Algoritmi za brzo učenje na neprijateljskim primjerima

Autor: Dominik Jambrović

Mentor: prof. dr. sc. Siniša Šegvić

Uvod

- sigurnost umjetne inteligencije
- neprijateljski primjeri
- zatrovani podatci

Robusno učenje

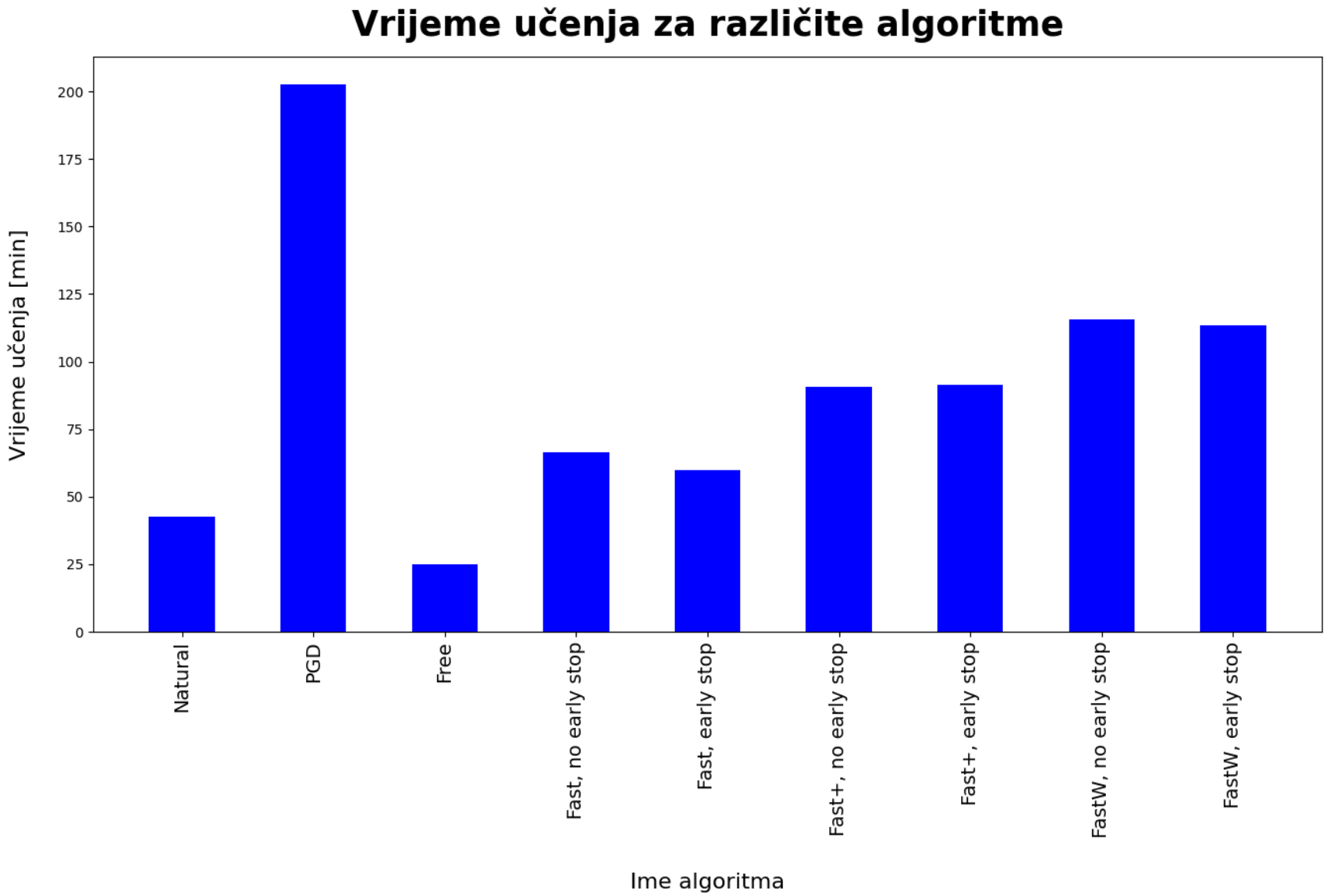
Klasični načini

- učenje metodom FGSM
- učenje metodom PGD

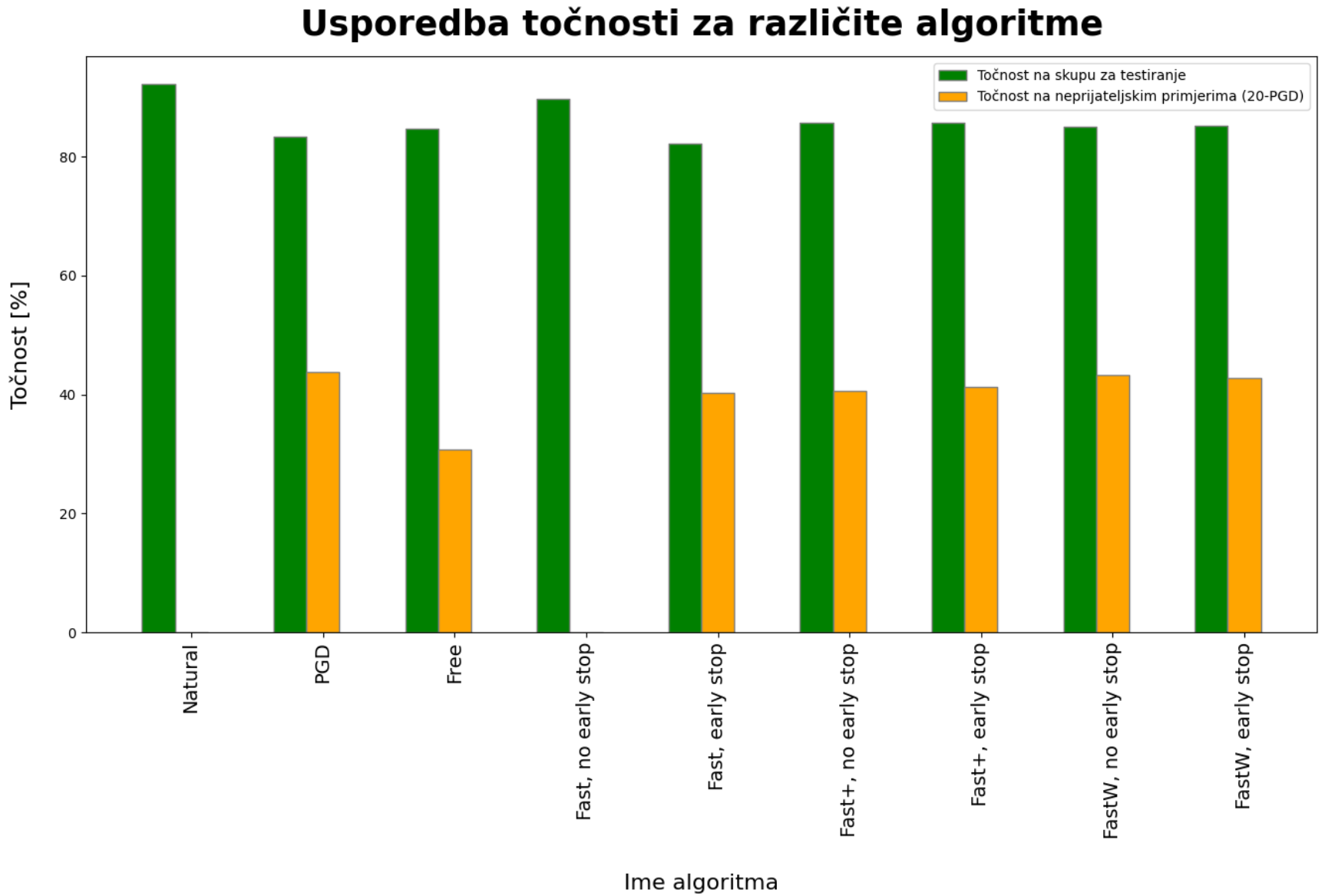
Novi načini

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

Usporedba vremena učenja



Usporedba točnosti



Generativna svojstva modela

Prirodne slike

Pravi razred: cat 

Pravi razred: ship 

Pravi razred: airplane 

Pravi razred: frog 

Pravi razred: automobile 

Prirodno učenje

Predviđeni razred: dog 

Predviđeni razred: automobile 

Predviđeni razred: ship 

Predviđeni razred: automobile 

Predviđeni razred: truck 

Algoritam PGD

Predviđeni razred: dog 

Predviđeni razred: airplane 

Predviđeni razred: bird 

Predviđeni razred: deer 

Predviđeni razred: dog 

Algoritam FreeAdv

Predviđeni razred: frog 

Predviđeni razred: automobile 

Predviđeni razred: ship 

Predviđeni razred: deer 

Predviđeni razred: truck 

Algoritam FastAdv, Early

Predviđeni razred: dog 

Predviđeni razred: airplane 

Predviđeni razred: bird 

Predviđeni razred: deer 

Predviđeni razred: truck 

Algoritam FastAdv+, Early

Predviđeni razred: dog 

Predviđeni razred: automobile 

Predviđeni razred: bird 

Predviđeni razred: cat 

Predviđeni razred: truck 

Algoritam FastAdvW, Early

Predviđeni razred: dog 

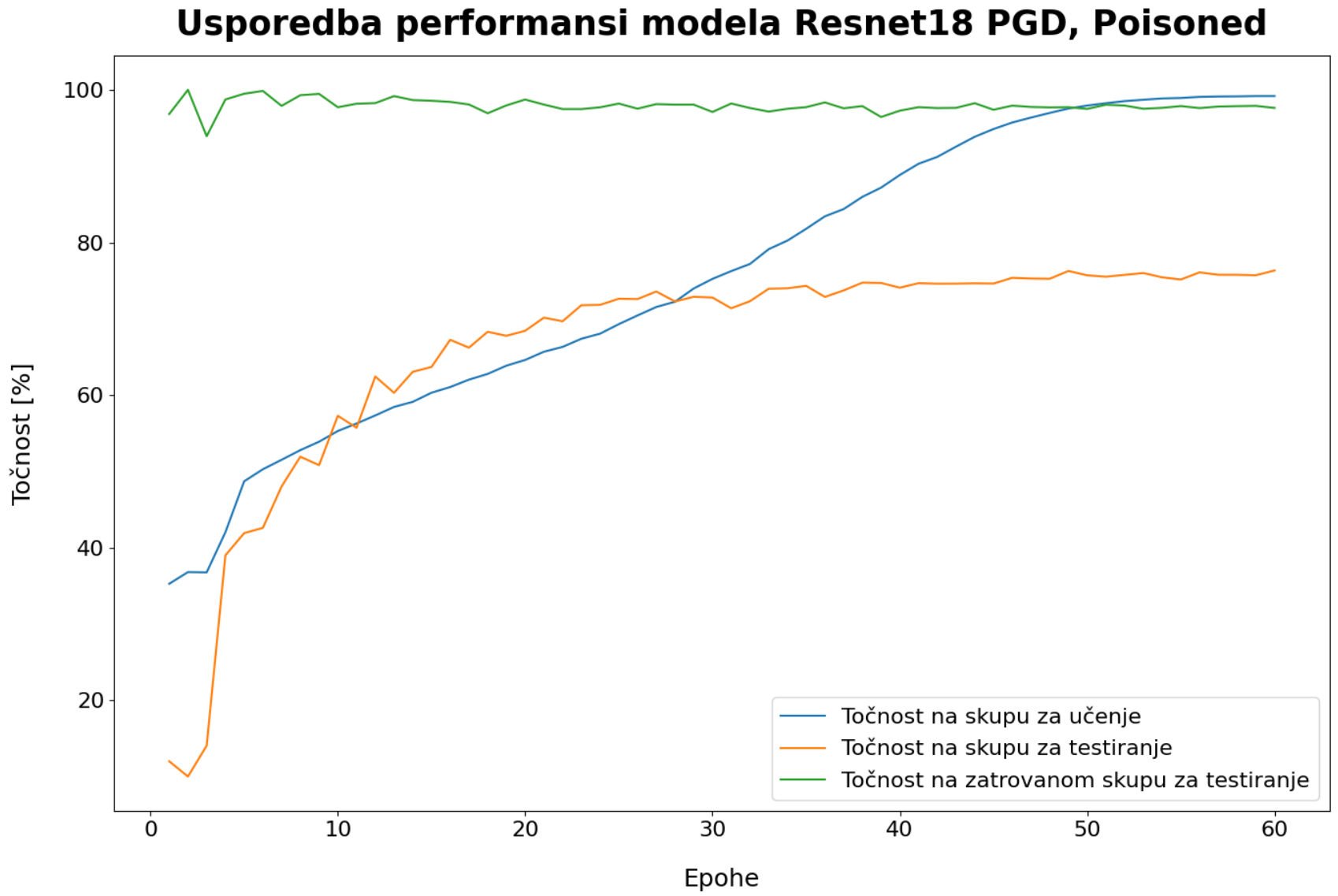
Predviđeni razred: automobile 

Predviđeni razred: bird 

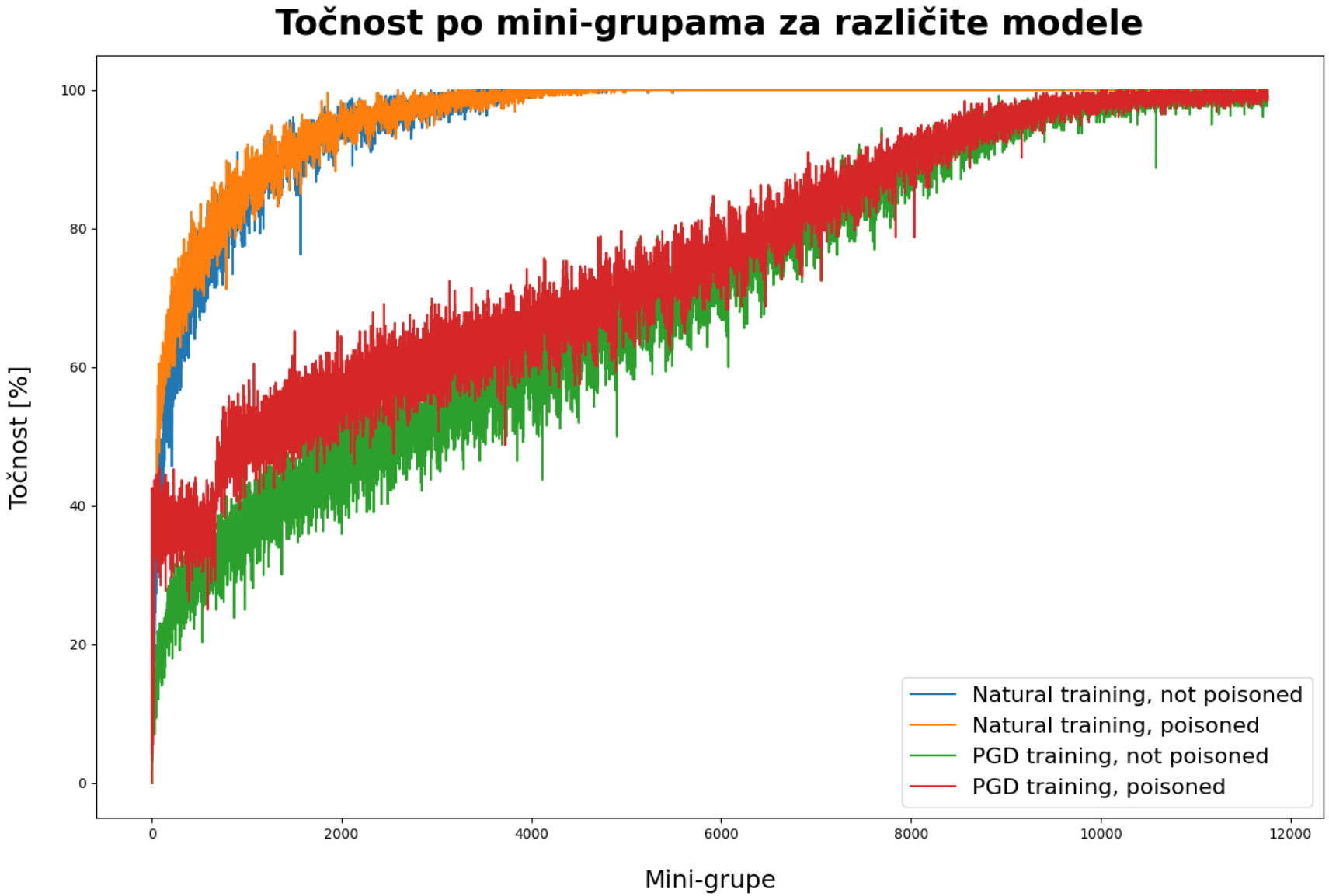
Predviđeni razred: deer 

Predviđeni razred: cat 

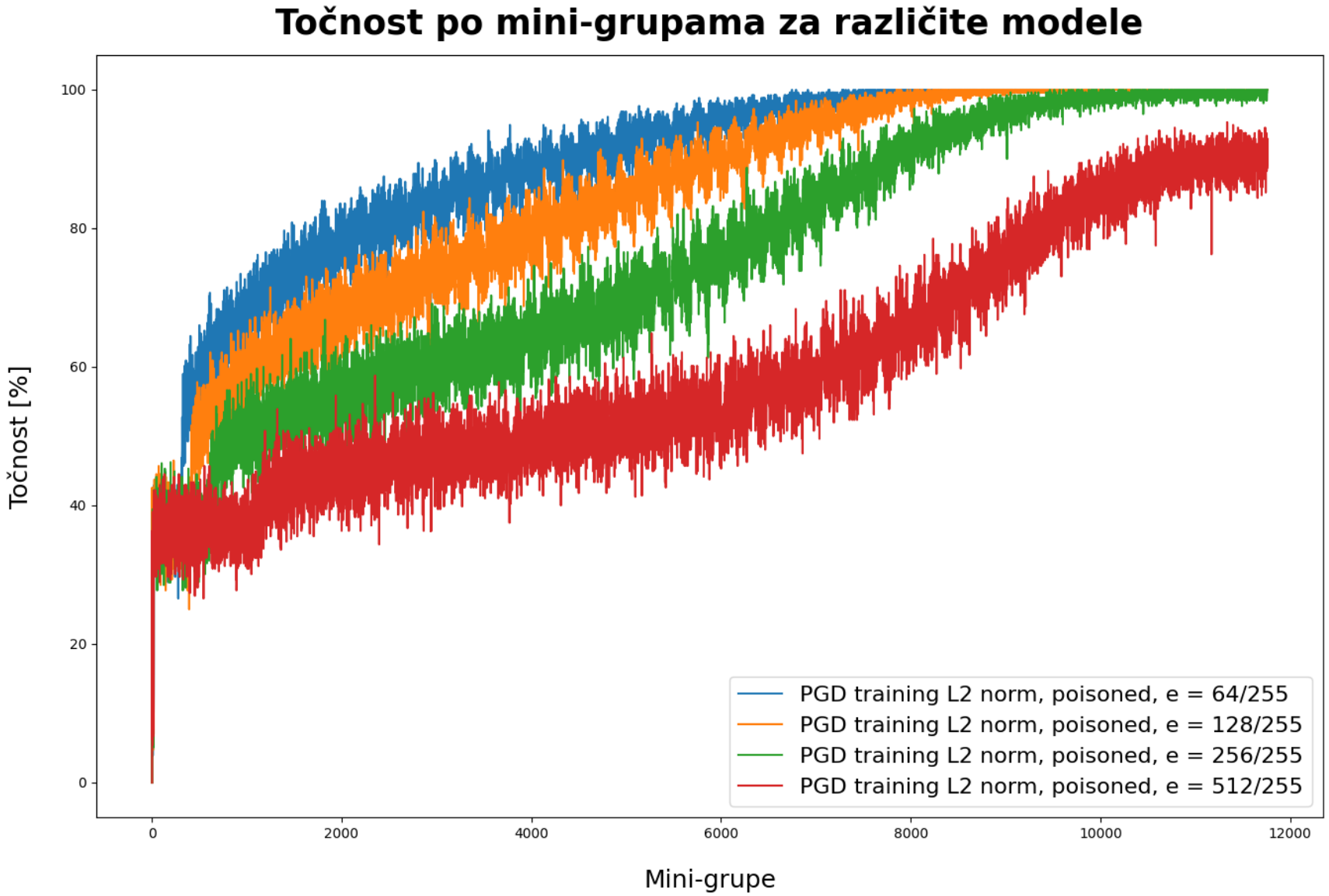
Usporedba performansi, PGD



Usporedba točnosti, mini-grupe



Usporedba točnosti, L2 norma



Neprijateljski primjeri

Zatrovane slike

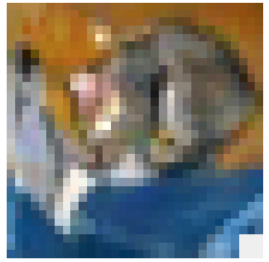
Pravi razred: cat

Pravi razred: ship

Pravi razred: airplane

Norma L_∞

Predviđeni razred: automobile

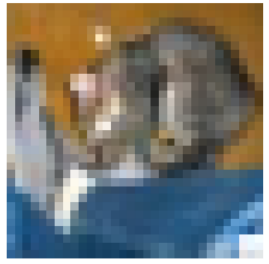


Predviđeni razred: ship

Predviđeni razred: automobile

Norma L2

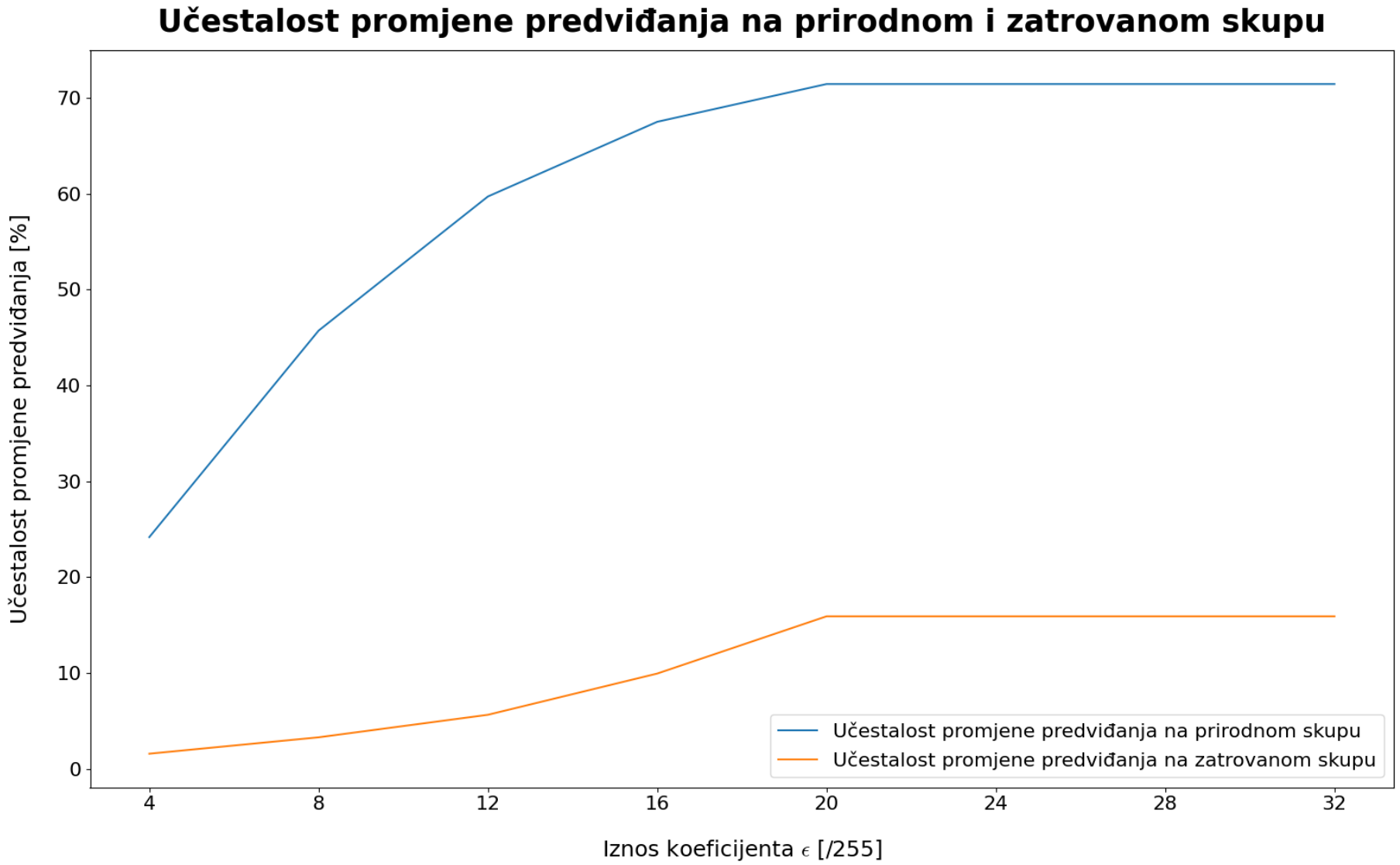
Predviđeni razred: automobile



Predviđeni razred: ship

Predviđeni razred: airplane

Učestalost promjene predviđanja



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

Hvala na pozornosti!