CLAIMS

1. A method for securing trained machine learning models in a computing environment by one or more processors, the method comprising:

providing one or more hardened machine learning models secured against adversarial attacks by adding adversarial protection to one or more trained machine learning model.

2. The method of claim 1, further including:

receiving the one or more trained machine learning models; and

retraining the one or more trained machine learning models to include the adversarial protection based on one or more adversarial protection protocols.

3. The method of either of the preceding claims, further including adding one or more preprocessing layers to one or more trained machine learning models, wherein a degree of adversarial protection strength is adjusted.

4. The method of any of the preceding claims, further including adding one or more neural network layers to one or more trained machine learning models.

5. The method of any of the preceding claims, further including adding one or more post-processing output layers to the one or more trained machine learning models that are retrained with the adversarial protection based on one or more adversarial protection protocols.

6. The method of any of the preceding claims, further including:

automatically implementing one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models; or

receiving, from a user, one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models.

7. The method of any of the preceding claims, further including:

monitoring and tracking each state of the one or more trained machine learning models while being retrained;

detecting a training collapse to the one or more trained machine learning models during the retraining; or

enabling one or more roll-back strategies for the one or more trained machine learning models during a retraining operation.

8. A system for securing trained machine learning models in a computing environment, comprising:

one or more computers with executable instructions that when executed cause the system to:

provide one or more hardened machine learning models secured against adversarial attacks by adding adversarial protection to one or more trained machine learning model.

9. The system of claim 8, wherein the executable instructions:

receive the one or more trained machine learning models; and

retrain the one or more trained machine learning models to include the adversarial protection based on one or more adversarial protection protocols.

10. The system of either of claims 8 or 9, wherein the executable instructions add one or more preprocessing layers to one or more trained machine learning models, wherein a degree of adversarial protection strength is adjusted.

11. The system of any of claims 8 to 10, wherein the executable instructions add one or more neural network layers to one or more trained machine learning models.

12. The system of any of claims 8 to 11, wherein the executable instructions add one or more post-processing output layers to the one or more trained machine learning models that are retrained with the adversarial protection based on one or more adversarial protection protocols.

13. The system of any of claims 8 to 12, wherein the executable instructions:

automatically implement one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models; or

receive, from a user, one or more adversarial protection protocols to be used for providing the one or more hardened machine learning models.

14. The system of any of claims 8 to 13, wherein the executable instructions:

monitor and track each state of the one or more trained machine learning models while being retrained; detect a training collapse to the one or more trained machine learning models during the retraining; or

enable one or more roll-back strategies for the one or more trained machine learning models during a retraining operation.

15. A computer program product for securing trained machine learning models in a computing environment by one or more processors, the computer program product comprising:

a computer readable storage medium readable by a processing circuit and storing instructions for execution by the processing circuit for performing a method according to any of claims 1 to 7.

16. A computer program stored on a computer readable medium and loadable into the internal memory of a digital computer, comprising software code portions, when said program is run on a computer, for performing the method of any of claims 1 to 7.