BAB I
PENDAHULUAN

1.1 Latar Belakang

Pembayaran transaksi keuangan melalui kartu kredit, merupakan pembayaran paling efisien dan sangat memudahkan, terutama semakin banyaknya layanan transaksi pembayaran *online* [1], [2]. Dengan meningkatnya penggunaan kartu kredit untuk transaksi keuangan, bermunculan pula teknik kecurangan baru dalam penipuan transaksi kartu kredit. Hal ini menyebabkan kerugian finansial bagi bank komersial dan perorangan [3]. Transaksi kartu kredit menampilkan perbedaan yang cukup tinggi antara transaksi normal dan transaksi *fraud*, sehingga menyebabkan masalah ketidakseimbangan (*imbalanced*) data [4].

Fraud atau penipuan dapat didefinisikan sebagai suatu kejadian yang melibatkan motif kriminal dari pelakunya. Hal ini menjadi perhatian khusus bagi para peneliti dan institusi keuangan. Di sisi lain, mendeteksi kecurangan merupakan sebuah pekerjaan yang sangat sulit jika menggunakan metode-metode standar. Perkembangan teknologi terkini yang semakin canggih menyebabkan deteksi *fraud* menjadi semakin sulit, karena pelaku mengembangkan teknik penipuan dengan memanfaatkan teknologi. Pengembangan metode deteksi *fraud* menjadi semakin penting bagi kalangan bisnis dan akademis [5]. Dengan meningkatnya jumlah transaksi yang menggunakan kartu kredit, maka setiap harinya transaksi kartu kredit menghasilkan volume data yang sangat besar. Pada data yang sangat besar ini, transaksi kartu kredit memiliki kompleksitas data yang sangat tinggi karena memiliki banyak jumlah variabel, seperti lokasi transaksi, *IP address* saat melakukan transaksi, jenis transaksi dan pola penggunaan kartu kredit. Hal ini menyebabkan kesulitan bagi pihak institusi untuk mendeteksi *fraud* secara efektif dan dalam waktu singkat karena masih menggunakan metode-metode yang standar.

Dalam melakukan deteksi *fraud*, para peneliti menggunakan metode *machine learning* atau *deep learning*. Penggunaan *machine learning*/*deep learning* dapat mengembangkan model yang bisa melakukan analisis pada data yang cukup besar, mengidentifikasikan pola dan mendeteksi perilaku abnormal yang mengarahkan kepada aktifitas ilegal dari sebuah transaksi. Metode *deep learning* juga dapat melakukan adaptasi pada data yang baru dan meningkatkan efisiensi dalam melakukan identifikasi pada tindakan *fraud*. Dalam melakukan deteksi *fraud* pada transaksi kartu kredit, kondisi tidak seimbang (*imbalanced*) selalu ditemukan. Mayoritas data transaksi memiliki kelas normal, sedangkan

data yang tergolong anomali/*fraud* menjadi bagian kecil dari data transaksi. Ketidakseimbangan data transaksi kartu kredit menghasilkan deteksi yang bias dan mengarah ke hasil data yang normal, sehingga mengurangi performa deteksi kecurangan dari model [6].

Beberapa pendekatan yang digunakan untuk mengatasi *imbalanced data*, yaitu teknik *resampling* dan augmentasi. Para peneliti telah memperkenalkan algoritma untuk mengatasi masalah ketidakseimbangan data [7], [8], [9], [10]. Metode seperti *oversampling* dan *undersampling* yang dipakai untuk menangani data yang tidak seimbang, sering kali menimbulkan masalah baru. *Oversampling* dengan pendekatan statistik menghasilkan sejumlah besar data redundan, sementara *undersampling* rentan terhadap hilangnya informasi pada data transaksi normal [6].

Keterbatasan teknik *resampling* dapat diatasi dengan menggunakan teknik augmentasi untuk memperbanyak data. Augmentasi data adalah proses untuk membangkitkan data baru berdasarkan data original. Teknik ini penting untuk meningkatkan keandalan dan kinerja model dan membuatnya menjadi sangat efisien dalam aplikasi di dunia nyata. Dengan menggunakan augmentasi, peneliti bisa meningkatkan ukuran dan diversitas data, membantu mengurangi *overfitting* dan meningkatkan kekokohan model dengan membuatnya sangat mudah beradaptasi dengan data baru [5].

Untuk mengatasi masalah *imbalanced data* yang dapat menyebabkan penurunan performa pada deteksi *fraud*, digunakan *Conditional Tabular Generative Adversarial Networks* (CTGAN). CTGAN [11] adalah metode *oversampling* berbasis GAN yang dapat mempelajari distribusi data kompleks dan menghasilkan sampel transaksi *fraud* yang sesuai dengan distribusi dunia nyata. CTGAN dapat menangani data campuran yang bersifat numerik dan kategorikal secara efektif. Dibandingkan metode *oversampling* yang lain seperti SMOTE, CTGAN unggul dalam menangkap ketergantungan kompleks antar fitur. Selanjutnya data sintetis hasil augmentasi dilakukan penerapan *Neural Network* (*Autoencoders*) untuk mendeteksi anomali/*fraud*. Penggunaan *deep learning* pada deteksi *fraud* menghasilkan model yang memiliki kemampuan menangani data yang kompleks dan memiliki skalabilitas yang besar jika dibandingkan dengan model *machine lerning* tradisional.

Berdasarkan uraian latar belakang di atas, penulis memberikan judul penelitian sebagai berikut: **"Deteksi Anomali pada Generatif Augmentasi Data Transaksi Keuangan dengan Menggunakan *Generative Adversarial Networks*".**

1.2 Rumusan Masalah

Jumlah data yang sangat besar pada transaksi kartu kredit, mengakibatkan meningkatnya kesulitan untuk mendeteksi transaksi *fraud* dengan menggunakan metode standar. Untuk itu dilakukan pendekatan deteksi kecurangan dengan menggunakan *deep learning*. Agar model *deep learning* bisa dilatih untuk mempelajari pola-pola kecurangan, dibutuhkan data dengan tingkat ketidakseimbangan yang tidak terlalu jauh antara data normal dan data *fraud* pada sebuah *dataset*. Untuk menangani ketidakseimbangan data ini, dibutuhkan teknik augmentasi data dengan tujuan memperbanyak data pada kelas *fraud*. Setelah data diseimbangkan, dilakukan penerapan model *neural network* untuk melakukan deteksi kecurangan pada *dataset* yang telah diaugmentasi yang pada penelitian ini digunakan model *neural network autoencoder* untuk mendeteksi *fraud*.

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dan pertanyaan penelitian adalah sebagai berikut:

1. Bagaimana cara melakukan *preprocessing* data agar menghasilkan data sintetis yang memiliki konsistensi, kualitas yang bagus dan mampu mereplika pola karakteristik dari data minoritas untuk dilatih dengan menggunakan CTGAN?
2. Apakah dengan melakukan penambahan data sintetis pada *dataset* yang memiliki ketidakseimbangan yang sangat tinggi, mampu meningkatkan performa model *autoencoder* dalam mendeteksi *fraud* terutama dalam hal klasifikasi dan performa model?

1.3 Tujuan

Tujuan dari penelitian ini adalah melakukan deteksi anomali/*fraud* menggunakan model *machine learning* dengan menambahkan data sintetis pada data dengan kondisi *highly imbalanced*. Penambahan data dilakukan dengan menggunakan CTGAN dan deteksi *fraud* menggunakan model *autoencoder*. Setelah menambahkan data sintetis, distribusi perbandingan data normal terhadap *fraud* yang lebih baik mampu meningkatkan performa model.

1.4 Manfaat

Manfaat dari penelitian ini adalah:

1. Model yang diusulkan dapat memberikan kontribusi dalam melakukan deteksi *fraud* dengan tingkat keakuratan yang lebih tinggi.

2. Model yang diusulkan dapat menjadi pertimbangan untuk menambah sarana ilmu pengetahuan yang dapat digunakan sebagai referensi dalam deteksi anomali.
3. Model yang diusulkan diharapkan memberikan dampak bagi industri terkait seperti peningkatan keamanan, peningkatan kepercayaan, peningkatan efisiensi dan kepatuhan regulasi. Industri yang rentan terhadap aktivitas *fraud* seperti perbankan, *e-commerce* dan asuransi mengalami penurunan reputasi dan tingkat kepercayaan jika terpapar aktivitas *fraud.*

1.5 Ruang Lingkup

Batasan dan ruang lingkup yang diterapkan dalam penelitian ini adalah sebagai berikut:

1. *Dataset* yang digunakan adalah *dataset* transaksi kartu kredit dari situs : https://www.datacamp.com/datalab/sample-*datasets*/*dataset*-python-credit-card-fraud?searchQuery=credit%20card, yang merupakan data transaksi kartu kredit yang berada di Amerika Serikat bagian barat.
2. *Dataset* menampilkan data transaksi yang muncul dalam rentang transaksi dimulai dari tanggal 1 Januari 2019 sampai 31 Desember 2020. Total data keseluruhan adalah 339.607 transaksi, dimana terdapat 1.782 data *fraud* dan 337.825 data normal. Perbandingan data normal terhadap fraud adalah 99,5 : 0,5.
3. *Dataset* terdiri 15 kolom yang berisi data transaksi, lokasi saat melakukan order, lokasi *merchant*, *amount*, tanggal/waktu transaksi dan biodata pemegang kartu (tanggal lahir, kota dan pekerjaan). *Dataset* tidak memiliki kolom nomor kartu kredit.