

**58. SREČANJE MLADIH RAZISKOVALCEV SLOVENIJE 2024**

**KLASIFIKACIJA GALAKSIJ Z UPORABO STROJNEGA  
UČENJA**

računalništvo ali telekomunikacije  
Raziskovalna naloga

Šola: Srednja elektro-računalniška šola Maribor

Avtor: Maj Korent

Mentor: Nataša Petelin

Šolsko leto: 2023/24

**Maribor, 2024**

## KAZALO

1.	POVZETEK .....	5
2.	ABSTRACT .....	5
3.	ZAHVALA .....	6
4.	UVOD .....	8
5.	HIPOTEZE .....	9
6.	TEORETIČNI DEL .....	10
6.1.	Strojno učenje .....	10
6.2.	Globoko učenje .....	11
6.3.	Nevronska mreža .....	12
6.4.	Konvolucijska nevrnska mreža .....	12
6.5.	Nevron .....	13
6.6.	Perceptron .....	14
6.7.	Množice .....	14
6.7.1.	Učna množica .....	14
6.7.2.	Validacijska množica .....	14
6.7.3.	Bias .....	15
6.7.4.	Testna množica .....	15
6.8.	Galaksija .....	15
6.9.	Morfologija galaksij .....	16
6.10.	Klasifikacija galaksij .....	18
7.	PRAKTIČNI DEL .....	20
7.1.	Metodologija .....	20
7.1.1.	Razvojno okolje in orodja .....	20
7.1.2.	Uporabljene knjižnice in pristopi .....	20
7.1.3.	Podatkovni viri in shranjevanje .....	21

7.2.	Metodološki pristop.....	21
7.2.1.	Instalacija knjižnic.....	22
7.2.2.	Prenos in priprava podatkovnega nabora .....	23
7.2.3.	Obdelava in shranjevanje podatkov .....	23
7.2.4.	Strukturiranje podatkov in organizacija .....	23
7.2.5.	Detajlna struktura podmap .....	24
7.2.6.	Kako je sestavljen podatkovni nabor .....	24
7.2.7.	Proces filtriranja podatkov .....	26
7.2.8.	Dodatna optimizacija podatkovne datoteke: .....	27
7.2.9.	Kako bomo dostopali do datotek.....	27
7.3.	Povezovanje podatkov iz CSV datoteke z bazo slik.....	28
7.3.1.	Izdelava podatkovnega nabora .....	29
7.3.2.	Povezovanje podatkov.....	29
7.4.	Učenje nevronske mreže .....	31
7.4.1.	Kako smo nevronske mreže učili.....	36
8.	REZULTATI.....	38
8.1.	Kaj se zgodi če model ne more prepoznati galaksije .....	40
9.	INTERPRETACIJA REZULTATOV .....	41
9.1.	Interpretacija rezultatov: .....	41
9.2.	Hipoteze .....	41
9.3.	Zanimivosti.....	41
10.	ZAKLJUČEK .....	43
11.	DRUŽBENA ODGOVORNOST .....	44
12.	VIRI IN LITERATURA .....	45
12.1.	Knjige in članki .....	45
12.2.	Spletni viri .....	45

12.3. Viri slik.....	47
13. PRILOGE .....	48

## KAZALO SLIK

Slika 1 Morfologija galaksij.....	17
Slika 2 Podatkovna zbirka shranjena v Google Drive.....	24
Slika 3 Primer a      Slika 4 Primer b      Slika 5 Primer c.....	26
Slika 6 Primer d      Slika 7 Primer e      Slika 8 Primer f.....	26
Slika 9 Primerjava velikosti CSV datotek.....	26
Slika 10 Razlika med originalno in procesirano sliko.....	36

## KAZALO GRAFIKONOV

Graf 1 Kako se je NM učila .....	39
----------------------------------	----

## KAZALO ENAČB

Enačba 1 Izračun nevrona .....	13
Enačba 2 Izračun uteži perceptrona .....	14
Enačba 3 Izračun konvolucijskega sloja .....	31
Enačba 4 Funkcija ReLU .....	34
Enačba 5 Enačba plasti za združevanje.....	34

## KAZALO IZZSEKOV KODE

Izrezek kode 1 Uporabljene knjižnice.....	22
Izrezek kode 2 Povezovanje podatkov iz csv in učnega nabora .....	29
Izrezek kode 3 Izdelava končnega podatkovnega nabora .....	30
Izrezek kode 4 CNN nastavitve.....	33

## **1. POVZETEK**

Raziskava se osredotoča na izboljšanje klasifikacije galaksij z uporabo strojnega učenja, cilj je razviti natančen model za morfološko klasifikacijo galaksij s podatkovne zbirke Galaxy Zoo DECaLS. Predstavljeni so postopki razvoja, testiranja modela in primerjava z obstoječimi metodami, pri čemer se poudarja pomen avtomatizacije v astronomiji zaradi nenehnega naraščanja števila odkritih galaksij. Rezultati kažejo na visoko učinkovitost modela v natančnosti klasifikacije, kar pomeni pomemben korak naprej v raziskavah vesolja. Raziskava poleg tehničnih podrobnosti obravnava tudi potencialne aplikacije in vpliv na prihodnje astronomske študije, s čimer prispeva k boljšemu razumevanju vesoljskih struktur in procesov.

Ključne besede: strojno učenje, klasifikacija galaksij, konvolucijske nevronske mreže, procesiranje in analiza podatkov, računalniški vid.

## **2. ABSTRACT**

The research focuses on improving the classification of galaxies using machine learning, the goal is to develop an accurate model for the morphological classification of galaxies from the Galaxy Zoo DECaLS database. The development processes, model testing, and comparison with existing methods are presented, emphasizing the importance of automation in astronomy due to the continuous increase in the number of discovered galaxies. The results show the high efficiency of the model in classification accuracy, which represents an important step forward in space research. In addition to technical details, the research also addresses potential applications and impact on future astronomical studies, thereby contributing to a better understanding of space structures and processes.

Keywords: machine learning, galaxy classification, convolutional neural networks, data processing and analysis, computer vision.

### **3. ZAHVALA**

Iskreno se želimo zahvaliti naši mentorici, katere neprestana podpora, strokovno vodenje in dragocene pripombe so nam bile v veliko pomoč pri raziskovanju. Rad bi se zahvalili tudi vsem ostalim za nenehno spodbudo.

## Seznam kratic

ML – strojno učenje (ang. Machine learning)

AI – Umetna inteligenca (ang. Artificial intelligence)

NM – Nevronska mreža

UM – Učna množica

CNN - konvolucijska nevrnska mreža (ang. convolutional neural network)

#### 4. UVOD

Raziskovanje nebesnih teles, predvsem galaksij, predstavlja eno izmed ključnih področij sodobne astronomije. Naše zanimanje za to tematiko se je pričelo med opazovanjem nočnega neba skozi teleskop, kjer smo se soočili z osupljivo raznolikostjo galaktičnih struktur. To nas je spodbudilo k razmisleku o metodah, ki jih znanstveniki uporabljajo za klasifikacijo tovrstnih objektov. Z razvojem teleskopov, kot je James Webb, ki prinaša revolucionarne možnosti za opazovanje vesolja, se je povečalo tudi število odkritih galaksij. Ta napredek postavlja pred nas vprašanje, kako učinkovito in natančno klasificirati naraščajoče število galaksij, med katerimi mnoge ostajajo neidentificirane zaradi omejenih kapacitet tradicionalnih metod klasifikacije.

V sodobnem času, ko se tehnologija razvija eksponentno, se odkrivanje vesolja neprestano širi, prinaša pa tudi vedno nove izzive za znanstveno skupnost.

Ta raziskovalna naloga ima namen raziskati, kako bi lahko strojno učenje pripomoglo k izboljšanju klasifikacije galaksij, ob upoštevanju nenehno naraščajočega števila novoodkritih nebesnih teles.

V raziskovalni nalogi bomo razvijali nov model za klasifikacijo morfologije galaksij, osredotočen na uporabo podatkovne zbirke Galaxy Zoo DECaLS. Cilj naloge je ustvariti model, ki bo z visoko natančnostjo lahko določil morfologijo galaksij, uporabljajoč tehnologije strojnega učenja z orodjem TensorFlow.

Galaxy Zoo DECaLS je del večjega projekta Galaxy Zoo, ki vključuje klasifikacijo galaksij s pomočjo množičnega sodelovanja javnosti. Podatkovna zbirka vsebuje slike galaksij z različnimi morfološkimi lastnostmi, kar omogoča strokovnjakom in algoritmom strojnega učenja, da analizirajo in razvrstijo te kompleksne astronomske objekte.



## 5. HIPOTEZE

**Morfološka Razlikovanja:** Model z visoko natančnostjo razlikuje med spiralnimi in eliptičnimi galaksijami na podlagi morfoloških značilnosti na slikah.

**Občutljivost za Značilnosti:** Model globokih učenj prepozna in razvrsti subtilne značilnosti galaksij, kot so prečke, obroči in plinski repi, ki so težko razpoznavni v standardnih astronomskih podatkovnih zbirkah.

**Avtomatizirano Odkrivanje Anomalij:** Model učinkovito identificira anomalne ali nepravilne galaksije, ki ne spadajo v standardne kategorije, kar omogoča odkrivanje unikatnih ali redkih galaktičnih pojavov, vključno s trčenjem galaksij.

## **6. TEORETIČNI DEL**

### **6.1. Strojno učenje**

Strojno učenje predstavlja temeljno področje računalništva, ki se ukvarja z razvojem algoritmov in statističnih modelov, ki omogočajo računalniškim sistemom, da se "učijo" in izvajajo predikcije ali odločitve na podlagi podatkov, brez da bi bili eksplicitno programirani za izvajanje specifičnih nalog. Temelji na ideji, da lahko sistemi najdejo skrite vzorce v podatkih in se na tej osnovi izboljšajo pri opravljanju določenih nalog. Strojno učenje je tesno povezano z umetno inteligenco, saj predstavlja eno od metod za izgradnjo inteligentnih sistemov.

Na področju strojnega učenja ločimo tri glavne vrste učenja: nadzorovano, nenadzorovano in učenje s krepitvijo.

Nadzorovano učenje vključuje učenje na podlagi vnaprej označenih podatkov, kjer model poskuša napovedati izhodno vrednost na podlagi vhodnih podatkov. Primeri uporabe nadzorovanega učenja vključujejo klasifikacijo e-pošte na neželena pošta in ne-neželena pošta ter napovedovanje cen nepremičnin.

Nenadzorovano učenje, nasprotno, deluje na nepoimenovanih podatkih in poskuša najti skrite vzorce ali strukture znotraj podatkovne množice. Uporablja se za segmentacijo trga ali identifikacijo skupin podobnih objektov v podatkovni zbirki. Učenje s krepitvijo je usmerjeno k temu, da model s poskusi in napakami ugotovi, kako izvesti nalogo, da bo maksimiziral neko nagrado, pri čemer se pogosto uporablja v navigaciji robotov in igralnih strategijah.

Poleg teh treh osnovnih tipov učenja se strojno učenje deli tudi na področja, glede na specifične pristope in tehnike, kot so nevronske mreže, globoko učenje, strojno učenje z omejenimi podatki, prenosno učenje in več. Nevronske mreže, zlasti globoko učenje, so se izkazale za izjemno močne pri obdelavi in razumevanju velikih in kompleksnih podatkovnih nizov, kot so slike, zvok in besedilo, kar je pripeljalo do znatnih napredkov na področjih, kot so prepoznavanje govora, vizualno prepoznavanje objektov in naravna obdelava jezika.

Strojno učenje najde uporabo v številnih aplikacijah in industrijah, vključno z zdravstvom, kjer se uporablja za diagnozo bolezni in personalizirano medicino, v financah za upravljanje tveganj in algoritmično trgovanje, v avtomobilski industriji za razvoj avtonomnih vozil, v marketingu za personalizacijo uporabniških izkušenj in v mnogih drugih. Njegova sposobnost, da iz

podatkov izlušči smiselne informacije in napovedi, je temeljna vrednost, ki jo prinaša v sodobno družbo in gospodarstvo.

Z razvojem tehnologije in vedno večjimi količinami zbranih podatkov postaja strojno učenje vse bolj dostopno in uporabno za raziskovalne namene ter inovacije v industriji, kar odpira nove možnosti za razumevanje sveta okoli nas in izboljšanje kakovosti življenja.

V naši raziskovalni nalogi bomo strojno učenje uporabili za razvoj in izvedbo naprednega modela za klasifikacijo galaksij.

## **6.2. Globoko učenje**

Globoko učenje, podkategorija strojnega učenja, predstavlja revolucionaren pristop k modeliranju podatkov, ki se navdihuje po delovanju človeških možganov, natančneje, pri strukturi in funkciji nevronske mreže. Zasnovano je na večplastnih (globokih) nevronske mrežah, ki omogočajo avtomatsko učenje visoko kompleksnih značilnosti iz velikih količin podatkov. Te značilnosti ali reprezentacije so organizirane v hierarhičnih plasteh, kjer vsaka plast transformira svoj vhod na način, ki postopoma povečuje abstrakcijo informacij.

Ena od ključnih prednosti globokih učnih modelov je njihova sposobnost, da iz neobdelanih podatkov samodejno izluščijo relevantne značilnosti, kar odpravlja potrebo po ročnem inženiringu značilnosti, ki je bila prej pogosto zahtevana v tradicionalnih pristopih strojnega učenja. Ta zmožnost globokih učnih modelov je pripomogla k prelomnim dosežkom na področjih, kot so prepoznavanje govora, vizualno prepoznavanje objektov in analiza naravnega jezika, kjer so modeli globokih nevronske mreže znatno presegle zmogljivosti prejšnjih metod.

V kontekstu naše raziskovalne naloge bo globoko učenje služilo kot temelj za razvoj modela, ki bo sposoben natančne klasifikacije galaksij. S pomočjo konvolucijskih nevronske mreže (CNN), specializirane vrste globokih nevronske mreže za obdelavo slik, bomo model usposobili, da samostojno identificira in razume kompleksne morfološke strukture galaksij. Ta pristop omogoča, da model ne le "vidi" slike galaksij, ampak tudi "razume" njihove značilnosti na način, ki presega zgolj vizualno analizo, s čimer omogoča bolj natančno in obsežno klasifikacijo.

Uporaba globokih učnih modelov v naši raziskavi predstavlja obetaven pristop k reševanju izziva avtomatizirane klasifikacije galaksij, saj ponuja možnost za razvoj modelov, ki lahko obvladujejo velike in kompleksne podatkovne nize, kot je Galaxy Zoo DECaLS. S tem pristopom nameravamo prispevati k bolj učinkovitemu in natančnemu razumevanju vesolja, kar odpira nova obzorja v astronomske raziskave in povečuje našo sposobnost interpretacije kozmičnih pojavov.

### **6.3. Nevronska mreža**

Nevronska mreža predstavlja osrednji koncept v globokem učenju in strojnem učenju, navdihnjen po bioloških nevronskih mrežah, ki sestavljajo človeške možgane. Sestavljena je iz vozlišč ali "nevronov", ki so med seboj povezani preko uteži, ki predstavljajo moč povezave med nevroni. Vsak nevron prejema vhodne signale, jih procesira in nato generira izhod, ki služi kot vhod za naslednje plasti nevronov v mreži.

Nevronske mreže so zasnovane za učenje z izkušnjami, prilagajajo se na podlagi podatkov brez potrebe po eksplicitnem programiranju za specifične naloge. Učenje poteka s prilagajanjem uteži med nevroni, kar omogoča mreži, da optimizira svoje odločitve glede na zastavljeno nalogo.

Postopek učenja vključuje vnaprej zastavljen cilj, ki skozi mrežo določa posamezne uteži.

Nevronske mreže so izjemno prilagodljive in zmogljive. Sposobne so učenja kompleksnih vzorcev in odnosov v podatkih, kar jih naredi primerne za širok spekter aplikacij, od prepoznavanja vzorcev, klasifikacije, regresije do generiranja podatkov. Uporabljajo se v različnih področjih, kot so vizualno prepoznavanje objektov, obdelava naravnega jezika, igre, napovedovanje finančnih trendov in še več.

### **6.4. Konvolucijska nevrnska mreža**

Konvolucijska nevrnska mreža (CNN) je posebna vrsta nevrnske mreže, ki je posebej zasnovana za obdelavo podatkov, ki imajo jasno mrežasto topologijo, kot so slike. CNN izkorišča prostorsko strukturo podatkov z uporabo konvolucijskih operacij, ki filtrirajo vhodne podatke za izluščenje pomembnih značilnosti, kot so robovi, texture in druge vizualne znake. Konvolucijske plasti omogočajo mreži, da se avtomatsko in učinkovito uči hierarhije značilnosti iz vhodnih podatkov.

Značilnost CNN je njihova sposobnost deljenja uteži in lokalnega sprejemanja polja, kar pomeni, da se isti filter (ali jedro) uporablja za celotno sliko ali vhodno matriko. To zmanjšuje število prostih parametrov v mreži, kar omogoča globlje in učinkovitejše mreže z manjšo potrebo po računski moči in manjšim tveganjem za prenačenje.

Poleg konvolucijskih plasti CNN običajno vključujejo tudi združevalne (pooling) plasti, ki zmanjšujejo dimenzionalnost značilnostnih zemljevidov, z združevanjem izhodov konvolucijskih operacij. To povečuje odpornost mreže na manjše variacije in premike v vhodnih podatkih. CNN so postale temelj modernega strojnega vida in so ključne pri različnih aplikacijah, vključno z avtomatsko klasifikacijo slik, prepoznavanjem obrazov, avtonomno vožnjo in drugih področjih, kjer je potrebno razumevanje vizualnih informacij.

V kontekstu naše raziskave bo uporaba CNN omogočila natančno analizo in klasifikacijo galaksij, izkoristila bo njihovo sposobnost učenja kompleksnih vzorcev iz vizualnih podatkov, kar bo privedlo do razvoja učinkovitega modela za avtomatizirano klasifikacijo galaksij.

## 6.5. Nevron

Nevron, v kontekstu nevronske mreže, je osnovna računski enota, ki posnema biološki nevron v človeških možganih. Sestavljen je iz več vhodov, ki prejemajo signale, te so povezani z temi vhodi, ki določajo pomen samega signala in aktivacijsko funkcijo, ki določa ali bo nevron poslal signal naprej. Vsak vhodni signal se pomnoži z svojo težo, vsi ti produkti se nato seštejejo skupaj, skupna vsota pa se nato preda aktivacijski funkciji. Iztopni signal nevrona je rezultat aktivacijske funkcije.

Enačba za izračun nevrona  $y$  je lahko predstavljena kot:

*Enačba 1 Izračun nevrona*

$$y = f\left(\sum_{i=1}^n w_i x_i + b\right)$$

Kjer je  $x_i$  vhodni signal,  $w_i$  je utež povezana z vhodom  $x_i$ ,  $b$  je »bias«, ki omogoča prilagoditev izhodom nevrona,  $f$  pa je aktivacijska funkcija.

## 6.6. Perceptron

Pravilo za učenje nevronske mreže temelji na prilagajanju uteži, na podlagi napak pri napovedih. Za vsak primer v naboru podatkov, če je napoved perceptrona napačna, se uteži prilagodijo, da se izboljša natančnost modela v naslednjih iteracijah. Prilagoditev uteži se izvede tako, da se uteži povečajo ali zmanjšajo, odvisno od napake in stopnje učenja  $\eta$ , ki določa velikost koraka prilagoditve:

*Enačba 2 Izračun uteži perceptrona*

$$w_i^{(nova)} = w_i + \eta(y - \hat{y})x_i$$

kjer je  $w_i$  trenutna teža,  $y$  dejanski izhod,  $\hat{y}$  pa je napovedani izhod perceptrona.

Perceptron je bil eden od prvih algoritmov, razvitih v zgodovini strojnega učenja in je služil kot temelj za razvoj bolj kompleksnih nevronskih mrež. Kljub svoji preprostosti in omejitvam (kot je nezmožnost reševanja problemov, ki niso linearno ločljivi) je pomemben zaradi svoje zgodovinske vloge in osnov, ki jih je postavil za nadaljnji razvoj na področju umetne inteligence.

## 6.7. Množice

V kontekstu strojnega učenja in globokih nevronskih mrež, razlikujemo med tremi ključnimi vrstami podatkovnih množic, ki so bistvene za uspešno usposabljanje, validacijo in testiranje modelov. Te so: učna množica, validacijska množica in testna množica.

### 6.7.1. Učna množica

Učna množica (ang. training set) je zbirka podatkov, ki se uporablja za usposabljanje modela strojnega učenja. Vsebuje primerke, za katere so znani tako vhodni podatki kot želene izhodne vrednosti (oznake). Model se "uči" tako, da prilagodi svoje parametre (npr. uteži v nevronske mreži) na način, da bo lahko čim bolj napovedal izhodne vrednosti na podlagi vhodnih podatkov. Učni proces vključuje več iteracij, kjer model poskuša zmanjšati napako med svojimi napovedmi in dejanskimi izhodnimi vrednostmi v učni množici.

### 6.7.2. Validacijska množica

Validacijska množica (ang. validation set) je zbirka podatkov, ki se uporablja za ocenjevanje modela med procesom usposabljanja, vendar se ne uporablja za neposredno učenje (prilagajanje

modelovih parametrov). Glavni namen validacijske množice je, da zagotovi nepristransko oceno učinkovitosti modela med usposabljanjem. To omogoča identifikacijo in preprečevanje prenaučnosti (ang. overfitting), kar se zgodi, ko model postane preveč specifičen za učno množico in izgubi sposobnost generalizacije na novih podatkih. Na podlagi rezultatov, pridobljenih na validacijski množici, se lahko odločimo o prilagoditvah modela ali o ustavitvi usposabljanja.

### **6.7.3. Bias**

Pristranskost ali bias je parameter, ki omogoča modelu, da se bolje prilagodi podatkom s premikom aktivacijske funkcije. Bias deluje kot konstanta dodana vsoti uteži in vhodnih vrednosti pred uporabo aktivacijske funkcije. Pristranskost omogoča, da nevronska mreža izvaja natančnejše prilagoditve, tudi ko so vhodni podatki enaki nič, in s tem izboljša sposobnost učenja kompleksnejših vzorcev.

### **6.7.4. Testna množica**

Testna množica (ang. test set) je zbirka podatkov, ki se uporablja za ocenjevanje končne učinkovitosti modela, po tem ko je bil že usposobljen. Testna množica je popolnoma neodvisna od podatkov, ki so bili uporabljeni med ucnim in validacijskim procesom, kar zagotavlja, da ocena učinkovitosti modela temelji na popolnoma novih podatkih. Uporaba testne množice omogoča realistično oceno, kako bo model deloval v praksi, na podatkih, ki jih model predhodno ni "videl".

## **6.8. Galaksija**

Galaksije predstavljajo enega od temeljnih gradnikov vesolja, masivne sisteme, ki lahko vsebujejo od nekaj milijonov do več kot bilijon zvezd, skupaj z oblaki plina, prahu in temne snovi, ki jih povezuje gravitacija. So predmet številnih raziskav, saj njihova struktura, dinamika in evolucija nudijo ključne vpoglede v razumevanje nastanka in razvoja vesolja. Galaksije obstajajo v različnih oblikah in velikostih, od veličastnih spiralnih galaksij, ki so pogosto predstavljene v javnih medijih, do nepravilnih in eliptičnih galaksij, ki kažejo manj strukturirane značilnosti.

Galaksije ne le, da predstavljajo osnovne strukturne enote vesolja, ampak tudi služijo kot laboratoriji za preučevanje procesov nastajanja zvezd, interakcij med galaksijami in vpliva supermasivnih črnih lukenj na svojo okolico. Njihova razporeditev in gibanje v vesolju sta ključna za razumevanje kozmoloških modelov, vključno z ekspanzijo vesolja in vplivom temne energije.

Vsebina galaksij je kompleksna; poleg zvezd, plina, prahu in temne snovi, vsebujejo tudi različne populacije zvezd. Od starih, zelo razvitih do mladih, masivnih zvezd, ki se šele oblikujejo v njihovih spiralnih rokavih ali jedrih. Medzvezdni medij galaksij je napolnjen z molekularnimi oblaki, ki so ključnega pomena za procese zvezdnega nastajanja, ter s toplo in vročo medzvezdno plazmo, ki kaže na dinamične procese, kot so supernove in galaktični vetrovi.

Galaksije so razporejene po vesolju bodisi kot posamezniki, bodisi v skupinah, jatah in nadjatah. Ta združevanja lahko povzročijo dramatične spremembe v morfologiji in zvezdni populaciji galaksij, kar vodi do pomembnih faz v njihovi evoluciji.

Študij galaksij je pomemben tudi za razumevanje temeljnih principov univerzalne gravitacije in distribucije temne snovi, saj se gibanje zvezd in plina v galaksijah ne ujema vedno s pričakovanji, ki temeljijo na vidni masi. To je privedlo do razvoja teorij o temni snovi, nevidni sestavini, ki prispeva k večini mase v vesolju.

## **6.9. Morfologija galaksij**

Morfologija galaksij predstavlja področje astronomske znanosti, ki se ukvarja s študijem oblik, struktur in razvrščanjem galaksij. Ta raznolikost oblik galaksij ne samo, da odraža različne zgodovine nastanka in evolucijske poti, ampak tudi ponuja vpogled v dinamične procese, ki oblikujejo vesolje. V središču morfologije galaksij je Hubblova klasifikacijska shema, ki jo je v 20. stoletju razvil Edwin Hubble in je še danes temelj za razvrščanje galaksij.

Hubblova klasifikacijska shema galaksije razdeli na več osnovnih tipov: spiralne (S), eliptične (E), lečaste (SO) in nepravilne (Irr). Vsaka od teh kategorij zajema široko paleto objektov z različnimi značilnostmi, ki se odražajo v njihovih oblikah, velikostih in zvezdnih populacijah.



Eliptične galaksije (E) predstavljajo enega od najbolj homogenih tipov galaksij, z značilnostmi, ki segajo od skoraj sferičnih do raztegnjenih elipsoidnih oblik. Brez jasno ločenih struktur, kot so spiralni rokavi. Eliptične galaksije vsebujejo starejše, rdeče zvezde in relativno malo medzvezdnega plina in prahu, kar kaže na omejeno trenutno zvezdno nastajanje.

Spiralne galaksije (S) so morda najbolj ikonične, z značilnimi spiralnimi rokavi, ki se vrtijo okoli sijočega jedra. Spiralni rokavi so prizorišča intenzivnega zvezdnega nastajanja in dom mladih, vročih, modrih zvezd. Spirale galaksije so razdeljene na spirale galaksije (Sa, Sb, Sc) in spirale galaksije s prečko - bariero (SBa, SBb, SBc), kjer "SB" označuje prisotnost prečke, ki prečka jedro in vpliva na dinamiko in strukturo spiralnih rokavov.



Slika 1 Morfologija galaksij

Vir: <https://sl.wikipedia.org/wiki/Galaksija>

Lečaste galaksije (S0) zavzemajo vmesni položaj med eliptičnimi in spiralnimi galaksijami, ki združujejo značilnosti obeh. Sestavljene so iz svetlega jedra, obdanega z diskom, vendar brez izrazitih spiralnih rokavov, kar nakazuje na staro zvezdno populacijo.

Npravilne galaksije (Irr) ne sledijo nobeni od zgoraj navedenih struktur in se pojavljajo v različnih, pogosto kaotičnih oblikah. Ta kategorija zajema galaksije, ki so bodisi bile preoblikovane zaradi gravitacijskih interakcij (trkov) z drugimi galaksijami, bodisi so ohranile

svojo prvotno, nepravilno obliko zaradi relativne izoliranosti. Nepravilne galaksije so pogosto bogate z medzvezdnim plinom in prahom, kar omogoča obsežno nastajanje zvezd.

V novejšem času so raziskave, z uporabo naprednih teleskopov in računalniških simulacij, privedle do dodatnega razumevanja in dopolnitev Hubblove sheme. To vključuje prepoznavanje podkategorij in posebnih primerov, kot so ultra kompaktne pritlikave galaksije (UCD) in ultra difuzne galaksije (UDG), ki širijo našo razumevanje raznolikosti galaktičnih struktur.

V tej raziskovalni nalogi se bomo posvetili le spiralnim in pa eliptičnim galaksijam.

#### **6.10. Klasifikacija galaksij**

V zadnjem desetletju so obsežne nebesne raziskave, kot je Sloan Digital Sky Survey (SDSS), prinesle ogromno količino podatkov. Klasifikacija teh podatkov je za astronome zamudna. Leta 2007 je bil uveden projekt državljanske znanosti "Galaxy Zoo", ki temelji na prostovoljcih, in je znatno skrajšal čas za klasifikacijo. Vendar pa v sodobni dobi globokih učenj avtomatizacija te naloge prinaša velike koristi, saj še dodatno skrajša potreben čas za klasifikacijo. V zadnjih letih je bilo predlaganih več algoritmov, ki so se izkazali za izjemno uspešne pri razvrščanju galaksij v več razredov. Vendar so vsi ti algoritmi galaksije klasificirali v manj kot 6 razredov. Ob upoštevanju podrobnih informacij, ki jih imamo o galaksijah, je postalo jasno, da je nujno galaksije klasificirati v več kot 8 razredov.

V tej študiji je predlagan model nevronske mreže, ki klasificira podatke SDSS v 10 razredov, ki izhajajo iz razširjene Hubbleove vilice za uglasitev. Posebna pozornost je namenjena galaksijam z diskom na robu in obrazu, razlikovanju med različnimi podstrukturami in drobnimi lastnostmi, ki so povezane z vsakim razredom. Predlagani model vključuje konvolucijske plasti za izvlečenje značilnosti, kar omogoča popolnoma avtomatizirano metodo. Dosežena testna natančnost je 84,73 %, kar je obetavno, upoštevajoč tako drobne podrobnosti v razredih. Poleg konvolucijskih plasti ima predlagani model še 3 plasti, odgovorne za klasifikacijo, kar omogoča, da algoritem porabi manj časa.

"V tej študiji je predlagan model nevronske mreže, ki klasificira podatke SDSS v 10 razredov, ki izhajajo iz razširjene Hubbleove vilice za uglasitev." (Galaxy Classification: A deep learning approach for classifying Sloan Digital Sky Survey images, Sarvesh Gharat, Yogesh Dandawate, Department of Electronics and Telecommunication, Vishwakarma Institute of Information Technology, Pune, India).

## **7. PRAKTIČNI DEL**

### **7.1. Metodologija**

V tej raziskovalni nalogi smo uporabili eksperimentalno metodologijo za razvoj in testiranje modela za klasifikacijo galaksij.

#### **7.1.1. Razvojno okolje in orodja**

Za razvoj in implementacijo modelov je bilo izbrano razvojno okolje Jupyter Notebook, gostovano na platformi Google Colab. Google Colab je bil izbran zaradi svoje sposobnosti zagotavljanja močne računske zmogljivosti preko oblaka, vključno z dostopom do naprednih grafičnih procesorskih enot (GPU), kot sta TESLA T4 in NVIDIA A100. Te grafične kartice so bile ključne za učinkovito treniranje modelov globokega učenja, saj omogočajo hitro obdelavo velikih količin podatkov, kar je bistvenega pomena za analizo in klasifikacijo slik galaksij.

#### **7.1.2. Uporabljene knjižnice in pristopi**

Za obdelavo in analizo podatkov so bile uporabljene različne knjižnice v Pythonu. Vključno s Pandas in NumPy za manipulacijo s podatkovnimi okviri in matrične operacije, Matplotlib in Seaborn za vizualizacijo podatkov, ter Scikit-learn za predobdelavo podatkov, razdelitev podatkovnega nabora in ocenjevanje modelov. Uporabljene so bile tudi tehnike strojnega učenja, kot so logistična regresija, naključni gozd in stroji podpornih vektorjev (SVM), za primerjavo različnih pristopov klasifikacije.

Osrednji del raziskave predstavlja uporaba knjižnic TensorFlow in Keras za gradnjo, usposabljanje in validacijo modelov globokega učenja. S pomočjo teh orodij je bil razvit zaporedni model, ki vključuje več slojev konvolucij, združevanja (MaxPooling) in popolnoma povezanih slojev (Dense), skupaj s tehnikami, kot so Dropout za preprečevanje prenaučeni in aktivacijske funkcije ReLU ter sigmoid za linearno in binarno klasifikacijo.

### **7.1.3. Podatkovni viri in shranjevanje**

Za vir podatkov je bila izbrana zbirka Galaxy Zoo DECaLS, dostopna preko projekta na GitHubu, ki je služila kot osnova za zbiranje, obdelavo in analizo slik galaksij. Učna množica podatkov je bila shranjena na Google Drive, kar omogoča enostaven dostop do podatkov iz različnih lokacij in naprav ter zagotavlja integracijo s Colabovim razvojnim okoljem.

## **7.2. Metodološki pristop**

Raziskovalni proces se je začel z zbiranjem in predobdelavo podatkov iz nabora Galaxy Zoo DECaLS, vključno s filtriranjem, standardizacijo in kodiranjem slik. Nato je bila izvedena razdelitev podatkov na učno, validacijsko in testno množico.

### 7.2.1. Instalacija knjižnic

Začeli smo z instalacijo potrebnih knjižnic, ki so osnova za naše delo z naborom podatkov in modeli strojnega učenja. Kot že omenjeno glavne uporabljene knjižnice so bile TensorFlow in Keras za gradnjo in treniranje modelov globokega učenja, ter Scikit-learn za različne predobdelave podatkov, izvedbo razdelitve podatkovnega nabora in ocenjevanje učinkovitosti modelov.

```
import pandas as pd
import numpy as np

import os

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC

import tensorflow as tf
from tensorflow import keras
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout, Conv2D, MaxPooling2D, Flatten

from google.colab import drive
drive.mount('/content/drive')

%matplotlib inline

%config InlineBackend.figure_format = 'retina'

plt.rcParams['figure.figsize'] = (10, 6)

#import warnings
#warnings.filterwarnings('ignore')
print(tf.__version__)
print("GPU Available:", tf.test.is_gpu_available())
```

*Izrezek kode 1 Uporabljene knjižnice*

### **7.2.2. Prenos in priprava podatkovnega nabora**

Celotni podatkovni nabor Galaxy Zoo DECaLS zajema impresivnih 100 GB podatkov s 314.000 galaksijami. Glede na omejitve, ki jih postavlja Google Colab, in praktične potrebe naše raziskave, smo se odločili uporabiti le delček te obsežne zbirke - približno 20 GB podatkov. Ta pristop nam je omogočil, da smo se izognili morebitnim omejitvam prostora in obenem zagotovili, da imamo dovolj obsežno učno množico za zanesljivo treniranje in testiranje naših modelov.

Nato smo prenesli prvi del podatkovnega nabora, ki je obsegal približno 25 GB, direktno s spletne strani projekta. Poleg slik galaksij smo prenesli tudi pripadajočo bazo podatkov v obliki CSV datoteke, ki vsebuje identifikacijske številke slik in njihove klasifikacije.

### **7.2.3. Obdelava in shranjevanje podatkov**

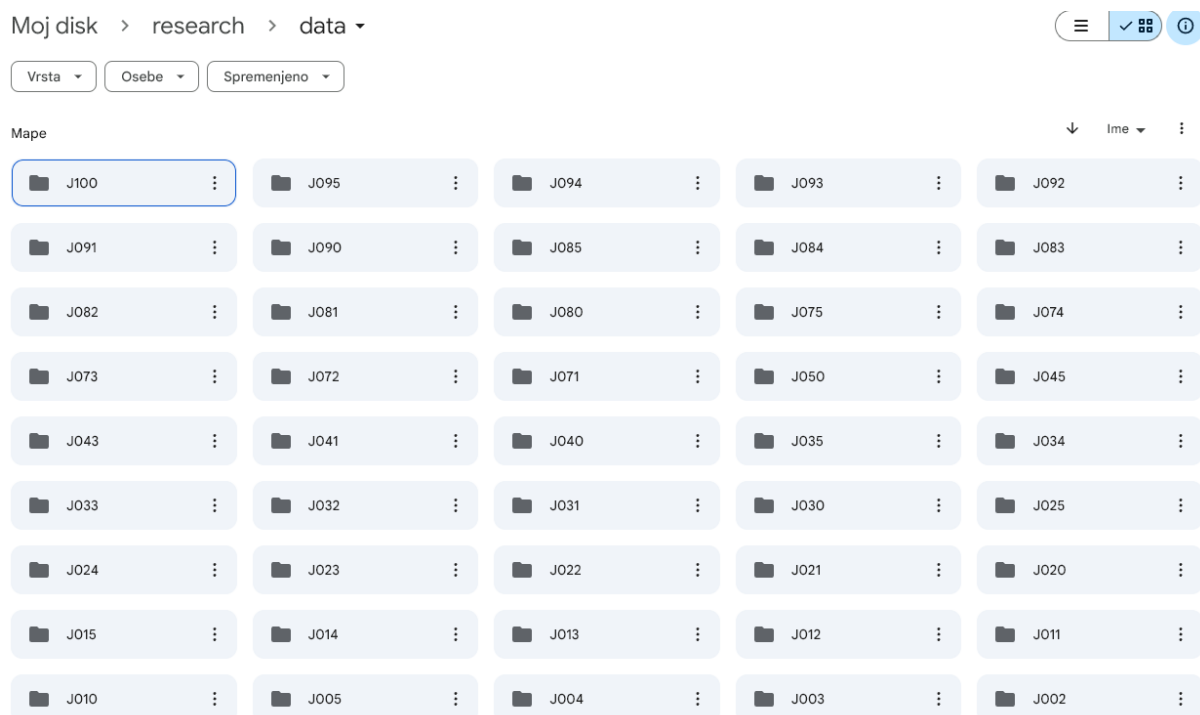
Po prenosu podatkov na virtualni računalnik v Google Colabu smo izvedli postopek ekstrakcije datotek iz ZIP arhiva. To je bil pomemben korak, saj smo s tem pripravili podatke za nadaljnje obdelovanje. Za zagotovitev enostavnega dostopa do podatkov iz različnih lokacij in naprav smo uporabili Google Drive, kar nam je omogočilo shranjevanje učne množice na način, ki podpira neprekinjeno delo v Colabovem okolju.

### **7.2.4. Strukturiranje podatkov in organizacija**

Za učinkovito delo in lažji dostop do podatkov smo vzpostavili jasno strukturo shranjevanja. Celoten dataset, ki zavzema približno 100 GB, je razdeljen na pet delov. Odločili smo se osredotočiti na prvi del tega obsežnega nabora, velik približno 25 GB. Ta del je po našem mnenju vseboval dovolj veliko učno množico za potrebe našega projekta, hkrati pa je upoštevanje omejitev Colabovega prostora za shranjevanje omogočilo bolj gladko delovanje.

V mapi research na Google Drive-u smo shranili CSV datoteko, ki vsebuje identifikacijske številke in klasifikacije galaksij. Ta organizacijski pristop nam omogoča, da imamo vse potrebne metapodatke na dosegu roke, ločeno od samih slik, kar poenostavi obdelavo in analizo podatkov. Ekstrahirane slike iz ZIP datoteke so bile shranjene v podmapo research/data, kjer

smo ohranjali enako strukturo, kot je bila zasnovana v originalnem naboru podatkov.



Slika 2 Podatkovna zbirka shranjena v Google Drive

Vir: Lasten vir

### 7.2.5. Detajlna struktura podmap

Znotraj mape research/data smo ustvarili 101 podmapo, označeno od J000 do J100. Vsaka od teh map vsebuje približno 1500 slik, kar pomeni, da smo za vsako podmapo obdelovali in analizirali veliko število galaksij. Slike so imele dimenzije 424 x 424 pikslov, kar zagotavlja visoko kakovost podrobnosti in omogoča natančno klasifikacijo.

### 7.2.6. Kako je sestavljen podatkovni nabor

Pri raziskovanju nebesnih teles s pomočjo strojnega učenja je ključnega pomena natančna identifikacija in organizacija podatkov. V našem projektu smo uporabili podatkovni nabor Galaxy Zoo DECaLS, ki vsebuje slike galaksij, zbrane z namenom klasifikacije njihovih morfologij. Imena slik v tem naboru so zasnovana po posebnem sistemu, ki temelji na koordinatnih sistemih, uporabljenih v astronomiji. Ta pristop omogoča izjemno natančnost pri identifikaciji in kategorizaciji galaksij, kar je bistvenega pomena za uspeh naših raziskav.

Vsaka slika v našem podatkovnem naboru je poimenovana po standardiziranem vzorcu, ki omogoča natančno določitev lokacije vsake galaksije na nebesnem svodu. Na primer, ime slike



J000009.64+140542.7.png vsebuje več ključnih informacij. Prva oznaka J se nanaša na referenčni okvir J2000.0, ki je široko sprejet v astronomski skupnosti. Ta referenčni okvir zagotavlja fiksno točko v času, glede na katero lahko izmerimo položaje nebesnih teles, kar omogoča znanstvenikom, da delijo in primerjajo opazovanja z visoko stopnjo natančnosti.

Naslednji segment, 000009.64, predstavlja rektascenzijo (RA) galaksije. Rektascenzija je nebesna dolžina - astronomski ekvivalent zemljepisne dolžine in se meri v urah, minutah in sekundah, od 0 do 24 ur. Ta koordinata nam pove, kako daleč vzhodno ali zahodno od pomladišča se nahaja objekt. Pomladišče je točka na nebesnem ekvatorju, kjer Sonce prečka ekvator 21. marca, označuje začetek pomladi na severni polobli. Je presečišče nebesnega ekvatorja in ekliptike (ravnine gibanja planetov).

Zadnji del, +140542.7, se nanaša na deklinacijo (Dec) - nebesno širino, ki je ekvivalent zemljepisne širine in se meri v stopinjah od - 90 do + 90 kotnih stopinj. Deklinacija pove, kako daleč severno ali južno od nebesnega ekvatorja se nahaja objekt na nebu. Deklinacija nebesnega ekvatorja je 0 kotnih stopinj. Pozitivne vrednosti označujejo objekte severno od nebesnega ekvatorja, negativne pa južno.

Ta sistem imenovanja ne le olajša iskanje in analizo specifičnih galaksij znotraj obsežnega podatkovnega nabora, ampak tudi omogoča raziskovalcem, da natančno določijo njihovo lokacijo v vesolju. To je še posebej pomembno pri primerjavi opazovanj, zbranih z različnih instrumentov ali v različnih časovnih obdobjih.

Poleg tega ta metoda imenovanja odraža globoko povezanost med astronomske podatke in koordinatne sisteme, ki se uporabljajo za kartiranje nebesnega svoda. V kontekstu našega projekta ta pristop ne omogoča le natančne identifikacije in klasifikacije galaksij, temveč tudi podpira kompleksne analize, kot so iskanje vzorcev v distribuciji galaksij, razumevanje njihove prostorske orientacije in preučevanje dinamike galaktičnih skupin.

Ko smo zagotovili dostop do vseh relevantnih podatkov za naš projekt klasifikacije galaksij, smo se soočili z novim izzivom: obsežnost in kompleksnost podatkov. Naša glavna .csv datoteka, ki je bila velika 2.5 GB, je vsebovala ogromno količino podatkov, od katerih mnogi niso bili neposredno relevantni za naše specifične potrebe. Zavedajoč se omejitev, ki jih prinaša

delo z obsežnimi podatkovnimi zbirkami, in željo po učinkovitejšem delovnem procesu, smo se odločili za korak, ki bi optimiziral našo delovno bazo.

Kako dejansko izgledajo fotografije v našem naboru podatkov:



Slika 3 Primer a



Slika 4 Primer b



Slika 5 Primer c



Slika 6 Primer d



Slika 7 Primer e





Slika 8 Primer f

Viri slik: Galaxy Zoo DECaLS

### 7.2.7. Proces filtriranja podatkov

Prvi korak v procesu optimizacije je bilo ustvarjanje nove datoteke, poimenovane `filtered_gz_decals_auto_posteriors.csv`. Namen te datoteke je bil shraniti le tiste podatke, ki so bili bistveni za naše analize in treniranje modelov strojnega učenja. Ta proces filtriranja ni pomenil le odstranjevanje nepotrebnih stolpcev iz izvirne datoteke, ampak tudi znatno zmanjšanje njene velikosti - s 2.5 GB na zgolj 50 MB. Takšno zmanjšanje je imelo dvojni

 filtered_gz_decals_auto_posteriors (1).csv	7. 02. 2024 00:33	Microsoft Excel Com...	56.259 KB
 gz_decals_auto_posteriors.csv	6. 02. 2024 23:06	Microsoft Excel Com...	2.518.711 KB

Slika 9 Primerjava velikosti CSV datotek

Vir: Lasten vir

učinek: po eni strani je olajšalo rokovanje s podatki in zmanjšalo čas, potreben za njihovo obdelavo, po drugi strani pa je izboljšalo učinkovitost naših analitičnih in trenirnih procesov.

### **7.2.8. Dodatna optimizacija podatkovne datoteke:**

Nadaljnja izboljšava je vključevala specifično funkcijo, ki je iz naše filtrirane datoteke odstranila vse zapise, ki niso bili del prvega dela našega dataset-a. Glede na to, da smo se osredotočili izključno na prvi del celotnega nabora podatkov, ki je obsegal okoli 25 GB slik galaksij, je bilo smiselno odstraniti vse reference, ki se nanašajo na ostale dele dataset-a. S pomočjo te funkcije smo izbrisali vsak zapis, kjer je bilo ime datoteke večje kot P100, s čimer smo še dodatno zmanjšali velikost naše delovne .csv datoteke in zagotovili, da vsebuje samo najbolj relevantne in uporabne informacije za naš projekt.

### **7.2.9. Kako bomo dostopali do datotek**

Za shranjevanje in dostop do učne množice smo izbrali Google Drive. Ta odločitev je izhajala iz potrebe po fleksibilnosti in dostopnosti podatkov iz različnih lokacij. Nismo se omejili le na delo z lokalnega računalnika; želeli smo zagotoviti, da so naši podatki dostopni kjerkoli in kadarkoli, kar je Google Drive omogočal z izjemno preprostostjo. Ta izbira je bila še posebej pomembna, saj je omogočila, da smo lahko delali na projektu iz različnih naprav brez potrebe po ponovnem prenašanju ali sinhroniziranju podatkov. Alternativa, postavitve lastnega strežnika doma, bi sicer ponudila nekaj prednosti v smislu nadzora nad podatki in možnosti prilagajanja, vendar bi hkrati prinesla dodatne izzive, kot so vzdrževanje strežnika, varnostni vidiki in omejitve dostopnosti.

Pri delu na projektu smo se soočili z določenimi omejitvami, ki jih postavlja Google Colab, predvsem s časovno omejitvijo sej. Google Colab samodejno prekine sejo, če ta ni aktivna več kot tri ure, kar pomeni, da vsi začasno shranjeni podatki na njegovem strežniku niso več dostopni. Kljub temu smo se odločili uporabiti začasni prostor, ki ga ponuja Google Colab, za končno učenje naše nevronske mreže. Glavna prednost te odločitve je bila, da smo imeli program in učno množico shranjeno na istem disku, kar je močno poenostavilo delo z velikimi količinami podatkov. Ta pristop je omogočil hitrejše branje in pisanje podatkov ter s tem učinkovitejše treniranje modelov.

Za zagotovitev učinkovitega procesa učenja smo učno množico razdelili na tri ključne segmente: učno, validacijsko in testno množico. Ta razdelitev je bila ključna za razvoj in ocenjevanje učinkovitosti naših modelov strojnega učenja. Učna množica je služila kot osnova za treniranje modelov, validacijska množica pa za preverjanje njihovega delovanja in prilagajanje hiperparametrov med procesom učenja. Testna množica, ločena od prvih dveh, je bila uporabljena za končno oceno učinkovitosti modela, s čimer smo zagotovili, da so naše ugotovitve zanesljive in da model dobro deluje na nevidnih podatkih

### **7.3. Povezovanje podatkov iz CSV datoteke z bazo slik**

Pri razvoju naprednih sistemov strojnega učenja, ki so namenjeni klasifikaciji nebesnih teles, kot so galaksije, temelji uspešnost modela na kakovosti in obsegu uporabljenih podatkov. Eden ključnih korakov pri pripravi podatkovne zbirke za takšen sistem je povezovanje identifikatorjev nebesnih teles, ki so shranjeni v strukturirani obliki, kot je CSV datoteka, z ustreznimi slikami, ki predstavljajo ta nebesna telesa. V primeru naše raziskave imamo na voljo 101 mapo, poimenovano od J000 do J100, kjer vsaka mapa vsebuje različno število slik galaksij. Smeri, v katerih so bile te galaksije posnete, se razlikujejo, kar pomeni, da nekatere mape vsebujejo več slik galaksij kot druge.

```

csv_path =
'/content/drive/MyDrive/research/filtered_gz_decals_auto_posteriors.csv'
df = pd.read_csv(csv_path)

threshold = 0.5
df['label'] = (df['has-spiral-arms_yes_fraction'] > threshold).astype(int)

df['file_loc'] = df['file_loc'].apply(lambda x: '/content/temp_extracted/' +
x.replace('dr5/png/', ''))
df_filtered = df[df['file_loc'].apply(os.path.exists)]

def parse_function(filename, label):
    image_string = tf.io.read_file(filename)
    image_decoded = tf.image.decode_png(image_string, channels=3)
    image_resized = tf.image.resize(image_decoded, [106, 106])
    return image_resized / 255.0, label

filenames = df_filtered['file_loc'].values
labels = df_filtered['label'].values

dataset = tf.data.Dataset.from_tensor_slices((filenames, labels))
dataset = dataset.map(parse_function)

```

*Izrezek kode 2 Povezovanje podatkov iz csv in učnega nabora*

### 7.3.1. Izdelava podatkovnega nabora

Za uspešno povezavo podatkov iz CSV datoteke z ustreznimi slikami v naši bazi, smo uporabili programski pristop, ki omogoča avtomatizacijo procesa združevanja. Vsaka vrstica v CSV datoteki vsebuje identifikacijske podatke, ki so povezani s posamezno sliko galaksije. Ti identifikacijski podatki vključujejo unikatne identifikatorje, ki ustrezajo imenom map in slik v naši podatkovni zbirki. Glavni izziv, s katerim smo se soočili, je bil zagotoviti, da ima vsaka vrstica v CSV datoteki svojo ustrezno sliko v mapi.

### 7.3.2. Povezovanje podatkov

Da bi zagotovili natančno povezavo med podatki v CSV datoteki in slikami, smo razvili algoritem, ki sledi naslednjim korakom:

1. Branje CSV datoteke: Najprej smo prebrali CSV datoteko in iz nje izluščili potrebne informacije, kot so identifikatorji in klasifikacije galaksij.
2. Dostop do map s slikami: Nato smo programsko prešli skozi vse mape, od J000 do J100, in preverili prisotnost slik, ki ustrezajo identifikatorjem iz CSV datoteke.
3. Preverjanje ujemanja: Za vsako vrstico v CSV datoteki smo preverili, ali obstaja ustrezna slika v eni od map. To je bilo ključno za zagotovitev, da bo podatkovni nabor, uporabljen za učenje nevronske mreže, popoln in brez manjkajočih slik.
4. Obvladovanje manjkajočih slik: Če slika, ki ustreza določenemu identifikatorju v CSV datoteki, ni bila najdena, smo to situacijo obravnavali kot potencialni problem, saj bi manjkajoče slike lahko negativno vplivale na učenje in natančnost nevronske mreže. V takih primerih smo se odločili za izključitev teh vrstic iz končnega podatkovnega nabora, da bi zagotovili njegovo integriteto.
5. Izdelava končnega podatkovnega nabora: Po uspešnem povezovanju identifikatorjev z ustreznimi slikami smo ustvarili končni podatkovni nabor, ki je bil pripravljen za uporabo v procesu strojnega učenja.

```
dataset = dataset.shuffle(buffer_size=len(df_filtered))

train_size = int(0.7 * len(df_filtered))
val_size = int(0.2 * len(df_filtered))
test_size = len(df_filtered) - train_size - val_size

train_dataset = dataset.take(train_size)
test_val_dataset = dataset.skip(train_size)
val_dataset = test_val_dataset.take(val_size)
test_dataset = test_val_dataset.skip(val_size)

BATCH_SIZE = 32
train_dataset =
train_dataset.batch(BATCH_SIZE).prefetch(tf.data.experimental.AUTOTUNE)
val_dataset =
val_dataset.batch(BATCH_SIZE).prefetch(tf.data.experimental.AUTOTUNE)
test_dataset =
test_dataset.batch(BATCH_SIZE).prefetch(tf.data.experimental.AUTOTUNE)
```

*Izrezek kode 3 Izdelava končnega podatkovnega nabora*

#### 7.4. Učenje nevronske mreže

Za treniranje modela za klasifikacijo galaksij smo izbrali konvolucijske nevronske mreže (CNN), ki so se izkazale kot zelo učinkovite pri obdelavi slik. Vhodne slike za treniranje so dimenzij 424 x 424 pikslov in vsebujejo tri barvne kanale (RGB). Ta visoka ločljivost omogoča, da model zazna fine morfološke značilnosti galaksij, kot so spiralne rokave, prečke in jedra.

Začeli bomo s treniranjem modela na zmogljivih grafičnih procesorskih enotah (GPU), kot sta NVIDIA Tesla T4 in NVIDIA A100. Model bomo sprva optimizirali in fino nastavili na T4, ki je cenejša opcija, preden preidemo na dražjo in zmogljivejšo A100. Ta pristop nam omogoča, da preverimo delovanje in prilagodljivost modela pri različnih konfiguracijah strojne opreme.

V primeru, da se treniranje na slikah polne velikosti izkaže za preveč zahtevno - bodisi zaradi časovnih omejitev ali omejitev strojne opreme - bomo razmislili o zmanjšanju vhodne resolucije slik na 106 x 106 pikslov. Ta ukrep bi zmanjšal računalniške zahteve, vendar bi lahko vplival na zmožnost modela, da natančno razlikuje med kompleksnimi morfološki značilnostmi galaksij. Vendar pa bi ohranjanje treh barvnih kanalov še vedno omogočalo, da model izkorišča informacije o barvi, ki so lahko ključne za razlikovanje med različnimi tipi galaksij.

Konvolucijske nevronske mreže so idealne za tovrstne naloge, saj lahko avtomatsko in učinkovito izluščijo značilnosti iz slik, ki so pomembne za klasifikacijo. S tem pristopom si prizadevamo doseči visoko stopnjo natančnosti pri klasifikaciji galaksij, kar bo pripomoglo k boljšemu razumevanju njihovih morfoloških lastnosti in razširitvi našega znanja o vesolju.

V kontekstu učenja konvolucijskih nevronskih mrež (CNN) za klasifikacijo galaksij, je ena ključnih enačb, ki se uporablja pri oblikovanju arhitekture mreže, enačba za izračun dimenzij izhoda posameznega konvolucijskega sloja. Ta enačba je pomembna za načrtovanje globine in kompleksnosti modela ter za zagotavljanje, da so vhodne in izhodne dimenzije slojev usklajene s pričakovanji. Enačba za izračun dimenzij izhoda ( $O$ ) konvolucijskega sloja, ob upoštevanju dimenzij vhoda ( $I$ ), velikosti filtra ali jedra ( $K$ ), koraka ( $S$ ) in oblazinjenja ( $P$ ), je naslednja

*Enačba 3 Izračun konvolucijskega sloja*

$$O = \frac{I - K + 2P}{S} + 1$$

Razčlenimo enačbo:

- I predstavlja dimenzijo ene stranice vhodne slike ali značilnostnega zemljevida. V primeru našega projekta, kjer začnemo z vhodnimi slikami dimenzij 424 x 424 pikselov, bi I bil 424.
- K je velikost filtra ali jedra, ki ga uporabljamo v konvolucijskem sloju. Filtri so ključni za izluščanje značilnosti iz slik, velikost pa običajno izbiramo glede na specifične potrebe projekta, npr. 3x3 ali 5x5.
- S je korak, s katerim premikamo filter preko vhodne slike. Korak določa, kako "gosto" filter preslika vhodno sliko. Večji kot je korak, manjši je izhodni značilnostni zemljevid.
- P je oblazinjenje, ki se doda okoli robov vhodne slike. To omogoča, da filter deluje tudi na robovih slike, s čimer se ohranja prostorska dimenzija značilnostnega zemljevida.

Primer: Če imamo vhodno sliko dimenzij 424 x 424, uporabimo filter velikosti 3x3, korak 1 in brez oblazinjenja ( $P=0$ ), potem bi izhodna dimenzija značilnostnega zemljevida po tem sloju bila:

$$O = \frac{424 - 3 + 2(0)}{1} + 1 = 422$$

To pomeni, da po uporabi konvolucijskega sloja z navedenimi parametri dobimo izhodni značilnostni zemljevid dimenzij 422 x 422. Ta izračun nam pomaga pri načrtovanju arhitekture CNN, saj moramo zagotoviti, da bodo vse dimenzije usklajene skozi celotno mrežo, kar omogoča učinkovito učenje in izogibanje napakam pri dimenzijah.



## CNN nastavitve

```
def ConvBlock(layers, model, filters):
    for i in range(layers):
        model.add(ZeroPadding2D((1,1)))
        model.add(Conv2D(filters, (3, 3), activation='relu'))

    model.add(MaxPooling2D((2,2), strides=(2,2)))

def FCBlock(model):
    model.add(Dense(4096, activation='relu'))
    model.add(Dropout(0.5))

def VGG_16():
    model = Sequential()

    model.add(Conv2D(64, (3, 3), activation='relu', input_shape=(106, 106,3)))

    ConvBlock(1, model, 64)
    ConvBlock(2, model, 128)
    ConvBlock(3, model, 256)
    ConvBlock(3, model, 512)
    ConvBlock(3, model, 512)

    model.add(Flatten())
    FCBlock(model)
    FCBlock(model)

    model.add(Dense(1, activation='sigmoid'))

    return model

optimizer = Adam(lr=1e-4)
model = VGG_16()
model.compile(loss='binary_crossentropy', optimizer=optimizer,
metrics=['accuracy'])
```

### *Izrezek kode 4 CNN nastavitve*

V procesu učenja nevronske mreže za klasifikacijo galaksij smo se osredotočili na implementacijo zaporednega modela, ki uporablja več plasti za izluščenje in učenje prostorskih informacij iz vhodnih slik. Začeli smo z definiranjem vhodne plasti modela, ki sprejema slike velikosti 424x424 pikslov, kar je optimalno za naše potrebe, saj želimo zagotoviti, da model

zajame čim več podrobnosti z vsake slike. Vsaka slika ima tri barvne kanale (RGB), ki omogočajo modelu, da izkoristi barvne informacije pri klasifikaciji galaksij.

Prva konvolucijska plast uporablja 32 filtrov velikosti 3x3, kar omogoča modelu, da zazna manjše vzorce in značilnosti, kot so robovi in teksture znotraj slik. Aktivacijska funkcija ReLU je izbrana zaradi njene sposobnosti vnašanja nelinearnosti v model, kar je ključno za učenje kompleksnih vzorcev. Sledi plast za združevanje z oknom 2x2, ki zmanjša dimenzionalnost izhodnih značilnostnih zemljevidov (»Feature Map«), s čimer se zmanjša tveganje za prenaučenos in poveča učinkovitost obdelave.

V konvolucijski plasti se na vhodnih slikah izvaja operacija filtriranja, da se identificirajo prostorske značilnosti, kot so robovi, teksture ali vzorci. Zapis enačbe za prvo konvolucijsko plast z 32 filtri velikosti 3x3 in aktivacijsko funkcijo ReLU je:

*Enačba 4 Funkcija ReLU*

$$F_{ij}^{(1)} = \text{ReLU} \left( \sum_{m=0}^2 \sum_{n=0}^2 \sum_{k=1}^3 W_{mnk} \cdot X_{i+m,j+n,k} + b \right)$$

Vsak filter te plasti se "spelje" čez vhodno sliko, pri čemer izvaja elementno množenje svojih uteži  $W_{mnk}$  z ustreznimi piksli vhodne slike  $X_{i+m,j+n,k}$ . Rezultati se nato seštejejo, k temu se prišteje še pristranskost (»bias«)  $b$ , in končni rezultat se preda funkciji ReLU.

Plast za združevanje sledi konvolucijski plasti in služi zmanjšanju dimenzionalnosti zemljevidov značilnosti, kar pripomore k zmanjšanju števila parametrov in izboljšanju učinkovitosti mreže ter pomoči pri preprečevanju prenaučenos. Za plast združevanja z oknom 2x2 enačba izgleda:

*Enačba 5 Enačba plasti za združevanje*

$$P_{ij}^{(1)} = \max_{a \in \{0,1\}, b \in \{0,1\}} \left( F_{i \cdot 2 + a, j \cdot 2 + b}^{(1)} \right)$$

Za vsako območje velikosti 2x2 na zemljevidu značilnosti, ki ga proizvede konvolucijska plast, ta operacija izbere maksimalno vrednost. Ta izbrana vrednost potem predstavlja celotno območje v zmanjšanem zemljevidu značilnosti. Z združevanjem (pooling) se bistveno zmanjša velikost vsakega zemljevida značilnosti, pri čemer se ohranja najpomembnejše informacije. To omogoča mreži, da postane bolj odporna na manjše variacije in translacije v vhodnih podatkih.

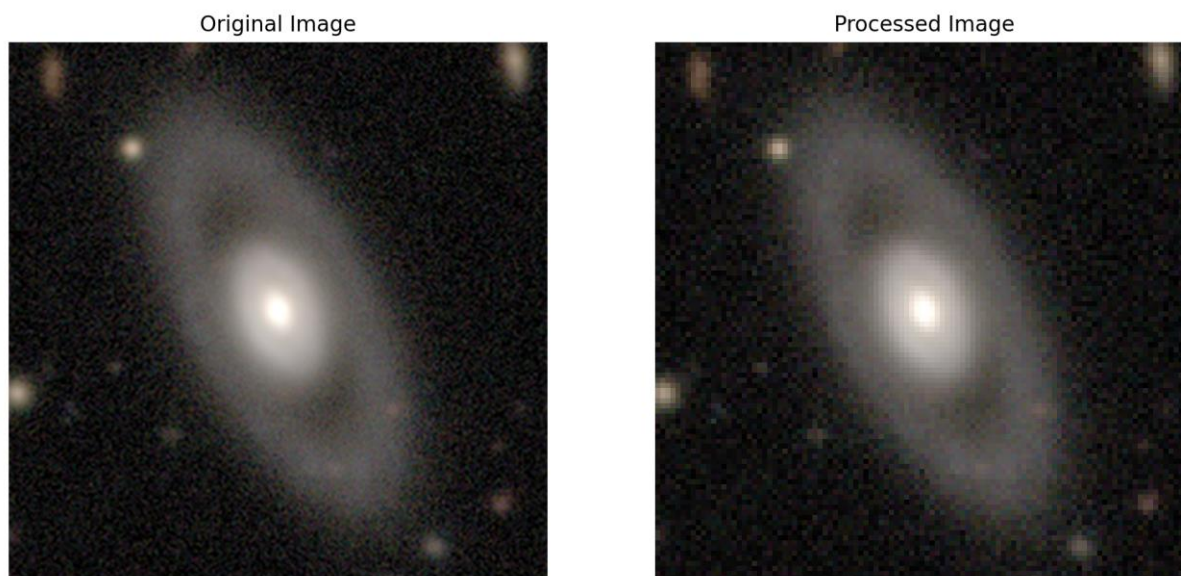
Z izbiro maksimalnih vrednosti pa se ohranja prisotnost določenih značilnosti, kot so robovi ali teksture, ki so ključni za prepoznavanje objektov.

Model nadaljuje z dodatnima konvolucijskima plastema, kjer druga uporablja 64 filtrov, tretja pa 128 filtrov, obakrat z jedrom  $3 \times 3$ . Ta postopno povečanje števila filtrov omogoča modelu, da postane občutljivejši na bolj specifične in kompleksne značilnosti v slikah, kot so strukture galaksij. Vsaki konvolucijski plasti sledi še ena plast za združevanje, ki nadalje zmanjšuje dimenzije značilnostnih zemljevidov in pomaga modelu osredotočiti se na najpomembnejše informacije.

Po več plasteh konvolucije in združevanja model vključuje ploščati sloj, ki pretvori večdimenzionalne značilnostne zemljevide v enodimenzionalni vektor. Ta korak je nujen za prehod od prostorskih značilnosti k polno povezanim plastem, ki omogočajo klasifikacijo. Prva polno povezana plast (Dense) uporablja 512 enot in aktivacijsko funkcijo ReLU, kar modelu omogoča, da združi naučene značilnosti v kompleksnejše vzorce, ki so pomembni za končno odločitev klasifikacije.

Zadnji sloj modela je polno povezana plast z eno enoto in sigmoidno aktivacijsko funkcijo, ki izračuna verjetnost, da slika pripada določeni kategoriji galaksij. Ta pristop omogoča, da model na koncu poda verjetnostno oceno, ki jo lahko interpretiramo kot konfidenčno merilo, ali slika prikazuje določen tip galaksije.

Kako izgleda fotografija po tem ko je sprocesirana:



*Slika 10 Razlika med originalno in procesirano sliko*

*Vir: Lasten vir*

Celoten proces učenja modela poteka na zmogljivih grafičnih procesorskih enotah, kot sta NVIDIA Tesla T4 in NVIDIA A100, kar zagotavlja hitro in učinkovito treniranje. Začetno učenje in optimizacija modela na T4 omogočata preverjanje in prilagajanje modela pred prehodom na dražjo A100, kar optimizira stroške in zagotavlja, da je model čim bolj natančen pred uporabo v produkciji. V primeru, da model na slikah polne resolucije ne dosega zelenih rezultatov, smo pripravljeni zmanjšati dimenzije vhodnih slik, kar bi lahko pripomoglo k hitrejšemu učenju, vendar s potencialnim tveganjem izgube nekaterih finih značilnosti, ki so ključne za natančno klasifikacijo.

#### **7.4.1. Kako smo nevronske mreže učili**

Ko smo začeli z učenjem nevronske mreže za klasifikacijo galaksij, smo se osredotočili na tri ključne aspekte, ki so bistvenega pomena za uspešnost in učinkovitost modela: velikost serije (batch size), število epoh in mehanizme za preprečevanje prenaučeniosti.

Izbrali smo batch size 32, kar pomeni, da naša mreža obdela 32 slik pred vsako posodobitvijo uteži. Ta pristop nam omogoča, da učinkovito izkoristimo računalniške vire, saj namesto obdelave celotne učne množice naenkrat, kar bi bilo zelo zahtevno in bi lahko vodilo do prekomerne porabe pomnilnika, obdelujemo manjše skupine slik. Poleg tega nam ta pristop

omogoča, da vnesemo element stohastičnosti<sup>1</sup> v proces učenja, kar nam pomaga pri iskanju globalnega minimuma funkcije stroškov, namesto da bi se ujel v lokalne minimume.

Nastavili smo tudi učenje modela na 10,000 epoh, kar pomeni, da bo naš model večkrat prešel čez celotno učno množico, kar mu omogoča, da se nauči in prilagodi na podlagi opazovanih vzorcev. Večje število epoh nam daje več priložnosti za optimizacijo modela, vendar smo hkrati zavedeni tveganja prenaučeniosti, ki se pojavi, ko se model preveč prilagodi učnim podatkom in posledično izgubi svojo sposobnost generalizacije na novih podatkih.

Da bi se izognili prenaučeniosti, smo uporabili tehniko zgodnjega ustavljanja s “patience” nastavitvijo na 7. Ta tehnika nam omogoča, da natančno spremljamo uspešnost modela na validacijski množici podatkov in če se uspešnost ne izboljša v sedmih zaporednih epohah, proces učenja avtomatsko prekinemo. Ta pristop nam pomaga ohraniti uravnoteženo sposobnost modela, da se uči iz podatkov, ne da bi pri tem izgubil svojo splošno učinkovitost. Zgodnje ustavljanje nam torej služi kot varovalka, ki zagotavlja, da naš model ostane robusten in sposoben generalizacije, ne glede na to, koliko epoh smo prvotno načrtoval izvesti.

---

<sup>1</sup> Stohastičnost pomeni uporabo naključnosti ali verjetnostnih procesov.

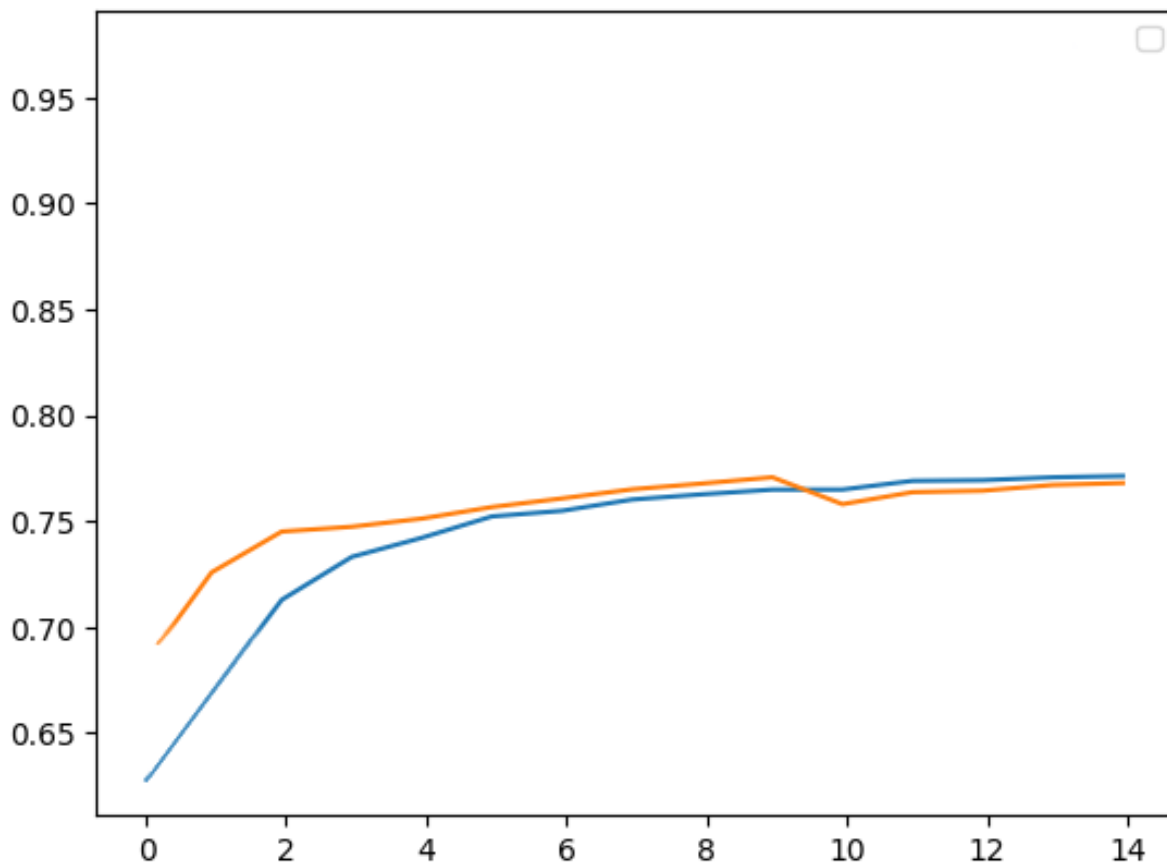
## 8. REZULTATI

Prvi pomemben korak je bila optimizacija podatkovne zbirke za lažje upravljanje v okolju za programiranje Colab. Razvili smo funkcijo, ki omogoča enostavno nalaganje podatkovne zbirke na katerikoli strežnik, kar pomeni znatno izboljšanje dostopnosti in uporabnosti podatkov za strojno učenje. Poleg tega smo ustvarili orodje, ki omogoča prilagajanje .csv datotek glede na določene parametre, kar uporabnikom omogoča odstranjevanje nepotrebnih podatkov iz katerekoli csv datoteke, ki se uporablja za strojno učenje.

Ta pristop ne samo da poenostavlja obdelavo in analizo podatkov, temveč tudi povečuje učinkovitost raziskovalnega procesa. Z zmožnostjo natančnega prilagajanja podatkovnih nizov so raziskovalci bolj opremljeni za osredotočanje na relevantne informacije, kar vodi k bolj točnim in zanesljivim rezultatom strojnega učenja.

Nadalje smo pri razvoju našega modela strojnega učenja dosegli pomembne tehnične izboljšave. Funkcija, ki smo jo razvili za lažje upravljanje podatkovnih zbirk, omogoča hitrejši in bolj fleksibilen dostop do podatkov, kar je ključno za učinkovito treniranje modelov strojnega učenja. Prav tako smo z orodjem za urejanje .csv datotek zagotovili, da so podatki, uporabljeni v procesu učenja, natančno prilagojeni potrebam našega modela, kar je povečalo njegovo natančnost in zmanjšalo čas potreben za njegovo treniranje.

Na podlagi podatkov, ki smo jih pridobili med raziskavo, smo ugotovili, da je naš model strojnega učenja za klasifikacijo galaksij dosegel 75% natančnost. To pomeni, da je bil model sposoben pravilno klasificirati tri četrtine vseh galaksij v testni množici, kar predstavlja pomemben uspeh glede na kompleksnost in raznolikost morfoloških značilnosti galaksij. Pri tem smo se omejili na uporabo le 25 GB podatkov iz celotne 100 GB velike podatkovne zbirke, kar kaže na to, da je naš model zmožen delovati učinkovito tudi pri manjših količinah podatkov.



*Graf 1 Kako se je NM učila*

Pomembno je izpostaviti, da smo pri izvajanju učenja modela opazili pomembne razlike v času obdelave med različnimi grafičnimi procesnimi enotami (GPU). Medtem ko je na GPU enoti T4 celoten proces učenja za eno epoho trajal približno 30 minut, je bila na naprednejši GPU enoti A100 potrebna le petina tega časa, torej približno 5 minut na epoho. Ta opažanja poudarjajo pomembnost izbire ustrezne strojne opreme za optimizacijo časa učenja, kar je ključnega pomena pri obdelavi velikih podatkovnih zbirk.

Vendar pa smo med raziskavo naleteli tudi na oviro, ki se nanaša na velikost vhodnih slik. Sprva smo predpostavljali, da uporaba visokokakovostnih slik ne bo predstavljala težave, saj smo imeli dostop do zmogljivih grafičnih kartic. Kljub temu se je izkazalo, da je bila naša izbira prevelikih slik neoptimalna, kar je vplivalo na učinkovitost modela. To nas uči, da je pri delu s podatkovno intenzivnimi projekti ključnega pomena optimizacija vhodnih podatkov, da se zagotovi hitra in učinkovita obdelava, ne glede na zmogljivost uporabljene strojne opreme.

Natančnost 75% modela v kontekstu strojnega učenja pomeni, da ima model dobro splošno zmožnost razlikovanja med eliptičnimi in spiralnimi galaksijami, vendar še vedno obstaja prostor za izboljšave. Ta natančnost omogoča modelu, da je v veliki meri zanesljiv pri identifikaciji in klasifikaciji galaksij, vendar pa je za nadaljnje izboljšave potrebno nadaljnje fine nastavitve modela, morda z vključitvijo dodatnih podatkovnih množic, izboljšanjem algoritmov za obdelavo slik ali prilagoditvijo arhitekture nevronske mreže.

### **8.1. Kaj se zgodi če model ne more prepoznati galaksije**

Ko model strojnega učenja naleti na galaksijo, ki je ne more natančno klasificirati v obstoječe kategorije, jo avtomatsko razvrsti kot "neopredeljeno". Ta pristop omogoča, da model ohrani natančnost in zanesljivost pri razvrščanju galaksij, ki jih lahko prepozna, medtem ko za tiste, ki predstavljajo izziv zaradi njihove »unikatnosti« ali zaradi pomanjkanja jasnih morfoloških značilnosti, vzpostavi posebno kategorijo. Takšna kategorizacija kot "neopredeljeno" služi kot varnostni mehanizem, ki preprečuje napačno klasifikacijo galaksij in omogoča nadaljnjo analizo in raziskavo teh posebnih primerov. S tem model zagotavlja, da ostanejo vse galaksije ustrezno obravnavane.



## **9. INTERPRETACIJA REZULTATOV**

### **9.1. Interpretacija rezultatov:**

- Optimizacija podatkovne zbirke: razvili smo funkcijo za enostavno nalaganje podatkovne zbirke, kar izboljšuje dostopnost in uporabnost podatkov.
- Dosežena natančnost: model je dosegel 75% natančnost pri klasifikaciji galaksij, kar kaže na zmožnost razlikovanja med tipi galaksij.
- Učinkovitost pri manjših količinah podatkov: model deluje učinkovito tudi z manjšimi količinami podatkov (25 GB od 100 GB)
- Vpliv strojne opreme na učenje: opažanja kažejo na pomembnost izbire ustrezne strojne opreme za optimizacijo časa učenja.
- Optimizacija vhodnih podatkov: ugotovili smo, da je optimizacija vhodnih slik ključna za izboljšanje učinkovitosti modela

### **9.2. Hipoteze**

Prvo hipotezo smo potrdili, saj je naš model z visoko natančnostjo razlikoval med spiralnimi in eliptičnimi galaksijami.

Drugo hipotezo smo ovrgli, ker naš model ni zmoženo natančno določiti prisotnosti subtilnih značilnosti galaksij, kot so prečke, obroči in plinski repi.

Tretjo hipotezo smo delno potrdili. Naš model je sposoben prepoznati anomalije galaksij, vendar njegova metoda prepoznavanja temelji na nezmožnosti odločitve o klasični morfološki kategorizaciji, kar lahko vključuje širok spekter anomalij, ne le tiste, ki so izrecno povezane z neobičajnimi galaktičnimi pojavi.

### **9.3. Zanimivosti**

- Naš dataset vsebuje 314.000 že klasificiranih galaksij, kar priča o obsežnosti in raznolikosti astronomskega opazovanja.
- Klasifikacijo galaksij v podatkovni zbirki so ročno opravili prostovoljci.
- Zaradi lažje uporabe smo dataset obrezali na približno 25 GB. Uporabili smo dataset s približno 100.000 slikami galaksij, kar omogoča učinkovitejše obdelovanje podatkov.

- Vse slike v datasetu so enotnih dimenzij 424 x 424 pikslov, kar zagotavlja konsistentnost pri obdelavi in analizi.
- V datasetu se lahko pojavljajo ponovitve istih galaksij na različnih slikah, vendar nobena slika ni popolnoma identična drugi, kar omogoča bogatejšo analizo morfoloških značilnosti.
- Bitna globina slik je 24 bitov, kar zagotavlja visoko kakovost in natančnost pri prikazu barv in detajlov galaksij.

## **10. ZAKLJUČEK**

V zaključku naše raziskovalne naloge smo ugotovili, da model strojnega učenja odpira nova obzorja v avtomatizaciji klasifikacije galaksij, hkrati pa izpostavlja potrebo po nadaljnjih izboljšavah in optimizaciji. V prihodnje bomo delali na zmanjšanju dimenzij vhodnih slik in prehodu na črno-belo sliko za hitrejše procesiranje. Raziskovali bomo tudi, kako različni spektri, kot je infrardeči, vplivajo na modelovo sposobnost detekcije morfoloških lastnosti galaksij. Naš program je dostopen na odprtokodnem portalu GitHub, kar spodbuja sodelovanje in nadaljnji razvoj na tem področju. [GitHub Repository povezava: <https://github.com/MZNM-Raziskovalna-naloga-KGUSU/raziskovalnanaloga-kgusu>]

## **11. DRUŽBENA ODGOVORNOST**

Naša raziskovalna naloga prispeva k družbeni odgovornosti z razvojem programa za avtomatizirano klasifikacijo galaksij, ki naslavlja potrebe znanstvene skupnosti in splošne javnosti. Z dosedanjim odkritjem približno 100 milijonov galaksij in potencialom za nadaljnje odkritja, naš program obeta revolucijo v načinu, kako pristopamo k analizi kozmičnih podatkov. Omogoča znanstvenikom, da prihranijo dragoceni čas in sredstva, ki so potrebni za ročno klasifikacijo, in hkrati odpira možnosti za boljše razumevanje vesolja.

Poleg tega je naš projekt zasnovan kot odprtokoden, kar zagotavlja, da je dostopen širši skupnosti raziskovalcev, razvijalcev in navdušencev nad astronomijo. Ta odprtost spodbuja sodelovanje, inovacije in nadaljnji razvoj orodij, ki lahko prispevajo k še hitrejšemu in učinkovitejšemu razumevanju vesolja.

## 12. VIRI IN LITERATURA

### 12.1. Knjige in članki

1. Géron, A. (2nd Edition). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.
2. Walmsley, M., Lintott, C., Géron, T., Kruk, S., Krawczyk, C., Willett, K. W., Bamford, S., Kelvin, L. S., Fortson, L., Gal, Y., Keel, W., Masters, K. L., Mehta, V., Simmons, B. D., Smethurst, R., Smith, L., Baeten, E. M., Macmillan, C. (30 September 2021). "Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies." Monthly Notices of the Royal Astronomical Society, Volume 509, Issue 3, January 2022, Pages 3966–3988. Dostopno na: <https://academic.oup.com/mnras/article/509/3/3966/6378289>
3. Zwitter, T. (2002). Pot skozi vesolje. Ljubljana: Modrijan.
4. Moore, P. (1999). Atlas vesolja. Prevod: Tomaž in Savina Zwitter. Ljubljana: Mladinska knjiga.
5. Ploj B. (2019) Bionska umetna inteligenca: Knjiga o napravah, ki se učijo in se samostojno odločajo, pri čemer lahko presežejo tudi človeštvo. Visoka šola na Ptuju.
6. Hope T., Resheff Y., Lieder I. (2017) Learning TensorFlow: A Guide to Building Deep Learning Systems. O'Reilly Inc.

### 12.2. Spletni viri

1. Gharat, S., Dandawate, Y. "Galaxy Classification: A deep learning approach for classifying Sloan Digital Sky Survey images." (4. 2. 2024) Dostopno na: <https://arxiv.org/ftp/arxiv/papers/2211/2211.00397.pdf>
2. Walmsley, M. et al. (2023). "Zoobot: Adaptable Deep Learning Models for Galaxy Morphology." Journal of Open Source Software, 8(85), 5312. (4. 2. 2024) Dostopno na: <https://joss.theoj.org/papers/10.21105/joss.05312>

3. kaggle\_galaxy\_zoo GitHub. (15. 11. 2023) Dostopno na: [https://github.com/jameslawlor/kaggle\\_galaxy\\_zoo/tree/master](https://github.com/jameslawlor/kaggle_galaxy_zoo/tree/master)
4. Wikipedia: Machine Learning. (15. 11. 2023) Dostopno na: [https://en.wikipedia.org/wiki/Machine\\_learning](https://en.wikipedia.org/wiki/Machine_learning)
5. TensorFlow. (30. 11. 2023) Dostopno na: <https://www.tensorflow.org/>
6. Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J. (September 2008). "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey." Monthly Notices of the Royal Astronomical Society, Volume 389, Issue 3, pp. 1179-1189. (16. 12. 2023) Dostopno na: <https://ui.adsabs.harvard.edu/abs/2008MNRAS.389.1179L/abstract>
7. Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C., Raddick, M. J., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J. (January 2011). "Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies." Monthly Notices of the Royal Astronomical Society, Volume 410, Issue 1, pp. 166-178. (7. 2. 2024) Dostopno na: <https://ui.adsabs.harvard.edu/abs/2011MNRAS.410..166L/abstract>
8. The Dark Energy Camera Legacy Survey (DECaLS). (1. 11. 2023) Dostopno na: <https://www.legacysurvey.org/decamls/>
9. Galaxy Zoo DECaLS: Detailed Visual Morphology Measurements from Volunteers and Deep Learning for 314,000 Galaxies. (5. 2. 2024) Dostopno na: <https://zenodo.org/records/4573248>
10. Zoobot GitHub rep. (1. 11. 2023) Dostopno na: <https://github.com/mwalmsley/zoobot>

11. OpenLLM (5. 2. 2024)Dostopno na: <https://github.com/bentoml/OpenLLM>

### 12.3. Viri slik

1. Slika - Dostopno na:

[https://upload.wikimedia.org/wikipedia/commons/thumb/8/8a/Hubble\\_sequence\\_photo.png/250px-Hubble\\_sequence\\_photo.png](https://upload.wikimedia.org/wikipedia/commons/thumb/8/8a/Hubble_sequence_photo.png/250px-Hubble_sequence_photo.png)

2. Slika - Lasten vir

3. Slika - Dostopno na: <https://astronn.readthedocs.io/en/latest/galaxy10.html>

4. Slika - Dostopno na: <https://astronn.readthedocs.io/en/latest/galaxy10.html>

5. Slika - Dostopno na: <https://astronn.readthedocs.io/en/latest/galaxy10.html>

6. Slika - Dostopno na: <https://astronn.readthedocs.io/en/latest/galaxy10.html>

7. Slika - Dostopno na: <https://astronn.readthedocs.io/en/latest/galaxy10.html>

8. Slika - Dostopno na: <https://astronn.readthedocs.io/en/latest/galaxy10.html>

9. Slika - Lasten vir

10. Slika - Lasten vir

### **13. PRILOGE**

GitHub Repozitorij z vso kodo, ki je bila uporabljena v projektu. Dostopno na:

<https://github.com/korentmaj/Classification-of-celestial-bodies-using-machine-learning>