ABSTRAK

Sistem pengenalan wajah berbasis *Deep Learning* seperti *Deep Face Recognition System* (DFRS) semakin banyak digunakan dalam aplikasi keamanan. Namun, sistem ini rentan terhadap serangan *adversarial attack*, yaitu serangan yang memanipulasi citra input dengan perturbation kecil namun mampu mengubah hasil prediksi model. Penelitian ini bertujuan untuk mengembangkan metode *Fast Numerical Gradient Sign Method* (FNGSM) sebagai pendekatan alternatif untuk meluncurkan *adversarial attack* tanpa memerlukan akses langsung ke dalam arsitektur sebuah model.

Metode FNGSM yang dikembangkan dalam penelitian ini mengadopsi pendekatan numerik *central difference* untuk melakukan estimasi gradien dari *loss function* model. Teknik ini memungkinkan perhitungan perkiraan gradien hanya dengan mengobservasi perubahan output model terhadap variasi input yang sangat minimal, tanpa memerlukan pengetahuan menyeluruh tentang arsitektur internal model. Untuk meningkatkan efisiensi komputasi dan efektivitas serangan, metode FNGSM ini diintegrasikan secara sinergis dengan ADAM *Optimizer*. ADAM *Optimizer* dipilih karena kemampuannya yang telah terbukti dalam menyesuaikan laju pembelajaran secara adaptif untuk setiap parameter, yang berkontribusi pada percepatan konvergensi dan peningkatan stabilitas proses optimasi dalam menghasilkan *perturbation* yang optimal. Serangan dilakukan dalam dua skema, *Dodging* (menghindari pengenalan identitas) dan *Impersonation* (meniru identitas target). Dataset **CFPW-Frontal** digunakan sebagai data uji, sementara model **InsightFace MS1MV2-ArcFace** digunakan sebagai target sistem pengenalan wajah.

Hasil menunjukkan bahwa integrasi metode FNGSM yang dikembangkan dengan ADAM *Optimizer* berhasil meningkatkan success rate hingga 73.5% pada skema *Impersonation* dan 62% pada *Dodging*, dengan kualitas visual tetap terjaga (SSIM $\geq$ 0.97, PSNR $\geq$ 40 dB) meskipun menggunakan jumlah iterasi lebih sedikit (rata-rata 1406 *steps* untuk $\epsilon = 5$). Selain itu, temuan menarik menunjukkan bahwa pada beberapa pasangan citra, metode FNGSM mampu menurunkan nilai *Cosine Similarity* hingga di bawah ambang batas hanya dalam satu langkah iterasi. Fenomena ini mengindikasikan adanya kerentanan inheren pada sistem terhadap *perturbation* yang sangat minimal sekalipun. Secara keseluruhan, hasil penelitian ini memberikan bukti bahwa sistem DFRS memiliki kerentanan terhadap *adversarial attacks* berbasis gradien numerik. Oleh karena itu, menjadi sangat krusial untuk mengembangkan mekanisme mitigasi yang lebih adaptif dan efektif, seperti implementasi *adversarial training* yang komprehensif atau pengembangan sistem deteksi otomatis terhadap penambahan *perturbation* pada citra input, guna memperkuat ketahanan DFRS terhadap ancaman yang terus berkembang.

Kata Kunci: *Deep Learning*, *Face Recognition*, *Adversarial Attacks*

ABSTRACT

Deep Learning-based face recognition systems, such as the Deep Face Recognition System (DFRS), are increasingly utilized in security applications. However, these systems are vulnerable to adversarial attacks, which are input image manipulations using small, often imperceptible, perturbations capable of altering the model's prediction results. This research aims to develop the Fast Numerical Gradient Sign Method (FNGSM) as an alternative approach for launching adversarial attacks without requiring direct access to a model's architecture.

The FNGSM method developed in this study adopts a numerical central difference approach to estimate the gradient of the model's loss function. This technique allows for gradient approximation by solely observing changes in the model's output in response to minimal input variations, without needing comprehensive knowledge of the model's internal architecture. To enhance computational efficiency and attack effectiveness, the FNGSM method is synergistically integrated with the ADAM Optimizer. The ADAM Optimizer was chosen for its proven ability to adaptively adjust learning rates for each parameter, contributing to accelerated convergence and increased stability in the optimization process for generating optimal perturbations. Attacks were conducted under two schemes: Dodging (evading identity recognition) and Impersonation (mimicking a target identity). The **CFPW-Frontal** dataset was used as test data, while the **MS1MV2-ArcFace** model from **InsightFace** served as the target face recognition system.

Results demonstrate that the integration of the developed FNGSM method with the ADAM Optimizer successfully increased the attack success rate to 73.5% in the Impersonation scheme and 62% in the Dodging scheme, while maintaining visual quality (SSIM $\geq$ 0.97, PSNR $\geq$ 40 dB) despite requiring fewer iterations (with an average of 1406 steps for ϵ=5). Furthermore, an interesting finding revealed that for several image pairs, the FNGSM method was capable of reducing the Cosine Similarity value below the threshold in just a single iteration step. This phenomenon indicates an inherent system vulnerability to even minimal perturbations. Overall, these research findings provide evidence that DFRS systems are vulnerable to numerical gradient-based adversarial attacks. Therefore, it is crucial to develop more adaptive and effective mitigation mechanisms, such as the implementation of comprehensive adversarial training or the development of automated systems for detecting perturbations added to input images, in order to strengthen DFRS resilience against ever-evolving threats.

Keywords: Deep Learning, Face Recognition, Adversarial Attacks