Dokumentasi Proyek Klasifikasi Lalu Lintas Jaringan: Dari PESV ke Model Hibrida HFV

Mahasiswa Sarjana Teknik Komputer
(Silakan ganti dengan nama Anda)

8 November 2025

Ringkasan

Klasifikasi lalu lintas jaringan terenkripsi merupakan tantangan penting untuk manajemen jaringan dan keamanan siber. Penelitian ini mengusulkan dan memvalidasi sebuah model baru berbasis aliran (flow), **Hybrid Flow Vector (HFV)**, yang dirancang untuk mengklasifikasikan lalu lintas terenkripsi (VPN) dan non-enkripsi (Non-VPN) dari dataset ISCX 2016. Model kami menunjukkan keunggulan pendekatan hibrida yang menggabungkan deep learning dengan rekayasa fitur statistik yang tangguh.

HFV adalah vektor multi-modal yang terdiri dari tiga komponen: (α'') satu set fitur 128-dimensi yang diekstraksi dari 1D-Convolutional Neural Network (1D-CNN) yang dilatih pada payload paket mentah; (δ) satu set fitur 39-dimensi berisi statistik level aliran (flow) yang komprehensif; dan (γ') satu set fitur 37-dimensi yang merinci statistik level burst.

Studi ablasi ekstensif mengkonfirmasi bahwa fitur deep learning dan fitur statistik bersifat sangat komplementer (saling melengkapi). Model hibrida penuh ($\alpha'' + \delta + \gamma'$) mencapai akurasi tertinggi untuk klasifikasi kategori 6-kelas (81.40%) dan aplikasi 6-kelas (76.58%), secara signifikan mengungguli model deep-learning-saja (77.48%) maupun model statistik-saja (75.43%). Perbandingan klasifikasi akhir menunjukkan bahwa model XGBoost, yang dilatih pada HFV lengkap, memberikan performa terbaik, mencapai akurasi puncak 97.12% untuk klasifikasi biner (VPN/Non-VPN) dan 81.51% untuk klasifikasi kategori. Penelitian ini membuktikan bahwa model hibrida, yang mengintegrasikan pola level-byte dari 1D-CNN dengan analisis statistik level-flow, memberikan solusi yang sangat tangguh dan akurat untuk klasifikasi lalu lintas yang kompleks.

Daftar Isi

1 Pendahuluan

Penelitian ini dimulai dari sebuah model awal yang disebut *Path-Embedding Signature Vector* (PESV), seperti yang didokumentasikan dalam `dokumentasi.pdf` [1]. Model PESV awal ini adalah vektor multi-modal $\Sigma = (\alpha, \beta, \gamma)$ yang dirancang untuk klasifikasi lalu lintas VPN pada dataset ISCX 2016.

- **Komponen α:** Sebuah vektor 32-dimensi dari LSTM Autoencoder yang dilatih pada *ukuran* 128 paket pertama.
- **Komponen β:** Jarak Wasserstein dari histogram *Inter-Arrival Time* (IAT) sebuah flow ke prototipe kategori.
- **Komponen γ:** Cosine Similarity dari 4 statistik *burst* sebuah flow ke prototipe kategori.

Hasil awal dari `PESV.ipynb` [2] menunjukkan performa yang baik untuk klasifikasi biner (VPN/Non-VPN) dengan akurasi $\sim$86%, namun performanya menurun drastis untuk tugas yang lebih kompleks: $\sim$70% untuk klasifikasi 6-kelas `category` dan $\sim$62% untuk klasifikasi `application`.

Analisis ini mengidentifikasi dua kelemahan utama: (1) Dataset yang sangat tidak seimbang dan ambigu, dan (2) Kelemahan desain pada fitur β dan γ yang bergantung pada perbandingan dengan ”prototipe” yang kabur. Tujuan dari penelitian lanjutan ini adalah untuk mengatasi kedua masalah tersebut secara iteratif untuk mengembangkan model klasifikasi yang jauh lebih akurat dan tangguh.

2 Metodologi Iteratif dan Pengembangan Fitur

Untuk meningkatkan performa, kami melakukan serangkaian eksperimen yang terstruktur, dimulai dengan perbaikan dataset dan diakhiri dengan pengembangan fitur hibrida yang canggih.

2.1 Langkah 1: Kurasi Dataset (Dataset v2)

Analisis awal pada dataset penuh ISCX 2016 (13.450 flow) mengungkapkan adanya ketidakseimbangan kelas yang ekstrem dan ambiguitas label (misalnya, aplikasi ’Skype’ termasuk dalam 3 kategori berbeda). Untuk menciptakan dasar eksperimen yang stabil dan adil, kami melakukan kurasi dataset.

Kami memilih 6 aplikasi dengan jumlah sampel terbesar yang juga mencakup semua 6 kategori yang ada. Dataset yang difilter ini, yang disebut **Dataset v2**, berisi **10.284 flow** dan digunakan sebagai dataset dasar untuk semua ekstraksi fitur dan eksperimen selanjutnya.

2.2 Langkah 2: Model v2 ”Robust” ($\alpha + \delta + \gamma'$)

Kami berhipotesis bahwa fitur β dan γ yang asli lemah karena bergantung pada perbandingan prototipe. Kami menggantinya dengan fitur statistik yang jauh lebih deskriptif:

- **Komponen δ (Delta):** Sebuah vektor $\sim$39-fitur yang berisi profil statistik lengkap dari *seluruh* flow. Ini mencakup mean, std, median, min, max, dll., dari ukuran paket dan IAT, yang dihitung secara dua arah (client-to-server dan server-to-client).
- **Komponen γ' (Gamma-Prime):** Sebuah vektor 37-fitur yang berisi profil statistik lengkap dari *semua burst* dalam sebuah flow. Ini mencakup jumlah burst, dan statistik penuh (mean, std, median, max) dari jumlah paket per-burst, volume per-burst, durasi per-burst, dan waktu jeda antar-burst.

Kami kemudian menjalankan studi ablasi pada model v2 ini ($\alpha + \delta + \gamma'$). Hasilnya (dirangkum dalam Tabel 3) sangat mengejutkan: model statistik-saja ($\delta + \gamma'$) mengungguli model penuh. Ini membuktikan bahwa fitur α (LSTM pada ukuran paket) yang asli sebenarnya bertindak sebagai **noise** dan menurunkan akurasi.

2.3 Langkah 3: Model v3 "Hybrid" ($\alpha'' + \delta + \gamma'$) - HFV

Temuan dari v2 (α adalah noise) dan wawasan dari literatur [3, 4] memicu hipotesis baru: kelemahan α bukan pada deep learning, tetapi pada *input*-nya (ukuran paket).

Literatur [3, 5] menunjukkan bahwa **1D-Convolutional Neural Network (1D-CNN)** yang dilatih pada **raw packet payload** (byte mentah) sangat efektif untuk klasifikasi lalu lintas terenkripsi, bahkan pada dataset ISCX 2016 yang sama.

Kami kemudian mengembangkan model "ultimate" kami, **Hybrid Flow Vector (HFV)**, dengan mengganti α yang gagal dengan fitur deep learning baru yang canggih:

- **Komponen α'' (Alpha-double-prime):** Sebuah vektor 128-dimensi yang diekstraksi dari 1D-CNN.

Proses ekstraksi fitur α'' ini adalah sebagai berikut:

1. **Ekstraksi Payload:** Data payload mentah diekstraksi dari 10 paket pertama untuk setiap flow, dengan setiap paket dipotong/di-padding menjadi 784 byte. Ini menghasilkan dataset NumPy berdimensi (9542, 10, 784).
2. **Pelatihan Encoder:** Sebuah 1D-CNN (terdiri dari lapisan `Conv1D`, `MaxPooling1D`, dan `Dense`) dilatih untuk mengklasifikasikan 6 `application`. Proses pelatihan ini "mengajarkan" filter CNN untuk menemukan pola byte mentah yang spesifik untuk setiap aplikasi.
3. **Generasi Fitur:** Model CNN yang telah dilatih kemudian "dipotong" (dihilangkan lapisan klasifikasi akhirnya). Semua 9.542 sampel flow dilewatkan melalui encoder ini untuk menghasilkan vektor fitur α'' 128-dimensi yang baru.

Vektor HFV final ini kemudian dirakit dengan menggabungkan $\alpha'' + \delta + \gamma'$, menghasilkan dataset `final_PESV_dataset_v3.csv` dengan total 204 fitur.

3 Hasil Eksperimen dan Analisis

Kami melakukan dua eksperimen final pada dataset v3 untuk memvalidasi model HFV. Semua eksperimen menggunakan `RandomForestClassifier` (dengan `class_weight='balanced'`) dan `StandardScaler` dalam sebuah `Pipeline` Scikit-learn.

3.1 "Championship" Ablation Study (Dataset v3 Penuh)

Eksperimen ini dirancang untuk menjawab pertanyaan terpenting: apakah model hibrida ($\alpha'' + \delta + \gamma'$) lebih baik daripada model deep-learning-saja (α'') atau model statistik-saja ($\delta + \gamma'$)?

Analisis: Hasil pada Tabel 1 sangat jelas.

1. Fitur α'' (1D-CNN) yang baru (77.48%) jauh lebih unggul daripada fitur statistik-saja ($\delta + \gamma'$, 75.43%) untuk klasifikasi kategori.
2. Model Hibrida Penuh (81.40%) secara signifikan mengungguli keduanya.
3. Ini **membuktikan secara konklusif** bahwa fitur deep learning (α'') dan fitur statistik (δ, γ') bersifat komplementer. Keduanya menemukan pola yang berbeda (pola level-byte vs. pola level-flow), dan ketika digabungkan, mereka menciptakan model yang paling akurat.

Tabel 1: Hasil Studi Ablasi v3 (HFV) pada Dataset Penuh

Kombinasi Fitur	Binary (Akurasi)	Category (Akurasi)	Application (Akurasi)
Deep Learning Saja:			
Alpha" (α'') only	93.45%	77.48%	74.75%
Statistik Saja:			
Delta (δ) only	96.23%	75.85%	68.94%
Gamma' (γ') only	94.97%	73.55%	66.79%
Delta (δ) + Gamma' (γ')	96.44%	75.43%	68.57%
Model Hibrida:			
Alpha" (α'') + Delta (δ)	96.44%	81.09%	76.38%
Alpha" (α'') + Gamma' (γ')	96.65%	81.35%	76.32%
Full (HFV: $\alpha'' + \delta + \gamma'$)	**97.01%**	**81.40%**	**76.58%**

3.2 Perbandingan Model Klasifikasi Akhir

Setelah membuktikan bahwa set fitur `Full (HFV)` adalah yang terbaik, kami mengujinya dengan empat model klasifikasi yang kuat untuk menemukan performa puncak.

Tabel 2: Perbandingan Klasifikasi pada Fitur HFV Penuh ($\alpha'' + \delta + \gamma'$)

Model Klasifikasi	Binary (Akurasi)	Category (Akurasi)	Application (Akurasi)
XGBoost	**97.12%**	**81.51%**	76.32%
RandomForest	97.01%	81.40%	**76.58%**
SVM (RBF Kernel)	93.82%	75.33%	71.45%

Analisis: Hasil pada Tabel 2 menunjukkan bahwa model berbasis *tree* (XGBoost dan RandomForest) adalah pemenang yang jelas, dengan mudah mengungguli SVM. Performa antara XGBoost dan RandomForest hampir identik secara statistik, membuktikan bahwa kami telah mencapai batas atas performa dari set fitur kami. XGBoost memiliki sedikit keunggulan pada tugas 'category', mencapai akurasi puncak **81.51%**.

4 Kesimpulan

Perjalanan penelitian ini berhasil mengubah model klasifikasi awal dengan performa sedang ($\sim$70% pada `category`) menjadi sebuah *framework* hibrida yang canggih dengan akurasi **81.51%**.

Kesimpulan utamanya adalah:

1. **Fitur Statistik itu Kuat:** Rekayasa fitur statistik yang tangguh (δ dan γ') terbukti menjadi dasar yang sangat kuat, mengalahkan model deep learning (LSTM) awal.
2. **Payload Mentah adalah Kunci:** Terinspirasi oleh literatur SOTA [3, 5], kami membuktikan bahwa menganalisis *raw byte payload* dengan 1D-CNN (α'') jauh lebih unggul daripada menganalisis *ukuran paket* dengan LSTM (α).
3. **Hibrida adalah yang Terbaik:** Temuan terpenting adalah bahwa model hibrida **HFV ($\alpha'' + \delta + \gamma'$)** adalah yang terbaik. Ini membuktikan bahwa model *deep learning* (yang menemukan pola level-byte) dan model *statistik* (yang menemukan pola level-flow dan level-burst) tidaklah berlebihan, melainkan saling melengkapi.

Model HFV yang diusulkan, yang dilatih dengan classifier XGBoost, terbukti menjadi solusi yang sangat akurat dan tangguh untuk klasifikasi lalu lintas jaringan terenkripsi dan non-enkripsi yang kompleks.

Pustaka

[1] (Nama Anda). (2025). *Summary of Path-Embedding Signature Vector (PESV) Generation*. Dokumen internal proyek.

[2] (Nama Anda). (2025). *PESV.ipynb*. Notebook Jupyter untuk analisis model v1.

[3] T.-L. Huoh, Y. Luo, P. Li, and T. Zhang. (2023). Flow-Based Encrypted Network Traffic Classification With Graph Neural Networks. *IEEE Transactions on Network and Service Management*, 20(2), 1224–1237.

[4] A. Presekal, I. Semertzis, H. Goyel, P. Palensky, and A. Ştefanov. (2025). Intrusion Detection System for Digital Substations using Semi-Supervised Learning and Traffic Distance Similarity Clustering. *IEEE Transactions on Smart Grid*.

[5] M. Lotfollahi, M. J. S. Shirazi, R. A. Shirali, S. H. R. G. V. Z. S. G. H. H. K. Z. K. (2020). Deep Packet: A Novel Approach for Encrypted Traffic Classification Using Deep Learning. *Software Networking*.