Sistem Data Loss Protection Berbasis Differencial Privacy Untuk Mitigasi Ancaman Ekstraksi Data Pada Model Generative AI

Riset PPMI Guru Besar - 2026

Suhardi [1], I Gusti Bagus Baskara Nugraha [1], Budiman Dabarsyah [1].

[1] Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung, Indonesia

Latar Belakang & Kondisi Mitra

Peningkatan adopsi AI dan sistem digital menempatkan data pada risiko eksfiltrasi yang masif. Ancaman keamanan kini tidak lagi bersifat tunggal, melainkan multi-vektor mulai dari kelemahan tata kelola privasi sistem hingga serangan *ransomware* dan penyusupan jaringan tingkat lanjut.

Awalnya, mitigasi riset ini difokuskan pada *Data Loss Prevention* (DLP) untuk AI Generatif. Namun, kondisi mitra di lapangan menunjukkan bahwa kebocoran data membutuhkan arsitektur perlindungan mendalam (*defense-in-depth*) yang holistik, mencakup kepatuhan hukum, pengamanan titik akhir (*endpoint*), hingga pelacakan rute transmisi.

Di ranah manajerial, institusi kesulitan menerjemahkan regulasi privasi yang abstrak menjadi kontrol teknis. Di ranah komputasi komersial, detektor keamanan konvensional lumpuh oleh peringatan palsu (*alert fatigue*) saat menghadapi *ransomware* modern yang menyamar sebagai proses administratif normal.

Di ranah infrastruktur, pelacakan pencurian data sering terputus ketika penyerang menggunakan proksi anonim terenkripsi. Simulator jaringan yang ada pun gagal membantu mitra karena tidak mampu memodelkan anomali fisik (*TCP re-segmentation*) pada arus data di dunia nyata secara akurat.

Keyword: Keamanan, Privasi, Ransomware.

Penelitian ini didukung oleh pendanaan dari Riset PPMI STEI 2026-Riset Guru Besar, Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung.

Tujuan & Solusi Program

Program riset multidisiplin ini bertujuan untuk merancang ekosistem pelindungan data ujung-ke-ujung (*end-to-end*). Arsitektur dirancang untuk mengatasi kerentanan sistem mitra mulai dari fase desain sistem (hulu) hingga pelacakan forensik pasca-insiden (hilir).

Pilar pertama adalah inovasi *Modular Privacy Engineering Framework* (MPEF). Kerangka ini mendekomposisi tata kelola rekayasa privasi menjadi lima kapabilitas operasional untuk menjembatani kepatuhan hukum menjadi bukti teknis yang dapat diaudit secara nyata dan otomatis.

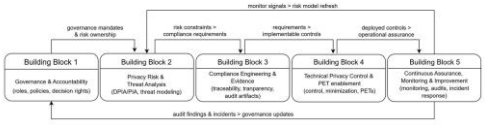
Pilar kedua berfokus pada pelindungan operasional melalui arsitektur AI *HIERA-TD*. Solusi ini mendeteksi *ransomware* pra-enkripsi secara instan menggunakan penalaran jalur-ganda (GNN dan SCM) yang mampu mengeliminasi peringatan palsu secara total bagi tim keamanan.

Pilar ketiga menghadirkan ekosistem *testbed* lintas-lapisan berbasis *Software-Defined Networking* (SDN) dan arsitektur *Network Watermark Tracing* (NWT). Sistem ini menyuntikkan waktu tunda (*Inter-Packet Delay*) untuk melacak rute eksfiltrasi, yang distabilkan guna menahan serangan gangguan waktu dari musuh.

Metode & Pelaksanaan Kegiatan

Pelaksanaan kegiatan dibagi menjadi tiga alur metodologi yang berjalan paralel dan saling melengkapi. Integrasi ketiganya memastikan bahwa kebijakan tata kelola selaras dengan ketahanan aplikasi kecerdasan buatan dan infrastruktur perutean jaringan secara fisik.

Pada dimensi privasi (MPEF), metode tinjauan literatur sistematis mengekstraksi parameter kapabilitas global. Kerangka yang terbangun kemudian divalidasi ketat melalui evaluasi empiris berorientasi praktisi yang melibatkan 34 profesional keamanan dari lintas sektor industri.

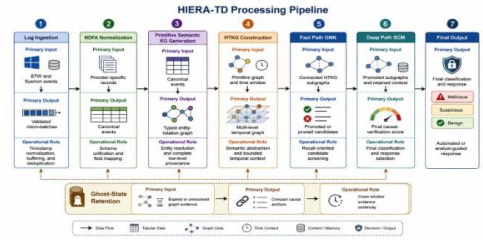


Picture 1. Arsitektur MPEF

Pada dimensi titik akhir (HIERA-TD), log sistem diubah menjadi *Hierarchical Temporal Knowledge Graph* (HTKG). Evaluasi dilakukan menggunakan emulator ancaman yang aman untuk memicu algoritma analisis AI (*Fast Path* dan *Deep Path*) tanpa risiko merusak infrastruktur.

Pada dimensi jaringan (NWT), *testbed* emulasi dibangun secara fisik menggunakan *Mininet* untuk menyediakan tumpukan TCP/IP secara utuh. Intersepsi paket disinkronkan guna menghadapi serangan *traffic shaping* musuh dan mutasi ukuran *payload* aplikasi secara bersamaan.

Seluruh purwarupa diuji secara rigor menggunakan beban kerja ekstrim. Metrik komparatif seperti analisis kesenjangan kelayakan (*feasibility gap*), *Bit Error Rate* (BER) forensik, dan margin keselamatan (*safety margin*) diukur guna memastikan keandalan seluruh inovasi di dunia nyata.



Picture 2. Pipeline pemrosesan HIERA-TD

Hasil & Peningkatan Keberdayaan Mitra

Pelaksanaan riset sukses menghadirkan ekosistem pertahanan berlapis yang tervalidasi secara empiris. Inovasi MPEF mencatatkan nilai penerimaan tinggi (4,196 dari skala 5) dan secara presisi berhasil memetakan titik lemah industri pada otomatisasi de-identifikasi data saat ini.

Model AI HIERA-TD beroperasi sempurna dengan latensi super cepat di bawah 1 milidetik. Sistem cerdas ini sukses memblokir proses *ransomware* 6 detik penuh sebelum kerusakan sistem terjadi, sekaligus membunuh angka peringatan palsu hingga mencapai nol mutlak.

Pada pengujian infrastruktur, arsitektur NWT sukses membendung serangan perturbasi waktu (*timing perturbation*) agresif dari musuh. Kesalahan pembacaan pelacakan forensik (*Bit Error Rate*) berhasil dijaga stabil di bawah 10% meskipun jaringan dihantam oleh *jitter* ekstrem sebesar 15 milidetik.

Keberdayaan mitra meningkat drastis dari postur keamanan reaktif menjadi ekosistem *defense-in-depth* prediktif. Tim operasional kini difasilitasi dengan kepatuhan pelindungan *by-design*, deteksi ancaman instan tanpa *alert fatigue*, serta kapabilitas pelacakan eksfiltrasi data jaringan yang kebal manipulasi.

Luaran & Keberlanjutan Progam

Janji luaran utama Riset PPMI telah dilunasi seratus persen melalui publikasi ilmiah yang telah terbit (*Published*) pada jurnal bereputasi tinggi IEEE Access (Q1) dengan judul "A Modular Privacy Engineering Framework for Regulatory-Compliant System Design: Capability Composition, Evidence Traceability, and Practitioner-Oriented Evaluation", pada bulan Juli 2026. Publikasi ini mendiseminasi cetak biru arsitektur rekayasa privasi ke komunitas *software engineering* global.

Riset ini juga menghasilkan nilai tambah inovatif berupa dua draf naskah publikasi tingkat internasional tambahan yang mengeksplorasi NWT dan HIERA-TD. Keberhasilan ini turut melahirkan artefak teknologi berupa purwarupa *testbed* SDN dan skrip operasional model deteksi AI.

Keberlanjutan program dipastikan melalui penyusunan peta jalan riset tingkat lanjut. Pengujian masa depan akan melibatkan aplikasi studi kasus di organisasi korporat, evaluasi model kecerdasan buatan pada lingkungan terisolasi (*malware sandbox*), serta optimasi ekstraksi forensik.

Secara makro, integrasi inovasi dari riset terapan ini diagendakan untuk dihilirisasi menjadi standar operasional Pusat Operasi Keamanan (*SOC*). Temuan ilmiah akan diintegrasikan ke dalam materi literasi dan pelatihan profesional untuk mencetak SDM siber nasional yang unggul.