BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan pesat dalam pembelajaran mesin, terutama di bidang jaringan saraf tiruan (*Neural Network*), telah membawa kemajuan luar biasa di berbagai sektor, termasuk komputer vision (Rahman et al., 2022). *Neural Network* dilatih sehingga dapat mengenali berbagai data visual, seperti *object detection*, *face recognition*, segmentasi citra, dan bahkan diagnosis medis berbasis citra (Musa et al., 2021). Kinerja *Neural Network* menunjukan kemampuan yang luar biasa, namun rentan terhadap serangan yang menyebabkan model tidak dapat mengenali objek menyebabkan kesalahan prediksi (Shtaiwi et al., 2022). Oleh karena itu, pemodelan *Neural Network* harus dioptimalkan supaya dapat mengurangi resiko serangan yang menyebabkan kesalahan prediksi.

Klasifikasi gambar menggunakan *Convolutional Neural Network* (CNN) menjadi metode yang dominan karena kemampuannya untuk mengekstraksi fitur spasial (Alzubaidi et al., 2021). CNN merupakan bagian dari teknik *deep learning* untuk pengolahan citra sehingga model dapat mengenali objek (Guo et al., 2023). CNN dirancang khusus untuk pemrosesan data grid seperti citra dan dapat melakukan pengenalan citra (Kumar et al., 2021). CNN menggunakan *dataset* CIFAR-10 menghasilkan akurasi 73% lebih tinggi daripada yang menggunakan *dataset* ImageNet dengan akurasi sebesar 69.7% (Remerscheid et al., 2022). Penelitian oleh (Antonio et al., 2023) menunjukan bahwa CNN lebih optimal dibandingkan AlexNet yang hanya menghasilkan akurasi 63.3% dengan

menggunakan data CIFAR-10. CNN memberikan kinerja yang optimal dalam melakukan tugas klasifikasi citra namun rentan terhadap *adversarial attack* (Shtaiwi et al., 2022). Pada penelitian (Musa et al., 2021) dan (Sen & Dasgupta, 2023) dilakukan pengujian model pengenalan citra dengan salah satu teknik *adversarial attack* yaitu *Fast Gradient Sign Method* (FGSM) mengakibatkan model mengalami penurunan akurasi.

Penelitian oleh (Sen & Dasgupta, 2023) model CNN menghasilkan akurasi 90% dalam memprediksi objeknya, tetapi mengalami penurunan hingga 66% dampak dari perubahan input data menggunakan FGSM. Parameter mengatur besarnya perubahan pada input pada FGSM ditentukan oleh *epsilon* (Hassan et al., 2022). Nilai *epsilon* pada penelitian (Sen et al., 2023) menunjukan semakin tinggi nilai *epsilon* membuat nilai akurasi model semakin berkurang. Pengujian model CNN menunjukkan akurasi awal sebesar 0.9077, yang menurun menjadi 0.4746 saat diuji menggunakan FGSM dengan nilai *epsilon* 0.1, dan semakin menurun menjadi 0.0499 ketika nilai *epsilon* ditingkatkan menjadi 0.9 (Waghela, 2024). FGSM merupakan salah satu teknik untuk menghasilkan *adversarial* yang dapat mengukur ketahanan model terhadap serangan (Han et al., 2023).

Adversarial adalah suatu pendekatan di mana input dimodifikasi secara halus sehingga tetap terlihat normal oleh manusia, namun dirancang untuk mengecoh model *machine learning* atau *deep learning* (Sen et al., 2023). Perubahan atau modifikasi ini biasanya sangat kecil sehingga tidak terdeteksi oleh penglihatan manusia, tetapi cukup untuk menyebabkan model menghasilkan prediksi yang salah

(Yang et al., 2022). Metode *adversarial* secara umum dikategorikan menjadi dua bagian, yaitu *adversarial attack* dan *adversarial defense.*

Adversarial attack adalah teknik untuk melakukan serangan terhadap kelemahan *machine learning* terutama pada *deep learning*. *Adversarial attack* dimanipulasi secara halus sehingga model dapat menghasilkan prediksi yang salah (Sen et al., 2023). Perubahan dilakukan secara spesifik untuk melakukan serangan model tertentu. Memanfaatkan bahwa pemodelan *machine learning* maupun *deep learning* bergantung pola pola numerik (Waghela, 2024). Terdapat beberapa algoritma *adversarial attack* yaitu, *Fast Gradient Sign Method* (FGSM), *Projected Gradient Descent (*PGD) dan *Carlini & Wagner* (C&W) *attack* (Li, 2021).

Dikarenakan FGSM tidak menghasilkan variasi serangan pada penelitian (Sen & Dasgupta, 2023) menambahkan *patch attack* untuk menghasilkan serangan beragam dengan menambahkan *patch* mencolok pada input gambar, sehingga menurunya akurasi model sedangkan nilai *perturbation* meningkat. Pendekatan FGSM dikombinasikan dengan *Project Gradient Descent* (PGD) telah dilakukan oleh (Zhao et al., 2022) sebagai *iterative attack* untuk menemukan serangan yang paling optimal dengan peningkatan nilai *pertubation*. Namun, serangan tersebut memiliki ketergantungan pada *step size* dan jumlah iterasi sehingga membutuhkan komputasi lebih banyak.

Random noise untuk sebuah gangguan acak yang diterapkan untuk data, salah satunya adalah data gambar (Attias, 2024). *Random noise* dapat digunakan untuk menambah variasi atau data *augmentasi*. *Random noise* bisa digunakan untuk *attack* model *Neural Network,* walaupun penggunaanya kurang optimal

dikarenakan tanpa memperhitungkan struktur data atau model (Zhang et al., 2022). Pada penelitian (Li, 2021) menunjukan *random noise* kurang optimal dibandingkan FGSM dalam melakukan serangan terhadap model, tetapi dapat membuat serangan lebih bervariasi. Semakin banyak *random noise* pada input data dapat terlihat perubahan inputnya. Pengujian oleh (Lin et al., 2024) serangan *random noise* terhadap model CNN tetap dapat mengenali karena dari kinerjanya yang mampu mengenali pola spasial.

Berdasarkan pekerjaan terdapat kesenjangan penelitian. Kesenjangan tersebut adalah FGSM dalam melakukan serangan model bergantung *gradien loss* menyebabkan serangan tidak dapat menghasilkan variasi serangan yang beragam untuk model CNN. Sehingga penelitian ini, FGSM menggunakan *random noise* untuk *adversarial attack* terhadap CNN. FGSM dikombinasikan dengan *random noise* dapat meningkatkan variasi serangan dalam menguji kerentanan CNN.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disajikan dapat dirumuskan beberapa masalah menjadi fokus penelitian, sebagai berikut :

1. Bagaimana pengaruh penambahan *random noise* terhadap FGSM dalam melakukan *adversarial attack* pada model CNN?
2. Bagaimana nilai evaluasi dari FGSM yang ditambahkan dengan *random noise* dibandingkan dengan FGSM tanpa *random noise*?

1.3 Tujuan Penelitian

Penelitian ini secara khusus bertujuan sesuai dengan perumusan masalah yang telah ditentukan, yaitu sebagai berikut:

1. Mengukur pengaruh FGSM ditambahkan dengan *random noise* untuk *adversarial attack* terhadap akurasi model CNN.
2. Menganalisis perbedaan nilai evaluasi dan *perturbation* yang dihasilkan antara FGSM dengan penambahan *random noise* dan FGSM tanpa *random noise.*

1.4 Manfaat Penelitian

Manfaat dari penelitian ini dijabarkan secara spesifik ke dalam beberapa poin berikut:

1. Mengetahui perubahan nilai akurasi model CNN saat diuji menggunakan FGSM dan *random noise* untuk mengevaluasi performa model dalam memprediksi data yang terkena *adversarial attack*.
2. Memberikan pemahaman yang lebih mendalam tentang bagaimana *random noise* memengaruhi efektivitas *adversarial attack* FGSM.

1.5 Batasan Masalah

Penelitian ini memiliki batasan-batasan masalah untuk memfokuskan ruang lingkup penelitian. Berikut adalah rincian batasan-batasan yang diterapkan dalam penelitian ini:

1. Fokus dari penelitian ini, yaitu penambahan *random noise* terhadap FGSM pada model CNN untuk mengetahui perubahan nilai akurasi dan *perturbation*.

2. Model CNN dilatih menggunakan data citra untuk melakukan pengklasifikasian untuk mengetahui nilai akurasi yang dihasilkan tanpa *adversarial attack*.