BAB V PENUTUP

5.1. Kesimpulan

Penelitian ini telah berhasil mengembangkan, membandingkan, dan mengintegrasikan sistem deteksi wallet phishing berbasis Machine Learning pada jaringan Ethereum dengan mekanisme pencatatan risk score ke dalam smart contract blockchain. Berdasarkan hasil eksperimen yang telah dilakukan, dapat ditarik kesimpulan yang menjawab setiap research question sebagai berikut:

- (RQ1 — Performa Model): XGBoost dengan kombinasi fitur MulDiGraph dan Etherscan (Skenario 2, 26 fitur) adalah model terbaik secara keseluruhan dengan F1-Score 91,02%, Accuracy 97,50%, Precision 88,62%, dan Recall 93,56%. GNN-GraphSAGE V2 pada kondisi yang sama mencapai F1-Score 89,60% dengan Recall tertinggi (96,14%) dan FN Rate terendah (3,9%). Random Forest sebagai strong baseline mencapai F1-Score 87,93% pada dataset EPTransNet 1:10. Kedua model advanced (XGBoost dan GNN-GraphSAGE V2) secara konsisten mengungguli kedua model baseline (Logistic Regression F1 78,50% dan Random Forest F1 86,85%) pada semua skenario, memvalidasi kontribusi utama penelitian ini.
- (RQ2 — Overhead Blockchain): Pencatatan risk score ke smart contract RiskRegistry pada Hardhat localhost terbukti feasible dengan latency commit rata-rata 13,6 ms (min 11,2 ms, maks 38,4 ms) dan gas rata-rata 35.462 unit (std = 5). Hasil ini membuktikan kelayakan teknis integrasi ML-blockchain pada lingkungan proof-of-concept, bukan biaya tetap untuk produksi. Dengan asumsi ilustratif gas price 20 Gwei dan harga ETH US$2.000, 35.462 gas setara sekitar US$1,42 per commit di Ethereum Mainnet. Jaringan berbiaya lebih rendah seperti Polygon PoS atau Layer-2 seperti Arbitrum One berpotensi menurunkan biaya, tetapi nilai aktual harus diukur pada jaringan target karena fee model, kondisi jaringan, dan harga aset berubah.
- (RQ3 — Sumber Fitur Optimal): Fitur Etherscan secara konsisten meningkatkan F1-Score semua model (rata-rata +3,66 poin dari Skenario 1 ke Skenario 2). GNN mendapat manfaat terbesar (+5,28 poin). Fitur BigQuery

tidak memberikan nilai tambah independen setelah Etherscan tersedia — menunjukkan redundansi informasi antara BigQuery dan fitur graph. Fitur paling informatif adalah in_out_ratio (importance 0,142), unique_senders (0,131), dan in_degree (0,118), semuanya berasal dari fitur graph MulDiGraph.

- (Eksperimen tambahan late fusion): Pada protokol leakage-controlled, GraphSAGE–Etherscan late fusion menghasilkan F1-Score 88,25%. Ablasi pada embedding GraphSAGE yang sama menunjukkan bahwa tujuh fitur Etherscan menambah 6,41 poin F1 (81,84% menjadi 88,25%), sedangkan mengganti kepala linear dengan XGBoost saja tidak meningkatkan F1. Hasil ini menguatkan kesimpulan RQ3 bahwa Etherscan adalah sumber pengayaan yang bernilai; namun late fusion belum melampaui XGBoost Skenario 2 (91,02%).
- (Pengujian Hipotesis): Keempat hipotesis yang diajukan terbukti secara empiris — H1 (Etherscan meningkatkan F1: terbukti, rata-rata +3,66 poin), H2 (XGBoost F1 tertinggi: terbukti, 91,02%), H3 (GNN Recall tertinggi: terbukti, 96,14%), dan H4 (overhead blockchain < 50 ms: terbukti, 13,6 ms rata-rata).
- (Sistem End-to-End): Sistem end-to-end berhasil dibangun dan diverifikasi mencakup antarmuka Streamlit (port 8502), API server FastAPI (port 8000), modul inferensi Machine Learning, dan smart contract RiskRegistry pada Hardhat (chainId 31337), terintegrasi melalui web3.py. Risk score dapat dikueri kembali dari blockchain melalui endpoint GET /risk/{wallet} dengan data yang immutable dan timestamped.

Kontribusi utama penelitian ini terhadap bidang keilmuan adalah: (1) kerangka evaluasi terkontrol yang memisahkan kontribusi data-centric dan model-centric menggunakan populasi wallet serta split yang sama; (2) bukti empiris bahwa transformasi graph-to-tabular dan pengayaan Etherscan meningkatkan kualitas representasi, sedangkan penambahan BigQuery tidak otomatis meningkatkan performa karena redundansi fitur; (3) bukti bahwa XGBoost paling efektif memanfaatkan representasi MultiDiGraph dan Etherscan dengan F1-Score 91,02%, sementara late-fusion ablation menunjukkan tujuh fitur Etherscan menambah 6,41 poin F1 pada embedding GraphSAGE yang sama; dan (4) implementasi model terpilih

pada sistem end-to-end dengan DApp RiskRegistry sebagai mekanisme pencatatan risk score. Kontribusi keempat bersifat implementatif dan tidak diposisikan sebagai sumber peningkatan kinerja Machine Learning.

5.2. Keterbatasan Penelitian

Meskipun penelitian ini telah berhasil menjawab ketiga research question dan membuktikan keempat hipotesis, terdapat sejumlah keterbatasan penting yang perlu diakui secara eksplisit untuk memberikan konteks yang tepat terhadap temuan-temuan ini:

- Keterbatasan Dataset: EPTransNet menggunakan data snapshot Ethereum hingga tahun 2019, sehingga pola phishing yang lebih modern dan sophisticated (misalnya rug-pull, sandwich attacks, dan flash loan phishing yang berkembang pasca-2020 seiring ledakan DeFi) tidak terwakili dalam data. Model yang dilatih pada data ini mungkin mengalami concept drift ketika diaplikasikan pada transaksi Ethereum terkini.
- Keterbatasan Sampling: Rasio 1:10 yang digunakan masih merupakan distribusi artifisial yang jauh dari rasio populasi asli (1:2.550). Metrik performa yang dilaporkan (F1, Recall, Precision) hanya valid untuk distribusi yang digunakan dalam eksperimen dan tidak dapat diinterpretasikan langsung sebagai performa pada distribusi wallet nyata di blockchain.
- Keterbatasan GNN: Subgraph untuk GNN sangat sparse. Pada eksperimen late fusion yang dikendalikan, audit MulDiGraph menemukan 528 entri edge mentah yang terhitung ganda, atau hanya 264 edge terarah unik untuk 8.595 alamat. Sparsity ini membatasi efektivitas message passing dan menjadi faktor utama mengapa embedding graf maupun late fusion belum mengungguli XGBoost tabular.
- Keterbatasan Integrasi Blockchain: Seluruh pengujian blockchain dilakukan pada Hardhat localhost (chainId 31337) sebagai proof-of-concept. Kondisi ini tidak merepresentasikan latency jaringan publik, variabilitas gas price, potensi

kongesti mempool, atau biaya nyata pada Ethereum mainnet. Hasil pengukuran overhead tidak dapat digeneralisasi langsung ke kondisi deployment produksi.

- Keterbatasan Evaluasi: Eksperimen menggunakan satu fixed split (random_state=42) tanpa cross-validation k-fold, terutama karena keterbatasan waktu komputasi training GNN (90-118 epoch per run). Tanpa variance estimate, tidak dapat dipastikan apakah hasil F1 91,02% untuk XGBoost konsisten di berbagai split. Penelitian lanjutan sebaiknya melakukan Stratified 5-Fold CV minimal untuk model tabular (LR, RF, XGBoost) dan melaporkan mean ± std F1 untuk memperkuat klaim stabilitas model.
- Keterbatasan Risk Score: Formula $risk_{score} = round\left(P\left(phishing\right) \times 100\right)$ menggunakan probabilitas dari model yang tidak terkalibrasi. Uncalibrated probability berarti nilai P=0,7 tidak secara literal berarti "70% kemungkinan phishing" — nilai aktual bisa lebih tinggi atau lebih rendah tergantung distribusi skor model.
- Keterbatasan Fitur: Penelitian ini tidak mengeksplorasi potensi data leakage antara fitur Etherscan/BigQuery dan label phishing. Jika fitur Etherscan dikumpulkan setelah alamat wallet teridentifikasi sebagai phishing, ada kemungkinan fitur-fitur tersebut mencerminkan kondisi post-phishing (misalnya saldo mendekati 0 setelah dana dikuras) bukan kondisi during-phishing.

5.3. Saran untuk Pengembangan Selanjutnya

Berdasarkan keterbatasan yang telah diidentifikasi dan hasil temuan penelitian, beberapa rekomendasi konkret untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

- Pembaruan dan Ekspansi Dataset: menggunakan data Ethereum post-2020 yang mencakup era DeFi modern, termasuk pola phishing yang lebih sophisticated. Dataset dapat diperkaya melalui integrasi langsung dengan The Graph Protocol atau Ethereum Full Node untuk crawling real-time. Penambahan label dari sumber terpercaya seperti Chainabuse, CertiK, dan PeckShield akan meningkatkan kualitas dan kuantitas sampel phishing. Untuk menghadapi concept drift, pipeline deployment perlu memantau pergeseran

distribusi fitur dan skor secara otomatis (misalnya dengan PSI atau uji distribusi) serta perubahan metrik ketika label baru tersedia. Jika ambang drift terlampaui atau performa menurun, sistem dapat memicu retraining berkala/berbasis kondisi dan menerbitkan model_version baru agar perubahan tetap dapat ditelusuri.

- Penanganan Imbalanced Data yang Lebih Baik: mengeksplorasi kombinasi SMOTE (Synthetic Minority Over-sampling Technique) dengan Tomek Links atau ADASYN sebagai alternatif undersampling murni. Evaluasi menggunakan k-fold stratified cross-validation (minimal k=5) untuk mendapatkan estimasi variance metrik yang lebih reliable. Implementasi threshold optimization pada model untuk memaksimalkan F1 atau recall sesuai kebutuhan use case.
- Pengembangan Arsitektur GNN: mengeksplorasi Heterogeneous GNN, khususnya HAN atau HGT, untuk membedakan tipe node seperti EOA dan smart contract serta tipe edge seperti transfer ETH, transfer token ERC-20, dan approval. Pendekatan ini relevan dengan studi heterogeneous Ethereum fraud detection [39] dan berpotensi menangkap pola phishing yang tidak terlihat pada homogeneous graph. Temporal GNN yang mempertimbangkan urutan waktu transaksi juga dapat diuji untuk menangkap pola yang bergantung pada timing, sedangkan Graph Attention Networks (GAT) dapat memberi bobot berbeda pada tetangga yang berbeda.
- Pengayaan Topologi dan Fusi Lanjutan: membangun subgraph 1-hop dari MulDiGraph penuh untuk meningkatkan konektivitas sebelum memperluas GNN. Setelah itu, evaluasi dapat menguji embedding graf bersama seluruh 26 fitur tabular Skenario 2 (vektor 155 dimensi) dengan validasi internal dan ablasi modalitas yang sama. Optimasi ambang keputusan pada data validasi juga perlu dilakukan sebelum menyimpulkan perubahan F1 pada classifier embedding-only.
- Deployment pada Jaringan Produksi: mengimplementasikan smart contract RiskRegistry pada jaringan publik dan membandingkan Ethereum Mainnet dengan jaringan berbiaya lebih rendah seperti Polygon PoS serta Layer-2 seperti Arbitrum One. Evaluasi harus mengukur biaya aktual per commit,

latency konfirmasi, variabilitas fee, dan dampak kongesti, bukan hanya menggunakan hasil localhost. Untuk volume tinggi, optimasi dapat dilakukan melalui batch commit atau pencatatan commitment/hash agregat agar biaya rata-rata per risk score menurun tanpa menghilangkan audit trail.

- Kalibrasi Probabilitas dan Threshold: mengimplementasikan Platt Scaling atau Isotonic Regression agar risk score memiliki interpretasi probabilistik yang lebih reliable, kemudian mengevaluasinya dengan reliability diagram dan Expected Calibration Error (ECE). Setelah probabilitas terkalibrasi, DApp dapat menyediakan threshold yang dapat diatur pengguna sesuai toleransi risiko: threshold lebih rendah memprioritaskan recall, sedangkan threshold lebih tinggi mengurangi false positive. Default threshold tetap ditentukan dari data validasi; pilihan pengguna hanya mengubah kebijakan peringatan dan tidak mengubah parameter model.
- Integrasi dengan Ekosistem Web3: mengembangkan browser extension atau API publik yang memungkinkan pengguna DeFi memeriksa risk score suatu alamat wallet secara real-time sebelum melakukan transaksi. Integrasi dengan MetaMask Snaps atau WalletConnect dapat memperluas jangkauan sistem ke pengguna crypto mainstream.
- Eksplorasi Model Advanced Lainnya: penelitian lanjutan dapat mengeksplorasi model advanced tambahan seperti LightGBM, TabNet, atau Graph Transformer Networks untuk memperkuat perbandingan dalam kategori model advanced, sehingga kesimpulan tentang superioritas XGBoost dan GNN menjadi lebih robust. Validasi Multi-seed dan Reproducibility: menjalankan seluruh eksperimen dengan minimal 5 nilai random_state berbeda dan melaporkan mean ± std dari setiap metrik. Ini akan meningkatkan validitas statistik dan memungkinkan peneliti lain untuk membandingkan hasil dengan lebih akurat. Publikasi seluruh kode dan data preprocessing di repository publik (GitHub) untuk mendukung reproducibility.
- Explainability (XAI): mengintegrasikan metode SHAP (SHapley Additive exPlanations) atau LIME untuk memberikan penjelasan per-prediksi — "mengapa wallet ini mendapat risk score tinggi?" Ini akan meningkatkan

kepercayaan pengguna dan membantu auditor dalam memeriksa keputusan model, sejalan dengan prinsip responsible AI.

Secara keseluruhan, penelitian ini telah meletakkan fondasi yang kuat untuk pengembangan sistem deteksi phishing blockchain yang lebih komprehensif. Kombinasi pendekatan Machine Learning multi-model, eksplorasi sumber fitur hybrid, dan integrasi blockchain melalui smart contract membuka jalur penelitian baru yang relevan di persimpangan antara AI, keamanan siber, dan teknologi blockchain.

UNIVERSITAS
NUSA MANDIRI