

Algoritmi za brzo učenje na neprijateljskim primjerima

Autor: Dominik Jambrović

Mentor: prof. dr. sc. Siniša Šegvić

Sadržaj

1. Uvod
2. Robusno učenje
3. Eksperimenti
4. Budući rad
5. Diskusija

Uvod

- sigurnost umjetne inteligencije
- neprijateljski primjeri
- zatrovani podatci

Robusno učenje

Klasični načini

- učenje metodom FGSM
- učenje metodom PGD

Novi načini

- „besplatno” učenje
- brzo učenje
- nadogradnje brzog učenja

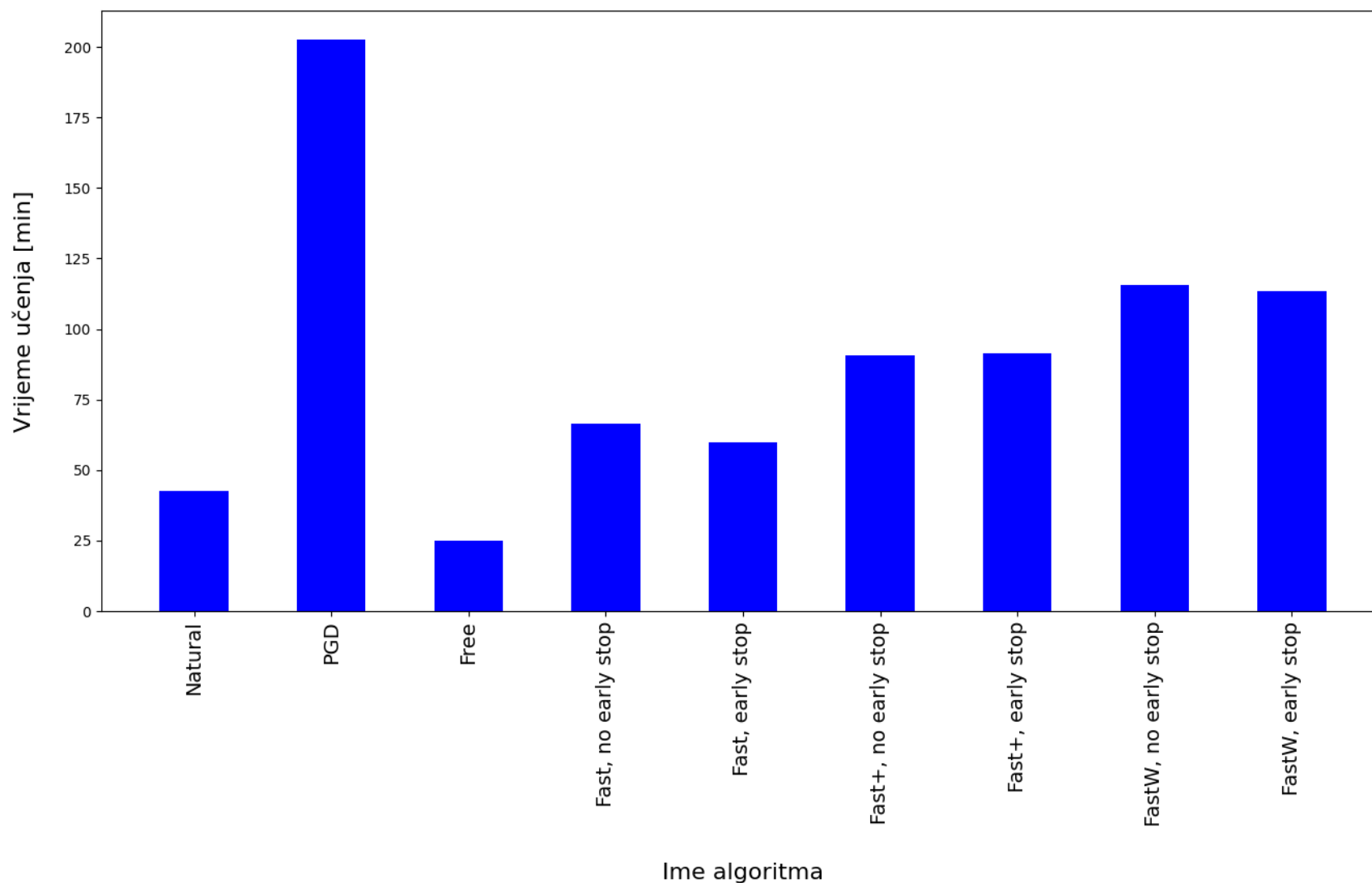
Eksperimenti

Izvođenje eksperimenata

- skup podataka CIFAR-10
- arhitektura ResNet-18
- korišćenje računanja u mješovitoj preciznosti
- izvođenje eksperimenata na platformi Kaggle (2x NVIDIA T4)

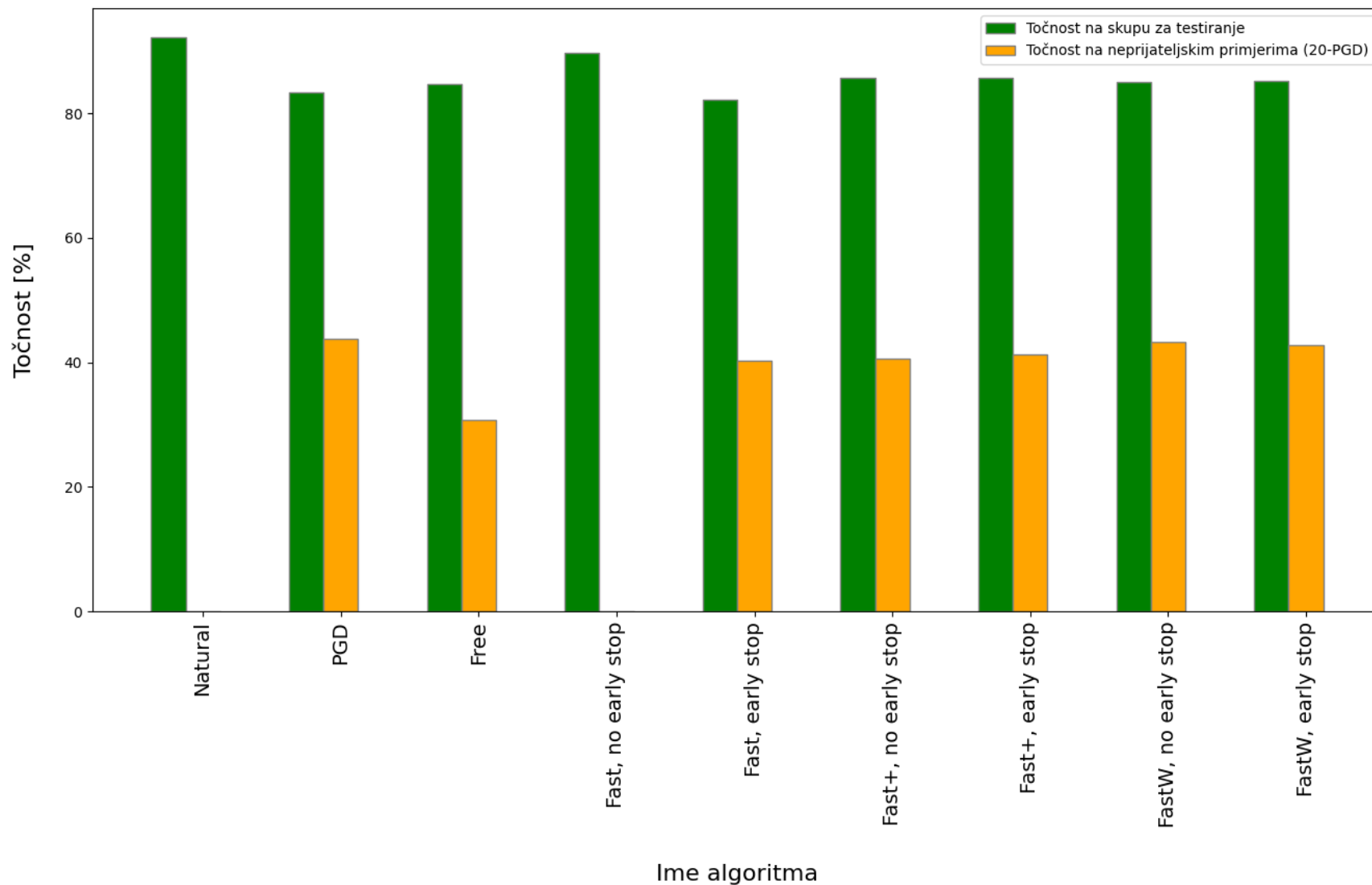
Usporedba vremena učenja

Vrijeme učenja za različite algoritme



Usporedba točnosti

Usporedba točnosti za različite algoritme



Generativna svojstva modela

Pravi razred: cat



Pravi razred: ship



Prirodne slike

Pravi razred: airplane



Pravi razred: frog



Pravi razred: automobile



Predviđeni razred: dog



Predviđeni razred: automobile



Prirodno učenje

Predviđeni razred: ship



Predviđeni razred: automobile



Predviđeni razred: truck



Predviđeni razred: dog



Predviđeni razred: airplane



Algoritam PGD

Predviđeni razred: bird



Predviđeni razred: deer



Predviđeni razred: dog



Predviđeni razred: frog



Predviđeni razred: automobile



Algoritam FreeAdv

Predviđeni razred: ship



Predviđeni razred: deer



Predviđeni razred: truck



Predviđeni razred: dog



Predviđeni razred: airplane



Algoritam FastAdv, Early

Predviđeni razred: bird



Predviđeni razred: deer



Predviđeni razred: truck



Predviđeni razred: dog



Predviđeni razred: automobile



Algoritam FastAdv+, Early

Predviđeni razred: bird



Predviđeni razred: cat



Predviđeni razred: truck



Predviđeni razred: dog



Predviđeni razred: automobile



Algoritam FastAdvW, Early

Predviđeni razred: bird



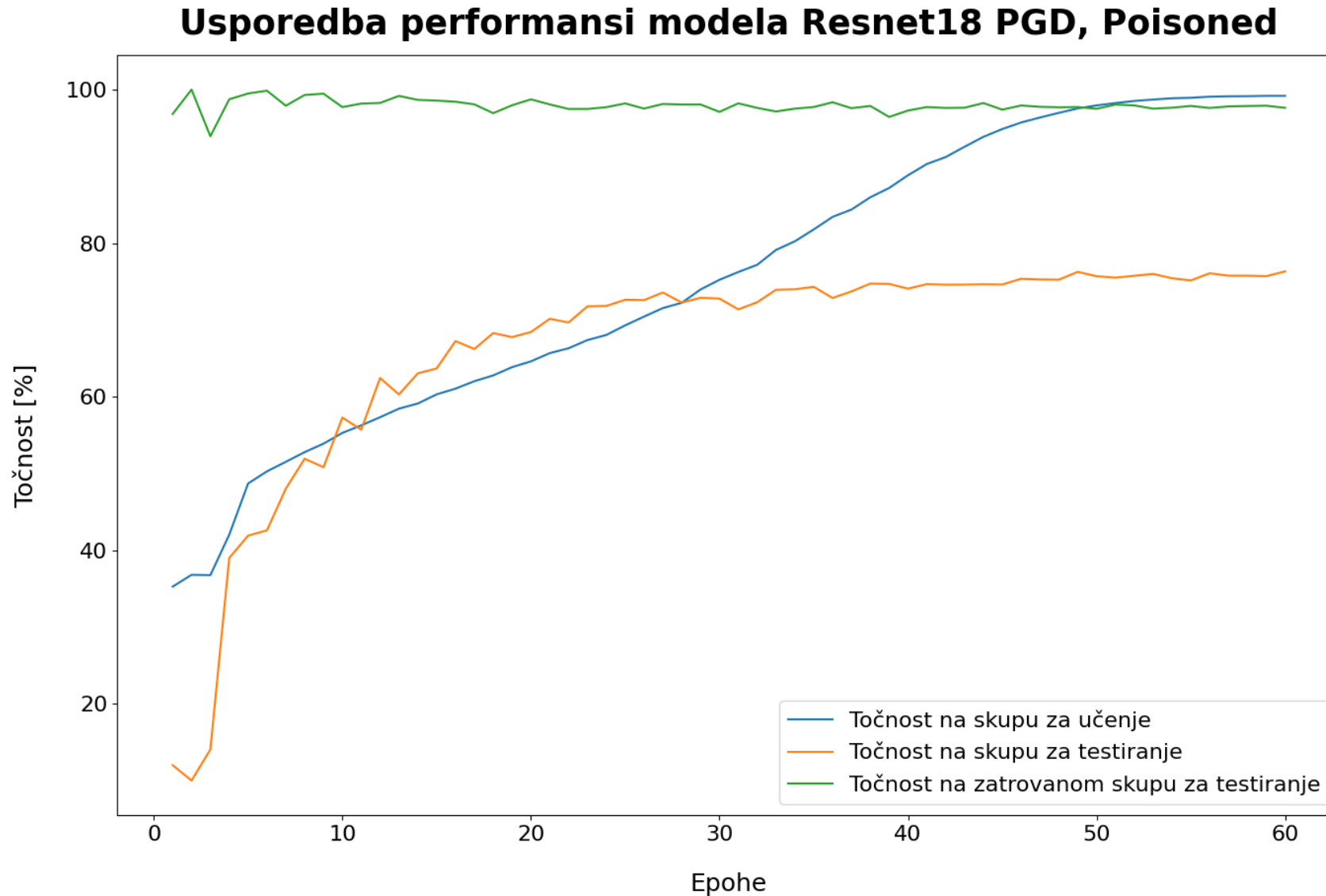
Predviđeni razred: deer



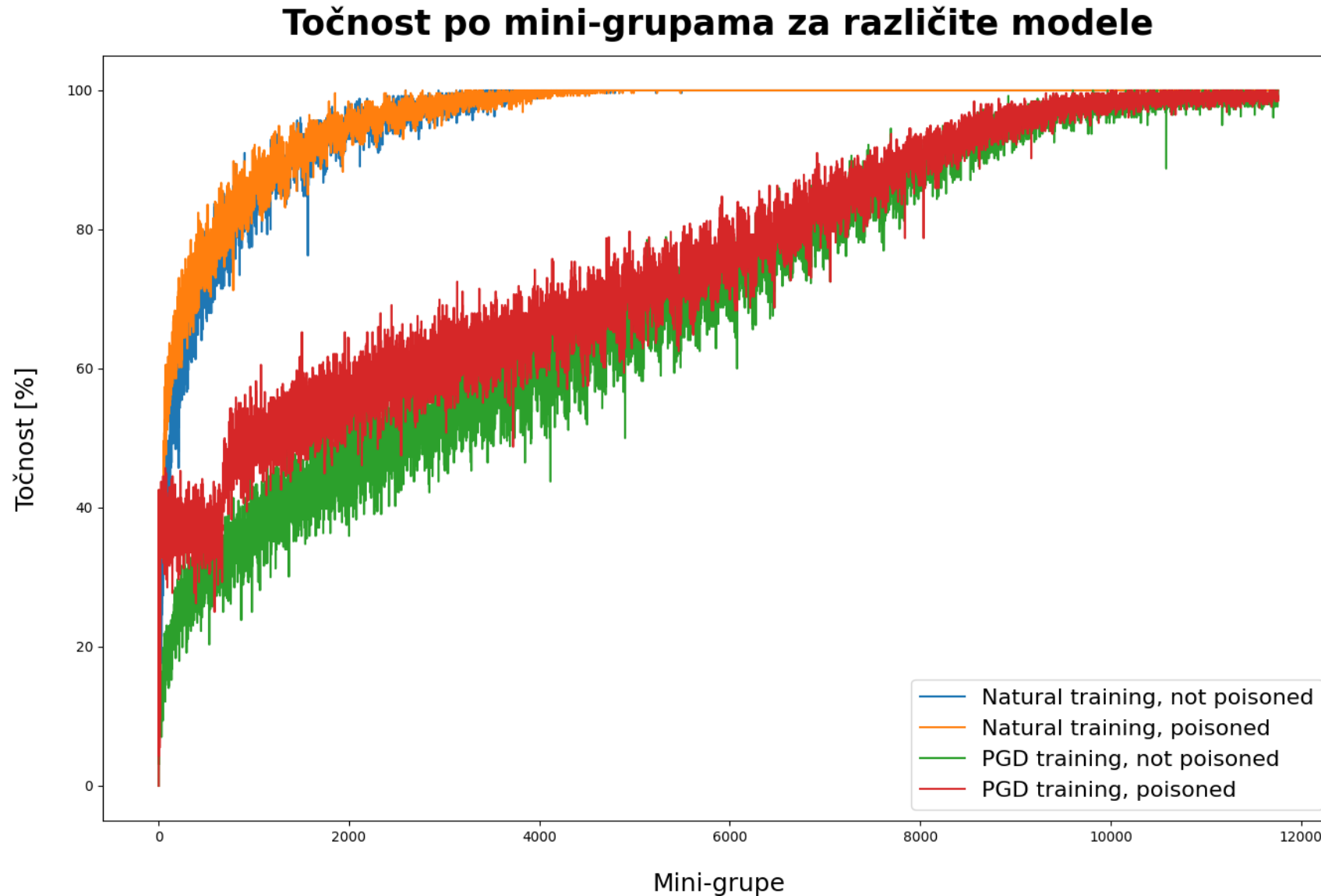
Predviđeni razred: cat



Usporedba performansi, PGD

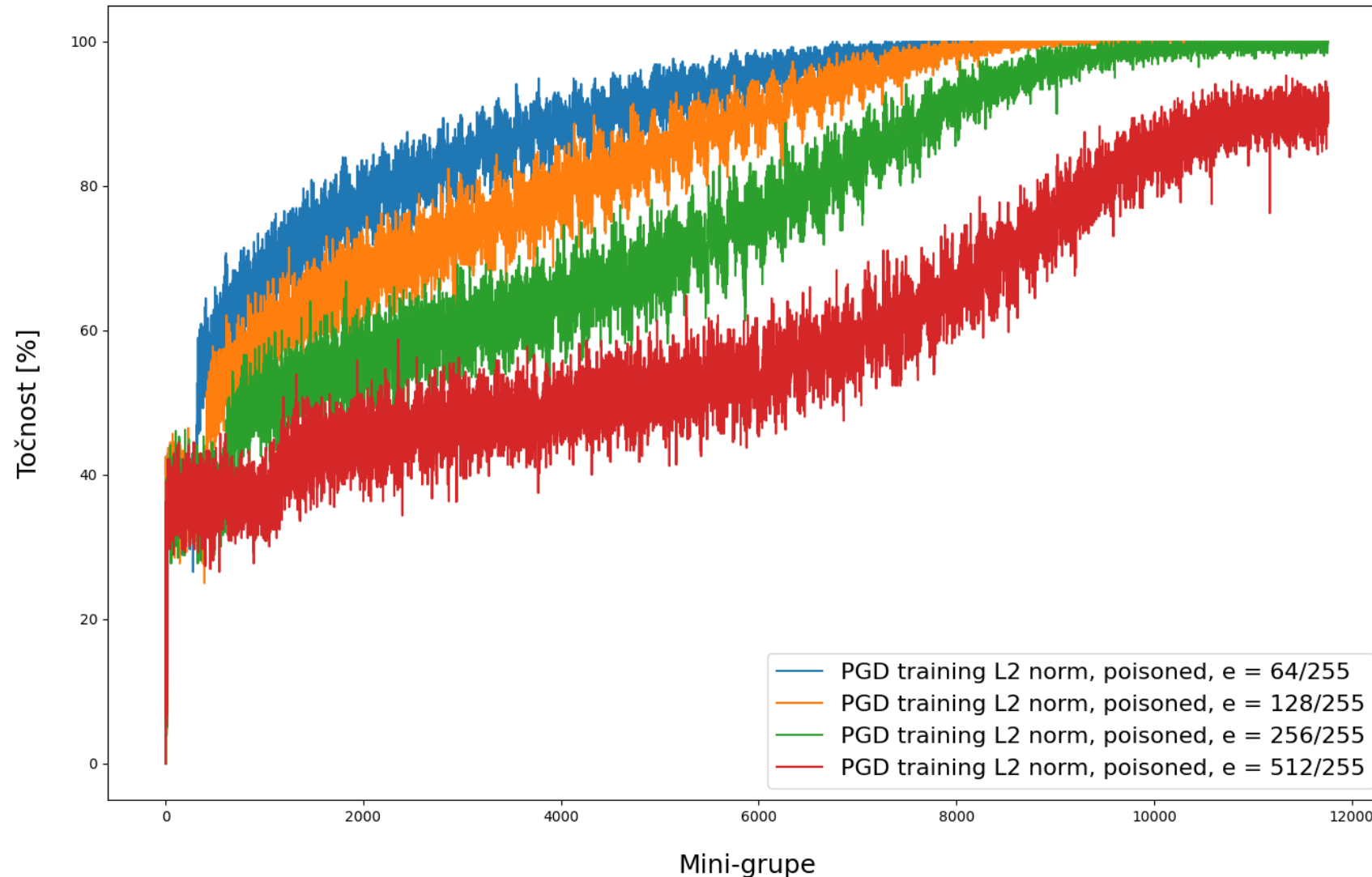


Usporedba točnosti, mini-grupe



Usporedba točnosti, L2 norma

Točnost po mini-grupama za različite modele



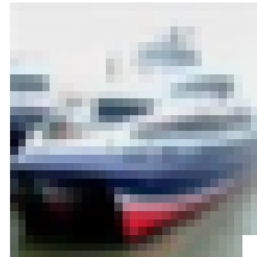
Neprijateljski primjeri

Zatrovane slike

Pravi razred: cat



Pravi razred: ship

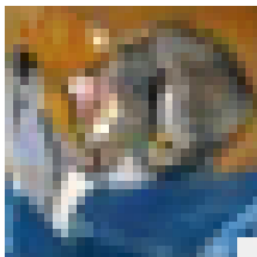


Pravi razred: airplane

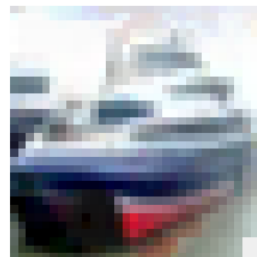


Norma L_{∞}

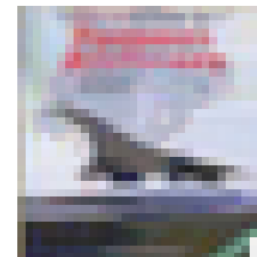
Predviđeni razred: automobile



Predviđeni razred: ship

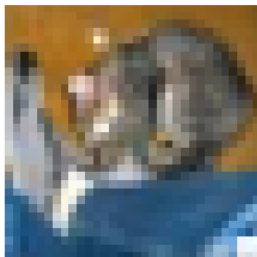


Predviđeni razred: automobile

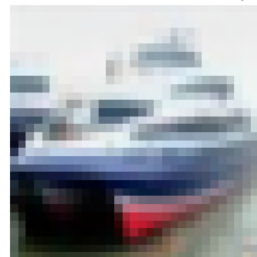


Norma L_2

Predviđeni razred: automobile



Predviđeni razred: ship

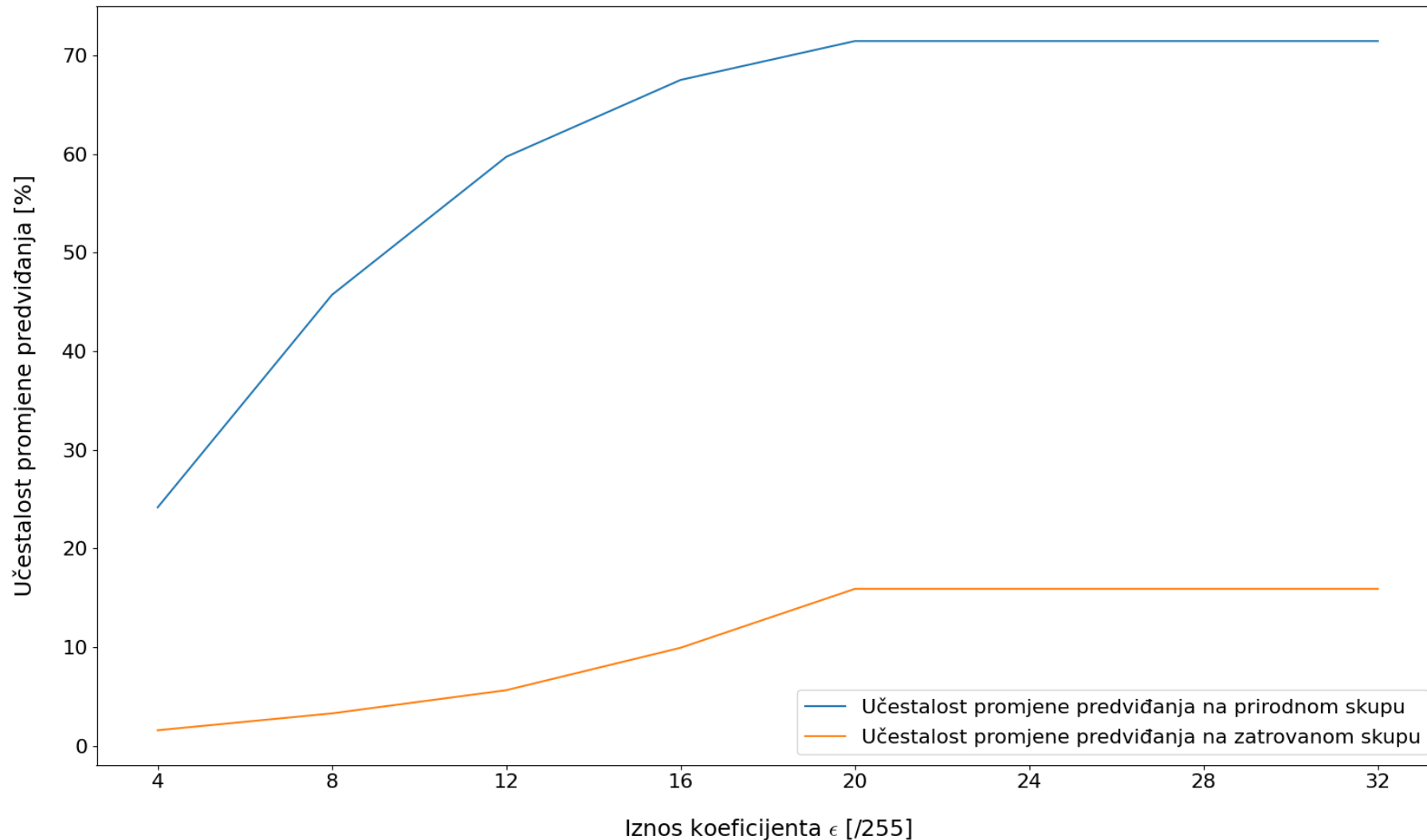


Predviđeni razred: airplane



Učestalost promjene predviđanja

Učestalost promjene predviđanja na prirodnom i zatrovanom skupu



Budući rad

Algoritmi za brzo robusno učenje

- primijeniti metode na kompleksnije arhitekture
- mogućnost kombiniranja „besplatnog” i brzog učenja

Detekcija zatrovanih podataka

- proučiti utjecaj korištenja L1 norme
- mjeriti uspješnost detekcije zatrovanih podataka praćenjem promjena predviđanja

Diskusija