Algoritmi za brzo učenje na neprijateljskim primjerima

Autor: Dominik Jambrović

Mentor: prof. dr. sc. Siniša Šegvić

Sadržaj

Uvod

- sigurnost umjetne inteligencije
- neprijateljski primjeri
- zatrovani podatci

Robusno učenje

Klasični načini

- učenje metodom FGSM
- učenje metodom PGD

Novi načini

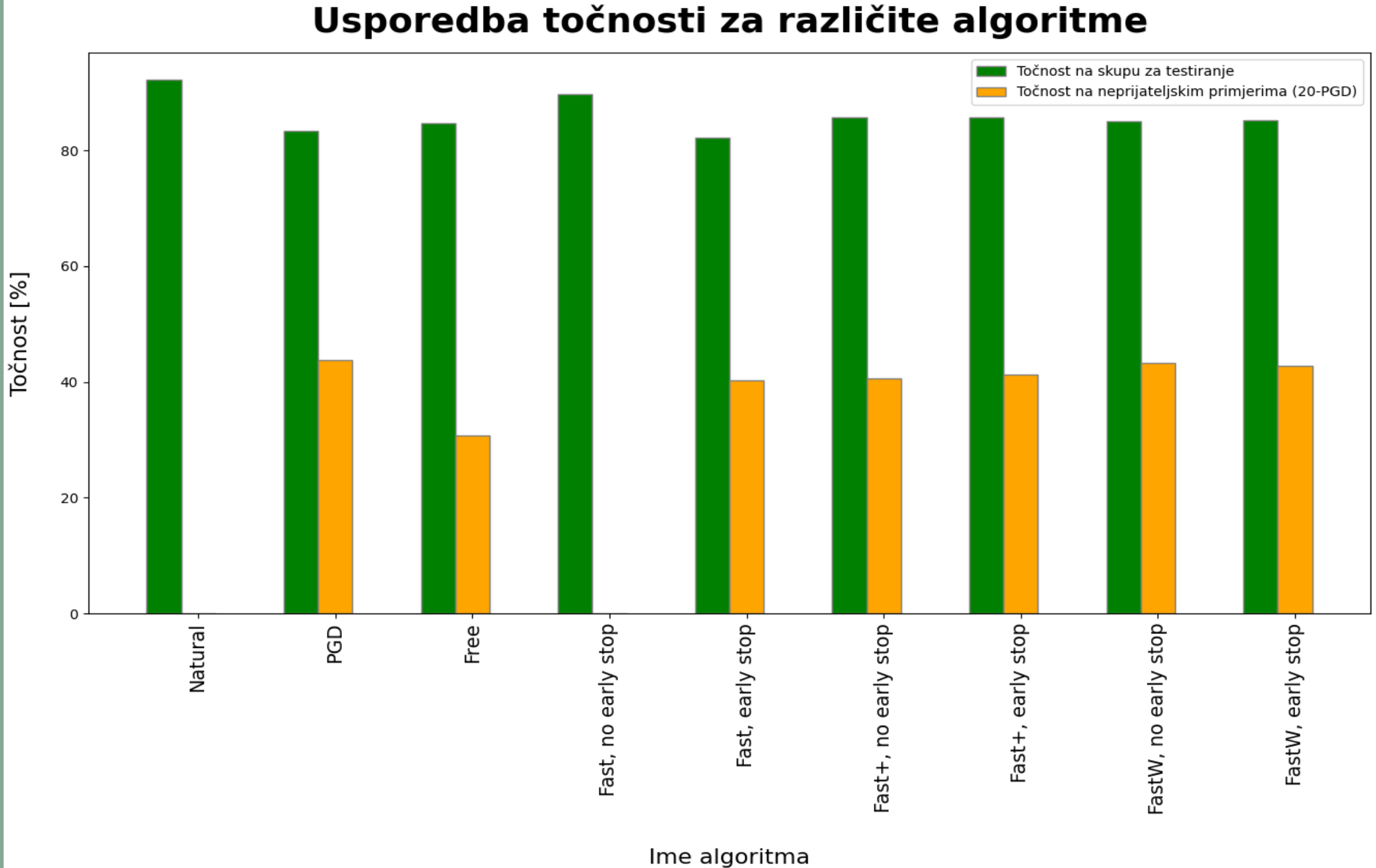
- „besplatno” učenje
- brzo učenje
- nadogradnje brzog učenja

Eksperimenti

Izvođenje eksperimenata

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


Vrijeme učenja za različite algoritme
Vrijeme učenja [min]
0
25
50
75
100
125
150
175
200
Natural
PGD
Free
Fast, no early stop
Fast, early stop
Fast+, no early stop
Fast+, early stop
FastW, no early stop
FastW, early stop
Ime algoritma


Usporedba točnosti za različite algoritme
Točnost na skupu za testiranje
Točnost na neprijateljskim primjerima (20-PGD)
Točnost [%]
0
20
40
60
80
Natural
PGD
Free
Fast, no early stop
Fast, early stop
Fast+, no early stop
Fast+, early stop
FastW, no early stop
FastW, early stop
Ime algoritma

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

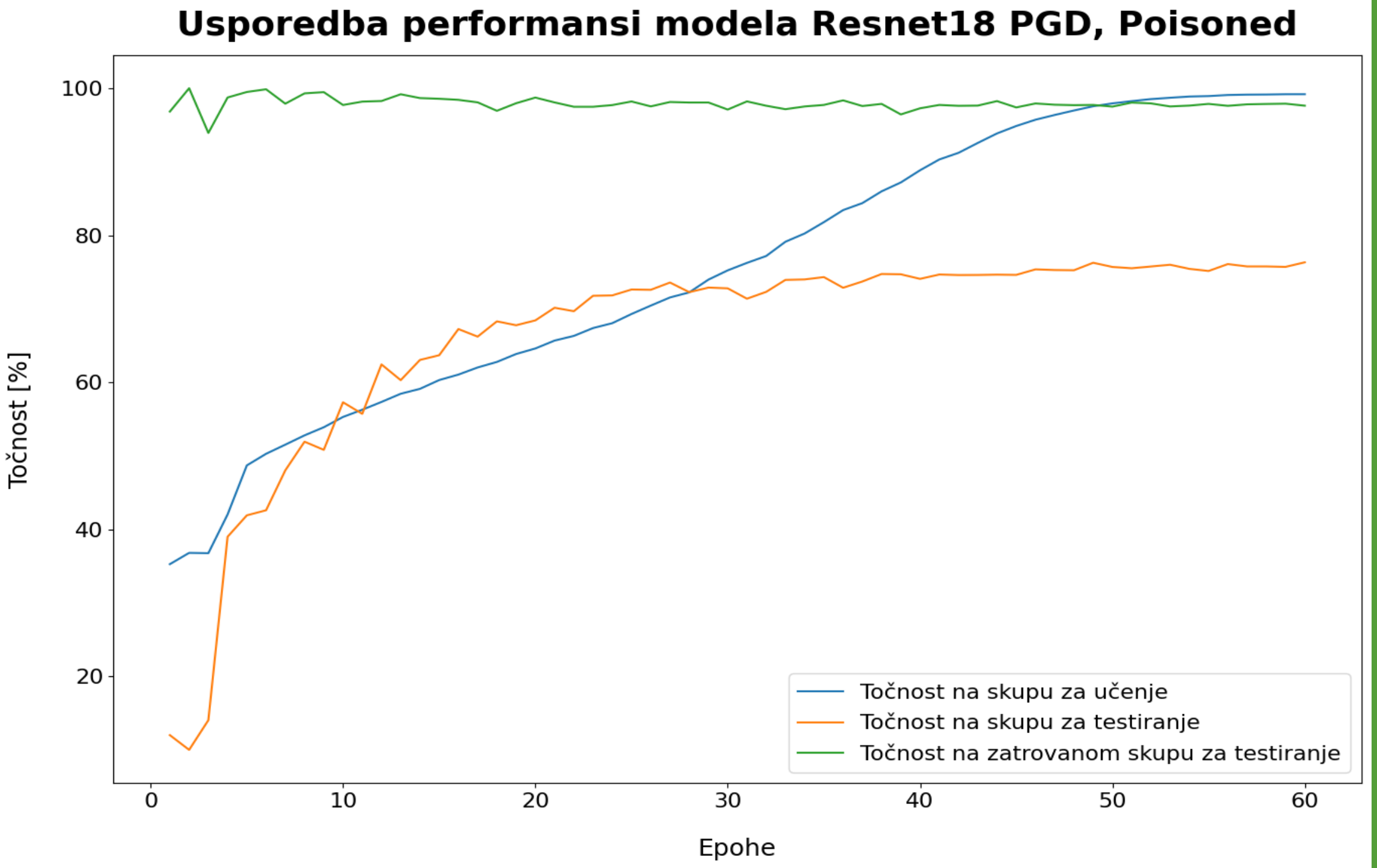
Predviđeni razred: dog | Predviđeni razred: airplane | Predviđeni razred: bird | Predviđeni razred: deer | Predviđeni razred: truck

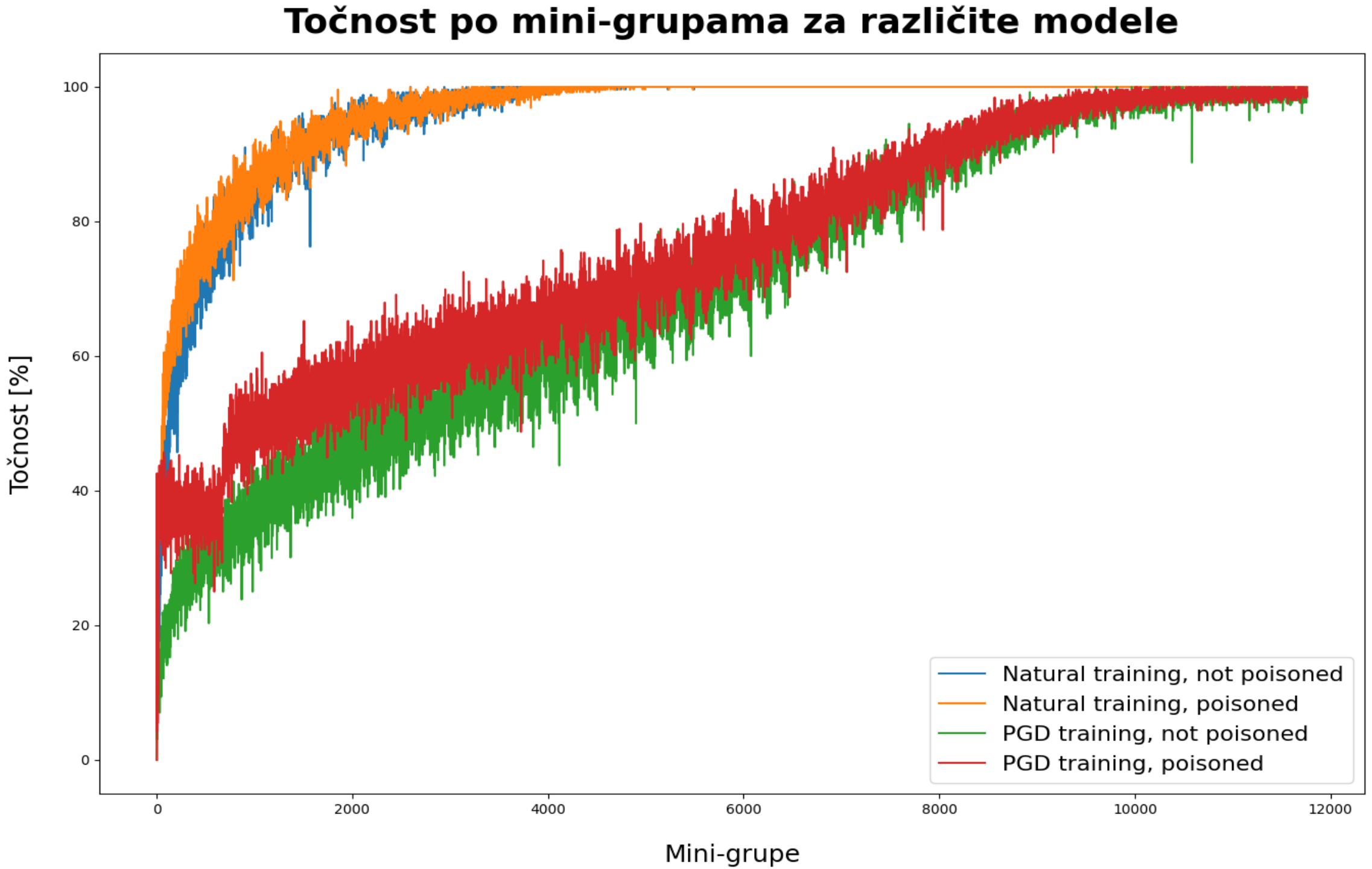
Algoritam FastAdv+, Early

Predviđeni razred: dog | Predviđeni razred: automobile | Predviđeni razred: bird | Predviđeni razred: cat | Predviđeni razred: truck

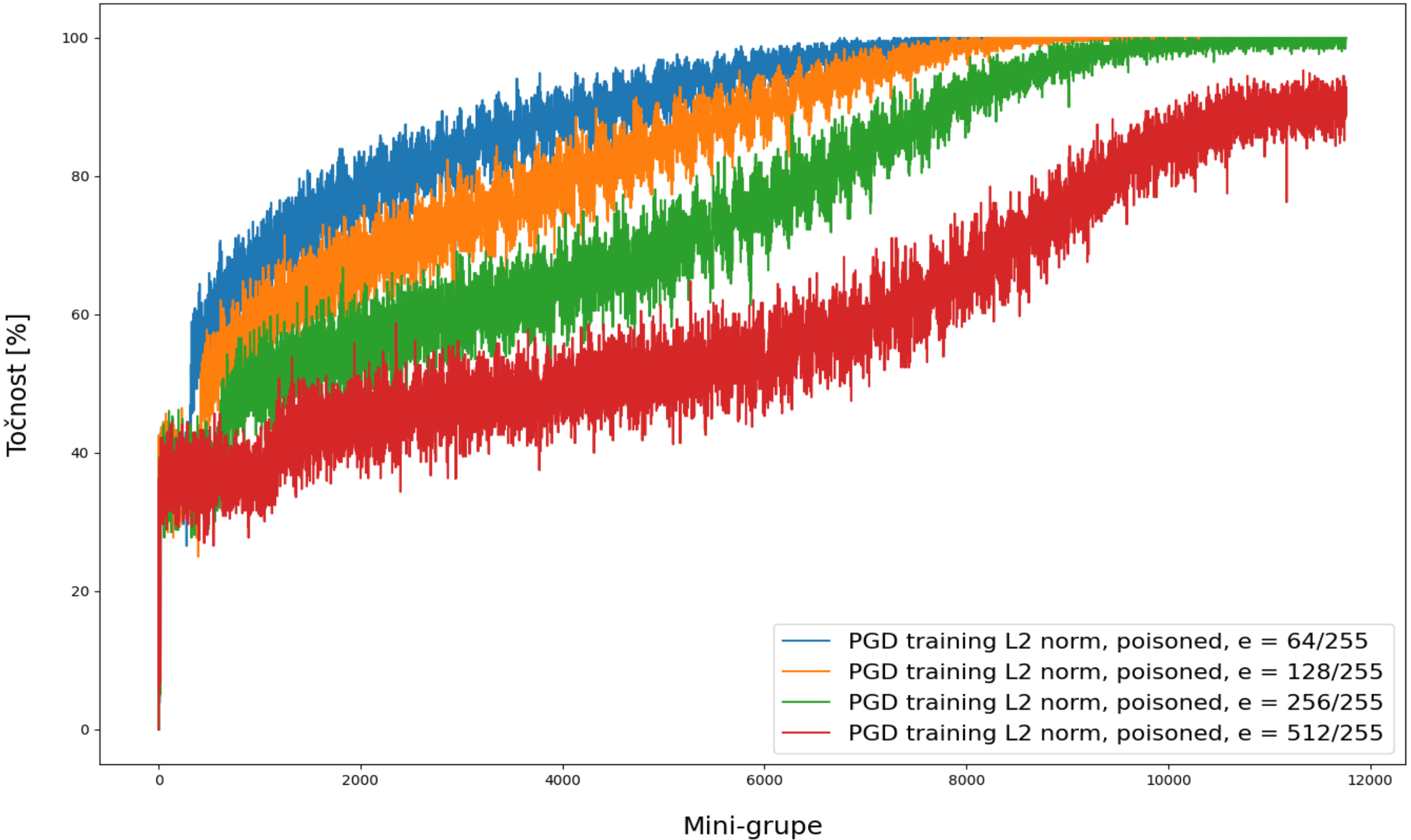
Algoritam FastAdvW, Early

Predviđeni razred: dog | Predviđeni razred: automobile | Predviđeni razred: bird | Predviđeni razred: deer | Predviđeni razred: cat


Usporedba performansi modela Resnet18 PGD, Poisoned
Točnost [%]
100
80
60
40
20
0
10
20
30
40
50
60
Epohe
Točnost na skupu za učenje
Točnost na skupu za testiranje
Točnost na zatrovanom skupu za testiranje


Točnost po mini-grupama za različite modele
Točnost [%]
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
12000
Mini-grupe
Natural training, not poisoned
Natural training, poisoned
PGD training, not poisoned
PGD training, poisoned

Točnost po mini-grupama za različite modele



Zatrovane slike

Pravi razred: cat

Pravi razred: ship

Pravi razred: airplane

Norma L_∞

Predviđeni razred: automobile

Predviđeni razred: ship

Predviđeni razred: automobile

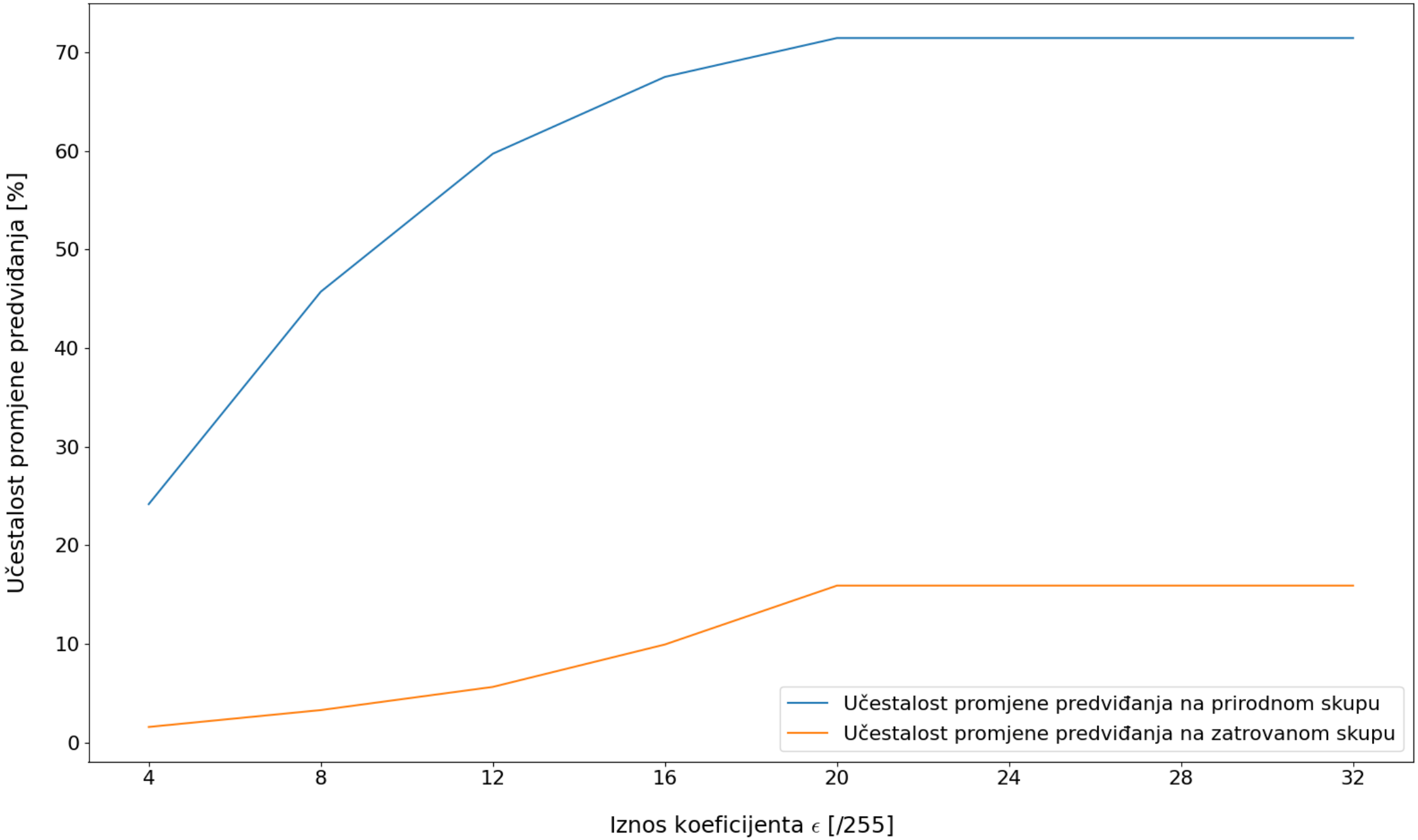
Norma L2

Predviđeni razred: automobile

Predviđeni razred: ship

Predviđeni razred: airplane

Učestalost promjene predviđanja 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