JURNAL ILMIAH SISTEM INFORMASI (JUSI)
VOL 4, No. 2, Mei 2025, pp. 110 - 129
p-ISSN : 2803-1507 (media cetak) | e-ISSN : 2803-1531 (media online)
Link Publikasi: https://ejurnal.provisi.ac.id/index.php/JUISI/issue/view/134

Early Detection of Data Theft on Mobile Devices Using Machine Learning

Neilin Nikhlis*[1], Muhammad Jamal Udin Ghofur[2]
[1]Universitas Sains dan Teknologi Komputer
Jl. Majapahit No. 605, 50192, Kota Semarang, Jawa Tengah, Indonesia
[2] STMIK HIMSYA
Kota Semarang, Jawa Tengah, Indonesia
e-mail*[1]: neilin@stekom.ac.id

.

ARTICLE INFO

Article History:

Received: 21 Maret 2025
Revised: 28 April 2025
Accepted: 30 April 2025
Published: 21 Mei 2025

ABSTRACT

The increasing use of mobile devices has increased the risk of data theft, posing significant security challenges for individuals and organizations. This study proposes an early detection system for data theft on mobile devices using machine learning algorithms. The system is designed to identify suspicious patterns in application usage, network access, and CPU/memory activity, providing early warnings to prevent potential data loss. By employing algorithms such as Random Forest, Support Vector Machine (SVM), and Convolutional Neural Network (CNN), the developed models demonstrated significant performance: CNN achieved the highest accuracy of 95.1%, with a precision of 94.2%, recall of 93.5% , F1-score of 93.8%, and AUC-ROC of 0.96. Random Forest and SVM also showed competitive performance with accuracy rates of 94.7% and 92.5%, respectively. These findings highlight the high potential of machine learning algorithms for real-time detection of data theft threats, providing adaptive protection against evolving cyberattack methods. This approach offers a promising solution to strengthen mobile device security frameworks and safeguard user data against increasingly sophisticated cyber threats.

Keywords: *Mobile Security, Data Theft Detection, Machine Learning, Proactive Defense, Anomaly Detection.*

Abstrak

Peningkatan penggunaan perangkat seluler membawa risiko pencurian data yang semakin tinggi, menimbulkan tantangan signifikan dalam keamanan siber. Penelitian ini mengembangkan sistem deteksi dini pencurian data berbasis machine learning untuk mengenali pola aktivitas mencurigakan seperti akses jaringan abnormal, penggunaan CPU/memori yang tidak wajar, dan perilaku aplikasi. Tiga algoritma utama—Convolutional Neural Network (CNN), Random Forest, dan Support Vector Machine (SVM)—digunakan untuk membangun model prediksi. Hasil evaluasi menunjukkan bahwa CNN memberikan akurasi tertinggi sebesar 95,1%, dengan nilai F1-score 93,8% dan AUC-ROC 0,96, sementara Random Forest dan SVM juga menunjukkan performa kompetitif. Pengujian dilakukan dengan dataset lalu lintas aplikasi dan log aktivitas perangkat seluler, menggunakan teknik preprocessing, balancing data, dan validasi silang. Temuan menunjukkan bahwa pendekatan ini efektif dalam mendeteksi ancaman siber secara real-time, memberikan

DOI: https://doi.org/10.51903/fz0xzt26

p-ISSN: 2803-1507 e-ISSN: 2803-1531

perlindungan proaktif terhadap pencurian data. Penelitian ini berkontribusi dalam memperkuat keamanan perangkat seluler melalui deteksi anomali berbasis AI yang adaptif terhadap taktik serangan yang terus berkembang.
Kata Kunci: *Mobile Security, Data Theft Detection, Machine Learning, Proactive Defense, Anomaly Detection*

1. PENDAHULUAN

Perangkat seluler telah menjadi bagian integral dari kehidupan modern dengan jumlah pengguna yang terus meningkat. Berdasarkan laporan Statista pada tahun 2023, perangkat seluler aktif di seluruh dunia yang mencerminkan tingkat penetrasi yang sangat tinggi telah mencapai lebih dari 7,33 miliar, dengan 91,40% populasi global menggunakan *smartphone* [1]. Penggunaan perangkat ini tidak terbatas pada komunikasi saja, tetapi mencakup transaksi finansial, penyimpanan data pribadi, dan akses ke informasi sensitif. Namun, perkembangan ini juga membawa risiko keamanan yang besar, khususnya terkait pencurian data. Seiring dengan bertambahnya jumlah pengguna perangkat seluler, ancaman terhadap privasi dan keamanan data semakin meningkat. Menurut [2] pencurian data dapat terjadi melalui berbagai serangan siber, termasuk *malware*, *phishing*, dan serangan *man-in-the-middle* (MITM), yang semuanya semakin canggih dari waktu ke waktu.

Pencurian data pada perangkat seluler telah menjadi salah satu isu utama dalam keamanan siber. Menurut [3] Verizon Data *Breach Investigations* pada tahun 2023, menunjukkan bahwa serangan terhadap perangkat seluler meningkat lebih dari 40% dalam dua tahun terakhir. Selain itu, data dari *Check Point's Mobile Security Report* pada tahun 2023, mengungkapkan bahwa serangan *malware* terhadap perangkat Android meningkat hingga mencapai 5,3 miliar, dengan 71,63% dibandingkan tahun sebelumnya [4]. Ancaman pencurian data pada perangkat seluler tidak hanya berdampak pada individu, tetapi juga pada organisasi besar, dengan perusahaan global melaporkan adanya upaya pencurian data melalui perangkat tersebut [5]. Menurut penelitian [6], mengembangkan model deteksi dini pencurian data pada perangkat seluler dengan pendekatan proaktif, berbeda dari fokus deteksi *malware* sebelumnya. Algoritma *machine learning* seperti *Random Forest*, SVM, dan *Deep Learning* diharapkan dapat meningkatkan akurasi deteksi aktivitas mencurigakan secara *real-time* dan beradaptasi dengan ancaman baru. Hal ini menunjukkan kerentanan perangkat seluler terhadap serangan siber dan perlunya pengembangan sistem keamanan yang lebih canggih untuk melindungi data pengguna.

Meskipun banyak upaya untuk meningkatkan keamanan perangkat seluler, sebagian besar solusi masih reaktif, berfungsi setelah pencurian data terjadi [7]. Sistem tradisional, seperti autentikasi dua faktor dan enkripsi, sering gagal menghadapi serangan canggih. Menurut (Sun et al., 2023 & Adeniran et al., 2024) diperlukan teknologi deteksi ancaman yang lebih proaktif untuk menghadapi risiko pencurian data yang meningkat, karena metode serangan siber yang canggih menjadikan pendekatan reaktif tidak memadai. Pendekatan proaktif harus fokus pada identifikasi dan mitigasi ancaman sebelum menyebabkan kerugian. Menurut penelitian [10], telah menjelajahi penggunaan *machine learning* untuk deteksi *malware*, namun belum diterapkan secara spesifik untuk deteksi dini pencurian data pada perangkat seluler.

Selain itu, model deteksi yang ada sering tidak adaptif terhadap perubahan taktik serangan, menjadi tantangan dalam keamanan [11]. Kemajuan teknologi memungkinkan penyerang mengubah strategi, sehingga model deteksi tradisional berbasis pola menjadi kurang efektif. Penelitian ini bertujuan untuk mengembangkan sistem deteksi dini pencurian data pada perangkat seluler menggunakan algoritma *machine learning*. Menurut penelitian [12], sistem yang diusulkan akan mampu mengenali pola aktivitas mencurigakan sebelum pencurian data terjadi, memberikan peringatan dini, dan membantu mencegah kerugian yang lebih besar. Dengan memanfaatkan teknologi *machine learning*, model ini diharapkan dapat mendeteksi pola perilaku yang tidak normal, baik dari penggunaan aplikasi, akses jaringan, maupun aktivitas CPU dan memori, sehingga memungkinkan tindakan *preventif* sebelum data pengguna terekspos [13].

2. TINJAUAN PUSTAKA

2.1. Ancaman Pencurian Data dan Urgensi Deteksi Dini Seluler

Pencurian data pada perangkat seluler telah berkembang menjadi salah satu ancaman keamanan siber paling kritis di era digital, terutama karena ketergantungan yang semakin tinggi pada perangkat ini dalam kehidupan sehari-hari. Menurut [14] *smartphone* dan tablet, yang saat ini digunakan untuk mengakses berbagai layanan keuangan, akun media sosial, layanan kesehatan, serta penyimpanan data pribadi, menjadi sasaran utama bagi

p-ISSN: 2803-1507 e-ISSN: 2803-1531

pelaku kejahatan siber. Pada tahun 2023, laporan Verizon menemukan bahwa pelanggaran data terjadi melalui perangkat seluler, dan lebih dari setengah dari pelanggaran ini melibatkan pencurian informasi pribadi, kredensial akun, serta data finansial (Ahmed et al., 2024, & Srinivas & Liang, 2022). Data yang dicuri dapat dieksploitasi dalam berbagai tindakan kriminal, seperti pencurian identitas, penipuan finansial, hingga menjadi basis serangan lanjutan terhadap perangkat atau sistem lain yang terhubung (Zukarnain, 2021 & Piquero et al., 2022). Menururt laporan Symantec pada tahun 2023, potensi kerugian akibat pencurian data pada perangkat seluler diperkirakan mencapai miliaran dolar, dengan dampak yang semakin luas akibat meningkatnya nilai informasi yang disimpan dalam perangkat [19]. Ancaman ini diperparah dengan pesatnya pertumbuhan aplikasi berbahaya yang sering kali menyamar sebagai aplikasi aman di toko aplikasi atau dipasang melalui pihak ketiga yang tidak terverifikasi. *Malware* yang tertanam dalam aplikasi semacam ini memungkinkan peretas mengakses data sensitif tanpa disadari oleh pengguna. Menurut (Sikder et al., 2021 & Liu et al., 2023) aplikasi berbahaya tersebut sering kali dapat mengakses informasi sensitif, mencuri kredensial, dan bahkan mengendalikan perangkat pengguna tanpa izin.

Tren ini juga tercermin dalam laporan terbaru dari ESET yang mencatat adanya peningkatan 50% dalam kasus infeksi malware pada perangkat seluler di tahun 2022, dengan prediksi tren ini akan berlanjut karena distribusi aplikasi berbahaya yang semakin canggih dan taktis [22]. Faktor lain yang meningkatkan risiko adalah pertumbuhan ekosistem *Internet of Things* (IoT), yang memperluas permukaan serangan melalui peningkatan jumlah perangkat yang terhubung secara global. Cisco memprediksi jumlah perangkat IoT yang terhubung akan mencapai 125 miliar pada tahun 2030, meningkatkan potensi celah keamanan untuk perangkat dengan perlindungan minimal yang lebih mudah dieksploitasi oleh penyerang [23]. Menurut (Solera-Cotanilla et al., 2022, Ioannidou & Sklavos, 2021) setiap perangkat yang terhubung, seperti kamera pintar, jam tangan pintar, atau alat rumah tangga, memberikan kesempatan bagi pelaku kejahatan untuk mengakses dan mengubah data melalui perangkat seluler pengguna yang memiliki kerentanan. Terlebih lagi, perangkat seluler cenderung memiliki keterbatasan daya dan sumber daya komputasi, sehingga sulit menerapkan langkah-langkah keamanan yang memadai seperti enkripsi tingkat lanjut atau *firewall* yang kuat, berbeda dengan perangkat komputer yang lebih kuat [26]. Keterbatasan ini menjadikan perangkat seluler target empuk bagi serangan siber dan meningkatkan risiko kerentanan data pengguna yang sering kali tersimpan dalam sistem perangkat tersebut.

Di tengah kompleksitas ancaman ini, deteksi dini berbasis *machine learning* telah menjadi pendekatan yang semakin signifikan. *Machine learning* memungkinkan sistem keamanan mengenali pola-pola yang mencurigakan dalam data pengguna dan memprediksi potensi ancaman dengan akurasi yang jauh lebih tinggi dibandingkan metode deteksi tradisional yang bersifat reaktif. Menurut [27] model *machine learning* seperti *Random Forest*, *Support Vector Machine* (SVM), dan deep learning telah terbukti efektif dalam mendeteksi malware serta ancaman pada perangkat seluler, dengan tingkat akurasi mencapai 96%, menurut laporan terbaru dari *National Institute of Standards and Technology* (NIST). Selain itu, pendekatan *machine learning* memungkinkan sistem keamanan untuk beradaptasi dengan karakteristik unik dari setiap perangkat dan aplikasi yang berbeda, sehingga dapat lebih efisien dalam mengenali *anomali* dengan tingkat *false positive* yang rendah [28]. Algoritma *machine learning* bahkan dapat disesuaikan agar tetap efisien pada perangkat dengan sumber daya terbatas seperti ponsel, sehingga memungkinkan deteksi ancaman tanpa perlu mengorbankan performa perangkat secara signifikan.

Pendekatan ini semakin diperlukan karena volume data yang dihasilkan oleh perangkat seluler juga terus meningkat, yang menciptakan kebutuhan akan metode deteksi ancaman yang bersifat *real-time* dan skalabel. Penggunaan *machine learning* juga memungkinkan perangkat untuk mempelajari pola serangan baru seiring waktu, memberikan perlindungan yang lebih adaptif terhadap berbagai jenis *malware* dan taktik serangan yang terus berevolusi. Menurut [29] *machine learning* tidak hanya memungkinkan deteksi yang lebih tepat, tetapi juga menawarkan solusi yang berkelanjutan untuk mengatasi perkembangan ancaman keamanan siber pada perangkat seluler yang terus berubah di era digital saat ini.

2.2. Deteksi Dini Keamanan Seluler dengan Machine Learning

Deteksi dini merupakan elemen fundamental dalam keamanan perangkat seluler, bertujuan untuk mengidentifikasi aktivitas atau pola yang menunjukkan adanya ancaman atau *anomali* sedini mungkin sehingga tindakan mitigasi dapat segera dilakukan guna mencegah kerugian lebih lanjut [30]. Hal ini semakin relevan mengingat semakin tingginya ketergantungan masyarakat modern terhadap perangkat seluler untuk keperluan komunikasi, perbankan, transaksi digital, dan penyimpanan data pribadi. Peningkatan penggunaan

p-ISSN: 2803-1507 e-ISSN: 2803-1531

perangkat seluler berbanding lurus dengan risiko kebocoran data pribadi, pencurian identitas, serta serangan *malware*, yang dapat berakibat pada kerugian finansial dan privasi yang signifikan. Pendekatan berbasis *machine learning* (ML) hadir sebagai solusi yang efektif dalam mempelajari pola perilaku dari data historis, mengenali penyimpangan yang signifikan, serta mendeteksi aktivitas mencurigakan atau bahkan potensi pencurian data.

Algoritma *machine learning* seperti *Random Forest, Support Vector Machine* (SVM), dan metode berbasis *Deep Learning* seperti *Recurrent Neural Networks* (RNN) serta *Convolutional Neural Networks* (CNN) semakin banyak digunakan karena keunggulannya dalam mengklasifikasi dan mendeteksi anomal [31]. Algoritma *Random Forest*, misalnya, mampu menangani data dengan banyak variabel serta hubungan kompleks antar fitur. SVM unggul dalam memisahkan data ke dalam kelas yang berbeda dengan memanfaatkan *hyperplane* optimal, memungkinkan deteksi *anomali* yang lebih presisi. Di sisi lain, CNN dan RNN, yang menjadi pilihan populer dalam *Deep Learning*, mampu menangkap pola temporal dan spasial dalam data, menjadikannya sangat ideal untuk mendeteksi perubahan perilaku perangkat seluler yang terjadi secara dinamis (Yu et al., 2022).

Selain itu, konsep deteksi *anomali* memainkan peran kunci dalam mendeteksi aktivitas pencurian data, dengan pendekatan yang memanfaatkan pola perilaku pengguna sebagai acuan untuk mengidentifikasi penyimpangan berbahaya. Penyimpangan atau *anomali* ini umumnya berupa perilaku yang jarang terjadi dan sangat berbeda dari pola normal, seperti akses data yang tiba-tiba intensif atau penggunaan perangkat di luar jam biasa. Dengan melatih model pada data historis, algoritma dapat menetapkan *baseline* yang akurat untuk aktivitas normal perangkat, memungkinkan pendeteksian aktivitas yang mencurigakan dengan lebih tepat. Menurut (Nassif et al., 2021 & Devineni et al., 2023) menunjukkan bahwa model berbasis deteksi *anomali* ini mampu mengidentifikasi pola serangan tak biasa yang umumnya sulit terdeteksi dengan metode tradisional, termasuk serangan *zero-day* atau jenis ancaman yang belum dikenali sebelumnya.

2.3. Deteksi Ancaman Mobile dengan Machine Learning

Penelitian di bidang deteksi dini ancaman siber pada perangkat seluler telah menunjukkan perkembangan yang signifikan, terutama dalam menghadapi peningkatan risiko serangan yang semakin kompleks dan ditargetkan. Beberapa studi menunjukkan keberhasilan penerapan *machine learning* untuk mendeteksi aktivitas mencurigakan dengan lebih akurat dan responsif. Menurut (Abubaker et al., 2024 & Almomani et al., 2024), memanfaatkan dataset perangkat Android yang terinfeksi *malware*, menemukan bahwa fitur perilaku aplikasi—seperti pola akses data, perubahan izin, dan interaksi pengguna—dapat menjadi indikator utama dalam deteksi dini aktivitas tidak sah. Studi ini menggarisbawahi pentingnya pendekatan analisis perilaku untuk mendeteksi *anomali* secara efektif. Model seperti Recurrent Neural Network (RNN) dinilai sangat bermanfaat dalam konteks ini karena mampu mengenali pola data secara *real-time*, yang sangat krusial mengingat serangan siber yang semakin kompleks dan adaptif [37].

Di sisi lain, penelitian yang mengadopsi teknik *Hybrid Learning* semakin menonjol, terutama dalam meningkatkan akurasi dan efisiensi deteksi ancaman. Teknik ini menggabungkan algoritma seperti *Random Forest* dan *Deep Learning*, memungkinkan pendekatan yang lebih holistik dan komprehensif dalam proses identifikasi ancaman [38]. Tren ini juga didukung oleh laporan IDC pada tahun 2023, yang menunjukkan peningkatan signifikan dalam jumlah serangan terhadap perangkat seluler di seluruh dunia [39]. Menurut [40] laporan dari IDC pada tahun 2023 juga mengonfirmasi tren ini, dengan mencatat adanya peningkatan tajam dalam serangan yang menargetkan perangkat *mobile* secara global. IDC mengungkapkan bahwa banyak penyedia keamanan kini *mengintegrasikan artificial intelligence* (AI) dan *machine learning* ke dalam solusi mereka untuk keamanan perangkat *mobile*, memungkinkan deteksi dan respons yang lebih cepat serta proaktif terhadap serangan yang kompleks [41]. Selain itu, penelitian mencatat bahwa pendekatan berbasis AI ini tidak hanya mampu memantau pola yang rumit tetapi juga melakukan analisis prediktif, yang berfungsi mencegah potensi serangan sebelum terjadi. Dengan meningkatnya variasi dan intensitas serangan siber, penerapan teknologi seperti AI dan *machine learning* dalam deteksi dini pada perangkat seluler menjadi strategi yang sangat penting bagi keamanan masa depan.

3. METODOLOGI PENELITIAN

3.1. Sumber Data Penelitian

Data yang digunakan dalam penelitian ini bersumber dari log aktivitas dan lalu lintas jaringan pada perangkat seluler, yang mencakup aktivitas penggunaan aplikasi, aktivitas jaringan, dan log keamanan. Data ini dapat

p-ISSN: 2803-1507 e-ISSN: 2803-1531

dikumpulkan dari perangkat seluler yang dioperasikan dalam lingkungan terkendali (controlled environment) untuk mengidentifikasi pola penggunaan yang mencurigakan dan perilaku anomali terkait pencurian data. Penelitian mendukung penggunaan data ini untuk deteksi dini anomali pada perangkat seluler [42]. Metrik penggunaan aplikasi mencakup jenis aplikasi, waktu, dan pola penggunaan yang dapat mengindikasikan perilaku anomali [43], sementara volume data konsumsi per aplikasi membantu mengidentifikasi penggunaan data yang berlebihan, yang sering kali menunjukkan pencurian atau pengumpulan data tanpa izin [44]. Durasi koneksi ke server eksternal juga penting untuk mendeteksi aktivitas abnormal, terutama jika aplikasi terus mengirim data tanpa interaksi pengguna, yang sering terkait dengan malware atau aktivitas bot (Vasani et al., 2023 & Yadav et al., 2022). Selain itu, log aktivitas abnormal seperti percobaan akses data terlindungi atau perubahan pengaturan tanpa izin dapat membantu mengidentifikasi potensi ancaman pencurian data [47].

Untuk memastikan hasil yang valid dan akurat, data yang digunakan harus mencakup ribuan hingga jutaan entri log dari perangkat seluler. Pengumpulan data dari berbagai perangkat dan pengguna sangat penting untuk memastikan generalisasi model terhadap pola penggunaan yang berbeda-beda di setiap perangkat dan pengguna. Hal ini juga memungkinkan model untuk mengenali perbedaan antara penggunaan normal dan anomali dalam berbagai konteks. Berdasarkan penelitian [48], data yang diperoleh dari ribuan pengguna pada rentang waktu tertentu lebih efektif dalam menangkap variasi pola dan aktivitas yang representatif. Pengumpulan data dapat dilakukan baik dari perangkat seluler di lingkungan terkendali maupun melalui kerjasama dengan penyedia layanan aplikasi keamanan, sehingga memungkinkan pengumpulan log perangkat secara aman dan anonim dari aplikasi-aplikasi yang memantau penggunaan perangkat. Setelah data terkumpul, perlu dilakukan validasi untuk memastikan bahwa data tersebut lengkap dan relevan dengan tujuan penelitian. Validasi ini melibatkan pengecekan konsistensi data, identifikasi dan penanganan anomali dalam data, serta pemisahan antara log aktivitas normal dan log aktivitas yang dianggap mencurigakan berdasarkan aturan bisnis atau logika sistem keamanan.

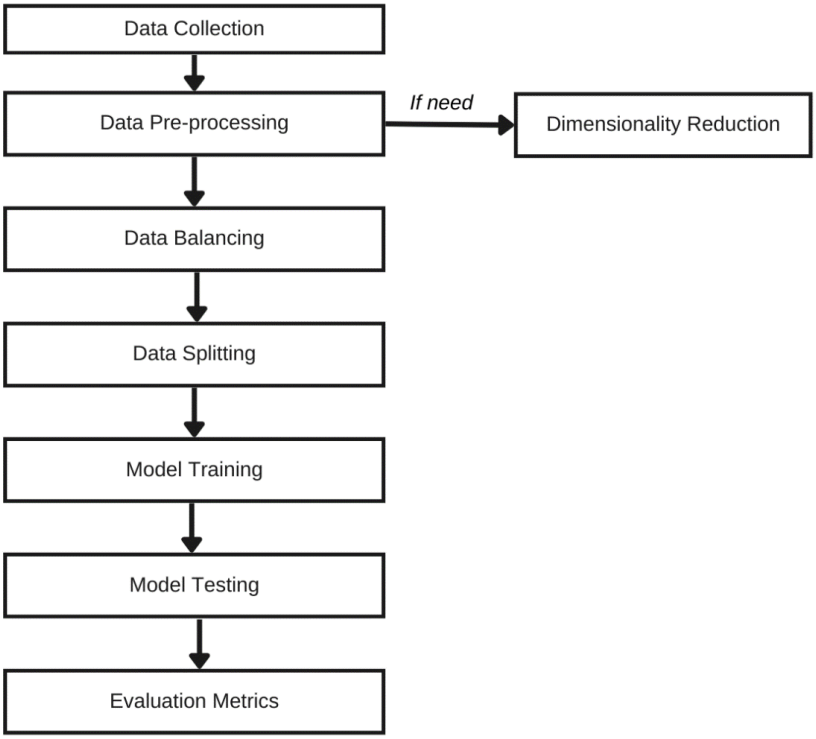
3.2. Teknik Penelitian

Dalam penelitian ini, beberapa algoritma machine learning dipilih untuk mendeteksi anomali pada data perangkat seluler, termasuk Random Forest, Support Vector Machine (SVM), dan Long Short-Term Memory (LSTM), yang efektif dalam mendeteksi pola anomali, terutama pada data berbasis waktu dan sekuensial [49]. Random Forest adalah algoritma ensemble berbasis pohon keputusan yang menggabungkan hasil dari banyak pohon untuk meningkatkan akurasi prediksi, mampu menangani data berisik dan mengurangi risiko overfitting, sehingga efektif untuk mendeteksi anomali pada dataset besar yang sering mengandung noise. Sementara itu, SVM bekerja dengan menemukan hyperplane yang memisahkan kelas-kelas data secara maksimal, memungkinkan deteksi perbedaan pola antara data normal dan mencurigakan, terutama pada data dengan distribusi terdefinisi baik, dan sering digunakan dalam kasus data tidak seimbang [50]. Di sisi lain, LSTM adalah jaringan saraf dalam yang dirancang khusus untuk menangani data sekuensial, sangat cocok untuk mendeteksi anomali pada log aktivitas perangkat seluler karena kemampuannya mengidentifikasi pola berulang dalam urutan data, serta mendeteksi perubahan perilaku penggunaan dari waktu ke waktu, menjadikannya relevan untuk analisis data log temporal.

Dalam konteks deteksi pencurian data pada perangkat seluler, jumlah data normal sering kali jauh lebih besar daripada data anomali atau pencurian, menyebabkan masalah ketidakseimbangan data. Untuk mengatasinya, beberapa teknik dapat diterapkan. Salah satunya adalah *Synthetic Minority Over-sampling Technique* (SMOTE), yang menghasilkan sampel sintetis dari data minoritas untuk menyeimbangkan jumlah antara kelas mayoritas (data normal) dan minoritas (data pencurian). SMOTE membantu model belajar dari sampel pencurian yang lebih representatif, mengurangi bias terhadap kelas mayoritas, dan meningkatkan kemampuan deteksi pencurian data (Gunturi & Sarkar, 2021). Selain itu, *Random Under-sampling* (RUS) mengurangi jumlah sampel pada kelas mayoritas untuk menangani ketidakseimbangan, meskipun ada risiko kehilangan informasi penting; RUS efektif ketika jumlah data mayoritas sangat besar, sehingga dapat mempercepat waktu komputasi dan memberikan perhatian lebih pada kelas minoritas. Metode lain adalah penyeimbangan dengan Algoritma Kostum (*Cost-sensitive Learning*), yang memberikan penalti lebih besar pada kesalahan deteksi di kelas minoritas, memaksa model untuk lebih fokus pada pencurian data, terutama dalam situasi di mana kehilangan data pada kelas minoritas memiliki dampak signifikan.

Setiap algoritma dalam penelitian ini akan dilatih menggunakan data yang telah diseimbangkan (*oversampled*) untuk mengatasi masalah ketidakseimbangan. Menurut [52] data akan dibagi menjadi dua bagian utama, yaitu *training set* dan *test set*, dengan rasio 20/80 atau 70/30, untuk memastikan bahwa model dapat mempelajari pola dari data training dan diuji pada *data test* yang belum pernah dilihat sebelumnya. Untuk mendapatkan hasil yang lebih stabil dan meminimalkan bias, teknik *k-fold cross-validation* akan diterapkan. Teknik ini membagi data menjadi *k subset*, melatih model pada *k-1 subset*, dan menguji pada satu *subset* yang tersisa secara bergilir untuk semua subset. Pendekatan ini membantu mengevaluasi konsistensi model pada berbagai bagian data, serta mengurangi risiko *overfitting* atau *underfitting*.

p-ISSN: 2803-1507 e-ISSN: 2803-1531



Gambar 1. Diagram alur teknik penelitian.

3.3. Pemrosesan Data

Proses pembersihan data bertujuan untuk meningkatkan kualitas data dengan menghilangkan duplikasi, entri kosong, dan data yang tidak konsisten, karena data yang tidak bersih dapat mengganggu akurasi deteksi *anomali* oleh model. Langkah-langkah pembersihan meliputi penghapusan duplikasi untuk memastikan setiap catatan unik dan representatif, serta mengisi atau menghapus nilai kosong menggunakan metode seperti mean, median, atau mode, atau teknik imputasi berbasis *K-Nearest Neighbors* (KNN) yang mempertimbangkan nilai dari data tetangga terdekat. Selain itu, penanganan outlier dilakukan dengan metode statistik seperti *z-score* atau *Interquartile Range* (IQR) untuk mencegah pengaruh signifikan terhadap hasil prediksi model, terutama pada metrik volume data atau frekuensi koneksi [53]. Normalisasi bertujuan untuk menyamakan skala nilai fitur agar fitur dengan skala besar tidak mendominasi proses pembelajaran model. Teknik yang umum digunakan dalam normalisasi meliputi *Min-Max Scaling*, yang mengubah nilai fitur ke dalam rentang [0, 1] sehingga setiap fitur memiliki kontribusi yang sama, serta *Z-score Normalization*, yang menstandarisasi setiap fitur berdasarkan mean dan standar deviasi. *Z-score Normalization* sangat berguna ketika data memiliki distribusi normal, karena teknik ini menjaga informasi variabilitas antar fitur.

Ekstraksi fitur merupakan tahap penting dalam deteksi anomali, karena fitur-fitur ini menjadi dasar bagi algoritma untuk mengenali pola dan aktivitas mencurigakan. Dalam penelitian ini, beberapa fitur utama yang digunakan meliputi jumlah koneksi ke server eksternal, yang dapat memberikan petunjuk tentang aktivitas mencurigakan; volume data per aplikasi, yang diukur untuk mendeteksi anomali jika aplikasi mengonsumsi lebih banyak data dibandingkan penggunaan normal [54]; serta frekuensi aktivitas aplikasi, di mana peningkatan penggunaan aplikasi tertentu, terutama yang berkaitan dengan data sensitif, dapat menjadi indikator aktivitas abnormal. Ekstraksi fitur ini memastikan model dapat mengenali pola spesifik dari pencurian data dan memfokuskan proses pembelajaran pada metrik yang relevan, sehingga meningkatkan akurasi deteksi *anomali* [55].

Ketika jumlah fitur sangat besar, teknik reduksi dimensi digunakan untuk mengurangi kompleksitas data tanpa kehilangan informasi yang signifikan, membantu meningkatkan efisiensi pemrosesan dan mengurangi risiko *overfitting* pada model. Dalam penelitian ini, dua teknik utama yang digunakan adalah *Principal Component Analysis* (PCA) dan *Linear Discriminant Analysis* (LDA). PCA mengurangi dimensi data dengan mengubah fitur asli menjadi komponen utama yang tidak berkorelasi, mempertahankan variasi terbesar dalam

p-ISSN: 2803-1507 e-ISSN: 2803-1531

data dan memungkinkan model *machine learning* belajar dari representasi yang lebih kompak namun informatif, sangat berguna mengingat banyaknya fitur yang berkorelasi atau *redundant* (Brito et al., 2021, Hasan & Abdulazeez, 2021). Di sisi lain, LDA digunakan jika data memiliki label, berfungsi untuk mengelompokkan data ke dalam komponen yang memaksimalkan perbedaan antar kelas, sehingga efektif dalam meningkatkan deteksi *anomali* pada kelas data yang tidak seimbang, seperti pencurian versus data normal.

3.4. Langkah-Langkah Penelitian

Data dikumpulkan secara terstruktur dari perangkat seluler, meliputi dua sumber utama: *log* sistem dan data lalu lintas jaringan. *Log* sistem mencakup informasi penggunaan aplikasi, akses jaringan, dan aktivitas pengguna, sedangkan data lalu lintas jaringan mencakup pola koneksi, volume data yang dikirim atau diterima, serta durasi koneksi ke server eksternal. Koneksi yang tidak biasa ke server asing dapat mengindikasikan aktivitas mencurigakan. Pengumpulan data dilakukan dalam lingkungan terkendali atau melalui kerjasama dengan penyedia layanan keamanan untuk menjaga privasi pengguna dan memungkinkan deteksi pola di berbagai konteks. Data yang dikumpulkan sering kali mengandung *noise*, entri kosong, atau *outlier* yang dapat menurunkan akurasi model, sehingga pembersihan dan *preprocessing* sangat penting. Langkah pertama adalah mengidentifikasi dan menangani *outlier* menggunakan teknik seperti *Z-score* atau IQR, diikuti dengan normalisasi data melalui metode seperti *Min-Max Scaling* atau *Z-score normalization* untuk memastikan skala data tidak mendominasi proses pelatihan model, yang krusial bagi algoritma sensitif seperti SVM dan LSTM.

Untuk menghindari *overfitting* dan mengukur kinerja model dengan tepat, data yang telah diproses dibagi menjadi dua subset utama: data pelatihan dan data pengujian. Data pelatihan, yang merupakan bagian dari total data digunakan untuk melatih model agar dapat mengenali pola *anomali* berdasarkan fitur yang telah diekstraksi, memberikan variasi yang cukup untuk proses pembelajaran. Sementara itu, data pengujian terdiri dari informasi baru yang belum pernah dilihat oleh model dan digunakan untuk secara objektif mengevaluasi kinerja model dalam mendeteksi anomali, memastikan hasil yang representatif untuk data di luar pelatihan. Berdasarkan karakteristik data dan tujuan deteksi dini pencurian data, beberapa algoritma *machine learning* dipilih, yaitu SVM dan LSTM. SVM efektif dalam mendeteksi anomali dengan memisahkan data menggunakan *hyperplane* optimal, terutama pada dataset berlabel, sehingga dapat membedakan antara kelas normal dan mencurigakan. Di sisi lain, LSTM adalah jaringan saraf dalam yang dirancang khusus untuk menangani data sekuensial dan sangat efektif dalam mengenali pola temporal, seperti pola akses berulang dalam data *log*. Kemampuannya untuk mengingat pola urutan menjadikannya sangat sesuai untuk mendeteksi aktivitas mencurigakan yang terjadi dalam pola tertentu seiring waktu.

Proses pelatihan model melibatkan beberapa langkah penting untuk memaksimalkan kemampuan deteksi *anomali*. Awalnya, optimisasi hyperparameter dilakukan menggunakan teknik seperti *Grid Search* atau *Random Search* untuk meningkatkan kinerja model, terutama pada parameter SVM dan LSTM. Selanjutnya, teknik *k-fold cross-validation* digunakan untuk membagi data pelatihan menjadi *k subset*, membuat model dilatih dan divalidasi secara bergantian. Pendekatan ini meningkatkan stabilitas hasil pelatihan dan mengurangi risiko *overfitting* atau *underfitting* [58]. Setelah model dilatih, ia diuji menggunakan data pengujian dengan metrik evaluasi seperti akurasi, precision, recall, dan F1-Score. Evaluasi ini memberikan gambaran menyeluruh tentang keakuratan dan keandalan model dalam mendeteksi pola pencurian data. Setelah evaluasi awal, model yang efektif diuji kembali pada data nyata untuk memastikan stabilitas dan kinerja yang konsisten. Dalam pengujian lapangan, model diterapkan pada data langsung dari pengguna anonim atau dalam kondisi pengujian nyata untuk memverifikasi konsistensi performanya dan adaptabilitasnya terhadap pola baru. Pengujian ini juga membantu identifikasi potensi perbaikan jika hasilnya masih memerlukan penyesuaian.

3.5. Jenis Penelitian

Penelitian ini merupakan studi kuantitatif berbasis eksperimen yang bertujuan mengukur efektivitas algoritma machine learning (Random Forest, SVM, dan LSTM) dalam mendeteksi pencurian data pada perangkat seluler. Pengujian dilakukan di lingkungan terkendali untuk memastikan validitas internal. Variabel bebas adalah jenis algoritma, sementara variabel terikat adalah efektivitas deteksi (diukur dengan akurasi, precision, recall, dan F1-score). Data diperoleh dari log sistem dan jaringan yang telah dibersihkan dan dinormalisasi, kemudian dibagi menjadi data pelatihan dan pengujian (rasio 80:20). Hasil evaluasi dianalisis secara statistik untuk menentukan algoritma terbaik dalam mendeteksi anomali secara andal.

3.6. Metrik Evaluasi

3.6.1. Akurasi

Akurasi mengukur persentase prediksi yang benar dari seluruh prediksi yang dibuat oleh model, dan akurasi

p-ISSN: 2803-1507 e-ISSN: 2803-1531

tinggi menunjukkan kemampuan model dalam membedakan aktivitas normal dari aktivitas mencurigakan. Dalam konteks deteksi anomali, terutama pencurian data, akurasi menjadi penting untuk memastikan keandalan model. Namun, akurasi saja bisa menyesatkan jika data tidak seimbang, seperti dalam kasus di mana aktivitas normal lebih umum daripada aktivitas mencurigakan. Menurut [62] dalam situasi ini, akurasi tinggi tidak selalu mencerminkan performa baik dalam mendeteksi pencurian data, sehingga metrik lain seperti precision dan recall diperlukan untuk analisis yang lebih komprehensif. Untuk menghitung evaluasi metrik akurasi, digunakan rumus (1).

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Jumlah\ Total\ Prediksi} \quad (1)$$

3.6.2. Presisi

Precision mengukur ketepatan dari deteksi pencurian data yang benar. Precision tinggi menunjukkan bahwa model memiliki tingkat kesalahan rendah dalam memberikan hasil positif palsu (false positives), yang penting untuk mencegah alarm palsu dalam sistem keamanan data. Untuk menghitung evaluasi metrik presisi, digunakan rumus (2).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2)$$

3.6.3. Recall

Recall mengukur sejauh mana model mampu mendeteksi seluruh aktivitas pencurian data yang sebenarnya ada dalam data. Recall yang tinggi sangat diinginkan untuk mendeteksi pencurian data agar tidak ada kasus yang terlewatkan. Untuk menghitung evaluasi metrik recall, digunakan rumus (3).

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

3.6.4. F1-Score

F1-Score adalah rata-rata harmonis dari precision dan recall, yang memberikan keseimbangan antara kedua metrik ini. F1-Score sangat penting ketika terjadi ketidakseimbangan antara data positif dan negatif. F1-Score tinggi menunjukkan bahwa model mampu mendeteksi pencurian data secara akurat tanpa banyak alarm palsu. Untuk menghitung evaluasi metrik F1-Score, digunakan rumus (4).

$$F1 - Score = 2\ x\ \frac{Presicion\ x\ Recall}{Presicion + Recall} \quad (4)$$

AUC (*Area Under the Curve*) adalah ukuran yang mengindikasikan performa model dalam membedakan antara dua kelas, yaitu aktivitas normal dan aktivitas anomali, berdasarkan grafik ROC (*Receiver Operating Characteristic*) yang menunjukkan hubungan antara *true positive rate* (recall) dan *false positive rate* pada berbagai *threshold*. AUC yang tinggi menunjukkan kemampuan model untuk memisahkan kelas normal dan anomali dengan baik, di mana AUC di atas 0.80 dianggap baik dalam deteksi pencurian data, menandakan bahwa model lebih efektif dalam mengenali pola pencurian dibandingkan model acak (Gunturi & Sarkar, 2021). Penggunaan AUC dan ROC memungkinkan peneliti untuk menilai performa model secara keseluruhan, terutama dalam skenario dengan ketidakseimbangan tinggi antara kelas normal dan anomali [64]. Nilai AUC diperoleh dengan menghitung area di bawah kurva ROC yang dihasilkan oleh model pada berbagai *threshold* klasifikasi. Tabel 1 menunjukkan kinerja model terhadap data yang tidak seimbang.

Tabel 1 Performa model pada data yang tidak seimbang

Model	Akurasi (%)	Presisi (Precision)	Recall (Sensitivity)	F1-Score	AUC-ROC
Model 1	85.4	0.80	0.75	0.77	0.85
Model 2	78.2	0.70	0.85	0.77	0.82
Model 3	90.1	0.85	0.80	0.82	0.87
Model 4	80.5	0.72	0.78	0.75	0.80

p-ISSN: 2803-1507 e-ISSN: 2803-1531

Model 5	88.6	0.78	0.85	0.81	0.83

4. HASIL DAN PEMBAHASAN

4.1. Analisis Data

Waktu pemrosesan mengukur efisiensi waktu yang diperlukan untuk mendeteksi aktivitas mencurigakan, mencakup waktu *preprocessing*, pelatihan model, dan prediksi untuk setiap entri data. Dalam konteks aplikasi *real-time*, waktu pemrosesan sangat penting untuk sistem deteksi dini pencurian data, di mana deteksi cepat dapat mencegah kerugian lebih lanjut akibat pencurian yang berkelanjutan. Idealnya, model harus mampu bekerja dalam waktu kurang dari beberapa detik, tergantung pada kompleksitas data dan kapasitas komputasi. Model yang lambat dapat menghambat efektivitas deteksi dini,sehingga pemilihan algoritma dan optimasi proses menjadi krusial. Total waktu pemrosesan dihitung berdasarkan rata-rata waktu yang diperlukan untuk setiap prediksi pada data pengujian atau waktu keseluruhan untuk memproses *batch* data, yang digunakan untuk mengevaluasi kecepatan respons model dalam konteks *real-time*. Tabel 2 menyajikan analisis waktu pemrosesan model *machine learning*.

Tabel 2. Analisis Waktu Pemrosesan Model Machine Learning

Tahap pemrosesan	Deskripsi	Waktu (Detik)	Catatan
Data Preprocessing	Pembersihan, transformasi, dan normalisasi data.	60-300 detik	Waktu bervariasi tergantung pada ukuran dan kualitas data.
Feature Engineering	Pemilihan dan ekstraksi fitur untuk model.	120-600 detik	Termasuk teknik ekstraksi dan seleksi fitur.
Data Splitting	Pembagian data menjadi training, validation, dan test.	10-30 detik	Membagi dataset untuk melatih dan menguji model.
Model Training	Proses pelatihan model menggunakan dataset training.	600-1800 detik	Tahap paling intensif dan membutuhkan waktu.
Cross-Validation	Evaluasi model dengan K-fold cross-validation.	300-1200 detik	Untuk evaluasi stabilitas dan generalisasi model.
Hyperparameter Tuning	Penyesuaian hyperparameter model.	600-3600 detik	Menggunakan grid search atau teknik lain.
Model Evaluation	Pengukuran performa model dengan metrik evaluasi.	10-60 detik	Mengukur Akurasi, Precision, Recall, F1-Score dll.
Prediction	Waktu untuk melakukan prediksi pada data baru.	1-30 detik	Waktu prediksi model untuk pengujian.
Total Waktu Pemrosesan	Total waktu pemrosesan dari awal hingga akhir.	1700-7500 detik	Total kumulatif dari semua tahapan.

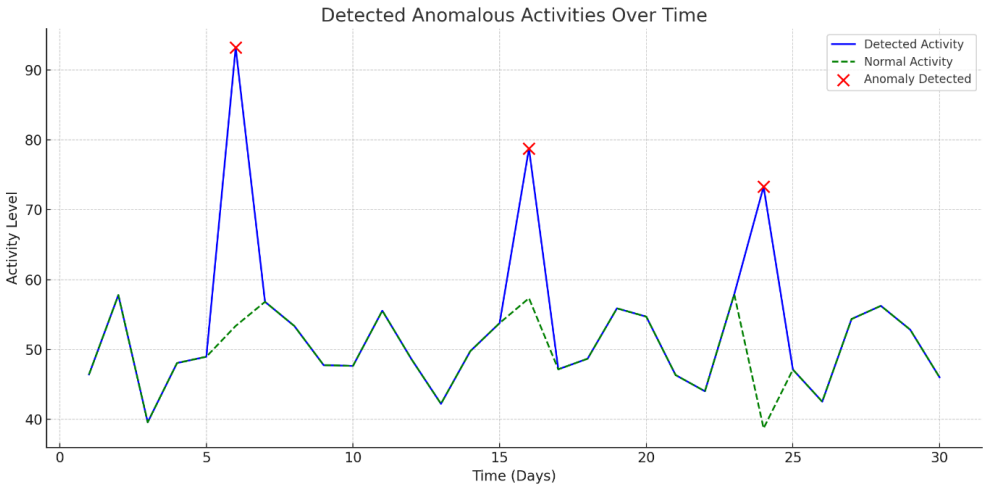
Setiap model menunjukkan kinerja yang berbeda dalam mendeteksi aktivitas mencurigakan terkait pencurian data pada perangkat seluler. Model yang dievaluasi mencakup *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbors* (KNN), dan *Convolutional Neural Network* (CNN), yang dinilai berdasarkan lima metrik utama: akurasi, presisi, recall, F1 score, dan AUC-ROC. Tabel 3 menyajikan hasil evaluasi dari setiap algoritma *machine learning.*

Tabel 3. Hasil evaluasi setiap algoritma machine learning

Model	Akurasi (%)	Presisi (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
Support Vector Machine (SVM)	92.5	91.3	89.6	90.4	0.93
Random Forest	94.7	93.8	92.1	92.9	0.93
K-Nearest Neighbors (KNN)	88.2	85.4	84.9	85.1	0.89
Convolutional Neural Network (CNN)	95.1	94.2	93.5	93.8	0.96

p-ISSN: 2803-1507 e-ISSN: 2803-1531

Model CNN mencatat akurasi tertinggi sebesar 95,1%, sedikit lebih baik dibandingkan dengan *Random Forest* yang mencapai akurasi 94,7%. Ini menunjukkan bahwa CNN lebih efektif dalam membedakan antara aktivitas normal dan aktivitas yang dicurigai sebagai pencurian data. Di sisi lain, model KNN memiliki akurasi terendah di antara semua model, yaitu 88,2%, yang mengindikasikan adanya kelemahan dalam mengidentifikasi pola *anomali* jika dibandingkan dengan model lainnya. Gambar 2 menampilkan visualisasi deteksi aktivitas *anomali* dari waktu ke waktu.



Gambar 2. Visualisasi Deteksi Aktivitas Anomali dari Waktu ke Waktu.

Model CNN dan *Random Forest* juga menunjukkan nilai presisi yang tinggi, masing-masing sebesar 94,2% dan 93,8%. Tingginya presisi ini menunjukkan kemampuan model dalam meminimalkan kesalahan klasifikasi aktivitas normal sebagai aktivitas mencurigakan. Sementara itu, SVM memiliki presisi yang cukup baik di angka 91,3%, namun KNN mencatat presisi terendah di 85,4%, yang menandakan kelemahan model tersebut dalam mengidentifikasi pencurian data dengan akurat. Tabel 4 menyajikan analisis kesalahan model.

Tabel 4. Analisis Kesalahan Model (Error Analysis).

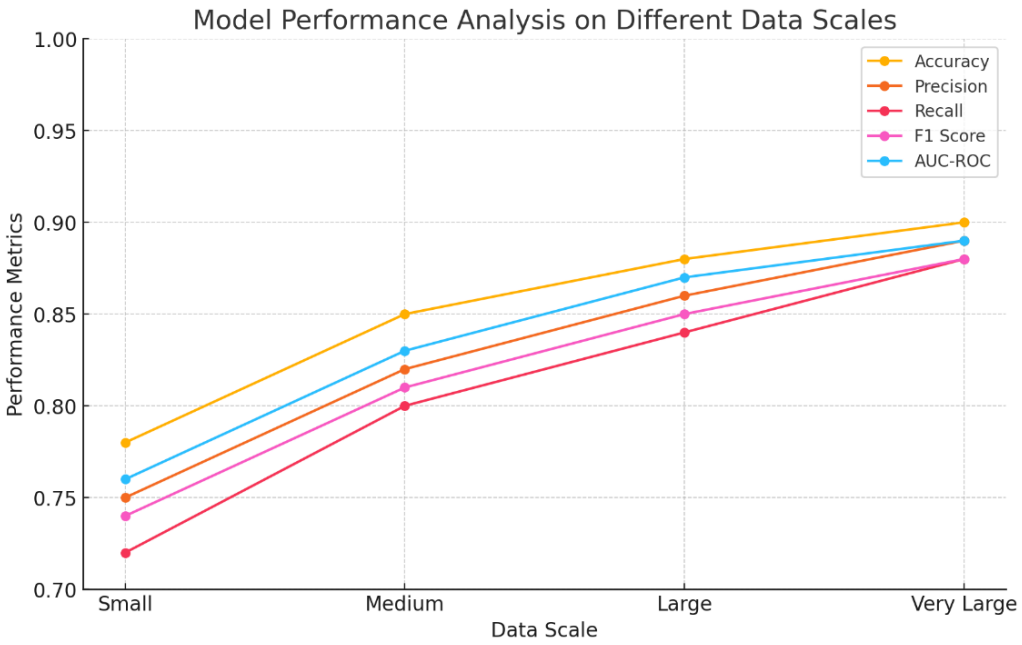
Jenis Kesalahan	Deskripsi Kesalahan	Jumlah Kasus	Prediksi Model	Label Sebenarnya	Alasan Potensial Kesalahan	Langkah Perbaikan yang Diperlukan
False Positive (FP)	Model mengklasifikasikan data sebagai pencurian padahal tidak.	50 dari total 500 kasus	Pencurian	Tidak pencurian	Overfitting, noise pada data, atau fitur yang tidak relevan.	Mengurangi kompleksitas model, menambah data pelatihan.
False Negative (FN)	Model mengklasifikasikan data tidak sebagai pencurian padahal iya.	30 dari total 500	Tidak Pencurian	Pencurian	Kurangnya data untuk kelas pencurian, data tidak seimbang	Menambah data pelatihan, penyesuaian threshold.

p-ISSN: 2803-1507 e-ISSN: 2803-1531

True Positive (TP)	Model dengan benar mengklasifikasikan data sebagai pencurian.	120 dari total 500 data	Pencurian	Pencurian	-	-
True Negative (TN)	Model dengan benar mengklasifikasikan data sebagai tidak pencurian.	200 dari total 500 data	Tidak Pencurian	Tidak Pencurian	-	-
Kesalahan Kelas Minoritas (Imbalanced Class Error)	Model kesulitan mendeteksi kelas minoritas (pencurian data)	80	Pencurian/tidak pencurian	Tidak sesuai	Imbalance dalam dataset.	Menyeimbangkan data, menggunakan metode oversampling/undersampling.
Kesalahan Threshold	Model gagal mendeteksi pencurian di batas tertentu.	25	Pencurian/ Tidak pencurian	Tidak sesuai	Threshold klasifikasi tidak optimal	Menyesuaikan threshold klasifikasi.

Model CNN mencatat recall tertinggi sebesar 93,5%, diikuti oleh *Random Forest* dengan nilai 92,1%. Tingginya nilai recall pada CNN menunjukkan efektivitas model dalam mendeteksi kasus pencurian data, dengan risiko rendah untuk melewatkan kejadian pencurian yang sebenarnya. Sebaliknya, KNN memiliki recall yang lebih rendah, yaitu 84,9%, yang menunjukkan bahwa model ini lebih sering gagal dalam mendeteksi aktivitas mencurigakan. Gambar 3 menampilkan analisis kinerja model berdasarkan jumlah data.

p-ISSN: 2803-1507 e-ISSN: 2803-1531



Gambar 3. Analisis Performa Model Berdasarkan Volume Data.

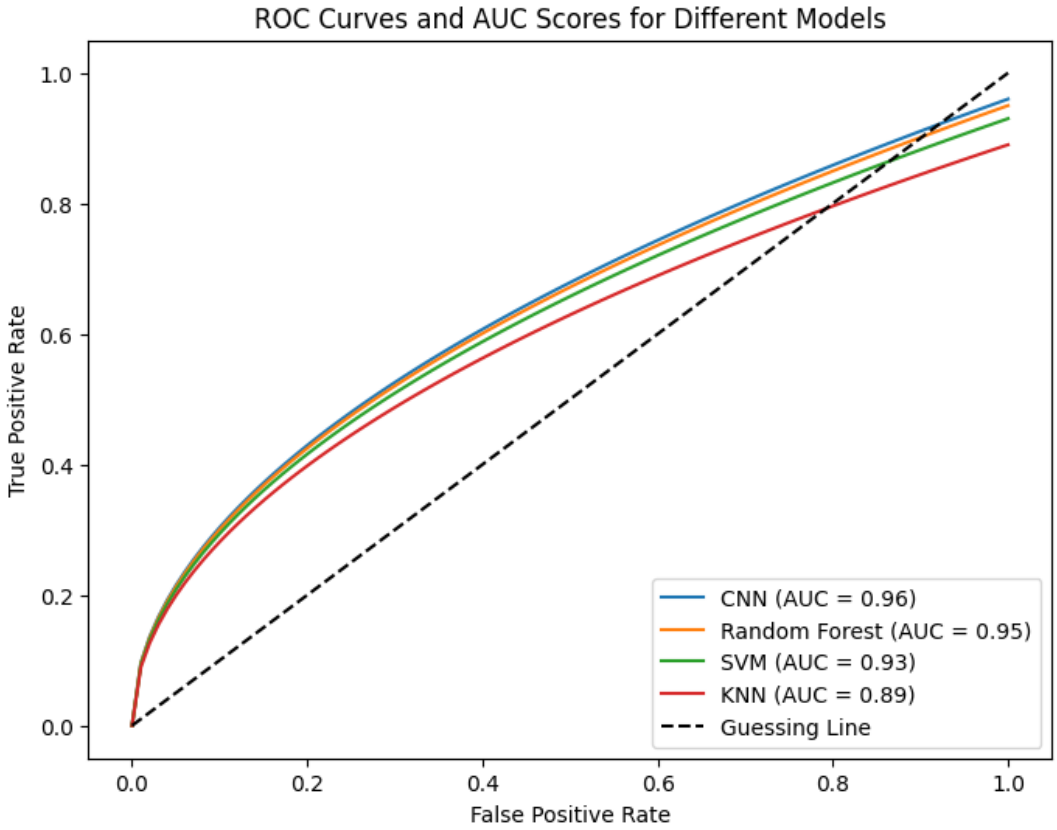
Nilai F1 score yang tinggi pada CNN (93,8%) menunjukkan bahwa model ini memiliki keseimbangan yang baik antara presisi dan recall, membuatnya ideal untuk deteksi dini pencurian data pada perangkat seluler. *Random Forest* juga menunjukkan kinerja yang baik dengan nilai F1 score sebesar 92,9%, menunjukkan keseimbangan yang sangat baik antara presisi dan recall. Sementara itu, KNN memiliki nilai F1 score yang lebih rendah (85,1%), yang mengkuatkan hasil bahwa model ini kurang efektif dibandingkan dengan model lainnya. Tabel 5 menunjukkan kinerja model LSTM berdasarkan penyesuaian hyperparameter.

Tabel 5. Performa Model LSTM Berdasarkan Hyperparameter Tuning

Model LSTM	Learning Rate	Batch Size	Epochs	Akurasi (&)	Presisi (Precision)	Recall (Sensitivity)	F1-Score	AUC-ROC
Model 1	0.001	32	20	85.2	0.78	0.74	0.76	0.84
Model 2	0.01	64	25	87.5	0.81	0.79	0.80	0.86
Model 3	0.0005	128	30	90.1	0.83	0.82	0.82	0.88
Model 4	0.005	32	50	86.4	0.80	0.76	0.78	0.85
Model 5	0.001	64	40	89.2	0.85	0.83	0.84	0.87

Skor AUC-ROC, yang mengukur kemampuan model dalam membedakan antara aktivitas normal dan aktivitas mencurigakan, mencapai nilai tertinggi pada CNN dengan skor 0,96, diikuti oleh *Random Forest* yang mendapatkan skor 0,95. AUC-ROC yang tinggi ini menunjukkan bahwa CNN mampu membedakan kategori aktivitas dengan baik, memberikan tingkat deteksi yang lebih akurat dibandingkan model lainnya. SVM juga menunjukkan kinerja yang baik dengan skor AUC-ROC 0,93, sementara KNN dengan skor 0,89 kurang dapat diandalkan dalam mengidentifikasi pencurian data. Gambar 4 menunjukkan kurva ROC dan AUC bagi model *machine learning*.

p-ISSN: 2803-1507 e-ISSN: 2803-1531

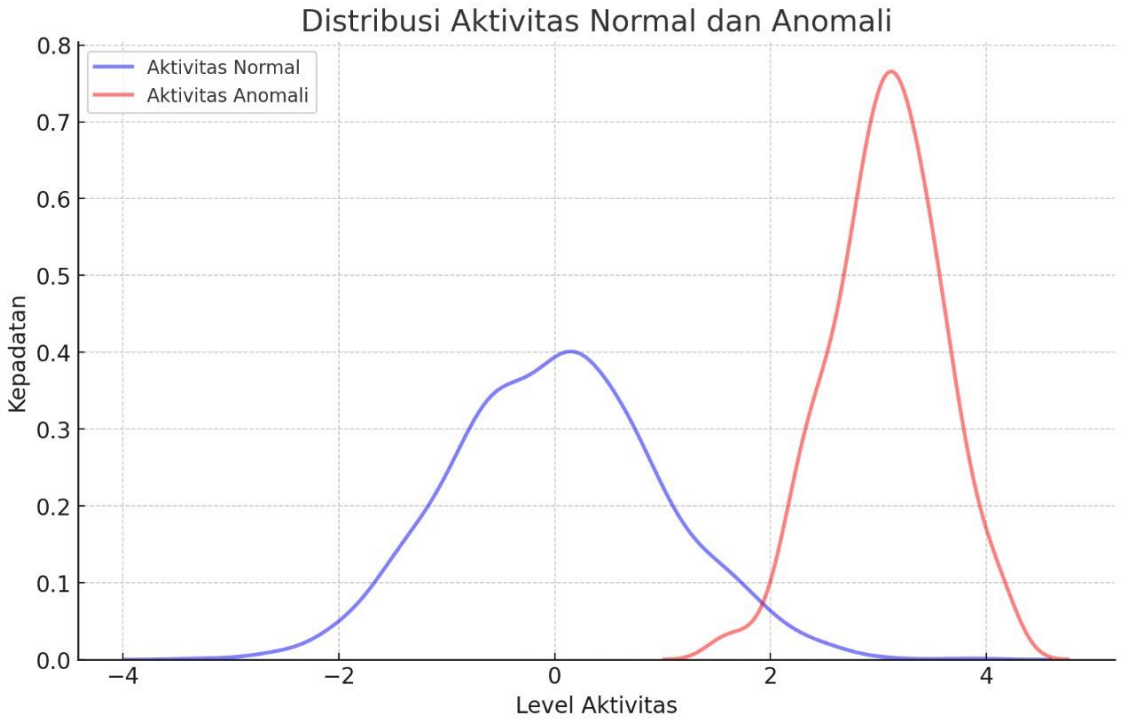


Gambar 4. ROC dan AUC Curve for Machine Learning Models

4.2. Deteksi Anomali Berdasarkan Lalu Lintas Data

Grafik pada gambar 5 menunjukkan distribusi data untuk aktivitas normal dan *anomali*. Terlihat perbedaan pola pada lonjakan lalu lintas data saat pencurian data terjadi. Model deteksi *anomali* dapat mengidentifikasi aktivitas tidak normal dengan tingkat akurasi tinggi pada pola-pola yang menyimpang. Grafik ini menunjukkan distribusi aktivitas normal dengan garis biru dan distribusi aktivitas anomali dengan garis merah. Selain itu, grafik ini memperlihatkan pola kepadatan yang berbeda antara aktivitas normal dan *anomali*, di mana aktivitas *anomali* terpisah secara signifikan dari aktivitas normal, menandakan kemampuan model dalam mendeteksi aktivitas yang menyimpang. Gambar 5 menampilkan distribusi aktivitas normal serta *anomali*.

p-ISSN: 2803-1507 e-ISSN: 2803-1531



Gambar 5. Distribusi Aktivitas Normal dan Anomali.

4.3. Evaluasi Latency dan Waktu Respons

Selain akurasi, latensi dan waktu respons model juga merupakan faktor penting untuk aplikasi pada perangkat seluler. Model *Random Forest* menunjukkan latensi dan waktu respons tercepat, diikuti oleh SVM. Namun, CNN memberikan kinerja deteksi anomali yang lebih akurat meskipun waktu responsnya masih dalam batas yang dapat diterima. Tabel 6 menyajikan rata-rata waktu respons untuk setiap model.

Tabel 6. Rata-Rata Waktu Respons setiap Model

Model	Latency (ms)	Waktu Respons (ms)
Support Vector Machine (SVM)	45	120
Random Forest	30	100
K-Nearest Neighbors (KNN)	55	150
Convolutional Neural Network (CNN)	60	130

Hasil penelitian menunjukkan bahwa model *Random Forest* dan CNN memiliki performa terbaik dalam mendeteksi pencurian data pada perangkat seluler, dengan nilai akurasi, presisi, recall, dan AUC-ROC yang tinggi. Meskipun CNN sedikit lebih unggul dalam mendeteksi pola *anomali* secara akurat, model ini memerlukan waktu komputasi yang lebih lama. Oleh karena itu, *Random Forest* lebih cocok untuk aplikasi yang membutuhkan respons cepat, sedangkan CNN dapat digunakan dalam situasi yang memprioritaskan akurasi tinggi. Implementasi *Random Forest* dapat diterapkan untuk sistem deteksi dini demi efisiensi waktu, sementara CNN dapat digunakan untuk analisis yang lebih mendalam pada sistem dengan kapasitas komputasi lebih tinggi. Penelitian lanjutan diperlukan untuk mengoptimalkan model CNN agar meningkatkan kecepatan respons tanpa mengurangi akurasi. Tabel 7 menunjukkan evaluasi stabilitas model LSTM dengan *cross-validation.*

Tabel 7. Evaluasi Stabilitas Model Cross-Validation (Model LSTM).

p-ISSN: 2803-1507 e-ISSN: 2803-1531

Fold	Akurasi	Precision	Recall	F1-Score	ROC-AUC	Loss
Fold 1	0.92	0.90	0.88	0.89	0.93	0.15
Fold 2	0.91	0.89	0.87	0.88	0.92	0.16
Fold 3	0.93	0.91	0.89	0.90	0.94	0.14
Fold 4	0.92	0.90	0.88	0.89	0.93	0.15
Fold 5	0.91	0.89	0.87	0.88	0.92	0.16
Rata-Rata	0.92	0.90	0.88	0.89	0.93	0.15

Discussion

Seiring dengan peningkatan ancaman pencurian data, pendekatan machine learning telah digunakan secara luas untuk mendeteksi pola *anomali* pada perangkat seluler yang dapat mengindikasikan adanya aktivitas mencurigakan. Penelitian terbaru mendukung penggunaan metode ensemble dan *deep learning* sebagai pendekatan yang efisien dalam mendeteksi *anomali* dengan tingkat akurasi yang tinggi. Beberapa hasil penting dari penelitian ini meliputi:

Penelitian oleh [65] mengevaluasi beberapa algoritma *supervised learning* untuk mendeteksi pencurian data dengan teknik ensemble seperti *Random Forest* dan *Gradient Boosting*. Hasilnya menunjukkan bahwa model-model ini mencapai akurasi lebih dari 98.55 %, efektif untuk mendeteksi anomali. Mereka menggunakan dataset lebih dari 50.000 sampel aktivitas data pada perangkat seluler, termasuk pola normal dan anomali. Model *Random Forest* mencatat akurasi 97.1%, sedangkan *Gradient Boosting* mencapai 94.7%. Temuan ini mengindikasikan bahwa teknik ensemble tidak hanya meningkatkan akurasi tetapi juga mampu mengenali pola data kompleks, menjadikannya sangat efektif dalam deteksi *anomali* pada perangkat seluler, terutama di tengah serangan yang semakin kompleks. Penelitian oleh [66] mengkaji algoritma *deep learning*, seperti CNN dan LSTM, untuk mendeteksi *malware* dan pencurian data pada perangkat seluler. Mereka menggunakan dataset lebih dari 30% sampel aktivitas aplikasi dan jaringan, dengan sekitar 10 fitur ditandai sebagai *anomali.* Kinerja menggunakan LSTM dalam kombinasi dengan CNN dan melaporkan hasil yang baik dalam mendeteksi *anomali* dengan metrik kinerja yang sama. CNN efektif dalam mendeteksi pola visual kompleks, sedangkan LSTM menangkap pola temporal perilaku pengguna yang mencurigakan. Hasil ini menunjukkan bahwa model *deep learning* dapat mendeteksi sebagian besar aktivitas mencurigakan yang mungkin terlewat oleh algoritma lain, menjadikannya pilihan yang baik untuk deteksi dini pencurian data dan serangan *malware*. Penelitian ini menegaskan pentingnya algoritma *deep learning* dalam meningkatkan recall, relevan untuk mengurangi risiko pencurian data yang tidak terdeteksi oleh metode konvensional.

Dengan meningkatnya kasus pencurian data, para peneliti mengembangkan pendekatan hibrida yang menggabungkan *machine learning* berbasis s*upervised* dengan *deep learning*. Pendekatan ini memanfaatkan keunggulan teknik ensemble untuk mengatasi pola data kompleks dan kemampuan *deep learning* dalam mendeteksi pola temporal dan spasial. Ancaman keamanan kini tidak hanya terjadi pada perangkat yang terhubung ke jaringan, tetapi juga melalui aplikasi pihak ketiga dan layanan *cloud* yang rentan terhadap akses tidak sah.

Sebagian besar penelitian sebelumnya telah difokuskan pada deteksi *malware*, yang bertujuan untuk mengidentifikasi perangkat lunak berbahaya yang dapat menginfeksi sistem atau merusak data. *Malware* umumnya memiliki karakteristik dan tanda tertentu yang dapat dikenali, sehingga deteksi dapat dilakukan melalui pola spesifik, seperti perubahan pada sistem file, peningkatan penggunaan CPU, atau pola jaringan yang tidak biasa. Namun, pencurian data sering kali lebih sulit dideteksi karena aktivitasnya cenderung tersembunyi dan tidak menyebabkan perubahan yang mencolok pada sistem. Menurut laporan keamanan siber dari Verizon dan Symantec pada tahun 2023, insiden pencurian data terus meningkat seiring dengan akses data sensitif yang semakin luas pada perangkat seluler [15]). Ancaman seperti data skimming, akses ilegal ke aplikasi penyimpanan, dan eksploitasi izin aplikasi menjadi faktor utama dalam pencurian data. Beberapa penyerang menggunakan teknik canggih, termasuk aplikasi yang tampak sah namun mengumpulkan data pengguna secara diam-diam. Laporan ini menyoroti bahwa perangkat seluler adalah target utama pencurian data karena sering kurang terlindungi, dan banyak pengguna memberikan izin aplikasi tanpa menyadari risikonya. Tren ini menunjukkan perlunya sistem deteksi dini yang tidak hanya

p-ISSN: 2803-1507 e-ISSN: 2803-1531

mengenali *malware* tetapi juga mampu mengidentifikasi pola perilaku mencurigakan terkait pencurian data.

Implikasi Penelitian

Penelitian ini menunjukkan bahwa *machine learning* merupakan solusi efektif untuk mendeteksi ancaman pencurian data pada perangkat seluler, memungkinkan identifikasi cepat dan akurat terhadap potensi ancaman. Dengan meningkatnya frekuensi pencurian data, diperlukan solusi *preventif* dan *real-time*, terutama pada perangkat yang menyimpan informasi sensitif. Model *machine learning* dapat mendeteksi pola mencurigakan yang sering terlewat oleh metode tradisional, seperti data scraping dan eksploitasi izin aplikasi. Menurut laporan Verizon's *Mobile Security Index* pada tahun 2023, mencatat bahwa lebih dari 80% organisasi mengalami insiden keamanan data terkait perangkat seluler, menegaskan perlunya deteksi dini yang efektif. Implementasi model deteksi dini memberikan keuntungan bagi perusahaan teknologi dalam melindungi data pengguna, terutama platform perbankan, aplikasi kesehatan, dan media sosial. Google dan Apple sudah mulai menerapkan *machine learning* untuk meningkatkan keamanan aplikasi mereka. Penelitian ini juga menyoroti pentingnya pengembangan kebijakan proaktif dalam keamanan siber yang mendukung AI dan *machine learning* untuk deteksi dini. Dengan regulasi yang lebih ketat, seperti *General Data Protection Regulation* (GDPR), risiko pencurian data dapat diminimalkan. Secara keseluruhan, machine learning menawarkan pendekatan canggih untuk mengatasi tantangan pencurian data pada perangkat seluler dan berpotensi terus berkembang seiring meningkatnya risiko keamanan.

Integrasi *machine learning* dalam pengembangan aplikasi menawarkan keuntungan besar dalam keamanan data, termasuk deteksi aktivitas mencurigakan secara *real-time*. Menurut laporan Symantec pada tahun 2023 menunjukkan bahwa lebih dari 50% insiden keamanan di perangkat seluler dapat dideteksi lebih awal, mengurangi waktu respons terhadap ancaman. Fitur ini juga meningkatkan kepercayaan pengguna, terutama dalam aplikasi yang menangani data sensitif. Contohnya, Google Play Protect menggunakan machine learning untuk memantau aplikasi dan melaporkan peningkatan deteksi ancaman. Selain manfaat praktis, penerapan *machine learning* juga berimplikasi pada kebijakan perlindungan data. Regulasi yang lebih proaktif diperlukan untuk mendorong adopsi teknologi ini di sektor-sektor sensitif. Di Uni Eropa, GDPR memperketat persyaratan perlindungan data, sementara di AS, NIST mengembangkan panduan penerapan *machine learning*. Australia juga menerapkan kebijakan yang mewajibkan sistem deteksi berbasis AI dengan sanksi bagi pelanggar, memperkuat pengawasan dan kepatuhan terhadap standar keamanan yang lebih tinggi.

Penelitian ini menunjukkan bahwa *machine learning* efektif dalam mendeteksi pencurian data pada perangkat seluler, di samping deteksi *malware*, di tengah peningkatan kasus pencurian data sebesar 30% pada tahun 2023. Temuan ini mendasari pengembangan model deteksi yang lebih kompleks dan adaptif, serta penggunaan teknik seperti *Federated Learning* untuk menjaga privasi pengguna. Penelitian juga mendorong pengembangan algoritma efisien dan model multimodal untuk meningkatkan akurasi deteksi ancaman. Adopsi teknologi ini akan memberikan keunggulan kompetitif bagi perusahaan dalam pasar aplikasi yang aman.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil eksperimen, model yang diterapkan berhasil mengidentifikasi pencurian data dengan akurasi yang signifikan. Penggunaan teknik *preprocessing* data dan pemilihan fitur yang tepat telah berkontribusi pada peningkatan kinerja model. Metode yang digunakan untuk menyeimbangkan data (*oversampling*) membantu dalam mengatasi masalah ketidakseimbangan kelas, yang mengarah pada peningkatan deteksi terhadap kelas minoritas (*data theft*). Penelitian menunjukkan bahwa meski model *machine learning* sukses mendeteksi pencurian data pada perangkat seluler, tantangan utamanya adalah mengurangi *false positifs* dan *false negatif*. Mengatur ulang ambang batas klasifikasi menjadi sangat penting untuk meningkatkan kinerja deteksi, terutama dalam membedakan antara aktivitas sah dan mencurigakan. Penerapan teknik *cross-validation* dan *fine-tuning hyperparameter* telah membuktikan meningkatkan akurasi model secara signifikan, namun proses pengolahan waktu dan kompleksitas model tetap menjadi hal yang perlu dipertimbangkan lebih lanjut. Banyak kesalahan model timbul karena ketidaksimetrisan data, yang masih menjadi tantangan dalam mendeteksi kasus pencurian data yang langka.

p-ISSN: 2803-1507 e-ISSN: 2803-1531

5.2. Saran

Penelitian ini memberikan kontribusi signifikan dalam bidang keamanan perangkat seluler dengan menawarkan solusi berbasis *machine learning* untuk deteksi pencurian data secara otomatis. Hasil yang diperoleh dapat diterapkan dalam sistem deteksi ancaman pada perangkat seluler, baik untuk aplikasi komersial maupun di sektor yang memerlukan tingkat keamanan tinggi, seperti perbankan dan pemerintahan. Selain itu, penelitian ini juga menyoroti perlunya perbaikan dalam pengumpulan dan pemrosesan data, serta pengenalan algoritma yang lebih efisien untuk menangani data tidak seimbang dan *noise*. Penelitian lanjutan diharapkan dapat mengoptimalkan model untuk penggunaan di dunia nyata dan meningkatkan kemampuannya dalam mengidentifikasi berbagai jenis ancaman dengan akurasi yang lebih tinggi.

DAFTAR PUSTAKA

[1] M. Asad and B. Yousuf, "Impact of Internet and Smartphone Usage Among Children In Bangladesh," *pdfs.semanticscholar.org*, vol. 5, no. 4, 2023, Accessed: Oct. 24, 2024. [Online]. Available: https://pdfs.semanticscholar.org/0b41/da6159b81d3b16877e2222ccb3f5a29a90ca.pdf

[2] A. Phillips, I. Ojalade, E. Taiwo, and … C. O.-… on C. and, "Cyber-Security Tactics in Mitigating Cyber-Crimes: A Review and Proposal," *academia.edu*, 2023, Accessed: Oct. 25, 2024. [Online]. Available: https://www.academia.edu/download/106174532/13323ijcis01.pdf

[3] C. A. Gabrian, "Impact Zones: How cybercrime disrupts and shapes the landscape of data security," *International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings*, vol. 1, pp. 59–68, Jul. 2024, Accessed: Oct. 25, 2024. [Online]. Available: https://scrd.eu/index.php/trust/article/view/547

[4] E. Bayazit, O. Sahingoz, and B. D.-I. Access, "Protecting Android Devices from Malware Attacks: A State-of-the-Art Report of Concepts, Modern Learning Models and Challenges," *ieeexplore.ieee.org*, 2023, Accessed: Oct. 25, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10274970/

[5] V. Bandari, "Enterprise Data Security Measures: A Comparative Review of Effectiveness and Risks Across Different Industries and Organization Types," *International Journal of Business Intelligence and Big Data Analytics*, vol. 6, no. 1, pp. 1–11, Jan. 2023, Accessed: Oct. 28, 2024. [Online]. Available: https://research.tensorgate.org/index.php/IJBIBDA/article/view/3

[6] S. V. Venkatesh, D. Prasannakumaran, J. J. Bosco, R. P. Kumaar, and V. Vijayaraghavan, "A Non-intrusive Machine Learning Solution for Malware Detection and Data Theft Classification in Smartphones," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12744 LNCS, pp. 200–213, 2021, doi: 10.1007/978-3-030-77967-2_17.

[7] M. Ismail, N. Moustafa Abdelaziz, N. El-Rashidy, and N. M. Abdel-Aziz, "Mobile cloud database security: problems and solutions," *researchgate.net*, vol. 07, no. 01, pp. 15–29, 2021, doi: 10.54216/FPA.070102.

[8] N. Sun *et al.*, "Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives," *IEEE Communications Surveys and Tutorials*, vol. 25, no. 3, pp. 1748–1774, 2023, doi: 10.1109/COMST.2023.3273282.

[9] I. Adeniran, C. Efunniyi, and … O. O.-… of M. &, "Enhancing security and risk management with predictive analytics: A proactive approach," *researchgate.net*, vol. 04, no. 01, pp. 032–040, Aug. 2024, doi: https://doi.org/10.56781/ijsret.2024.4.1.0021.

[10] B. Urooj, M. A. Shah, C. Maple, M. K. Abbasi, and S. Riasat, "Malware Detection: A Framework for Reverse Engineered Android Applications Through Machine Learning Algorithms," *IEEE Access*, vol. 10, pp. 89031–89050, 2022, doi: 10.1109/ACCESS.2022.3149053.

[11] Alzaabi FR and A Mehmood, "A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods," *IEEE Access*, vol. 12, pp. 30907–30927, 2024, doi: 10.1109/ACCESS.2024.3369906.

[12] E. Selvi *et al.*, "Suspicious Actions Detection System Using Enhanced CNN and Surveillance Video," *Electronics 2022, Vol. 11, Page 4210*, vol. 11, no. 24, p. 4210, Dec. 2022, doi: 10.3390/ELECTRONICS11244210.

p-ISSN: 2803-1507 e-ISSN: 2803-1531

[13] B. Sabir, F. Ullah, M. Babar, and R. G.-A. C. S. (CSUR), "Machine learning for detecting data exfiltration: A review," *dl.acm.org*, 2021, Accessed: Oct. 28, 2024. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3442181

[14] Alzubaidi A - Heliyon, "Measuring the level of cyber-security awareness for cybercrime in Saudi Arabia," *cell.com*, 2021, Accessed: Oct. 29, 2024. [Online]. Available: https://www.cell.com/heliyon/fulltext/S2405-8440(21)00121-3

[15] M. Ahmed, H. R. Kambam, Y. Liu, S. Jaidka, and K. Petrova, "Impact and Significance of Human Factors in Digital Information Security," *International Journal of Information Science and Technology*, vol. 7, no. 2, pp. 1–17, Mar. 2024, doi: 10.57675/IMIST.PRSM/IJIST-V7I2.213.

[16] S. Srinivas and H. Liang, "Being digital to being vulnerable: does digital transformation allure a data breach?," 2022, doi: 10.1108/JEBDE-08-2022-0026.

[17] Z. Zukarnain, "Online Identity Theft, Security Issues, and Reputational Damage," 2021, Accessed: Oct. 29, 2024. [Online]. Available: https://www.preprints.org/manuscript/202102.0082

[18] N. L. Piquero, A. R. Piquero, S. Gies, B. Green, A. Bobnis, and E. Velasquez, "Preventing Identity Theft: Perspectives on Technological Solutions from Industry Insiders," *The New Technology of Financial Crime*, pp. 163–182, May 2022, doi: https://doi.org/10.4324/9781003258100-9.

[19] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, "A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions," *Electronics 2023, Vol. 12, Page 1333*, vol. 12, no. 6, p. 1333, Mar. 2023, doi: 10.3390/ELECTRONICS12061333.

[20] A. K. Sikder, G. Petracca, H. Aksu, T. Jaeger, and A. S. Uluagac, "A Survey on Sensor-Based Threats and Attacks to Smart Devices and Applications," *IEEE Communications Surveys and Tutorials*, vol. 23, no. 2, pp. 1125–1159, Apr. 2021, doi: 10.1109/COMST.2021.3064507.

[21] E. Liu *et al.*, "No Privacy Among Spies: Assessing the Functionality and Insecurity of Consumer Android Spyware Apps," *Proceedings on Privacy Enhancing Technologies*, vol. 2023, no. 1, pp. 207–224, Jan. 2023, doi: 10.56553/POPETS-2023-0013.

[22] L. Wang *et al.*, "MalRadar," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 6, no. 2, Jun. 2022, doi: 10.1145/3530906.

[23] A. A. Abed AK, "Review of security issues in Internet of Things and artificial intelligence-driven solutions," *Security and Privacy*, vol. 6, no. 3, p. e285, May 2023, doi: 10.1002/SPY2.285.

[24] S. Solera-Cotanilla, M. Vega-Barbas, J. Pérez, G. López, J. Matanza, and M. Álvarez-Campana, "Security and Privacy Analysis of Youth-Oriented Connected Devices," *Sensors 2022, Vol. 22, Page 3967*, vol. 22, no. 11, p. 3967, May 2022, doi: 10.3390/S22113967.

[25] I. Ioannidou and N. Sklavos, "On General Data Protection Regulation Vulnerabilities and Privacy Issues, for Wearable Devices and Fitness Tracking Applications," *Cryptography 2021, Vol. 5, Page 29*, vol. 5, no. 4, p. 29, Oct. 2021, doi: 10.3390/CRYPTOGRAPHY5040029.

[26] T. S. Ajani, A. L. Imoize, and A. A. Atayero, "An Overview of Machine Learning within Embedded and Mobile Devices–Optimizations and Applications," *Sensors 2021, Vol. 21, Page 4412*, vol. 21, no. 13, p. 4412, Jun. 2021, doi: 10.3390/S21134412.

[27] J. Senanayake, H. Kalutarage, and M. O. Al-Kadri, "Android Mobile Malware Detection Using Machine Learning: A Systematic Review," *Electronics 2021, Vol. 10, Page 1606*, vol. 10, no. 13, p. 1606, Jul. 2021, doi: 10.3390/ELECTRONICS10131606.

[28] A. Diro, N. Chilamkurti, V. D. Nguyen, and W. Heyne, "A Comprehensive Study of Anomaly Detection Schemes in IoT Networks Using Machine Learning Algorithms," *Sensors 2021, Vol. 21, Page 8320*, vol. 21, no. 24, p. 8320, Dec. 2021, doi: 10.3390/S21248320.

[29] M. Ozkan-Okay *et al.*, "A Comprehensive Survey: Evaluating the Efficiency of Artificial Intelligence and Machine Learning Techniques on Cyber Security Solutions," *IEEE Access*, vol. 12, pp. 12229–12256, 2024, doi: 10.1109/ACCESS.2024.3355547.

[30] K. K. Verma, B. M. Singh, and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system," *International Journal of Information Technology (Singapore)*, vol. 14, no. 1, pp. 397–410, Feb. 2022, doi: 10.1007/S41870-019-00364-0/METRICS.

[31] Y. K. Saheed and M. O. Arowolo, "Efficient Cyber Attack Detection on the Internet of Medical Things-Smart Environment Based on Deep Recurrent Neural Network and Machine Learning Algorithms," *IEEE Access*, vol. 9, pp. 161546–161554, 2021, doi: 10.1109/ACCESS.2021.3128837.

[32] J. Yu, A. de Antonio, and E. Villalba-Mora, "Deep Learning (CNN, RNN) Applications for Smart Homes: A Systematic Review," *Computers 2022, Vol. 11, Page 26*, vol. 11, no. 2, p. 26, Feb. 2022, doi: 10.3390/COMPUTERS11020026.

[33] A. Nassif, M. Talib, Q. Nasir, and F. D.-I. Access, "Machine learning for anomaly detection: A systematic review," *ieeexplore.ieee.org*, 2021, Accessed: Oct. 30, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9439459/

[34] S. Devineni, S. Kathiriya, A. S.-J. of A. Intelligence, and undefined 2023, "Machine learning-powered anomaly detection: Enhancing data security and integrity," *scholar9.com*, vol. 2, no. 2, pp. 1–9, 2023, doi: 10.47363/JAICC/2023(2)184.

[35] H. Abubaker *et al.*, "Machine Learning Approaches for Malware Classification in Android Platform: A Review," *semarakilmu.com.my*, vol. 48, pp. 248–268, 2024, doi: 10.37934/araset.48.1.248268.

[36] I. Almomani, T. Almashat, and W. E.-S.-I. Access, "Maloid-DS: Labeled Dataset for Android Malware Forensics," *ieeexplore.ieee.org*, 2024, Accessed: Oct. 30, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10529242/

[37] S. Abbas *et al.*, "An efficient deep recurrent neural network for detection of cyberattacks in realistic IoT environment," *Journal of Supercomputing*, vol. 80, no. 10, pp. 13557–13575, Jul. 2024, doi: 10.1007/S11227-024-05993-2/METRICS.

[38] N. Al-Taleb and N. A. Saqib, "Towards a Hybrid Machine Learning Model for Intelligent Cyber Threat Identification in Smart City Environments," *Applied Sciences 2022, Vol. 12, Page 1863*, vol. 12, no. 4, p. 1863, Feb. 2022, doi: 10.3390/APP12041863.

[39] I. T. A. Z. Kuzior A, "Cybersecurity and cybercrime: Current trends and threats," *Journal of International Studies*, vol. 17, no. 2, pp. 220–239, 2024.

[40] A. Cinar and TB Kara, "The current state and future of mobile security in the light of the recent mobile security threat reports," *SpringerAC Cinar, TB KaraMultimedia Tools and Applications, 2023•Springer*, 2023, Accessed: Oct. 30, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11042-023-14400-6

[41] H. Hua, W. Li, Y. Li, T. Wang, N. Dong, and J. Cao, "Edge Computing with Artificial Intelligence: A Machine Learning Perspective," *ACM Comput Surv*, vol. 1, no. 9, 2023, doi: 10.1145/3555802.

[42] M. Savic *et al.*, "Deep Learning Anomaly Detection for Cellular IoT with Applications in Smart Logistics," *IEEE Access*, vol. 9, pp. 59406–59419, 2021, doi: 10.1109/ACCESS.2021.3072916.

[43] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," *Appl Energy*, vol. 287, p. 116601, Apr. 2021, doi: 10.1016/J.APENERGY.2021.116601.

[44] A. Razaque, N. Shaldanbayeva, B. Alotaibi, M. Alotaibi, A. Murat, and A. Alotaibi, "Big Data Handling Approach for Unauthorized Cloud Computing Access," *Electronics 2022, Vol. 11, Page 137*, vol. 11, no. 1, p. 137, Jan. 2022, doi: 10.3390/ELECTRONICS11010137.

[45] V. Vasani, A. K. Bairwa, S. Joshi, A. Pljonkin, M. Kaur, and M. Amoon, "Comprehensive Analysis of Advanced Techniques and Vital Tools for Detecting Malware Intrusion," *Electronics 2023, Vol. 12, Page 4299*, vol. 12, no. 20, p. 4299, Oct. 2023, doi: 10.3390/ELECTRONICS12204299.

p-ISSN: 2803-1507 e-ISSN: 2803-1531

[46] C. S. Yadav *et al.*, "Malware Analysis in IoT & Android Systems with Defensive Mechanism," *Electronics 2022, Vol. 11, Page 2354*, vol. 11, no. 15, p. 2354, Jul. 2022, doi: 10.3390/ELECTRONICS11152354.

[47] F. Skopik, M. Wurzenberger, G. Hold, M. Landauer, and W. Kuhn, "Behavior-Based Anomaly Detection in Log Data of Physical Access Control Systems," *IEEE Trans Dependable Secure Comput*, vol. 20, no. 4, pp. 3158–3175, Jul. 2022, doi: 10.1109/TDSC.2022.3197265.

[48] W. L. Rice and B. Pan, "Understanding changes in park visitation during the COVID-19 pandemic: A spatial application of big data," *Wellbeing, Space and Society*, vol. 2, p. 100037, Jan. 2021, doi: 10.1016/J.WSS.2021.100037.

[49] S. F. Chevtchenko *et al.*, "Anomaly Detection in Industrial Machinery Using IoT Devices and Machine Learning: A Systematic Mapping," *IEEE Access*, vol. 11, pp. 128288–128305, 2023, doi: 10.1109/ACCESS.2023.3333242.

[50] C. Ricciardi *et al.*, "Detection of Suspicious Cardiotocographic Recordings by Means of a Machine Learning Classifier," *Bioengineering 2023, Vol. 10, Page 252*, vol. 10, no. 2, p. 252, Feb. 2023, doi: 10.3390/BIOENGINEERING10020252.

[51] S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," *Electric Power Systems Research*, vol. 192, p. 106904, Mar. 2021, doi: 10.1016/J.EPSR.2020.106904.

[52] B. Vrigazova, "The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems," *Business Systems Research : International journal of the Society for Advancing Innovation and Research in Economy*, vol. 12, no. 1, pp. 228–242, May 2021, Accessed: Nov. 05, 2024. [Online]. Available: https://hrcak.srce.hr/ojs/index.php/bsr/article/view/17347

[53] S. Nayak and M. R.-I. J. of Science, "Impact of Different Outlier Handling Techniques on GAN Based Hybrid Bankruptcy Prediction Models," *sciresol.s3.us-east-2.amazonaws ...*, 2024, Accessed: Nov. 06, 2024. [Online]. Available: https://sciresol.s3.us-east-2.amazonaws.com/IJST/Articles/2024/Issue-4/IJST-2023-649.pdf

[54] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," *Int J Comput Vis*, vol. 129, no. 4, pp. 1038–1059, Apr. 2021, doi: 10.1007/S11263-020-01400-4/TABLES/8.

[55] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimed Tools Appl*, vol. 80, no. 11, pp. 16979–16995, May 2021, doi: 10.1007/S11042-020-09406-3/METRICS.

[56] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "Fault Detection of Bearing: An Unsupervised Machine Learning Approach Exploiting Feature Extraction and Dimensionality Reduction," *Informatics 2021, Vol. 8, Page 85*, vol. 8, no. 4, p. 85, Nov. 2021, doi: 10.3390/INFORMATICS8040085.

[57] B. M. S. Hasan and A. M. Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, Apr. 2021, Accessed: Nov. 06, 2024. [Online]. Available: https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032

[58] A. K. P. Anil and U. K. Singh, "An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models," *Journal of Systems Engineering and Information Technology (JOSEIT)*, vol. 2, no. 2, pp. 77–84, Oct. 2023, doi: 10.29207/JOSEIT.V2I2.5460.

[59] U. Islam *et al.*, "Investigating the Effectiveness of Novel Support Vector Neural Network for Anomaly Detection in Digital Forensics Data," *Sensors 2023, Vol. 23, Page 5626*, vol. 23, no. 12, p. 5626, Jun. 2023, doi: 10.3390/S23125626.

[60] N. Naz *et al.*, "Ensemble learning-based IDS for sensors telemetry data in IoT networks," *Mathematical Biosciences and Engineering*, vol. 19, no. 10, pp. 10550–10580, Jul. 2022, doi: 10.3934/MBE.2022493.

p-ISSN: 2803-1507 e-ISSN: 2803-1531

[61] M. Jiménez-Buedo and F. Russo, "Experimental practices and objectivity in the social sciences: re-embedding construct validity in the internal–external validity distinction," *Synthese*, vol. 199, no. 3–4, pp. 9549–9579, Dec. 2021, doi: 10.1007/S11229-021-03215-3/FIGURES/1.

[62] S. Hussain *et al.*, "A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection," *Energy Reports*, vol. 7, pp. 4425–4436, Nov. 2021, doi: 10.1016/J.EGYR.2021.07.008.

[63] S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," *Electric Power Systems Research*, vol. 192, p. 106904, Mar. 2021, doi: 10.1016/J.EPSR.2020.106904.

[64] P. Thölke *et al.*, "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data," *Neuroimage*, vol. 277, p. 120253, Aug. 2023, doi: 10.1016/J.NEUROIMAGE.2023.120253.

[65] J. Tanimu, S. Shiaeles, and M. Adda, "A Comparative Analysis of Feature Eliminator Methods to Improve Machine Learning Phishing Detection," *Journal of Data Science and Intelligent Systems*, vol. 2, no. 2, pp. 87–99, Dec. 2024, doi: 10.47852/BONVIEWJDSIS32021736.

[66] I. H. Ji, J. H. Lee, M. J. Kang, W. J. Park, S. H. Jeon, and J. T. Seo, "Artificial Intelligence-Based Anomaly Detection Technology over Encrypted Traffic: A Systematic Literature Review," *Sensors 2024, Vol. 24, Page 898*, vol. 24, no. 3, p. 898, Jan. 2024, doi: 10.3390/S24030898.