BAB II

TINJAUAN PUSTAKA

2.1 Machine Learning

Machine Learning merupakan bagian dari bidang kecerdasan buatan yang memungkinkan komputer untuk belajar dan bertindak seperti manusia. Proses ini juga memungkinkan komputer untuk meningkatkan kemampuannya secara otomatis melalui pengalaman, dengan menerima data dan informasi dari pengamatan serta interaksi langsung dengan dunia nyata (Mahesh, 2020).

Machine Learning adalah proses komputasi yang memanfaatkan data input untuk menyelesaikan tugas tertentu tanpa harus diprogram secara manual. Algoritma yang digunakan bersifat "*soft coded*," yang artinya dapat mengubah atau menyesuaikan struktur internalnya secara otomatis melalui proses berulang. Proses ini dikenal sebagai pelatihan (*training*), di mana data *input* dan *output* yang diharapkan diberikan kepada algoritma untuk dipelajari. Kemampuan *training* yang dimiliki memungkinkan algoritma untuk mengoptimalkan konfigurasinya sehingga tidak hanya dapat menghasilkan hasil yang diinginkan pada data *training*, tetapi juga mampu memproses data baru yang belum pernah dilihat sebelumnya dengan hasil yang sesuai (Naqa and Murphy, 2015).

Di era industri 4.0, dunia digital saat ini menyimpan sejumlah besar data, seperti data dari perangkat *IoT*, media sosial, kesehatan, dan lainnya. Salah satu pendekatan untuk menganalisis data dalam jumlah besar tersebut adalah dengan menggunakan aplikasi berbasis kecerdasan buatan, khususnya *Machine Learning*.

Machine Learning terdiri dari kumpulan algoritma dan model statistik yang dirancang untuk menyelesaikan tugas tertentu tanpa perlu instruksi pemrograman eksplisit (Mahesh, 2020). Algoritma *Machine Learning* secara umum dapat diklasifikasikan ke dalam empat kategori utama, yaitu *supervised*, *unsupervised*, *semi-supervised*, dan *reinforcement* (Mohammed, dkk., 2016).

Efektivitas dan efisiensi dari *Machine Learning* bergantung pada karakteristik data yang digunakan selama pelatihan, serta kinerja algoritma yang diterapkan. *Machine Learning* memiliki berbagai tugas yang bisa dilakukan seperti klasifikasi, pengelompokan (*clustering*), regresi, *feature engineering*, pengurangan dimensi, pembelajaran aturan asosiatif, dan pembelajaran *reinforcement* dirancang untuk memanfaatkan data secara optimal (Mohammed, dkk., 2016).

2.1.1 *Supervised Learning*

Supervised learning adalah salah satu jenis algoritma *Machine Learning* yang memungkinkan mesin mempelajari fungsi yang menghubungkan *input* dengan *output* berdasarkan contoh data *input-output* (Mahesh, 2020). Secara sederhana, *supervised learning* merupakan metode pembelajaran menggunakan contoh data yang sudah diberi label. Metode ini memanfaatkan data *training* berlabel dalam jumlah besar untuk memprediksi suatu fungsi. Supervised learning cocok digunakan ketika tujuan tertentu telah ditentukan dan dapat dicapai melalui serangkaian *input* atau contoh (Sarker dkk., 2020).

2.1.1.1 *Ensemble Learning*

Ensemble Learning adalah teknik *supervised learning* di mana beberapa model dilatih secara terpisah dan kemudian digabungkan untuk meningkatkan

performa secara keseluruhan (Ennaji, dkk., 2021). Pendekatan ini melibatkan pengelolaan *bias* dan *variance*.

Metode *ensemble* bisa dikelompokkan dalam beberapa macam tipe. Tipe tersebut terdapat tiga jenisnya, yaitu :

1. *Bagging*

Bagging adalah metode yang menggunakan beberapa model, masing-masing dilatih dengan subset data yang berbeda. Hasil prediksi dari setiap model kemudian dirata-rata untuk menghasilkan *output* akhir. Teknik ini sangat efektif dalam mengurangi *variance* hasil dan membantu mengatasi masalah *overfitting*.

2. *Boosting*

Boosting bekerja dengan menggabungkan sejumlah model sederhana yang disebut sebagai model lemah. Proses ini berlangsung dalam beberapa iterasi, di mana setiap model baru dilatih untuk memperbaiki kesalahan yang dibuat oleh model sebelumnya. Hasil akhirnya adalah model yang lebih kuat dengan tingkat akurasi prediksi yang lebih baik. Selain itu, metode ini lebih efektif dalam mengurangi bias dibandingkan *bagging*.

3. *Stacking* (*blending*)

Stacking menggunakan pendekatan *ensemble* dua tingkat. Pada tahap pertama, beberapa model dilatih secara terpisah menggunakan data *training*. Hasil prediksi dari semua algoritma tersebut kemudian digabungkan untuk menghasilkan prediksi akhir yang lebih akurat pada data baru.

2.1.2 *Classification Model*

Classification adalah salah satu jenis *supervised learning* yang berfokus pada tugas prediktif di mana model mencoba menentukan label dari data berdasarkan contoh yang diberikan. Model ini bekerja dengan memetakan fungsi dari variabel *input* ke *output*, yang biasanya berupa target, label, atau kategori. Berikut ini adalah beberapa algoritma *Machine Learning* untuk klasifikasi yang sering digunakan dalam pendekatan *ensemble* :

1. AdaBoost

AdaBoost, singkatan dari *Adaptive Boosting*, merupakan algoritma *ensemble* yang membangun serangkaian *Decision Tree* berbobot. Biasanya, *Decision Tree* yang digunakan berukuran kecil (sering kali berupa *stump* dengan satu tingkat). Setiap *Decision Tree* dilatih menggunakan seluruh *dataset*, tetapi bobot pada data sampel diubah secara adaptif untuk lebih menekankan pada data yang salah diklasifikasikan pada iterasi sebelumnya (Baladram, 2024).

2. *Gradient Boosting*

Gradient Boosting bekerja dengan membangun model secara bertahap pada *dataset training*. Setiap model berikutnya dilatih untuk memperbaiki kesalahan yang dibuat oleh model sebelumnya dengan cara mengurangi *error* dengan meminimalkan *loss function* dari model yang telah dilatih sebelumnya. Proses pelatihan model ini diulangi hingga kesalahan model menjadi seminimal mungkin, sehingga prediksi menjadi lebih akurat. Algoritma ini menggabungkan beberapa *weak learner* untuk membentuk satu *strong learner* yang lebih andal (Anshul, 2024).

3. *Random Forest*

Random Forest adalah algoritma tipe *ensemble* yang menggunakan teknik *parallel ensembling*. Algoritma ini menggabungkan beberapa *Decision Tree* untuk meningkatkan akurasi prediksi, meminimalkan risiko *overfitting*, dan memberikan kontrol yang lebih baik terhadap hasil (Pedregosa dkk., 2011).

2.1.3 *Feature Engineering dan Feature Selection*

Machine Learning dan *Data Science* memiliki tantangan besar berupa pengolahan data berdimensi tinggi. *Feature selection* dapat mengurangi dimensi data dan dapat mempermudah interpretasi data, menurunkan biaya komputasi, dan mengurangi risiko *overfitting* serta redundansi dengan model yang lebih sederhana. *Feature engineering* juga dapat mengurangi dan menambah dimensi data, bisa disesuaikan dengan tujuan yang ingin dicapai.

Feature selection adalah proses memilih fitur atau variabel prediktif yang relevan untuk membangun model *Machine Learning*. Proses ini mengurangi kompleksitas model dengan menghilangkan fitur-fitur yang tidak relevan atau kurang penting, sehingga dapat mempercepat proses pelatihan model.

Feature engineering adalah proses untuk mengurangi jumlah fitur atau variabel prediktif dengan menciptakan fitur baru dari fitur yang ada. Setelah fitur baru dibuat, fitur lama yang tidak lagi diperlukan biasanya dihapus dari *dataset* (Sarker dkk., 2020).

2.2 *Internet of Things* (IoT)

Internet of Things (*IoT*) adalah paradigma baru yang memungkinkan perangkat elektronik saling terhubung melalui internet untuk bertukar informasi

secara otomatis. Penelitian dan pengembangan terus dilakukan untuk meningkatkan teknologi yang berbasis *IoT*. Perangkat *IoT* dirancang dengan prinsip skalabilitas, modularitas, interoperabilitas, dan keterbukaan. Faktor-faktor ini menjadi kunci utama dalam menciptakan arsitektur *IoT* yang efisien, terutama dalam lingkungan yang beragam dan heterogen (Kumar, dkk., 2019).

Salah satu tantangan utama dalam pengembangan perangkat *IoT* adalah memastikan keamanan dan privasi, karena perangkat ini sangat rentan terhadap ancaman siber. Kerentanan ini sering kali disebabkan oleh kurangnya mekanisme otentikasi dan otorisasi, penggunaan perangkat lunak dan *firmware* yang tidak aman, antarmuka web yang rentan, serta enkripsi yang tidak memadai pada lapisan transportasi data (Babovic, dkk., 2016).

2.3 Forensik Digital

Forensik digital adalah proses sistematis yang bertujuan untuk memperoleh dan mengamankan bukti digital dalam bentuk aslinya. Proses ini melibatkan pengumpulan, identifikasi, dan validasi informasi digital guna merekonstruksi aktivitas yang terjadi di masa lalu. Dalam setiap tahapannya, bukti yang diperoleh harus dijaga integritasnya dan dilengkapi dengan dokumentasi rinci yang mencatat lokasi asal bukti hingga saat bukti digunakan dalam proses hukum di pengadilan (Mukherjee dan Haque, 2018).

2.4 Penelitian Terkait

Terdapat banyak penelitian yang membahas topik terkait *Machine Learning* dan forensik digital. Penelitian-penelitian tersebut dapat dijadikan acuan dan panduan untuk membantu memperlancar proses penelitian ini.

Penelitian pertama berjudul *AD-IoT: Anomaly Detection of IoT Cyberattacks Smart City Using Machine Learning* mengimplementasikan algoritma *Random Forest* dalam pengembangan kerangka kerja bernama *Anomaly Detection-IoT* (*AD-IoT*) yang berfungsi untuk mendeteksi anomali. Kerangka ini terdiri atas tiga lapisan, yaitu *IoT layer*, *Fog layer*, dan *Cloud layer*, di mana proses pelatihan dan pengujian model dilakukan pada *Fog layer*. Pada tahap pra-pemrosesan, dilakukan seleksi fitur menggunakan algoritma *ExtraTreesClassifier* yang mampu memilih variabel-variabel penting dari suatu *dataset*. Model yang dibangun diarahkan untuk melakukan klasifikasi biner, yakni memprediksi antara trafik *benign* dan *malicious*, dengan tingkat akurasi sebesar 99,34% (Ahmad dkk., 2021).

Penelitian kedua yang berjudul *Cyberattacks Detection in IoT-Based Smart City Applications Using Machine Learning Techniques* menerapkan beberapa algoritma untuk melakukan pelatihan model, dengan tujuan membandingkan performa masing-masing algoritma pada dua *dataset* yang berbeda, yaitu UNSW-NB15 dan CICIDS2017. Teknik *cross-validation* pada tahap praproses berfungsi untuk memilih variabel. Tugas klasifikasi dalam penelitian ini bersifat *multiclass classification*, dengan algoritma yang digunakan meliputi LR, SVM, DT, *Random Forest*, ANN, dan KNN. Tingkat akurasi yang dicapai oleh masing-masing algoritma pada *dataset* UNSW-NB15 berturut-turut adalah 72,32%, 71,49%, 80,69%, 81,77%, 78,89%, dan 78,23%, sedangkan pada *dataset* CICIDS2017 adalah 93,60%, 92%, 99,7%, 99,7%, 94,2%, dan 99,7% (Rashid dkk., 2020).

Penelitian ketiga yang berjudul *Hybrid Intelligent Intrusion Detection System for Internet of Things* mengimplementasikan dua algoritma *deep learning*,

yakni CNN dan LSTM. Penelitian ini memanfaatkan dua *dataset* untuk menguji adaptasi model, yaitu UNSW-NB15 dan NSL-botnet. Salah satu kontribusi penting dari penelitian ini adalah rendahnya nilai *False-Positive Rate* (FPR), yang merupakan metrik krusial dalam evaluasi model yang dilatih dengan dataset tidak seimbang (*imbalanced*). Rata-rata nilai FPR yang diperoleh adalah 0,76 untuk *binary classification* dan 0,53 untuk *multiclass classification* (Basar dan Wang, 2020).

Penelitian keempat yang berjudul *Realguard: A Lightweight Network Intrusion Detection System for IoT Gateways* menggunakan algoritma *deep learning*, yaitu *Convolutional Neural Network (CNN)*, dalam pengembangan model deteksinya. Model dilatih menggunakan *dataset* CICIDS2017 dan menerapkan teknik *K-cross validation* untuk memilih fitur terbaik. Proses *feature selection* ini memerlukan waktu yang cukup lama karena pelatihan model dilakukan secara iteratif pada subset variabel dari *dataset*. Model yang dikembangkan untuk tugas *multiclass classification* ini memperoleh tingkat akurasi sebesar 95,85% (Gyamfi dan Jurcut, 2022).

Penelitian kelima yang berjudul *An Ensemble of Deep Recurrent Neural Networks for Detecting IoT Cyber Attacks Using Network Traffic* mengembangkan sistem *ensemble* yang terdiri atas beberapa model kombinasi CNN+LSTM yang diintegrasikan dengan algoritma *Random Forest*. Sistem ini menggantikan penggunaan *Decision Tree* sebagai *weak learner* dalam struktur *Random Forest*. *Dataset* yang digunakan untuk pelatihan model adalah *Modbus*, yang diekstrak menggunakan program *CICFlowMeter*. Model yang dibangun berhasil mencapai tingkat akurasi sebesar 99% (Saharkhizan dkk., 2020).

Penelitian keenam yang berjudul *A Novel Wide & Deep Transfer Learning Stacked GRU Framework for Network Intrusion Detection* melatih model *deep learning* berbasis algoritma GRU (*Gated Recurrent Unit*). Dua *dataset* digunakan, yaitu UNSW-NB15 dan 10% subset dari KDDCup 99, dua *dataset* tersebut digunakan untuk keperluan komparasi performa. Seluruh model yang dikembangkan untuk menyelesaikan tugas *multiclass classification*. Model yang dilatih menggunakan *dataset* KDDCup 99 berhasil mencapai akurasi sebesar 99,92%, sedangkan model yang menggunakan *dataset* UNSW-NB15 memperoleh akurasi sebesar 94,22%. Model yang dikembangkan memiliki kemampuan *memorization* setara dengan model *regression* serta kapasitas *generalization* khas dari arsitektur GRU (Singh dkk., 2021).

Penelitian ketujuh yang berjudul *A Machine Learning Security Framework for IoT Systems* mengembangkan sebuah kerangka kerja berbasis *Machine Learning* yang bertujuan untuk memperluas aspek keamanan pada sistem *IoT*. Kerangka ini memanfaatkan teknologi *Software Defined Networking* (SDN) dan *Network Function Virtualization* (NFV). *Dataset* yang digunakan dalam pelatihan model adalah NSL-KDD, dan tahapan praproses data mencakup penerapan teknik *discretization*, yang berfungsi untuk mengubah nilai kontinyu menjadi diskrit guna mengurangi kardinalitas. Model yang dibangun untuk menyelesaikan tugas *multiclass classification* dan berhasil mencapai tingkat akurasi sebesar 99,71% (Bagaa dkk., 2020).

Penelitian kedelapan yang berjudul *An Efficient Deep-Learning-Based Detection and Classification System for Cyber-Attacks in IoT Communication Networks* menghasilkan sebuah sistem deteksi dan klasifikasi serangan siber

berbasis *deep learning* yang dinamakan *IoT-IDCS-CNN*. Model ini dilatih menggunakan *dataset* NSL-KDD, dengan penyesuaian label sesuai dengan jenis tugas klasifikasi yang dijalankan (*binary* dan *multiclass*). Tahapan praproses data dilakukan menggunakan teknik K-*cross validation*. Model yang dilatih untuk tugas *binary classification* memperoleh akurasi sebesar 99,3%, sedangkan model untuk tugas *multiclass classification* mencapai akurasi sebesar 98,2% (Al-Haija dan Zein-Sabatto, 2020).

Penelitian kesembilan yang berjudul *Intrusion Detection Systems Using Long Short-Term Memory (LSTM)* mengembangkan model berbasis *deep learning* dengan mengintegrasikan algoritma CNN dan LSTM untuk menangani tugas klasifikasi *binary* dan *multiclass*. Teknik *Principal Component Analysis* (PCA) dan *Mutual Information* digunakan pada tahap praproses data untuk seleksi fitur, di mana PCA menyederhanakan *dataset* menjadi dua komponen utama. Model untuk klasifikasi *binary* berhasil mencapai akurasi sebesar 99,44%, sedangkan model untuk klasifikasi *multiclass* memperoleh akurasi sebesar 99,39% (Laghrissi dkk., 2021).

Penelitian kesepuluh yang berjudul *Anomaly Detection Using Deep Neural Network for IoT Architecture* mengembangkan sejumlah model berbasis algoritma *deep learning*, seperti CNN, RNN, GRU, LSTM, dan DNN, untuk melakukan *benchmarking* terhadap *dataset IoT-Botnet* 2020. Teknik *Mutual Information* digunakan sebagai metode *feature selection* pada tahap praproses data. Model yang dilatih tanpa *feature selection* menghasilkan akurasi berturut-turut sebesar 98,44% (RNN), 98,88% (CNN), 98,39% (GRU), 94,41% (LSTM), dan 99,01% (DNN). Sementara itu, model yang dilatih menggunakan 32 fitur hasil *feature selection*

dengan *Mutual Information* menunjukkan akurasi sebesar 98,31% (RNN), 98,68% (CNN), 98,31% (GRU), 96,26% (LSTM), dan 98,75% (DNN) (Ahmad dkk., 2021b).

Penelitian kesebelas yang berjudul *Intrusion Detection in Internet of Things Using Supervised Machine Learning Based on Application and Transport Layer Features Using* UNSW-NB15 *Dataset* mengembangkan model deteksi berbasis algoritma *Random Forest* menggunakan aplikasi Weka. *Dataset* yang digunakan adalah UNSW-NB15, dan teknik *undersampling* diterapkan untuk mengatasi permasalahan ketidakseimbangan data (*imbalanced dataset*). *Feature selection* dilakukan dengan metode *ranking* berdasarkan nilai *Information Gain*, serta beberapa variabel tambahan dipilih berdasarkan hasil klasterisasi yang disusun melalui *knowledge base* para peneliti, dengan *Flow*, TCP, dan MQTT sebagai kelompok utama (Ahmad dkk, 2021a).

Penelitian kedua belas yang berjudul *A Machine Learning Approach for Improving the Performance of Network Intrusion Detection Systems* mengembangkan tiga algoritma *Machine Learning*, yaitu *Random Forest*, *Support Vector Machine* (SVM), dan DJ. Evaluasi performa dilakukan dengan menerapkan 10 variasi *train-test split ratio* pada masing-masing model, di mana setiap rasio disertai tahap praproses data yang mengikuti metodologi KDD. Berdasarkan hasil evaluasi menggunakan *dataset* CICIDS2017, nilai akurasi rata-rata yang diperoleh dari model SVM, *Random Forest*, dan DJ secara berturut-turut adalah 98,18%, 96,76%, dan 96,50%, sedangkan nilai *recall* rata-rata masing-masing adalah 95,63%, 97,62%, dan 95,77% (Azizan dkk., 2021).

Penelitian ketiga belas yang berjudul *IoT-Fog-Cloud Model for Anomaly Detection Using Improved Naïve Bayes and Principal Component Analysis* mengembangkan sistem deteksi intrusi berbasis anomali dengan memanfaatkan metode *Improved Naïve Bayes* (INB) yang dikombinasikan dengan teknik *Principal Component Analysis* (PCA) untuk proses *feature selection*. Model yang dilatih dengan UNSW-NB15 mampu mencapai akurasi sebesar 92,48% dan tingkat deteksi sebesar 95,35%, yang menunjukkan peningkatan efisiensi serta performa dalam mengidentifikasi serangan siber. Kerangka kerja *IoT-Fog-Cloud* ini dirancang untuk mendukung aplikasi *smart city*, dengan fokus utama pada isu keamanan dan privasi sistem *IoT*, serta memiliki kemampuan untuk melakukan klasifikasi terhadap 9 jenis serangan (Manimurugan, 2021).

Penelitian keempat belas yang berjudul *A Smart Anomaly-Based Intrusion Detection System for the Internet of Things (IoT) Network Using GWO–PSO–RF Model* mengusulkan sistem *Intrusion Detection System* (IDS) berbasis kecerdasan buatan untuk jaringan *IoT* yang dirancang guna mengatasi keterbatasan akurasi dan lambatnya waktu deteksi pada sistem IDS tradisional. Model yang dikembangkan memanfaatkan kombinasi teknik optimasi *Grey Wolf Optimization* (GWO) dan *Particle Swarm Optimization* (PSO) untuk melakukan tahapan *feature selection*, yang kemudian diklasifikasikan menggunakan algoritma *Random Forest*. Evaluasi dilakukan terhadap tiga *dataset*, yaitu KDDCup 99, NSL-KDD, dan CICIDS2017, dengan hasil akurasi rata-rata sebesar 99,66% pada tugas klasifikasi *multiclass*. Penggunaan teknik GWO dan PSO juga berfungsi sebagai pendekatan untuk mengatasi permasalahan ketidakseimbangan data (*imbalanced dataset*) (Keserwani dkk., 2021).

Penelitian kelima belas yang berjudul *Multi-Stage Optimized Machine Learning Framework for Network Intrusion Detection* mengusulkan kerangka kerja NIDS berbasis *Machine Learning* yang telah dioptimalkan guna menurunkan kompleksitas komputasi tanpa mengorbankan performa deteksi. Studi ini mengevaluasi efektivitas teknik *oversampling* terhadap ukuran data pelatihan dan membandingkan dua pendekatan *feature selection*: *Information Gain* dan korelasi. Selain itu, dilakukan optimasi *hyperparameter* untuk meningkatkan kinerja model. Evaluasi terhadap *dataset* CICIDS2017 dan UNSW-NB15 menunjukkan bahwa model ini mampu mengurangi ukuran data *training* hingga 74% dan jumlah variabel hingga 50%, dengan tingkat akurasi deteksi melebihi 99%, serta berhasil melampaui studi-studi terdahulu dalam hal akurasi dan tingkat *false alarm* (Injadat dkk., 2021).

Penelitian keenam belas yang berjudul *Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning* melatih beberapa model algoritma, salah satunya adalah model yang dilatih dengan algoritma *Random Forest*. Penelitian ini menggunakan *dataset* NSL-KDD dan CICIDS2018 untuk melatih model. Teknik *Difficult Set Sampling Technique* (DSST) digunakan dalam tahapan praproses data untuk menangani permasalahan *dataset* yang *unblanced* (Liu dkk., 2021).

Tabel II.1 Penelitian Terkait

No.	Penulis	Judul	Algoritma	Hasil
1.	(Ahmad dkk., 2021)	*AD-IoT: Anomaly Detection of IoT Cyberattacks in Smart*	*Random Forest*	Membuat sistem bernama *Anomaly Detection-IoT* (*AD-IoT*) yang

No.	Penulis	Judul	Algoritma	Hasil
		City Using Machine Learning		dapat mendeteksi anomali dengan algoritma *Random Forest*. Sistem ini mendapat akurasi tertinggi sebesar 99.34%
2.	(Rashid dkk., 2020)	*Cyberattacks Detection in IoT-Based Smart City Applications Using Machine Learning Techniques*	LR, SVM, DT, *Random Forest*, ANN dan KNN	Menggunakan 2 *dataset* yaitu UNSW-NB15 dan CICIDS2017. Pada *dataset* UNSW-BC15 akurasi yang didapatkan dari algoritma LR, SVM, DT, *Random Forest*, ANN dan KNN secara berturut-turut adalah 72.32%, 71.49%, 80.69%, 81.77%, 78.89% dan 78.23%. Sedangkan pada *dataset* CICIDS2017 adalah 93.60%, 92%, 99.7%, 99.7%, 94.2% dan 99.7%.
3.	(Basar dan Wang, 2020)	*Hybrid Intrusion Detection System for Internet of Things (IoT)*	*Deep Learning*	Penelitian ini mengajukan model *Deep Learning* yang diberi nama HCNN untuk mendeteksi intrusi (dijadikan *Intrusion Detection System*). Penelitian ini juga melakukan komparasi HCNN dengan RNN, tiap model dilatih dengan dataset UNSW-NB15 dimana RNN mendapat akurasi 95.7% sedangkan

No.	Penulis	Judul	Algoritma	Hasil
				HCNN mendapatkan 98.6%.
4.	(Gyamfi dan Jurcut, 2022)	*Realguard: A Lightweight Network Intrusion Detection System for IoT Gateways*	CNN	Model CNN yang dilatih dengan *dataset* hasil praproses dengan *K-cross validation* ini dikembangkan untuk tugas klasifikasi *multiclass* dan memperoleh tingkat akurasi sebesar 95,85%
5.	(Saharkhizan dkk., 2020)	*An Ensemble of Deep Recurrent Neural Networks for Detecting IoT Cyber Attacks Using Network Traffic*	*Decision Tree*	Model diintegrasikan dengan *Long Short-Term Memory* (LSTM), model dilatih dengan *Modbus network traffic dataset* dan bisa mendapatkan 99% akurasi.
6.	(Singh dkk., 2021)	*A novel wide & deeptransfer learning stacked GRU framework for network intrusion detection*	GRU	Model algoritma *Deep Learning* yang berupa GRU. Dua *dataset* digunakan untuk melakukan komparasi, dua *dataset* tersebut adalah UNSW-NB15 dan 10% data dari KDDCup 99. Model-model yang dilatih ini akan melakukan tugas klasifikasi *multiclass*. Model yang dilatih dengan *dataset* KDDCup 99 mendapatkan akurasi sebesar 99,92% dan model yang dilatih

No.	Penulis	Judul	Algoritma	Hasil
				dengan UNSW-NB15 mendapatkan akurasi sebesar 94.22%.
7.	(Bagaa dkk., 2020)	*A Machine Learning Security Framework for Iot Systems*	SVM	Membuat sebuah *framework* yang dikombinasikan dengan *Support Vector Machine* (SVM) dan berhasil mendapatkan akurasi 99.71% dalam mengklasifikasi anomali.
8.	(Al-Haija dan Zein-Sabatto, 2020)	*An Efficient Deep-Learning-Based Detection and Classification System for Cyber-Attacks in IoT Communication Networks*	*Deep Learning*	Membuat model *Deep Learning* yang diberi nama *IoT Communication Networks that Leverage the Power of Convolutional Neural Network* (*IoT-IDCS-CNN*), model ini dilatih dengan *dataset* NSL-KDD dan mendapatkan akurasi paling tinggi sebesar 99.3%.
9.	(Laghrissi dkk., 2021)	*Intrusion detection systems using long short-term memory (LSTM)*	CNN	Model dengan algoritma D*eep Learning* yaitu CNN dan LSTM untuk melakukan klasifikasi biner dan multi kelas. PCA dan *Mutual Information* digunakan sebagai teknik untuk melakukan *feature selection* dalam proses praproses

No.	Penulis	Judul	Algoritma	Hasil
				data. Model dengan tugas klasifikasi *binary* mendapatkan akurasi sebesar 99,44% dan akurasi untuk klasifikasi *multiclass* sebesar 99,39%.
10.	(Ahmad dkk., 2021b)	*Anomaly Detection Using Deep Neural Network for IoT Architecture*	*CNN, RNN, GRU, LSTM, DNN*	Model yang dilatih dengan 32 variabel (hasil dari *feature selection* menggunakan teknik *Mutual Information*) terhadap algoritma RNN, CNN, GRU, LTSM, dan DNN didapatkan akurasi berturut-turut sebesar 98,31%, 98,68%, 98,31%, 96,26%, dan 98,75%.
11.	(Ahmad dkk, 2021a)	*Intrusion detection in internet of things using supervised machine learning based on application and transport layer features using UNSW-NB15 data-set*	*Random Forest*	*Undersampling* dilakukan untuk mengatasi permasalahan *dataset unbalanced* dan I*nformation Gain* digunakan sebagai teknik *feature selection*. Beberapa fitur juga dipilih berdasarkan *knowledge base* seperti variabel Flow, TCP dan MQQT dijadikan sebagai *cluster*.
12.	(Azizan dkk., 2021)	*A Machine Learning Approach for Improving the Performance of*	*Random Forest*, DJ, SVM	Melakukan perbandingan performa berdasarkan 10 *train-test split*

No.	Penulis	Judul	Algoritma	Hasil
		Network Intrusion Detection Systems		*ratio* terhadap tiap model dan tiap *ratio* dilakukan praproses data berdasarkan metodologi KDD. Dari model SVM, *Random Forest* dan DJ berturut-turut didapatkan akurasi sebesar 98,18%, 96,76%, dan 96,50% yang merupakan rata-rata performa dari 10 *ratio* yang dilakukan terhadap *dataset*.
13.	(Manimurugan, 2021)	*IoT-Fog-Cloud model for anomaly detection using improved Naïve Bayes and principal component analysis*	*INB*	Teknik *Principal Component Analysis* (PCA) digunakan untuk melakukan *feature selection* terhadap *dataset* UNSW-NB15 untuk melatih model INB dan akurasi yang didapatkan mencapai 92,48%.
14.	(Keserwani dkk., 2021)	*A smart anomaly-based intrusion detection system for the Internet of Things (IoT) network using GWO–PSO–RF model*	*Random Forest*	Model *Random Forest* ini dilatih dengan *dataset* yang dipilih menggunakan teknik *Grey Wolf Optimization* (GWO) dan *Particle Swarm Optimization* (PSO) terhadap 3 *dataset* NIDS dan rata-rata akurasi yang didapatkan oleh model adalah 99,66%.
15.	(Injadat dkk., 2021)	*Multi-Stage Optimized Machine Learning*	*KNN*, *Random Forest*	Mengevaluasi pengaruh teknik *oversampling*

No.	Penulis	Judul	Algoritma	Hasil
		Framework for Network Intrusion Detection		terhadap ukuran data *training* dan membandingkan dua metode *feature selection*: I*nformation Gain* dan korelasi. Selain itu, dilakukan optimasi *hyperparameter* untuk meningkatkan kinerja model. Model yang dilatih dengan KNN dan *Random Forest* semuanya mendapatkan akurasi di atas 99%.
16.	(Liu dkk., 2021)	*Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning*	*Random Forest, SVM, XGBoost, LSTM, AlexNet, Mini-VGGNet*	Membandingkan performa 6 algoritma yang dilatih (*Random Forest, SVM, XGBoost, LSTM, AlexNet, dan Mini-VGGNet*) dengan *dataset* NSL-KDD dan CICIDS2018, tidak ada model dengan performa mencolok untuk model-model yang dilatih dengan NSL-KDD, namun untuk model yang dilatih dengan CICIDS2018 didapatkan bahwa *Random Forest* mendapatkan performa paling baik untuk 4 dari 5 kali pelatihan dengan teknik praproses yang berbeda.

Berdasarkan Tabel II.1, secara keseluruhan, model-model yang dilatih untuk melakukan klasifikasi pada *dataset* NIDS sangatlah baik, tidak ada model yang mendapatkan akurasi kurang dari 90% dan model yang dilatih dengan algoritma *Random Forest* dapat menghasilkan model dengan performa paling baik. Beberapa peneletian juga menggunakan tahapan praproses sederhana dengan melibatkan teknik umum seperti PCA dan SMOTE.

2.5 Matriks Penelitian

Berdasarkan penelitian terkait yang telah dirangkum sebelumnya, dapat disimpulkan dalam bentuk matriks penelitian seperti yang ditunjukkan dalam tabel berikut.

Tabel II.2 Matriks Penelitian

No.	Penulis	Tahun	Dataset				Praproses					Tipe Klasifikasi		Algoritma Klasifikasi						*Deployment*
			KDDCup99	UNSW-NB15	CICIDS2017	CICIDS2018	*Undersampling*	*Oversampling*	PCA	*Knowledge base*	EDA	*Binary*	*Multiclass*	*Random Forest*	*AdaBoost*	*Gradient Boosting*	SVM	KNN	CNN	
1.	(Ahmad dkk., 2021)	2021		✓								✓		✓						

No.	Penulis	Tahun	Dataset				Praproses					Tipe Klasifikasi		Algoritma Klasifikasi						*Deployment*
			KDDCup99	UNSW-NB15	CICIDS2017	CICIDS2018	*Undersampling*	*Oversampling*	PCA	*Knowledge base*	EDA	*Binary*	*Multiclass*	*Random Forest*	*AdaBoost*	*Gradient Boosting*	SVM	KNN	CNN	
2.	(Rashid dkk., 2020)	2020		✓	✓								✓	✓			✓	✓		
3.	(Basar dan Wang, 2020)	2020		✓				✓				✓	✓						✓	
4.	(Gyamfi dan Jurcut, 2022)	2022			✓							✓	✓						✓	

No.	Penulis	Tahun	Dataset				Praproses					Tipe Klasifikasi		Algoritma Klasifikasi						*Deployment*
			KDDCup99	UNSW-NB15	CICIDS2017	CICIDS2018	*Undersampling*	*Oversampling*	PCA	*Knowledge base*	EDA	*Binary*	*Multiclass*	*Random Forest*	*AdaBoost*	*Gradient Boosting*	SVM	KNN	CNN	
5.	(Saharkhizan dkk., 2020)	2020											✓	✓					✓	
6.	(Singh dkk., 2021)	2021	✓	✓					✓				✓							
7.	(Bagaa dkk., 2020)	2020															✓			

No.	Penulis	Tahun	Dataset				Praproses					Tipe Klasifikasi		Algoritma Klasifikasi						*Deployment*
			KDDCup99	UNSW-NB15	CICIDS2017	CICIDS2018	*Undersampling*	*Oversampling*	PCA	*Knowledge base*	EDA	*Binary*	*Multiclass*	*Random Forest*	*AdaBoost*	*Gradient Boosting*	SVM	KNN	CNN	
8.	(Al-Haija dan Zein-Sabatto, 2020)	2020										✓	✓						✓	
9.	(Laghrissi dkk., 2021)	2021	✓						✓			✓	✓						✓	
10.	(Ahmad dkk., 2021b)	2021										✓							✓	

No.	Penulis	Tahun	Dataset				Praproses					Tipe Klasifikasi		Algoritma Klasifikasi						*Deployment*
			KDDCup99	UNSW-NB15	CICIDS2017	CICIDS2018	*Undersampling*	*Oversampling*	PCA	*Knowledge base*	EDA	*Binary*	*Multiclass*	*Random Forest*	*AdaBoost*	*Gradient Boosting*	SVM	KNN	CNN	
11.	(Ahmad dkk, 2021a)	2021		✓			✓			✓		✓		✓						
12.	(Azizan dkk., 2021)	2021			✓					✓		✓		✓			✓			
13.	(Manimurugan, 2021)	2021		✓					✓				✓							

No.	Penulis	Tahun	Dataset				Praproses					Tipe Klasifikasi		Algoritma Klasifikasi						*Deployment*
			KDDCup99	UNSW-NB15	CICIDS2017	CICIDS2018	*Undersampling*	*Oversampling*	PCA	*Knowledge base*	EDA	*Binary*	*Multiclass*	*Random Forest*	*AdaBoost*	*Gradient Boosting*	SVM	KNN	CNN	
14.	(Keserwani dkk., 2021)	2021	✓		✓								✓	✓						
15.	(Injadat dkk., 2021)	2021		✓	✓			✓				✓		✓				✓		
16.	(Liu dkk., 2021)	2021				✓							✓	✓			✓		✓	

No.	Penulis	Tahun	Dataset				Praproses					Tipe Klasifikasi		Algoritma Klasifikasi						*Deployment*
			KDDCup99	UNSW-NB15	CICIDS2017	CICIDS2018	*Undersampling*	*Oversampling*	PCA	*Knowledge base*	EDA	*Binary*	*Multiclass*	*Random Forest*	*AdaBoost*	*Gradient Boosting*	SVM	KNN	CNN	
17.	**Penelitian ini**	**2025**				✓				✓	✓		✓	✓	✓	✓				✓

Pada Tabel II.2 dapat dilihat jika yang paling mencolok adalah seluruh penelitian yang dilakukan oleh peneliti lain tidak ada yang sampai ke tahap *deployment* atau memberikan tahapan secara eksplisit bagaimana model bisa digunakan atau diimplementasikan kedalam sebuah sistem.

2.6 Matriks Penelitian Terdekat

Berdasarkan beberapa penelitian terkait, ada dua penelitian yang sangat dekat dengan penelitian ini, berikut dua penelitian terdekat yang disajikan dalam bentuk matriks penelitian terdekat:

Tabel II.3 Matriks Penelitian Terdekat

Peneliti dan tahun				**Liu dkk., 2021**	**Keserwani dkk., 2021**
Judul Naskah				*Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning*	*A smart anomaly-based intrusion detection system for the Internet of Things (IoT) network using GWO–PSO–RF model*
Ruang Lingkup	**Sistem Operasi**		**Linux**		
			Windows		
			Unix		
	Algoritma	**Supervised Learning**	**Random Forest**	✓	✓
			AdaBoost		
			Gradient Boosting		
			XGBoost	✓	
			SVM	✓	
			KNN		
		Unsupervised Learning	**K-means clustering**		
			PCA		

Peneliti dan tahun				Liu dkk., 2021	Keserwani dkk., 2021
			Neural Network		
			Hierarchical Clustering		
		Reinforcement Learning	**Q-Learning**		
			Deep Q-Network		
			Policy Gradient methods		
			Proximal Policy Optimization		
		Neural Network	**ANN**		
			CNN	✓	
			RNN		
			LSTM	✓	
			GRU		
Keterangan				Penelitian ini menggunakan algoritma *Random Forest* dan juga *dataset* CICIDS2018. Dengan teknik praproses data yang sederhana yaitu data cleaning dan DSSTE. Gap penelitian ini adalah tidak melakukan *deploy* terhadap model yang sudah dilatih dan juga minimnya teknik praproses data.	Penelitian ini menggunakan teknik *Grey Wolf Optimization* (GWO) dan *Particle Swarm Optimization* (PSO) dalam melakukan tahapan feature selection untuk 3 dataset yang dipilih dan rata-rata akurasi yang didapatkan oleh model yang dilatih dengan algoritma *Random Forest* adalah 99,66%. Gap penelitian ini adalah penggunaan teknik umum dalam praproses dan juga tidak dilakukannya *deploy* terhadap model yang sudah dilatih.

Penelitian terdekat dekat pertama adalah penelitian (Liu, dkk., 2021) dimana *dataset* yang digunakan adalah CICIDS2018, namun pada penelitian tersebut menggunakan teknik *Difficult Set Sampling Technique* (DSST) untuk menangani permasalahan data yang *unbalanced*, Algoritma ini menggunakan

pendekatan *Edited Nearest Neighbor* (ENN) untuk membagi data pelatihan menjadi *near-neighbor set* (sampel sulit) dan *far-neighbor set* (sampel mudah) dengan melakukan kompresi pada kelas mayoritas dan memperbesar variasi atribut kontinu dari kelas minoritas dalam sampel sulit agar lebih mencerminkan distribusi aslinya, mirip dengan teknik SMOTE namun teknik ini melakukan pengurangan pada kelas mayoritas. Penelitian (Liu, dkk., 2021) memiliki kemiripan pada tahap praproses data, namun mereka hanya melakukan *feature engineering* sederhana berupa *one-hot encoding* untuk mentransformasikan variabel "*protocol*" dan melakukan *data cleaning* dengan cara menghapus variabel konstan saja. Pada (Liu, dkk., 2021) didapatkan hasil evaluasi model *Random Forest* dengan akurasi sebesar 96,92% dan nilai metrik *precision, recall* dan *F1-score* berturut-turut sebesar 97,39%, 96,92% dan 96,98%. Seperti kebanyakan penelitian, model yang telah dilatih tidak di-*deploy* ke sebuah aplikasi.

Gap penelitian (Liu, dkk., 2021) dengan penelitian ini adalah penggunaan sebuah teknik untuk menghadapi masalah data yang *unbalanced* dimana pada penelitian ini tidak dilakukan sama sekali, melainkan hanya menggunakan algoritma *boosting* yang dapat mengurangi dampak negatif dari data yang *unblanced.*

Penelitian terdekat kedua adalah penelitian (Keserwani dkk., 2021) yang menggunakan 3 *dataset* NIDS dan salah satunya adalah dataset CICIDS2017 dan algoritma *Random Forest*, walaupun *dataset* CICIDS2017 mencakup label yang berbeda, tapi memiliki kesamaan dalam bentuk format karena dibuat dengan *tool CICFlowMeter*. Teknik *oversampling* digunakan untuk menghadapi masalah

dataset yang *unbalanced*, teknik *Grey Wolf Optimization* (GWO) dan *Particle Swarm Optimization* (PSO) digunakan untuk melakukan *feature selection*.

Gap penelitian (Keserwani dkk., 2021) dengan penelitian ini adalah penggunaan teknik *oversampling* sebagai salah satu pendekatan dalam menangani *dataset* yang *unbalanced*. Penelitian (Keserwani dkk., 2021) ini cukup unik karena pengguaan teknik GWO dan PSO untuk melakukan *feature selection* yang merupakan teknik yang tidak umum.

Dari penelitian yang dilakukan oleh (Liu dkk., 2021) dan (Keserwani dkk., 2021) memiliki kesamaan gap dengan penelitian ini, yaitu model yang tidak di-*deploy* menjadi sebuah aplikasi siap pakai dimana tidak adanya tahapan teknis untuk menggunakan model yang sudah dilatih. Gap ini menjadi salah satu kontribusi yang bisa diberikan oleh penelitian ini.