SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1071

Analiza napada kroz stražnja vrata s obzirom na razdvojivosti zatrovanih podataka

Luka Glavinić

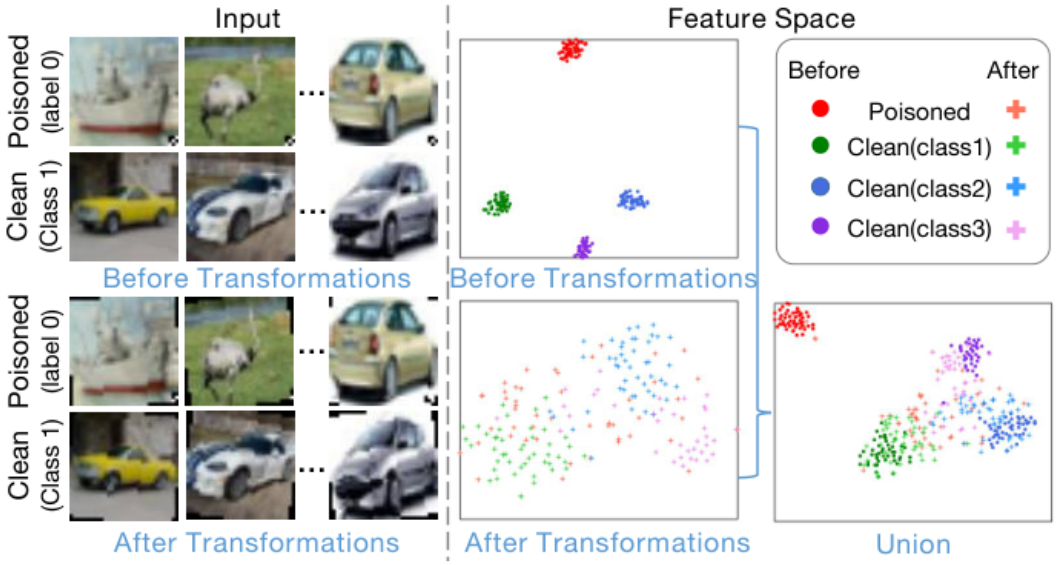
Zagreb, srpanj 2025.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`.

SADRŽAJ

1. Uvod

Učenje dubokih neuronskih mreža često zahtijeva veliku količinu podataka za treniranje, koji se ponekad dobiju od nepouzdanih izvora. No, nepovjerljivi izvori podataka mogu prouzročiti ozbiljne sigurnosne prijetnje. Jedna od tipičnih prijetnji je napad kroz stražnja vrata koji se temelji na trovanju podataka (Gu et al. (2019)). Napadač umeće zatrovane podatke u skup za učenje i ovim putem povećava korelaciju između okidača i ciljanog razreda. Ovakav napad je vrlo opasan jer je malen broj ovakvih otrovanih uzoraka dovoljan da model potpuno drugačije klasificira otrovane podatke. Zapaženo je u (Huang et al. (2022)) da se podaci otrovani s okidačima vrlo često skupljaju u prostoru značajki modela otrovanim napadom kroz stražnja vrata kao što se vidi na slici 1.1. Otrovani uzorci sadrže različite objekte na slikama, iz različitih razreda, ali pokazuje se da model ignorira informacije s tih podataka. Drugim riječima reprezentacijama značajki otrovanih uzoraka dominiraju okidači, a ne sami objekti s tih slika.



Slika 1.1: Prikaz čistih i otrovanih podataka iz CIFAR-10 prema (Alex Krizhevsky (2009)), i t-SNE vizualizacije (Laurens Van der Maaten (2008)) njihovih značajki otrovanog modela. Može se primijetiti u "Union" odsječku da su promjene na čistim podacima puno manje od promjena na otrovanim podacima. Slika preuzeta iz rada (Chen et al. (2022)).

Posljedično ovome model će vrlo dobro predviđati razrede čistih uzoraka, no svaki otrovani uzorak će opredijeliti ciljanom razredu.
Zbog ovoga motivacija ovog rada bila je razviti klasifikator koji s visokom točnošću može prepoznati otrovani podatak i razlikovati ga od čistog podatka, te time omogućiti filtraciju odnosno pouzdano razdvajanje otrovanih podataka od čistih.
U sljedećim poglavljima će biti objašnjena teorija iza treniranja dubokih modela, vrste i implementacije raznih napada kroz stražnja vrata. Nakon toga će biti objašnjene dvije mjere koje su se koristile za prepoznavanje otrovanog podatka i time omogućile treniranje binarnog klasifikatora. Na kraju će se objasniti i prikazati rezultati obavljenih eksperimenata i dati zaključak.

2. Duboko učenje

Duboko učenje je jedna od podgrana strojnog učenja i umjetne inteligencije. Prema (LeCun et al. (2015)) duboko učenje omogućava računalnim modelima obrađivanje podataka i učenje reprezentacije tih podataka s više različitih razina apstrakcije. Ovakve su metode dramatično "podigle letvicu" problemima: prepoznavanja govora, prepoznavanja vizualnih objekata, detektiranju objekata i raznim drugim domenama poput otkrivanju lijekova i genomike. Duboko učenje koristi algoritme unatrašnje propagacije za ažuriranje unutarnjih parametara dubokog modela kojima pokušava otkriti skrivene uzorke i značenja u velikim skupovima podataka (LeCun et al. (2015)). Duboke konvolucijske mreže (Aloysius i Geetha (2017)) omogućile su nove proboje u obrađivanju slika, videa, govora i zvuka, a pri tome su povratne mreže (Grossberg (2013)) obasjale svjetlo na uspjeh obrade sekvencijalnih podataka poput teksta i govora. Prije obrade rada metode obrane od napada kroz stražnja vrata bitno je znati pojmove koji će biti objašnjeni u sljedećem potpoglavlju.

2.1. Nadzirano učenje

U dubokom i strojnom učenju, nadzirano učenje je paradigma gdje se model trenira koristeći ulazne objekte i željene izlaze koje su često oznake kojima su ljudi ručno označili te objekte. Proces treniranja gradi funkciju koja preslikava nove podatke u očekivane izlazne vrijednosti. Optimalni scenarij dozvoljava algoritmu da točno predviđa izlazne vrijednosti neviđenih objekata. Ovo zahtijeva da algoritam učenja dobro generalizira od podataka za učenje. Treniranje se postiže matematičkim modelom koji traži minimum funkcije gubitka koja predstavlja koliko model griješi, odnosno koliko krivo predviđa izlazne vrijednosti. Da bi model dobro generalizirao kroz treniranje se mora raditi unakrsna evaluacija sa skupom podataka koje model nikad nije vidio i pratiti grešku na tim podacima.
Jedna od najpoznatijih i najkorištenijih funkcija gubitka je gubitak unakrsne entropije prikazan jednadžbom 2.1. U ovoj jednadžbi je prikazana formula za izračun gubitka

za jedan podatak pod indeksom j, gdje je z_j "logit" tog podatka, odnosno vrijednost ne-normaliziranog izlaza zadnjeg sloja modela. C u funkciji označava broj razreda, a z_y predstavlja "logit" pod indeksom od ispravnog razreda.

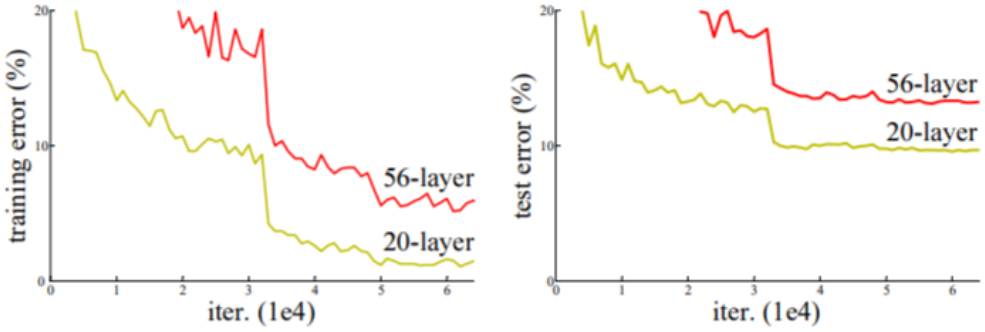
$$\mathcal{L} = -\log\left(\frac{e^{z_y}}{\sum_{j=1}^{C} e^{z_j}}\right) = -z_y + \log\left(\sum_{j=1}^{C} e^{z_j}\right) \tag{2.1}$$

Ova funkcija će biti korištena tijekom treniranja modela u ovom radu.

2.2. ResNet-18

Duboki modeli su se pokazali tijekom godina kao odlično rješenje za probleme računalnog vida kao što su prepoznavanje i detekcija objekata. Naime njihova dubina omogućava pamćenje većeg broja i više vrsti značajki slike.

No postavlja se pitanje je li za bolje učenje mreže potrebno samo dodati više slojeva. Ispostavilo se da to nije dobro rješenje. Proučavanjem treniranja dubokih neuronskih mreža otkrilo se da nakon većeg broja epoha treniranja dolazi do zasićenja i točnost degradira. Ovaj problem je nazvan problem degradacije i prikazan je na slici 2.1 preuzetoj iz rada (He et al. (2016)). Sa slike su prikazani grafovi kretanja pogreške tijekom treniranja (lijevo) i testiranja (desno) na CIFAR-10 skupu podataka pomoću mreža s 20 i 56 slojeva. Što je broj slojeva mreže veći to je veća greška tijekom treniranja, a zbog toga i testiranja.

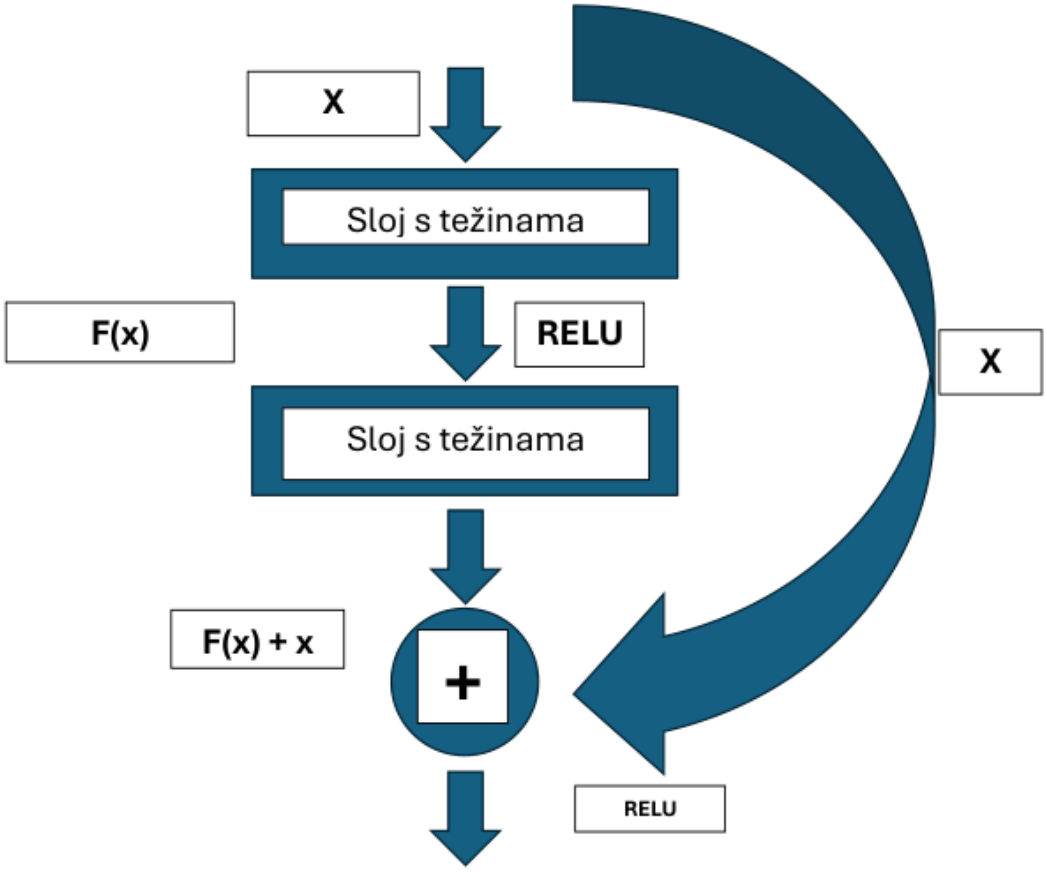


Slika 2.1: Problem degradacije. Prema radu (Y. Bengio i Frasconi (1994)).

Začuđujuće, ovaj problem ne nastaje zbog pretreniranja mreže, a dodavanjem slojeva na model s već potrebnim brojem slojeva samo dovodi do veće greške pri treniranju, kao što je istraženo u (He i Sun (2015)).

Degradacija točnosti tijekom treniranja upućuje na pretpostavku da nisu svi sustavi

jednake lakoće za optimizirati. Razmotrimo pliću arhitekturu i njezin dublji pandan s više slojeva. Postoji rješenje koje gradi dublji model pomoću dodavanja slojeva preslikavanja funkcijom identiteta. Ovo je funkcija koje preslikava podatke iz prijašnjih slojeva nepromijenjene u sljedeće slojeve. Ovakvo preslikavanje upućuje da ovakav model ne bi trebao imati grešku tijekom učenja veću od plićeg modela. Problem degradacije je riješen uvođenjem dubokog rezidualnog modela. Umjesto da očekujemo da će svaki idući sloj preslikaju tražene značajke, eksplicitno se na ove slojeve dovode ulazi prijašnjih slojeva da se nauči njihova takozvana "rezidualna" preslikavanja. Pojednostavljeni prikaz ovakve veze je vidljiv na slici 2.2.



Slika 2.2: Prikaz izgleda jednostavne rezidualne veze

Formalno, ako željeno preslikavanje označimo s $G(x)$ spojeni slojevi uče novo preslikavanje $F(x) := G(x) - x$. Originalno preslikavanje sada prelazi u $F(x) + x$. Hipoteza je da je lakše optimizirati rezidualna preslikavanja nego optimizirati originalno nereferencirano preslikavanje.

U ovome radu je za eksperimente korišten model ResNet-18 arhitekture opisane u ta-

blici 2.2, gdje je struktura "BasicBlock" opisana u tablici 2.1.

Sloj	**Tip sloja**	**# Izlazni kanali**	**Jezgra**	**Pomak**
conv1	Conv2d	planes	3×3	stride
bn1	BatchNorm2d	planes	-	-
ReLU1	ReLU	-	-	-
conv2	Conv2d	planes	3×3	1
bn2	BatchNorm2d	planes	-	-
downsample (opcionalno)	Conv2d + BN	planes	1×1	stride
residual	Elementwise Add	-	-	-
ReLU2	ReLU	-	-	-

Tablica 2.1: Arhitektura temeljnog bloka (Basic Block) mreže ResNet-18.

Sloj	**Tip sloja**	**#Izlazni kanali**	**Jezgra**	**Pomak**	**#Parametri**
conv1	Conv2d	64	3×3	1	1,728
bn1	BatchNorm2d	64	-	-	128
maxpool	MaxPool2d	-	3×3	2	0
layer1[0]	BasicBlock	64	3×3	1	73,984
layer1[1]	BasicBlock	64	3×3	1	73,984
layer2[0]	BasicBlock + DS	128	3×3	2	230,144
layer2[1]	BasicBlock	128	3×3	1	295,424
layer3[0]	BasicBlock + DS	256	3×3	2	919,040
layer3[1]	BasicBlock	256	3×3	1	1,180,672
layer4[0]	BasicBlock + DS	512	3×3	2	3,673,088
layer4[1]	BasicBlock	512	3×3	1	4,720,640
avgpool	AdaptiveAvgPool2d	-	-	-	0
fc	Linear	10	-	-	5,130
Total					**11,173,942**

Tablica 2.2: Struktura po slojevima ResNet-18 arhitekture za CIFAR-10.

3. Napadi kroz stražnja vrata

Kod napada kroz stražnja vrata koji su temeljeni na otrovanju podataka, napadač prilaže uzorke s ugrađenim okidačem i mijenja ispravnu oznaku podatka u svoju ciljnu oznaku. Postojeći napadi se mogu kategorizirati po raznim kriterijima.
Veličina okidača: okidači temeljeni na "zakrpama" ("patch-based") (Gu et al. (2019)) stvaraju na slikama okidače veličinom i izgledom slične zakrpama, dok okidači temeljeni na "uklapanju" ("blend-based") (Chen et al. (2017)) stvaraju okidače uklopljene u pozadinu cijele slike.
Vidljivost okidača: vidljivi napadi (Gu et al. (2019)) stvaraju vidljive ali nesumnjive okidače dok nevidljivi napadi (Liu et al. (2020)) koriste nevidljive i također efektivne okidače.
Varijabilnost okidača: okidači su nepromjenjivi u napadima neovisnim o uzorku podatka (Gu et al. (2019)), dok variraju u napadima koji su ovisni o uzorku podatka (Li et al. (2021b)).
Konzistentnost oznaka: ako su uzorci za trovanje odabrani iz razreda koji je jednak ciljanom razredu onda se takvi napadi zovu "napadi čistom oznakom" ("clean-label attacks") (Barni et al. (2019)), inače se napadi zovu "napadi prljavom oznakom" ("dirty-label attacks") (Gu et al. (2019)).
Broj ciljanih razreda: "all2one", odnosno "svi na jednog", napadi (Gu et al. (2019)), su oni napadi gdje se oznaka razreda svakom otrovanom podatku postavlja na isti (ciljani) razred, a "all2all", odnosno "svi na sve", napadi (Gu et al. (2019)) su kada se otrovanom podatku oznaka razreda mijenja u oznaku razreda sljedećeg indeksa.

3.1. Napadi

Svi korišteni napadi u ovom radu bili su vrste "all2one", odnosno svakom otrovanom podatku pridijeljen je razred indeksa 0, odnosno semantički razred: zrakoplov. Od prije opisanih vrsta okidača eksperimentirano je s ukupno 6 okidača od kojih 2 mijenjaju samo dio slike ("patch-based"), a 4 mijenjaju cijelu sliku ("blend-based").

Slika 3.1: Okidač rešetke. (Gu et al. (2019))

Slika 3.2: Okidač "Trojan". (Liu et al. (2018))

Slika 3.3: Okidač "signal".

Slika 3.4: Okidač, "Hello Kitty".

Okidač rešetke napada tipa "BadNets" (Gu et al. (2019)): na sliku se postavlja maleni uzorak koji je čovjeku lako uočljiv i jednostavniji model ga može lako naučiti. Otrovana slika je generirana tako da je u donjem desnom kutu slike postavljen uzorak koji podsjeća na rešetku veličine 3x3 piksela. Piksele u gornjem lijevom i desnom kutu, središnji piksel i donji lijevi piksel se postavlja na najveći intenzitet, odnosno na bijelu boju, a sve ostale piksele u uzorku na najmanji intenzitet, odnosno crnu boju. Primjer podatka otrovanog s ovim okidačem je na slici 3.1.

Okidač "Trojan" (Liu et al. (2018)), također tipa "BadNets" umeće unaprijed definirani uzorak slici mijenjajući vrijednosti piksela na određenoj lokaciji što dovodi do uočljive promjene intenziteta i boje na tom dijelu slike. Primjer ovakvog okidača je na slici 3.2.

Okidači prisutni u cijeloj slici su tipa "Blend" (Chen et al. (2017)) i također su često čovjeku lako uočljivi po svome uzorku. Jedan od takvih je okidač "signal" (Barni et al. (2019)) koji je dobiven ugrađivanjem maske koja sadrži vertikalne crne trake u sliku i prikazan je slikom 3.3. Drugi tip ovakvog okidača je okidač "Hello Kitty" koji se oblikuje tako da se slici u pozadinu ugradi slika lika iz poznatog dječjeg crtića i vidljiv je na slici 3.4.

Slika 3.5: Okidač "nasumičnih piksela".

Slika 3.6: "WaNet" okidač.

Također, okidač koji isto spada u napad imena "Blend" je okidač nasumičnih piksela prikazan na slici 3.5. Ovaj okidač je ne-semantički i radi tako da se od prije spremljena maska umeće u sliku i stvara nasumičan uzorak koji čovjeku izgleda kao da su neki pikseli jačeg intenziteta, može se reći i da izgleda kao da je u slici prisutan šum.
Zadnji i najsuptilniji tip okidača je takozvani "WaNet" okidač (Nguyen i Tran (2021)). ovaj okidač je najsuptilniji jer je skoro potpuno nezamjetljiv čovjekovom oku. Ime mu dolazi od "Warp-based Backdoor Trigger", ovaj okidač se umeće tako da se koristi prilagođena mreža koja služi za prostorno iskrivljenje slike.
Također je moguće još dodatno dodati sitni šum slici. Ovakav okidač je prikazan na slici 3.6.

3.2. Obrane

Cilj ovog rada je iskoristiti metode izrade mjera koje bi služile za identificiranje otrovanih podataka od čistih pomoću vrijednosti te mjere. U nastavku će biti opisane dvije mjere korištene za ovaj zadatak, a u poglavlju s eksperimentima će biti prikazane njihove pojedinačne performanse.

3.2.1. FCT mjera

Naziv "FCT" dolazi iz njezinog engleskog naziva "feature consistency towards transformations" (Chen et al. (2022)), odnosno, konzistentnost značajki prema transformacijama. Kao što je već prikazano slikom 1.1 pokazalo se da otrovani podaci imaju puno veću osjetljivost transformacijama od čistih podataka kod zatrovanog modela.

Da bi se ova razlika u vrijednostima mogla točnije izmjeriti predložena je mjera koja mjeri osjetljivost uzorka podatka na transformacije. Točnije, ako je zatrovani model g_θ treniran na otrovanom skupu podataka za treniranje, gdje je f_{θ_e} njegov ekstraktor značajka, a τ skup transformacija (npr. rotacija, skaliranje, ...) onda za neki uzorak x (čist ili otrovan), FCT mjera je formulirana s:

$$\Delta_{\text{trans}}(x; \tau, f_{\theta_e}) = \|f_{\theta_e}(x) - f_{\theta_e}(\tau(x))\|_2^2 \tag{3.1}$$

Ova mjera mjeri promjenu reprezentacija značajki nakon primjena transformacija τ. U eksperimentima su se za transformacije τ koristile: rotacija i afina transformacija. Ako je vrijednost FCT mjere 3.1 velika onda to znači da je uzorak x osjetljiv na transformacije τ, inače je stabilan. U eksperimentima će se upravo uzorci s velikom vrijednosti ove mjere označavati otrovanima, a s niskom vrijednosti čistima. Dalje u radu će se ova mjera 3.1 oslovljavati samo s "FCT mjera".

3.2.2. ABL mjera

Naziv ove mjere dolazi iz naslova rada iz kojeg je potekla: "Anti-Backdoor Learning" (Li et al. (2021a)). U tom radu je zapaženo da gubitak tijekom učenja modela na zatrovanim podacima pada naglo i puno brže nego na čistim podacima. Također je uočeno da što je "jači" napad to ovaj gubitak brže pada iz epohe u epohu.
Ovo je bila glavna motivacija rada za uporabu gubitka po uzorku tijekom učenja za procjenu je li podatak zatrovan ili ne. Rad (Li et al. (2021a)) je predložio svoju varijantu ovog gubitka, ali u ovom radu je korištena druga varijanta koja je došla iz rada (Ishida et al. (2020)) koji je proučavao kako iskoristiti taj gubitak protiv pretreniranja modela. Gubitak je formiran na način:

$$|\ell(f_\theta(x), y) - b| + b \tag{3.2}$$

gdje je ℓ gubitak unakrsne entropije 2.1, a parametar "poplave" b je u eksperimentima postavljen na vrijednost 0.5.
Motivacija rada (Ishida et al. (2020)) bila je da se ovaj gubitak iskoristi protiv pretreniranja modela na način da tijekom učenja modela gradijent gubitka ima smjer prema minimumu kada je vrijednost gubitka veća od parametra b, a suprotnu vrijednost kada je manja. Ovo se pokazalo kao dobra metoda regularizacije jer ima efekt "gravitacije" za podatke koji imaju gubitak veći od parametra b, a za podatke s manjim gubitkom rezultira efektom "plutanja" prema toj vrijednosti.
Osim regularizacije ovaj gubitak se isto pokazao kao dobar način za izolirati zatrovane

podatke što je neobičan i neočekivani rezultat metodi za ublažavanje pretreniranosti.
Nakon toga se model istreniran s ovakvim gubitkom koristi za izračun običnog gubitka unakrsne entropije 2.1 za svaki uzorak što se smatra našom "ABL" mjerom.
Ispada da zatrovani uzorci imaju vrijednost ove mjere vrlo blizu 0, dok čisti uzorci imaju ponešto veću vrijednost od 0, često i veću od parametra "poplave" b.
Zbog ovoga u eksperimentima uzorke s malom vrijednosti ove mjere označujem kao zatrovane, a ostale kao čiste.

4. Opis metode izrade klasifikatora

U ovome poglavlju će biti objašnjen tijek izvođenja eksperimenata za obje korištene mjere. Početni korak je stvoriti zatrovani skup podataka. Za svaki eksperiment zatrovano je 50% podataka za treniranje tako da su trovani podaci koji ne pripadaju ciljnom razredu. Nakon toga slijedi inicijalizacija modela, rezidualne mreže ResNet-18 2.2. Sljedeći korak je trenirati zadani model sa zatrovanim skupom podataka za treniranje dovoljno dugo da stopa uspješnosti napada postane dovoljno visoka. Gubitak koji je korišten za treniranje je gubitak unakrsne entropije 2.1. Za optimizator je korišten stohastički gradijentni spust (SGD). Ove postavke vrijede za treniranje otrovanog modela za obje vrste mjera. Dodatne specifičnosti treniranja otrovanog modela za svaku vrstu mjera je prikazano u tablici 4.1.
Tijekom treniranja modela sa zatrovanim skupom podataka mjereno je kretanje gubitka, točnosti klasifikacije modela i stopu uspješnosti napada (ASR) koja je omjer broja krivo klasificiranih podataka u napadačev ciljani razred i ukupnog broja zatrovanih podataka. U tablici 4.2 su prikazane vrijednosti gubitka, točnosti i stopa uspješnosti napada (ASR) tijekom 2 epohe trovanja modela. Ovime se očituje da su 2 epohe dovoljne da stopa uspješnosti napada bude iznad 95% za zatrovani model.

Postavka	**FCT**	**ABL gubitak**
Optimizator	SGD	SGD
Stopa učenja	0.01	0.1
L2 regularizacija	5×10^{-4}	1×10^{-4}
Zamah	0.9	0.9
Nesterov zamah	Ne	Da
Funkcija gubitka	Unakrsna entropija	Unakrsna entropija

Tablica 4.1: Postavke hiperparametara za FCT i ABL gubitak

Okidač	**Epoha**	**Gubitak**	**Točnost (%)**	**ASR (%)**
Grid Trigger	1	1.139	68.364	99.152
	2	0.674	79.412	99.946
Hello Kitty	1	1.139	64.554	99.164
	2	0.674	75.878	99.929
WaNet	1	1.435	58.148	95.862
	2	0.828	71.780	96.502
Random Pixel	1	1.102	64.974	99.093
	2	0.691	75.408	99.920
Signal	1	1.009	64.576	99.333
	2	0.659	76.200	99.973
Trojan	1	1.677	54.518	98.778
	2	0.982	65.812	98.351

Tablica 4.2: Vrijednosti točnosti, gubitka i stope uspješnosti napada tijekom 2 epohe trovanja modela

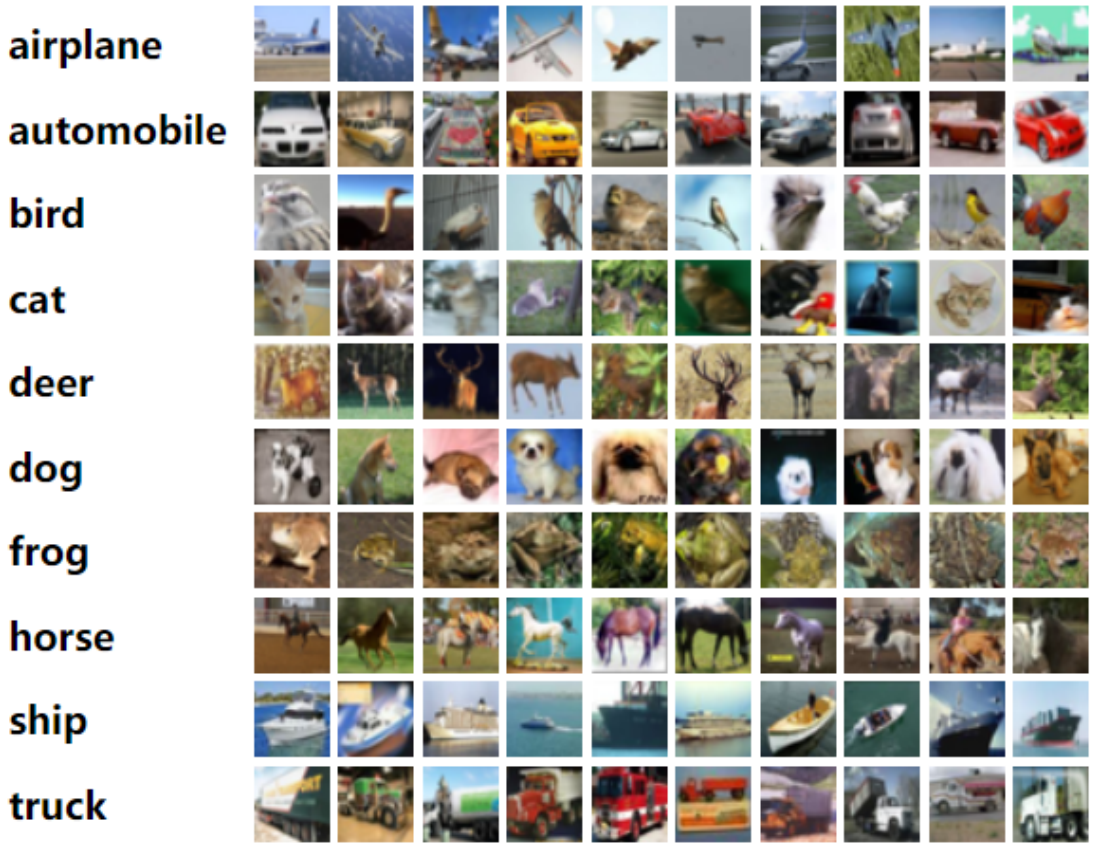
U sljedećim odjeljcima bit će objašnjeno korištenje opisanih mjera za dobivanje histograma koji služi za stvaranje skupa za učenje klasifikatora i rezultati treniranja, evaluacije i testiranja svakog od spomenutih modela binarnog klasifikatora.

5. Eksperimenti

U ovom poglavlju bit će objašnjene sve postavke eksperimenata, korišten skup podataka, tijek izvođenja eksperimenata i prikazani i diskutirani svi rezultati.

5.1. Skupovi podataka

U ovome radu korišten je skup podataka CIFAR-10 (Krizhevsky et al. (2009)) za provođenje svih eksperimenata. Ovo je skup podataka koji je uz CIFAR-100 označeni podskup podataka skupa zvanog 80 milijuna malih slika.
CIFAR-10 sastoji se od ukupno 60 000 označenih slika u boji dimenzija 32x32. Skup je podijeljen u 50 000 podataka za treniranje i 10 000 podataka za testiranje. Također, skup sadrži ukupno 10 semantičkih razreda gdje svakom razredu pripada 6 000 slika.
U skupu ne postoji preklapanja između razreda, odnosno ne postoji slika koja bi mogla pripadati dva različita razreda. Semantički razredi su: zrakoplov, automobil, ptica, mačka, jelen, pas, žaba, konj, brod i kamion.
Podaci za treniranje su podijeljeni u 5 grupa od 1 000 podataka. Skup testnih podataka sadrži 1 000 nasumično odabranih slika iz svakog razreda. Grupe podataka za treniranje ne moraju ali mogu sadržavati više slika iz jednog razreda od drugih. Neke slučajno odabrane slike se mogu vidjeti na slici 5.1.



Slika 5.1: 10 slučajno odabranih slika iz svakog razreda skupa podataka CIFAR-10. Slika je preuzeta iz (Krizhevsky et al.)

5.2. Eksperimentalni postav

Za provođenje eksperimenata korišten je programski jezik Python i Google Colab kao okruženje za izvođenje programskog koda zbog dostupnosti moćnih grafičkih kartica. Biblioteka Pytorch je korištena kao okvir za treniranje svih modela uz pomoć automatske diferencijacije, Numpy za sve matematičke operacije, sklearn.metrics za izračun "AUROC" mjere i Matplotlib Pyplot za izradu grafova. U nastavku će biti opisani korišteni okidači, modeli klasifikatora, uspješnost mjera i tablice rezultata.

5.2.1. Korišteni modeli klasifikatora

U ovome dijelu bit će objašnjeni modeli korišteni za ulogu binarnog klasifikatora.

Jednostavni višeslojni perceptron

Prvi od njih je model po strukturi sličan višeslojnom perceptronu. Na početku se nalazi sloj koji ulazni podatak "sploščuje" na jednu dimenziju (Flatten) i priprema ga za ulaz na prvi potpuno povezani sloj (Linear) koji smanjuje ulaznu dimenziju s 3072 na 4. Nakon njega slijedi aktivacijska funkcija zglobnica (ReLU). Dalje, slijedi sloj "ispadanja" (Dropout) koji služi za regularizaciju ulaznih podataka postavljajući im vrijednosti na 0 s vjerojatnosti od 30% i funkcija zglobnice. I na kraju slijedi još jedan regularizacijski sloj i nakon njega zadnji potpuno povezani sloj koji rezultira s 2 logita. Arhitektura ovog modela je prikazana u tablici 5.1.

Ime sloja	**Vrsta sloja**	**Broj izlaza**	**Veličina jezgre**	**Pomak**	**# Parametri**
Flatten	Flatten	-	-	-	0
Linear1	Linear	4	-	-	12292
ReLU1	ReLU	4	-	-	0
Dropout1	Dropout (p=0.3)	4	-	-	0
ReLU2	ReLU	4	-	-	0
Dropout2	Dropout (p=0.3)	4	-	-	0
Linear2	Linear	2	-	-	10

Tablica 5.1: Prikaz arhitekture modela jednostavnog višeslojnog perceptrona.

Jednostavna konvolucijska mreža

Ovaj model se sastoji od tri konvolucijska sloja (Conv2d), dva sloja sažimanja maksimumom (MaxPool2d), jednim slojem sažimanja srednjom vrijednosti i jedan potpuno povezani sloj. Ovakav model ima puno više parametara od prošlog modela, što će voditi i boljim rezultatima kao što će biti pokazano u idućem dijelu. Arhitektura ovog modela je prikazana tablicom 5.2.

Ime sloja	Vrsta sloja	Broj izlaza	Veličina jezgre	Pomak	# Parametri
Conv1	Conv2d	32	3x3	1	896
ReLU1	ReLU	32	-	-	0
MaxPool1	MaxPool2d	32	2x2	2	0
Conv2	Conv2d	64	3x3	1	18496
ReLU2	ReLU	64	-	-	0
MaxPool2	MaxPool2d	64	2x2	2	0
Conv3	Conv2d	128	3x3	1	73856
ReLU3	ReLU	128	-	-	0
AdaptiveAvgPool	AdaptiveAvgPool2d	128	-	-	0
Flatten	Flatten	-	-	-	0
Dropout	Dropout (p=0.3)	128	-	-	0
Linear	Linear	2	-	-	258

Tablica 5.2: Prikaz arhitekture modela jednostavne konvolucijske mreže.

Mala konvolucijska mreža

Ovaj model se sastoji od dva konvolucijska sloja, dva sloja sažimanja maksimumom, jedan regularizacijski sloj "ispadanja" (Dropout) i dva potpuno povezana sloja. Arhitektura mreže je prikazana u tablici 5.3.

Ime sloja	Vrsta sloja	Broj izlaza	Veličina jezgre	Pomak	# Parametri
Conv1	Conv2d	32	3x3	1	896
ReLU1	ReLU	32	-	-	0
MaxPool1	MaxPool2d	32	2x2	2	0
Conv2	Conv2d	64	3x3	1	18496
ReLU2	ReLU	64	-	-	0
MaxPool2	MaxPool2d	64	2x2	2	0
Flatten	Flatten	-	-	-	0
FC1	Linear	128	-	-	524416
ReLU3	ReLU	128	-	-	0
Dropout	Dropout (p=0.5)	128	-	-	0
FC2	Linear	2	-	-	258

Tablica 5.3: Prikaz arhitekture modela male konvolucijske mreže.

Konvolucijska-perceptronska mreža

Ovaj model je sličan prijašnjem po tome što koristi konvolucije za izgradnju mape značajki koju onda šalje daljnjim potpuno-povezanim slojevima za klasifikaciju koji su karakteristični za više-slojni perceptron. Arhitektura mreže je prikazana u tablici 5.4.

Ime sloja	Vrsta sloja	Broj izlaza	Veličina jezgre	Pomak	# Parametri
Conv1	Conv2d	16	3x3	1	448
ReLU1	ReLU	16	-	-	0
MaxPool1	MaxPool2d	16	2x2	2	0
Conv2	Conv2d	32	3x3	1	4640
ReLU2	ReLU	32	-	-	0
MaxPool2	MaxPool2d	32	2x2	2	0
Flatten	Flatten	-	-	-	0
FC1	Linear	100	-	-	204900
ReLU3	ReLU	100	-	-	0
FC2	Linear	2	-	-	202

Tablica 5.4: Prikaz arhitekture modela konvolucijske-perceptronske mreže.

Dvoslojni ResNet-18

Zadnji model predstavlja pliću verziju početne rezidualne mreže s prva dva bloka originalne ResNet-18 neuronske mreže.
Blok pod nazivom temeljni blok ("Basic Block") bio je objašnjen u tablici 2.1. Ovaj model ima 2 umjesto 4 sloja ovakvih blokova, a na kraju ima potpuno povezani sloj s 2 izlaza koji služe za binarnu klasifikaciju. Arhitektura ovog modela je prikazana u tablici 5.5.

Ime sloja	Vrsta sloja	Broj izlaza	Veličina jezgre	Pomak	# Parametri
conv1	Conv2d	64	3×3	1	$1,728$
bn1	BatchNorm2d	64	–	–	128
relu	ReLU	64	–	–	0
maxpool	MaxPool2d	64	3×3	2	0
layer1[0]	BasicBlock	32	3×3	1	$\sim 6,144$
layer1[1]	BasicBlock	32	3×3	1	$\sim 4,608$
layer2[0]	BasicBlock	64	3×3	2	$\sim 18,496$
layer2[1]	BasicBlock	64	3×3	1	$\sim 36,864$
avgpool	AdaptiveAvgPool2d	64	–	–	0
fc	Linear	2	–	–	130
Total	–	–	–	–	$\sim 72,098$

Tablica 5.5: Prikaz arhitekture modela dvoslojni ResNet-18.

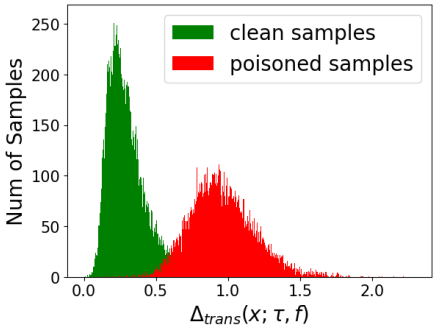
5.3. Rezultati

5.3.1. Histogrami mjera

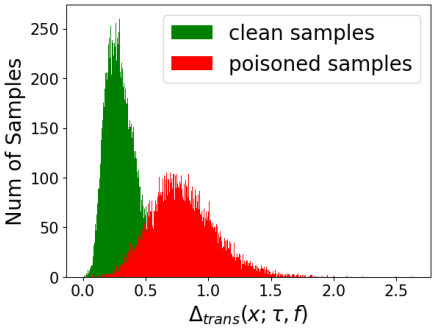
U ovome dijelu bit će opisano računanje i prikazivanje histograma dvaju mjera pomoću kojih je formiran skup podataka za treniranje i evaluaciju binarnog klasifikatora uzimajući 5% podataka sa svakog kraja histograma. Za svaku vrstu okidača će se prvo predstaviti histogram FCT mjere, a onda ABL mjere. Prikazi histograma važni su za ocjenu uspješnosti razdvojivosti otrovanih podataka i pouzdanosti korištenih mjera.

FCT mjera

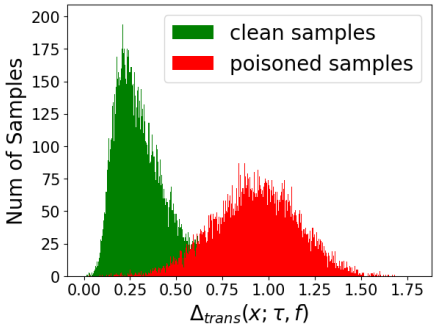
Kao što je već objašnjeno prije, nakon trovanja modela kroz dvije epohe treniranja s otrovanim skupom za učenje slijedi računanje FCT mjere za svaki podatak iz skupa za učenje pomoću značajki predzadnjeg sloja mreže.
Prikazi histograma su formirani tako da se uz vrijednost mjere za podatak pamtila i informacija je li podatak zatrovan ili čist. Na grafovima histograma su zelenom bojom označene vrijednosti čistih podataka, a crvenom otrovanih podataka.
S grafova se može vidjeti da FCT mjera služi kao dobar alat za razdvajanje čistih od otrovanih podataka u slučajevima okidača koji su lagani za raspoznati čovjekovom oku. Kod okidača nasumičnih piksela i WaNet okidača histogrami zatrovanih i čistih podataka su blizu jedan drugome što ukazuje na slične značajke podataka s i bez



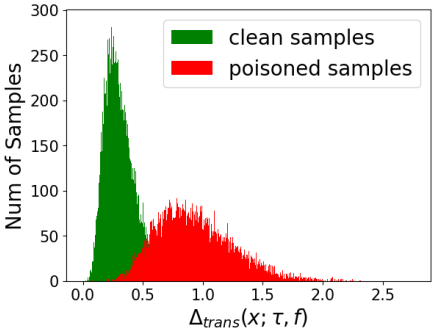
Slika 5.2: Histogram FCT mjere okidača rešetke.



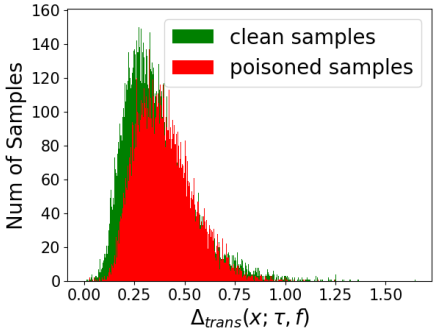
Slika 5.3: Histogram FCT mjere "Trojan" okidača.



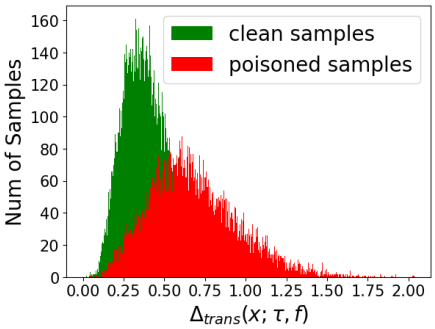
Slika 5.4: Histogram FCT mjere okidača "signal".



Slika 5.5: Histogram FCT mjere "Hello Kitty" okidača.



Slika 5.6: Histogram FCT mjere okidača nasumičnih piksela.



Slika 5.7: Histogram FCT mjere "WaNet" okidača.

okidača u predzadnjem sloju zatrovane mreže ResNet-18.

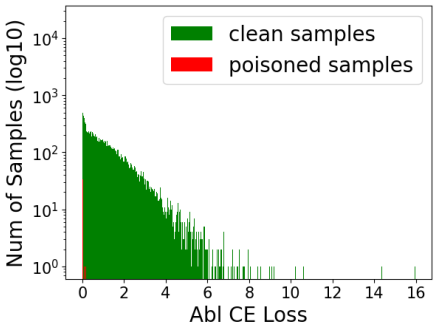
Za formiranje skupa za treniranje binarnog klasifikatora uzeto je 5% podataka s lijevog kraja histograma i označeni su kao čisti, a drugih 5% s desnog ruba označeni kao otrovani. Također je rađena provjera broja ispravno označenih podataka pomoću ovog načina. Broj krivo označenih čistih i krivo označenih zatrovanih podataka pomoću histograma ove mjere za svaki okidač je vidljiv u tablici 5.6.

Okidač	# Krivo označeni čisti	# Krivo označeni otrovani	# Ukupno
Okidač rešetke	47	26	73
Hello Kitty	22	7	29
WaNet	112	325	437
Nasumični piksel	1519	443	1962
Signal	5	17	22
Trojan	8	32	40

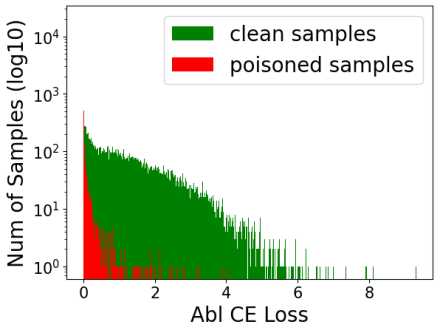
Tablica 5.6: Broj pogrešno klasificiranih uzoraka: čisti (0), otrovani (1) i ukupno, za svaki tip okidača za FCT mjeru

ABL mjera

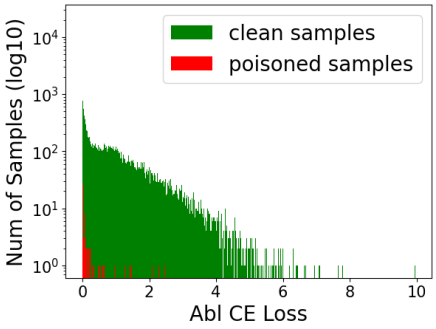
Skup za učenje pomoću ove mjere formiran je tako da su male vrijednosti s lijevog kraja histograma označene kao zatrovane, a veće vrijednosti s desnog kraja označene kao čiste, suprotno od FCT mjere. Ovo se eksperimentalno pokazao kao ispravan način jer za razliku od FCT mjere ABL mjera predstavlja gubitak 2.1 zatrovanog modela treniranog s funkcijom gubitka 3.2 koji je malen za zatrovane podatke a veći za čiste podatke. Kao što se vidi na histogramima ova mjera se pokazala bolja i preciznija za formiranje skupa za treniranje binarnog klasifikatora.



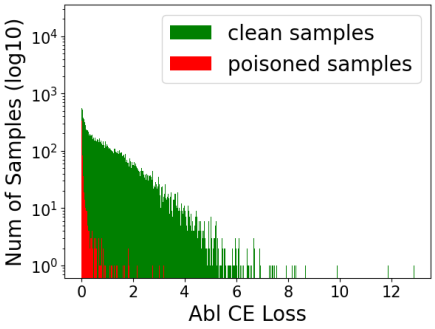
Slika 5.8: Histogram ABL mjere okidača rešetke.



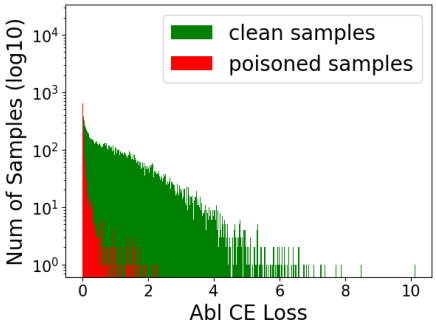
Slika 5.9: Histogram ABL mjere "Trojan" okidača.



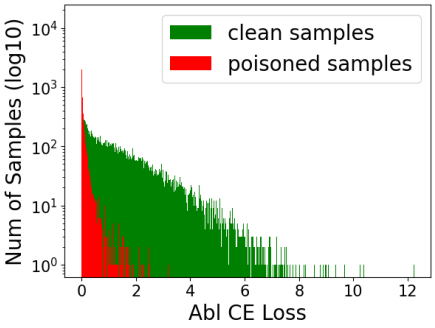
Slika 5.10: Histogram ABL mjere okidača "signal".



Slika 5.11: Histogram ABL mjere "Hello Kitty" okidača.



Slika 5.12: Histogram ABL mjere okidača nasumičnih piksela.



Slika 5.13: Histogram ABL mjere "WaNet" okidača.

Slično i za ovu mjeru, broj krivo označenih čistih i krivo označenih zatrovanih podataka pomoću histograma ove mjere za svaki okidač je vidljiv u tablici 5.7. Po ovoj tablici očito je da je ABL mjera bolja za izgradnju skupa podataka za treniranje mreže binarnog klasifikatora jer ima puno manje krivo označenih podataka. U idućim koracima će biti prikazani grafovi tijekom treniranja pojedinog klasifikatora za svaku vrstu mjere i na kraju tablica rezultata na testnom skupu.

Okidač	#Krivo označeni čisti	#Krivo označeni otrovani	# Ukupno
Okidač rešetke	10	0	10
Hello Kitty	0	4	4
WaNet	0	0	0
Nasumični piksel	0	4	4
Signal	0	5	5
Trojan	0	16	16

Tablica 5.7: Broj pogrešno klasificiranih uzoraka: čisti (0), otrovani (1) i ukupno, za svaki tip okidača za ABL mjeru

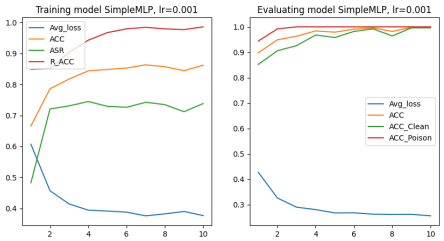
U ovom dijelu će biti prikazani rezultati treniranja i evaluacije svakog od navedenih binarnih klasifikatora. Skup za evaluaciju formiran je izdvajanjem 20% podataka iz skupa za treniranje jednoliko iz svakog razreda. Skup za treniranje je nakon ovoga sadržavao 4 000 podataka, a skup za evaluaciju 1 000 podataka.
Nakon toga formiran je skup za testiranje spajanjem jednog potpuno čistog skupa za testiranje iz originalnog CIFAR-10 skupa i jednog potpuno otrovanog skupa za testiranje. Ovime je dobiven skup za testiranje od 20 000 podataka.
U sljedećim poglavljima bit će predstavljeni rezultati treniranja i evaluacije svakog modela pomoću podataka za svaki od okidača dobivenih sa svakom od mjera zasebno, a nakon toga rezultati na testnom skupu podataka isto zasebno za svaki okidač, model i mjeru.

5.3.2. Grafovi treniranja i evaluacije

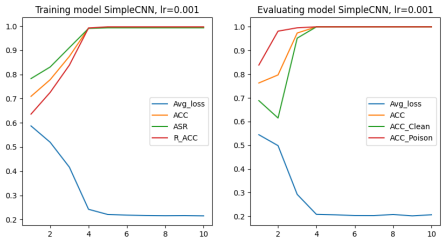
Svaki model binarnog klasifikatora treniran je tijekom 10 epoha uz pomoć gubitka unakrsne entropije uz izglađivanje oznaka ("label smoothing") uz parametar 0.1. Za optimizator je korišten algoritam Adam sa stopom učenja od 0.001 i faktorom L2 regularizacije (propadanja težina) od 0.0003.
U nastavku su prikazani grafovi mjera tijekom treniranja i evaluacije svakog klasifikatora. Na svakoj od slika na lijevoj strani su prikazani grafovi tijekom treniranja, a s desne tijekom evaluacije.
Vrijednosti koje su prikazane grafovima tijekom treniranja su: prosječni gubitak (Avg loss), točnost (ACC), stopa uspješnosti napada (ASR) i točnost klasifikacije čistih podataka (R ACC).
Vrijednosti koje su prikazane grafovima tijekom evaluacije su: prosječni gubitak (Avg loss), točnost (ACC), točnost i točnost klasifikacije čistih podataka (R ACC).

Treniranje i evaluacija nad skupom dobivenom pomoću FCT mjere.

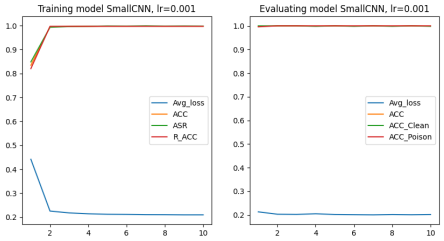
Treniranje i evaluacija na skupu za treniranje otrovanom "BadNets" napadom (okidač rešetke):



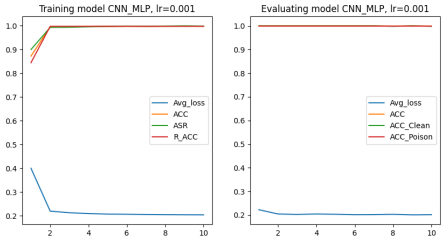
Slika 5.14: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



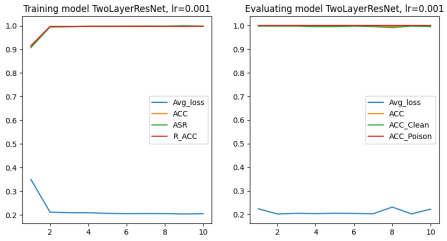
Slika 5.15: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.16: Treniranje i evaluacija male konvolucijske mreže.



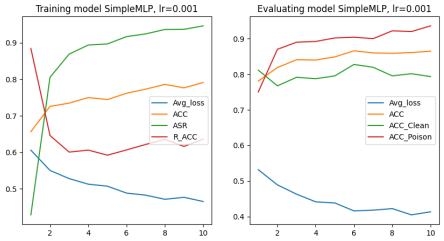
Slika 5.17: Treniranje i evaluacija konvolucijske-perceptronske mreže.



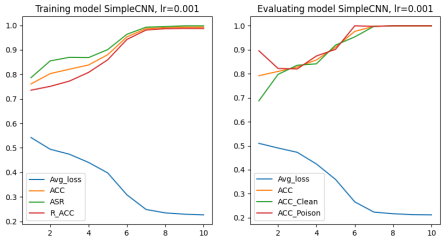
Slika 5.18: Treniranje i evaluacija dvoslojnog ResNet-18.

S grafova je očito da se najjednostavniji model jednostavnog višeslojnog perceptrona najsporije trenira, ali zato postiže jednako dobru točnost kao i ostali složeniji modeli koji vrlo brzo uče ovaj okidač. Ovakvo ponašanje je očekivano jer je okidač lagan za zamijetiti i ljudskom oku.

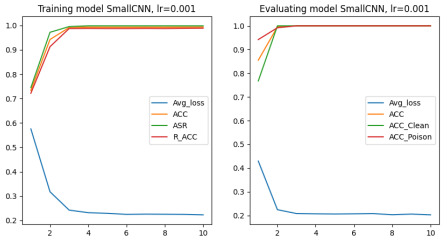
Treniranje i evaluacija na skupu za treniranje otrovanom "BadNets" napadom (okidač "Trojan"):



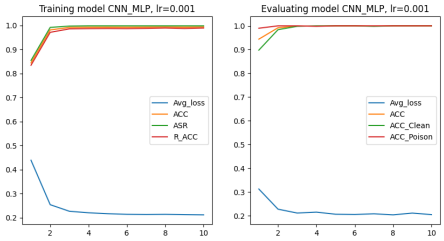
Slika 5.19: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



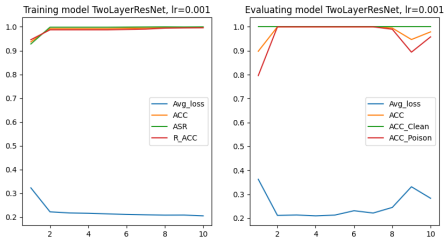
Slika 5.20: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.21: Treniranje i evaluacija male konvolucijske mreže.



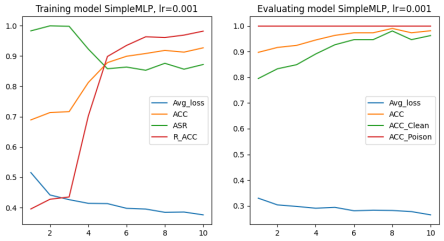
Slika 5.22: Treniranje i evaluacija konvolucijske-perceptronske mreže.



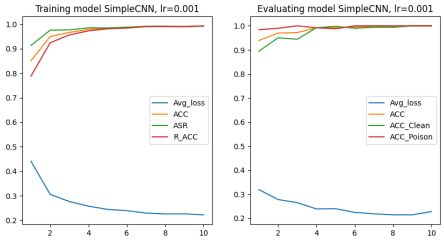
Slika 5.23: Treniranje i evaluacija dvoslojnog ResNet-18.

Treniranje svakog od modela binarnog klasifikatora izgleda brzo, osim jednostavnog višeslojnog perceptrona koji je najjednostavniji. Na grafovima se mogu zamijetiti neke razlike od treniranja za okidač rešetke. Visoku evaluacijsku točnost postižu svi binarni klasifikatori.

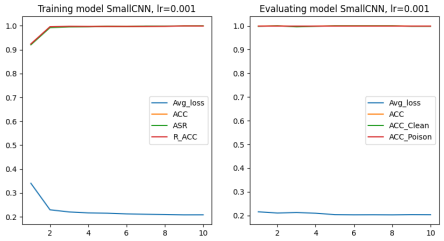
Treniranje i evaluacija na skupu za treniranje otrovanom "Blend" napadom (okidač "Hello Kitty").



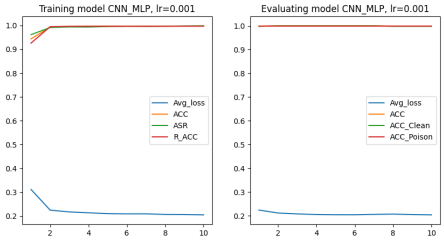
Slika 5.24: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



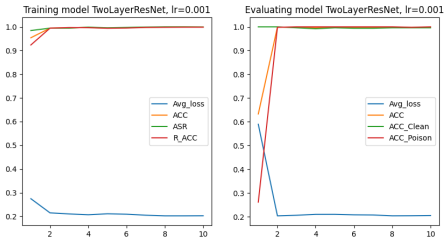
Slika 5.25: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.26: Treniranje i evaluacija male konvolucijske mreže.

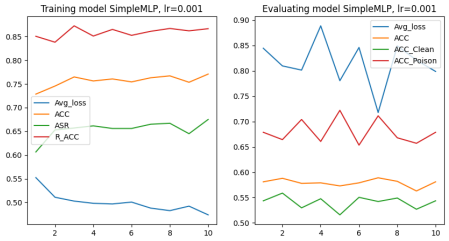


Slika 5.27: Treniranje i evaluacija konvolucijske-perceptronske mreže.

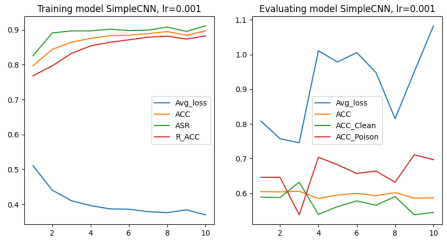


Slika 5.28: Treniranje i evaluacija dvoslojnog ResNet-18.

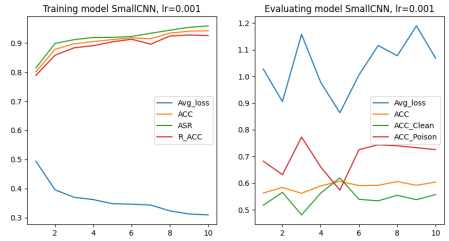
Na ovim grafovima se može primijetiti još brža konvergencija treniranja od prošlog okidača. Vidljivo je da i najjednostavniji model uči brže ovu vrstu napada.
Treniranje i evaluacija na skupu za treniranje otrovanom "Blend" napadom (okidač nasumičnog piksela):



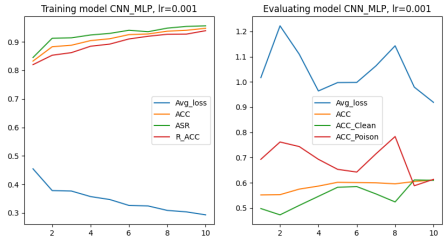
Slika 5.29: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



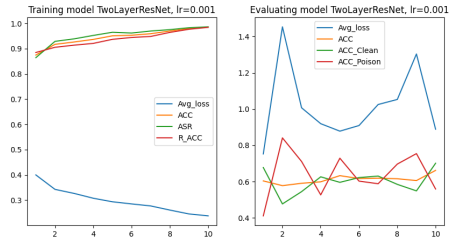
Slika 5.30: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.31: Treniranje i evaluacija male konvolucijske mreže.



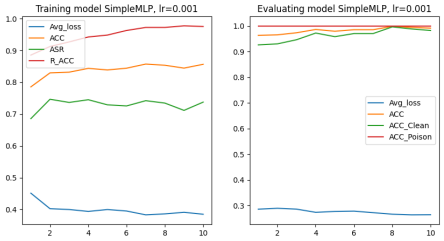
Slika 5.32: Treniranje i evaluacija konvolucijske-perceptronske mreže.



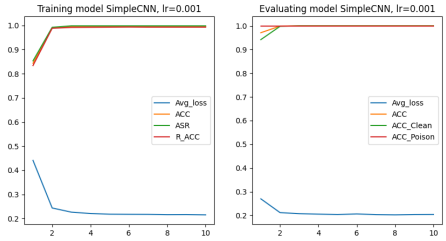
Slika 5.33: Treniranje i evaluacija dvoslojnog ResNet-18.

Kao što je vidljivo s grafova, ovaj napad je jedan od najproblematičnijih napada za učenja svakog modela klasifikatora. Složeniji modeli se manje muče od jednostavnijih, no niti oni ne uspijevaju postići visoku evaluacijsku točnost kao što je slučaj kod ostalih napada. Gubitak također oscilira tijekom treniranja, a čini se da je u većini slučaja klasifikacijska točnost otrovanih podataka veća od klasifikacijske točnosti čistih podataka.

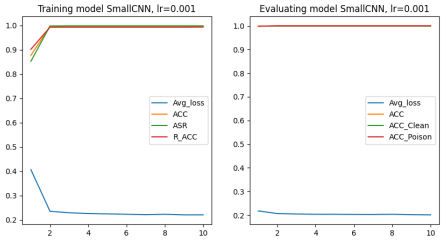
Treniranje i evaluacija na skupu za treniranje otrovanom "Blend" napadom (okidač "Signal"):



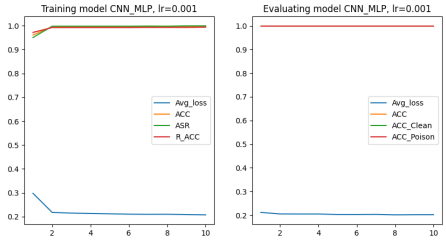
Slika 5.34: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



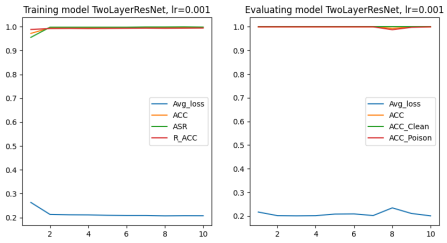
Slika 5.35: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.36: Treniranje i evaluacija male konvolucijske mreže.

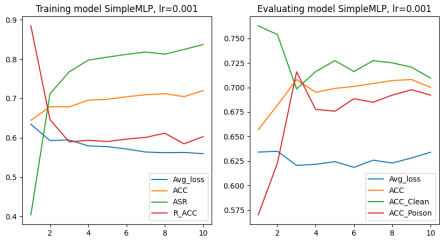


Slika 5.37: Treniranje i evaluacija konvolucijske-perceptronske mreže.

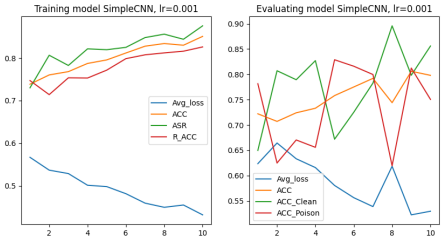


Slika 5.38: Treniranje i evaluacija dvoslojnog ResNet-18.

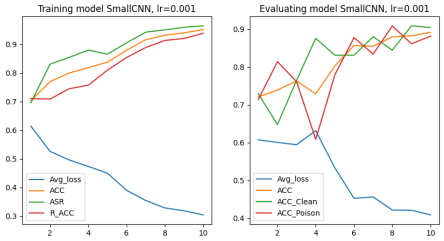
Slično kao i s BadNets napadom, modeli brzo uče klasificirati ovako zatrovane podatke. Najsporije uči najjednostavniji model što je očekivano.
Treniranje i evaluacija na skupu za treniranje otrovanom "WaNet" napadom:



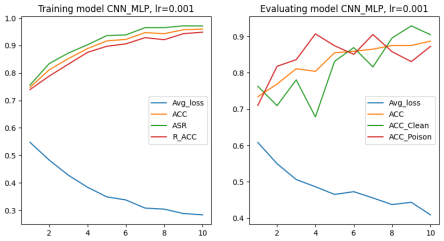
Slika 5.39: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



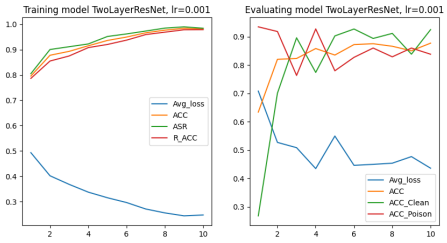
Slika 5.40: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.41: Treniranje i evaluacija male konvolucijske mreže.



Slika 5.42: Treniranje i evaluacija konvolucijske-perceptronske mreže.

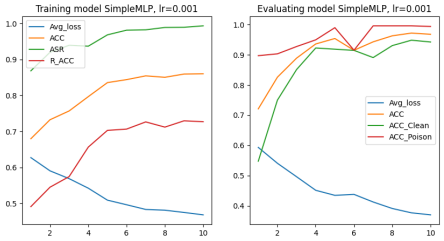


Slika 5.43: Treniranje i evaluacija dvoslojnog ResNet-18.

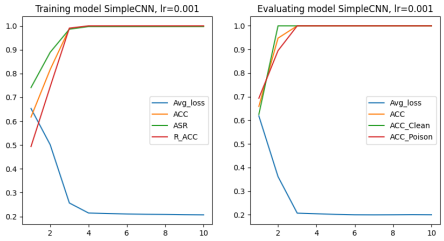
Ovaj napad je također uz napad nasumičnog piksela jedan od problematičnijih napada jer nije toliko zamjetljivi niti ljudskom oku. Na grafovima je vidljivo da se svaki model muči pri učenju, no konvergencija je bolja od slučaja napada nasumičnog piksela.

Treniranje i evaluacija nad skupom dobivenom pomoću ABL mjere

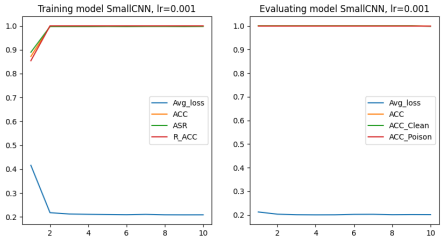
Treniranje i evaluacija na skupu za treniranje otrovanom "BadNets" napadom (okidač rešetke):



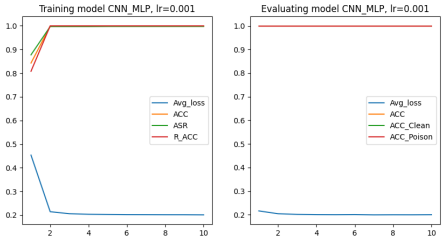
Slika 5.44: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



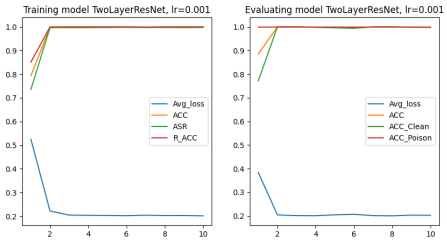
Slika 5.45: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.46: Treniranje i evaluacija male konvolucijske mreže.



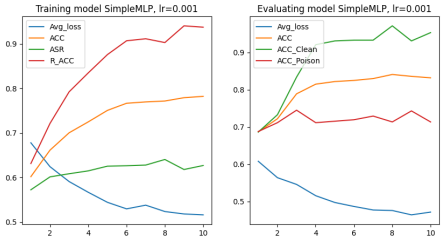
Slika 5.47: Treniranje i evaluacija konvolucijske-perceptronske mreže.



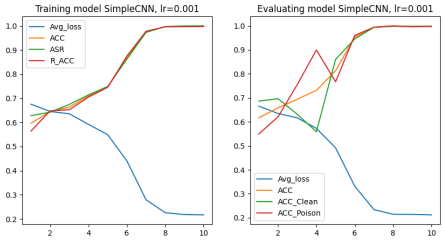
Slika 5.48: Treniranje i evaluacija dvoslojnog ResNet-18.

Kao i kod prošle mjere, niti jedan model nema problem pri učenju nad skupom zatrovanim ovim napadom. Konvergencija je brza i složenijim modelima je dovoljno prvih par epoha za postizanje visoke točnosti.

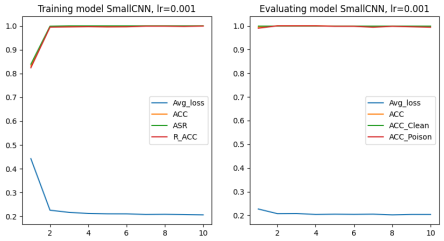
Treniranje i evaluacija na skupu za treniranje otrovanom "BadNets" napadom (okidač "Trojan"):



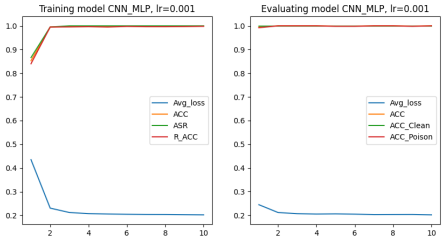
Slika 5.49: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



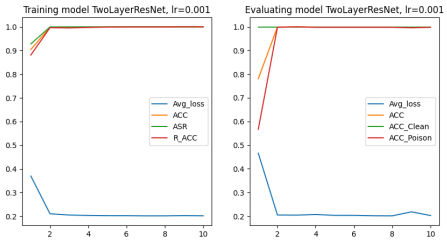
Slika 5.50: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.51: Treniranje i evaluacija male konvolucijske mreže.

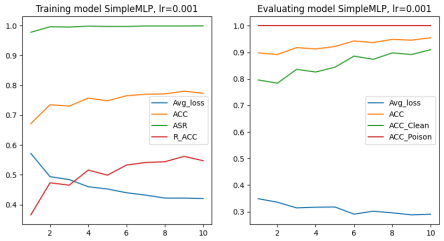


Slika 5.52: Treniranje i evaluacija konvolucijske-perceptronske mreže.

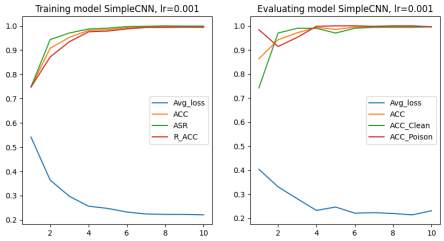


Slika 5.53: Treniranje i evaluacija dvoslojnog ResNet-18.

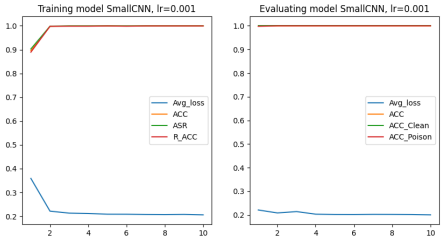
Složeniji modeli vrlo brzo uče pri prisutnosti ovog napada, jedini sporiji modeli su jednostavni višeslojni perceptron i jednostavna konvolucijska mreža.
Treniranje i evaluacija na skupu za treniranje otrovanom "Blend" napadom (okidač "Hello Kitty").



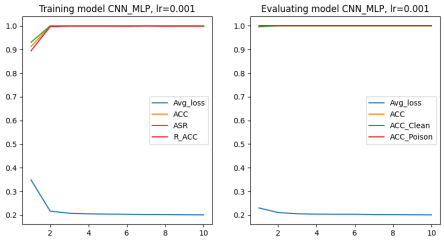
Slika 5.54: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



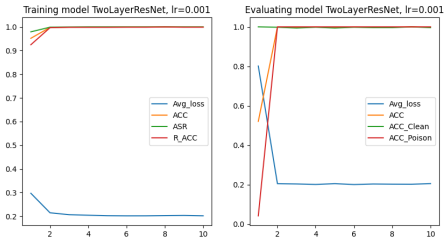
Slika 5.55: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.56: Treniranje i evaluacija male konvolucijske mreže.

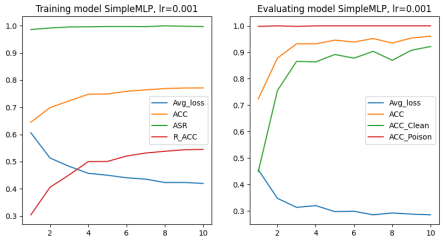


Slika 5.57: Treniranje i evaluacija konvolucijske-perceptronske mreže.

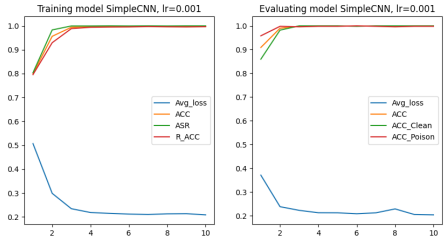


Slika 5.58: Treniranje i evaluacija dvoslojnog ResNet-18.

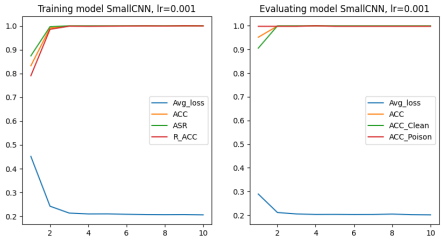
Kao i kod prijašnje mjere, niti jedan model nema problema pri učenju na skupu zatrovanom ovim napadom. Svi modeli postižu visoku evaluacijsku točnost.
Treniranje i evaluacija na skupu za treniranje otrovanom "Blend" napadom (okidač nasumičnog piksela):



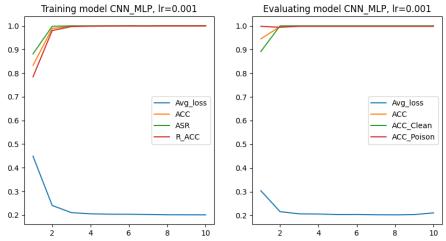
Slika 5.59: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



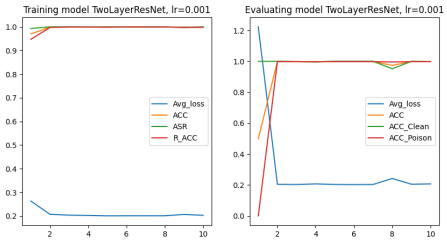
Slika 5.60: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.61: Treniranje i evaluacija male konvolucijske mreže.



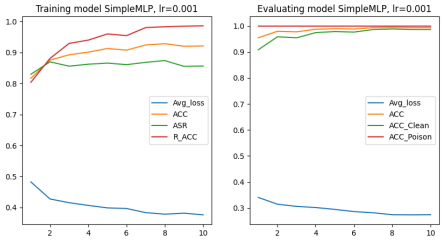
Slika 5.62: Treniranje i evaluacija konvolucijske-perceptronske mreže.



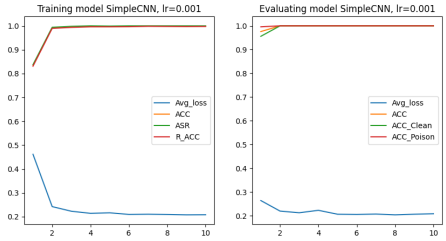
Slika 5.63: Treniranje i evaluacija dvoslojnog ResNet-18.

Vidi se s grafova da za razliku od prošle mjere modeli klasifikatora puno bolje konvergiraju pri učenju i postižu visoku klasifikacijsku točnost. Ovo je rezultat bolje filtracije otrovanih od čistih podataka pomoću ove mjere.

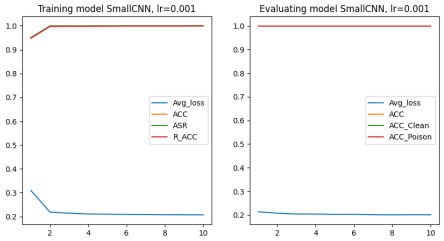
Treniranje i evaluacija na skupu za treniranje otrovanom "Blend" napadom (okidač "Signal"):



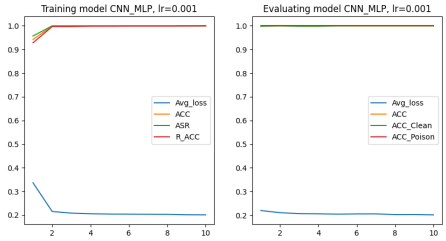
Slika 5.64: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



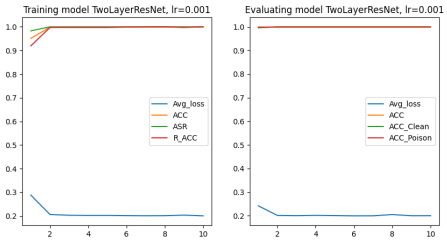
Slika 5.65: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.66: Treniranje i evaluacija male konvolucijske mreže.



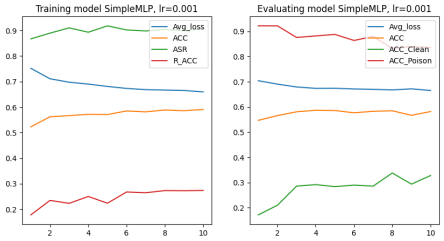
Slika 5.67: Treniranje i evaluacija konvolucijske-perceptronske mreže.



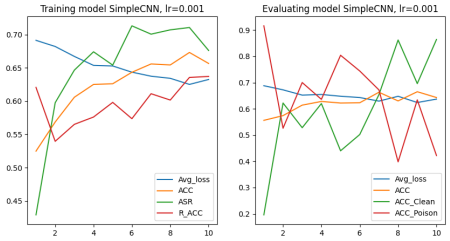
Slika 5.68: Treniranje i evaluacija dvoslojnog ResNet-18.

Kao i prije, modeli nemaju problema pri učenju klasifikacije podataka zatrovanih ovim tipom napada.

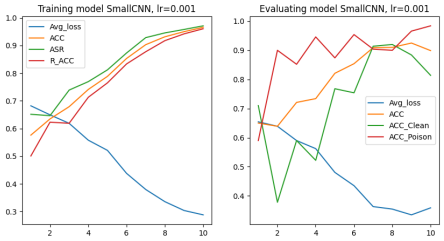
Treniranje i evaluacija na skupu za treniranje otrovanom "WaNet" napadom:



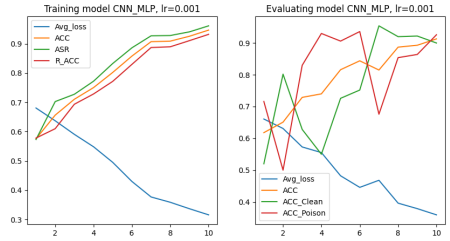
Slika 5.69: Treniranje i evaluacija jednostavnog višeslojnog perceptrona.



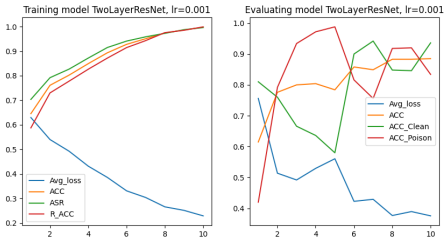
Slika 5.70: Treniranje i evaluacija jednostavne konvolucijske mreže.



Slika 5.71: Treniranje i evaluacija male konvolucijske mreže.



Slika 5.72: Treniranje i evaluacija konvolucijske-perceptronske mreže.



Slika 5.73: Treniranje i evaluacija dvoslojnog ResNet-18.

Složeniji modeli većinom uspijevaju bolje konvergirati od jednostavnijih modela. Može se primijetiti postizanje više točnosti naprema istog napada pomoću prošle mjere.

5.3.3. Rezultati nad testnim skupom

U ovom poglavlju će biti tablično prikazani rezultati na testnom skupu i grafički prikazane pripadajuće ROC krivulje za svaku mjeru zasebno. Testni skup, kao što je prije objašnjeno, čini jedan čisti i jedan potpuno zatrovani testni skup CIFAR-10 skupa podataka.

Tablice su prikazane za svaki okidač, u redcima se nalaze korišteni modeli, a u stupcima praćene testne mjere: testni gubitak ($\mathcal{L}$), ukupna točnost (T_{uk}), točnost na čistim

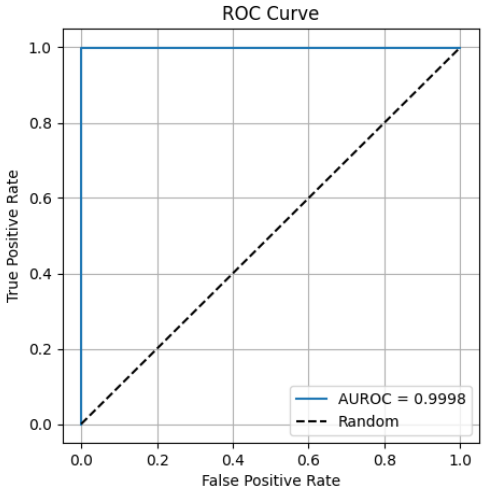
podacima ($T_{čis}$), točnost na otrovanim podacima (T_{zat}) i površina ispod ROC krivulje (AUROC). Modeli nad kojima se vršilo testiranje su modeli koji su tijekom treniranja imali najveću točnost na evaluacijskom skupu podataka.

Rezultati modela trenirani nad skupom dobivenom pomoću FCT mjere

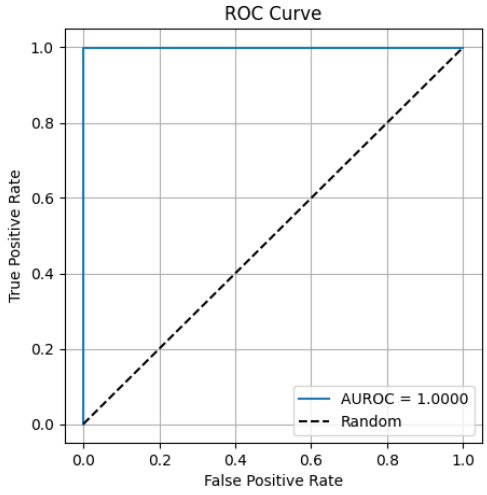
Tablica 5.8: Rezultati binarne klasifikacije za napad BadNets (okidač rešetke)

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.31	0.98	0.97	0.99	0.99
Jednostavna konvolucijska mreža	0.21	0.99	0.99	1.00	1.00
Mala konvolucijska mreža	0.20	0.99	0.99	1.00	1.00
Konvolucijska-perceptronska mreža	0.21	0.99	0.99	1.00	1.00
Dvoslojni ResNet-18	0.23	0.99	0.99	0.99	1.00

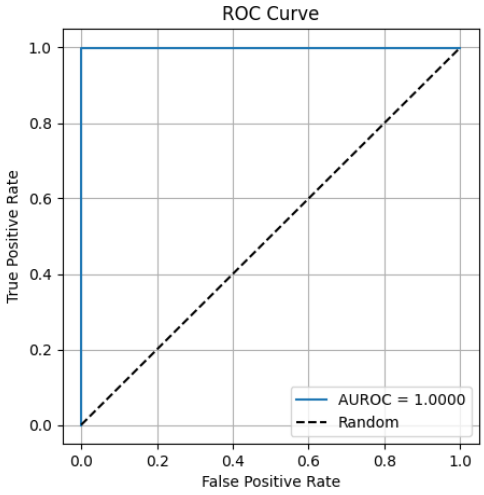
Iz tablice je vidljivo da svi modeli imaju vrlo dobre performanse na testnim podacima otrovani napadom BadNets okidačem rešetke.



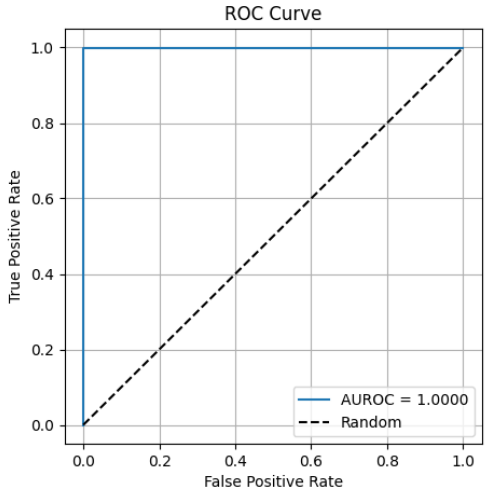
Slika 5.74: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim BadNets okidačem rešetke.



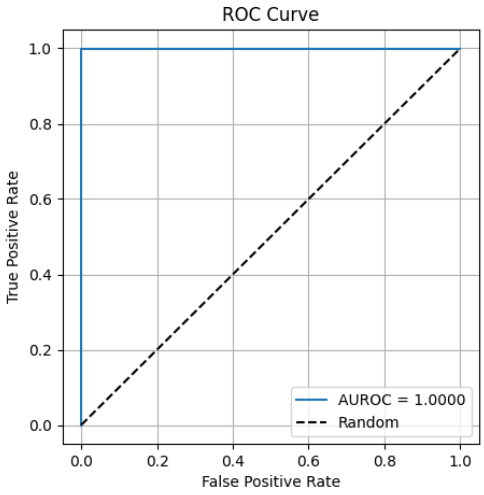
Slika 5.75: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim BadNets okidačem rešetke.



Slika 5.76: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim BadNets okidačem rešetke.



Slika 5.77: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim BadNets okidačem rešetke.



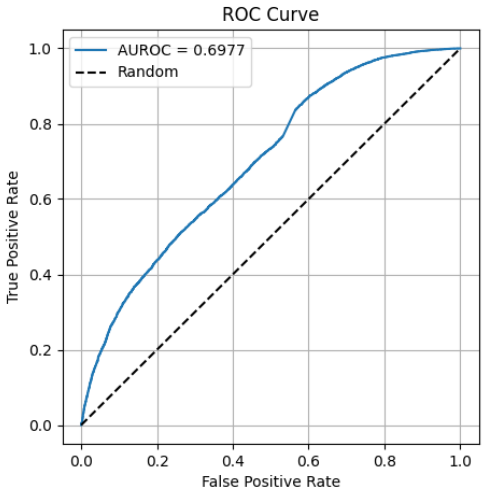
Slika 5.78: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim BadNets okidačem rešetke.

Kao što se vidi iz grafova ROC krivulja, svi modeli imaju vrlo dobre performanse za ovu vrstu napada.

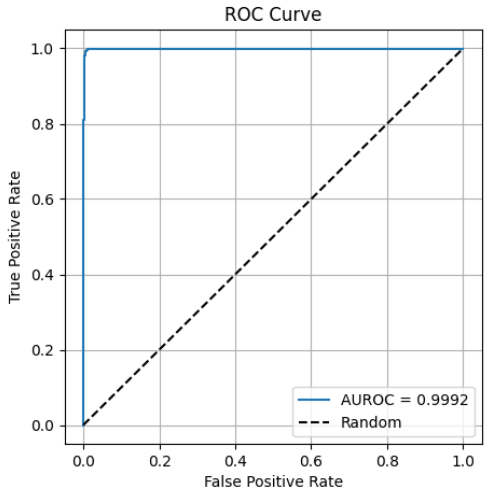
Tablica 5.9: Rezultati binarne klasifikacije za napad BadNets (okidač "Trojan")

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.63	0.66	0.49	0.82	0.70
Jednostavna konvolucijska mreža	0.25	0.99	0.99	0.98	0.99
Mala konvolucijska mreža	0.22	0.99	1.00	0.99	0.99
Konvolucijska-perceptronska mreža	0.23	0.99	0.99	0.99	0.99
Dvoslojni ResNet-18	0.39	0.87	0.99	0.74	0.99

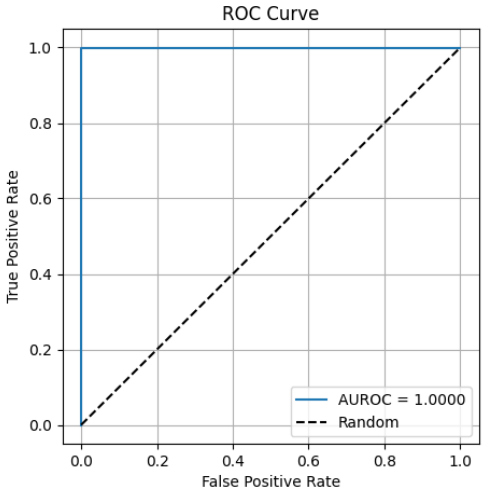
Slično kao i za prijašnji napad svi modeli imaju visoke performanse, oko 99% ukupne točnosti osim jednostavnog višeslojnog perceptrona koji postiže oko 66% ukupnu točnost.



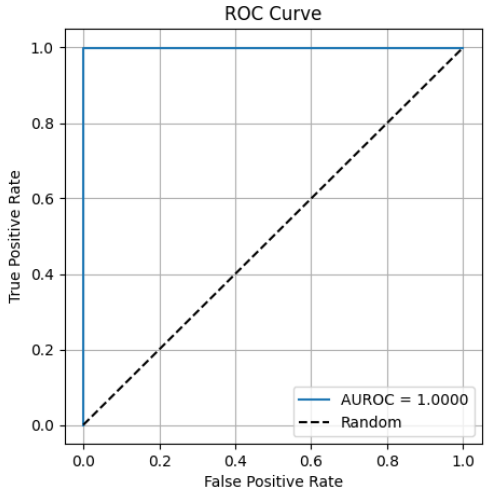
Slika 5.79: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim BadNets okidačem Trojan.



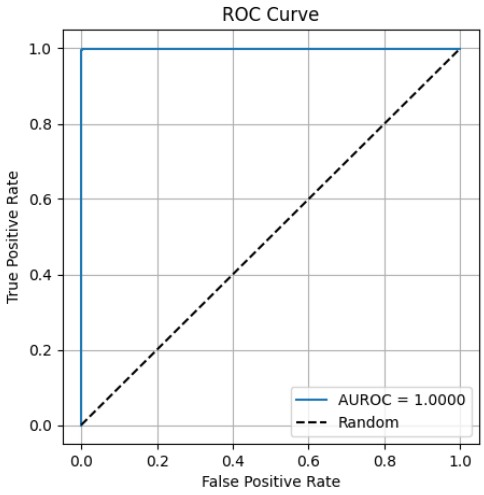
Slika 5.80: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim BadNets okidačem Trojan.



Slika 5.81: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim BadNets okidačem Trojan.



Slika 5.82: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim BadNets okidačem Trojan.



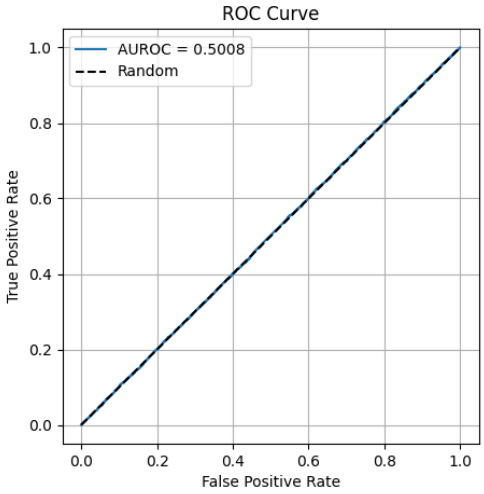
Slika 5.83: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim BadNets okidačem Trojan.

Iz grafova krivulja se vidi kako jedino model jednostavnog višeslojnog perceptrona ima lošije performanse.

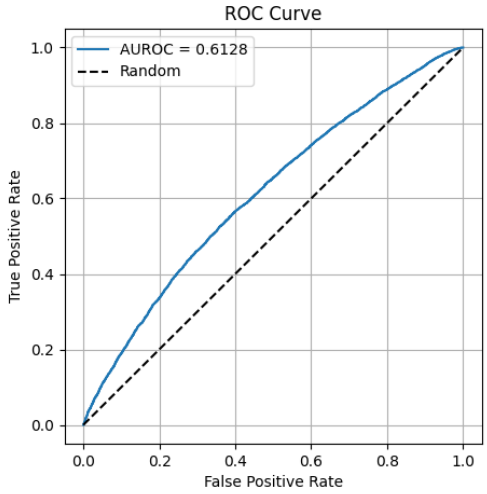
Tablica 5.10: Rezultati binarne klasifikacije za napad "WaNet"

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.82	0.50	0.40	0.60	0.50
Jednostavna konvolucijska mreža	0.78	0.60	0.62	0.58	0.61
Mala konvolucijska mreža	0.80	0.61	0.73	0.49	0.69
Konvolucijska-perceptronska mreža	0.80	0.60	0.75	0.45	0.66
Dvoslojni ResNet-18	0.86	0.64	0.84	0.44	0.64

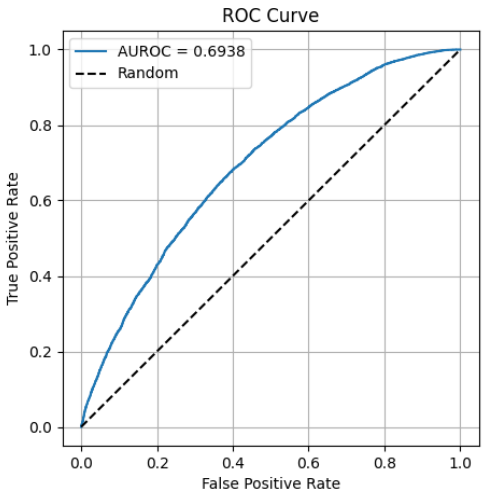
Rezultati za ovaj napad su zanimljivi jer se iz tablice vidi da je drugi po ukupnoj točnosti nakon modela dvoslojnog ResNet-18 model jednostavnog višeslojnog perceptrona. Napad je vrlo teško detektirati i čovjekovom oku što opravdava razlog da niti jedan model ne postiže točnost veću od 86%.



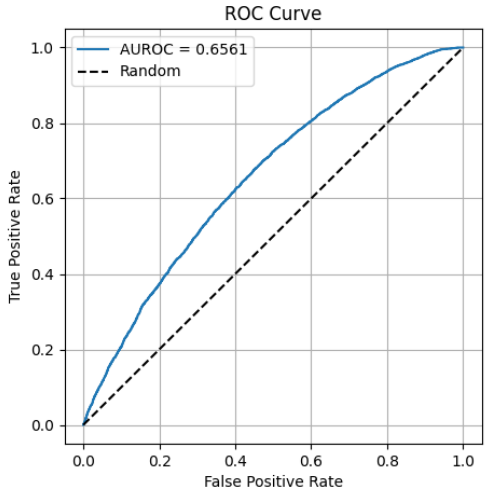
Slika 5.84: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim napadom WaNet.



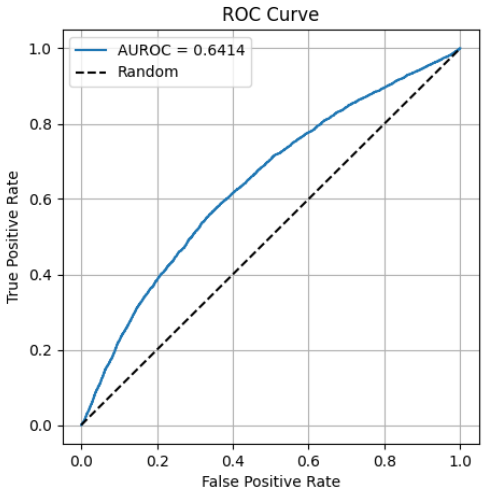
Slika 5.85: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim napadom WaNet.



Slika 5.86: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim napadom WaNet.



Slika 5.87: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim napadom WaNet.



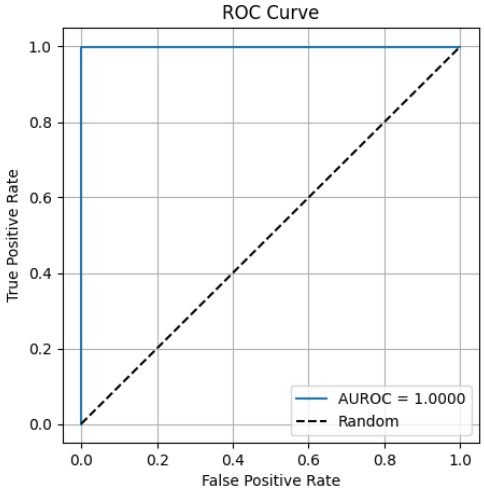
Slika 5.88: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim napadom WaNet.

Iz prikazanih krivulja je očito da je model jednostavnog višeslojnog perceptrona potpuno nasumičan, dok su ostali modeli nešto bolji po performansama.

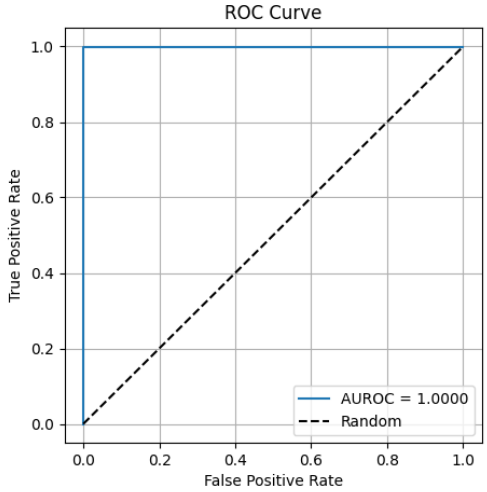
Tablica 5.11: Rezultati binarne klasifikacije za napad Blend (okidač "Signal")

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.31	0.97	0.94	0.99	0.99
Jednostavna konvolucijska mreža	0.21	0.99	0.99	0.99	0.99
Mala konvolucijska mreža	0.21	0.99	1.00	0.99	0.99
Konvolucijska-perceptronska mreža	0.21	0.99	0.99	0.99	0.99
Dvoslojni ResNet-18	0.21	0.99	0.99	0.99	0.99

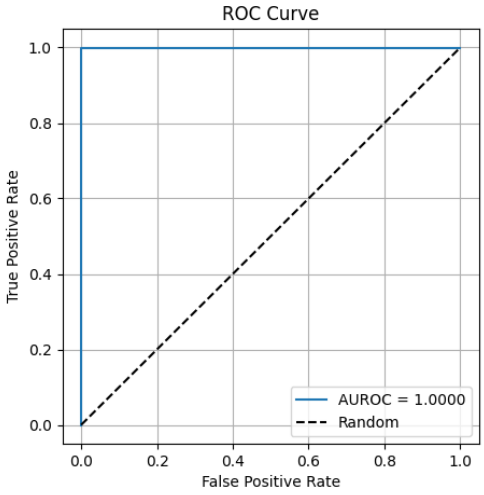
Za ovaj model svi modeli daju visoke točnosti od 99%, dok model jednostavnog višeslojnog perceptrona daje točnost od 97%.



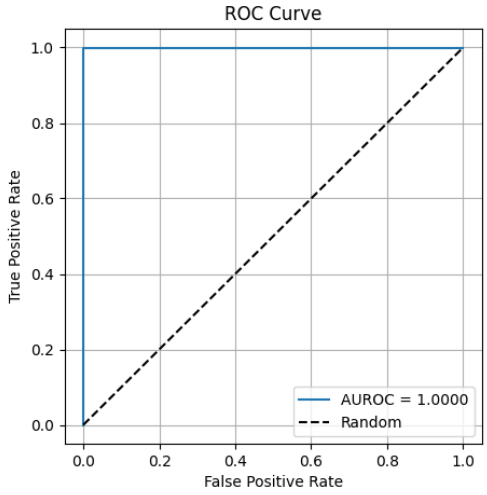
Slika 5.89: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim napadom Blend okidačem signal.



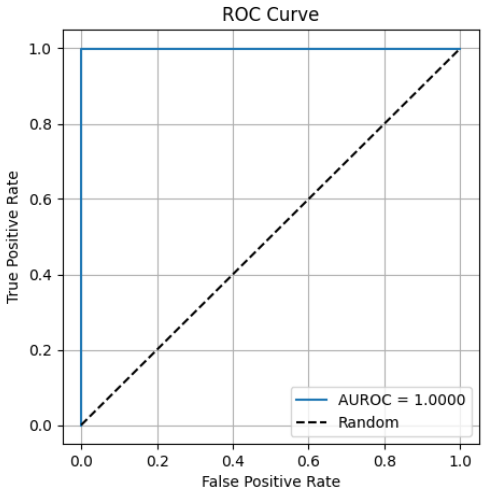
Slika 5.90: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim napadom Blend okidačem signal.



Slika 5.91: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim napadom Blend okidačem signal.



Slika 5.92: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim napadom Blend okidačem signal.



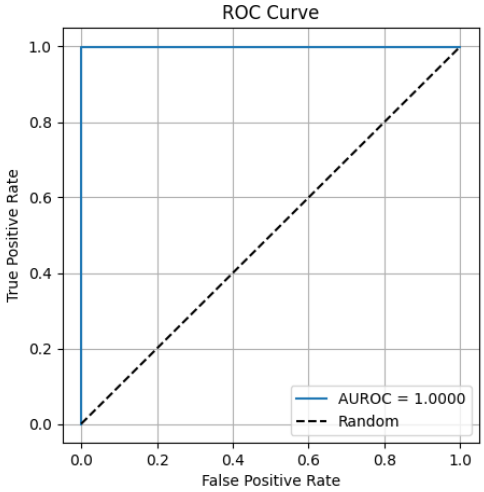
Slika 5.93: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim napadom Blend okidačem signal.

Grafovi krivulja pokazuju najviši stupanje performansi svih modela.

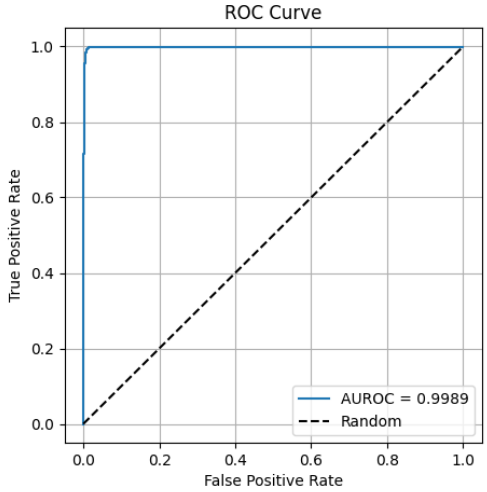
Tablica 5.12: Rezultati binarne klasifikacije za napad Blend (okidač "Hello Kitty")

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.34	0.93	0.85	1.00	1.00
Jednostavna konvolucijska mreža	0.28	0.98	0.99	0.95	0.99
Mala konvolucijska mreža	0.23	0.99	0.99	0.99	0.99
Konvolucijska-perceptronska mreža	0.22	0.99	0.99	0.99	0.99
Dvoslojni ResNet-18	0.24	0.98	0.99	0.96	0.99

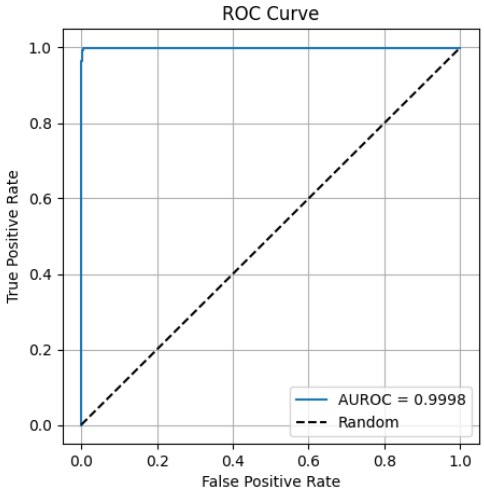
Za ovaj okidač svi modeli imaju točnost oko 98%, jednostavni višeslojni perceptron ima 93%. Ovo je očekivano, okidač je vrlo zamjetljiv na slici i čovjeku.



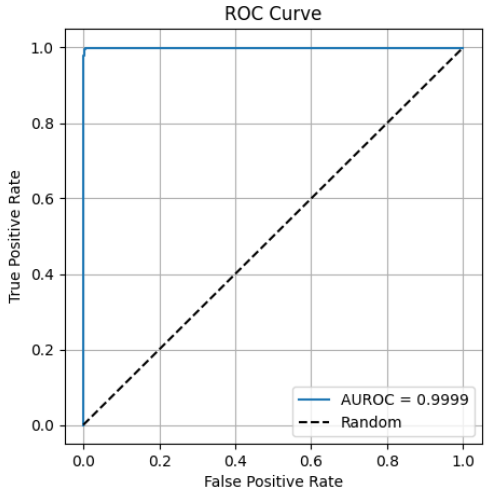
Slika 5.94: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim napadom Blend okidačem "Hello Kitty".



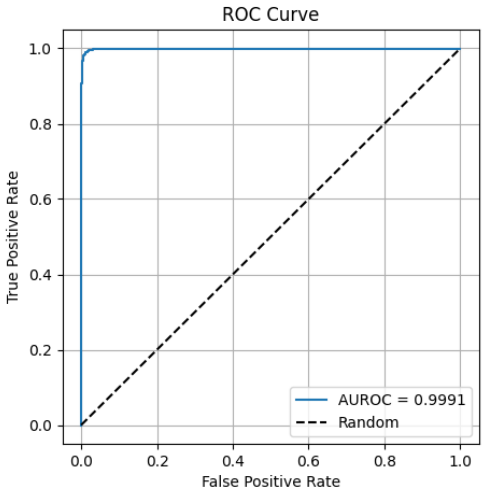
Slika 5.95: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim napadom Blend okidačem "Hello Kitty".



Slika 5.96: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim napadom Blend okidačem "Hello Kitty".



Slika 5.97: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim napadom Blend okidačem "Hello Kitty".



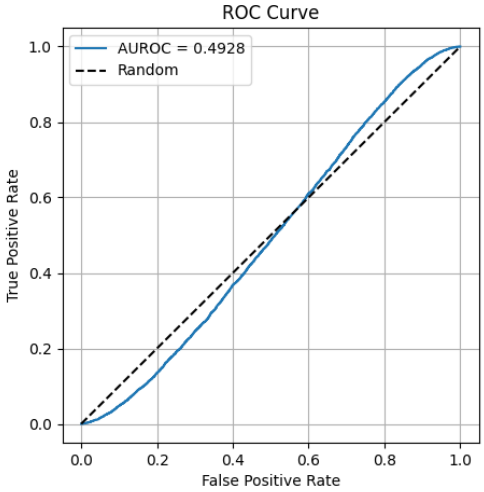
Slika 5.98: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim napadom Blend okidačem "Hello Kitty".

Prema grafovima iznad, svi modeli daju blizu optimalne performanse.

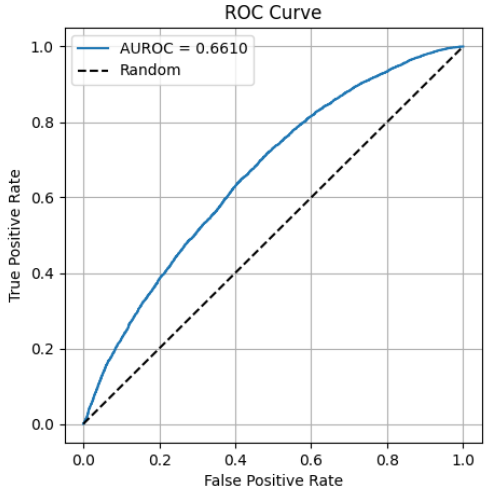
Tablica 5.13: Rezultati binarne klasifikacije za napad Blend (okidač nasumičnog piksela)

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.83	0.52	0.51	0.52	0.49
Jednostavna konvolucijska mreža	0.86	0.60	0.52	0.68	0.66
Mala konvolucijska mreža	0.94	0.56	0.59	0.52	0.56
Konvolucijska-perceptronska mreža	1.03	0.53	0.71	0.35	0.57
Dvoslojni ResNet-18	1.26	0.54	0.84	0.24	0.66

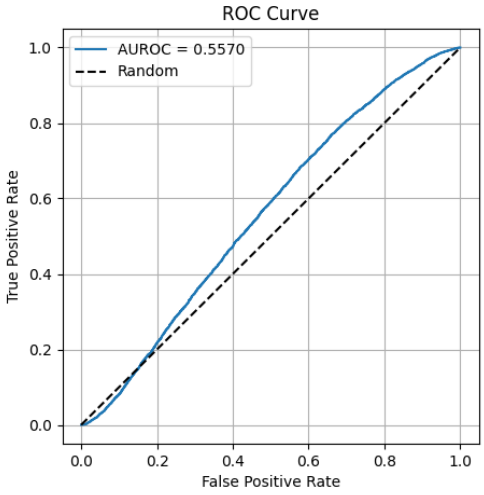
Iz tablice se iščitava kako svi modeli imaju dosta nisku ukupnu točnost. Ovo nije slučajno, napad nasumičnih piksela je dosta teško prepoznatljiv i ljudskom oku.



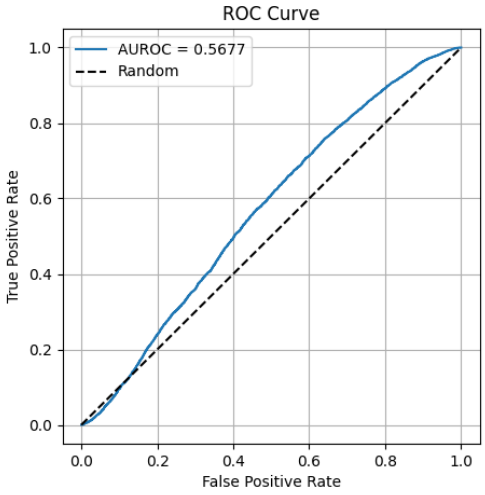
Slika 5.99: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.



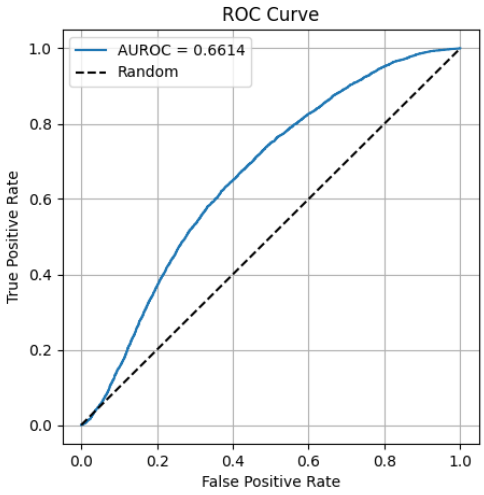
Slika 5.100: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.



Slika 5.101: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.



Slika 5.102: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.



Slika 5.103: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.

Iz prikazanih grafova krivulja se vidi kako modeli jednostavne konvolucijske mreže i dvoslojnog ResNet-18 imaju najbolje performanse, dok ostali modeli su malo bolji od nasumičnih klasifikatora. Najgori model je model jednostavnog perceptrona koji je lošiji od nasumičnog klasifikatora i ima točnost nižu od 50%, što znači da bi bio točniji kada bi zamijenili oznake predviđenim razredima.

Iz ovih podataka možemo napraviti tablicu s najboljim modelima i njihovim ukupnim točnostima na testnom skupu za svaku vrstu korištenih okidača 5.14.

Tablica 5.14: Najbolji model po ukupnoj točnosti za svaku vrstu okidača

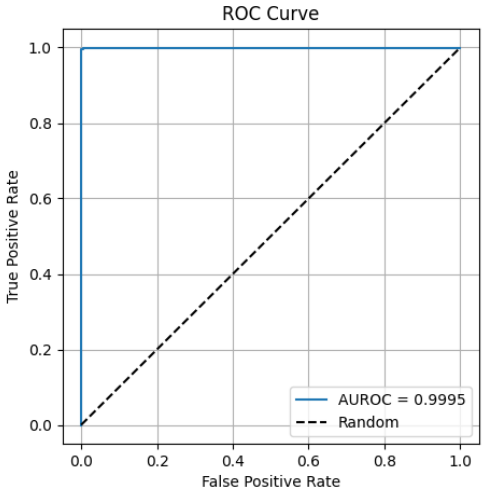
Okidač	**Model**	**Ukupna točnost**	**AUROC**
Rešetke	Jednostavna konvolucijska mreža	0.99	1.00
Trojan	Konvolucijska-perceptronska mreža	0.99	0.99
WaNet	Dvoslojni ResNet-18	0.64	0.64
Signal	Jednostavna konvolucijska mreža	0.99	0.99
Hello Kitty	Konvolucijska-perceptronska mreža	0.99	0.99
Nasumični piksel	Jednostavna konvolucijska mreža	0.60	0.66

Rezultati modela trenirani nad skupom dobivenom pomoću ABL mjere

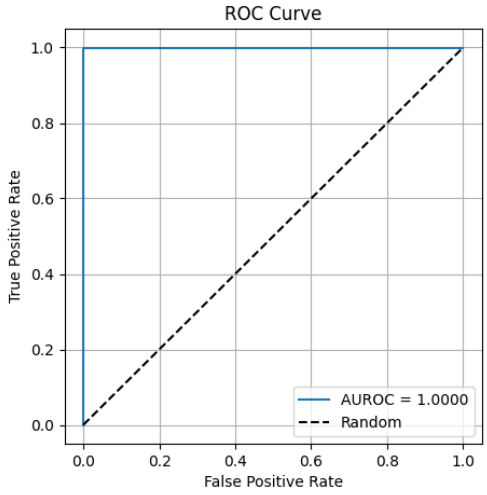
Tablica 5.15: Rezultati binarne klasifikacije za napad BadNets (okidač rešetke)

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.38	0.96	0.92	0.99	0.99
Jednostavna konvolucijska mreža	0.20	1.00	1.00	1.00	1.00
Mala konvolucijska mreža	0.20	1.00	1.00	1.00	1.00
Konvolucijska-perceptronska mreža	0.20	1.00	1.00	1.00	1.00
Dvoslojni ResNet-18	0.20	1.00	1.00	1.00	1.00

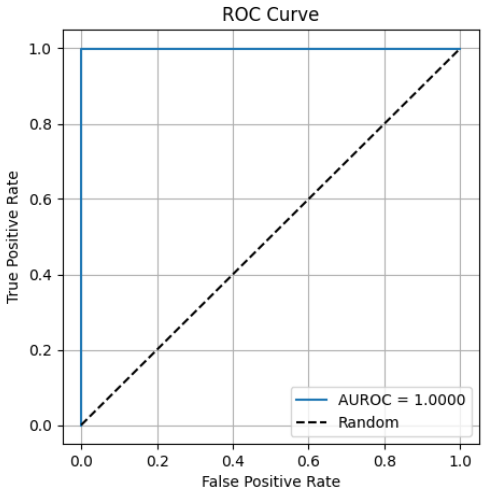
Kao i kod prošle mjere, svi modeli imaju visoku uspješnost za ovaj napad. Ovo je očekivano, napad je vrlo lako uočljiv.



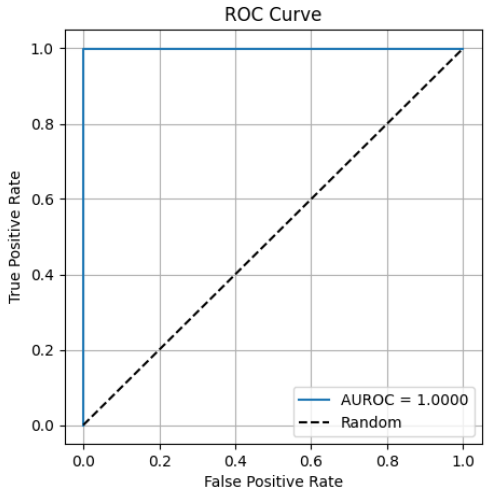
Slika 5.104: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim BadNets okidačem rešetke.



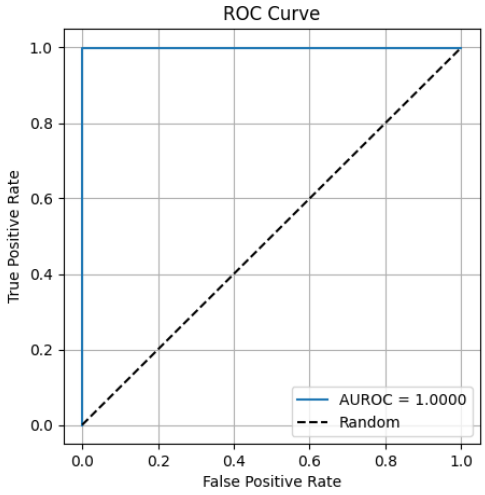
Slika 5.105: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim BadNets okidačem rešetke.



Slika 5.106: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim BadNets okidačem rešetke.



Slika 5.107: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim BadNets okidačem rešetke.



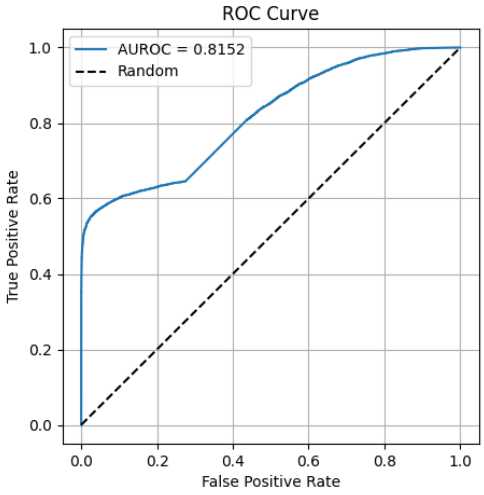
Slika 5.108: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim BadNets okidačem rešetke.

Po grafovima je vidljivo da svi modeli imaju optimalne performanse.

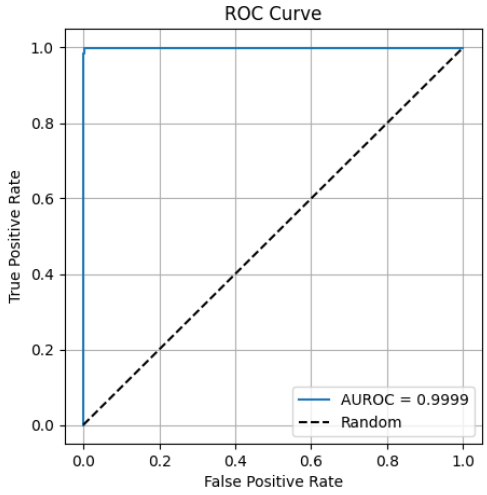
Tablica 5.16: Rezultati binarne klasifikacije za napad BadNets (okidač "Trojan")

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.57	0.75	0.95	0.56	0.82
Jednostavna konvolucijska mreža	0.24	0.99	0.99	0.97	0.99
Mala konvolucijska mreža	0.22	0.99	1.00	0.99	1.00
Konvolucijska-perceptronska mreža	0.21	0.99	1.00	0.99	0.99
Dvoslojni ResNet-18	0.21	0.99	1.00	0.99	0.99

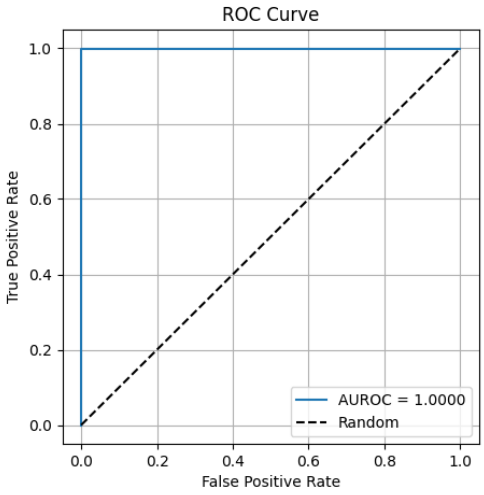
Za razliku od prošle mjere, sada svi modeli imaju točnost od 99% osim modela jednostavnog višeslojnog perceptrona koji ima točnost od 75%.



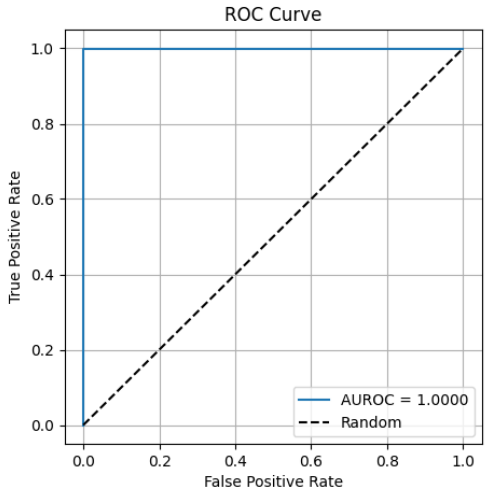
Slika 5.109: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim BadNets okidačem Trojan.



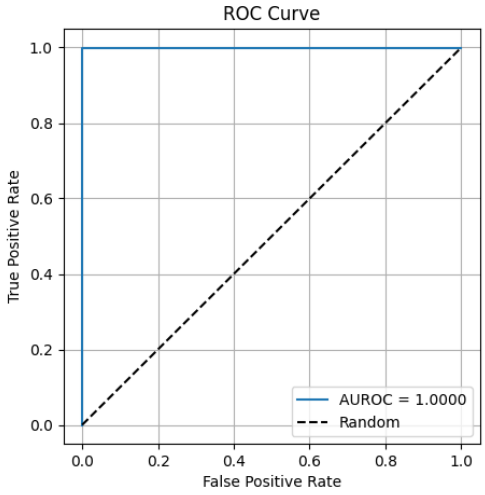
Slika 5.110: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim BadNets okidačem Trojan.



Slika 5.111: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim BadNets okidačem Trojan.



Slika 5.112: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim BadNets okidačem Trojan.



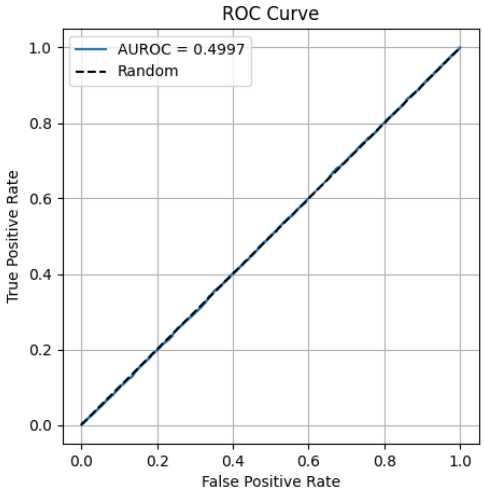
Slika 5.113: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim BadNets okidačem Trojan.

Na grafovima ROC krivulja su također vidljive savršene performanse svih modela osim prvoga koji je nešto slabiji.

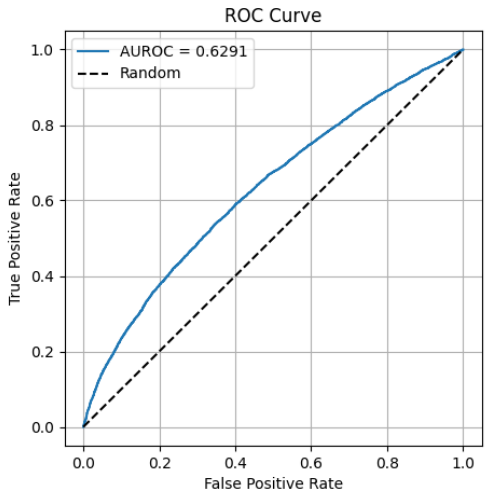
Tablica 5.17: Rezultati binarne klasifikacije za napad "WaNet"

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.74	0.50	0.36	0.64	0.50
Jednostavna konvolucijska mreža	0.75	0.53	0.87	0.20	0.63
Mala konvolucijska mreža	0.93	0.59	0.96	0.23	0.81
Konvolucijska-perceptronska mreža	1.20	0.54	0.97	0.11	0.73
Dvoslojni ResNet-18	1.18	0.55	0.98	0.12	0.74

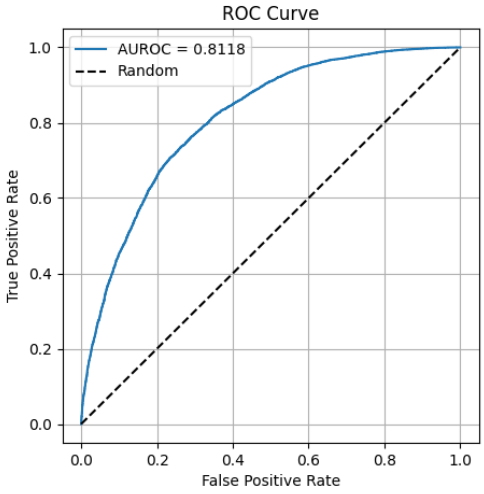
Ova tablica je jedina koja ima ponešto lošije mjere točnosti u usporedbi na prijašnju FCT mjeru. Modeli imaju točnost oko 50%, a najbolji model je mala konvolucijska mreža s 59% točnosti. Ovo je prihvatljivo jer je ovaj napad jedan od težih za detekciju kao što je i prije rečeno.



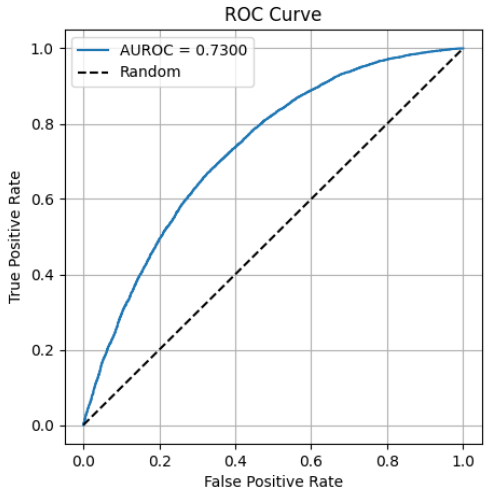
Slika 5.114: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim napadom WaNet.



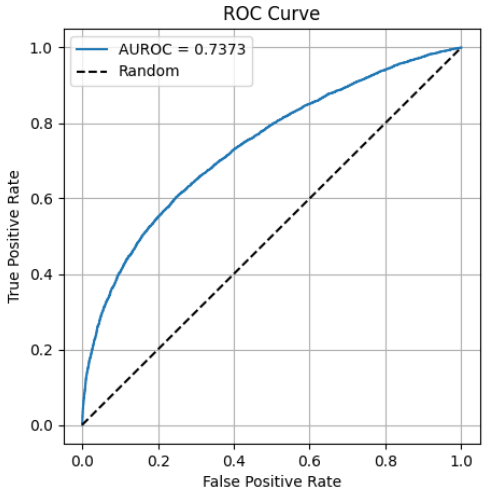
Slika 5.115: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim napadom WaNet.



Slika 5.116: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim napadom WaNet.



Slika 5.117: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim napadom WaNet.



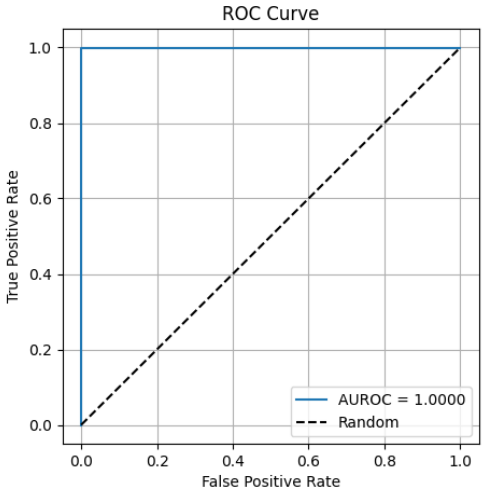
Slika 5.118: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim napadom WaNet.

S grafova ROC krivulja su također očite slabije performanse modela, gdje je vidljivo kako se model jednostavnog višeslojnog perceptrona ponaša kao potpuno nasumičan model.

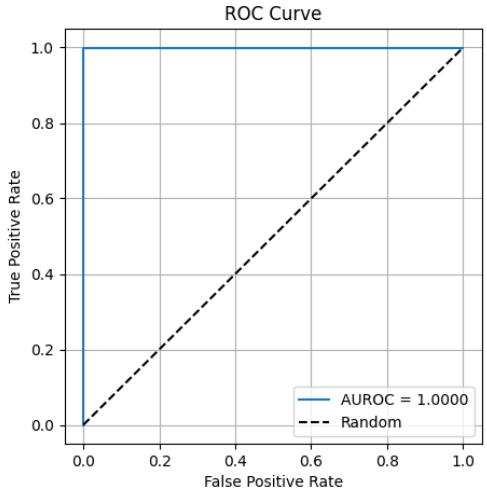
Tablica 5.18: Rezultati binarne klasifikacije za napad Blend (okidač "Signal")

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.28	0.99	0.99	1.00	0.99
Jednostavna konvolucijska mreža	0.21	0.99	1.00	0.99	0.99
Mala konvolucijska mreža	0.20	0.99	1.00	0.99	1.00
Konvolucijska-perceptronska mreža	0.20	0.99	1.00	0.99	1.00
Dvoslojni ResNet-18	0.20	1.00	1.00	0.99	0.99

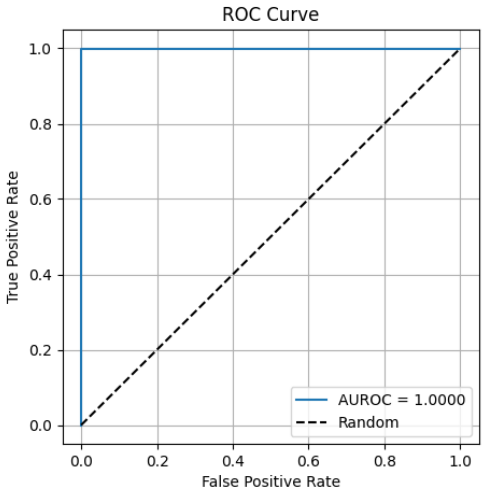
Kao i kod prijašnje FCT mjere, ovaj napad je lagano uočljiv i rezultira visokim točnostima svih modela.



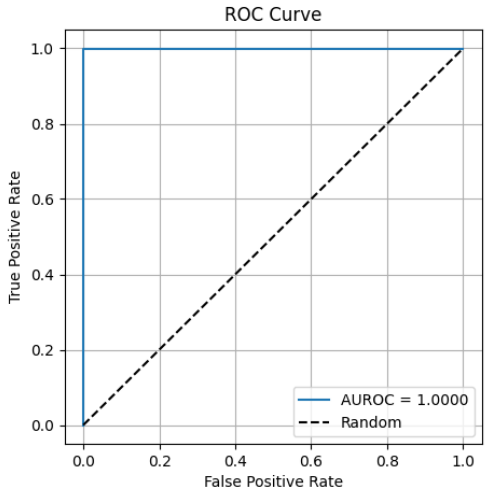
Slika 5.119: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim napadom Blend okidačem signal.



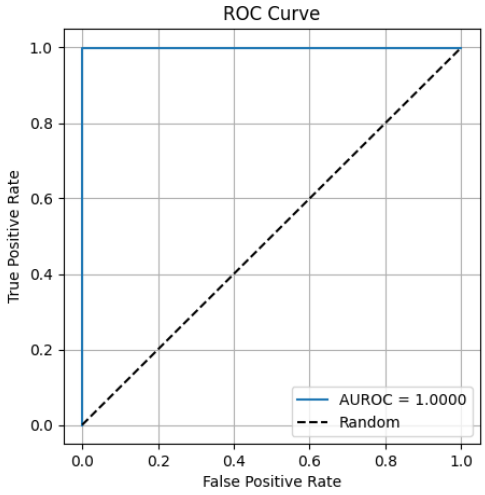
Slika 5.120: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim napadom Blend okidačem signal.



Slika 5.121: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim napadom Blend okidačem signal.



Slika 5.122: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim napadom Blend okidačem signal.



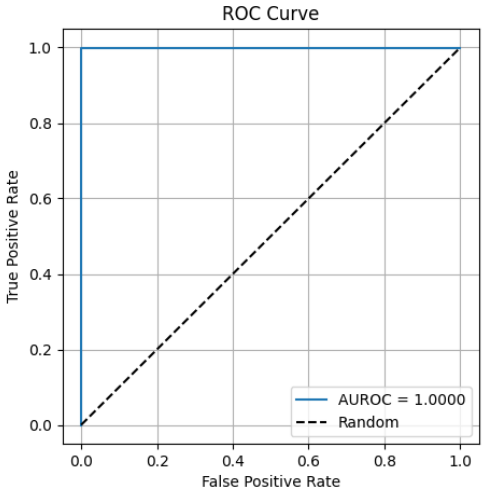
Slika 5.123: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim napadom Blend okidačem signal.

S grafova ROC krivulja se također vide optimalne performanse.

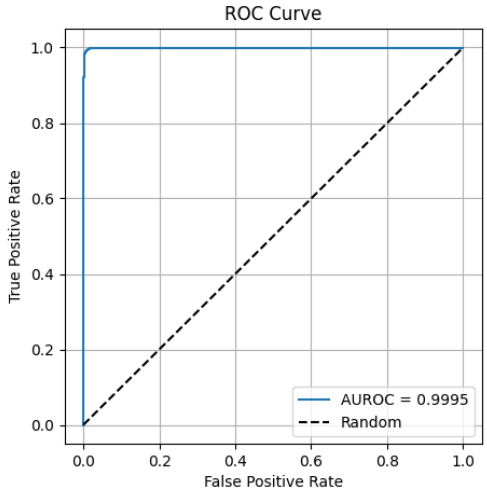
Tablica 5.19: Rezultati binarne klasifikacije za napad Blend (okidač "Hello Kitty")

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.30	0.95	0.89	1.00	0.99
Jednostavna konvolucijska mreža	0.28	0.97	0.99	0.93	0.99
Mala konvolucijska mreža	0.21	1.00	1.00	0.99	1.00
Konvolucijska-perceptronska mreža	0.21	0.99	1.00	0.99	1.00
Dvoslojni ResNet-18	0.21	0.99	0.99	0.99	0.99

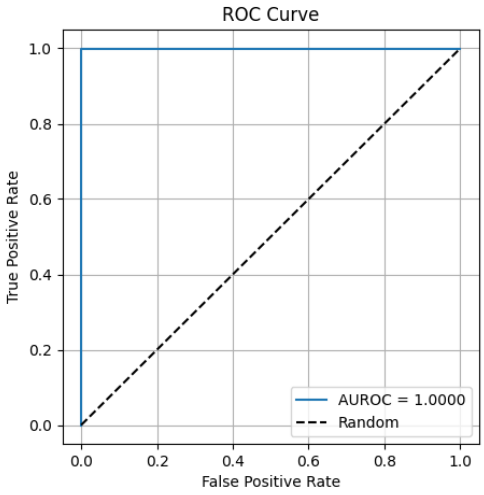
Vidi se iz tablice da za ovaj napad svi modeli postižu visoku razinu točnosti kao i kod prijašnje mjere.



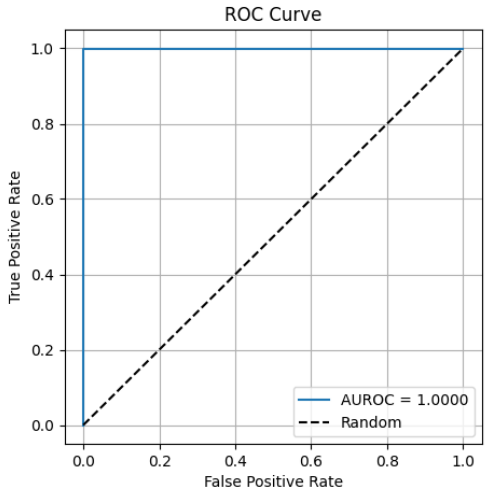
Slika 5.124: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim napadom Blend okidačem "Hello Kitty".



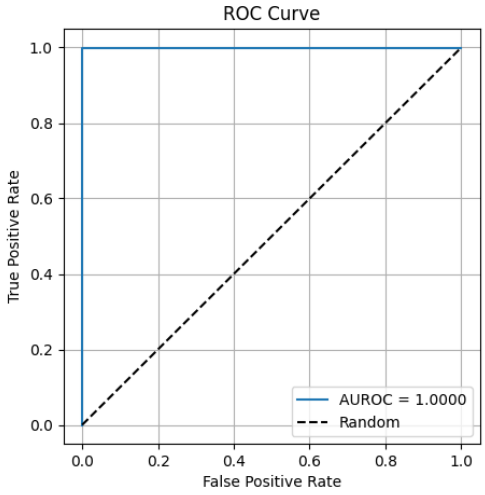
Slika 5.125: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim napadom Blend okidačem "Hello Kitty".



Slika 5.126: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim napadom Blend okidačem "Hello Kitty".



Slika 5.127: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim napadom Blend okidačem "Hello Kitty".



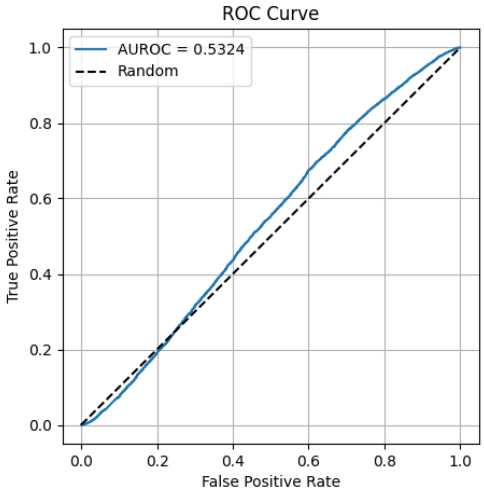
Slika 5.128: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim napadom Blend okidačem "Hello Kitty".

Grafovi ROC krivulja prikazuju odlične performanse modela.

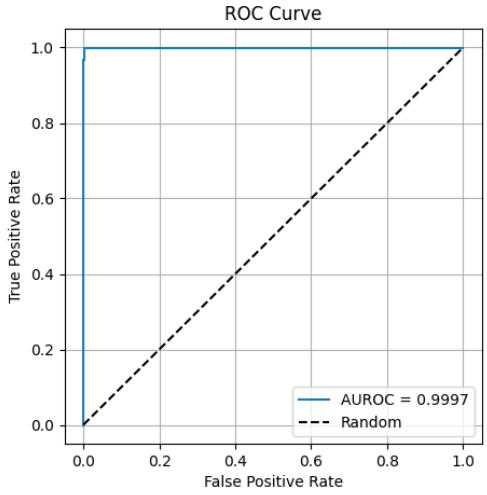
Tablica 5.20: Rezultati binarne klasifikacije za napad Blend (okidač nasumičnog piksela)

Model	$\mathcal{L}$	T_{uk}	$T_{čis}$	T_{zat}	AUROC
Jednostavni višeslojni perceptron	0.89	0.50	0.91	0.10	0.53
Jednostavna konvolucijska mreža	0.26	0.98	0.99	0.96	0.99
Mala konvolucijska mreža	0.49	0.75	1.00	0.50	0.99
Konvolucijska-perceptronska mreža	0.94	0.53	1.00	0.06	0.99
Dvoslojni ResNet-18	0.29	0.94	0.99	0.89	0.99

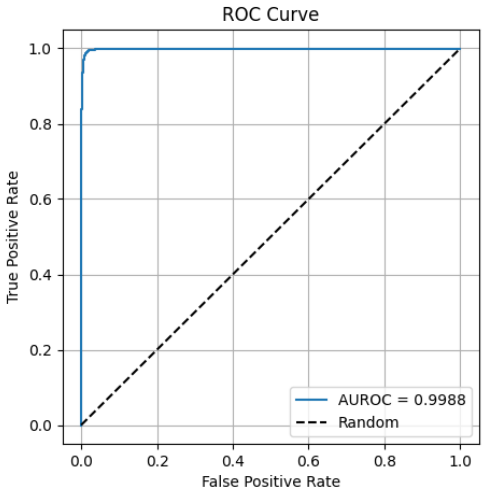
Tablica prikazuje daleko bolje performanse svih modela s usporedbom na performanse iz prethodnog dijela vezanih uz prijašnju mjeru. Najbolji model po ukupnoj točnosti je model jednostavne konvolucijske mreže s 98% točnosti.



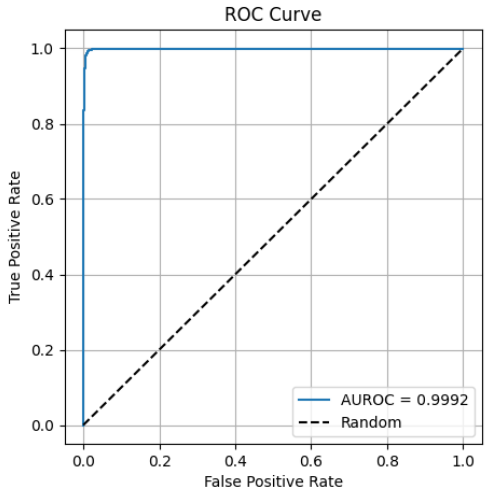
Slika 5.129: ROC krivulja dobivena testiranjem modela jednostavnog višeslojnog perceptrona nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.



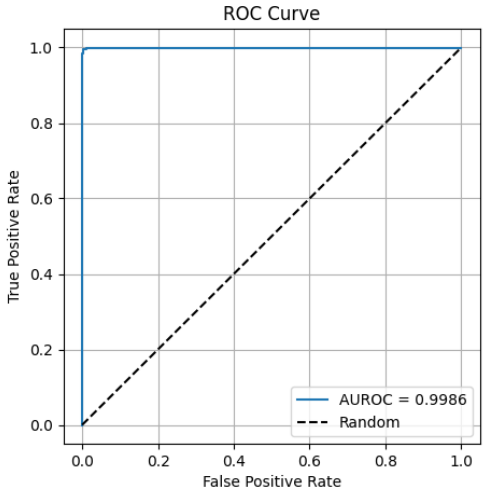
Slika 5.130: ROC krivulja dobivena testiranjem modela jednostavne konvolucijske mreže nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.



Slika 5.131: ROC krivulja dobivena testiranjem modela male konvolucijske mreže nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.



Slika 5.132: ROC krivulja dobivena testiranjem modela konvolucijske-perceptronske mreže nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.



Slika 5.133: ROC krivulja dobivena testiranjem modela dvoslojnog ResNet-18 nad podacima napadnutim napadom Blend okidačem nasumičnog piksela.

Iz grafova ROC krivulje se također očituju izvrsne performanse svih modela osim jednostavnog višeslojnog perceptrona koji je nešto bolji od nasumičnog modela, no sigurno bolji od rezultata dobivenog uz pomoć prve mjere.
Iz ovih podataka možemo napraviti tablicu s najboljim modelima i njihovim ukupnim točnostima na testnom skupu za svaku vrstu korištenih okidača 5.14.

Tablica 5.21: Najbolji model po ukupnoj točnosti za svaku vrstu okidača

Okidač	**Model**	**Ukupna točnost**	**AUROC**
Rešetka	Jednostavna konvolucijska mreža	1.00	1.00
Trojan	Konvolucijska-perceptronska mreža	0.99	0.99
WaNet	Mala konvolucijska mreža	0.59	0.81
Signal	Dvoslojni ResNet-18	1.00	0.99
Hello Kitty	Mala konvolucijska mreža	1.00	1.00
Nasumični piksel	Jednostavna konvolucijska mreža	0.98	0.99

5.3.4. Povezani rad

U ovom poglavlju će biti uspoređene AUROC mjere metoda detekcije zatrovanih podataka iz tri rada koji su imali sličnu motivaciju kao ovaj rad. Pri usporedbi će biti navedena vrijednost najboljeg modela za korišteni okidač i za korištenu mjeru iz pri-

jašnjih tablica.
Rad (Guo et al. (2023)) je imao motivaciju proizvesti crnu kutiju za detekciju i praćenje otrovanih podataka koja bi funkcionirala kao vatrozid pri ulazu na model. Prvo su proučavali povećanje na razini piksela na čistim i zatrovanom skupu podataka motiviranim razumijevanjem da ako se povećavaju vrijednosti okidača da to neće utjecati, niti pojačati, stopu uspješnosti napada na model. Demonstriraju kako predikcije od napadnutih slika, generiranih od klasičnih i naprednijih napada su vrlo više konzistentnije od onih od čistih podataka kada se povećaju vrijednosti svih piksela. Ovaj fenomen je nazvan skalirana konzistencija predikcija ("scaled prediction consistency", SPC). Prema ovom fenomenu su razvili efektivnu metodu nazvanu SCALE-UP koja radi nad slobodnim i ograničenim podacima. Nad slobodnim podacima metoda mjeri SPC vrijednost za svaki uzorak, koja je udio oznaka skaliranih slika koje su konzistentne s ulaznom slikom. Što je veća ova vrijednost, vjerojatnije je da je ulaz otrovan. Kod postavke ograničenih podataka pretpostavlja se da branitelj ima manje čistih podataka iz svakog razreda, čime se mogu smanjiti nuspojave razlika među razredima, što poboljšava metodu SCALE-UP.
U tablici 5.22 se može vidjeti usporedba vrijednosti AUROC njihove metode i metode ovog rada.

Tablica 5.22: AUROC vrijednosti za postavke metode SCALE-UP (Guo et al. (2023)) i metode obrane mjerom FCT i ABL od napada BadNets i WaNet.

Metoda	**BadNets**	**WaNet**
SCALE-UP (data-free)	0.97	0.92
SCALE-UP (data-limited)	0.97	0.93
(Ours) FCT	1.00	0.64
(Ours) ABL	1.00	0.81

Drugi rad (Pal et al. (2024)) , slično kao i ovaj rad, predlaže gubitak koji služi za procjenu zatrovanosti podatka. Gubitak su nazvali SPC gubitak, motivirani već spomenutom SPC mjerom. Gubitak su formulirani na način:

$$\ell_{\text{SPC}}(\mathbf{x}) = \sum_{n \in S} \frac{\mathbb{1}\!\!\!\!\mathbb{1} \left(\arg\max \mathcal{F}_\theta(\mathbf{x}) = \arg\max \mathcal{F}_\theta(n \cdot \mathbf{x}) \right)}{|S|}$$

, gdje je $\mathbb{1}\!\!\!\!\mathbb{1}$ indikatorska funkcija, S predstavlja skup skala, a n je konstanta skaliranja. Ako je dan uzorak x i skup skala S onda će vrijednost ovog gubitka mjeriti slaganje između predikcije modela uzorka x i skalarnog multiplikatora od x. Prema ovome

otrovani uzorci bi trebali imati nisku vrijednost ovog SPC gubitka jer će otrovani pikseli nestati ili se stopiti s pozadinom kada se množe s vrijednostima visokih skala. U drugu ruku, čisti podaci će imati visoku vrijednost SPC gubitka jer će predikcijske značajke glavnog objekta ostati netaknute čak i s visokim vrijednostima skala.
U tablici 5.23 se može vidjeti usporedba vrijednosti AUROC njihove metode i metode ovog rada.

Tablica 5.23: AUROC vrijednosti za postavke metode SCP gubitak (Pal et al. (2024)) i metode obrane mjerom FCT i ABL od napada BadNets i WaNet.

Metoda	**BadNets**	**Blend**	**Trojan**	**WaNet**
SCP loss	0.95	0.95	0.95	0.93
(Ours) FCT	1.00	0.99	0.99	0.64
(Ours) ABL	1.00	1.00	0.99	0.81

Treći rad (Li et al. (2025)) je razvio metodu zvanu skraćeno PSBD (Prediction Shift Backdoor Detection) koja je bila inspirirana fenomenom posmaka predikcija (prediction shift) PS, koji je zapažen kada predikcije otrovanog modela na čistim podacima odstupaju od točnih oznaka, krećući se drugačijim oznakama pogotovo kada se uvede sloj propadanja tijekom odluke modela. U drugu ruku opaženo je da otrovani podaci su manje osjetljivi na ovaj fenomen.
U tablici 5.24 se može vidjeti usporedba vrijednosti AUROC njihove metode i metode ovog rada.

Tablica 5.24: AUROC vrijednosti za postavke metode PSDB (Li et al. (2025)) i metode obrane mjerom FCT i ABL od napada BadNets i WaNet.

Metoda	**BadNets**	**Blend**	**Trojan**	**WaNet**
PSDB	1.00	1.00	0.98	1.00
(Ours) FCT	1.00	0.99	0.99	0.64
(Ours) ABL	1.00	1.00	0.99	0.81

6. Zaključak

Motivacija rada je bila analiza i detekcija napada trovanjem podataka kroz stražnja vrata. Objašnjena je ranjivost modela na takve napade i opisana su svojstva napada koja se mogu iskoristiti za obranu od istih. Opisan je korišteni skup podataka, 6 vrsti okidača korištenih za implementaciju napada i model kojeg se napada. Predložene su dvije mjere koje se koriste za filtriranje otrovanih podataka od čistih. U radu je opisana metoda za razvoj kvalitetnog binarnog klasifikatora koji može uspješno raspoznati čiste i otrovane podatke. Ideja je bila iskoristiti histograme mjera za formiranje skupa podataka za treniranje binarnog klasifikatora tako da se određeni postotak krajnjih vrijednosti histograma iskoristi kako bi ih s visokim pouzdanjem mogli označiti kao otrovane ili kao čiste. Prikazane su tablice s 5 različitih arhitektura za implementaciju binarnog klasifikatora. U eksperimentalnom dijelu rada je prikazan postupak trovanja temeljnog modela ResNet-18. Nakon toga je prikazano generiranje histograma korištenih mjera za svaku vrstu okidača. Dalje, prikazani su grafovi treniranja i evaluacije svakog modela nad skupom podataka za treniranje i evaluaciju sastavljenih pomoću dobivenih histograma. Prikazane su tablice u kojima su navedene mjere performansi svakog modela nad skupu za testiranje. Pokazano je da je po ukupnom broju dobro označenih slika i po performansama svih modela binarnog klasifikatora ABL mjera bolja od FCT mjere za rješenje ovog problema.

Na kraju su prikazane tablice s rezultatima svakog modela binarnog klasifikatora na testnim skupom. Potvrđuje se da je dovoljna mala konvolucijska mreža za vrlo uspješnu klasifikaciju otrovanih i čistih podataka za većinu isprobanih okidača. Za budući rad bi bilo dobro isprobati još različitih vrsta okidača i arhitektura binarnog klasifikatora.

LITERATURA

Geoffrey Hinton Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

Neena Aloysius i M Geetha. A review on deep convolutional neural networks. U *2017 international conference on communication and signal processing (ICCSP)*, stranice 0588–0592. IEEE, 2017.

Mauro Barni, Kassem Kallas, i Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. U *2019 IEEE International Conference on Image Processing (ICIP)*, stranice 101–105. IEEE, 2019.

Weixin Chen, Baoyuan Wu, i Haoqian Wang. Effective backdoor defense by exploiting sensitivity of poisoned samples. *Advances in Neural Information Processing Systems*, 35:9727–9737, 2022.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, i Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Stephen Grossberg. Recurrent neural networks. *Scholarpedia*, 8(2):1888, 2013.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, i Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, i Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023.

Kaiming He i Jian Sun. Convolutional neural networks at constrained time cost. U *Proceedings of the IEEE conference on computer vision and pattern recognition*, stranice 5353–5360, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, i Jian Sun. Deep residual learning for image recognition. U *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, i Kui Ren. Backdoor defense via decoupling the training process. *arXiv preprint arXiv:2202.03423*, 2022.

Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, i Masashi Sugiyama. Do we need zero training loss after achieving zero training error? *arXiv preprint arXiv:2002.08709*, 2020.

Alex Krizhevsky, Vinod Nair, i Geoffrey Hinton. The cifar-10 dataset. `https://www.cs.toronto.edu/~kriz/cifar.html`. Accessed: 2023-06-02.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Geoffrey Hinton Laurens Van der Maaten. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

Yann LeCun, Yoshua Bengio, i Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Wei Li, Pin-Yu Chen, Sijia Liu, i Ren Wang. Psbd: Prediction shift uncertainty unlocks backdoor detection. U *Proceedings of the Computer Vision and Pattern Recognition Conference*, stranice 10255–10264, 2025.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, i Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021a.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, i Siwei Lyu. Invisible backdoor attack with sample-specific triggers. U *Proceedings of the IEEE/CVF international conference on computer vision*, stranice 16463–16472, 2021b.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, i Xiangyu Zhang. Trojaning attack on neural networks. U *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

Yunfei Liu, Xingjun Ma, James Bailey, i Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. U *Computer vision–ECCV 2020: 16th*

European conference, Glasgow, UK, August 23–28, 2020, proceedings, part X 16, stranice 182–199. Springer, 2020.

Anh Nguyen i Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.

Soumyadeep Pal, Yuguang Yao, Ren Wang, Bingquan Shen, i Sijia Liu. Backdoor secrets unveiled: Identifying backdoor data with optimized scaled prediction consistency. *arXiv preprint arXiv:2403.10717*, 2024.

P. Simard Y. Bengio i P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks, 5(2):157–166*, 1994.

Analiza napada kroz stražnja vrata s obzirom na razdvojivosti zatrovanih podataka

Sažetak

U ovome radu su proučene ranjivosti rezidualnog modela ResNet-18 na napade kroz stražnja vrata i predložena metode razdvajanja otrovanih podataka od čistih. Objašnjeni su i opisani korišteni okidači za implementaciju napada i skup za učenje. Predložene su dvije mjere koje se koriste za filtraciju otrovanih podataka od čistih. Eksperimentalno je prikazano treniranje binarnog klasifikatora i grafički i tablično su prikazani svi rezultati eksperimenata. U zaključku su diskutirani rezultati i dane ideje za daljnji rad.

Ključne riječi: Nadzirano učenje, ResNet-18, CIFAR-10, napadi kroz stražnja vrata, mjere za odvajanje podataka, binarni klasifikator.

Backdoor attack analysis considering the separability of poisoned data

Abstract

In this paper, the vulnerabilities of the residual model ResNet-18 to backdoor attacks were studied and methods were proposed to separate the poisoned data from the clean data. The triggers used to implement the attack and the learning set are explained and described. Two metrics are proposed that are used to filter the poisoned data from the clean data. The training of the binary classifier is shown experimentally, and all the results of the experiments are presented graphically and tabularly. In the conclusion, the results are discussed and ideas for further work are given.

Keywords: Supervised learning, ResNet-18, CIFAR-10, backdoor attacks, data separation metrics, binary classifier.