RELIABILITAS KLASIFIKASI SERANGAN BERBASIS TELEMETRI SENSOR IoT DENGAN XAI

M. Muntaha[1], Muhammad Khamim Asy'ari[2]
[1,2] Program Studi Teknik Fisika, Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo, Surabaya 60111, Indonesia
[1]mmuntaha79792@gmail.com, , [2] mkasyari@its.ac.id

Abstract

This study evaluates the reliability of attack classification decisions from ToN_IoT telemetry, defined as predictive performance and its stability under controlled evaluation conditions, rather than hardware reliability or sensor accuracy. Weather, GPS Tracker, Thermostat, Modbus, and Fridge are assessed in multiclass and binary tasks. The protocol combines string normalization, removal of time fields and exact duplicates, five-fold predictor-signature group separation, validation-based model selection, dummy baselines, class-wise chronological sensitivity testing, signal perturbations, and permutation importance. Mean±sample standard deviation of multiclass macro-F1 is 95.32±0.31% for Weather, 73.84±1.19% for GPS, 14.07±0.19% for Thermostat, 15.03±0.60% for Modbus, and 13.14±0.67% for Fridge. Weather and GPS exceed their dummy baselines, whereas the other three devices provide no comparable improvement. Binary macro-F1 is 96.99±0.31%, 94.03±0.21%, 50.12±0.37%, 49.47±1.25%, and 46.15±0.00%, respectively. All 151 Fridge predictor signatures occur with conflicting labels, limiting class separability in the deduplicated representation. Its 85.71% binary accuracy therefore coexists with only 50.00% balanced accuracy. Chronological testing reduces Weather and GPS multiclass macro-F1 to 23.35% and 8.05%; missingness, noise, and drift also reduce performance. Permutation importance describes model reliance for Weather and GPS, with Thermostat included as a baseline-level comparator. These results show substantial differences in predictive capability across devices and limited temporal generalization. Reliable telemetry-based security evaluation requires complementary group, temporal, disturbance, and explanatory analyses; the findings do not establish real-time detection performance.

Keywords: *attack classification, IoT telemetry, classification decision reliability, group-disjoint and temporal evaluation, XAI*

Abstrak

Penelitian ini mengevaluasi reliabilitas keputusan klasifikasi serangan berbasis telemetri ToN_IoT, yaitu kinerja prediktif dan kestabilannya pada kondisi evaluasi yang dikendalikan, bukan reliabilitas perangkat keras atau akurasi sensor. Lima perangkat, yakni Weather, GPS Tracker, Thermostat, Modbus, dan Fridge, diuji pada tugas multikelas dan biner. Protokol menggabungkan normalisasi string, penghapusan kolom waktu dan duplikat tepat, pemisahan grup tanda tangan prediktor dalam lima lipatan, pemilihan model berdasarkan validasi, baseline dummy, uji sensitivitas kronologis per kelas, perturbasi sinyal, dan permutation importance. Rerata±simpangan baku sampel macro-F1 multikelas berturut-turut adalah 95,32±0,31% untuk Weather, 73,84±1,19% untuk GPS, 14,07±0,19% untuk Thermostat, 15,03±0,60% untuk Modbus, dan 13,14±0,67% untuk Fridge. Weather dan GPS melampaui baseline dummy, sedangkan tiga perangkat lainnya tidak menunjukkan peningkatan yang sebanding. Macro-F1 biner kelima perangkat berturut-turut sebesar 96,99±0,31%, 94,03±0,21%, 50,12±0,37%, 49,47±1,25%, dan 46,15±0,00%. Seluruh 151 tanda tangan prediktor Fridge memiliki konflik label, sehingga keterpisahan kelas terbatas pada representasi hasil deduplikasi. Accuracy biner Fridge sebesar 85,71% hanya disertai balanced accuracy 50,00%. Pada uji kronologis, macro-F1 multikelas Weather dan GPS turun menjadi 23,35% dan 8,05%; missing, noise, dan drift juga menurunkan kinerja. Permutation importance menjelaskan ketergantungan model Weather dan GPS, dengan Thermostat sebagai pembanding berskor setara baseline. Temuan menunjukkan kemampuan prediktif yang berbeda antarperangkat dan keterbatasan generalisasi temporal. Evaluasi keamanan berbasis telemetri memerlukan pemeriksaan grup, waktu, gangguan, dan penjelasan model secara bersama; hasil ini belum membuktikan kinerja deteksi waktu nyata.

Kata kunci: : *klasifikasi serangan, telemetri IoT, reliabilitas keputusan klasifikasi, evaluasi antarkelompok dan temporal, XAI*

1. PENDAHULUAN

Telemetri Internet of Things (IoT) mengubah besaran fisik dan status proses menjadi data untuk pemantauan dan keputusan. Dalam keamanan siber, perubahan operasi akibat serangan mungkin meninggalkan jejak pada kanal tersebut. Reliabilitas dalam penelitian ini secara khusus berarti kinerja keputusan klasifikasi dan kestabilannya ketika artefak data serta kondisi evaluasi dikendalikan; istilah ini tidak mengacu pada probabilitas kegagalan perangkat keras atau ketelitian sensor dalam satuan fisik. Model dapat memakai waktu perekaman, inkonsistensi string, pengulangan nilai, atau dominasi kelas sebagai jalan pintas yang tidak bertahan ketika kondisi operasi berubah

ToN_IoT menggabungkan telemetri, trafik jaringan, dan log sistem operasi dari testbed IoT/IIoT [1], [2], [3]. Studi deteksi serangan umumnya menekankan performa algoritma; contoh pada JIRE melaporkan akurasi ANN untuk log Apache [4]. Pada kajian peramalan cuaca berbasis data sensor, missing value dan noise dilaporkan sebagai masalah kualitas data yang memerlukan prapemrosesan [5]. Karena itu, kajian berbasis telemetri perlu memisahkan tiga konsep: akurasi sensor dalam satuan fisik, reliabilitas perangkat keras, dan reliabilitas keputusan klasifikasi. Penelitian ini hanya menilai konsep ketiga.

Le dkk. menggabungkan ensemble trees dan SHAP untuk menjelaskan keputusan IDS [6], pengembangan XAI pada anomali IoT dan kajian trafik jaringan juga menekankan interpretabilitas model [7], [8]. Sementara itu, kajian leakage menyoroti kontaminasi evaluasi [9], dan Bouke serta Abdullah menguji dampak pattern leakage pada beberapa dataset IDS [10]. Fokus tersebut berbeda dari pertanyaan apakah fitur telemetri yang sama tetap mampu membedakan label setelah input identik dipisahkan, urutan waktu diubah menjadi blok pengujian, dan sinyal diberi gangguan. Concept drift dan ketidakseimbangan kelas menambah alasan untuk tidak mengandalkan accuracy tunggal [11], [12]. Celah yang dikaji adalah keterkaitan validitas prediksi dan penjelasannya pada kanal telemetri, sehingga audit leakage, validasi antarkelompok, sensitivitas temporal, perturbasi, dan XAI perlu dianalisis dalam satu rangkaian evaluasi.

Penelitian ini bertujuan menguji keterpisahan label serangan dari prediktor telemetri setelah artefak nonfisik dan pengukuran identik dikendalikan, menilai kestabilan terhadap waktu dan gangguan sinyal, serta menjelaskan ketergantungan model pada fitur. Kontribusinya adalah integrasi audit string/date leakage dan konflik label, validasi luar berbasis tanda tangan prediktor dengan pemilihan model pada validation, serta evaluasi temporal dan perturbasi yang dibaca bersama baseline dan XAI. Integrasi ini membedakan keterpisahan dalam distribusi, kestabilan temporal, dan ketergantungan fitur, sehingga skor atau penjelasan yang tinggi tidak langsung dianggap sebagai bukti keputusan yang andal. Keterdeteksian prediktif berarti keterpisahan empiris label pada fitur dan protokol tersebut, bukan observabilitas formal dalam teori kontrol.

2. TINJAUAN PUSTAKA

2.1. Telemetri dan reliabilitas keputusan

Dataset ToN_IoT dirancang untuk evaluasi analitik keamanan pada sumber data heterogen[1] [2]. Booij dkk. [3] menunjukkan bahwa heterogenitas fitur dan tipe serangan memengaruhi evaluasi. Pada penelitian ini, reliabilitas keputusan didefinisikan sebagai kestabilan macro-F1 dan balanced accuracy ketika pengukuran identik dipisahkan antarsplit, urutan waktu diuji secara konservatif, dan input mengalami gangguan terkontrol. Definisi ini tidak menyatakan probabilitas kegagalan sensor, ketertelusuran kalibrasi, atau uncertainty metrologis.

2.2. Leakage, imbalance, robustness, dan XAI

Leakage dapat terjadi ketika informasi uji memengaruhi pelatihan atau ketika pola identik tersebar pada train dan test [9], [10]. Pada kelas tidak seimbang, accuracy dapat didominasi kelas mayoritas dan menghasilkan penilaian yang terlalu optimistis [5], [12]. Karena itu, penelitian ini menggunakan macro-F1 dan balanced accuracy sebagai ringkasan performa per kelas. Drift menggeser distribusi fitur sehingga model yang valid pada data bersih belum tentu stabil [11]. Studi IoT IDS dan klasifikasi trafik juga menunjukkan kegunaan XAI untuk mengidentifikasi atau memeringkat fitur yang memengaruhi prediksi [13], [14]. Permutation importance dipilih karena dapat diterapkan pada berbagai model dan mengukur penurunan skor uji setelah satu fitur diacak [15]; nilainya tetap

asosiatif dan interpretasinya dapat terganggu oleh korelasi antarprediktor [16]

TABEL I. PERBANDINGAN FOKUS PENELITIAN TERDAHULU DAN PENELITIAN INI

Studi	Data	Split	Audit leakage	Metrik imbalance	Robustness	XAI
Alsaedi dkk. [1]	Telemetri ToN_IoT	Train/test + k-fold tuning	Tidak khusus	F1/recall/accuracy	—	—
Booij dkk. [3]	ToN_IoT dan Aposemat	Komparasi/cross-training	Heterogenitas	F1/AUC/accuracy	—	—
Le dkk. [6]	Trafik IDS	Hold-out	Tidak khusus	F1/accuracy	—	SHAP
Bouke & Abdullah [10]	3 dataset IDS	Eksperimen leakage	Ya	F1/accuracy	—	—
Alani dan Miri [13]	Beberapa dataset IoT	Komparatif	Tidak khusus	F1/accuracy	—	XAI fitur
Penelitian ini	5 telemetri sensor	Outer group + temporal	Date, string, duplikat	Macro-F1/BA/dummy	Missing/noise/drift	PI ± SD

Tabel 1 menempatkan kontribusi pada rancangan evaluasi, bukan algoritma baru atau klaim bahwa setiap komponennya belum pernah digunakan. Benchmark ToN_IoT [1] menjadi dasar evaluasi telemetri, sedangkan kajian heterogenitas [3] membahas kesesuaian fitur dan tipe serangan antardataset. Studi pattern leakage [10] menjelaskan risiko prapemrosesan, sementara Alani dan Miri [13] menggabungkan fitur trafik yang dapat digunakan lintas dataset dengan penjelasan SHAP. Penelitian ini menghubungkan persoalan tersebut pada lima kanal telemetri: audit dan pemisahan grup memeriksa artefak evaluasi; pengujian kronologis dan perturbasi memeriksa kestabilan; dan XAI menilai ketergantungan model setelah kemampuan prediktif diperiksa. Hasil negatif pada tiga perangkat serta kesenjangan grup–waktu menjadi bagian dari kontribusi empiris, bukan sekadar pelengkap skor tinggi.

3. METODOLOGI PENELITIAN

3.1. Data dan batas penelitian

Lima berkas train-test ToN_IoT digunakan: Weather, GPS Tracker, Thermostat, Modbus, dan Fridge. Tugas multikelas memprediksi normal atau tipe serangan; tugas biner menggabungkan seluruh serangan sebagai Attack. Prediktor masing-masing adalah temperature–pressure–humidity; latitude–longitude; current_temperature–thermostat_status; empat register/coil Modbus; dan fridge_temperature–temp_condition. Analisis merupakan benchmark offline dan tidak mencakup packet capture langsung, respons otonom, atau kerusakan perangkat keras.

TABEL II. AUDIT DATASET SETELAH *STRIP* DAN DEDUPLIKASI TEPAT

Perangkat	Sumber	Analisis	Dihapus (%)	Prediktor	Kelas	Tanda tangan konflik
Weather	39.260	37.988	3,2	3	8	0
GPS	38.960	38.661	0,8	2	8	1
Thermostat	32.774	30.971	5,5	2	7	3
Modbus	31.106	8.868	71,5	4	6	1

Perangkat	Sumber	Analisis	Dihapus (%)	Prediktor	Kelas	Tanda tangan konflik
Fridge	39.944	1.057	97,4	2	7	151

Tanda tangan konflik adalah kombinasi prediktor identik yang memiliki lebih dari satu label setelah deduplikasi tepat.

3.2. Audit dan prapemrosesan

Nama kolom dan seluruh string dinormalisasi dengan strip sebelum langkah lain. Kolom ts, date, dan time dikeluarkan karena menunjukkan jadwal perekaman, bukan respons fisik. Duplikat tepat kemudian dihapus atas gabungan seluruh prediktor, label biner, dan tipe serangan. Penyertaan target mencegah penghapusan observasi konflik secara sepihak. Selanjutnya, hash hanya atas prediktor membentuk tanda tangan grup. Dengan desain ini, dua baris dengan input identik—meskipun labelnya berbeda—tidak boleh tersebar antara train, validation, dan test.

Deduplikasi menghapus 71,5% baris Modbus dan 97,4% Fridge (Tabel 2). Pada Fridge, hanya terdapat 151 tanda tangan prediktor dan seluruhnya berkonflik antarlabel; 1.057 baris deduplikasi tepat merupakan tujuh label pada masing-masing tanda tangan. Temuan tersebut menunjukkan bahwa frekuensi pengulangan tidak boleh disamakan dengan informasi baru. Imputasi median/modus, standardisasi numerik, dan one-hot encoding di-fit hanya pada bagian latih setiap iterasi.

TABEL III. DISTRIBUSI KELAS TOTAL/LATIH/VALIDASI/UJI PADA *OUTER FOLD* 1

Perangkat	Kelas: total/latih/validasi/uji
Weather	Back: 4847/2908/969/970; Ddos: 4825/2895/965/965; Inje: 4827/2896/966/965; Norm: 14458/8674/2892/2892; Pass: 4874/2924/975/975; Rans: 2776/1666/555/555; Scan: 525/315/105/105; Xss: 856/514/171/171
GPS	Back: 5000/3000/1000/1000; Ddos: 4999/2999/1000/1000; Inje: 5000/3000/1000/1000; Norm: 14709/8826/2942/2941; Pass: 4995/2997/999/999; Rans: 2832/1699/567/566; Scan: 550/330/110/110; Xss: 576/345/115/116
Thermostat	Back: 4741/2845/948/948; Inje: 4749/2850/950/949; Norm: 14121/8473/2824/2824; Pass: 4738/2843/948/947; Rans: 2144/1287/428/429; Scan: 58/34/12/12; Xss: 420/253/83/84
Modbus	Back: 1544/927/308/309; Inje: 1323/793/265/265; Norm: 3752/2251/751/750; Pass: 1712/1027/343/342; Scan: 292/175/58/59; Xss: 245/147/49/49
Fridge	Back: 151/90/30/31; Ddos: 151/90/30/31; Inje: 151/90/30/31; Norm: 151/90/30/31; Pass: 151/90/30/31; Rans: 151/90/30/31; Xss: 151/90/30/31

Singkatan kelas: Back = *backdoor*, Ddos = DDoS, Inje = *injection*, Norm = normal, Pass = *password*, Rans = *ransomware*, Scan = *scanning*, dan Xss = XSS. Urutan angka: total/latih/validasi/uji..

3.3. Validasi antarkelompok dan temporal

Protokol awal yang utama memakai StratifiedGroupKFold lima lipatan dengan shuffle=True, random_state=42, dan tanda tangan prediktor sebagai grup. Pada iterasi ke-i, lipatan ke-i menjadi test, lipatan berikutnya secara siklik menjadi validation, dan tiga lipatan sisanya menjadi train. Setiap kandidat LR, DT, RF, dan GB beserta prapemrosesnya dilatih hanya pada train yang sama. Kandidat dipilih berdasarkan macro-F1 validation tertinggi pada iterasi tersebut; jika seri, urutan prioritasnya LR, DT, RF, lalu GB. Parameter pada Bagian 3.4 ditetapkan tanpa pencarian hiperparameter tambahan. Pipeline terpilih yang sudah di-fit pada train langsung digunakan untuk memprediksi test; tidak dilakukan refit menggunakan gabungan train dan validation. Seed model adalah 42 + 17(i−1). Skor test tidak digunakan dalam aturan pemilihan. Setiap grup menjadi test sekali, dan hasil primer merupakan rerata serta simpangan baku sampel dari lima skor test model yang terpilih per iterasi. Model modal pada Tabel 4 hanya merangkum frekuensi pemilihan, bukan satu model akhir yang dilatih pada seluruh data.

Audit timestamp menemukan satu blok kronologis per tipe serangan pada berkas ini. Pemisahan global 60:20:20 menghasilkan cakupan kelas train dan test yang berbeda; leave-one-episode-out per kelas juga tidak dapat dilakukan karena hanya tersedia satu episode. Uji sensitivitas mengurutkan setiap kelas menurut

waktu, lalu membaginya menjadi 60% train, 5% embargo, 15% validation, 5% embargo, dan 15% test. Arsitektur RF dipertahankan dari pemilihan pada validasi grup; pipeline temporal di-fit ulang hanya pada 60% train dengan seed=42. Bagian validation 15% disisihkan dan tidak digunakan. untuk tuning atau refit. Karena itu, hasil temporal merupakan uji sensitivitas arsitektur terpilih pada bagian akhir episode yang sama, bukan evaluasi pemilihan model temporal yang independen atau validasi lintas episode.

3.4. Model, baseline, dan metrik

LR menggunakan max_iter 700 dan class_weight balanced; DT memakai max_depth 24, min_samples_leaf 5, dan class_weight balanced; RF memakai 150 pohon, min_samples_leaf 2, dan balanced_subsample; GB memakai parameter baku dengan seed iterasi. Baseline terdiri atas DummyClassifier most-frequent dan stratified. Balanced accuracy teoretis untuk tebakan tidak informatif adalah 1/K, sedangkan baseline macro-F1 harus diestimasi dari dummy karena tidak selalu sama dengan 1/K.

Metrik primer adalah macro-F1 dan balanced accuracy (BA). Accuracy tetap dilaporkan sebagai pembanding. Untuk K kelas, macro-F1 = (1/K)ΣF1_k dan BA = (1/K)ΣRecall_k. Kalibrasi probabilitas tidak dievaluasi karena pertanyaan penelitian berfokus pada keputusan label dan kestabilan performa. Implementasi memakai Python 3.12.13, NumPy 2.3.5, pandas 2.2.3, dan scikit-learn 1.8.0.

3.5. Perturbasi dan XAI

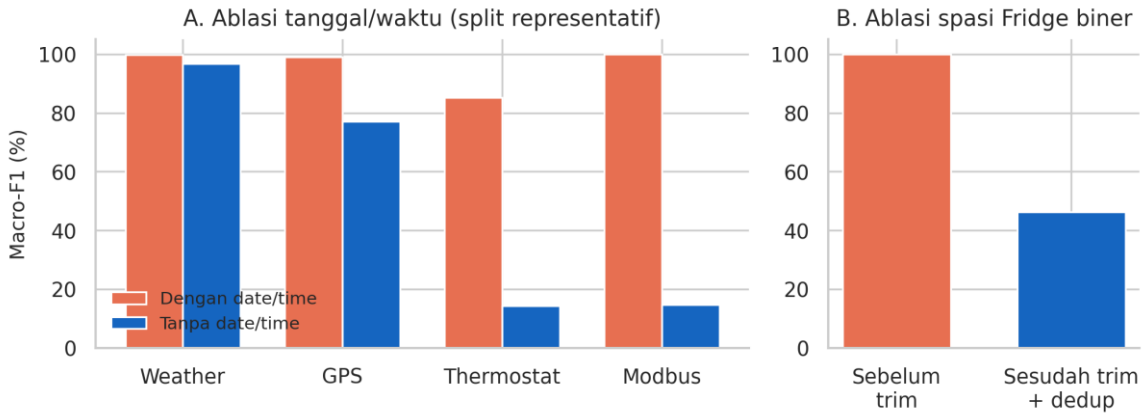
Uji perturbasi difokuskan pada Weather dan GPS multikelas karena keduanya memiliki skor antarkelompok yang melampaui baseline. RF terpilih pada seluruh lipatan kedua perangkat. Pipeline diagnostik dilatih kembali hanya pada train dengan parameter dan seed yang sama, tanpa memasukkan validation atau test. Missing completely at random diterapkan pada 5% dan 10% sel; noise Gaussian nol-rerata memakai simpangan baku 5% dan 10% dari SD latih; drift menambahkan offset ±5%, ±10%, dan ±20% SD latih. Nilai numerik dibatasi pada rentang min–maks data latih. Missing dan noise diulang sepuluh realisasi per outer fold; drift dievaluasi pada kedua arah pada lima fold. Persentase merupakan skala sensitivitas ternormalisasi antarsensor, bukan klaim toleransi fisik lapangan.

Permutation importance merupakan analisis diagnostik terbatas pada Weather, GPS, dan Thermostat multikelas. Weather dan GPS mewakili model yang melampaui baseline; Thermostat disertakan sebagai pembanding berskor setara baseline untuk memperlihatkan batas penafsiran importance pada model lemah. Ketiganya memilih RF pada seluruh lipatan. Modbus juga berskor setara baseline, sedangkan Fridge memiliki konflik label pada seluruh tanda tangan prediktor; importance keduanya tidak dihitung dalam eksperimen ini. Pembatasan tersebut bukan bukti bahwa fitur keduanya tidak relevan dan tidak mendukung perbandingan peringkat fitur kelima perangkat. Untuk setiap lipatan, RF di-fit kembali hanya pada train dengan parameter dan seed yang sama; maksimal 2.000 sampel test terstratifikasi digunakan dengan sepuluh pengacakan per fitur dan scoring macro-F1. Gambar 6 melaporkan rerata dan simpangan baku antar-fold; variasi dalam fold tersimpan pada JSON. Importance mengukur ketergantungan model, bukan pengaruh kausal; korelasi antarprediktor membatasi interpretasinya [16] [17].

4. HASIL DAN PEMBAHASAN

4.1. Ablasi artefak nonfisik

Gambar 1A menunjukkan bahwa date/time menghasilkan shortcut kuat pada split representatif: Weather dan GPS mendekati 100%, sedangkan penghapusannya terutama meruntuhkan Thermostat dan Modbus. Pada data mentah, kelas serangan memang direkam pada periode berbeda; model dapat mengidentifikasi jadwal eksperimen tanpa membaca respons sensor. Gambar 1B menunjukkan leakage kedua: spasi tersembunyi pada temp_condition Fridge menghasilkan macro-F1 biner 100%, tetapi strip dan deduplikasi menurunkannya menjadi 46,19%. Ablasi ini diperlakukan sebagai audit sensitivitas, bukan hasil primer outer-fold.



Gambar 1. Ablasi Date/Time Dan Spasi Tersembunyi Pada Split Representatif.

4.2. Kinerja validasi antarkelompok

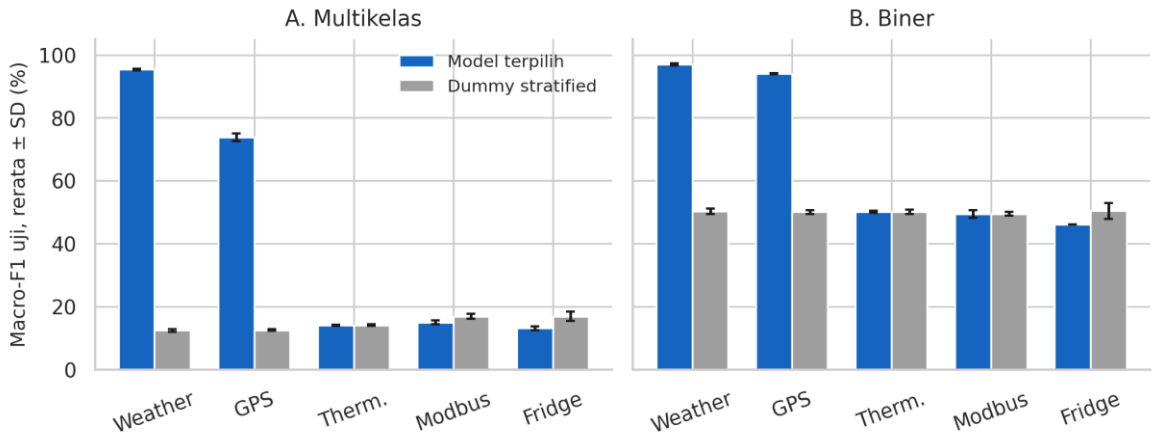
Tabel 4 dan Gambar 2 melaporkan hasil primer. RF terpilih pada seluruh fold Weather, GPS, Thermostat, dan Fridge multikelas. Weather mempertahankan macro-F1 95,32±0,31% (multikelas) dan 96,99±0,31% (biner). GPS mencapai 73,84±1,19% dan 94,03±0,21%. Keduanya jauh di atas dummy stratified, sehingga nilai tinggi bukan sekadar akibat proporsi kelas atau duplikat prediktor lintas split.

Thermostat, Modbus, dan Fridge tidak menunjukkan peningkatan yang meyakinkan dibanding dummy stratified. Macro-F1 multikelasnya masing-masing 14,07±0,19%, 15,03±0,60%, dan 13,14±0,67%, dibanding baseline 14,15±0,33%, 16,88±0,80%, dan 16,93±1,44%. Macro-F1 biner Thermostat dan Modbus sebesar 50,12±0,37% dan 49,47±1,25%, sedangkan Fridge 46,15±0,00%. Perbandingan ini bersifat deskriptif, tanpa klaim signifikansi statistik. Hasil rendah membatasi bukti keterpisahan pada representasi fitur dan kandidat model yang diuji; hasil tersebut tidak membuktikan bahwa semua serangan tidak memengaruhi perangkat.

TABEL IV. PERFORMA *OUTER GROUP-DISJOINT* LIMA *FOLD* (%)

Perangkat	Tugas	Model modal	Macro-F1	BA	Accuracy	Dummy F1
Weather	Multi	RF	95,32 ± 0,31	95,17 ± 0,27	95,80 ± 0,22	12,44 ± 0,39
Weather	Biner	RF	96,99 ± 0,31	97,14 ± 0,28	97,15 ± 0,29	50,34 ± 0,87
GPS	Multi	RF	73,84 ± 1,19	73,68 ± 1,40	80,84 ± 0,54	12,59 ± 0,21
GPS	Biner	RF	94,03 ± 0,21	94,44 ± 0,17	94,33 ± 0,21	50,06 ± 0,61
Thermostat	Multi	RF	14,07 ± 0,19	14,25 ± 0,21	24,59 ± 0,43	14,15 ± 0,33
Thermostat	Biner	RF	50,12 ± 0,37	50,13 ± 0,37	50,49 ± 0,35	50,11 ± 0,75
Modbus	Multi	DT (3/5)	15,03 ± 0,60	16,43 ± 0,94	24,45 ± 8,21	16,88 ± 0,80
Modbus	Biner	LR (2/5)	49,47 ± 1,25	49,93 ± 1,20	51,32 ± 1,78	49,48 ± 0,66
Fridge	Multi	RF	13,14 ± 0,67	14,29 ± 0,00	14,29 ± 0,00	16,93 ± 1,44
Fridge	Biner	LR (4/5)	46,15 ± 0,00	50,00 ± 0,00	85,71 ± 0,00	50,41 ± 2,49

Model modal menunjukkan model yang paling sering terpilih; angka dalam kurung adalah jumlah *fold* jika kurang dari lima. Semua nilai adalah rerata ± SD *test*.

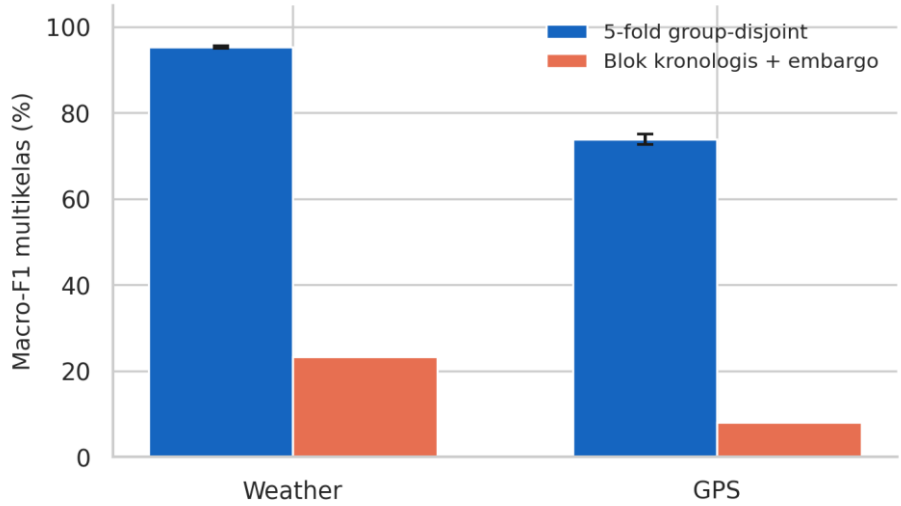


Gambar 2. Macro-F1 Outer Group-Disjoint Dan Dummyclassifier Stratified.

4.3. Sensitivitas temporal mengubah kesimpulan

Kinerja tinggi Weather dan GPS tidak bertahan pada blok kronologis (Gambar 3). Macro-F1 Weather turun dari 95,32% menjadi 23,35% dan GPS dari 73,84% menjadi 8,05%; balanced accuracy temporalnya 30,53% dan 20,40%. Penurunan bukan bukti bahwa serangan tidak memengaruhi sensor, tetapi menunjukkan distribusi bagian akhir episode berbeda dari bagian awal yang dipelajari. Dengan demikian, klaim yang didukung data adalah sinyal kuat dalam distribusi antarkelompok, bukan generalisasi waktu atau deployment.

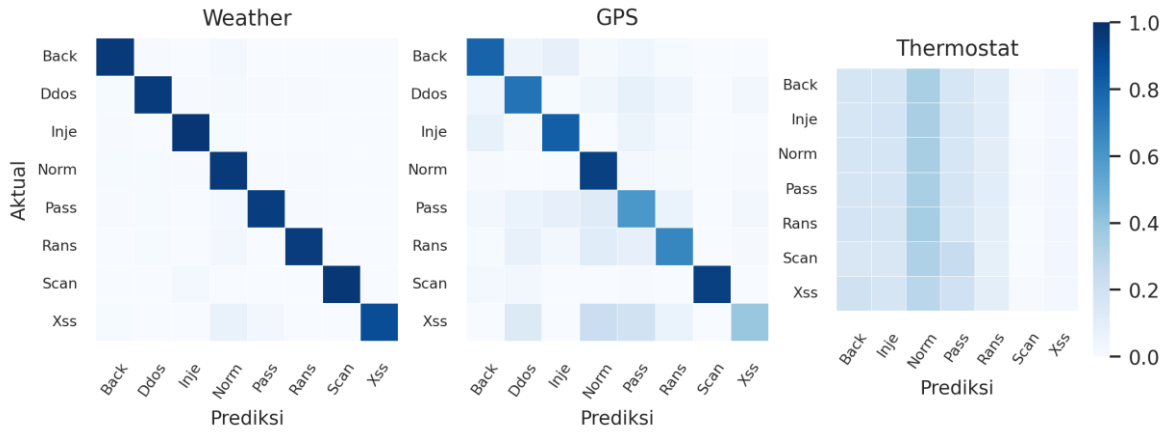
Pemisahan waktu global murni juga tidak layak untuk tugas ini. Pada Weather, 60% awal hanya memuat normal, scanning, injection, dan sebagian DDoS, sedangkan 20% akhir hanya backdoor dan ransomware. GPS menunjukkan pola serupa. Struktur satu episode per kelas harus dinyatakan sebagai keterbatasan dataset; penelitian lanjutan memerlukan beberapa episode serangan pada waktu, lokasi, dan perangkat berbeda.



Gambar 3. Perbandingan Validasi Antarkelompok Dan Blok Kronologis Multikelas.

4.4. Pola kesalahan antarkelas

Confusion matrix out-of-fold pada Gambar 4 menegaskan perbedaan antarperangkat. Weather menunjukkan diagonal dominan untuk delapan kelas. GPS mempertahankan pola diagonal, tetapi lebih banyak pertukaran antartipe serangan. Thermostat tidak membentuk diagonal yang stabil dan banyak kelas dialihkan ke normal atau kelas dominan. Visual ini konsisten dengan macro-F1, sekaligus mencegah satu skor agregat menutupi kelas yang gagal.



Gambar 4. Confusion Matrix Out-Of-Fold Ternormalisasi: Weather, GPS, dan Thermostat.

4.5. Accuracy dapat menyesatkan

Pada Fridge, 1.057 baris hasil deduplikasi terbentuk dari 151 tanda tangan prediktor, masing-masing muncul sekali untuk setiap tujuh label multikelas. Dengan demikian, input fridge_temperature–temp_condition yang identik berpasangan dengan normal dan enam tipe serangan. Klasifikator deterministik yang hanya memakai kedua fitur memberikan satu prediksi untuk tujuh target berbeda pada setiap tanda tangan; paling banyak satu dari tujuh baris dapat benar. Hal ini menjelaskan accuracy dan balanced accuracy multikelas sebesar 1/7 atau 14,29% pada split yang mempertahankan seluruh anggota grup, meskipun macro-F1 dapat berbeda menurut distribusi prediksi. Pada tugas biner, enam label serangan digabung menjadi Attack sehingga proporsinya 6/7. Prediksi Attack untuk seluruh baris menghasilkan accuracy 6/7 atau 85,71%, balanced accuracy 50,00%, dan macro-F1 46,15%, sesuai model terpilih serta dummy most-frequent. Konflik yang dipertahankan selama deduplikasi bukan kesalahan klasifikasi yang dapat dihilangkan melalui pergantian algoritma saja: label tidak teridentifikasi secara unik dari fitur yang tersedia. Kesimpulan ini berlaku pada representasi hasil deduplikasi, yang mengubah bobot frekuensi, dan bukan estimasi prevalensi serangan di lapangan. Temuan ini juga tidak membuktikan sensor rusak atau label salah; fitur tambahan dan konteks waktu perlu diuji pada data independen.

4.6. Missing, noise, dan drift

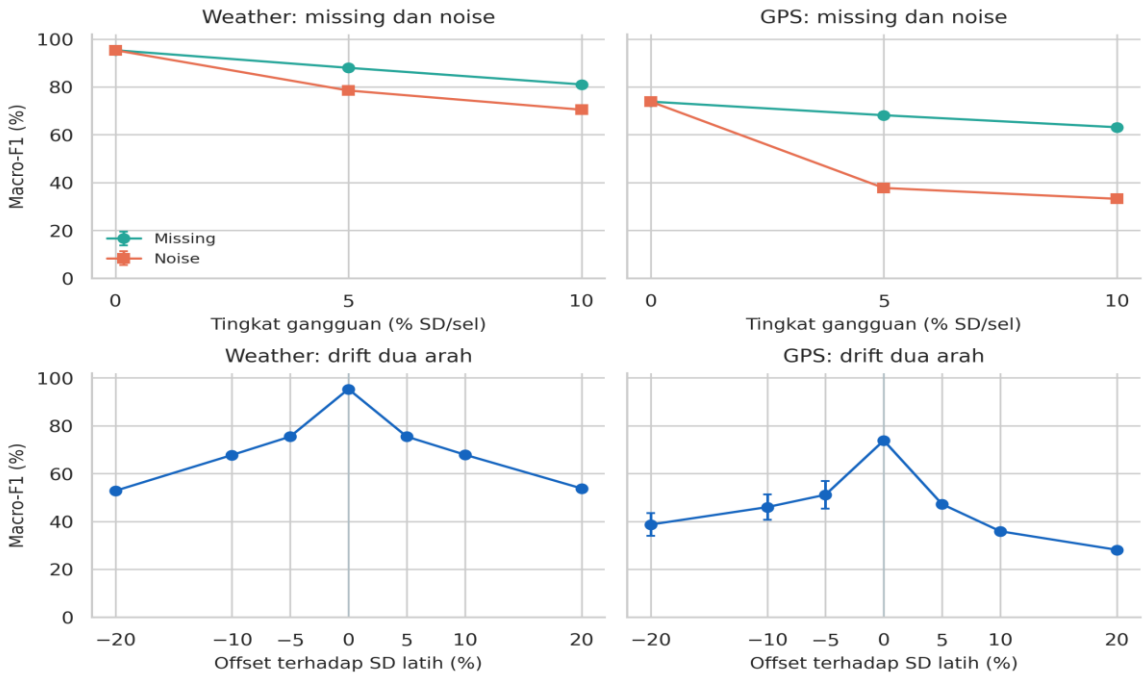
Tabel 5 dan Gambar 5 memperlihatkan degradasi yang konsisten pada dua perangkat berkinerja tinggi. Missing 5% menurunkan macro-F1 Weather menjadi 87,95±0,42% dan GPS menjadi 68,15±1,20%. Noise 5% lebih berat: Weather 78,47±0,45% dan GPS 37,75±0,93%. Drift ±10% menghasilkan Weather 67,75–67,89% dan GPS 35,90–46,00%. Respons dua arah tidak simetris, terutama pada GPS, sehingga satu arah offset saja tidak cukup untuk menggambarkan risiko.

Skala 5–20% SD dipakai untuk membangun kurva sensitivitas yang dapat dibandingkan antarfitur, bukan sebagai batas kalibrasi sensor. Kebutuhan operasional harus mengganti persentase ini dengan spesifikasi noise, drift, rentang, dan kondisi lingkungan perangkat sebenarnya. Meski demikian, hasil saat ini cukup untuk menolak asumsi bahwa skor clean otomatis stabil terhadap integritas pengukuran.

TABEL V. MACRO-F1 PERTURBASI MULTIKELAS, RERATA ± SD (%)

Perangkat	Clean	Miss. 5	Miss. 10	Noise 5	Noise 10	Drift −10 / +10
Weather	95,32 ± 0,31	87,95 ± 0,42	81,02 ± 0,56	78,47 ± 0,45	70,46 ± 0,46	67,75 ± 0,87 / 67,89 ± 0,13
GPS	73,84 ± 1,19	68,15 ± 1,20	63,13 ± 1,28	37,75 ± 0,93	33,22 ± 1,01	46,00 ± 5,27 / 35,90 ± 0,50

Missing dan *noise*: 10 realisasi per *fold*; *drift*: dua arah pada lima *fold*. Fitur numerik dibatasi pada rentang latih.

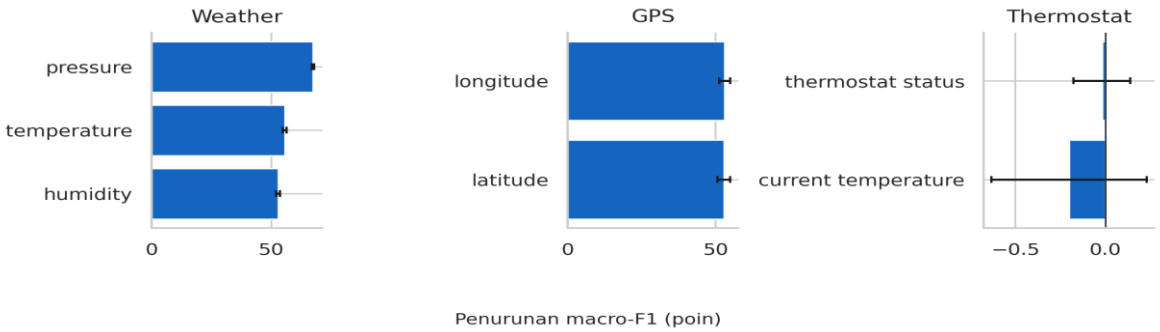


Gambar 5. Kurva Sensitivitas Missing, Noise, Dan Drift Dua Arah.

4.7. XAI sebagai audit ketergantungan fitur

Permutation importance pada Gambar 6 menunjukkan pressure sebagai fitur Weather paling dominan (penurunan macro-F1 67,57±0,56 poin antar-fold), diikuti longitude dan latitude pada GPS yang hampir seimbang (53,09±1,89 dan 52,86±2,24 poin). Pada Thermostat, kedua fitur bernilai sekitar nol atau negatif, bersamaan dengan kinerja setara baseline. Nilai tersebut menunjukkan bahwa pengacakan tidak memperburuk skor secara konsisten pada model ini, bukan membuktikan bahwa fitur secara intrinsik tidak mengandung informasi atau tidak berhubungan dengan serangan. Karena Modbus dan Fridge tidak dianalisis dengan metode ini, temuan XAI dibatasi pada tiga perangkat yang ditampilkan.

Importance Weather yang tinggi pada seluruh fitur tidak boleh dijumlahkan sebagai kontribusi kausal karena temperature, pressure, dan humidity dapat berkorelasi. Pengacakan satu fitur memutus struktur bersama dan dapat menghasilkan penurunan yang tumpang tindih [17]. Pelaporan sepuluh pengacakan per fold dan error bar membuat ketidakpastian penjelasan eksplisit, tetapi belum menggantikan eksperimen kausal.



Gambar 6. Permutation Importance Multikelas, Rerata ± SD Antar-Fold (10 Pengacakan/Fold).

4.8. Implikasi dan perbandingan

Dibanding benchmark telemetri ToN_IoT [1]. hasil ini memperlihatkan bahwa skor klasifikasi perlu dibaca bersama pemisahan tanda tangan prediktor dan urutan waktu. Temuan ini sejalan dengan peringatan Bouke dan Abdullah mengenai inflasi evaluasi akibat pattern leakage, tetapi menambahkan bukti konflik label pada Fridge serta penurunan Weather dan GPS pada uji kronologis dan gangguan. Studi fitur universal berbasis SHAP[8],[14]mendukung penggunaan penjelasan model; penelitian ini menempatkan penjelasan tersebut setelah pemeriksaan baseline dan kestabilan. Kebaruan integrasinya bersifat

metodologis dan empiris: protokol yang sama membedakan perangkat yang diskriminatif dalam distribusi dari perangkat berskor baseline, sekaligus menunjukkan bahwa keduanya memerlukan batas klaim yang berbeda. Skor antarpublikasi tidak dibandingkan langsung karena fitur, kelas, dan skema pemisahannya berbeda

Implikasi bagi rantai pengukuran-keputusan adalah perlunya menilai mutu fitur dan rancangan evaluasi sebelum memilih model untuk keamanan. Weather dan GPS diskriminatif pada validasi antarkelompok, tetapi kinerjanya menurun pada bagian akhir episode dan saat diberi gangguan. Pada Thermostat, Modbus, dan Fridge, pengayaan modalitas seperti arus, getaran, status aktuator, atau trafik jaringan layak diuji, tanpa menganggap manfaatnya telah terbukti. Konflik label Fridge secara khusus menuntut pemeriksaan konteks pengukuran dan anotasi; optimasi algoritma saja tidak dapat membuat input identik membedakan target yang berlainan.

4.9. Keterbatasan

Dataset merupakan benchmark sekunder offline dengan satu episode waktu per kelas dan tanpa grup perangkat fisik independen. Validasi kronologis membandingkan bagian awal dan akhir episode yang sama menggunakan arsitektur hasil validasi grup; bagian validation temporal tidak dimanfaatkan untuk pemilihan model. Hasilnya tidak membuktikan generalisasi lintas episode, lokasi, perangkat, atau firmware. Deduplikasi mengubah bobot frekuensi observasi, sehingga distribusi kelas hasil audit tidak merepresentasikan prevalensi operasional. Perturbasi dinormalisasi terhadap SD latih dan belum didasarkan pada sertifikat kalibrasi atau spesifikasi sensor. Permutation importance hanya dihitung pada tiga perangkat, bersifat prediktif, dan tidak membuktikan sebab-akibat atau ketiadaan informasi pada fitur berskor rendah. Penelitian juga tidak mengukur latensi streaming, respons terhadap serangan adaptif, atau probabilitas kegagalan perangkat keras.

5. KESIMPULAN

Reliabilitas keputusan klasifikasi berbeda antarperangkat dan bergantung pada protokol evaluasi. Setelah audit leakage dan pemisahan tanda tangan prediktor, macro-F1 multikelas Weather dan GPS mencapai 95,32±0,31% dan 73,84±1,19%, melampaui baseline dummy. Thermostat, Modbus, dan Fridge hanya mencapai 14,07±0,19%, 15,03±0,60%, dan 13,14±0,67%, tanpa peningkatan yang sebanding atas baseline. Seluruh 151 tanda tangan prediktor Fridge memiliki konflik tujuh label, sehingga fitur yang tersedia tidak mengidentifikasi kelas secara unik; accuracy biner 85,71% juga hanya disertai balanced accuracy 50,00%. Keunggulan antarkelompok Weather dan GPS tidak bertahan pada uji kronologis, dengan macro-F1 turun menjadi 23,35% dan 8,05%, serta menurun pada missing, noise, dan drift. Dengan demikian, hasil mendukung keterpisahan dalam distribusi pada dua perangkat, tetapi belum mendukung generalisasi temporal yang stabil atau penerapan pada perangkat baru. Audit leakage, baseline, validasi grup/waktu, uji gangguan, dan XAI perlu. dibaca bersama untuk membatasi klaim; hasil penelitian ini tidak mengukur reliabilitas perangkat keras, akurasi sensor, atau kinerja deteksi waktu nyata.

PERNYATAAN ETIK DAN KONFLIK KEPENTINGAN

Penelitian menggunakan benchmark publik sekunder tanpa intervensi manusia atau hewan dan tidak memiliki konflik kepentingan. Kode analisis dan manuskrip dikembangkan secara mandiri. Selain ToN_IoT, tidak digunakan data rahasia, kode proprietari, atau komputasi eksternal. Dataset tetap diatribusikan sesuai[1] [2] [3].

KETERSEDIAAN DATA DAN KODE

Dataset ToN_IoT tersedia melalui portal resmi UNSW[1][2][3]. Sumber data, fitur, prapemrosesan, parameter, random seed, validasi, dan hasil agregat dijelaskan dalam artikel. Kode sumber dan aplikasi merupakan kekayaan intelektual penulis sehingga tidak didistribusikan secara terbuka. Bukti numerik dapat disampaikan secara privat melalui kanal editorial jika diperlukan untuk verifikasi akademik.

DAFTAR PUSTAKA

[1] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and Adna N Anwar, “TON-IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems,” *IEEE Access*, vol. 8, pp. 165130–165150, 2020, doi: 10.1109/ACCESS.2020.3022862.

[2] N. Moustafa, “A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets,” *Sustain. Cities Soc.*, vol. 72, Sep. 2021, doi: 10.1016/j.scs.2021.102994.

[3] T. M. Booij, I. Chiscop, E. Meeuwissen, N. Moustafa, and F. T. H. D. Hartog, “ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack

Types in IoT Network Intrusion Data Sets," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 485–496, Jan. 2022, doi: 10.1109/JIOT.2021.3085194.

[4] A. Wicahyanto, "PENERAPAN ARTIFICIAL NEURAL NETWORK DALAM DETEKSI SERANGAN PADA WEB SERVER APACHE," 2025. [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jire ISSN.2620-6900

[5] A. A. Nugroho and M. Haris, "ANALISIS EFEKTIVITAS TEKNIK IMPUTASI PADA LSTM UNTUK MENINGKATKAN KUALITAS DATA PADA PERAMALAN CURAH HUJAN," 2024. [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jire ISSN.2620-6900

[6] T. T. H. Le, H. Kim, H. Kang, and H. Kim, "Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method," *Sensors*, vol. 22, no. 3, Feb. 2022, doi: 10.3390/s22031154.

[7] A. Namrita Gummadi, J. C. Napier, and M. Abdallah, "XAI-IoT: An Explainable AI Framework for Enhancing Anomaly Detection in IoT Systems," *IEEE Access*, vol. 12, pp. 71024–71054, 2024, doi: 10.1109/ACCESS.2024.3402446.

[8] A. Nascita, G. Aceto, D. Ciuonzo, A. Montieri, V. Persico, and A. Pescape, "A Survey on Explainable Artificial Intelligence for Internet Traffic Classification and Prediction, and Intrusion Detection," 2025, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/COMST.2024.3504955.

[9] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[10] M. A. Bouke and A. Abdullah, "An empirical study of pattern leakage impact during data preprocessing on machine learning-based intrusion detection models reliability," *Expert Syst. Appl.*, vol. 230, Nov. 2023, doi: 10.1016/j.eswa.2023.120715.

[11] M. Amin, F. Al-Obeidat, A. Tubaishat, B. Shah, S. Anwar, and T. A. Tanveer, "Cyber security and beyond: Detecting malware and concept drift in AI-based sensor data streams using statistical techniques," *Computers and Electrical Engineering*, vol. 108, May 2023, doi: 10.1016/j.compeleceng.2023.108702.

[12] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[13] M. M. Alani and A. Miri, "Towards an Explainable Universal Feature Set for IoT Intrusion Detection," *Sensors*, vol. 22, no. 15, Aug. 2022, doi: 10.3390/s22155690.

[14] P. Khani, E. Moeinaddini, N. D. Abnavi, and A. Shahraki, "Explainable artificial intelligence for feature selection in network traffic classification: A comparative study," *Transactions on Emerging Telecommunications Technologies*, vol. 35, no. 4, Apr. 2024, doi: 10.1002/ett.4970.

[15] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," 2019. [Online]. Available: http://jmlr.org/papers/v20/18-760.html.

[16] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," 2010. [Online]. Available: http://www.biomedcentral.com/1471-2105/11/110

[17] C. Molnar, G. König, B. Bischl, and G. Casalicchio, "Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach," *Data Min. Knowl. Discov.*, vol. 38, no. 5, pp. 2903–2941, Sep. 2024, doi: 10.1007/s10618-022-00901-9.