Assessing the vulnerability of single-cell RNA-sequencing classifiers to adversarial attacks with adverSCarial

Ghislain Fievet, Julien Broséus, David Meyre, Sébastien Hergalant

Assessing the vulnerability of single-cell RNA-sequencing classifiers to adversarial attacks with *adverSCarial*

Ghislain Fiévet[1], **Julien Broséus**[1,2], **David Meyre**[1,3] and **Sébastien Hergalant**[1]

[1]INSERM U1256, Nutrition, Genetics, and Environmental Risk Exposure (NGERE), University of Lorraine, Nancy, France
[2]Department of Biological Hematology, Laboratory Center, University Hospital of Nancy, Nancy, France
[3]Department of Molecular Medicine, Division of Biochemistry, Molecular Biology, and Nutrition, University Hospital of Nancy, Nancy, France

eccb2024

Several **machine learning** (ML) algorithms dedicated to the detection of healthy and diseased cell types from single-cell RNA-sequencing data (scRNA-seq) have been proposed for biomedical purposes. Before establishing clinical routines and diagnostic decision support tools based on these methods, concerns about their security must be surveyed. Among the possible threats are **adversarial attacks**, the deliberate introduction of carefully crafted input data into ML models, taking advantage of the vulnerabilities in the models' decision-making processes to cause misclassifications or faulty outputs. With **transcriptomic data**, we propose to explore this by modulating gene expression values to **simulate** a wide range of causes, whether technical biases, biological and experimental design variations, or malicious events.

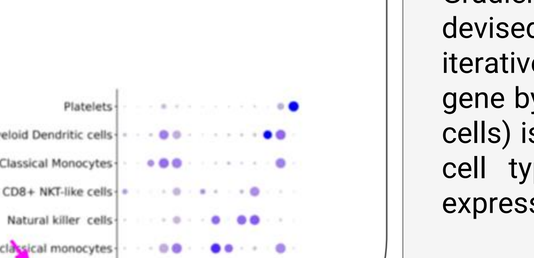
AdverSCarial *an original* R *package available in* Bioconductor *that can be used by researchers, developers, and end-users interested in surveying ML tools that cluster and annotate cells from scRNA-seq data. The package provides a toolkit for the assessment of these algorithms' robustness/vulnerabilities and for evaluating the important variables/features used by the models to reach a final decision and produce an output.* AdverSCarial *is resource-optimized and potentially works on all kinds of single-cell classifiers.*

Adversarial attacks on medical data

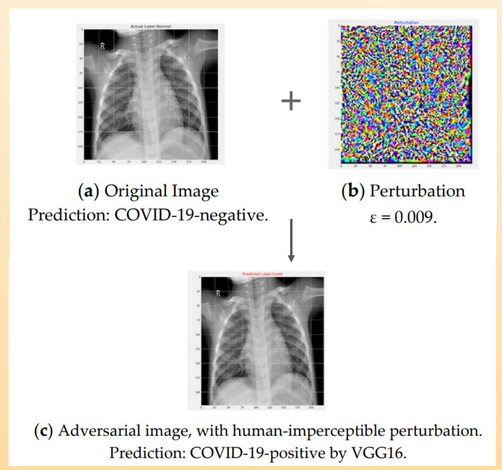
Within the field of medical research, numerous examples of successful adversarial attacks on dedicated ML classifiers are found in the scientific literature: data from electrocardiogram (ECG), X-ray, pathology slides. These attacks are based on the following sequence:

1) Start from raw data (images or signals)
2) Add small / imperceptible perturbations
3) Produce a change of classification

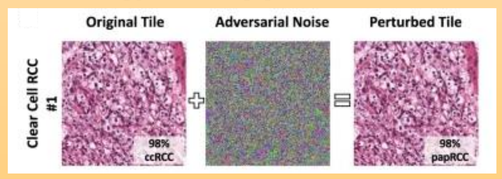
ECG [1]



X-ray [2]



Pathology slides [3]



Attack modes

Single-gene
Produces cell misclassification and mislabelling after altering the expression of one gene only.

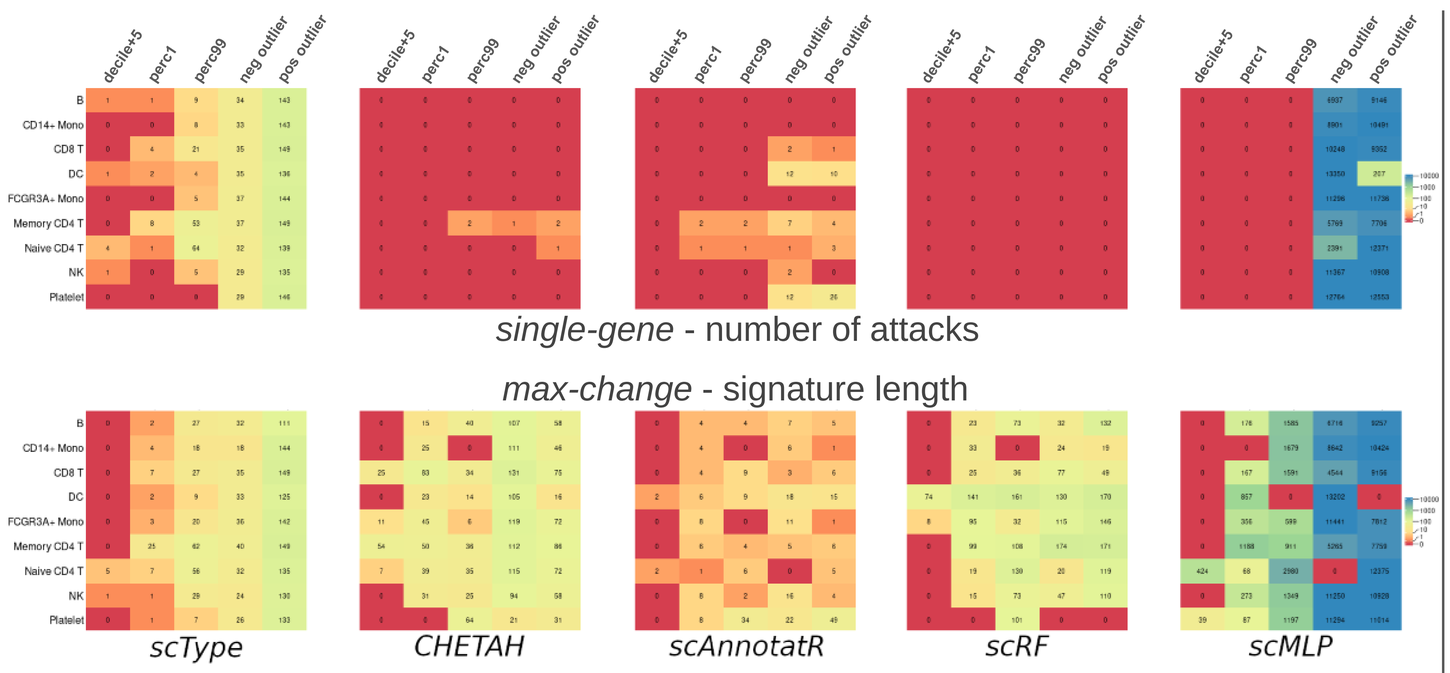
Max-change
Tries to modify as many genes as possible without affecting the cell clustering output. The final complementary gene list consisting of the untouchable genes is referred to as the cell cluster signature.

CGD – Cluster-based Gradient Descent
Gradient-derived adversarial perturbation method devised for black-box scRNA-seq classifiers. This iterative, cluster-based modification process (one gene by step, whole cluster perturbation instead of cells) is guided by an approximated gradient of the cell type likelihood with respect to a gene j expression.

$$slope = \left(\nabla_{j'}^{fd}\hat{y}_{x_l}\right)_{ctype2_{x^n_{clust\,m}}} - \left(\nabla_{j'}^{fd}\hat{y}_{x_l}\right)_{ctype1_{x^n_{clust\,m}}},$$

$$\left(x_l^{n+1}\right)_j = x_{lj}(1-\delta_{jj'}) + \left(\left(1+\alpha.sign(x_{lj})sign(slope)\right)x_{lj} + \epsilon.sign(slope)\right)\delta_{jj'},$$

$$X^{adv}_{clust\,m} = \left(x_l^{n+1}\right)_{l\,\epsilon clust\,m}$$

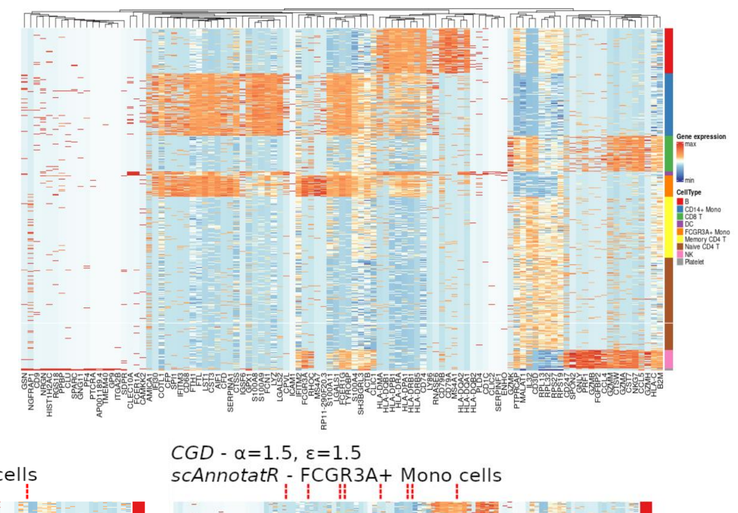


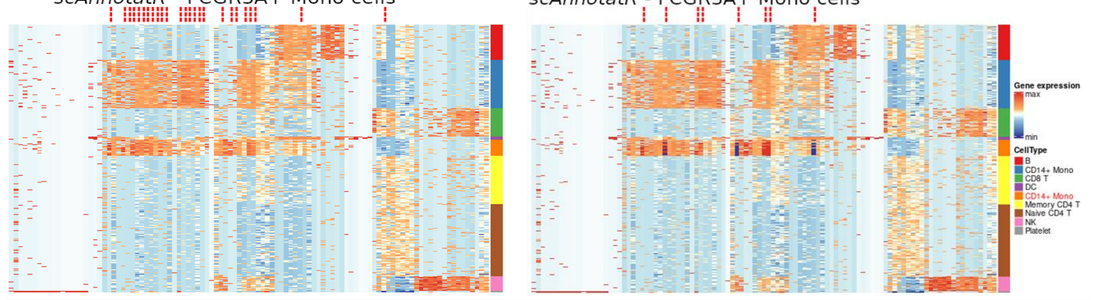
Evaluation of different types of scRNA-seq classifiers

We next conducted a comprehensive examination of the robustness of five distinct ML models dedicated to classifying scRNA-seq data. Three of these classifiers were sourced from existing literature: scType (marker-based model) [4], CHETAH (hierarchical classification) [5], and scAnnotatR (SVM) [6]. For the purpose of this evaluation, we implemented and trained two new classifiers, one based on random forests, hereby called scRF, and one deep learning neural network model based on a multilayer perceptron, called scMLP.

Detectability

Allowing for control over modification intensity using α and ε parameters, *CGD* yields invisible modifications with low values such as **α** = **ε** = 0.3 to visible ones with **α** = **ε** = 1.5. However, with low values the attacks have to impact many genes to fool the classifier, when far less genes are needed with higher values. In the example below, *CGD* imperceptibly adjusted 125 genes to alter *scAnnotatR* classification of the FCGR3A+ Mono cell cluster. Only 19 genes had to be ostensibly modified for the same result.





With the continuous expansion of **single-cell transcriptomics** in many domains, including clinical use and precision medicine, we believe *adverSCarial* can be of help for the development of more **reliable** scRNA-seq **models**, with improved **interpretability**.

Contact ghislain.fievet@univ-lorraine.fr

adverSCarial https://github.com/GhislainFievet/adverSCarial

References

[1] Han X et al. Deep learning models for electrocardiograms are susceptible to adversarial attack. Nat Med. 2020 Mar;26(3):360-363. doi: 10.1038/s41591-020-0791-x. Epub 2020 Mar 9. PMID: 32152582; PMCID: PMC8096552.
[2] Pal, B. et al. Vulnerability in Deep Transfer Learning Models to Adversarial Fast Gradient Sign Attack for COVID-19 Prediction from Chest Radiography Images. Appl. Sci. 2021, 11, 4233. https://doi.org/10.3390/app11094233
[3] Ghaffari Laleh, N. et al. Adversarial attacks and adversarial robustness in computational pathology. Nat Commun 13, 5711 (2022). https://doi.org/10.1038/s41467-022-33266-0
[4] Ianevski, A., Giri, A.K., Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun 2022;13:1246. https://doi.org/10.1038/s41467-022-28803-w
[5] de Kanter, JK., Lijnzaad, P., Candelli, T. et al. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Research, 2019;47(16):e95. https://doi.org/10.1093/nar/gkz543
[6] Nguyen, V., Griss, J. scAnnotatR: framework to accurately classify cell types in single-cell RNA-sequencing data. BMC Bioinformatics, 2022. 23(44). https://doi.org/10.1186/s12859-022-04574-5