

# 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. Zaključak
5. Budući rad
6. Diskusija

# Uvod

- sigurnost umjetne inteligencije
- neprijateljski primjeri
- zatrovani podatci

# Robusno učenje

## Klasični načini

- učenje metodom FGSM
- učenje metodom PGD

## Novi načini

- „besplatno” učenje PGD-om
- brzo učenje pažljivo inicijaliziranim FGSM-om
- kombinirano robusno učenje

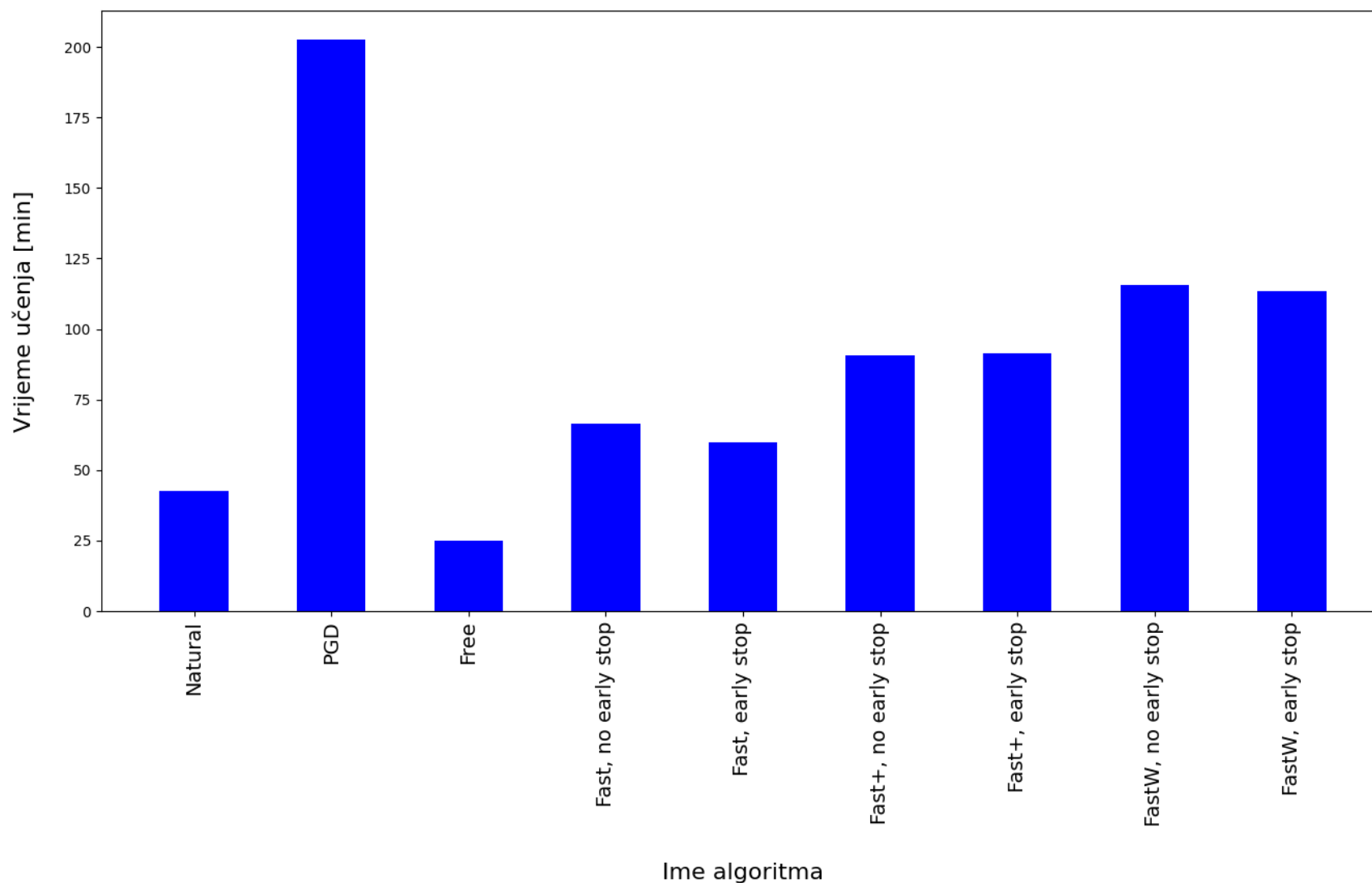
# Eksperimenti

# Izvođenje eksperimenata

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

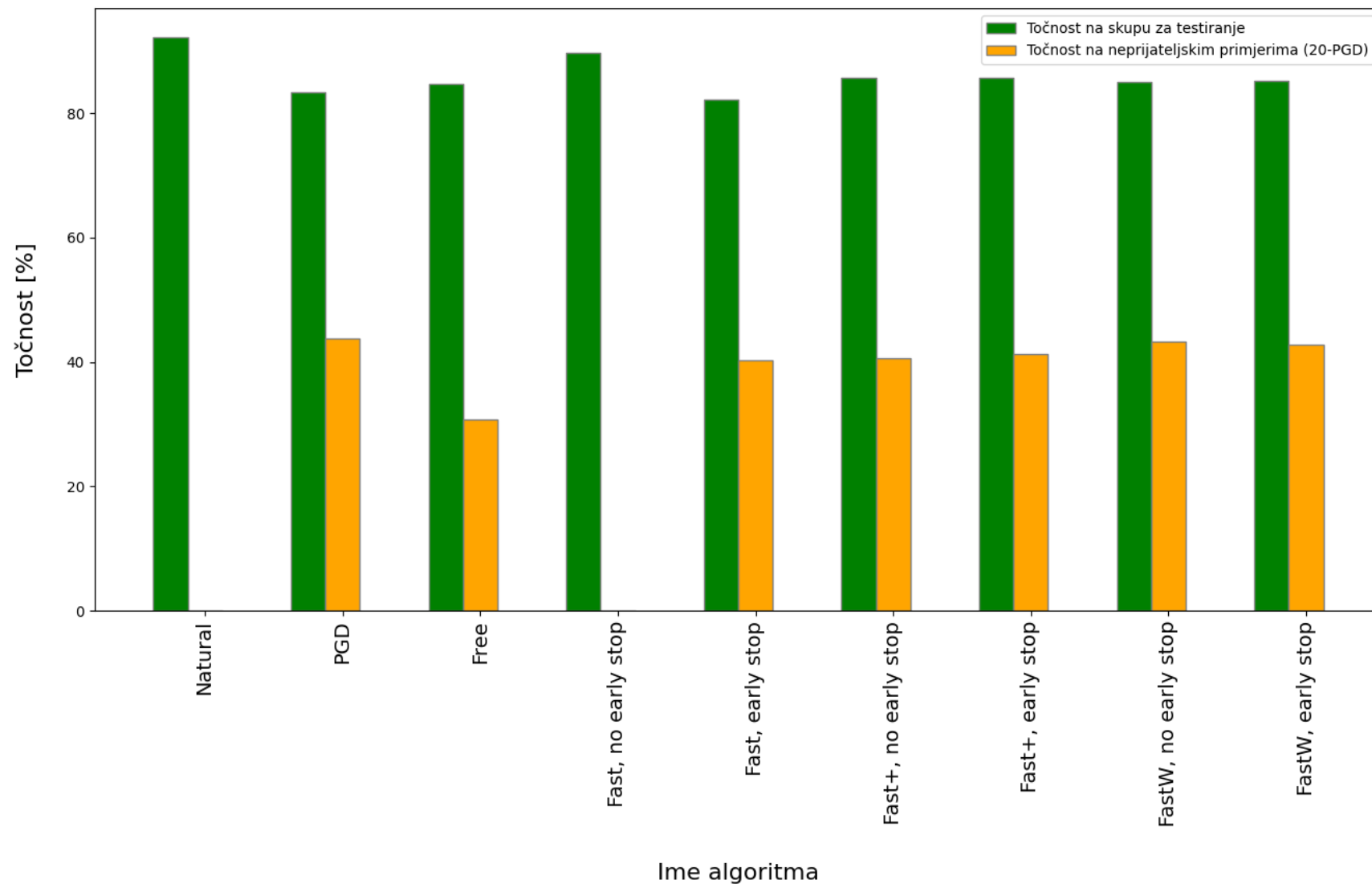
# Usporedba vremena učenja

Vrijeme učenja za različite algoritme



# Usporedba točnosti

## Usporedba točnosti za različite algoritme





# Generativna svojstva modela

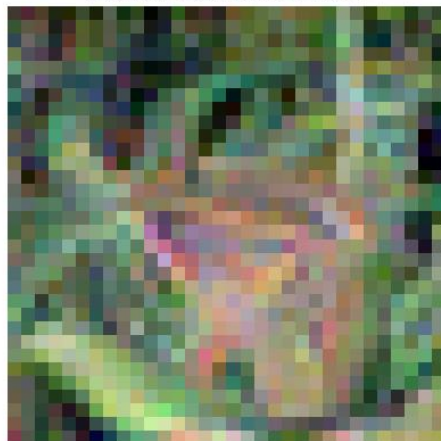
**Prirodna slika**

Pravi razred: frog



**Prirodno učenje**

Predviđeni razred: automobile



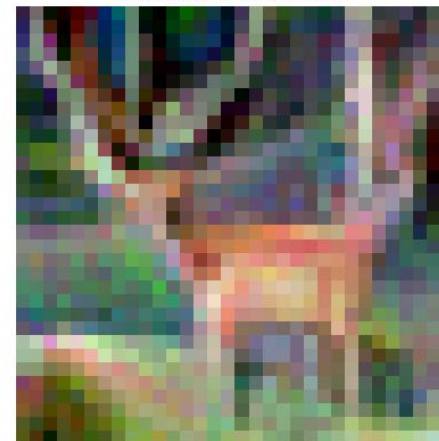
**Algoritam PGD**

Predviđeni razred: deer



**Algoritam FreeAdv**

Predviđeni razred: deer



**Algoritam FastAdv, Early**

Predviđeni razred: deer



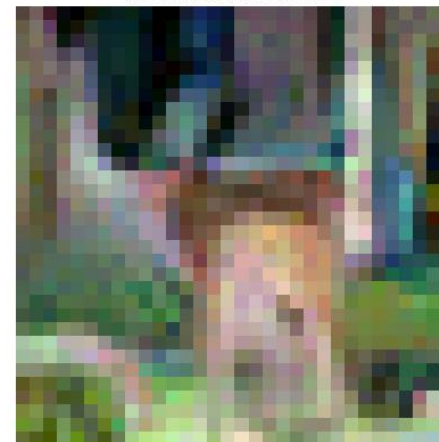
**Algoritam FastAdv+, Early**

Predviđeni razred: cat

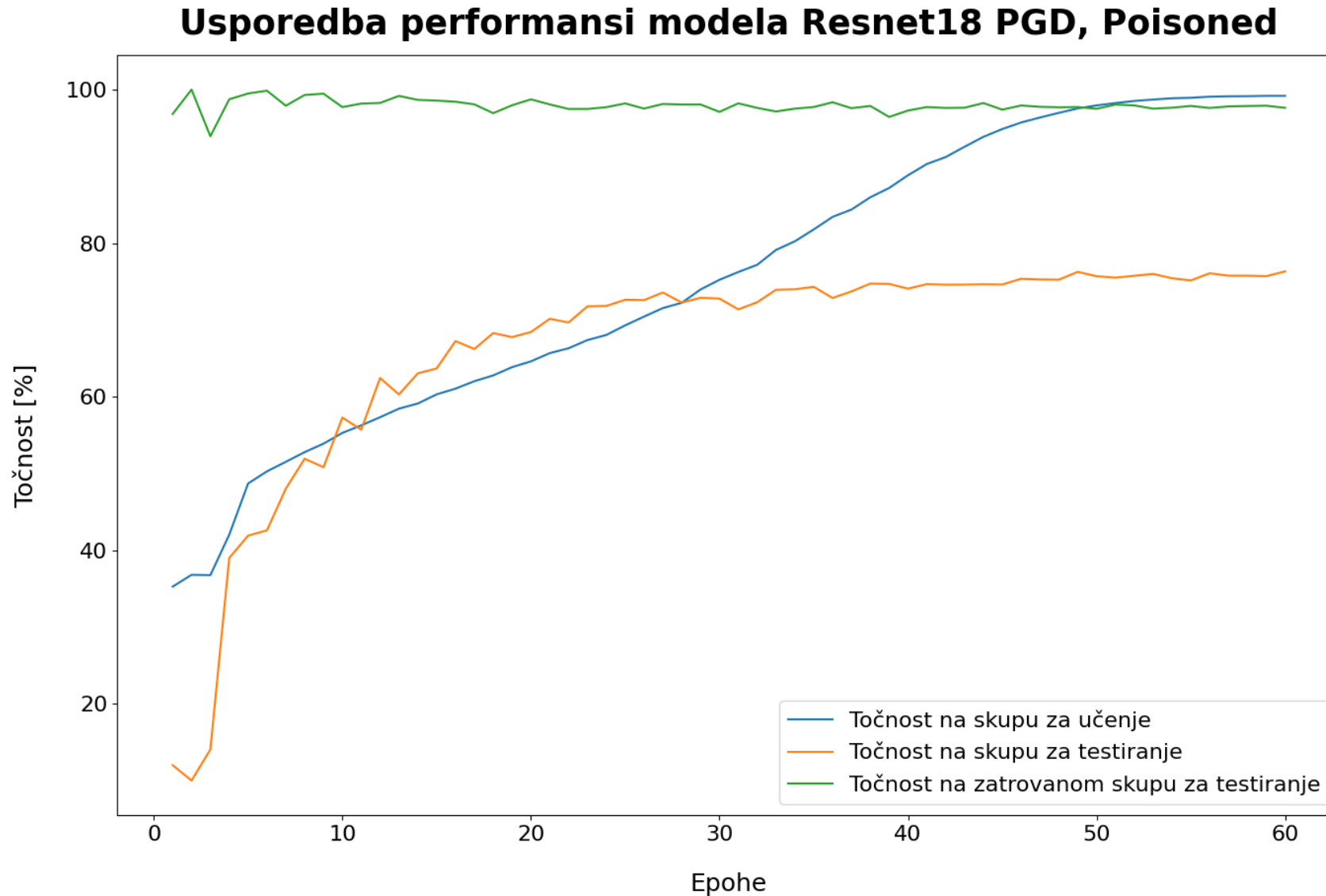


**Algoritam FastAdvW, Early**

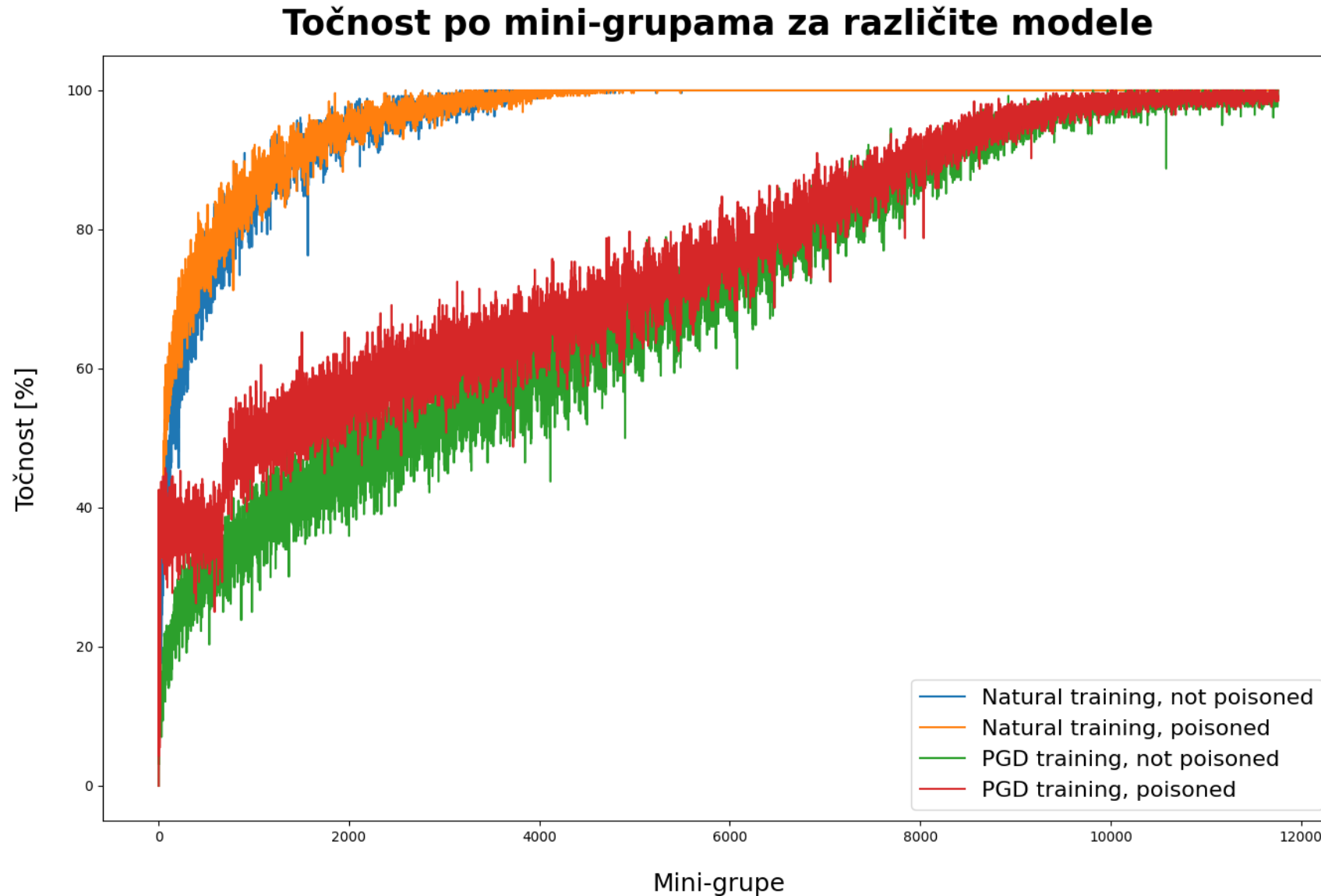
Predviđeni razred: deer



# Usporedba performansi, PGD

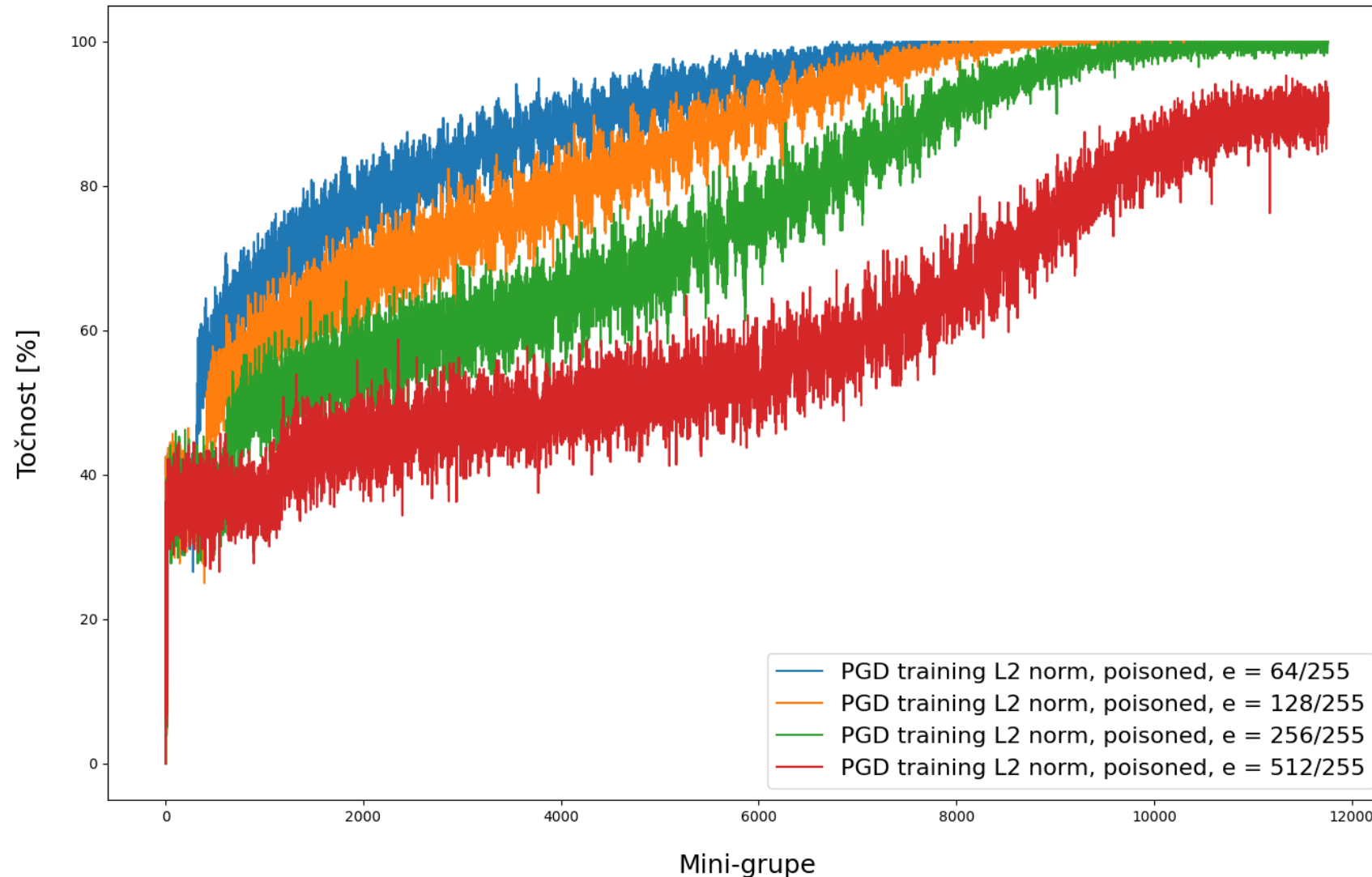


# Usporedba točnosti, mini-grupe



# Usporedba točnosti, L2 norma

Točnost po mini-grupama za različite modele



# Usporedba Lp normi

## Zatrovane slike

Pravi razred: cat



Pravi razred: ship

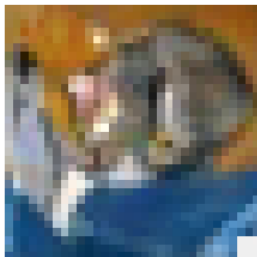


Pravi razred: airplane



## Norma $L_{\infty}$

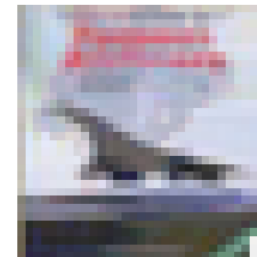
Predviđeni razred: automobile



Predviđeni razred: ship



Predviđeni razred: automobile

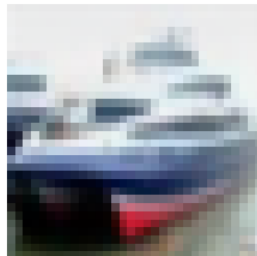


## Norma $L_2$

Predviđeni razred: automobile



Predviđeni razred: ship

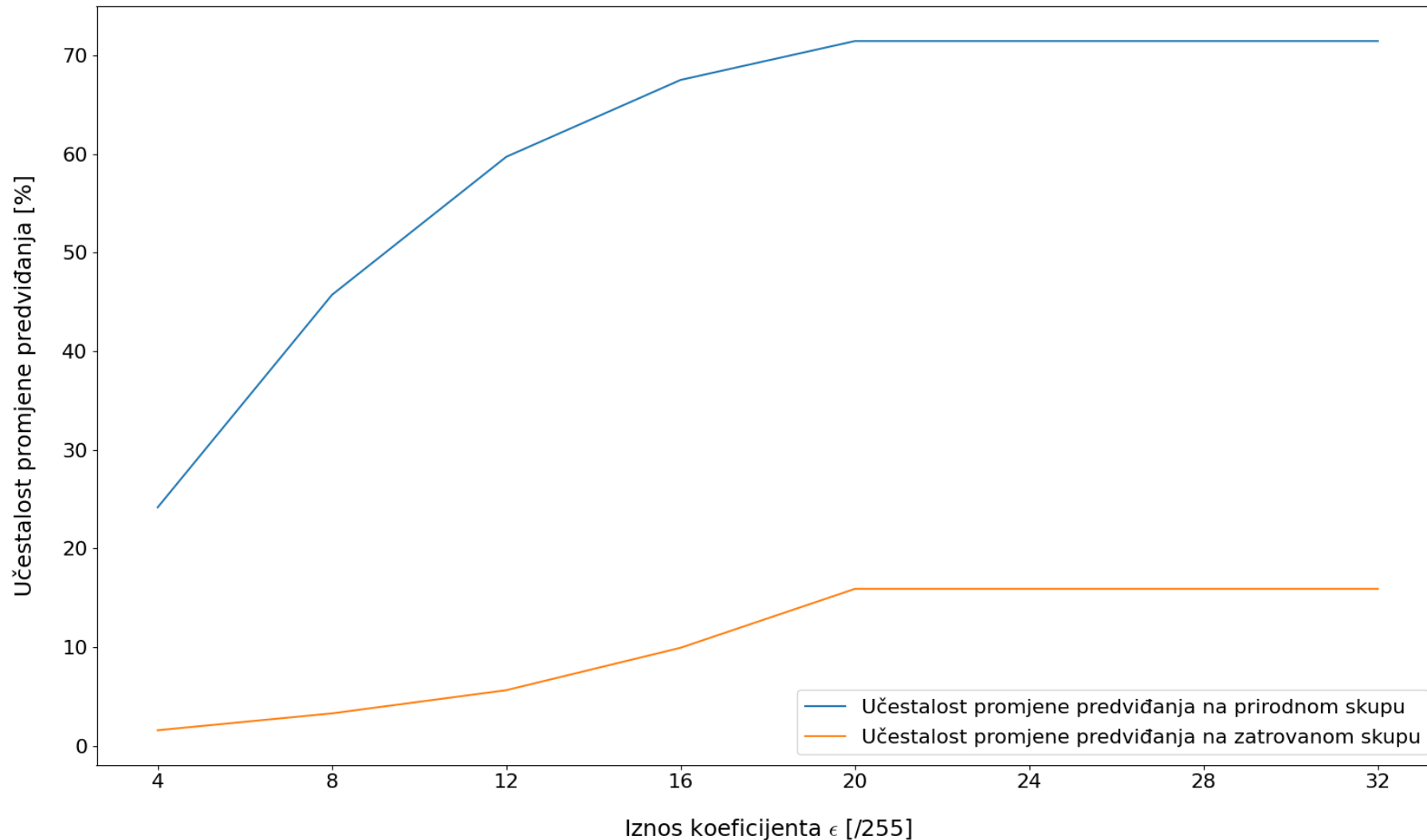


Predviđeni razred: airplane



# Učestalost promjene predviđanja

Učestalost promjene predviđanja na prirodnom i zatrovanom skupu



# Zaključak

Algoritmi za brzo robusno učenje

- algoritam FastAdvW pruža točnost i robusnost podjednake onima koje pruža algoritam PGD, ali uz skoro dvostruko kraće vrijeme učenja

Detekcija zatrovanih podataka

- mogućnost korištenja mjerenja točnosti na skupu za učenje po mini-grupama kao detektora postojanja zatrovanih podataka
- alternativno, mogućnost korištenja praćenja promjena predviđanja modela

# 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