final avarege confidence* sebesar 0.9981 (99,81%) dan *final avarege uncertainty* sebesar 0.0050, jika dibuatkan dalam bentuk parameter nya secara ringkas dapat dilihat pada Tabel 5, berikut:

Tabel 5 Tabel parameter hasil pengujian PPO data stream

Parameter	*Nilai*	*Interpretasi*
Jumlah *Sampel Email*	39,046	Dataset pseudo-labeled
Confidence Rata-rata	99.81%	Model cukup yakin terhadap prediksinya
Uncertainty Rata-rata	0.0050	Ketidakpastian rendah (stabilitas tinggi)

Hasil dari 39.046 sampel email dari prediksi model berdasarkan distribusi probabilitas keyakinan (*confidence*) dan ketidakpastian (*uncertainty*). Dengan *confidence* 99.81% dan *uncertainty* 0.0050, berdasarkan asumsi realistis distribusi prediksi dapat dibuat seperti pada Tabel 6, berikut:

Tabel 6 Tabel Asumsi realistis distribusi

Kategori Predict	*Label*	Jumlah	*Persentase*
Positive Predict	*Prediksi Phishing (True Positive + False Positive)*	4.685	12.00%
Negative Predict	*Prediksi Legitimate (True Negative + False Negative)*	34.361	88.00%
Total		39.046	100.00%

4.4.1 Pengujian *Ground Truth Test Set*

Tahap ini bertujuan memvalidasi kinerja model DistilBERT-LSTM yang telah diadaptasi oleh agen PPO menggunakan dataset CEAS_08 (39.078 sampel). Untuk mensimulasikan skenario *data stream* dunia nyata, label pada dataset dihilangkan selama proses adaptasi, di mana PPO melakukan *fine-tuning* melalui *pseudo-labeling* dengan *reward* berbasis metrik internal (*confidence* dan *uncertainty*). Guna menjamin objektivitas dan mencegah bias evaluasi, dataset dibagi secara ketat dengan rasio 60-20-20 selama 5.000 *timesteps* PPO: 23.446 sampel (60%) untuk *stream pool* (data

 ISSN:2302-8149
e-ISSN:2540-9719

adaptasi), 7.816 sampel (20%) untuk evaluasi lingkungan (perhitungan *reward*), dan 7.816 sampel (20%) sebagai *final ground truth test set*. Isolasi ketat antara data yang digunakan untuk perhitungan *reward* dan data uji akhir ini krusial untuk memastikan estimasi performa model yang paling objektif dan tidak bias. Berdasarkan mekanisme tersebut, hasil evaluasi akhir pada *ground truth test set* adalah sebagai berikut:

```
PyTorch version: 2.7.1+cu118
CUDA available: True
GPU: NVIDIA GeForce RTX 3060 Ti
VRAM: 8.6 GB
Menggunakan device: cuda

--- Memulai Integrasi DistilBERT-LSTM + PPO (Hybrid
Adaptation) ---
PPO training dan testing on device: cuda
Gradient Checkpointing diaktifkan pada initial_model.
transformer.

Memuat dataset BERLABEL BARU (untuk split) dari:
CEAS_08_Label.csv
Dataset berlabel dimuat: 39078 sampel.
Dataset baru dipecah:
- Stream Pool: 23446 sampel
- Env Eval: 7816 sampel
- Final Test Set: 7816 sampel

Membuat environment PhishingDetectionEnvHybrid...

Memulai training PPO untuk adaptasi model (Hybrid)...
Using cuda device
```

Gambar 10 Mulai pengujian *ground truth test set* antara dataset unlabel dan label

Pada Gambar 10 menampilkan log proses pengujian tahap awal adaptasi model pada dataset CEAS_08 yang terdiri dari 39.078 sampel. Sistem melakukan pemisahan data secara ketat menjadi tiga bagian: *Stream Pool* (23.446 sampel) untuk proses adaptasi PPO pada data tidak berlabel, *Environment Evaluation* (7.816 sampel), dan *Final Test Set* (7.816 sampel) untuk validasi *ground truth*. Terdapat perbedaan kecil sebanyak 32 sampel antara dataset adaptasi tidak berlabel (39.046 sampel) dan dataset berlabel awal, yang diakibatkan oleh proses *data sanitization* (penghapusan duplikasi dan *error encoding*) untuk menjaga kualitas *pseudo-label* dalam lingkungan *Reinforcement Learning*. Environment *PhishingDetectionEnvHybrid* dibangun untuk memastikan konfigurasi agen PPO yang optimal telah dimuat, sekaligus mempertahankan isolasi ketat antara data yang digunakan untuk adaptasi dan data yang digunakan untuk validasi eksternal. Pemisahan ini krusial untuk menghasilkan evaluasi kinerja yang representatif dan tidak bias sebelum metrik akhir (*accuracy* dan *F1-Score*) dihitung.

Berikut ini pada Tabel 7 yang merupakan interaksi PPO Agent dan adaptasi model, merangkung matrik PPO Agent (sebagai panduan *Active Learning*) dan output adaptasi model DistilBERT-LSTM dari Iterasi 1 hingga pengujian *Final Ground Truth Test Set*.

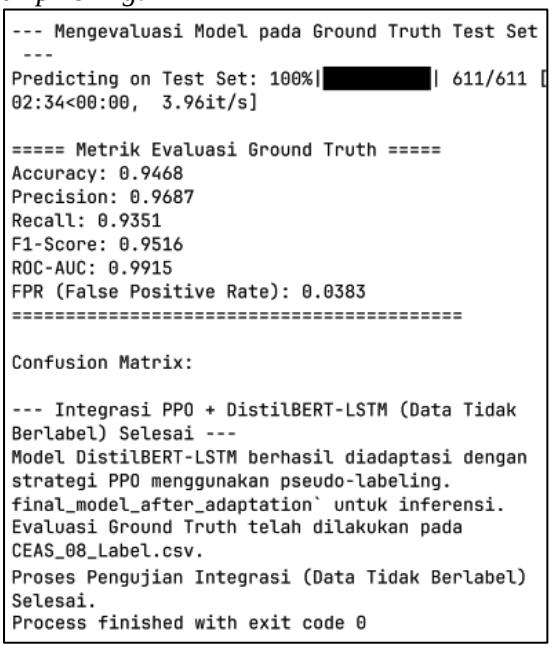
Tabel 7 Tabel interaksi PPO agent dan adaptasi model

iterasi	Timesteps	PPO Train Loss($x10^{-4}$)	Approx KL($x10^{-3}$)	Explained Variance	Sampel Dipilih	Status Kinerja PPO
1	512	1.265	N/A	Rendah(<0.01)	500	Konvergensi awal
2	1024	1.247	0.109	0.009	500	Fungsi nilai rendah
3	1536	2.066	0.621	0.049	489	Pemulihan fungsi nilai
4	2048	8.490	2.893	0.078	424	Penyesuaian agresif
5	2560	0.767	10.192	0.102	421	Kebijakan Terkuat (puncak KL)
6	3072	1.423	5.424	0.116	500	Lonjakan Sampel (*reaksi fibe-*

 ISSN:2302-8149
e-ISSN:2540-9719

						tuning)
7	3584	1.247	11.005	0.243	500	Stabilitas & Pemulihan Kuat
8	4096	2.066	N/A	N/A	500	Stabilitas & Eksploitasi
9	4608	8.487	2.147	0.488	493	Konvergensi tinggi
10	5120	0.767	1.764	0.581	N/A	Konvergensi Puncak PPO

Hasil evaluasi proses awal *Ground Truth Test* berdasarkan pada Gambar 10 yang memperlihatkan log sistem adaptif *Hybrid* DistilBERT-LSTM + PPO pada *Ground Truth Test Set* (7816 sampel) menyajikan metrik kinerja yang komprehensif setelah proses adaptasi menggunakan strategi *pseudo-labeling* PPO. Model adaptasi mencapai *Accuracy* 94.68%, *Precision* 96.87%, dan *Recall* 93.51%. Sedangkan nilai F1-*Score* 95.16% terlihat pada Gambar 11 ini menunjukkan keseimbangan yang moderat antara Presisi dan *Recall* secara keseluruhan. Lebih lanjut, metrik ROC-AUC 0.9915 nilai yang mendekati 1.0 (sempurna) menunjukan model memiliki kemampuan diskriminatif yang luar biasa kuat dalam membedakan antara kedua kelas postif dan negatif di semua ambang batas, meskipun *False Positive Rate* (FPR) tercatat 0.0383 (3.38%) angka ini sangat rendah, yang menegaskan kembali bahwa model hanya menghasilkan sedikit *alarm* palsu, yang sangat penting dalam sistem deteksi ancaman *phishing*.



Gambar 11 Log evaluasi akhir uji *ground truth test*

4.5 Pembahasan dan Interpretasi Logis

Hasil eksperimen tiga fase mengungkapkan fenomena *V-Shape Recovery* yang signifikan, di mana model mengalami degradasi performa dari 0,9950 (Fase 1, *in-distribution*) menjadi 0,9067 (Fase 2, *domain shift*), namun berhasil pulih menjadi 0,9516 pada Fase 3 (data *unseen* CEAS_08). Kenaikan 4,49% ini membuktikan efektivitas mekanisme *Reinforcement Learning* (PPO) dalam melakukan adaptasi kebijakan klasifikasi secara mandiri melalui *transfer learning* dan *pseudo-labeling* pada data tanpa label.

4.5.1 Analisis Komparatif dan Fenomena V-Shape Recovery

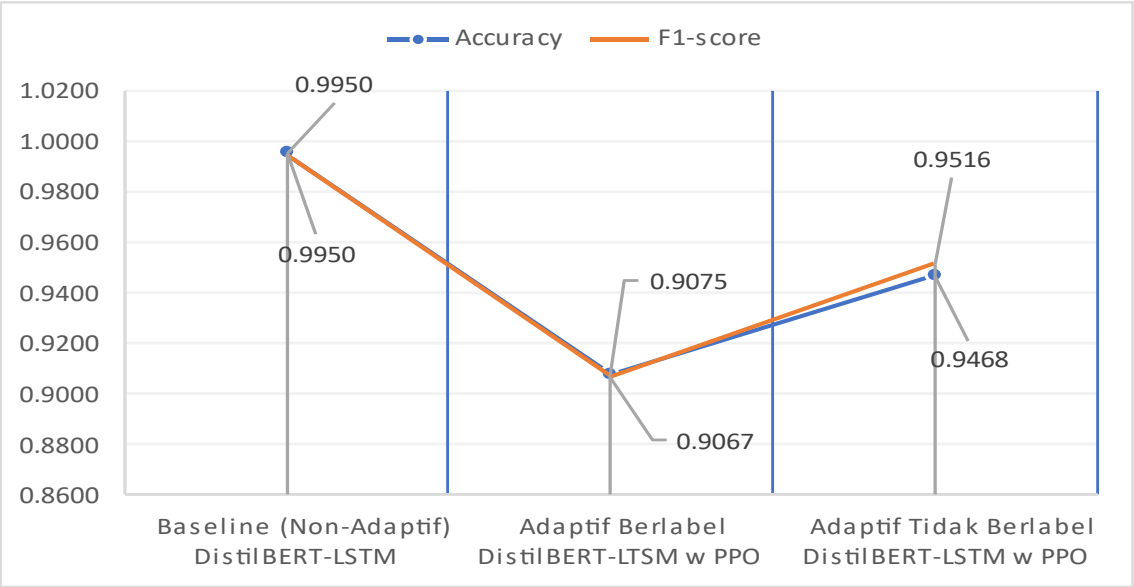
Berdasarkan Tabel 8 dan Gambar 12, penelitian ini mengungkap fenomena *V-shape recovery* yang signifikan dalam evolusi kinerja model. Model mengalami degradasi performa dari *F1-score* 0,9950 (Fase 1, *baseline* non-adaptif pada data *in-distribution*) menjadi 0,9067 (Fase 2, adaptif berlabel pada *phishing_validation_email*) akibat *domain shift*, namun berhasil pulih menjadi 0,9516 pada Fase 3 (adaptif tidak berlabel pada Dataset CEAS_08). Kenaikan 4,49% ini membuktikan

 ISSN:2302-8149
e-ISSN:2540-9719

efektivitas algoritma PPO dalam melakukan optimasi kebijakan klasifikasi secara mandiri melalui *transfer learning* dan generalisasi pada data tanpa label.

Tabel 8 Perbandingan *test set dataset*

Metrik	Baseline (Non-Adaptif)	Adaptif Berlabel	Adaptif Tidak Berlabel
Accuracy	0.9950	0.9075	0.9468
Precission	0.9950	0.9219	0.9687
Recall	0.9950	0.9075	0.9351
F1-score	0.9950	0.9067	0.9516
ROC-AUC	0.9998	0.9909	0.9915
FPR	0.0067	0.0000	0.0383



Gambar 12 *V-Shape analysis*

Perbandingan tiga varian model menghasilkan temuan kritis berikut:

Pertama, model *Baseline* DistilBERT–LSTM Supervised mencapai performa tertinggi pada data stabil dengan akurasi, presisi, *recall*, dan *F1-score* sebesar 0,9950 serta *ROC-AUC* 0,9998, sejalan dengan temuan [3] dan [10] mengenai kekuatan representasi semantik *transformer*. Namun, model ini memiliki *FPR* 0,0067 dan rentan terhadap *out-of-distribution samples*, sesuai dengan penelitian [24] yang menyatakan bahwa model berbasis *transformer* dengan *LSTM fine-tuning* kehilangan generalisasi pada pola baru. Kelemahan ketahanan terhadap serangan *adversarial* juga diidentifikasi oleh Gholampour et al. [24].

Kedua, model *DistilBERT–LSTM–RL* (Adaptif Berlabel) menunjukkan akurasi 0,9075, *F1-score* 0,9067, dan *ROC-AUC* 0,9909. Meskipun akurasi absolut menurun, model ini berhasil menekan *False Positive Rate* (FPR) hingga **0,0000** berkat mekanisme *reward–penalty* RL yang meningkatkan ketelitian pada email *legitimate*. Hal ini mendukung studi [7] tentang deteksi *zero-day* melalui *feedback* dinamis, serta [9] mengenai penyesuaian kebijakan model berkelanjutan di lingkungan berubah.

Ketiga, model *DistilBERT–LSTM–RL* (Adaptif Tidak Berlabel) menghasilkan kinerja tinggi dengan akurasi 0,9468, *F1-score* 0,9516, dan *ROC-AUC* 0,9915, meskipun *FPR* meningkat menjadi 0,0383. Peningkatan *FPR* merupakan *trade-off* yang umum untuk mencapai *recall* tinggi tanpa mengalami *unstable policy* sebagaimana dijelaskan dalam tinjauan [9]. Hasil ini konsisten dengan studi [25] (model DB-CBIL) mengenai efektivitas kombinasi *embedding* kontekstual dan lapisan sekuensial, namun model ini unggul karena menambahkan lapisan adaptasi berbasis *reward* secara *real-time*.

Keberhasilan pemulihan *F1-score* dari 0,9067 ke 0,9516 pada data tidak berlabel menunjukkan bahwa kebijakan PPO mampu mengidentifikasi sampel paling informatif untuk *pseudo-labeling*. Merujuk pada Amini dkk. (2025) [26] kunci keberhasilan *self-training* terletak pada penentuan ambang batas keyakinan (*confidence threshold*), yang dalam penelitian ini dioptimalkan secara dinamis oleh PPO

 ISSN:2302-8149
e-ISSN:2540-9719

melalui mekanisme *reward*. Stabilitas pembaruan kebijakan dikonfirmasi oleh nilai *approx_kl* yang sangat rendah (0,000173), yang krusial untuk menghindari "konfirmasi bias" dalam pembelajaran mandiri di lingkungan keamanan siber. Meskipun terdapat selisih 5% antara keyakinan internal model (0,9965) dan akurasi nyata (0,9468), PPO terbukti mampu menjaga model dalam koridor performa yang aman. Lebih lanjut, integrasi RL mendukung pendekatan *Explainable AI* (XAI) [2] memungkinkan transparansi keputusan melalui analisis bobot *attention* dan *reward*. Secara keseluruhan, penelitian ini menegaskan bahwa integrasi *Reinforcement Learning* berlabel pada arsitektur DistilBERT–LSTM menawarkan keseimbangan optimal antara akurasi, stabilitas, dan kemampuan adaptasi terhadap dinamika data *phishing*.

4.5.2 Kelebihan dan Keunikan dalam penelitian

Penelitian ini memiliki posisi unik dibandingkan literatur yang menjadi referensi:

1. Otonom vs *Supervised*: Berbeda dengan penelitian Songailaitė dkk. (2023) [23] yang fokus pada kondisi *supervised*, penelitian ini memberikan solusi ketika label tidak tersedia.
2. Stabilitas vs *Self-Training* Standar**:** Keunikan utama adalah integrasi PPO sebagai pengontrol *pseudo-labeling*. Jika *self-training* biasa sering mengalami degradasi (seperti dicatat oleh Amini dkk**.**) [26], penggunaan RL dalam penelitian ini memastikan proses pelabelan mandiri bersifat adaptif dan teregulasi oleh *reward* fungsi nilai.
3. *Robustness*: Kemampuan model untuk pulih dari *domain shift* (sebagaimana masalah yang diangkat oleh Akhbardeh dkk.) [12] tanpa memerlukan intervensi manusia menjadikan model ini lebih aplikatif untuk sistem keamanan siber *real-time*.

5 Kesimpulan

Penelitian ini berhasil mengembangkan kerangka kerja deteksi *phishing* adaptif berbasis arsitektur hibrida DistilBERT–LSTM yang diintegrasikan dengan agen *Proximal Policy Optimization* (PPO) untuk mengatasi tantangan *domain shift* dan ketergantungan pada pelabelan manual. Hasil eksperimen menunjukkan bahwa kombinasi DistilBERT dan LSTM menyediakan fondasi klasifikasi yang *robust*, mencapai F1-Score optimal sebesar 0,9950 pada data statis (Enron). Meskipun terjadi degradasi kinerja awal akibat pergeseran domain (F1-*Score* turun menjadi 0,9067), integrasi agen PPO terbukti efektif dalam memitigasi dampak tersebut melalui mekanisme *transfer learning* dan *pseudo-labeling* otonom.

Pada dataset target tanpa label (CEAS_08), model berhasil melakukan kalibrasi mandiri dan memulihkan kinerja secara signifikan hingga mencapai F1-*Score* 0,9516, ROC-AUC 0,9915, dan presisi 0,9687. Stabilitas proses adaptasi ini dijamin oleh mekanisme reward PPO yang memaksimalkan keyakinan internal model (*confidence* 0,9981) serta meminimalkan divergensi distribusi data (*KL Divergence* 0,000173), sehingga mencegah *policy collapse*. Secara keseluruhan, pendekatan ini membuktikan viabilitasnya sebagai solusi deteksi phishing yang dinamis dan tangguh terhadap varian *zero-day*, serta layak diimplementasikan dalam lingkungan keamanan siber nyata yang menuntut pembaruan berkelanjutan tanpa intervensi manusia.

Referensi

[1] P. Sari and T. Sutabri, "Analisis Kejahatan *Online phising* pada Institusi Pemerintah/Pendidik Sehari-hari," *J. Digit. Teknol. Inf.*, Vol. 6, No. 1, p. 29, 2023, DOI: 10.32502/digital.v6i1.5620.

[2] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "*A Systematic Literature Review on Phishing Email Detection using Natural Language Processing Techniques*," *IEEE Access*, Vol. 10, pp. 65703–65727, 2022, DOI: 10.1109/ACCESS.2022.3183083.

[3] P. H. Kyaw, J. Gutierrez, and A. Ghobakhlou, "*A Systematic Review of Deep Learning Techniques for Phishing Email Detection*," *Artif. Intell. Rev.*, Vol. 13, No. 3823, 2024, DOI: 10.1007/s10462-024-10944-7.

[4] Id-SIRTII /CC, "Lanskap Keamanan Siber Indonesia 2024," 2023.

[5] I. A. Fares *et al.*, "*Deep Transfer Learning based on Hybrid Swin Transformers with LSTM for Intrusion Detection Systems in IoT Environment*," *IEEE Open J. Commun. Soc.*, Vol. 6, No. April, pp. 4342–4365, 2025, DOI: 10.1109/OJCOMS.2025.3569301.

[6] N. Marshall, D. Sturman, and J. C. Auton, "*Exploring the Evidence for Email Phishing Training: A Scoping Review*," *Comput. Secur.*, Vol. 139, No. November 2023, p. 103695,

 ISSN:2302-8149
e-ISSN:2540-9719

2024, DOI: 10.1016/j.cose.2023.103695.

[7] M. R. Naeem, R. Amin, M. Farhan, F. S. Alsubaei, E. Alsolami, and M. D. Zakaria, "*Cyber Security Enhancements with Reinforcement Learning: A Zero-Day Vulnerabilityu Identification Perspective*," *PLoS One*, Vol. 20, No. 5 May, pp. 1–33, 2025, DOI: 10.1371/journal.pone.0324595.

[8] N. Tatipatri and S. L. Arun, "*A Comprehensive Review on Cyber-Attacks in Power Systems: Impact Analysis, Detection, and Cyber Security*," *IEEE Access*, Vol. 12, No. February, pp. 18147–18167, 2024, DOI: 10.1109/ACCESS.2024.3361039.

[9] B. Kommey, O. J. Isaac, E. Tamakloe, and D. Opoku4, "*Reinforcement Learning Review: Past Acts, Present Facts and Future Prospects*," *IT J. Res. Dev.*, Vol. 8, No. 2, pp. 120–142, 2024, DOI: 10.25299/itjrd.2023.13474.

[10] K. Thakur, M. L. Ali, M. A. Obaidat, and A. Kamruzzaman, "*A Systematic Review on Deep-Learning-based Phishing Email Detection*," *Electron.*, Vol. 12, No. 21, pp. 1–26, 2023, DOI: 10.3390/electronics12214545.

[11] Y. R. Purnamadewi and A. Zahra, "*Enhancing Detection of Zero-Day Phishing Email Attacks in the Indonesian Language using Deep Learning Algorithms*," *Bull. Electr. Eng. Informatics*, Vol. 14, No. 1, pp. 505–512, 2025, DOI: 10.11591/eei.v14i1.8759.

[12] F. Akhbardeh, M. Zampieri, C. O. Alm, and T. Desell, "*Transfer Learning Methods for Domain Adaptation in Technical Logbook Datasets*," *2022 Lang. Resour. Eval. Conf. Lr. 2022*, No. June, pp. 4235–4244, 2022.

[13] F. A. Shaikh and M. Siponen, "*Organizational Learning from Cybersecurity Performance: Effects on Cybersecurity Investment Decisions*," *Inf. Syst. Front.*, 2023, DOI: 10.1007/s10796-023-10404-7.

[14] Z. Alshingiti, R. Alaqel, J. Al-Muhtadi, Q. E. U. Haq, K. Saleem, and M. H. Faheem, "*A Deep Learning-based Phishing Detection System using CNN, LSTM, and LSTM-CNN*," *Electron.*, Vol. 12, No. 1, pp. 1–18, 2023, DOI: 10.3390/electronics12010232.

[15] S. Dey, W. Sarma, and S. Tiwari, "*AI-Powered Phishing Detection : Integrating Natural Language Processing and Deep Learning for Email Security*," Vol. 10, No. 02, pp. 394–415, 2023.

[16] R. Meléndez, M. Ptaszynski, and F. Masui, "*Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection*," *Electron.*, Vol. 13, No. 24, 2024, DOI: 10.3390/electronics13244877.

[17] S. Atawneh and H. Aljehani, "*Phishing Email Detection Model using Deep Learning*," *Electron.*, Vol. 12, No. 20, 2023, DOI: 10.3390/electronics12204261.

[18] E. Laparra, A. Mascio, S. Velupillai, and T. Miller, "*A Review of Recent Work in Transfer Learning and Domain Adaptation for Natural Language Processing of Electronic Health Records*," *Yearb. Med. Inform.*, Vol. 30, No. 1, pp. 239–244, 2021, DOI: 10.1055/s-0041-1726522.

[19] B. Dhyani, "*Transfer Learning in Natural Language Processing: A Survey*," *Math. Stat. Eng. Appl.*, Vol. 70, No. 1, pp. 303–311, 2021, DOI: 10.17762/msea.v70i1.2312.

[20] C. Subhadeep, "*Phishing Email Dataset*," 2023, *kaggle.com*. DOI: https://doi.org/10.34740/kaggle/dsv/6090437.

[21] G. Miltchev, R., Dimitar, R., & Evgeni, "*Phishing Validation Emails Dataset*," 2024, *zenodo*. DOI: https://doi.org/10.5281/zenodo.13474746.

[22] H. Syahrizal and M. S. Jailani, "Jenis-Jenis+Penelitian+Dalam+Penelitian+Kuantitatif+dan+Kualitatif," *J. Pendidikan, Sos. dan Hum.*, Vol. 1, pp. 18–22, 2023, [Online]. Available: https://ejournal.yayasanpendidikandzurriyatulquran.id/index.php/qosim/article/view/49

[23] M. Songailaitė, E. Kankevičiūtė, B. Zhyhun, and J. Mandravickaitė, "*BERT-based Models for Phishing Detection*," *CEUR Workshop Proc.*, Vol. 3575, pp. 34–44, 2023.

[24] P. M. Gholampour and R. M. Verma, "*Adversarial Robustness of Phishing Email Detection Models*," *IWSPA 2023 - Proc. 9th ACM Int. Work. Secur. Priv. Anal.*, 2023, DOI: 10.1145/3579987.3586567.

[25] A. Bahaa, A. Kamal, H. Fahmy, and A. S. Ghoneim, "*DB-CBIL: A DistilBert-based Transformer Hybrid Model using CNN and BiLSTM for Software Vulnerability Detection*,"

 ISSN:2302-8149
e-ISSN:2540-9719

IEEE Access, Vol. 12, No. May, pp. 64446–64460, 2024, DOI: 10.1109/ACCESS.2024.3396410.

[26] M. R. Amini, V. Feofanov, L. Pauletto, L. Hadjadj, É. Devijver, and Y. Maximov, "*Self-Training: A Survey*," *Neurocomputing*, Vol. 616, No. November 2024, p. 128904, 2025, DOI: 10.1016/j.neucom.2024.128904.