BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini berfokus pada pengembangan model deteksi anomali berbasis data untuk mendukung sistem *Predictive Maintenance* (PdM) di lingkungan industri. Objek utama penelitian adalah dataset time-series sensor mesin produksi, yang digunakan untuk melatih dan mengevaluasi model *Machine Learning*. Mengingat kerahasiaan data operasional PT Saka Farma, penelitian ini memanfaatkan dataset public yang relevan sebagai representasi data asli perusahaan, isi dataset ditunjukkan pada tabel 3.2.

Tabel 3.1 Dataset Statistik

Komponen	Nilai
Total Data	4.993
Jumlah Fitur	9 fitur
Interval	30 Menit
Periode	1 Januari 2022 – 14 April 2022
Label	Tersedia

Tabel 3.2 Deskripsi Atribut Dataset Sensor

Fitur	Deskripsi
timestamp	Waktu dan tanggal spesifik saat data sensor direkam (interval rekaman 30 menit).
sensor_id	Identitas unik sensor yang merekam data,
temperature	Suhu lingkungan atau komponen mesin yang terukur, dalam derajat Celcius
vibration	Intensitas getaran mesin (dalam satuan Unit), menunjukkan keausan atau ketidakseimbangan komponen.
pressure	Tingkat tekanan yang terukur, penting untuk monitoring kesehatan mesin Utility (seperti sistem HVAC atau air).

operational_status	Status operasional mesin pada saat perekaman data (misalnya: Operational, Maintenance, atau Failure).
energy_consumption	Data konsumsi energi mesin saat itu, dalam kilowatt-hour (kWh).
fault_flag	Indikator biner (True/False) yang menandakan terjadinya kerusakan atau kegagalan pada mesin.
decision_label	Label akhir yang ditetapkan pada data (sering digunakan sebagai *ground truth* untuk melatih model

Dataset terpakai yang dicantum pada tabel 3.1 terdiri dari 4.993 baris data dengan interval pencatatan setiap 30 menit, mencakup periode 1 Januari 2022 hingga 14 April 2022. Pendekatan ini memungkinkan penelitian tetap valid secara akademis, sekaligus menghasilkan model dan rekomendasi yang dapat diimplementasikan langsung pada data aktual perusahaan di masa mendatang.

3.2 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif di mana data yang dikumpulkan dianalisis menggunakan metode statistik untuk mengukur dan menginterpretasi hubungan antara variabel. Pendekatan ini dipilih karena memungkinkan pengujian hipotesis secara objektif dan terukur, dengan hasil yang dapat digeneralisasi. Metode kuantitatif dalam penelitian ini melibatkan pengolahan data menggunakan teknik analisis berbasis metrik, seperti evaluasi performa model menggunakan nilai-nilai numerik. Pendekatan ini mendukung validitas hasil penelitian serta memberikan dasar yang kuat untuk kesimpulan yang dibuat.

3.2.1 Metodologi

Terdapat dua jenis metodologi yang dibandingkan dalam table 3.1, yaitu *SEMMA* dan *CRISP-DM*. Keduanya merupakan metodologi yang umum digunakan dalam proyek *data mining*. Berikut adalah tabel yang membandingkan kedua metodologi tersebut:

Tabel 3.3 Perbandingan Metodologi

Metodologi	*SEMMA*	*CRISP-DM*
Tujuan Utama	Berfokus pada tahapan teknis analitik untuk membangun, mengeksplorasi, dan mengevaluasi model data.	Memandu seluruh siklus proyek data mining dari pendefinisian masalah bisnis hingga rekomendasi implementasi.
Struktur	Cenderung linear dan lebih berfokus pada tahap teknis seperti eksplorasi, modifikasi data, dan pelatihan model.	Iteratif, mencakup seluruh siklus proyek dengan perhatian pada pemahaman masalah bisnis dan implementasi.
Tahapan Utama	1. Sample 2. Explore 3. Modify 4. Model 5. Assess	1. Business Understanding 2. Data Understanding 3. Data Preparation 4. Modeling 5. Evaluation 6. Deployment
Fokus pada Bisnis	Tidak memiliki tahapan eksplisit untuk mendefinisikan kebutuhan bisnis atau konteks operasional mesin HVAC.	Tahapan pertama berfokus pada latar belakang bisnis potensi kerugian finansial (Rp 1.75 M/hari) dan urgensi implementasi PdM di PT Saka Farma.
Kemampuan Iterasi	Kurang menekankan perulangan, tahapannya cenderung diselesaikan secara linear.	Sangat iteratif. Memungkinkan revisi pada Data Preparation dan Modeling berulang kali setelah Evaluation.
Kelebihan untuk Penelitian Ini	Cocok untuk uji coba model sederhana secara cepat.	Sangat cocok karena mendukung siklus percobaan model yang ekstensif (*Unsupervised*, *Supervised*, dan *Hybrid*) serta mengikat hasil evaluasi model langsung ke tujuan bisnis.
Kekurangan untuk Penelitian Ini	Mengabaikan konteks bisnis kritis dan tidak ideal untuk membandingkan berbagai jenis arsitektur model secara iteratif.	Membutuhkan alokasi waktu yang signifikan di awal untuk tahap Business Understanding yang komprehensif.

Penelitian ini menggunakan kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining) karena pendekatannya yang menyeluruh dan lebih adaptif dibandingkan SEMMA. Metodologi SEMMA (Sample, Explore, Modify, Model, Assess) bersifat linear dan cenderung fokus pada aspek teknis data saja, sehingga kurang tepat untuk skripsi yang harus mempertimbangkan konteks bisnis secara mendalam. Sebaliknya, CRISP-DM memulai dengan Business Understanding, sebuah tahap krusial yang memastikan tujuan model deteksi anomali pada sensor HVAC selaras

dengan upaya PT Saka Farma untuk memitigasi kerugian besar (Rp 1.75 Miliar per hari) akibat downtime. Aspek Business Understanding ini tidak terakomodasi dalam SEMMA.

Dalam aspek fleksibilitas dan eksperimen, CRISP-DM jauh lebih unggul karena sifatnya yang iteratif (berulang-ulang). Fleksibilitas ini sangat dibutuhkan mengingat penelitian ini menguji berbagai algoritma dalam tiga kategori: Individual *Unsupervised* (DBSCAN, Autoencoder), Individual *Supervised* (SVM, RF, XGBoost, termasuk penanganan imbalance data dengan SMOTE), serta kombinasi *Hybrid*. Ketika hasil Evaluation menunjukkan kinerja salah satu model (misalnya, Autoencoder + XGBoost) kurang optimal, CRISP-DM memudahkan peneliti untuk segera kembali ke tahap Data Preparation guna melakukan feature engineering atau data balancing tambahan. Dengan kerangka kerja ini, setiap algoritma, dari *Unsupervised* hingga *Hybrid*, dapat dioptimalkan secara terstruktur. Pada akhirnya, meskipun model tidak di-deploy secara fisik, CRISP-DM mencakup tahap Deployment yang dalam konteks skripsi ini diwujudkan sebagai rekomendasi model terbaik. Proses ini memastikan bahwa hasil evaluasi model yang paling superior secara teknis dan paling menguntungkan secara bisnis, siap menjadi panduan implementasi nyata bagi perusahaan

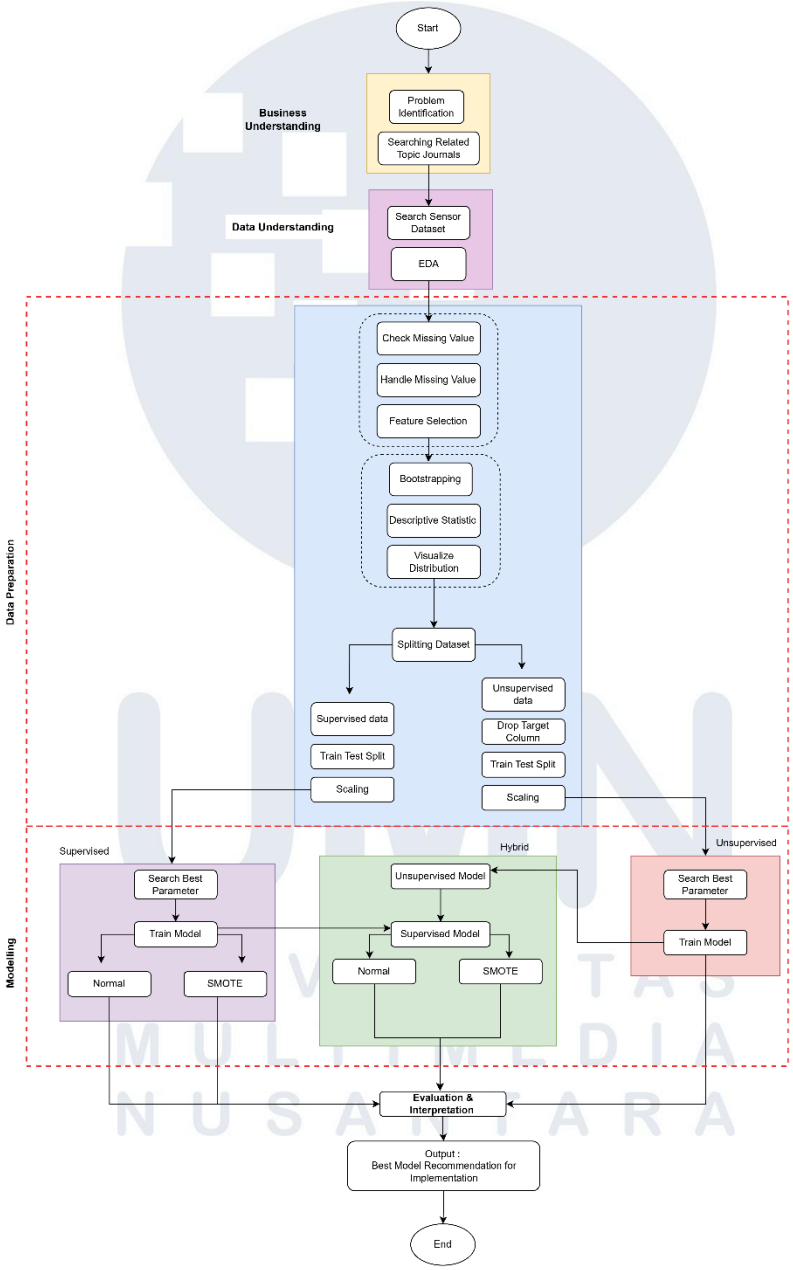
3.2.2 Alur Penelitian

Sebelum memulai penelitian ini, diperlukan penyusunan rancangan dan kerangka yang dirancang secara sistematis serta berkesinambungan. Hal ini bertujuan untuk memastikan bahwa setiap langkah dalam penelitian memberikan dasar yang kokoh dalam mencapai tujuan yang telah ditetapkan. Rancangan alur yang baik tidak hanya memandu jalannya penelitian, tetapi juga menjaga agar proses analisis berjalan secara terarah dan konsisten. Dengan pendekatan ini, setiap tahapan penelitian dapat saling mendukung, sehingga temuan yang

diperoleh memiliki nilai akademis dan relevansi yang tinggi. Adapun alur penelitian yang dirancang dalam penelitian ini adalah sebagai berikut.

3.2.2.1 CRISP-DM

Alur dan penerapan tahapan CRISP-DM yang akan diimplementasikan dalam penelitian ini dapat dilihat pada gambar 3.2 di bawah ini.



Gambar 3.1 Alur Penelitian

3.2.1.1.1 Business Understanding

Tahap Business Understanding adalah langkah awal yang paling fundamental dalam penelitian ini. Tujuannya adalah untuk memahami secara mendalam tujuan dan persyaratan bisnis dari sudut pandang PT. Saka Farma. Masalah utama yang diidentifikasi adalah tingginya kerugian akibat downtime mesin yang tidak terduga, yang berdampak pada biaya, waktu, dan manajemen rantai pasok. Penelitian ini bertujuan untuk mengatasi masalah tersebut melalui penerapan sistem perawatan prediktif.

Dengan demikian, tujuan penelitian diterjemahkan menjadi kebutuhan bisnis: mengembangkan model *Machine Learning* yang efektif untuk memprediksi anomali pada mesin produksi. Tahap ini juga menetapkan kriteria keberhasilan proyek, yaitu menghasilkan rekomendasi model terbaik yang dapat mengurangi risiko kegagalan mesin dan mengoptimalkan jadwal perawatan.

3.2.1.1.2 Data Understanding

Pada tahap Data Understanding, penelitian berfokus pada pengumpulan, eksplorasi, dan pemahaman data yang relevan. Karena isu kerahasiaan, data aktual dari PT. Saka Farma tidak dapat digunakan secara langsung. Oleh karena itu, akan digunakan dataset publik yang berfungsi sebagai representasi data, yang memiliki karakteristik serupa dengan data sensor mesin di perusahaan.

Validasi data dalam penelitian ini dilakukan secara kualitatif menggunakan metode *domain expert* [76]. Domain expert yang terlibat merupakan Production Engineer di PT Saka Farma, yang memiliki pengalaman langsung dalam pengoperasian dan pemantauan mesin produksi serta

interpretasi data sensor mesin dalam aktivitas harian pabrik. *Domain expert* ini terlibat secara langsung dalam proses produksi berbasis shift dan bertanggung jawab terhadap evaluasi kondisi mesin berdasarkan parameter suhu, getaran, tekanan, dan konsumsi energi.

Pada kondisi operasional nyata, data sensor mesin industri umumnya tidak memiliki label anomali, sehingga pendekatan unsupervised diperlukan sebagai tahap awal untuk mendeteksi pola penyimpangan dengan pseudo-labeling. Oleh karena itu, penelitian ini mempertahankan pembelajaran unsupervised untuk merepresentasikan kondisi nyata industri. Namun, untuk keperluan evaluasi kuantitatif dan perbandingan performa model, penelitian ini menggunakan dataset publik yang telah dilengkapi label anomali sebagai ground truth eksperimental agar kinerja berbagai pendekatan dapat diukur secara objektif menggunakan metrik evaluasi. Demikan pula pendekatan supervised dalam penelitian ini tidak menggantikan kondisi nyata industri, melainkan digunakan sebagai alat evaluasi dan pembanding, serta sebagai kombinasi komponen classifier dalam pendekatan *hybrid* untuk meningkatkan akurasi deteksi anomali. Maka dari itu dilakukan banyak pemodelan untuk mencari kombinasi model *hybrid unsupervised – supervised* terbaik.

Dataset ini akan dieksplorasi untuk memahami jenis fitur yang tersedia (e.g., suhu, getaran, tekanan), format data (time-series), distribusi nilai, keseimbangan data serta potensi adanya data yang hilang atau anomali yang sudah ada di dalamnya. Eksplorasi visual dan analisis statistik sederhana akan dilakukan untuk mendapatkan wawasan

awal tentang data, yang akan menjadi panduan untuk langkah pra-pemrosesan berikutnya.

3.2.1.1.3 Data Preparation

Pada tahap data preprocessing, pertama-tama dilakukan pengecekan terhadap dataset untuk memastikan apakah terdapat missing values. Selanjutnya, dilakukan feature selection untuk memilih fitur-fitur utama dari dataset, yang terdiri dari temperature, vibration, pressure, energy_consumption, dan fault_flag. Setelah itu, dilakukan bootstrap resampling sebanyak 10.000 baris untuk menambah jumlah data dan meningkatkan variasi data. Proses bootstrap ini diperlukan untuk menangani dataset yang mungkin tidak mencukupi untuk model training, serta memastikan model dapat belajar dari data yang lebih beragam. Setelah bootstrap, dilakukan descriptive statistics untuk membandingkan data awal dan data bootstrap, di mana hasilnya menunjukkan kemiripan yang signifikan, baik dari segi rata-rata maupun distribusi data. Visualisasi seperti histogram dan kernel density estimates juga menunjukkan bahwa distribusi data pada dataset bootstrap sangat mirip dengan data asli, yang menandakan validitas hasil bootstrap.

Dataset yang telah dibootstrap kemudian dibagi menjadi dua bagian: *supervised* dan *unsupervised*. Pada data *unsupervised*, kolom target fault_flag dihapus, dan dataset dibagi menjadi data training dan testing menggunakan time based train-test split. Data numerik pada kedua dataset kemudian dilakukan scaling, dengan proses fit-transform dilakukan pada data training dan transform pada data testing. Data yang telah diproses ini kemudian digunakan untuk melatih model *supervised* (Random Forest, SVM, XGBoost) serta *unsupervised* (DBSCAN, Isolation Forest,

Autoencoder) dalam rangka mendeteksi anomali, khususnya pada data sensor yang belum memiliki label anomali.

3.2.1.1.4 Modelling

Setelah data melalui proses preprocessing dan scaling pada fitur numerik, tahap selanjutnya adalah pemodelan yang bertujuan untuk membangun dan menguji berbagai algoritma deteksi anomali. Proses ini dilakukan dengan tiga pendekatan utama, yaitu *Supervised, Unsupervised,* dan *Hybrid Model.* Seluruh model dilatih menggunakan data latih yang telah dipisahkan sebelumnya.Pada pendekatan *Supervised* Learning, model dilatih secara langsung menggunakan kolom *fault_flag* sebagai variabel target, sehingga model dapat belajar mengklasifikasikan kondisi operasional normal dan anomali pada mesin. Algoritma yang diterapkan dalam pendekatan ini meliputi Support Vector Machine (SVM), *Random Forest* (RF), dan XGBoost. Selain kondisi dasar, kinerja model juga diuji setelah penerapan teknik SMOTE (Synthetic Minority Over-sampling Technique) untuk mengatasi masalah ketidakseimbangan kelas.Sementara itu, pada pendekatan *Unsupervised* Learning, variabel target *fault_flag* sengaja dihapus. Model hanya bergantung pada pola distribusi data untuk mengidentifikasi titik-titik data yang menyimpang dari pola umum sebagai indikasi anomali. Algoritma yang digunakan meliputi DBSCAN *(Density-Based Spatial Clustering of Applications with Noise*), Autoencoder, dan Isolation Forest. Selain pemodelan individual, penelitian ini juga mengembangkan *Hybrid Approaches* dengan menggabungkan kekuatan dari algoritma *unsupervised* dan *supervised*. Contoh kombinasi yang dikembangkan adalah DBSCAN dengan Random

Forest, Autoencoder dengan RF, atau Isolation Forest dengan XGBoost. Tujuan dari pendekatan *hybrid* ini adalah memanfaatkan kemampuan deteksi pola atau klaster anomali yang dilakukan oleh metode *unsupervised*, lalu memperkuat tahap klasifikasi akhir menggunakan algoritma *supervised* agar hasil deteksi menjadi lebih akurat dan *robust*.

3.2.1.1.5 Evaluation

Tahap evaluation dilakukan untuk menilai performa model deteksi anomali yang telah dibangun. Proses evaluasi menggunakan metrik yang relevan dengan klasifikasi data tidak seimbang, yaitu accuracy, *Precision*, *recall*, dan *F1-score* , serta analisis confusion matrix untuk melihat distribusi prediksi antara kelas normal dan anomali. Pada model *supervised* (SVM, Random Forest, XGBoost), evaluasi dilakukan dengan membandingkan hasil prediksi terhadap label fault_flag yang tersedia. Untuk model *unsupervised* (DBSCAN, Autoencoder, Isolation Forest), evaluasi dilakukan dengan mengukur kemampuan algoritma dalam mengidentifikasi pola outlier dan membandingkannya dengan label referensi.

Selain itu, penelitian ini juga membandingkan performa individual models dengan *hybrid*. *Hybrid* model dievaluasi untuk melihat apakah kombinasi metode *unsupervised* dan *supervised* mampu meningkatkan akurasi deteksi anomali dibandingkan penggunaan algoritma tunggal. Hasil evaluasi ini menjadi dasar dalam menentukan model terbaik yang paling sesuai untuk diterapkan pada data sensor industri.

3.2.1.1.6 Deployment

Dalam konteks penelitian ini, tahap deployment tidak dilakukan dalam bentuk implementasi langsung ke sistem produksi, melainkan difokuskan pada penyusunan rekomendasi model terbaik yang dapat diadaptasi oleh PT Saka Farma. Berdasarkan hasil evaluasi, model dengan performa paling konsisten dan akurat dalam mendeteksi anomali dipilih sebagai kandidat utama untuk digunakan dalam sistem *Predictive Maintenance*. Rekomendasi ini mencakup penjelasan mengenai algoritma yang digunakan, keunggulan model dalam menangani data sensor industri, serta potensi penerapannya pada data aktual perusahaan.

3.3 Teknik Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data time-series dari parameter sensor mesin yang diperoleh dari repositori data public github dengan nama Industrial_IoT_Sensor_Dataset. Pemilihan dataset eksternal ini didasari oleh pertimbangan kerahasiaan dan keamanan data operasional PT. Saka Farma. Dataset publik ini dipilih karena memiliki karakteristik serupa dengan data asli, seperti format time-series dan keberadaan anomali yang relevan dengan kasus kegagalan mesin. Data tersebut mencakup berbagai parameter sensor seperti suhu, tekanan, getaran, dan arus listrik yang direkam secara berkala. Teknik pengumpulan data ini memungkinkan peneliti untuk melakukan studi valid dan mendalam tanpa mengakses data internal perusahaan.

3.4 Variabel Penelitian

Dalam penelitian ini, terdapat dua jenis variabel utama, yaitu variabel independen dan variabel dependen, yang didefinisikan sebagai berikut:

1) **Variabel Independen** adalah variabel yang menjadi input atau faktor pengaruh utama dalam penelitian ini, dan tidak dipengaruhi oleh variabel lain. Dalam konteks deteksi anomali, variabel independen adalah seluruh data fitur dari parameter sensor mesin kecuali target label. Data ini

mencakup seri waktu dari berbagai sensor yang merekam kondisi temperature, vibrasi, tekanan dan konsumsi energi.

2) **Variabel Dependen** adalah variabel yang menjadi hasil atau output dari penelitian ini, yang dipengaruhi oleh variabel independen. Dalam penelitian ini, variabel dependen adalah status anomali mesin (*fault_flag*), yang merupakan hasil prediksi dari model *Machine Learning*. Status anomali ini dapat berupa nilai biner (normal atau anomali) atau skor anomali yang menunjukkan tingkat deviasi dari kondisi normal. Kualitas dari variabel dependen ini dievaluasi dengan menggunakan metrik *Precision*, *Recall*, dan *F1-score* .

3.5 Teknik Analisis

Ada beberapa *tools* yang berguna untuk memilih teknik analisis data, misalnya dengan menggunakan *Jupyter Notebook* yang paling sering digunakan, ataupun *Google Collab*. Berikut adalah perbandingan antara kedua tools yang dapat dilihat pada tabel 3.2

Tabel 3.4 Perbandingan *Tools*

Kriteria	**Jupyter Notebook**	**Google Colaboratory**
Kemudahan Penggunaan	Antarmuka interaktif dan mudah digunakan, ideal untuk eksplorasi data dan visualisasi langsung.	Antarmuka berbasis Jupyter Notebook yang identik dan ramah pengguna, mudah diakses melalui browser.
Akses Hardware	Terbatas pada spesifikasi hardware lokal.	Menyediakan akses gratis ke GPU (Graphics Processing Unit) dan TPU (Tensor Processing Unit) yang powerful.
Kolaborasi	Terbatas pada satu pengguna, memerlukan platform lain seperti Git untuk berbagi.	Terintegrasi dengan Google Drive, memungkinkan kolaborasi real-time dan berbagi notebook dengan mudah.
Manajemen Proyek	Terutama digunakan untuk analisis ad-hoc dan eksperimen cepat, kurang optimal untuk proyek skala besar.	Sangat cocok untuk eksperimen dan prototyping model *deep learning* yang memerlukan sumber daya komputasi besar.

Kelebihan	Cocok untuk tahapan eksplorasi data, pra-pemrosesan, dan analisis cepat.	Mendukung pelatihan model yang kompleks dan membutuhkan waktu lama tanpa biaya hardware, sangat ideal untuk penelitian *deep learning*.
Kekurangan	Memerlukan konfigurasi lingkungan dan keterbatasan komputasi untuk proyek berat.	Keterbatasan waktu sesi dan koneksi internet yang stabil untuk akses.

Dalam pelaksanaan penelitian ini, Google Colaboratory (Colab) dipilih sebagai lingkungan utama untuk analisis dan pemodelan, menggantikan Jupyter Notebook lokal. Keputusan ini didasarkan pada keunggulan Colab, terutama dalam aksesibilitas sumber daya komputasi yang vital untuk pengembangan model *Machine Learning* yang kompleks. Colab menyediakan akses gratis ke GPU (Graphics Processing Unit) dan TPU (Tensor Processing Unit) yang jauh lebih powerful dibandingkan spesifikasi perangkat keras lokal (hardware). Ketersediaan sumber daya ini menjadi faktor penentu, mengingat penelitian ini melibatkan pelatihan model yang membutuhkan komputasi berat dan waktu yang lama, seperti Autoencoder dan berbagai algoritma kompleks lainnya (SVM, RF, XGBoost).

Keunggulan Colab juga terasa dalam efisiensi waktu dan manajemen data. Lingkungan Colab sudah terinstal dengan sebagian besar library *Machine Learning* dan Deep Learning yang dibutuhkan (seperti scikit-learn, TensorFlow, atau PyTorch), sehingga peneliti tidak perlu lagi melakukan proses pip install atau konfigurasi lingkungan yang rumit. Hal ini mempercepat alur kerja pemodelan dan pengujian. Selain itu, Colab dapat langsung terintegrasi dan terhubung dengan Google Drive, memungkinkan seluruh dataset sensor dan hasil pemodelan disimpan ke layanan cloud secara efisien, sehingga tidak membebani memori internal perangkat yang digunakan. Dengan demikian, Colab dinilai sangat ideal untuk eksperimen dan prototyping model deteksi anomali skala besar, memberikan solusi yang efisien, cepat, dan powerful tanpa dibatasi oleh spesifikasi hardware local.