Adversarial Attacks and Defense Analysis

1. Objective of the experiments.

The goal of this experimental work is to evaluate the effectiveness of defensive techniques against security threats in a machine learning model. The end state of this study is to quantify the impact of security threats on machine learning models and demonstrate how defensive techniques can mitigate these threats. The effectiveness of the defensive techniques will be evaluated by comparing the percentage of correctly labeled instances in the model with and without the application of defensive techniques.

2. Experiment setup. E.g., describe the software/programming language you use, the input you have for the experiments, the basic test scenarios (like how many nodes in the system.)

This experiment will start with a machine learning model that labels pictures of animals and tests the model against new raw data getting a percentage of correctness. Then I will introduce a security attack utilizing the Fast Gradient Sign Attack and measuring correctness of labels. Finally, I will introduce a defensive technique to measure correctness compared to the control and after the attack.

3. Description of the program. Briefly explain the implementation you have done, e.g., modules and functions. Please also submit your program as part of the deliverables.

4. Experiment results. Describe the measurements have you made, the variables that change.  
Present numerical data in tables or graphs. Discuss the results you obtained, and what they mean in practice.