

Problem Solving

1. Business Understanding.

1.1 Background

Efisiensi pengeluaran finansial merupakan aspek krusial dalam pengelolaan perusahaan. Sayangnya, masih sering ditemukan kasus kecurangan internal seperti manipulasi proses reimbursement, penggunaan kartu kredit perusahaan untuk keperluan pribadi, serta pemalsuan dokumen perjalanan dinas. Hal-hal ini menyebabkan kerugian finansial dan melemahkan sistem pengawasan perusahaan.

Saat ini, proses verifikasi dilakukan secara manual melalui cross-check yang memakan waktu dan tidak efisien. Oleh karena itu, perusahaan membutuhkan solusi berbasis Artificial Intelligence (AI) dan Machine Learning (ML) untuk secara otomatis memantau, mendeteksi, dan memperingatkan adanya transaksi yang mencurigakan.

1.1 Problem Statement.

Kecurangan (fraud) dalam proses reimbursement dapat menimbulkan kerugian langsung maupun tidak langsung bagi perusahaan. Beberapa modus yang umum terjadi antara lain:

- Pengajuan biaya fiktif
- Duplikasi atau manipulasi nota pembelian
- Pemalsuan aktivitas perjalanan dinas
- Pembelian kebutuhan pribadi dengan fasilitas kantor

Masalah ini mengindikasikan kurangnya sistem pengawasan otomatis dan sulitnya mendeteksi anomali dalam pengajuan reimbursement secara real-time.

1.2. Goal

Tujuan utama dari inisiatif ini adalah untuk mengurangi potensi kerugian finansial akibat penyalahgunaan proses reimbursement, sekaligus meningkatkan akuntabilitas dan integritas internal.

Dengan memanfaatkan AI/ML, perusahaan menargetkan untuk:

- Mengidentifikasi pola transaksi tidak wajar secara otomatis (anomali)
- Pencegahan duplikasi pada nota pembelian / Tiket perjalanan/ sejenis.

1.3. Manfaat

Adopsi teknologi AI/ML dalam sistem reimbursement memberikan berbagai manfaat strategis, antara lain:

- Deteksi otomatis dan cepat terhadap klaim reimbursement yang tidak wajar
- Proses audit yang lebih efisien dengan fokus pada klaim berisiko tinggi
- Pengurangan kerugian finansial akibat kecurangan
- Peningkatan kepercayaan karyawan terhadap sistem reimburse yang lebih adil dan objektif

2. Usulan Solusi AI / ML

Berdasarkan permasalahan yang dihadapi perusahaan berikut beberapa pendekatan solusi yang dapat diimplementasikan dengan menggunakan teknologi AI/ML.

1. ML Anomaly Detection .

Machine learning untuk deteksi anomali merupakan sebuah pendekatan teknologi berbasis machine learning untuk mendeteksi perilaku yang tidak wajar dengan mengenali pola data menggunakan algoritma statistika.

1.1. Cara Kerja.

- Data collecting
Tahapan ini yaitu mengumpulkan semua data user termasuk data transaksi seperti (jumlah klaim, kategori pengeluaran, waktu pengajuan, lokasi, dan identitas karyawan) selama periode tertentu.
- Training Data
Tahapan ini yaitu dengan menggunakan data yang telah dikumpulkan. proses pembelajaran data dilakukan dengan melakukan proses data manipulation seperti (cleansing, transform dsb) yang kemudian dengan menggunakan algoritma statistika pola transaksi dipelajari yang nantinya hasilnya digunakan sebagai sebuah detector anomaly yang disebut model dengan mempelajari fitur data user tersebut.
- Anomaly Detection
Ketika ada transaksi baru, model menghitung "anomaly score". Transaksi dengan skor tinggi dianggap menyimpang dan akan di-flag untuk investigasi lebih lanjut.

1.2. Model ML yang Direkomendasikan.

- Isolation forrest.
- Autoencoder (Neural Network)
- One-Class SVM (Support Vector Machine)
- Random forest.

2. OCR & Image Forgery Detection.

OCR + Image Forgery Detection merupakan kombinasi machine learning yang dapat digunakan untuk mengatasi permasalahan pada transaksi pada nota palsu. OCR dengan kemampuannya mengekstrak image menjadi data text bermanfaat untuk melakukan ekstrak data, dengan kombinasi Image forgery data yang diekstrak dengan OCR dapat analisis untuk diverifikasi keasliannya.

1.1. Cara Kerja.

- Upload Gambar Bukti Reimbursement oleh karyawan.
- OCR Processing untuk ekstraksi teks.
- Image Forensics dijalankan untuk mengecek kemungkinan pemalsuan gambar.
- Cross-validation hasil OCR dengan data sistem (misal: tanggal perjalanan dinas, nominal batas pengeluaran).
- Jika terdapat ketidaksesuaian atau indikasi editan digital, sistem akan mengirim alert ke tim audit.

1.1. Model

- OCR
Tesseract OCR, Google Vision API, EasyOCR
- Image Forgery Detection
CNN-based model, ResNet pretrained, Image Error Level Analysis (ELA), Noise Map

3. NLP (Natural Language Processing)

Beberapa formulir reimbursement mewajibkan karyawan menuliskan deskripsi kegiatan atau keperluan pengeluaran. Di sinilah potensi kecurangan dapat terjadi, misalnya:

- Klaim pribadi ditulis seolah-olah kegiatan dinas.
- Penggunaan istilah kabur atau tidak sesuai konteks pengeluaran.
- Pemakaian narasi berulang untuk klaim yang berbeda-beda.

1.1. Cara Kerja.

- Memahami isi deskripsi klaim.
- Mendeteksi inkonsistensi atau pola bahasa mencurigakan.
- Membandingkan deskripsi dengan kebijakan perusahaan dan histori klaim lainnya.

1.2. Model

- Preprocessing Teks
NLTK, spaCy, scikit-learn
- Text Similarity Analysis
TF-IDF, cosine similarity, Sentence-BERT
- Named Entity Recognition (NER)
spaCy NER, Hugging Face transformers
- Text Classification / Clustering
KMeans, DBSCAN, LDA, Random Forest, BERT
- Text Anomaly Detection
Autoencoder, Isolation Forest, GPT embeddings

3. Data.

3.1 Data Transaksi Reimbursement

Penjelasan:

Data ini mencakup semua informasi yang dimasukkan oleh karyawan saat mengajukan klaim reimbursement.

Tujuan:

Digunakan untuk mendeteksi pola normal vs tidak normal (anomali), klasifikasi pengeluaran, dan identifikasi kebiasaan mencurigakan.

Contoh elemen penting:

- **ID transaksi:** Untuk identifikasi unik dan pelacakan.
- **Jumlah pengeluaran (amount):** Digunakan dalam anomaly detection.
- **Jenis biaya (expense type):** Untuk clustering berdasarkan kategori (makan, transport, dll).
- **Tanggal:** Validasi temporal apakah pengeluaran sesuai dengan waktu dinas.
- **Deskripsi:** Input NLP untuk mengecek keabsahan klaim

3.2 Data Bukti Transaksi (Struk / Nota / Kwitansi)

Penjelasan:

Merupakan bukti visual dari transaksi yang dilakukan karyawan, biasanya berupa foto struk atau nota.

Tujuan:

Digunakan dalam dua sistem:

- OCR untuk mengekstrak teks dari gambar.
- Image Forgery Detection untuk mengecek apakah gambar telah dimanipulasi (edit tanggal, harga, dll).

Komponen penting:

- File gambar asli.
- Metadata gambar (timestamp, lokasi GPS).
- Hasil ekstraksi OCR.
- Skor deteksi manipulasi (dari model ELA/CNN).

3.3 Data Perjalanan Dinas

Penjelasan:

Merupakan informasi resmi perjalanan dinas yang disetujui perusahaan.

Tujuan:

Untuk mencocokkan apakah pengeluaran memang terjadi saat dan di tempat yang sesuai dengan dinas resmi.

Komponen penting:

- Tanggal berangkat dan kembali.
- Tujuan lokasi.
- Bukti tiket dan booking hotel.
- Agenda perjalanan.

3. 4. Riwayat Reimbursement Karyawan

Penjelasan:

Rekaman historis pengajuan reimbursement setiap karyawan selama periode waktu tertentu.

Tujuan:

- Mendeteksi pola individual (normal atau mencurigakan).
 - Membandingkan antar karyawan dalam satu divisi.
 - Menilai risiko fraud berbasis frekuensi, jumlah, dan jenis pengajuan.
-

3. 5. Log Aktivitas Sistem

Penjelasan:

Catatan aktivitas teknis pengguna di dalam sistem reimbursement.

Tujuan:

Digunakan untuk mengidentifikasi:

- Penggunaan sistem di luar jam kerja.
- Perangkat atau lokasi tidak biasa.
- Aktivitas pengeditan mencurigakan (misal mengubah nominal pengeluaran).

3. 6. Data Referensi Eksternal

Penjelasan:

Data pihak ketiga yang digunakan sebagai pembandingan atau validasi terhadap data internal.

Tujuan:

- Mengecek kewajaran harga (misalnya apakah wajar makan siang Rp800.000 di kota tertentu).
- Validasi apakah pengeluaran dilakukan di hari libur nasional.
- Validasi kurs mata uang jika pembelian dilakukan di luar negeri.

4. Technology Lifecycle.

Data Collection

Deskripsi: Mengumpulkan data dari berbagai sumber internal dan eksternal.

Sumber Data:

- Sistem reimbursement internal (database SQL).
- Upload bukti transaksi (gambar struk).
- Log sistem.
- Jadwal perjalanan dinas (Google Sheets, aplikasi HRIS).
- Harga referensi (API publik atau scrapping harga).

Data Preprocessing

Deskripsi: Membersihkan, memformat, dan menyiapkan data agar bisa digunakan untuk pelatihan model.

Aktivitas:

- Normalisasi angka & teks.
- OCR struk → Tesseract / Google Vision API.
- Anomali removal (misalnya data kosong, duplikat).
- Feature engineering.

Model Training (AI/ML Modeling)

Deskripsi: Pelatihan model deteksi fraud dari data yang telah dibersihkan.

Jenis Model:

- Anomaly Detection → Isolation Forest, Autoencoder
- OCR + Image Forgery → ResNet / CNN + ELA
- NLP untuk deskripsi → BERT / TF-IDF + Cosine Similarity
- Text Classification / Clustering → KMeans, RandomForest, BERT

Model Evaluation

Deskripsi: Mengukur performa model sebelum deployment.

Metric:

- Precision, Recall, F1-Score (fraud detection imbalance)
- Confusion Matrix
- ROC-AUC

Deployment (Model Serving)

Deskripsi: Menyediakan API model yang bisa diakses oleh sistem reimbursement.

Teknologi:

- FastAPI / Flask (REST API)
- ONNX (untuk mengoptimalkan model inference)
- Docker + Docker Compose
- Gunicorn + Nginx (untuk prod deployment)

MLOps (Monitoring & Maintenance)

Deskripsi: Memastikan model berjalan stabil, bisa diupdate, dan mendeteksi drift (perubahan data).

Aktivitas:

- Logging prediksi model (dengan timestamp dan input)
- Monitoring performa real-time
- Alert jika model mendeteksi banyak anomali dalam waktu singkat
- Jadwal retraining berdasarkan data baru

Infrastruktur Minimal yang Dibutuhkan

Komponen	Minimal Setup
Data Storage	PostgreSQL, MongoDB, File Storage (Local/S3)
Compute	Server dengan GPU optional (NVIDIA GTX/RTX)
Pipeline Scheduler	Apache Airflow / Prefect
ML Environment	Python (venv), Docker, Jupyter
Model Hosting	FastAPI + Gunicorn, hosted on VPS / cloud
Monitoring	Prometheus + Grafana / MLflow
Version Control	Git + GitHub / GitLab
CI/CD	GitHub Actions + Docker