

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1312

**USPOREDBA METODA STROJNOG I DUBOKOG UČENJA U
RJEŠAVANJU PROBLEMA NEURAVNOTEŽENOSTI KLASA
NA TABLIČNIM I SLIKOVNIM SKUPOVIMA PODATAKA**

Andrea Milanović

Zagreb, lipanj 2026.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1312

**USPOREDBA METODA STROJNOG I DUBOKOG UČENJA U
RJEŠAVANJU PROBLEMA NEURAVNOTEŽENOSTI KLASA
NA TABLIČNIM I SLIKOVNIM SKUPOVIMA PODATAKA**

Andrea Milanović

Zagreb, lipanj 2026.

DIPLOMSKI ZADATAK br. 1312

Pristupnica: **Andrea Milanović (0036543226)**

Studij: Računarstvo

Profil: Računarska znanost

Mentor: prof. dr. sc. Alan Jović

Zadatak: **Usporedba metoda strojnog i dubokog učenja u rješavanju problema neuravnoteženosti klasa na tabličnim i slikovnim skupovima podataka**

Opis zadatka:

Algoritmi strojnog i dubokog učenja najčešće klasificiraju najtočnije kod približno jednakog broja primjeraka u svim klasama, no to u praksi nije uvijek slučaj. Neuravnoteženost klasa u skupovima podataka predstavlja veliki problem u područjima detekcije rijetkih događaja, poput medicinske dijagnoze i detekcije anomalija. U ovom diplomskom radu potrebno je implementirati i ispitati različite metode uravnoteživanja broja primjeraka po klasama na većem broju tabličnih skupova podataka. Metode ponovnog uzorkovanja potrebno je usporediti statističkim testovima radi pouzdanijeg vrednovanja. Također je potrebno ispitati neke metode za rješavanje istog problema na slikovnim skupovima podataka, uključujući različite funkcije gubitka te metode generiranja sintetskih primjeraka. U radu je potrebno koristiti javno dostupne skupove podataka dostupne na internetskim repozitorijima (npr. Kaggle, PhysioNet). Konačno, potrebno je grafički prikazati i analizirati sve dobivene rezultate te usporediti s rezultatima iz literature.

Rok za predaju rada: 26. lipnja 2026.

*Hvala mentoru prof. dr. sc. Alanu Joviću na lijepoj suradnji i vodstvu prilikom izrade ovog
diplomskog rada.*

Sadržaj

1. Uvod	3
2. Opis problema	4
2.1. Metode na tabličnim skupovima podataka	5
2.1.1. Poduzorkovanje	5
2.1.2. Naduzorkovanje	6
2.1.3. Hibridne metode	7
2.1.4. Statistički testovi	8
2.2. Metode na slikovnim skupovima podataka	8
2.2.1. Funkcije gubitka	8
2.2.2. Generativna suparnička mreža	9
3. Izrada rješenja	11
3.1. Tablični skupovi podataka	11
3.2. Slikovni skupovi podataka	13
4. Rezultati	16
4.1. Tablični skupovi podataka	16
4.2. Slikovni skupovi podataka	21
5. Zaključak	23
Literatura	25
Sažetak	27
Abstract	28

A: Pseudo kod algoritama	29
A1. Algoritam SMOTE	29
A2. Algoritam ADASYN	30
A3. Arhitektura modela DCGAN	31

1. Uvod

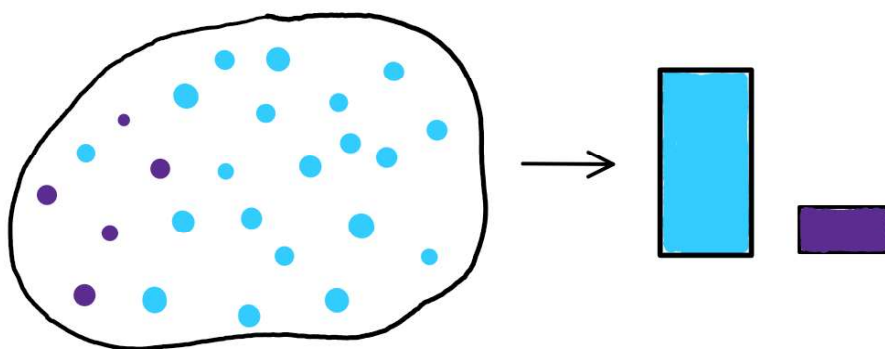
Modeli strojnog i dubokog učenja pretpostavljaju jednake udjele zastupljenosti razreda kod problema klasifikacije, odnosno balansiranost skupova podataka. Međutim to nije uvijek slučaj, neuravnoteženi podaci često se javljaju u problemima detekcije rijetkih događaja koji mogu pripadati u područje medicinske dijagnostike, detekcije transakcijskih prijevara, anomalija u industrijskoj proizvodnji, itd.

Cilj ovog diplomskog rada jest ispitati i usporediti različite metode rješavanja problema neuravnoteženosti razreda na tabličnim i slikovnim skupovima podataka. Pritom će se primjenjivati različite tehnike ponovnog uzorkovanja, raznolike funkcije gubitka te metode generiranja sintetskih primjeraka preko generativnih mreža.

Rad se sastoji od pet poglavlja nakon kojih slijedi popis literature, sažetak i privitak. U poglavljima je dan opis problema te pregled mogućih metoda rješavanja, zatim opis izrade rješenja, pregled dobivenih rezultata i potom je iznesen zaključak.

2. Opis problema

Problem nebalansiranosti podataka javlja se u skupovima podataka u kojima postoji neuravnoteženost između broja primjeraka razreda ciljne varijable. Nebalansiranost otežava izgradnju modela strojnog ili dubokog učenja koji će jednako dobro klasificirati primjerke manje zastupljenog razreda i više zastupljenog razreda. Na slici 2.1. demonstriran je problem neuravnoteženosti gdje je plavi razred značajno više zastupljen od ljubičastog, a međusobno su izmiješani primjerci.



Slika 2.1. Vizualizacija problema neuravnoteženosti

Algoritmi koji nemaju modifikacije za rad s nebalansiranim skupovima podataka često tijekom učenja dolaze u lokalni minimum optimizacijske funkcije u kojem sve ili gotovo sve primjerke svrstavaju u većinsku klasu. Time se postiže visoka točnost, međutim preciznost manjinskog razreda jako je mala ili jednaka nuli. Stoga je važno kod rada s ovakvim podacima gledati odgovarajuće mjere uspješnosti poput matrice konfuzije, preciznosti, odziva i F1-mjere koje daju više informacija o uspješnosti na pojedinom razredu klasifikacije.

Neki od mogućih pristupa rješavanju ovakvih problema uključuju tehnike ponovnog uzorkovanja, modifikacije u optimizacijskom postupku učenja ili čak korištenje modela

koji su robusniji na ovakve podatke.

Također je važno naglasiti da se svi postupci koji se provode nad nebalansiranim skupovima podataka izvode isključivo na skupu za učenje. U slučaju provedbe bilo kakvih balansiranja nad skupom na kojim se testira model, gubi se informacija o uspješnosti modela u stvarnim uvjetima, odnosno prirodnoj distribuciji razreda koje klasificiramo.

2.1. Metode na tabličnim skupovima podataka

Tablični skupovi podataka sastoje se od primjeraka zapisanih u redcima tablice pri čemu stupci predstavljaju različite značajke, odnosno atribute, tih primjeraka. Ako se takvi podaci koriste u svrhu klasifikacije, svaki primjerak sadrži i dodatni stupac s informacijom o pripadnosti razredu. U kontekstu ovog rada razmatraju se skupovi podataka koji imaju isključivo numeričke i kategorijske značajke. Glavna skupina metoda koja se koristi na ovakvim skupovima podataka jest ponovno uzorkovanje. Riječ je o tehnikama kojima je cilj izmijeniti razdiobu primjeraka ciljnog razreda tako da su razredi približno jednako zastupljeni u skupu podataka za učenje. Metode imaju tri različita pristupa: poduzorkovanje (engl. *undersampling*), naduzorkovanje (engl. *oversampling*) i hibridni pristupi.

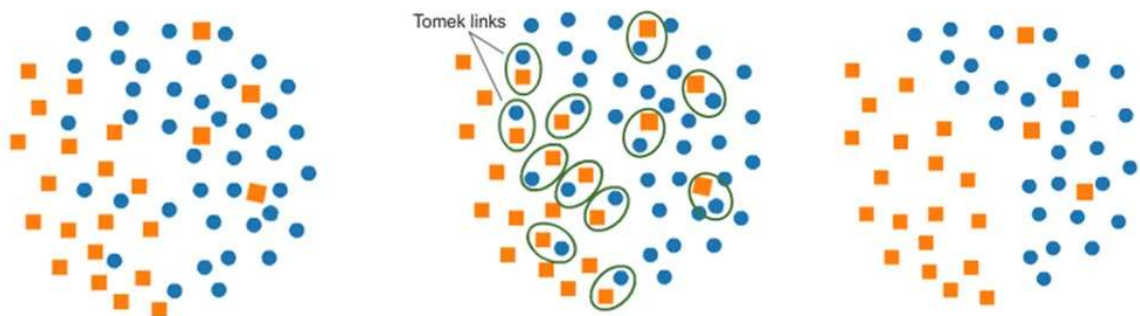
2.1.1. Poduzorkovanje

Poduzorkovanje podrazumijeva smanjenje broja primjeraka većinskog razreda njihovim uklanjanjem iz skupa podataka tako da se barem približno izjednači broju primjeraka manjinskog razreda. Glavna kritika metoda poduzorkovanja jest gubitak informacija iz početnog skupa podataka.

Jedna od primitivnijih metoda poduzorkovanja je slučajno poduzorkovanje u kojem se, kao što i ime daje naslutiti, slučajno odabrani primjerci većinskog razreda uklanjaju iz skupa podataka dok se ne postigne ravnoteža. Problem kod ove metode je potpuno slučajni gubitak informacije budući da ne postoji nikakav odabir primjeraka po korisnosti za učenje modela kao kod drugih metoda poduzorkovanja.

Metoda Tomekovih poveznica (engl. *Tomek links*, TL) uklanja primjerke većinskog razreda koji tvore Tomekovu poveznicu s nekim primjerkom manjinskog razreda. Dva primjerka tvore poveznicu ako su međusobno najbliži primjerci, to jest njihova udalje-

nost je manja od udaljenosti bilo kojeg od svih ostalih primjeraka u skupu. Cijeli postupak demonstrira slika 2.2.



Slika 2.2. Tomekove poveznice [1]

Metoda uređenih najbližih susjeda (engl. *Edited Nearest Neighbors*, ENN) uklanja primjerak većinskog razreda ako bi on kod korištenja kNN klasifikatora bio svrstan u manjinski razred. Odnosno, u slučaju da je većina od njegovih k najbližih susjeda manjinski razred, primjerak se uklanja, pri čemu je k hiperparametar algoritma.

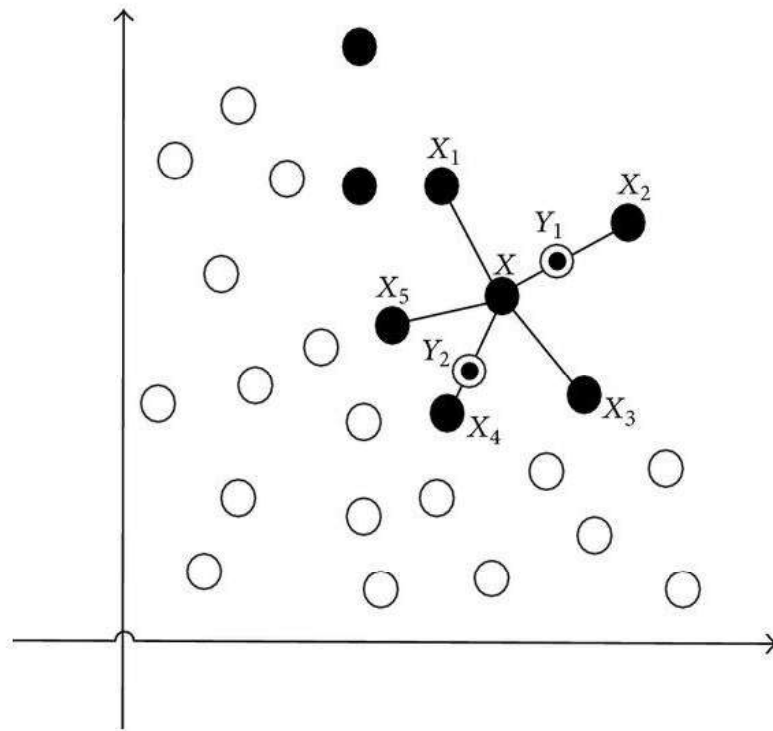
2.1.2. Naduzorkovanje

Naduzorkovanje podrazumijeva povećanje broja primjeraka manjinskog razreda generiranjem novih tako da se njihov broj izjednači s brojem primjera većinskog razreda. Prednost naduzorkovanja jest što ne dolazi do gubitka informacija, no moguće je povećanje šuma u podacima tijekom generiranja novih primjeraka iz postojećih.

Najprimitivnija metoda naduzorkovanja je slučajno naduzorkovanje s ponavljanjem u kojem se slučajnim odabirom u skup podataka dodaju duplikati primjeraka manjinskog razreda dok se ne postigne ravnoteža. Problem ovog pristupa je što će se modeli učeni na ovakvim podacima često prenaučiti zbog manjka unosa nove informacije u skup podataka.

Umjesto repliciranja primjeraka koriste se neke metode koje generiraju sintetske primjerke poput tehnike SMOTE (engl. *Synthetic Minority Over-sampling Technique*) [2]. Njome se stvaraju novi primjerci manjinskog razreda preko interpolacije između k najbližih susjeda. Na slici 2.3. primjer je stvaranja novih primjeraka Y_i za postojeći primjerak X uz 200 % naduzorkovanja i pretpostavku korištenja dvije značajke (2D-prostor). U

privitku dan je pseudokod cijelog postupka, algoritam A1. SMOTE se pokazao kao bolje rješenje od naduzorkovanja s ponavljanjem jer unosi novu informaciju, međutim, ako postoji šum ili pogrešno označeni primjerci, SMOTE pojačava taj šum stvaranjem mosta između ispravnih i stršećih primjeraka.



Slika 2.3. Primjer izračuna za algoritam SMOTE [3]

Metoda ADASYN (engl. *adaptive synthetic*) nastala je kao inačica tehnike SMOTE koja dinamički odlučuje koliko novih primjeraka generira za svakog postojećeg člana manjinskog razreda [4]. ADASYN za svaki primjerak slabije zastupljenog razreda, na temelju okoline k najbližih susjeda, određuje broj generiranih sintetskih primjeraka. Takvom tehnikom dolazi do povećanja gustoće primjeraka na granici između dva razreda. Pseudokod algoritma može se pronaći u privitku, algoritam A2.

2.1.3. Hibridne metode

Hibridni postupci kombiniraju tehnike naduzorkovanja i poduzorkovanja. U sklopu ovog rada razmatrat će se kombinacije tehnike SMOTE kao naduzorkovanja i Tomekovih poveznica ili algoritma ENN kao poduzorkovanja.

2.1.4. Statistički testovi

S obzirom na to da je jedan od ciljeva ovog rada ispitivanje ostvaruje li neka od ranije spomenutih metoda bolje rezultate od ostalih, nad dobivenim rezultatima bit će provedeni neki statistički testovi. Za to su prikladni neparametarski testovi jer ne zahtijevaju stroge pretpostavke o teorijskoj distribuciji podataka. Friedmanov test je neparametarski ekvivalent testa analize varijance za ponovljena mjerenja (ANOVA). Riječ je o testu koji ispituje postoji li statistički značajna razlika između rangova više povezanih uzoraka. Osim Friedmanovog testa, ovdje je primjenjiv i Wilcoxonov test rangova za povezane uzorke. To je neparametarski test koji se koristi za usporedbu dva povezana uzorka, u ovom slučaju prije i nakon primjene metoda za uravnotežavanje.

2.2. Metode na slikovnim skupovima podataka

Slikovni skupovi podataka sastoje se od digitalnih zapisa slika u nekom formatu te ako se koriste za klasifikaciju sadrže dodatnu datoteku koja kodira pripadnost razredima ili se labela odražavaju kroz strukturu direktorija. Kod slikovnih podataka, za razliku od tabličnih, koriste se isključivo duboki modeli zbog količine informacije koja proizlazi iz slika. Kao rješenje neuravnoteženosti stoga je moguće provoditi metode ponovnog uzorkovanja poput slučajnog naduzorkovanja, ali i dodavanja augmentacija na slike poput rotacija i zrcaljenja.

2.2.1. Funkcije gubitka

Osim modificiranja skupa podataka moguće je provesti izmjene u učenju algoritma korištenjem alternativnih funkcija gubitka. Takav pristup je dobar jer model prilikom učenja se manje fokusira na jednostavne primjerke, a teži primjerci, tipično iz manjinskog razreda, dobivaju veći utjecaj.

Umjesto korištenja uobičajenog gubitka unakrsne entropije može se koristiti težinska inačica. U tom slučaju, umjesto da se svaki primjerak kažnjava jednako kod krive klasifikacije, kazna se skalira ovisno o stupnju neuravnoteženosti skupa tako da se manje zastupljeni razred više kažnjava kod krive klasifikacije. U slučaju binarne klasifikacije koristi se formula (2.1) pri čemu je y točna oznaka primjerka, p vjerojatnosni izlaz modela, a w_0 i w_1 dodijeljene težine za svaki razred.

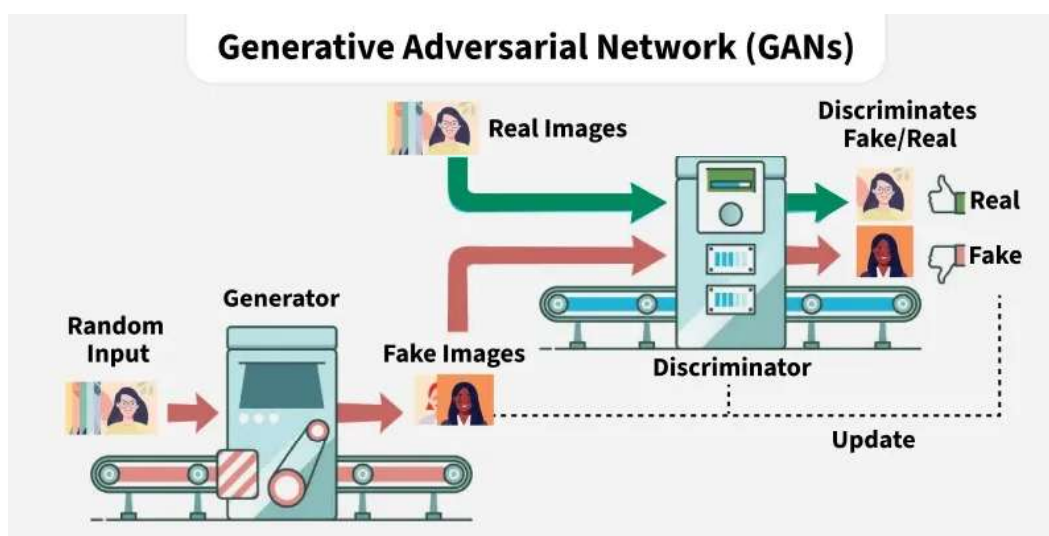
$$L = -(w_1 y \log(p) + w_0 (1 - y) \log(1 - p)) \quad (2.1)$$

Također je moguće koristiti žarišni gubitak (engl. *focal loss*) koji funkcionira na način da se fokusira na primjerke koje model klasificira krivo [5]. Umjesto da iteracijama povećava sigurnost u točno klasificirane primjerke, smanjuje težine takvih i koncentrira se na to da se poboljšaju predviđanja na neispravnim primjercima. Tipično se koristi varijanta žarišnog gubitka koja zadržava parametar o težini razreda i stoga istovremeno sadrži elemente fokusa na teškim primjerima i uravnoteživanja između razreda. U konačnici dobije se formula (2.2) gdje je α_t faktor težine za razred t , γ parametar fokusa i p_t vjerojatnosni izlaz modela za točni razred.

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (2.2)$$

2.2.2. Generativna suparnička mreža

Generativne suparničke mreže (engl. *Generative Adversarial Network*, GAN) vrsta su dubokih modela koje se koriste za generiranje novih podataka koji nalikuju podacima iz skupa za učenje. Arhitektura GAN-a sastoji se od dvije neuronske mreže, generatora (G) i diskriminatora (D) [6]. Generator je zadužen za generiranje novih slika, odnosno podataka, a diskriminator se uči raspoznavati sintetske i stvarne podatke. Cijeli tok prolaska slika kroz ova dva modela prikazan je na slici 2.4.



Slika 2.4. Arhitektura modela GAN [7]

Generator prima slučajni vektor šuma z iz unaprijed definiranog latentnog prostora te ga transformira s ciljem generiranja uvjerljivih sintetskih uzoraka. S druge strane, diskriminator djeluje kao klasifikator koji prima stvarne i lažne uzorke iz skupa podataka te procjenjuje vjerojatnost da je pojedini uzorak stvaran. Proces obučavanja temelji se na teoriji igara, preciznije na igri dvaju igrača s nultom sumom (igra *minimax*). Tijekom učenja, mreže se neprestano nadmeću: generator pokušava minimizirati sposobnost diskriminatora da prepozna lažne podatke (odnosno nastoji ga "prevariti"), dok diskriminator nastoji maksimizirati svoju točnost u razlikovanju stvarnih uzoraka od sintetskih. Ova kompetitivna dinamika postupno podiže kvalitetu generiranih podataka do točke konvergencije, u kojoj generator proizvodi visokokvalitetne uzorke koje diskriminator više ne može pouzdano razlikovati od stvarnih (teži vjerojatnosti od 0.5).

3. Izrada rješenja

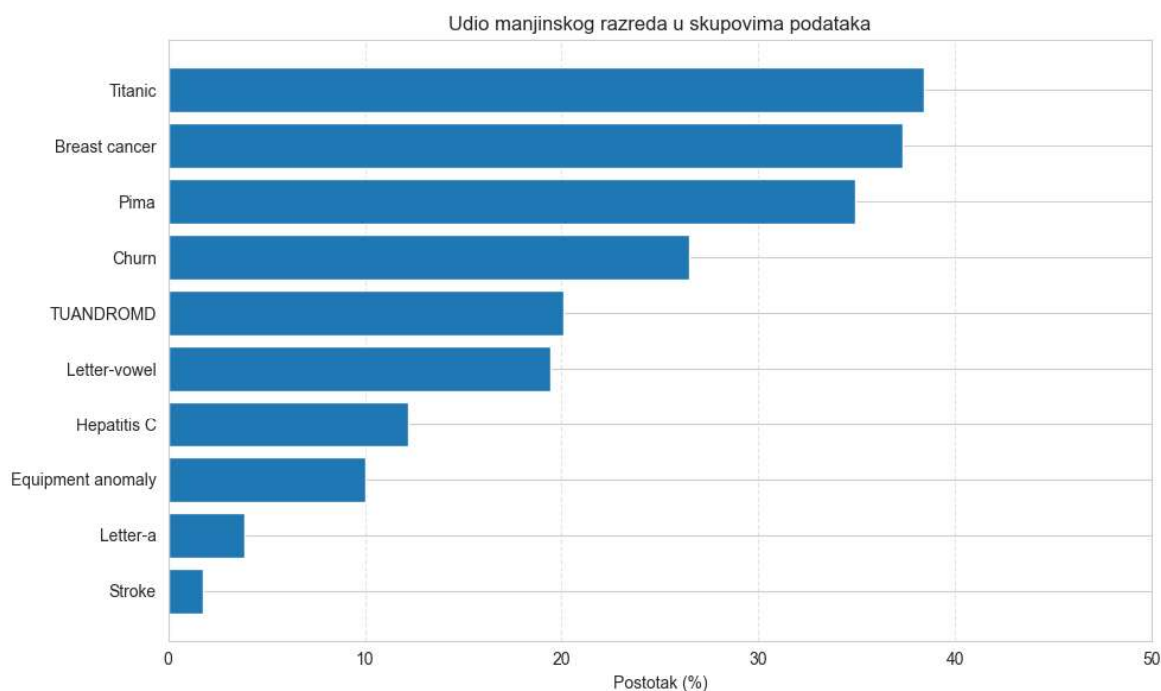
Svrha ovog rada bila je usporedba uspješnosti rada različitih metoda rješavanja problema nebalansiranosti, a u sklopu izrade rješenja korišten je programski jezik Python. Kod tabličnih skupova podataka korištene su biblioteke *pandas* za baratanje samim tablicama, *matplotlib* za iscrtavanje grafova, *sklearn* za algoritme strojnog učenja i mjere uspješnosti, *scipy.stats* za provođenje statističkih testova i *imblearn* za tehnike ponovnog uzorkovanja provedene nad neuravnoteženim podacima. Tijekom izrade rješenja za slikovne skupove podataka osim nekih ranije spomenutih korištene su i biblioteke *torch* i *torchvision* za rukovanje dubokim modelima te *numpy* za matrične operacije.

3.1. Tablični skupovi podataka

U prvom dijelu istraživanja postupak je proveden na ukupno deset skupova podataka koji su prikazani u tablici 3.1. s udjelima manjinskog razreda za svaki skup i brojem primjeraka u skupu. Kolekcija skupova podataka pokušava pokriti što više različitih stupnjeva neuravnoteženosti, a vizualizacija postotaka nalazi se na slici 3.1.

Tablica 3.1. Popis tabličnih skupova podataka

Skup podataka	Udio manjinskog razreda	Broj primjeraka
Stroke	1,8 %	43400
Letter-a	3,9 %	20000
Equipment anomaly	10,0 %	7672
Hepatitis C	12,2 %	615
Letter-vowel	19,4 %	20000
TUANDROMD	20,1 %	4465
Churn	26,5 %	7043
Pima	34,9 %	768
Breast cancer	37,3 %	569
Titanic	38,4 %	889



Slika 3.1. Postoci manjinskog razreda u skupovima podataka

Svi korišteni skupovi podataka bave se binarnom klasifikacijom uz iznimku skupa *letter* koji se sastoji od 26 razreda od kojih svaki predstavlja jedno slovo pa je podijeljen na skup slovo a i ostatak te skup samoglasnika i ostatak po uzoru na referentni rad [8]. Radi konzistentnosti u svim skupovima podataka stupac s ciljnom varijablom nazvan je *target*, a manjinski razred je pod oznakom 1 dok je većinski označen s 0. Na svakom skupu podataka provedena je uobičajena predobrada uključujući uklanjanje nedostajućih značajki, normalizaciju i kodiranje kategorijskih značajki. Zatim je provedena desetorostruka unakrsna validacija s algoritmima stabla odluke i logističke regresije za svaki od pristupa balansiraju navedenih u tablici 3.2.

Nakon učenja modela uspješnost je izmjerena F1-mjerom i PR-AUC-om s kojima su potom provedeni statistički testovi. Korištena je F1 izračunata na manjinskom razredu, a za izračun površine ispod krivulje korištena je krivulja preciznosti i odziva. Friedmanovim testom testirano je postoji li razlika uspješnosti neke metode u odnosu na sve ostale, a Wilcoxonovim testom provjereno je postoji li kod ikoje metode značajna statistička razlika između uspješnosti prije i nakon provedenog uravnotežavanja.

Tablica 3.2. Popis provedenih metoda ponovnog uzorkovanja

Bez modifikacija
Slučajno naduzorkovanje
SMOTE
ADASYN
Slučajno poduzorkovanje
Tomekove poveznice
ENN
SMOTE + Tomek
SMOTE + ENN

3.2. Slikovni skupovi podataka

U drugom dijelu istraživanja korišten je samo jedan skup podataka koji sadrži slike ispravnih i neispravnih vijaka. Ukupno je sadržano 480 slika, od čega su njih 119 vijci s greškom što čini 24,8 % cijelog skupa podataka. Primjeri jednog dobrog i jednog lošeg vijka nalaze se na slici 3.2. Na tim slikama učen je model ResNet18 s prednaučenim težinama skupa ImageNet i zamrznutim slojevima, a provedene su modifikacije na skupovima podataka za učenje i funkcijama gubitka. Testirani pristupi prikazani su tablicom 3.3., a za sve korištene metode slučajnim odabirom odvojeno je 20 % početnog skupa podataka u skup za testiranje uz zadržavanje distribucije između razreda.



(a) Ispravan vijak



(b) Neispravan vijak

Slika 3.2. Primjeri slika iz skupa podataka [9]

Tablica 3.3. Popis provedenih metoda na slikovnom skupu podataka

Uzorkovanje	Gubitak
Bez modifikacija	Unakrsna entropija
Slučajno naduzorkovanje	Unakrsna entropija
Augmentacije	Unakrsna entropija
Bez modifikacija	Težinska unakrsna entropija
Augmentacije	Težinska unakrsna entropija
Bez modifikacija	Žarišni
Augmentacije	Žarišni
GAN	Unakrsna entropija

U testiranjima s augmentacijama provedene su sljedeće modifikacije:

- Slučajno horizontalno zrcaljenje, $p = 50 \%$
- Slučajno vertikalno zrcaljenje, $p = 50 \%$
- Slučajna rotacija, uniformna razdioba na $[-180^\circ, 180^\circ]$

Kod računanja težinske unakrsne entropije korištene su težine 1.0 za većinski razred i 3.0 za manjinski razred budući da je približno tri puta više ispravnih vijaka u odnosu na neispravne. Za žarišni gubitak korištene su iste te težine za parametar α , a za parametar γ korištene su vrijednosti 2.0 i 1.0. Zbog konzistentnosti kod usporedbe metoda, svi modeli ućeni su na 10 epoha što je procijenjeno kao dovoljno vremena da model dođe u neko stabilno stanje, uz iznimku modela sa slikama generiranim modelom GAN koji je ućen na 20 epoha. Također, konaćno spremljeno stanje modela je ono u kojem je postignuta najveća vrijednost mjere F1 na manjinskom razredu.

Model GAN strukturiran je kao duboka konvolucijska generativna suparnićka mreža (engl. *Deep Convolutional GAN*, DCGAN). Generator prima latentni vektor dimenzije $z = 100$ i zatim kroz niz slojeva postupno povećava prostorne dimenzije do konaćnog formata slike 64×64 piksela s tri kanala boja (RGB). Pritom se koriste serijske normalizacije po mini-grupama i aktivacijske funkcije ReLU osim na izlaznom sloju gdje se koristi tangens hiperbolni ($\tanh$). Diskriminator je binarni klasifikator koji prima sliku i reducira dimenzije preko konvolucija do konaćne vjerojatnosti dobivene korištenjem sigmoidne aktivacijske funkcije. Arhitektura modela kreirana je u skladu s mnogobrojnim smjernicama iz literature, a konaćni model napravljen je po uzoru na referentni rad

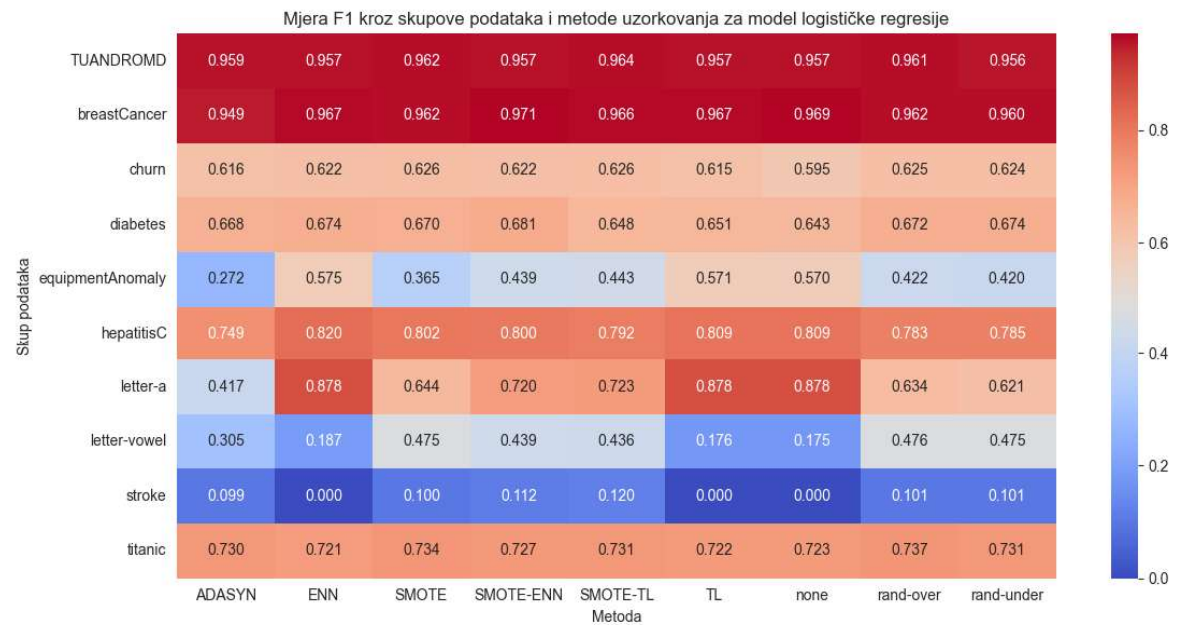
[10]. Detaljniji tablični prikaz cjelokupne arhitekture nalazi se u privitku, tablica A3. Učenje ovog modela provedeno je isključivo na skupu neispravnih vijaka korištenih za učenje. Prije učenja klasifikacijskog modela naučenim GAN-om generirano je 50 slika, čime je omjer pravih i sintetskih slika vijaka dostigao omjer od približno 2:1.

4. Rezultati

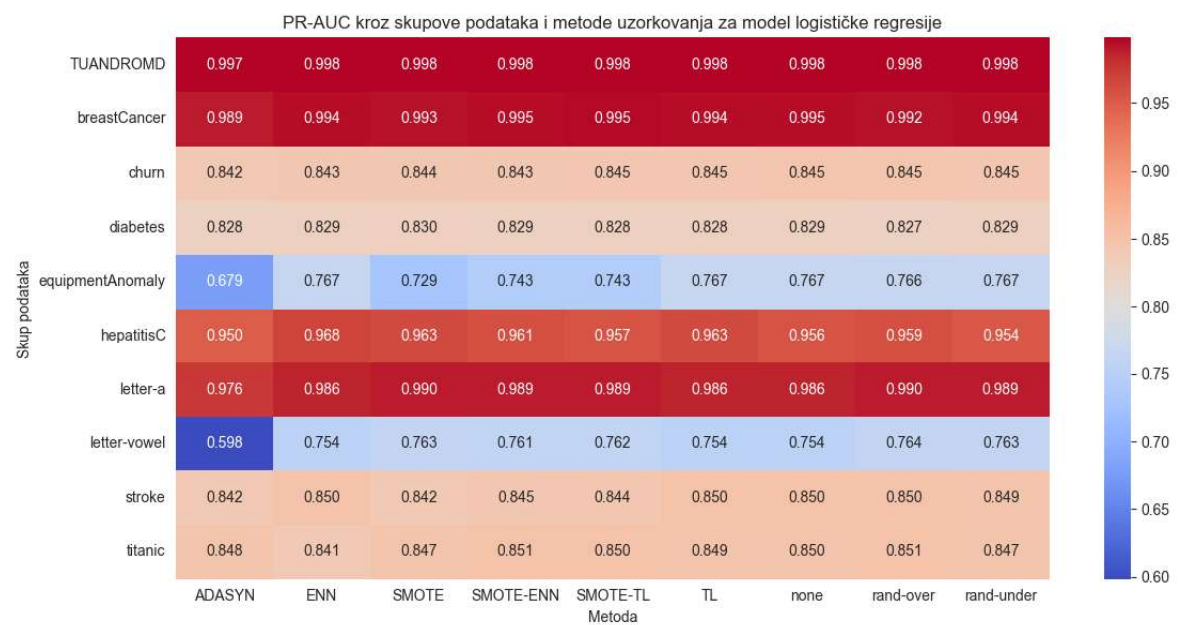
4.1. Tablični skupovi podataka

Sveukupni rezultati na svim skupovima podataka, za sve metode ponovnog uzorkovanja za model logističke regresije prikazani su na slici 4.1. Kod modela stabala odluke vidljive su puno manje razlike u uspješnosti između metoda ponovnog uzorkovanja, a rezultati su demonstrirani na slici 4.2. Gornja slika prikazuje uspješnost mjerenu F1 na isključivo manjinskom razredu, a donja slika prikazuje mjeru PR-AUC. Teško je procijeniti iz ovakvih prikaza radi li neka od metoda bolje na svim podacima. Na primjer, ENN i TL na skupu *equipmentAnomaly* malo poboljšavaju mjeru F1, ali te iste dvije metode rade lošije od svih ostalih na skupu *stroke*. Isto tako, na nekim skupovima podataka učinak svih metoda poprilično ujednačen, a intervencije balansiranja ne donose značajne pomake.

Ovakve varijacije u uspješnosti sugeriraju da uspješnost metode jako ovisi o samom skupu podataka i njegovoj inicijalnoj neuravnoteženosti, ali i korištenom modelu strojnog učenja. Budući da iz ovakvih prikaza nije moguće donijeti globalne zaključke o superiornosti nekih strategija, u nastavku su provedeni statistički testovi kako bi se utvrdilo postoji li statistički značajna razlika između korištenih metoda.

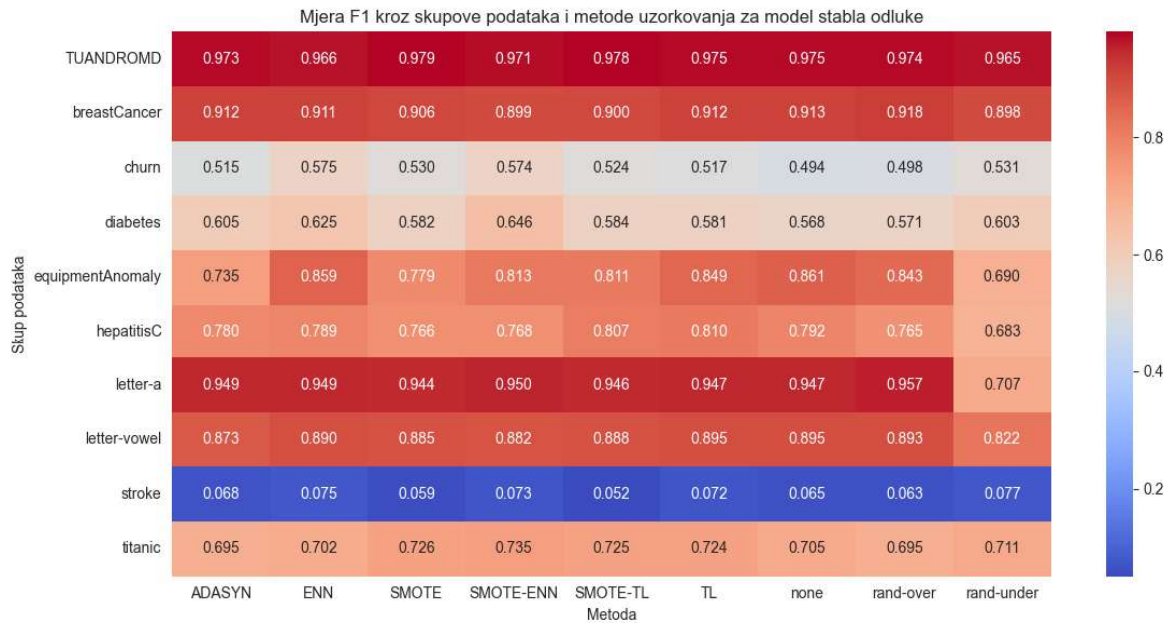


(a) F1



(b) PR-AUC

Slika 4.1. Rezultati metoda ponovnog uzorkovanja za logističku regresiju



(a) F1



(b) PR-AUC

Slika 4.2. Rezultati metoda ponovnog uzorkovanja za stablo odluke

Rezultati Friedmanovog testa prikazani su tablicom 4.1. u kojoj je vidljivo da se samo u kombinaciji logističke regresije i mjere PR-AUC može sa statističkom značajnošću odbaciti nulta hipoteza, odnosno ustvrditi da postoji značajna razlika u rangovima između metoda. Logistička regresija ocijenjena mjerom F1 također pokazuje mogućnost postojanja razlike, međutim p-vrijednost je na 11,76 % naspram standardnog praga od 10 %

koji je potrebno zadovoljiti da bi nulta hipoteza bila odbačena.

Tablica 4.1. Rezultati Friedmanovog testa

Algoritam	Mjera	p-vrijednost
Stablo odluke	F1	0.3266
Stablo odluke	PR-AUC	0.4695
Logistička regresija	F1	0.1176
Logistička regresija	PR-AUC	0.0060

Wilcoxonov test rangova za povezane uzorke na mjeri F1 dao je rezultate prikazane tablicom 4.2. U tablici je označeno koje metode odbacuju nultu hipotezu na razini značajnosti 10 %. Na algoritmu stabla odluke odbacuje se jedino za Tomekove poveznice, a na modelu logističke regresije za metodu ENN. Na ostalim metodama nije utvrđena statistički značajna razlika u odnosu na rezultate dobivene bez primjene ponovnog uzorkovanja. Zanimljivo je primijetiti da su obje metode za koje je utvrđena statistički značajna razlika metode poduzorkovanja, što pokazuje da uklanjanje izabranih primjeraka većinskog razreda ima veći utjecaj na vrijednost mjere F1 od generiranja novih primjeraka.

Tablica 4.2. Rezultati Wilcoxonovog testa na mjeri F1

Model	Metoda uzorkovanja	p-vrijednost	Odbacuje H_0
Stablo odluke	ADASYN	0.6953	Ne
	ENN	1.0000	Ne
	SMOTE	0.7695	Ne
	SMOTE + ENN	0.8457	Ne
	SMOTE + Tomek	0.6953	Ne
	Tomekove poveznice	0.0929	Da
	Slučajno nadzuzorkovanje	0.6953	Ne
	Slučajno poduzorkovanje	0.2324	Ne
Logistička regresija	ADASYN	1.0000	Ne
	ENN	0.0663	Da
	SMOTE	0.6250	Ne
	SMOTE + ENN	0.5566	Ne
	SMOTE + Tomek	0.6953	Ne
	Tomekove poveznice	0.3105	Ne
	Slučajno naduzorkovanje	0.6953	Ne
	Slučajno poduzorkovanje	0.8457	Ne

Wilcoxonov test rangova za povezane uzorke na mjeri PR-AUC dao je rezultate prikazane u tablici 4.3. Za model stabla odluke jedina metoda koja dovodi do odbacivanja nulte hipoteze na razini značajnosti od 10 % su Tomekove poveznice, što je isti obrazac

kao i kod mjere F1. Nasuprot tome, kod logističke regresije većina metoda uzorkovanja dovodi do statistički značajne promjene PR-AUC vrijednosti. Nulta hipoteza odbacuje se za metode ADASYN, SMOTE, SMOTE + ENN, SMOTE + Tomek, slučajno naduzorkovanje i slučajno poduzorkovanje. Jedine metode koje ne pokazuju statistički značajnu razliku u odnosu na izvorni skup podataka su ENN i Tomekove poveznice.

Tablica 4.3. Rezultati Wilcoxonovog testa na mjeri PR-AUC

Model	Metoda uzorkovanja	p-vrijednost	Odbacuje H_0
Stablo odluke	ADASYN	0.9219	Ne
	ENN	1.0000	Ne
	SMOTE	0.7695	Ne
	SMOTE + ENN	0.9219	Ne
	SMOTE + Tomek	1.0000	Ne
	Tomekove poveznice	0.0929	Da
	Slučajno naduzorkovanje	1.0000	Ne
	Slučajno poduzorkovanje	0.4922	Ne
Logistička regresija	ADASYN	0.0195	Da
	ENN	0.3750	Ne
	SMOTE	0.0840	Da
	SMOTE + ENN	0.0195	Da
	SMOTE + Tomek	0.0371	Da
	Tomekove poveznice	0.9528	Ne
	Slučajno naduzorkovanje	0.0488	Da
	Slučajno poduzorkovanje	0.0020	Da

Rezultati upućuju na to da je logistička regresija osjetljivija na promjene distribucije razreda uzrokovane metodama ponovnog uzorkovanja od stabla odluke. Dok kod stabla odluke većina metoda ne dovodi do statistički značajnih promjena performansi, kod logističke regresije postupci uzorkovanja često rezultiraju značajnim promjenama PR-AUC vrijednosti. Posebno je zanimljivo da su značajne razlike prisutne i kod metoda naduzorkovanja i kod metoda poduzorkovanja, što sugerira da je za ovaj model ravnoteža razreda važan čimbenik pri izgradnji klasifikatora. Međutim, kod mjere F1 nije toliko vidljiva razlika te stoga isplativost provedbe postupaka ponovnog uzorkovanja ovisi o tome koji su prioriteti kod performansi klasifikacijskog modela.

4.2. Slikovni skupovi podataka

Rezultati svih 9 modela akumulirani su tablicom 4.4., pri čemu se preciznost, odziv i F1-mjera odnose isključivo na manjinski razred, a točnost obuhvaća cijeli testni skup podataka.

Tablica 4.4. Rezultati eksperimenata na testnom skupu

Uzorkovanje	Gubitak	Preciznost	Odziv	F1-mjera	Točnost
Bez modifikacija	BCE	0.75	0.75	0.75	0.88
Slučajno naduzorkovanje	BCE	0.58	0.75	0.65	0.80
Augmentacija	BCE	0.80	0.50	0.62	0.84
Bez modifikacija	Težinski BCE	0.69	0.83	0.75	0.86
Augmentacija	Težinski BCE	0.42	0.88	0.57	0.67
Bez modifikacija	Žarišni, $\gamma = 1$	0.43	0.83	0.57	0.69
Bez modifikacija	Žarišni, $\gamma = 2$	0.43	0.88	0.58	0.68
Augmentacija	Žarišni, $\gamma = 2$	0.27	1.0	0.43	0.33
GAN	BCE	0.93	0.58	0.72	0.89

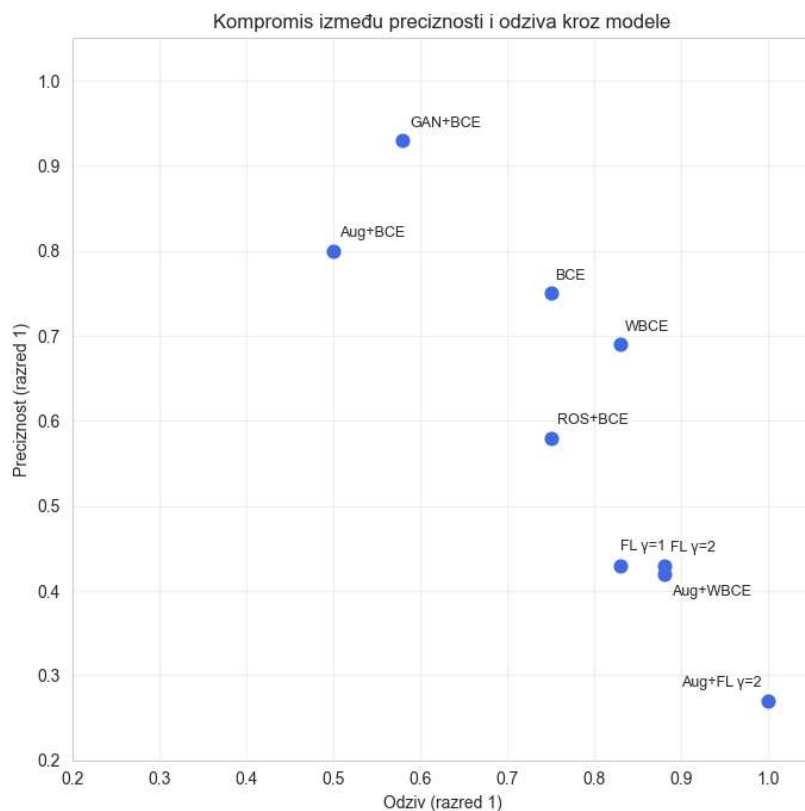
Iz rezultata je vidljiv kompromis između preciznosti i odziva koji dolazi na izražaj korištenjem različitih tehnika uravnoteživanja podataka. Prilikom analize, nužno je uzeti u obzir da su konačne težine modela one gdje je postignuta najbolja vrijednost mjere F1 i stoga je moguće da postoji neka pristranost. Drugim riječima, ako je cilj klasifikacijskog modela maksimizirati neku drugu mjeru moguće je da bi rezultati bili malo drugačiji.

Dodavanje faktora težine u funkciju gubitka poboljšava osjetljivost modela na manjinski razred što je vidljivo na vrijednostima odziva te na slici 4.3. Žarišni gubitak u ovom slučaju evidentno preveliki fokus stavlja na manjinski razred čime dolazi do drastičnog pada preciznosti unatoč jako dobrom odzivu. S druge strane, model koji koristi sintetske primjerke generirane GAN-om pokazuje obrnute trendove. Ovim modelom postignuta je najbolja preciznost uz nešto lošiji odziv, odnosno smanjen je broj lažno pozitivnih primjera, a povećan broj lažno negativnih.



Slika 4.3. Rezultati preciznosti, odziva i F1 na svim metodama

Cijeli koncept kompromisa između preciznosti i odziva jako je dobro vidljiv na grafu 4.4. Dakle, odgovor na pitanje isplati li se koristiti metode za uravnotežavanje podataka jest da ovisi o tome koju mjeru je cilj maksimizirati. Pretpostavka je da u slučajevima detekcije rijetkih događaja želja poboljšati odziv, odnosno imati što manje lažno negativnih primjeraka, no to možda ne vrijedi za svaki problem.



Slika 4.4. Krivulja preciznosti i odziva

5. Zaključak

Problem neuravnoteženosti razreda jedan je od čestih izazova u primjeni metoda strojnog i dubokog učenja jer može dovesti do pristranosti modela prema većinskom razredu i smanjenja sposobnosti prepoznavanja rijetkih događaja. U ovom radu analizirane su različite metode rješavanja tog problema na tabličnim i slikovnim skupovima podataka.

Na tabličnim skupovima podataka uspoređene su metode naduzorkovanja i poduzorkovanja korištenjem modela stabla odluke i logističke regresije. Rezultati su pokazali da ne postoji univerzalno najbolja metoda ponovnog uzorkovanja te da njezina uspješnost ovisi o karakteristikama skupa podataka i odabranom klasifikacijskom modelu. Friedmanov test pokazao je statistički značajne razlike među metodama jedino za logističku regresiju promatranu kroz PR-AUC mjeru. Wilcoxonov test dodatno je pokazao da pojedine metode mogu značajno utjecati na performanse modela, pri čemu je logistička regresija osjetljivija na promjene distribucije razreda od stabla odluke. Dobiveni rezultati potvrđuju da primjena metoda ponovnog uzorkovanja može biti korisna, ali da njihov učinak nije moguće unaprijed generalizirati na sve probleme.

Na slikovnom skupu podataka analizirani su utjecaji različitih funkcija gubitka, augmentacije podataka i generiranja novih primjeraka uz pomoć generativnih suparničkih mreža. Pokazano je da metode uravnoteživanja podataka utječu na kompromis između preciznosti i odziva. Uvođenje težinskog i žarišnog gubitka povećalo je odziv modela, odnosno sposobnost prepoznavanja manjinskog razreda, ali je istovremeno dovelo do smanjenja preciznosti. S druge strane, korištenje sintetskih slika generiranih modelom GAN rezultiralo je najvećom preciznošću među promatranim pristupima, uz niži odziv. Time je potvrđeno da različite metode uravnoteživanja mogu biti prikladne za različite ciljeve klasifikacije.

Glavni zaključak rada jest da ne postoji jedinstveno rješenje problema neuravnoteženosti podataka. Odabir odgovarajuće metode ovisi o prirodi podataka, korištenom modelu i kriterijima uspješnosti koji su važni u konkretnoj primjeni.

Literatura

- [1] L. Guadagnolo, “Imbalanced Data: an Extensive Guide on How to Deal with Imbalanced Classification Problems”, <https://medium.com/eni-digitalks/imbalanced-data-an-extensive-guide-on-how-to-deal-with-imbalanced-classification-problems-6c8df0bc2cab>, 2020., [posjećeno: 16. lipnja 2026.].
- [2] N. V. Chawla, K. W. Bowyer, L. O. Hall, i W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique”, *Journal of Artificial Intelligence Research*, sv. 16, str. 321–357, 2002. <https://doi.org/10.1613/jair.953>
- [3] Y. Xie, Y. Liu, i Q. Fu, “Imbalanced Data Sets Classification Based on SVM for Sand-Dust Storm Warning”, *Discrete Dynamics in Nature and Society*, 2015. <https://doi.org/10.1155/2015/562724>
- [4] H. He, Y. Bai, E. A. Garcia, i S. Li, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning”, u *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2008., str. 1322–1328. <https://doi.org/10.1109/IJCNN.2008.4633969>
- [5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, i P. Dollár, “Focal Loss for Dense Object Detection”, u *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017., str. 2980–2988.
- [6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, i Y. Bengio, “Generative Adversarial Networks”, *Communications of the ACM*, sv. 63, br. 11, str. 139–144, November 2020. <https://doi.org/10.1145/3422622>

- [7] GeeksforGeeks, “Generative Adversarial Network (GAN)”, <https://www.geeksforgeeks.org/deep-learning/generative-adversarial-network-gan/>, [posjećeno: 13. lipnja 2026.].
- [8] G. E. A. P. A. Batista, R. C. Prati, i M. C. Monard, “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data”, *SIGKDD Explorations*, sv. 6, br. 1, str. 20–29, 2004. <https://doi.org/10.1145/1007730.1007735>
- [9] P. Bergmann, M. Fauser, D. Sattlegger, i C. Steger, “MVTec AD: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection”, u *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019., str. 9584–9592. <https://doi.org/10.1109/CVPR.2019.00982>
- [10] A. Radford, L. Metz, i S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”, 2016. [Mrežno]. Adresa: <https://arxiv.org/abs/1511.06434>

Sažetak

Usporedba metoda strojnog i dubokog učenja u rješavanju problema neuravnoteženosti klasa na tabličnim i slikovnim skupovima podataka

Andrea Milanović

Rad prikazuje neke od metoda rješavanja problema neuravnoteženosti skupova podataka te njihovu usporedbu. Korištene su različite metode ponovnog uzorkovanja na tabličnim skupovima podataka. Na slikovnim skupovima podataka korištene su različite funkcije gubitaka te neke metode augmentacije i generiranja novih primjeraka generativnim suparničkim mrežama. Analizirana je uspješnost svih metoda, a na uspješnosti ponovnog uzorkovanja provedeni su i statistički testovi, Friedmanov test i Wilcoxonov test rangova za povezane uzorke.

Ključne riječi: neuravnoteženi skupovi podataka, ponovno uzorkovanje, žarišni gubitak, generativne suparničke mreže

Abstract

Comparison of Machine and Deep Learning Methods in Solving the Class Imbalance Problem on Tabular and Image Datasets

Andrea Milanović

This thesis presents several methods for addressing the problem of imbalanced datasets and their comparison. Different resampling techniques were applied to tabular datasets. For image datasets, various loss functions were used, along with several augmentation methods and approaches for generating new samples using generative adversarial networks. The performance of all methods was analyzed, and statistical tests, namely the Friedman test and the Wilcoxon signed-rank test, were also conducted to evaluate the resampling methods.

Keywords: imbalanced datasets, resampling, focal loss, generative adversarial networks

Privitak A: Pseudo kod algoritama

U nastavku je dan pseudokod algoritama SMOTE i ADASYN te tablični prikaz korištenog modela GAN.

A1. Algoritam SMOTE

Algoritam 1 SMOTE

Ulaz: Skup manjinskog razreda S_{min} , razina naduzorkovanja N , broj susjeda k

Izlaz: Prošireni balansirani skup podataka

- 1: **Za sve** $\mathbf{x}_i \in S_{min}$ **radi**
- 2: Pronađi k najbližih susjeda iz manjinske klase
- 3: Odredi broj sintetičkih uzoraka n_i koje treba generirati
- 4: **Za** $j = 1$ do n_i **radi**
- 5: Nasumično odaberi susjeda $\mathbf{x}_{nn}$
- 6: Generiraj slučajni broj $\lambda \sim U(0, 1)$
- 7: Kreiraj novi sintetički uzorak:

$$\mathbf{x}_{novi} = \mathbf{x}_i + \lambda (\mathbf{x}_{nn} - \mathbf{x}_i)$$

- 8: Dodaj $\mathbf{x}_{novi}$ u skup podataka
 - 9: **Kraj**
 - 10: **Kraj**
 - 11: **Vrati** Prošireni skup podataka
-

A2. Algoritam ADASYN

Algoritam 2 ADASYN

Ulaz: Skup manjinskog razreda S_{min} , skup većinskog razreda S_{maj} , broj susjeda k

Izlaz: Prošireni balansirani skup podataka

1: Izračunaj stupanj neuravnoteženosti:

$$d = \frac{m_s}{m_l}$$

2: **Ako** $d < d_{th}$ **onda**

3: Izračunaj ukupan broj sintetskih uzoraka:

$$G = (m_l - m_s) \cdot \beta$$

4: **Za sve** $\mathbf{x}_i \in S_{min}$ **radi**

5: Pronađi k najbližih susjeda

6: Odredi broj susjeda iz većinskog razreda Δ_i

7: Izračunaj

$$r_i = \frac{\Delta_i}{k}$$

8: **Kraj**

9: Normaliziraj sve vrijednosti r_i u r'_i

10: **Za sve** $\mathbf{x}_i \in S_{min}$ **radi**

11: Izračunaj broj sintetskih uzoraka:

$$g_i = r'_i \cdot G$$

12: **Za** $j = 1$ do g_i **radi**

13: Nasumično odaberi manjinskog susjeda $\mathbf{x}_{zi}$

14: Generiraj slučajni broj $\lambda \sim U(0, 1)$

15: Stvori sintetski uzorak:

$$\mathbf{s}_i = \mathbf{x}_i + (\mathbf{x}_{zi} - \mathbf{x}_i)\lambda$$

16: Dodaj $\mathbf{s}_i$ u skup podataka

17: **Kraj**

18: **Kraj**

19: **Kraj**

20: **Vrati** Prošireni skup podataka

A3. Arhitektura modela DCGAN

Tablica A1. Detaljna arhitektura generatora i diskriminatora unutar modela DCGAN

Sloj	Tip sloja / Parametri	Izlazna dimenzija	Aktivacija
Generator			
Ulaz	Latentni vektor z	$100 \times 1 \times 1$	–
1. sloj	ConvTranspose2d (kernel=4, stride=1, pad=0) + BatchNorm	$512 \times 4 \times 4$	ReLU
2. sloj	ConvTranspose2d (kernel=4, stride=2, pad=1) + BatchNorm	$256 \times 8 \times 8$	ReLU
3. sloj	ConvTranspose2d (kernel=4, stride=2, pad=1) + BatchNorm	$128 \times 16 \times 16$	ReLU
4. sloj	ConvTranspose2d (kernel=4, stride=2, pad=1) + BatchNorm	$64 \times 32 \times 32$	ReLU
Izlaz	ConvTranspose2d (kernel=4, stride=2, pad=1)	$3 \times 64 \times 64$	Tanh
Diskriminator			
Ulaz	Slika (stvarna ili generirana)	$3 \times 64 \times 64$	–
1. sloj	Conv2d (kernel=4, stride=2, pad=1)	$64 \times 32 \times 32$	LeakyReLU (0.2)
2. sloj	Conv2d (kernel=4, stride=2, pad=1) + BatchNorm	$128 \times 16 \times 16$	LeakyReLU (0.2)
3. sloj	Conv2d (kernel=4, stride=2, pad=1) + BatchNorm	$256 \times 8 \times 8$	LeakyReLU (0.2)
4. sloj	Conv2d (kernel=4, stride=2, pad=1) + BatchNorm	$512 \times 4 \times 4$	LeakyReLU (0.2)
Izlaz	Conv2d (kernel=4, stride=1, pad=0)	$1 \times 1 \times 1$	Sigmoid