

UNIVERZITET U NOVOM SADU PRIRODNO-MATEMATIČKI FAKULTET DEPARTMAN ZA FIZIKU



## Potraga za aksionima na eksperimentu ATLAS korišćenjem metoda mašinskog učenja

- master rad -

*Mentori:* dr Nenad Vranješ prof. dr Jovana Nikolov

Student: Olivera Vujinović

Novi Sad, Septembar 2019.

Veliku i iskrenu zahvalnost dugujem svojim mentorima, dr Nenadu Vranješu i prof. dr Jovani Nikolov, na podršci, savetima i strpljenju koje su mi pružili tokom pisanja rada. Ovaj rad je realizovan u okviru Laboratorije za fiziku visokih energija Instituta za fiziku u Beogradu. U izradi ovog rada značajnu ulogu imala je i grupa sa Univerziteta Johannes Gutenberg u Majncu, koju vodi prof. dr Matthias Schott, a koja mi je pružila ogromno znanje i poverenje.

Zahvaljujem se i svojoj porodici, kao i prijateljima, na ljubavi, razumevanju i uvek prisutnoj bezuslovnoj podršci.

## Sadržaj

U	Uvod 4				
1	Pot	raga za	a aksionima	5	
	1.1	Naruše	enje CP simetrije u jakim interakcijama	5	
		1.1.1	QCD lagranžijan	5	
		1.1.2	P, C i T-simetrija	6	
		1.1.3	Efekti narušenja CP simetrije	8	
		1.1.4	Peccei-Quinn teorija	10	
	1.2	Dosada	ašnja potraga za aksionima na Velikom sudaraču hadrona u CERN-u	11	
		1.2.1	Veliki sudarač hadrona	11	
		1.2.2	Potrage za česticama poput aksiona	13	
<b>2</b>	Ma	šinsko 1	učenje u fizici visokih energija	16	
	2.1	Šta je 1	mašinsko učenje?	16	
		2.1.1	Algoritam mašinskog učenja	19	
		2.1.2	Vrste mašinskog učenja	20	
	2.2	Metode	e mašinskog učenja	21	
		2.2.1	Algoritam $k$ najbližih suseda	22	
		2.2.2	Stabla odlučivanja i Random Forest algoritam	23	
		2.2.3	Algoritam potpornih vektora	24	
	2.3	Veštač	ke neuronske mreže	25	
		2.3.1	Struktura i princip rada veštačke neuronske mreže	25	
		2.3.2	Funkcije aktivacije	27	
		2.3.3	Cross entropija i gubitak pri treningu	32	
3	AT]	LAS de	etektor i identifikacija fotona	33	
	3.1	ATLAS	S detektor	33	
		3.1.1	Unutrašnji detektor	35	
		3.1.2	Kalorimetri	35	
		3.1.3	Mionski spektrometar	37	
		3.1.4	Triger	37	
	3.2	Identifi	ikacija fotona na eksperimentu ATLAS	38	
		3.2.1	Rekonstrukcija fotona	38	
		3.2.2	Identifikacija fotona	38	

<b>4</b>	$\mathbf{Rez}$	Rezultati i diskusija 41				
	4.1	Korišć	eni Monte Karlo uzorci	41		
		4.1.1	Uzorak $h \to Za$	41		
		4.1.2	Uzorak $h \to Z\gamma$	45		
		4.1.3	Uzorak $\pi^0 \to \gamma \gamma$	48		
	4.2	Impler	nentacija veštačke neuronske mreže	51		
		4.2.1	Priprema i analiza podataka iz uzoraka	51		
		4.2.2	Podela podataka na trening, validacioni i test skup podataka	55		
		4.2.3	Treniranje modela	55		
		4.2.4	Zaključni plotovi	56		
	4.3	Prime	na metoda mašinskog učenja	57		
Zaključak						
Bi	3ibliografija 6					

## Uvod

Tokom poslednjih nekoliko godina, mašinsko učenje i njegove metode sve više dobijaju na značaju. Njihova primena zalazi u mnogobrojne naučne, ali i druge, discipline. U eri koja je nastupila nakon otkrića Higsovog bozona, iskorišćenje potencijala Velikog sudarača hadrona, kao i njegove unapređene verzije visoke luminoznosti - HL-LHC, postavlja se za novi cilj. Time se očekuje otvaranje sasvim novih prozora u oblast Nove fizike, će sada biti dostupno mnogo više podataka nego ranije. Međutim, porast količine podataka usled povećanja integralne luminoznosti 20 puta, postaviće neke nove izazove pred stručnjake u oblasti fizike visokih energija. Pre svega, misli se na veličinu događaja, zapreminu podataka, ali i na njihovu kompleksnost. Sve to upućuje na izvesnu ograničenost sa kojom će dosadašnji kompjuterski softver, korišćen u fizici, morati da se izbori. Upravo ovde na scenu stupa mašinsko učenje, čije tehnike obećavaju napredak na polju rešavanja datih problema u fizici čestica.[1]

Eksperimentalna fizika visokih energija svodi se na dve glavne stavke [1] - ispitivanje Standardnog modela merenjima visoke preciznosti i potraga za novim česticama koja se često svrstava u fiziku izvan Standardnog modela. Oba ova zadatka zahtevaju identifikaciju retkih signala, što će biti znatno otežano povećanim *pile-up*-om usled dodatnih sudarajućih protona na HL-LHC.

Svrha ovog rada jeste upravo ispitivanje uspešnosti različitih algoritama mašinskog učenja u potrazi za aksionima, pri čemu je centralni deo rezultata zasnovan na modelu veštačke neuronske mreže. Rad je podeljen u nekoliko celina, pri čemu se prve dve bave teorijom aksiona kao hipotetičkih čestica i mašinskog učenja. Zatim je dat kratak opis ATLAS detektora, kao i osnovne varijable korišćene u identifikaciji fotona koji potiču od aksiona. Na samom kraju, prikazani su korišćeni uzorci, kao i dobijeni rezultati, uz koje je data i odgovarajuća diskusija.

## Glava 1

## Potraga za aksionima

## 1.1 Narušenje CP simetrije u jakim interakcijama

Jedna od još uvek nerešenih misterija u fizici jeste upravo očuvanje CP-simetrije u kvantnoj hromodinamici (QCD<sup>1</sup>). Naime, tokom sedamdesetih godina 20. veka, primećeno je da bi jaka interakcija teorijski mogla da krši simetriju koja bi predstavljala kombinovane operaciju prostorne inverzije i konjugaciju naboja [2]. Sa druge strane, eksperimenti su pokazali da se CP-simetrija ipak održava u procesima jake interakcije. Međutim, ostaje pitanje zašto je to tako. Radi lakšeg razumevanja, biće objašnjene osnovne karakteristike lagranžijana u kvantnoj elektrodinamici (QED<sup>2</sup>) i kvantnoj hromodinamici, kao i definisane P, C i T simetrije.

#### 1.1.1 QCD lagranžijan

Veličina iz koje se izvode jednačine kretanja u kvantnoj teoriji polja jeste lagranžijan. Usled postojeće analogije između QCD i QED, paralelno će biti date osobine oba lagranžijana, kao i posledice njhovih oblika na održanje određenih simetrija [2]. U QED, lagranžijan je dat kao:

$$\mathcal{L}_{QED} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \bar{\psi} (i\gamma^{\mu} D_{\mu} - m)\psi, \qquad (1.1)$$

gde prvi član karakteriše polje foton<br/>a $A_{\mu}(x)$ iz kog proizilazi tenzor snage elektromagnetnog polja

$$F_{\mu\nu} = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu}. \tag{1.2}$$

Drugi član u lagranžijanu sadrži polje fermiona  $\psi(x)$  mase m. Tu se takođe javlja i kovarijantni izvod

$$D_{\mu} = \partial_{\mu} + ieQA_{\mu},\tag{1.3}$$

gde je e jačina sprezanja u QED, dok je Q broj koji je karakteristika polja  $\psi$  i prema tome zavisi od vrste datog polja.

<sup>&</sup>lt;sup>1</sup>eng. Quantum Chromodynamics

<sup>&</sup>lt;sup>2</sup>eng. Quantum Electrodynamics

Prateći prikazani primer za QED, može se napisati i lagranžijan za QCD. Jaka interakcija se odvija između polja kvarkova q(x) i gluona  $G^a_{\mu}(x)$ . Polja kvarkova mogu imati tri naboja boje - crveni, zeleni i plavi, stoga ih zapisujemo na sledeći način:

$$q = \begin{pmatrix} q_R \\ q_G \\ q_B \end{pmatrix} \tag{1.4}$$

Kvarkovi predstavljaju fundamentalnu reprezentaciju SU(3) grupe. Za polja gluona, potrebna je i jačina sprezanja jake interakcije g, kao i konstante strukture SU(3) simetrije  $f_{abc}$  koje su neophodne prilikom definisanja tenzora snage gejdž polja:

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g f_{bca} G^b_\mu G^c_\nu.$$
(1.5)

Jasno je da se tenzori polja u QED i QCD razlikuju samo u jednom dodatnom članu koji se ovde javlja. Takođe, može se primetiti da ovde postoji još jedan indeks. Naime, za razliku od QED slučaja gde postoji jedan foton, u QCD postoji 8 gluona koje broji dodatni indeks. Sada se može konačno napisati i QCD lagranžijan:

$$\mathcal{L}_{QCD} = -\frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a + \bar{q}_j (i\gamma^\mu D_\mu - m_j) q_j, \qquad (1.6)$$

gde indeks j ide po svim ukusima kvarkova i gde je  $D_{\mu}$  kovarijantni izvod definisan preko generatora SU(3) simetrije  $T_a$ :

$$D_{\mu} = \partial_{\mu} + igT_a \mathcal{G}^a_{\mu}. \tag{1.7}$$

Generatori se obično definišu ka<br/>o $T_a=\frac{1}{2}\lambda_a,$ gde su  $\lambda_a$  Gel-Manove matrice<sup>3</sup>.

#### 1.1.2 P, C i T-simetrija

**P-simetrija.** Ova simetrija označava invarijantnost na operaciju parnosti, pod kojom se pre svega misli na transformaciju ogledalske refleksije prostornih koordinata (x, y, z) u odnosu na tri uzajamno normalne ravni [4]. To je dakle refleksija svake prostorne koordinate ponaosob u odnosu na koordinatni početak:

$$\begin{array}{l} x \to -x \\ y \to -y \\ z \to -z \end{array}$$

Za vektor  $\vec{r}$  to znači inverziju u odnosu na koordinatni početak:  $\vec{r} \to -\vec{r}$ . Osim vektoru položaja, ova transformacija će se desiti i vektoru impulsa, koji sadrži prvi izvod prostornih koordinata:  $\vec{p} \to -\vec{p}$ . Ovo neće važiti za pseudo, odnosno aksijalne vektore, kao što je spin (vektor momenta impulsa, odnosno angularnog momenta), koji predstavlja vektorski prozivod dve prethodno pomenute veličine. Spin  $\vec{L}$  ovom operacijom ne menja svoj znak.

 $<sup>^3 \</sup>rm Ove$ matrice je razvio Marej Gel-Man u proučavanju jake interakcije i one predstavljaju skup od 8 linearno nezavisnih 3  $\times$  3 ermitskih matrica.

Za procese koji su invarijantni na opisanu transformaciju - operaciju parnosti, kažemo da odlikuje P-simetrija (simetrija parnosti). Prilikom eksperimentalne provere, pokazano je da se parnost održava u slučaju procesa koji se odigravaju pod dejstvom elektromagnetne i jake interakcije, dok do narušenja P-simetrije dolazi kod beta raspada - procesu koji se odvija pod uticajem slabe interakcije.

Dakle, ako se radi o prostor-vremenu čiji vektor predstavljamo kao  $x^{\mu} = (t, \boldsymbol{x})$ , ova transformacija, obeležena slovom  $\mathcal{P}$  će uticati samo na prostorne koordinate, dok će vremenska ostati nepromenjena [2]. Transformisani vektor je dat kao

$$\tilde{x}^{\mu} = (t, -\boldsymbol{x}) \tag{1.8}$$

Da bi se dokazala invarijantnost lagranžijana na promenu parnosti, potrebno je naći transformacije koje će zadovoljiti relaciju:

$$\mathcal{PLP}^{-1} = \mathcal{L}(\tilde{x}), \tag{1.9}$$

gde treba uzeti u obzir da su sama polja takođe funkcije prostorno-vremenskih koordinata. Za transformacije polja kvarkova i gluona u QCD, koje će održati parnost, se dobija:

$$\mathcal{P}q(x)\mathcal{P}^{-1} = \eta_P \gamma_0 q(\tilde{x})$$
  

$$\mathcal{P}\bar{q}(x)\mathcal{P}^{-1} = \eta_P \bar{q}(\tilde{x})\gamma_0$$
  

$$\mathcal{P}G_0(x)\mathcal{P}^{-1} = G_0(\tilde{x})$$
  

$$\mathcal{P}G_i(x)\mathcal{P}^{-1} = -G_i(\tilde{x}),$$
  
(1.10)

gde je  $\eta_P$  parametar koji može uzimati vrednosti  $\eta_P = \pm 1$ . Dokazivanje invarijantnosti lagranžijana se svodi na zamenu svih polja u relaciji (1.6) novim oblicima dobijenim primenom transformacija (1.10). U slučaju invarijantnosti, dobio bi se isti rezultat kao i kada bi se svako x zamenilo sa  $\tilde{x}$  i kada bi se takođe promenio znak na odgovarajućim mestima isperd članova sa izvodima prostornih koordinata.

**C-simetrija.** Ova simetrija označava invarijatnost na konjugaciju naboja, odnosno konjugaciju čestica-antičestica. Njome se ne menjaju prostorne koordinate, vreme i spin, već električnog naboja, barionskog broja, leptonskog broja i svih aditivnih kvantnih brojeva koji karakterišu datu česticu [4]. Kod slabe interakcije se narušava i C-simetrija (elektroni iz beta raspada imaju prvenstveno levu polarizaciju - spin i impuls su antiparalelni, dok pozitroni iz naboj-konjugovanih raspada imaju prvenstveno desnu polarizaciju - spin i impuls paralelni).

Transformacije koje moraju da važe za polja gluona i kvarkova da bi QCD lagranžijan bio invarijantan na konjugaciju naboja (operacija obeležena sa C) su date kao

$$C\bar{q}qC^{-1} = \bar{q}q$$

$$C\bar{q}\gamma_{\mu}\frac{\lambda_{a}}{2}qC^{-1} = -\eta(a)\bar{q}\gamma_{\mu}\frac{\lambda_{a}}{2}q$$

$$CG^{a}_{\mu}C^{-1} = -\eta(a)G^{a}_{\mu},$$
(1.11)

gde je

$$\eta(a) \begin{cases} +1 & \text{ako je } a = 1, 3, 4, 6, 8\\ -1 & \text{ako je } a = 2, 5, 7. \end{cases}$$
(1.12)

Koeficijent  $\eta(a)$  je posledica SU(3) strukture kvantne hromodinamike. Da bi se dobila transformaciona prafila za kvantnu elektrodinamiku, potrebno bi bilo zanemariti  $\eta(a)$  koeficijente, kao i generatore SU(3) simetrije,  $\frac{1}{2}\lambda_a$ .

**T-simetrija.** Treća simetrija jeste T-simetrija koja se odnosi na invarijantnost procesa na vremensku inverziju. To znači da promena smera vremena  $t \to -t$  neće uticati na koordinate. Invarijantnost na promenu smera vremena zahteva jednake amplitude verovatnoće direktnog i inverznog procesa [4]. Operacija vremenske inverzije je obeležena sa  $\mathcal{T}$  i njen efekat na vektor u prostor-vremenu  $x_{\mu} = (t, \boldsymbol{x})$  [2] se može zapisati kao:

$$-\tilde{x}_{\mu} = (\boldsymbol{x}). \tag{1.13}$$

Lagranžijan će biti invarijantan na ovu transformaciju ako važi

$$\mathcal{TL}(x)\mathcal{T}^{-1} = \mathcal{L}(-\tilde{x}). \tag{1.14}$$

#### 1.1.3 Efekti narušenja CP simetrije

Dobija se da je QCD lagranžijan invarijantan u odnosu na C, P i T transformacije ponaosob. Međutim, nakon detaljnijeg izvođenja, uz korišćenje svih prethodno navedenih pravila i definicija, dolazi se do relacije koja potiče od polja gluona (pri čemu indeksi 0, i, j, k sada označavaju prostorno-vremenske komponente):

$$G\tilde{G}_{\text{cubic}} = -2g\epsilon^{0ijk}f_{bca}\left(\partial_0 \mathbf{G}_i^a\right)\mathbf{G}_j^b\mathbf{G}_k^c - (\partial_i \mathbf{G}_0^a)\mathbf{G}_j^b\mathbf{G}_k^c - (\partial_j \mathbf{G}_i^a)\mathbf{G}_0^b\mathbf{G}_k^c - (\partial_k \mathbf{G}_0^a)\mathbf{G}_j^b\mathbf{G}_0^c\right).$$
(1.15)

Direktnom primenom transformacije parnosti na ovaj izraz, sva polja bi doživela promenu  $G^a_{\mu}(x) \rightarrow G^a_{\mu}(\tilde{x})$ . Takođe, poslednja tri člana bi promenila znak, jer sadrže prostorne izvode. Međutim, pravila za transformaciju polja gluona iz (1.10) neće dati isti rezultat, jer bi u slučaju njihove primene došlo do promene znakova sva četiri člana. Odavde se zaključuje da je parnost narušena. Na isti način se može pokazati da dolazi i do narušenja T-simetrije, dok se prema pravilima transformacije u slučaju konjugacije naboja iz (1.11) ispostavlja da QCD lagranžijan ne krši C-simetriju. Stoga, kombinacijom, CP simetrija jeste narušena.

Prilikom pomenutog izvođenja, polazilo se od nešto drugačijih relacija za QED i QCD lagranžijane - zapisani su preko dualnih oblika tenzora snage odgovarajućih polja:

$$\tilde{\mathcal{L}}_{Q\mathcal{ED}} = \alpha F_{\mu\nu} \tilde{F}^{\mu\nu} 
\tilde{\mathcal{L}}_{Q\mathcal{CD}} = \alpha G^a_{\mu\nu} \tilde{G}^{\mu\nu}_a,$$
(1.16)

gde  $\alpha$  predstavlja numeričku konstantu. Redefinisanjem te konstante nakon utvrđivanja invarijantnosti i neinvarijantnosti QCD lagranžijana na različite operacije, dobija se:

$$\tilde{\mathcal{L}}_{QCD} = \frac{\theta_{QCD}g^2}{32\pi^2} G^a_{\mu\nu} \tilde{G}^{\mu\nu}_a, \qquad (1.17)$$

gde je sada uveden parametar  $\theta_{QCD}$ .

Može se pokazati da uvođenjem još jedne transformacije (aksijalne transformacije) na sve ukuse kvarkova, QCD lagranžijan dobija još jedan član, istog oblika kao i u (1.17):

$$\frac{g^3\beta}{16\pi^2}G^a_{\mu\nu}\tilde{G^{\mu\nu}}.$$
(1.18)

Pošto opisana transformacija daje U(1) grupu, može se reći da QCD ima  $U(1)_A$  anomaliju. Pošto se ovaj elektroslabi efekat smatra QCD doprinosom, može se dodati već poznatom parametru  $\theta_{QCD}$ , čime se dobija efektivni  $\theta$  parametar

$$\theta_{\text{eff}} = \theta_{QCD} + \arg \det \left( M^{(u)} M^{(d)} \right), \tag{1.19}$$

gde su  $M^{(u)}$  i  $M^{(d)}$  matrice masa kvarkova koji pripadaju različitim generacijama (da bi se dobili članovi koji sadrže oba kvarka iz iste generacije, ove matrice moraju da budu dijagonalizovane).

**Eksperimentalni rezultati.** Vršena su precizna merenja električnog dipolnog momenta neutrona [2], koji se dobija iz izraza

$$d_N \sim \frac{e\theta_{\text{eff}} m_\pi^2}{m_N^3},\tag{1.20}$$

gde je e konstanta elektromagnetne interakcije,  $m_{\pi}$  masa piona i  $m_N$  masa neutrona. Dakle, ukoliko bi kod jake interakcije došlo do narušenja CP simetrije, došlo bi do nastanka električnog dipola kod neutrona za koji je izmerena gornja granica od  $|d_N| < 2.9 \cdot 10^{-26}$  e · cm sa nivoom pouzdanosti od 90% [3].



Slika 1.1: Doprinos narušenja CP simetrije električnom dipolnom momentu [2]

Za gornju granicu vrednosti $\theta_{\rm eff}$ se dobija

$$|\theta_{\text{eff}}| \le \mathcal{O}(10^{-10}),\tag{1.21}$$

što je izuzetno mala vrednost, s obzirom na to da ne postoji očigledan razlog da ta vrednost ne bude  $\theta_{\text{eff}} \sim \mathcal{O}(1)$  [2]. Iz eksperimentalnog rezultata, moglo bi se zaključiti da je  $\theta_{\text{eff}}$  jako malo usled određenog poništavanja dva člana u relaciji (1.19), što je veoma malo verovatno, sudeći po tome da potiču iz različitih sektora Standardnog modela. Jedno od mogućih objašnjenja bilo bi ukoliko barem jedan od kvarkova ne bi imao masu, jer bi u tom slučaju mogla da se izvrši hiralna rotacija polja kvarkova za iznos dovoljan da se poništi  $\theta$  član. Međutim, eksperimentalno je potvrđeno da kvarkovi imaju masu, te se time ova opcija isključuje.

#### 1.1.4 *Peccei-Quinn* teorija

Prema ovoj teoriji, QCD lagranžijanu se dodaje još jedan član koji sadrži i novu skalarnu česticu nazvanu aksion:

$$\mathcal{L}_{a} = \xi \frac{a}{f_{a}} \frac{g^{2}}{32\pi^{2}} G^{a}_{\mu\nu} \tilde{G}^{\mu\nu}_{a}, \qquad (1.22)$$

gde je  $\xi$  koeficijent koji zavisi od modela, a  $f_a$  konstanta raspada aksiona. Može se primetiti sličnost između ovog i  $G\tilde{G}$  člana iz 1.17, pa ih stoga možemo sjediniti u novi član koji će opisivati efektivan potencijal polja aksiona:

$$\mathcal{L}_{G\tilde{G}} = \left(\theta_{\text{eff}} + \xi \frac{a}{f_a}\right) \frac{g^2}{32\pi^2} G^a_{\mu\nu} \tilde{G}^{\mu\nu}_a. \tag{1.23}$$

Za minimum ovog polja, očekivana vakuumska vrednost aksiona iznosi

$$\langle a \rangle = -\frac{\theta_{\text{eff}} f_a}{\xi},\tag{1.24}$$

za koju čitav član nestaje. Ukoliko se fizički aksion zapiše kao  $a_{\text{phys}} = a - \langle a \rangle$ , za  $G\tilde{G}$  član se dobija

$$\mathcal{L}_a = \xi \frac{a_{\text{phys}}}{f_a} \frac{g^2}{32\pi^2} G^a_{\mu\nu} \tilde{G}^{\mu\nu}_a. \tag{1.25}$$

Na ovaj način parametar  $\theta_{\text{eff}}$  se potpuno gubi iz lagranžijana i CP-simetrija će biti održana u jakim interakcijama. U neku ruku, ovo rešenje asocira parametar  $\theta_{\text{eff}}$  sa ne statičnim, već dinamičnim poljem, čije su ekscitacije oko nulte vrednosti predstavljene aksionom.



Slika 1.2: Doprinos narušenja CP simetrije električnom dipolnom momentu [2]

Većina eksperimenata koji se bave potragom za aksionima zapravo istražuju mogućnost njegovog raspada na dva fotona (prikazano na slici 1.2). Međutim, da bi vreme njegovog života bilo duže od starosti univerzuma, njegova masa bi trebalo da iznosi  $m_a \leq 20$  eV [2]. To znači da, ukoliko mu je masa dovoljno mala, aksioni bi se možda još uvek nalazili u univerzumu. To takođe otvara i pitanje tamne materije čije bi oni bili komponente.

## 1.2 Dosadašnja potraga za aksionima na Velikom sudaraču hadrona u CERN-u

#### 1.2.1 Veliki sudarač hadrona

Prvi akceleratori čestica u CERN-u su proradili 1957. godine, odnosno 1959. Najsnažniji akcelerator danas je Veliki hadronski sudarač - LHC (eng. Large Hadron Collider). U njemu se ubrzavaju protoni, i ima obim 27 km. Smešten je u kružnom tunelu na oko 100 m ispod zemlje. Dva protonska snopa se kreću u suprotnim smerovima u ultra-visokom vakuumu kroz kružne cevi, i nakon dostizanja odgovarajuće energije oni se sudaraju, pri čemu dolazi do kreiranja ogromnog broja različitih čestica. Istraživački rad LHC-a je počeo 2010. godine, nakon perioda testiranja. Započeo je rad sa energijama protona od 7 TeV, što znači da je na raspolaganju bilo 7 TeV po svakom paru sudarajućih protona. Prvi radni ciklus je završio 2013. godine. Nakon servisiranja, novi ciklus je započeo 2015. godine, sa energijama od 13 TeV. Prednost kolajderskih eksperimenata u odnosu na sudare čestica sa fiksnom metom je ta da se može iskoristiti celokupna energija zbira kinetičkih energija sudarajućih čestica za kreiranje novih. Ukupan impuls dve čestice koje se kreću jedna ka drugoj je jednak nuli, a onda i ukupan impuls nastalih čestica mora biti takav nakon sudara.

Visoka kinetička energija u LHC-u se ostvaruje primenom električnih polja u akceleratoru na pojedinim delovima putanje protona, putem, tzv. radio-frekventnih šupljina [4, 5], tako da proton prilikom svakog prolaska kroz ovakav segment dobija određeni iznos energije, dok magnetna polja duž putanje obezbeđuju preciznost njihove kružne putanje upotrebom magnetnih dipola i fokusiranjem snopova, pomoću magnetnih kvadrupola. Tako jaka magnetna polja obezbeđuju superprovodni elektromagneti koji veoma efikasno provode elektricitet bez otpora ili gubitka energije. Da bi se proizvela jaka magnetna polja, potrebno je ostvariti protok veoma jakih struja kroz pomenute superprovodnike. Struja jačine 11 850 A protiče kroz dipole kako bi se kreirala magnetna polja indukcije 8.33 T. Da bi ovo bilo moguće, neophodno je ohladiti magnete do izuzetno niskih temperatura, čineći njihovo okruženje hladnijim od svemira - one iznose i do -271.3 °C. Iz tog razloga, veći deo akceleratora je povezan sa sistemom koji distribuira hladni helijum, kojim je omogućeno hlađenje magneta. Maksimalna energija koju protoni konačno dostižu u LHC-u (7 TeV po svakom protonu u jednom snopu) ostvaruje se postupno, prethodnim ubrazavanjima u manjim akceleratorima velikog akceleratorskog kompleksa u CERN-u.

Važan aspekt koji se mora razmotriti prilikom konstrukcije akceleratora kod kojih se čestice kreću po kružnoj putanji jeste emisija sinhrotronskog zračenja. Dakle, ovaj pojam se generalno odnosi na elektromagnetno zračenje koje se emituje kada se naelektrisane čestice kreću kružnom ili nekom drugom zakrivljenom putanjom. Tada se vektor brzine menja po pravcu, što predstavlja ubrzano kretanje. Ova vrsta zračenja predstavlja energijski gubitak za čestice, što dalje znači da se dodatna energija mora obezbediti od strane akceleratora, kako bi se energija snopa održala konstantnom. Međutim, u kružnim akceleratorima, kao što je LHC, teške čestice poput protona imaju mnogo manji energijski gubitak putem sinhrotronskog zračenja nego lake čestice kao što su elektroni.

Cilj protonskih sudara u LHC-u je reprodukcija i istraživanje uslova koji su vladali u ranom kosmosu:  $10^{-12}$  s posle Velikog praska, kada je odgovarajuća temperatura kosmosa bila  $\approx 10^{16}$  K.

Protoni su u snopovima tokom kretanja po obimu LHC-a grupisani u "pakete", gde u jednom paketu ima  $\approx 10^{11}$  protona. Luminoznost svakog snopa iznosi  $10^{34}$  protona/(cm<sup>2</sup>s), dok se paketi protona sudaraju svakih 25 ns. Pri tome se odigrava oko 600 miliona protonproton sudara svake sekunde, dok se po jednoj sekundi selektuje samo oko 100 događaja radi čuvanja podataka i naknadne analize. Inače, protoni se kreću u ultra-visokom vakuumu koji iznosi  $\approx 10^{-9}$  Pa.

Najveći tehnički izazov za dizajniranje LHC-a predstavljali su magnetni dipoli, postavljeni duž kružne putanje protona, tj. po obimu snopova, čiji je cilj da jakim magnetnim poljima savijaju snop protona energije 7 TeV i održe ih na kružnoj putanji. LHC dipoli koriste niobijumtitanijum provodnike (kablove) koji postaju superprovodni na temperaturama ispod 10 K. Ustvari, ovi dipoli bivaju ohlađeni pomoću tečnog helijuma na temperaturu od 1.9 K. Može se primetiti da je ovo temperatura niža od današnje temperature kosmosa (2.7 K), pa je interesantna činjenica da su to najhladnija područja u kosmosu.



Slika 1.3: CERN-ov akceleratorski kompleks [6]

Na mestima koja su predviđena za interakciju snopova (mesta sudara protona) nalaze se moćni detektori: CMS (*Compact Muon Solenoid*) i ATLAS (*A Toroidal LHC ApparatuS*) čiji je zadatak da registruju putanje novoformiranih čestica i mere njihovu energiju, kako bi mogao da se utvrdi njihov identitet. Osim navedenih detektora, operativna su još 4 manja detektora specijalne namene. Detektori CMS i ATLAS imaju približno cilindričan oblik i veoma složenu, slojevitu unutrašnju strukturu, a unutar njih se generišu jaka magnetna polja i tako utiče na putanje kreiranih čestica kroz detektore. Na osnovu karakteristika putanje u magnetnom polju može se odrediti impuls čestice. Detektori predstavljaju kombinaciju velikog broja pojedinačnih manjih detektora koji čine aktivne slojeve i pasivnih slojeva materijala velike gustine (kao što su gvožđe i olovo), gde čestice deponuju deo svoje energije. Upravo se iznos deponovane energije u pojedinim apsorbujućim slojevima koristi da bi se utvrdila energija kojo su čestice posedovale.

Veliko otkriće od strane kolaboracija CMS i ATLAS usledilo je u julu 2012. godine, kada

je objavljeno otkriće Higsovog bozona, što je predstavljalo veliki trijumf fizike visokih energija i izvrsnu potvrdu da su teorijska predviđanja korektna. Postoje neka značajna pitanja na koja se čekaju odgovori, a tiču se supersimetrije, ekstra dimenzija, tamne materije i tamne energije, narušenja simetrije između materije i antimaterije itd. Veliki izazov kod analize sudarnih procesa je uočavanje nekolicine događaja za kojima tragamo među milijardama drugih. Razvijena je posebna, snažna računarska mreža za preuzimanje i čuvanje ogromne količine prethodno filtriranih podataka, kako bi se oni mogli naknadno analizirati.

Znanja stečena pri realizaciji ovog velikog eksperimenta i tehnologije koje su razvijene pri rešavanju brojnih problema, nalaze svoju primenu u najrazličitijim oblastima ljudske delatnosti, i doprinose opštem napretku civilizacije.

#### 1.2.2 Potrage za česticama poput aksiona

Sve češće se javljaju pitanja konceptualne i praktične prirode sa svrhom provere i usavršavanja Standardnog modela u fizici čestica [7]. Da bi se pronašli odgovori na njih, neophodno je zaći u tzv. fiziku izvan Standardnog modela, koja postaje glavna tema u većini naučnih razgovora. Vodeći princip koji Standardni model zastupa kaže da je neophodno da su svi procesi koje vidimo konzistentni sa odgovarajućim simetrijama. Jedan takav proces generiše električni dipolni moment neutrona. Međutim, kao što smo ranije videli, eksperimenti su pokazali da je on barem 10 milijardi puta manji od predviđene vrednosti. Jedina okolnost koja bi opravdala ovaj slučaj jeste postojanje nekakvog, do sada neuočenog, mehanizma koji je Standardni model prevideo. Kako se ovo dešava samo u kvantnoj hromodinamici, čitav problem se naziva problemom narušenja CP simetrije u jakim interakcijama.

Videli smo da se ovaj problem može rešiti uvođenjem nove hipotetičke čestice, kvantnohromodinamičkog aksiona, koji je obuhvaćen u znatno složenijoj *Peccei-Quinn* teoriji [7]. Pored aksiona, postoje i čestice poput njih (ALP<sup>4</sup>) koje su zapravo hipotetički laki pseudo-Nambu-Goldstounovi bozoni<sup>5</sup>. Dakle, čestice poput aksiona ne rešavaju nužno problem narušenja CP simetrije u jakim interakcijama i pojavljuju se prilikom spontanog narušenja globalne simetrije. One mogu interagovati sa svim česticama Standardnog modela, dok su njihove mase i jačine sprezanja teorijski slobodni parametri čije vrednosti se mogu prostirati duž velikog raspona.

Potraga za ovim česticama zavisi pre svega od njihove mase, kao i od njihovog sprezanja sa česticama Standardnog modela. Ukoliko imaju masu čija je vrednost niža od dvostruke mase elektrona, ALP-ovi se mogu raspadati jedino na fotone, sa odgovarajućom stopom raspada koja ima vrednost jednaku trećem stepenu mase [7]. To znači da bi laki ALP-ovi trebalo da budu dugoživeće čestice koje prelaze velike udaljenosti pre raspada. Eksperimenti koji odgovaraju ovom slučaju su oni koji tragaju za dugoživećim ALP-ovima koji bi poticali od Sunca i drugih udaljenih objekata. Dodatno ograničenje se javlja usled merenja dužine praska neutrina poteklih iz Supernove SN1987a, koja bi bila kraća da je postojao i gubitak energije usled emisije ALP-ova. U prilog ovome ide i nedostatak eksperimentalnog zapažanja praska fotona iz SN1987s nastalih raspadom emitovanih ALP-ova. Takođe, postoji i niz kosmoloških ograničenja koja potiču od proučavanja nukleosinteze u Velikom prasku, kosmičkog mikrota-

<sup>&</sup>lt;sup>4</sup>eng. Axion-Like Particles

<sup>&</sup>lt;sup>5</sup>Nambu-Goldstounovi bozoni su bozoni koji nastaju prilikom spontanog narušenja kontinualnih simetrija i ne poseduju spin. Ukoliko su Nambu-Goldstounovi bozoni nastali spontanim narušenjem simetrije, oni su bezmaseni; međutim, ukoliko su nastali spontanim i ujedno eksplicitnim narušenjem simetrije, tada imaju relativno malu masu i nazivaju se pseudo-Nambu-Goldstounovim bozonima.

lasnog pozadinskog i vangalaktičkog pozadinskog zračenja. Time je isključena velika oblast u prostoru definisanom masom i jačinom sprezanja ALP-ova. Za ALP-ovima mase ispod GeV-a tragaju eksperimenti bazirani na modelu snop-meta i osetljivi su na ALP-ove nastale iz visokoenergijskih fotona prilikom interakcije upadnog snopa i jezgara mete (Primakofov efekat), nakon čega se ALP-ovi raspadaju ponovo na fotone van mete. Predložen je ShiP <sup>6</sup> eksperiment koji bi pokrio istraživanja ALP-ova mase u rangu od 1 MeV do 1 GeV, ali i aktuelne potrage na NA62 eksperimentu takođe mogu biti od koristi u datom rangu. Oba su u sklopu CERN-a. U potrazi za ALP-ovima mase od nekoliko MeV-a pa do nekoliko stotine GeV-a, eksperimenti na sudaračima čestica postaju relevantni. Ukoliko je ova čestica za kojom se traga dugoživeća, ostavila bi signaturu sa velikom nedostajućom energijom u detektoru, i za ovakvim signalima bi tragali BaBar, CLEO, LEP Tevatron i LHC. ALP-ovi veće mase i/ili sa većim jačinama sprezanja postaju kratkoživeće čestice i mogu se raspasti unutar samog detektora. Ovim se otvaraju sasvim nove mogućnosti za date potrage na LHC-u i nekim budućim sudaračima. U zavisnosti od mase, ALP se može raspasti na dva fotona, leptona ili džet.

Na samom Velikom hadronskom sudaraču, ALP-ovi se mogu proizvesti u procesima tipa  $pp \rightarrow a$ , ili prilikom egzotičnih raspada čestica poput Z ili Higsovog bozona [7]. Neki od relevantnih procesa bili bi  $Z \rightarrow \gamma a$ ,  $h \rightarrow aa$ . Nakon njenog nastanka, postoje dve mogućnosti - u slučaju da je ALP dugoživeća čestica, napustila bi detektor u vidu signatura nedostajuće energije, dok bi za kratkoživeće ALP-ove ATLAS i CMS eksperimenti mogli da rekonstruišu odgovarajuće produkte raspada. U poslednjem slučaju, laki ALP-ovi se mogu raspadati jedino na dva fotona. Za mase dvostruko veće od mase elektrona, javlja se i mogućnost raspada na dva elektrona, pri čemu je slična situacija i za slučaj sa dva miona ili tau leptona. Masa ALP-a bi morala da bude veća od trostruke mase piona kako bi došlo do hadronskih raspada na dva džeta. Pomenuti modovi raspada se mogu istraživati na LHC-u.



Slika 1.4: Ograničenja na ALP masu i jačinu sprezanja za ALP-foton interakciju dobijena na osnovu mnogobrojnih eksperimenata [7]

<sup>&</sup>lt;sup>6</sup>Search for Hidden Particles



Slika 1.5: Ograničenja na ALP masu i jačinu sprezanja za ALP-foton interakciju dobijena na osnovu mnogobrojnih eksperimenata, gde je uključena i oblast koja bi se dobila iz raspada Higsa  $h \to Za \to l^+ l^- \gamma \gamma$  [8]. Na levoj slici je prikazan rezultat dobijen iz podataka prikupljenih na LHC-u prilikom Run 2 faze sa integralnom luminoznošću od 300 fb<sup>-1</sup> (prikazano svetlo zelenom oblašću), pri čemu nam je potrebno barem 100 događaja sa izmerenim signalom. Desna slika pokazuje tamnom zelenom regijom oblast koja je isključena dosadašnjim potragama za  $h \to Z\gamma$ .



Slika 1.6: Ograničenja na ALP masu i jačinu sprezanja za ALP-foton interakciju dobijena na osnovu mnogobrojnih eksperimenata, gde je uključena i oblast koja bi se dobila iz raspada Higsa  $h \to aa \to 4\gamma$  [8]. Na levoj slici je prikazan rezultat dobijen iz podataka prikupljenih na LHC-u prilikom Run 2 faze sa integralnom luminoznošću od 300 fb<sup>-1</sup> (prikazano svetlo zelenom oblašću), pri čemu nam je potrebno barem 100 događaja sa izmerenim signalom. Desna slika pokazuje tamnom zelenom regijom oblast koja je isključena dosadašnjim potragama za  $h \to \gamma\gamma$  i  $h \to 4\gamma$ .

## Glava 2

## Mašinsko učenje u fizici visokih energija

Danas, kompjuteri mogu da učine za nas, i umesto nas, mnogo više stvari nego ikada pre. Osim što rešavaju neke trivijalne zadatke za koje nemamo vremena, imaju i potencijal da zaštite naše podatke, pa tako i nas same, mnogo bolje nego što bismo to sami učinili. U ovom odeljku će biti data definicija mašinskog učenja, kao i pregled nekih njegovih osnovnih metoda.

## 2.1 Šta je mašinsko učenje?

Kao ustaljena definicija mašinskog učenja navodi se citat Artura Semjuela iz 1959. godine, prema kome je ono "naučno polje koje kompjuterima daje mogućnost da uče, bez prethodnog eksplicitnog programiranja" [16]. Kao najprostiji odgovor na pitanje *kako?*, daje se činjenica da je neophodno da algoritam mašinskog učenja dobije što više konkretnih i odgovarajućih podataka kako bi uspeo da reši konkretan i odgovarajući problem. To bi opet značilo da je na neki način neophodno programiranje. Međutim, proces poboljšanja performanse algoritma se zasniva na istom mehanizmu na kome i čovekov proces učenja - koriste se podaci kao ulazne informacije i porede sa željenim izlazom, a zatim se koriguju sve dok se ne dobije željeni rezultat. Što više podataka, odnosno iskustva, kompjuter sretne, to će bolje raditi zadati problem [16]. Jedan od primera koji bi ilustrovao ovo jeste naš telefon, koji "pamti" poruke koje pišemo i ukoliko ih ima dovoljno, može na osnovu njih da predvidi ustaljeni rečnik kojim se služimo, na osnovu kog nam daje svoje predloge, pa tako ubrzava kucanje poruka sa manje grešaka.

U nešto širem naučnom smislu, mašinsko učenje predstavlja podskup u polju veštačke inteligencije i veoma usko je povezano sa matematikom i statistikom.

**Primena mašinskog učenja.** Stručnjaku koji radi sa podacima je neophodna klasifikacija, za šta je najbolje primeniti neku od metoda mašinskog učenja. Veoma često se to može primeniti u najraznolikijim oblastima i zadacima, kao što su:

- pronalazak polja sa velikom količinom nafte, zlata ili arheoloških značajnosti;
- pretraživanje teksta radi pronalaženja imena mesta ili osobe;
- identifikacija ljudi na osnovu fotografija ili glasovnih snimaka;

- identifikacija ptica na osnovu njihovog cvrkuta;
- identifikacija profitabilnih mušterija;
- identifikacija delova automobila za koje postoji velika verovatnoća da će se pokvariti;
- prepoznavanje tumora i bolesti;
- predikcija količine novca koju bi osoba potrošila na dati proizvod;
- predviđanje broja vulkanskih erupcija u određenom periodu;
- predviđanje godišnjih prihoda date kompanije;
- predviđanje tima koji bi pobedio u datom sportskom takmičenju.

Ponekad bi naučnici koji rade sa podacima izgradili model, koji bi u tom trenutku bio apstraktan i ne bi savršeno odgovarao realnosti, ali bi im dao uvid u procese na kojima se proučavani fenomen zasniva. Kada model nije usmeren ka predikciji već razumevanju nekog problema, onda se kaže da se radi o analizi uzroka, odnosno korena problema <sup>1</sup>[16]. Nekoliko primera ovakvog tipa istraživanja su:

- Razumevanje i optimizacija poslovnog plana, kao i određivanje proizvoda koji doprinose vrednosti celokupnoj liniji proizvoda;
- Određivanje uzroka dijabetesa;
- Određivanje uzroka saobraćajnih gužvi.

Ova lista mogućih primena mašinkog učenja je samo mali deo onoga što mašinsko učenje zaista može - naime, ono je sveprisutno u nauci kao i radu sa podacima. Dva osnovna problema kojima se ono bavi jesu regresija i klasifikacija, pa se tako svaki problem koji želimo da rešimo uz pomoć mašinskog učenja mora odrediti i svrstati u jednu od ove dve mogućnosti. Zatim se pristupa problemu odgovarajućim tehnikama, definisanim na osnovu toga da li je u pitanju regresija ili klasifikacija<sup>2</sup>.

Kao što je već pomenuto, mašinsko učenje se koristi u naučnom procesu istraživanja podataka, koji se može podeliti na neke fundamentalne korake [16]:

1. Definisanje istraživačkog cilja - u ovoj fazi se imenuje predmet istraživanja i definišu korist za kompaniju, tip potrebnih podataka, tip konačnog rezultata i rokovi. Takođe, savetuje se i formiranje *project charter*-a, kako bi se efikasnije prikazao plan rada i proverio napredak. U ovoj fazi je generalno važniji rad sa ljudima, jer se tu dele zadaci, procenjuju tehničke veštine, ali i prezentuje klijentima plan, dokaz o uspešnosti i cena samog projekta.

<sup>&</sup>lt;sup>1</sup>eng. root cause analysis

<sup>&</sup>lt;sup>2</sup>Problemi sa kojima se surećemo u mašinskom učenju se svode na dva tipa - regresione i klasifikacione probleme. Suštinski, razlikuju se u tome što klasifikacioni predviđaju klasu kojoj ulazni podatak pripada, dok regresioni predviđaju kvantitativnu veličinu za odgovarajući ulazni podatka.

- 2. Skupljanje podataka iako kompanije uglavnom imaju svoje podatke sa kojim dolaze kod tima stručanjaka, često se dešava da postoji nekoliko različitih izvora podataka koje je potrebno uskladiti. Tu se pre svega misli na proveru njihove dostupnosti, konzistentnosti i formata. Osim toga, danas je veliki problem i privatnost podataka, na šta se mora obratiti posebna pažnja.
- 3. Priprema podataka s obzirom na to da se na dosadašnje korake potroši i do 80% vremena, a da se do modeliranja ni ne dođe, ovaj korak se može raščlaniti na odgovarajuće faze:
  - (a) čišćenje podataka uklanjaju se greške nastale pri unosu, višak razmaka na početku ili kraju vrednosti (što se može veoma lako otkloniti primenom komande strip() u Python-u), velika/mala slova, nerealne vrednosti, *outlier*-i<sup>3</sup>, nedostajuće vrednosti, odstupanja od predviđenih vrednosti i različite merne jedinice.
  - (b) ispravljanje podataka najbolje bi bilo ispravljati podatke prilikom samog unosa, međutim to često nije moguće, te se tada prepravke vrše u samom kodu. Iz tog razloga bi bilo dobro čuvati kopiju podataka u inicijalnom stanju pre prečišćavanja.
  - (c) kombinovanje različitih skupova podataka u slučaju tabelarnih podataka postoje dva načina za to: *joining* spajanje skupova na osnovu zajedničkih obeležja i *appending* dodavanje jednog skupa na kraj drugog.
  - (d) upotreba materijalizovanih pogleda umesto fizičkog spajanja tabela na taj način se koristi manje prostora za skladištenje jer se podaci ne dupliraju. Mana ovog postupka jeste što se svaki put prilikom provere ili upita ono mora iznova kreirati, dok se prilikom spajanja tabela to vrši samo jednom.
  - (e) dodavanje novih obeležja na osnovu agregiranih vrednosti često nam je potrebno da dodamo ,,kolonu" koja bi sadržala nove informacije koje smo dobili izračunavanjem iz već postojećih podataka (zbir, procenat, udeo, itd)
  - (f) transformisanje podataka podrazumeva se da podaci moraju biti u odgovarajućem formatu kako bi bili pogodni za modelovanje. To znači da bi bilo potrebno izbaciti atribute sa visokom korelacijom, primeniti logaritam na nelinearne zavisnosti ili pretvoriti orginalne vrednosti u *dummy* vrednosti (neme varijable).
- 4. Istraživanje podataka u ovoj fazi, služimo se raznim tehnikama kako bismo bolje razumeli podatke, pa tako i problem sa kojim se susrećemo. Tu se pre svega misli na vizualizaciju podataka pomoću raznih tipova grafika, preseka i ostalih tehnika.
- 5. Kreiranje modela sam model je srž rešenja problema, i on je taj koji treba da obezbedi predikciju ili klasifikaciju novih elemenata na osnovu postojećih. Tu do izražaja dolazi primena mašinskog učenja, *data mining*-a i statistike. Sastoji se iz izbora tehnike za modelovanje i promenljivih koje ulaze u izgradnju modela, implementacije i izvršavanja modela, kao i analize dobijenih podataka.
- 6. Prezentacija rezultata i automatizacija analize podataka kada se dobiju rezultati, potrebno ih je na odgovarajući način vizualizovati, kako bi se klijentu lakše prikazala uspešnost

 $<sup>^{3}\</sup>mathrm{U}$ stati<br/>tistici, outlieroznačava vrednosti koje se značajno razlikuju od ostalih izmer<br/>enih vrednosti.

modela. Osim toga, neophodno je i izvršiti automatizaciju čitavog procesa, kako bi u u budućnosti čitav isti postupak mogao biti ponovljen sa drugim vrednostima.

### 2.1.1 Algoritam mašinskog učenja

Tom Mičel je 1997. godine u svojoj knjizi dao jednu veoma praktičnu definiciju mašinskog učenja [18] - "Mašinsko učenje je proces u kome kompjuterski program uči iz iskustva E koje odgovara klasi zadataka T i meri performanse P, ukoliko se njegove performanse P u rešavanju zadataka T povećavaju sa iskustvom E." Razlozi zbog kojih često pribegavamo mašinskom učenju su pre svega to da rešavanje problema zahteva postojanje mnogo pravila, nepostojanje tradicionalnog načina za rešavanje veoma složenog problema i često menjanje ulaznih parametara.

Napokon dolazimo i do postupka kreiranja modela, gde pod terminom *model* smatramo pre svega funkciju kojom se modeluju zavisnosti između datih astributa. Postupak se može podeliti u 4 faze, pri čemu je svaka detaljnije objašnjena u sledećem tekstu.

**Izbor atributa.** Atributima nazivamo osobine<sup>4</sup> podataka, koje je ponekad potrebno modifikovati ili čak kombinovati više u jedan, ne bi li se dobio odgovarajući atribut, a sve sa svrhom što realnijeg opisa problema.



Slika 2.1: Vizualizacija podele skupa podataka na trening, validacioni i test podskup podataka

**Treniranje modela.** Ova faza započinje pokretanjem algoritma mašinskog učenja na odabranom trening skupu (ovo važi samo za nadgledano učenje, što će biti objašnjeno u daljem tekstu). Naravno, posebna pažnja se posvećuje i odabiru algoritma, koji zavisi od vrste problema, karakteristika skupa atributa, tipa atributa, stepena homogenosti tipova i opsega atributa, stepena korelacije između atributa, obima raspoloživih podataka, itd.

Validacija modela. Dobrim modelom se smatra onaj model koji ima veliku stopu tačnosti predviđanja na treniranom skupu ali i na potpuno novim podacima. Upravo na ovom prvom delu definicije se zasniva validacija, koja ima za ulogu proveru valjanosti modela na poznatim podacima. To znači da je isto tako neophodno odvojiti deo podataka koje ćemo smatrati nepoznatim, i koje ćemo iskoristiti za definitivno utvrđivanje valjanosti modela. Metrike koje se koriste za validaciju modela su sledeće:

(a) greške pri klasifikaciji (eng. *error classification rate*) - procenat podataka u validacionom skupu koje je algoritam pogrešno klasifikovao;

 $<sup>^4</sup>$ eng. features

(b) srednja standardna greška (standardna devijacija) - pokazuje koliko u proseku koliko elementi validacionog skupa odstupaju od aritmetičke sredine skupa. Koristi se kod regresije.

Pored ovoga, bitne su i strategije validacije, gde postoje dva osnovna tipa:

- (a) podela skupa podataka u određenom odnosu na trening, validacioni i test podskup;
- (b) unakrsna validacija skup podataka se deli na k delova i prilikom validacije se svaki deo koristi jednom kao test skup a ostali u treniranju. Treniranje i validacija se ponavljaju k puta, nakon čega se nađe srednja vrednost dobijenih rezultata.



Slika 2.2: Vizualizacija unakrsne validacije sa skupom podataka podeljenim na k podskupova [19]

**Predviđanje.** Ukoliko smo kreirali dobar model, možemo ga primeniti na neobeleženim podacima (to je naš test *dataset*) i tako dobiti predviđanje. Ovo je korak u kome se model primenjuje u realnim uslovima, i neophodno je kvantifikovati stepen njegove tačnosti.

#### 2.1.2 Vrste mašinskog učenja

Prema tipu primenjenog algoritma, ulaznih i izlaznih podataka, kao i tipu problema kome se pristupa, mašinsko učenje može biti nadgledano i nenadgledano. Ono što je takođe bitan faktor u ovoj podeli jeste i mera u kojoj se zahteva ljudska angažovanost.

Nadgledano učenje. Ovde postoji tačno definisan problem, kao i skup tehnika kojim se on može rešiti. Obezbeđeni su:

- skup ulaznih podataka;
- skup željenih/tačnih izlaznih vrednosti.

Ovo znači da za svaki obezbeđeni ulazni podatak, postoji definisan jedan željeni/tačan podatak. Zadatak kompjuterskog programa je da neobeleženom ulaznom podatku dodeli tačnu izlaznu vrednost, tako što će pronaći optimalnu funkciju koja mapira ulazne podatke na izlazne. Ta realna funkcija je nepoznata, pa se stoga funkcija koja se optimizuje naziva hipotezom h(x). Njen opšti oblik određuje model nadgledanog učenja, te se tako dobija odgovarajući prostor mogućih hipoteza (to npr. može biti linearnost). Osim toga, postoji i funkcija greške, tzv. loss funkcija. Ona meri odstupanje predviđenih i stvarnih vrednosti ciljne promenljive.

Uspešnost ovog algoritma se meri njegovom sposobnošću generalizacije, odnosno uspešnošću u radu sa novim podacima. U ovu grupu spadaju regresija i klasifikacija, pri čemu su rezultati dobijeni u ovom radu zasnovani na primeni klasifikacionih metoda.

Nenadgledano učenje. U ovom slučaju su za naš rad obezbeđeni:

- samo skup ulaznih podataka;
- nema informacija o željenoj izlaznoj vrednosti.

U ovoj situaciji pristupamo rešavanju zadatka, iako ne znamo šta to rešenje može biti. Problem sa kojim se srećemo se svodi na pronalazak strukture u dostupnim podacima. U neku ruku, to je grupisanje podataka, bez znanja o tome kakve sve grupe postoje. Primer koji bi to ilustrovao jeste određivanje konfekcijske veličine na osnovu visine i mase ljudi. Neki od algoritama koji bi se ovde mogli svrstati su klasterovanje primenom hijerarhije, vizualizacija, kao i smanjenje dimenzionalnosti.

**Polunadgledano učenje.** Ova vrsta mašinskog učenja predstavlja kombinaciju nadgledanog i nenadgledanog učenja. To znači da za manji deo ulaznih podataka  $x_i$  postoje obeležene izlazne vrednosti  $y_i$ , dok za veći deo one ne postoje. Cilj ovog pristupa jeste iskorišćavanje neobeleženih ulaznih podataka radi boljeg treniranja modela. On često pronalazi svoju primenu u praksi, jer nije neuobičajeno da podaci većinom budu neobeleženi.

**Učenje sa podrškom.** Kod ove vrste mašinskog učenja postoji određena interakcija između kompjuterskog programa i okruženja. On, dakle, određenim akcijama deluje na stanje svog okruženja, koje zatim povratno utiče na program, dajući mu povratne informacije koje mogu biti ili "nagrada" ili "kazna". Cilj ove metode jeste da program nauči na koji način deluje u datom okruženju tako da vremenom maksimizuje nagrade, odnosno minimizuje kazne. To znači da će u procesu učenja, program birati one akcije koje daju najveće "nagrade". Ovo nalazi čestu primenu u kompjuterskim igricama i robotici, dok se ne može koristiti u nekim previše složenim situacijama poput igranja partije šaha, gde bi bilo neophodno za svaku moguću poziciju odrediti najbolji mogući potez.

## 2.2 Metode mašinskog učenja

Ovde će biti date najčešće metode nadgledanog mašinskog učenja, koje su primenjene u dobijanju rezultata u ovom radu. Pored toga, biće opisane i osnovne mere uspešnosti klasifikatora.

#### 2.2.1 Algoritam k najbližih suseda

Ovo je jedan od najjednostavnijih algoritama mašinskog učenja, kako u primeni tako i za razumevanje. Jedan od zahteva jeste da su sve vrednosti atributa numeričke vrednosti, što znači da je one koje to nisu potrebno pretvoriti u numeričke odgovarajućim metodama kako bi se sa njima moglo raditi. Algoritam k najbližih suseda  $(KNN^5)$  se primenjuje prilikom rešavanja problema i klasifikacije i regresije.

Koraci u ovom algoritmu se svode na:

- 1. Za svaki element u skupu podataka potrebno je izračunati udaljenost od neoznačenog elementa;
- 2. Zatim se dobijena rastojanja sortiraju prema rastućem poretku;
- 3. Uzima se prvih k elemenata sa najmanjim rastojanjima;
- 4. Neoznačeni element će, prema tome, pripadati onoj klasi kojoj pripada većina odabranih k elemenata.

Prilikom računanja rastojanja se najčešće koristi Euklidsko rastojanje:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$
(2.1)

ili Menhetn nejednakost:

$$d(x,y) = \left|\sum_{i=1}^{n} (x_i - y_i)^2\right|.$$
(2.2)

Može se desiti da atributi imaju vrednosti u opsezima koji se međusobno znatno razlikuju, pri čemu se javlja problem u kome atributi sa većim opsegom imaju veći uticaj na rezultat. Tada na scenu stupa normalizacija, koja daje nove vrednosti u intervalu 0 < x < 1 na sledeći način:

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)}.$$
(2.3)

Prednosti ovog algoritma su u tome što daje prilično dobru tačnost predviđanja s obzirom na svoju jednostavnost, kao i to što *outlier*-i ne utiču na nju. Međutim, zahteva dosta memorije s obzirom na to da se radi o dosta računski zahtevnom algoritmu, što može biti problem samo po sebi.

<sup>5</sup>eng. K Nearest Neighbours

#### 2.2.2 Stabla odlučivanja i Random Forest algoritam

Stabla odlučivanja predstavljaju algoritam koji klasifikuje elemente u dve ili više klasa na osnovu njihovih atributa. Sastoje se iz dva osnovna segmenta - čvora i grana koje povezuju čvorove. Čvor pri vrhu nema roditelja, dok svi ostali imaju po tačno jednog roditelja. Čvorovi koji nemaju potomke nazivaju se listovima - to su sva moguća rešenja koja se mogu dobiti iz datog skupa podataka, te se listovi nazivaju još i čvorovima odgovora. Svi ostali čvorovi su zapravo čvorovi odluke. Osnovna ideja ovog algoritma može se prikazati sledećim koracima:

- 1. Odabir atributa za čvor stabla;
- 2. Kreitanje grana za svaku vrednost atributa;
- 3. Podela instanci na podskupove, tako da imamo po jedan podskup za svaku vrednost atributa;
- 4. Proces ponavljati za svaku granu;
- 5. Stati kada sve instance u datoj grani pripadaju istoj klasi ili ukoliko su potrošeni svi atributi za dalje razvrstavanje.

Kao cilj postavlja se kreiranje najmanjeg stabla sa što većom prediktivnom moći. Iz tog razloga se uvode određene metrike preuzete iz teorije informacija - entropija i informaciona dobit. Entropija predstavlja veličinu koja kvantifikuje stepe neizvesnosti po pitanju vrednosti neke slučajne promenljive ili po pitanju ishoda nekog slučajnog događaja. Informaciona dobit je veličina koja kvantifikuje smanjenje entropije, odnosno neizvesnosti, putem sticanja znanja o o slučajnoj promenljivoj, ili u našem slučaju - prema mogućnosti pristupa podacima. Entropija se definiše kao

$$H(S) = -\sum_{i=1}^{N} p_i \log p_i, \tag{2.4}$$

gde je S skup datih podataka, N broj mogućih klasa i  $p_i$  verovatnoća da dati element pripada *i*-toj klasi. Informaciona dobit<sup>6</sup> Gain(A, S) se definiše kao

$$Gain(A,S) = H(S) - H(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} H(S_j),$$
(2.5)

gde je v broj mogućih vrednosti atributa A, |S| ukupan broj instanci u skupu S,  $|S_j|$  broj instanci u skupu S koji imaju vrednost j za atribut A, dok je  $H(S_j)$  entropija podskupa instanci sa j-tom vrednošću atributa A. Na osnovu ovih definicija, može se reći da čvor sa najvećom informacionom dobiti postaje korenski čvor.

Prednosti ovog algoritma jesu mogućnosti grafičkog prikaza kao i jednostavne interpretacije. Može se primeniti i na klasifikacione i na regresione probleme, kao i u slučaju atributa sa nedostajućim vrednostima. Za njega nije prepreka ukoliko atributi, osim numeričkih imaju i kategoričke vrednosti, za razliku od algoritma k najbližih suseda. Neki od nedostataka su manja tačnost predikcije u odnosu na ostale algoritme mašinskog učenja, kao i sklonost ka *overfitting*-u, o čemu će biti reči kasnije.

<sup>&</sup>lt;sup>6</sup>eng. gain

**Random Forest algoritam.** Ovaj algoritam predstavlja klasifikator koji je sačinjen iz kolekcije međusobno nezavisnih stabala odlučivanja, pri čemu svako stablo predstavlja jedan glas u procesu većinskog donošenja odluke. Može se svesti na sledeće korake:

- 1. Izbor k slučajnih atributa iz skupa podataka;
- 2. Kreiranje stabla odlučivanja za jedan podskup podataka, prema već opisanom značenju;
- 3. Izbor broja željenih stabala N, nakon čega se ponavljaju prvi i drugi korak N puta;
- 4. U procesu predikcije, slučaj prolazi kroz sva stabla odlučivanja koja nezavisno klasifikuju instancu. Nakon toga se bira ona klasa koju je većina stabala odlučivanja predvidela.

Ovaj algoritam je dobar jer osim što je precizan, takođe je i efikasan kada se radi sa velikim skupom podataka i/ili velikim brojem atributa, jer se može "paralelizovati". Pored toga, njegovu tačnost ne umanjuju moguće nedostajuće vrednosti atributa. Takođe, problem *overfitting*-a se eliminiše nezavisnim odlučivanjem većeg broja stabala. Može se koristiti i za klasifikaciju i za regresiju.

#### 2.2.3 Algoritam potpornih vektora

Support Vector Machine (SVM), kako se još naziva, se zasniva na pronalaženju hiperravni koja razdvaja podatke koji pripadaju različitim klasama, pri čemu je glavni kriterijum što veća "margina" između klasa. Potporni vektori su zapravo podaci koji se nalaze najbliže ivici margine i oni su dovoljni za uspešnu klasifikaciju, čak i kada bi se svi ostali podaci uklonili, jer ne doprinose radu algoritma.

Da bi se pronašla razdvajajuća hiperravan, potrebno je pronaći njenu jednačinu, koja je data hipotezom:

$$h(x) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n,$$
(2.6)

odnosno:

$$h(x) = W^T X + w_0 = 0, (2.7)$$

gde su n b<br/>toj atributa, X vektor vrednosti osobina i W vektor normale na hiperravan razdvajanja Udaljenost ne<br/>obeleženog podatka  $X^{(a)}$  od hiperravni se dobija kao:

$$\operatorname{dist}(X^{(a)}, P) = \frac{|W^T X^{(a)} + w_0|}{||W||_2}.$$
(2.8)

Udaljenost potpornog vektora  $X^{(sv)}$  se računa kao:

$$\operatorname{dist}(X^{(sv)}, P) = \frac{|W^T X^{(sv)} + w_0|}{||W||_2} = \frac{1}{||W||_2}.$$
(2.9)

Pošto se potporni vektori nalaze sa obe strane hiperravni, širina margine će biti  $\frac{2}{||W||_2}$ . Tražimo da ova veličina bude što je veća moguća, što znači da je potrebno minimizovati  $||W||_2$ . Veličina koju je potrebno minimizovati se može zapisati kao:

$$||W||_2 = \sqrt{w_1^2 + w_2^2 + \dots + w_n^2}.$$
(2.10)

### 2.3 Veštačke neuronske mreže

Covečanstvo je oduvek sanjalo o izumima koji bi na neki način oponašali ljudski inteligenciju. Prvi nagoveštaj ove teže pokazao se oko 800 godina p.n.e. kada su u zapisima Starih Grka pomenuti "mehanički" ljudi koje su stvorili bogovi [20]. Tokom poslednjeg veka, tehnologija je napredovala toliko da je omogućila prelazak te ideje iz sna u stvarnost. Dakle, sada robote koji poseduju vid primitivne svesti povezujemo sa vrhuncem tehnologije i nauke, a ne magijom i bogovima. Stvaranje veštačke inteligencije nam je omogućilo kreiranje neverovatnih programa koji mogu daleko da nadmaše ljudske sposobnosti, u određenim aspektima. Ključni korak u opisanom napretku na polju veštačke inteligencije je, pored nekih svakodnevnih izuma poput *chat-bot*-ova, otkriće veštačkih neuronskih mreža koje su se pokazale veoma uspešnim u detekciji malignih tumora [20]. Pre ovoga, programeri su mogli da kreiraju regresione i klasifikacione modele, koji su u neku ruku ograničeni.

Neuronske mreže su upravo ono što njihov sam naziv i kaže - to su (složene) mreže sačinjene od neurona čija je svrha obrada informacija [20]. Da bi ih napravili, naučnici su se poslužili najboljom mogućom inspiracijom - ljudskim mozgom. Naš mozak funkcioniše tako što obrađuje informacije pomoću mreža neurona. Oni dobiju određeni *input*, obrade ga i zatim daju kao *output* električne signale neuronima sa kojima su povezani. Prateći ovaj složen proces, naučnici su uspeli da, imitacijom (Slika 2.3), veštački izgrade arhitekturu sličnu biološkoj.



Slika 2.3: Prikaz inspiracije veštaškog neurona biološkim

#### 2.3.1 Struktura i princip rada veštačke neuronske mreže

Veoma je bitno razlikovati strukturu prave neuronske mreže i veštačke, odnosno kompjuterske simulacije. Naime, ova druga predstavlja čist skup matematičkih jednačina, te koncept neuronske mreže više znači ljudima koji ih prave nego samim računarima.

Neuronska mreža se može sastojati od nekoliko pa i do milion neurona, koji zajedno čine mrežu izdeljenu na nekoliko slojeva [21]. Glavni delovi neuronske mreže su prema tome:

- 1. ulazni sloj;
- 2. sakriveni sloj/slojevi;

#### 3. izlazni sloj;

Svaki neuron je u njoj povezan sa oba sloja koja ga okružuju, osim onih koji se nalaze u početnom, odnosno ulaznom sloju. Oni su programirani tako da mogu da prihvataju informacije koje im se dodeljuju iz okoline, dok neuroni koji se nalaze na suprotnom kraju mreže izlaznom sloju, proizvode signal kao naučenu reakciju čitave mreže na ulaznu informaciju. Ova dva opisana sloja ograđuju treću celinu, koja se nalazi u sredini i može biti sačinjena iz jednog ili više skrivenih slojeva. To čini deo neuronske mreže koji je najviše odgovoran za učenje. Praktično, što više skrivenih slojeva u neuronskoj mreži ima, to će ona biti tačnija [20]. Svaki sloj je sačinjen iz neurona, odnosno čvorova - nodova, pri čemu je svaki nod povezan sa svakim nodom iz prethodnog i narednog sloja (Slika 2.4).



Slika 2.4: Struktura neuronske mreže [20]

Veze među nodovima se predstavljaju funkcijom koja kao nezavisno promenljivu uzima informaciju iz neurona prethodnog sloja, množi je *težinom* i na dobijeni broj dodaje *pristrasnost*, odnosno *bajas*<sup>7</sup> [20]. Uticaj jednog neurona na drugi se meri vrednošću koju poseduje težina što je težina veća, to je veći uticaj.

Što se tiče toka informacija kroz ovakav sistem, postoje dva načina. Jedan predstavlja neuronsku mrežu propagacije unapred (eng. *feedforward neural network*) [21] i predstavlja unos informacija u neuronsku mrežu putem ulaznog sloja, odakle se one prenose dalje kroz skrivene slojeve, aktivirajući njihove nodove, nakon čega konačno stižu i do izlaznog sloja. Dalje, nodovi dobijaju informaciju prema već opisanom računu iz nodova prethodnih slojeva. Svaki nod sabira sve ulazne vrednosti koje ovako dobija, i ukoliko je dobijena suma veća od neke granične vrednosti - praga, dati nod se aktivira i prenosi dalje pobuđivanje nodovima sa kojima je povezan. Ovaj tip neuronske mreže radi sa već utvrđenim vrednostima težina, te nam nije potrebno njihovo ispravljanje nakon dobijanja izlaznih vrednosti.

Drugi način toka informacija kroz mrežu jeste tzv. propagacija unazad (eng. backpropagating) [21]. Ovde veliku ulogu igra povratna informacija, slično kao i kod ljudi kada uče - ta povratna informacija im pokazuje da li nešto rade dobro ili loše. Dakle, kada se na izlaznom sloju dobije određena vrednost, ona se uporedi sa očekivanom nakon čega je potrebno ustanoviti da li je potrebna korekcija težina ili ne. Ukoliko jeste, prvo se ispravljaju težine bliže izlaznom sloju, krećući se odatle unazad sve do ulaznog sloja. Kada se ovaj ciklus ponovi više puta, naš algoritam uči i tako smanjuje odstupanje stvarnih od željenih rezultata, sve dok ne dođe do prihvatljivog poklapanja.

<sup>&</sup>lt;sup>7</sup>eng. bias

Mera greške se naziva *cost* funkcijom (funkcijom cene [21]) neuronske mreže i računa se na sledeći način: prvo se pronađe razlika između dobijene i željene vrednosti, ta razlika se kvadrira a potom se svi kvadrati sumiraju (Slika 2.5). Dakle, ona se dobija primenom metode najmanjih kvadrata.



Slika 2.5: Kvadrati razlike između dobijene i željene vrednosti [20]; crvena linija predstavlja predviđanja koja daje naša neuronska mreža, dok su plave tačke tačna (željena) predviđanja. Neuronska mreža potom računa udaljenosti linije od datih tačaka i kvadrira ih (zeleni kvadrati), da bi ih potom sabrala i dala vrednost *cost* funkcije.

#### 2.3.2 Funkcije aktivacije

Kada se govori o veštačkim neuronskim mrežama, često se postavlja pitanje odabira funkcija aktivacija, kao i velikih broj mogućih opcija. Kao što je već ranije pomenuto, veštački neuron (nod) računa otežanu sumu svojih ulaza, dodaje *bias* i procenjuje da li će reagovati ili ne. To je u suštini uloga funkcije aktivacije, međutim posmatrajmo jedan nod za početak [22], koji daje:

$$Y = \sum(weight * input) + bias.$$
(2.11)

Prema neuronu, Y može uzeti bilo koju vrednost od  $-\infty$  do  $\infty$ . Dakle, on ne određuje ni na koji način granice unutar kojih se ova vrednost može javiti. Međutim, kako sada odlučiti da li će neuron reagovati ili ne? Iz tog razloga na scenu stupaju funkcije aktivacije - one određuju da li će se dati neuron, kao i veze koje polaze od njega, aktivirati. Ispod će biti opisani neki od glavnih tipova ovih funkcija.

Step funkcija. Ova funkcija je verovatno prva koja bi nam pada na pamet, a ujedno i najjednostavnija. Dakle, ako je Y iznad neke određene vrednosti, neuron se proglašava aktiviranim

[22]. U suprotnom, on ne prosleđuje dalje informaciju. Njen izlaz je ili 1 (aktivirani neuron) ili 0 (neaktivirani neuron), i ovo radi odlično za slučaj jednog neurona.



Slika 2.6: Step funkcija

Postavlja se sledeće pitanje - šta se dešava kada je potrebno povezati više neurone u klase, koji bi radili po istom principu? Recimo, može se desiti da više od jednog neurona na izlazu da 1, te je na osnovu toga nemoguće odlučiti kojoj klasi bi *input* pripadao. Ono što bi rešilo ovaj problem situacija u kojoj bismo imali mrežu u kojoj bi se aktivirao samo jedan neuron, dok se ostali ne bi aktivirali. Međutim, ovakav sistem je znatno teže istrenirati da dovede do konvergencije ka jednom odgovoru. Zato bi bilo bolje ukoliko sama aktivacija ne bi bila binarna (ili 0 ili 1). Umesto toga, kao rezultat bismo dobijali da je neki neuron 50% aktiviran ili 20% aktiviran. Tada bi u slučaju aktivacije više neurona bilo moguće pronaći neuron koji ima najviši nivo aktivacije [22]. Problem nastaje kada bi više od jednog neurona bilo 100% aktivirano. Svakako, ovakva mreža je i dalje pogodnija za trening, pogotovo ako se uzme u obzir da je mala verovatnoća da se ovaj problem dogodi.

Linearna funkcija. Zaključuje se da je najbolja opcija funkcija koja ima proizvoljne vrednosti aktivacionog nivoa u određenim granicama [22], te bi sledeća funkcija po jednostavnosti bila linearna funkcija:

$$A(x) = cx. (2.12)$$

Dakle, nivo aktivacije bio bi proporcionalan ulaznoj vrednosti (otežanoj sumi). Tako, dobijamo čitav niz vrednosti, i u slučaju da je više neurona aktivirano bilo bi potrebno pronaći maksimum. To se vrši primenom gradijentnog spusta, o kome će biti reči kasnije. Ono što je bitno jeste da je potrebno izračunati izvod funkcije, a izvod linearne funkcije daje konstantu - c. To znači da gradijent neće imati vezu sa x. Ukoliko bi se desila neka greška, bilo bi potrebno primeniti propagaciju unazad. Međutim, svaka promena koja bi se uvela ne bi dovela do napretka jer ona ne bi zavisila od promene u ulazu  $\Delta x$ . Takođe, ukoliko bismo imali više slojeva, kao što je to obično slučaj sa neuronskim mrežama, i ukoliko se svaki aktivira linearnom funkcijom, konačna aktivacija poslednjeg sloja bila bi zapravo linearna funkcija ulaza iz prvog sloja [22]. Kaskadnim aktivacijama pomoću linearnih funkcija od prvog sloja do poslednjeg, izgubili smo mogućnost slaganja slojeva. Tako se čitava mreža može svesti na jedan sloj sa linearnom aktivacijom (kombinacija linearnih funkcija je i dalje linearna funkcija).

**Sigmoidna funkcija.** Ova funkcija izgledom podseća na *step* funkciju, s tim da je više glatka [22]. Razlog uvođenja sigmoidne funkcije je pre svega njena nelinearna priroda, pri čemu je kombinacija ovakvih funkcija takođe nelinearna funkcija. Ovo znači da nam se mogućnost ređanja slojeva vratila, kao i da smo se rešili isključivo binarne aktivacije. Pored toga, njen gradijent je takođe glatka funkcija.



Slika 2.7: Sigmoidna funkcija

Kao što se na Slici 2.7 može videti, za x u intervalu od -2 do 2, ova funkcija je veoma strma. To ima za posledicu dosta drastične promene u Y sa malim promenama u x, odnosno Y će veoma lako biti blizu jedne od dve krajnosti. Ova osobina je veoma poželjna kada se radi o klasifikaciji, jer obezbeđuje lakšu predikciju. Još jedna pogodnost ove funkcije aktivacije je i to što, za razliku od linearne, daje izlaz u intervalu (0, 1) a ne  $(-\infty, \infty)$ . Dakle, aktivacija je na neki način ograničena.

Problem sa sigmoidnom funkcijom je što na njenim krajevima Y veoma slabo reaguje na promene u x [22]. To znači da će gradijent u tim "skoro horizontalnim" oblastima krive biti veoma mali, ili će potpuno nestati. Dakle, više se neće javljati velike promene na izlazu sa malom promenom na ulazu, što dovodi do znatno sporijeg učenja neuronske mreže. Uprkos tome, sigmoidna funkcija je danas veoma rasprostranjena u upotrebi i prilično popularna u klasifikacionim problemima.

Hiperbolički tangens. Oblik ove funkcije je dat kao:

$$f(x) = tanh(x) = \frac{2}{1 + e^{-2x}} - 1.$$
(2.14)

Vidi se da je ova funkcija veoma slična sigmoidnoj [22], što znači da poseduje i svojstva slična njenim:

$$tanh(x) = 2 * sigmoid(2x) - 1.$$

$$(2.15)$$



Slika 2.8: tanh funkcija

Dakle, nelinearne je prirode, što nam omogućava ređanje slojeva. Isto tako, ograničena je na interval (-1, 1) pa je i aktivacija ograničena. Ono što je bitno jeste da je njen gradijent jači nego kod sigmoidne funkcije - njeni izvodi su strmiji. Prilikom odabira između ove dve funkcije, presudni kriterijum se uglavnom tiče željenog oblika izvoda. Međutim, tanh funkcija ima problem sa nestajanjem gradijenta, kao i sigmoidna funkcija [22]. To je svakako ne sprečava da bude jedna od najpopularnijih i najrasprostranjenijih funkcija kada je reč o funkcijama aktivacije.

**ReLu funkcija.** ReLu funkcija je data kao:

$$A(x) = \max(0, x).$$
 (2.16)

ReLu funkcija zapravo daje izlaz ukoliko x ima pozitivnu vrednost, a u suprotnom daje 0 [22]. Na prvi pogled se čini da će se javiti isti problemi kao kod linearne funkcije, s obzirom na to da je ReLu linearna na pozitivnom delu x-ose. Međutim, ReLu je nelinearne prirode, i njene kombinacije su takođe nelinearne - time je obezbeđeno slaganje slojeva. Međutim, ostaje problem ograničenja aktivacije, s obzirom na to da je interval vrednosti koje ova funkcija daje  $[0, \infty)$ . Jedna ključna prednost ove funkcije u odnosu na ostale jeste razređenost u aktivaciji [22]. U slučaju sigmoidne i tanh funkcije, skoro svi neuroni će se aktivirati u određenoj meri. Sve te aktivacije će dalje biti analizirane i iskorišćene u opisu izlaza neuronske mreže. Dakle, aktivacije će biti guste, što ima veliku cenu. Ono što bi bilo idealno jeste da se neki od neurona u mreži ne aktiviraju i tako omoguće "razređeniju" aktivaciju, što bi takođe doprinelo većoj efikasnosti neuronske mreže. Kod ReLu funkcije, to bi izgledalo ovako: početne težine bi imale nasumične vrednosti, i skoro 50% neuronske mreže će dati 0 prilikom aktivacije (0 je *output* za negativne vrednosti x); to bi značilo da će se aktivirati manje neurona (razređena aktivacija) i da je mreža postala lakša.



Slika 2.9: ReLu funkcija

Međutim, javlja se i problem - horizontalni deo ReLu funkcije (za negativno x) dovodi do toga da gradijent postaje 0. Dakle, za aktivacije u toj oblasti ReLu funkcije, gradijent će biti 0 što dalje znači da se težine neće prilagoditi prilikom procesa gradijentnog spusta. Ovo dovodi do toga da neuroni koji se nađu u takvom stanju prosto prestanu da reaguju na varijacije u odnosu greška - input (gradijent je 0, te se ništa ne menja). Ovaj problem se naziva *dying ReLu problem* [22], jer može da prouzrokuje smrt nekoliko neurona i time učini ključan deo mreže pasivnim. Postoji nekoliko rešenja za opisani problem. Jedno od njih jeste modifikacija horizontalnog dela ReLu funkcije, tako da on više ne bude horizontalan već da bude oblika y = 0.01x za x < 0 [22]. Takav oblik ReLu funkcije se naziva *leaky* ReLu. Naravno, postoje još neke opcije, ali je ideja svake sprečavanje gradijenta da bude 0.

ReLu funkcija je manje računski zahtevna od *tanh* i sigmoidne funkcije jer obuhvata jednostavnije matematičke operacije. Međutim, sigmoidna funkcija omogućava znatno brže učenje i treniranje u slučaju problema klasifikacije. Svakako, preporučuje se i korišćenje nekih drugih funkcija koje bi bile kombinacija navedenih, u cilju pronalaska optimalne funkcije aktivacije. Svi prikazani grafici dobijeni su pomoću softvera *Wolfram Alpha*.

#### 2.3.3 Cross entropija i gubitak pri treningu

Jedna od najčesćih funkcija aktivacije, koja nije opisana u prethodnom delu ali može dobro da posluži prilikom objašnjavanja ovog, jeste *Softmax* funkcija. Za nju je karakteristično to što pretvara brojeve, odnosno cifre, u verovatnoće čiji krajnji zbir daje 1. Ova funkcija kao *output* daje vektor koji predstavlja raspodelu verovatnoće u vidu liste mogućih rezultata. Njen oblik je dat kao

$$S(y_i) = \frac{\mathrm{e}^{y_i}}{\sum_j \mathrm{e}^{y_j}}.$$
(2.17)

Recimo da vektor raspodele verovatnoće koji model daje nakon primene *Softmax* funkcije daje sledeće verovatnoće: [0.7, 0.2, 0.1]. Vidimo da je njihova suma 100%, pri čemu je prva vrednost najverovatnija. Tačne vrednosti bi bile: [1,0,0]. Razlika između dva vektora se u linearnoj algebri računa skalarnim proizvodom, koji je jednak sumi proizvoda pojedinačnih elemenata vektora - prvi sa prvim, drugi sa drugim i treći sa trećim: 0.7\*1+0.2\*0+0.1\*0 = 0.7. *Softmax* funkcija je idealna slučaju klasifikacije, pri čemu postoji više od dve klase.

Još jedan način da se izmeri koliko je tačna predikcija dobijena *Softmax* funkcijom jeste cross entropija koja predstavlja prirodan način da se izmeri "rastojanje" između dva vektora verovatnoće i definisana je kao

$$D(S,L) = -\sum_{i} L_i \log(S_i).$$
(2.18)

 $S_i$  su elementi *output* vektora dobijenog *Softmax* funkcijom, dok su  $L_i$  elementi tačnog *outputa*:

$$S(Y) = \begin{bmatrix} 0.7\\ 0.2\\ 0.1 \end{bmatrix}; \ L = \begin{bmatrix} 1\\ 0\\ 0 \end{bmatrix}.$$
(2.19)

Krajnji *output* je dat u vidu *one-hot encoded* vektora, čiji svi elementi imaju vrednost 0, osim jednog koji ima vrednost 1 a koji upućuje na klasu kojoj instanca pripada. Vidi se da je log primenljiv samo na izlaz dobijen *Softmax* funkcijom, jer u slučaju tačnog *output-*a gde imamo puno nula on nije definisan. Dakle, prvo se ulaz množi težinom i dodaje se *bias* (Y = wx + b). Zatim se na njega primenjuje *Softmax* funkcija aktivacije i dobija S(Y). Na posletku, traži se *cross* entropija D(S(wx + b), L).

Ono što je ključno u procesu treniranja, a pomenuto je ranije, jeste gubitak pri treningu<sup>8</sup> koji predstavlja usrednjenu *cross* entropiju po čitavom trening setu. On je dat kao

$$\mathcal{L} = \frac{1}{N} \sum_{i} D(S(wx_i + b), L_i).$$
(2.20)

Odatle sledi funkcija datog gubitka<sup>9</sup>, čiji minimum je neophodno pronaći. Upravo na tome se zasniva proces učenja neuronske mreže.

<sup>8</sup>eng. training loss

<sup>9</sup>eng. loss function

## Glava 3

## ATLAS detektor i identifikacija fotona

## 3.1 ATLAS detektor

ATLAS je detektor namenjen za istraživanja u fizici elementarnih čestica, proveravajući predviđanja Standardnog modela koji obuhvata naše trenutno razumevanje gradivnih blokova materije i njihovu međusobnu interakciju. Njegov zadatak je da registruje sve čestice koje nastanu prilikom neelastičnih protonskih sudara na veoma visokoj energiji LHC-a i pri velikom broju istovremenih interakcija. Kako bi to postigao, mora posedovati sledeće karakteristike [9, 10]:

- Efikasno praćenje tragova čestica pri visokoj luminoznosti;
- Identifikacija elektrona i fotona elektromagnetnim kalorimetrom; identifikacija džetova i merenje nedostajuće transverzalne energije hadronskim kalorimetrom;
- Precizna merenja impulsa miona za koja je pri visokoj luminoznosti dovoljan samo mionski spektrometar;
- Velika pokrivenost oblasti prostornog ugla;
- Trigerovanje i merenja vezana za čestice sa niskim transverzalnim impulsom, što obezbeđuje visoku efikasnost za većinu fizičkih procesa.



Slika 3.1: Trodimenzionalni presek ATLAS detektora [9]

ATLAS detektor je dužine 44 m i prečnika 22 m, dok mu je masa oko 7000 tona (Slika 3.1). Čine ga sledeće komponente, konstruisane tako da zadovoljavaju prethodno navedene uslove: unutrašnji detektor, elektromagnetni i hadronski kalorimetarski sistem, mionski spektrometar i magnetni sistem. On takođe prati simetrično cilindričnu geometriju i pokriva ugao od skoro  $4\pi$  [12]. Pre svega, ovaj detektor se oslanja na desni koordinatni sistem sa svojim početkom u tački interakcije (IP<sup>1</sup>) koja se nalazi u središtu detektora, dok je z-osa usmerena duž pravca snopa. x-osa je usmerena od tačke interakcije ka centru prstena LHC-a, a y-osa je usmerena naviše. Cilindrične koordinate  $(r, \phi)$  određuju transverzalnu ravan, gde je  $\phi$  azimutalni ugao opisan oko cevi sa snopom. Takođe, definiše i tzv. pseudorapiditet preko polarnog ugla  $\theta$  kao

$$\eta = -\ln \tan(\theta/2). \tag{3.1}$$

Ugaono rastojanje se definiše kao

$$\Delta R = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}.$$
(3.2)

Još jedna od bitnih veličina kojom se opisuje kinematike procesa koji se odvijaju unutar detektorskog sistema jeste i transverzalni impuls (u ovom slučaju fotonskih kandidata) kao

$$E_T = E/\cosh(\eta),\tag{3.3}$$

gde je E energija posmatranog kandidata.

<sup>&</sup>lt;sup>1</sup>eng. *interaction point* 

#### 3.1.1 Unutrašnji detektor

Unutrašnji detektor (*Inner detector*) [11] je smešten najbliže tački interakcije. To je prvi deo ATLAS detektora koji ima kontakt sa produktima interakcije, te je neophodno da on bude veoma kompaktan i visoke osetljivosti. Sačinjen je iz tri različita sistema senzora (*Pixel Detector, Semiconductor Tracker* (SCT) i *Transition Radiation Tracker* (TRT)), koji se nalaze u magnetnom polju usmerenom paralelno osi snopa jačine 2 T. Cilindrične je geometrije, dužine od 7 m ograničene *end-cap* delovima kalorimetra, i radijusa 1.15 m određenog unutrašnjom površinom kriostata u kome se nalazi (*Liquid Argon*) elektromagnetni kalorimetar.



Slika 3.2: Trodimenzionalni prikaz unutrašnjeg detektora [11]

S obzirom na mesto na kome se nalazi unutrašnji detektor, njegovi delovi moraju da ispunjavaju uslov visoke granularnosti kako bi se obezbedila dobra rezolucija u merenjima. Iz tog razloga njega čine fini silikonski objekti (*silicon pixels*) koji leže uz samu osu snopa i silikonske tračice (*silicon strips*) - ova dva dela čine SCT. Pored toga, tu su i cevčice u spoljašnjoj oblasti unutrašnjeg detektora. Prostorna oblast unutrašnjeg detektora se, međutim, može podeliti na tri dela: *barrel* deo koji pokriva oko 80 cm radijusa oko mesta interakcije i dva *end-cap* dela koja se nalaze na krajevima detektora. Ono što je karakteristično za *barrel* deo jeste to što su detektorski slojevi raspoređeni cilindrično oko linije snopa, dok su u *end-cap* delovima oni poređani normalno na istu liniju. Unutrašnji detektor daje informacije o pravcu, impulsu i električnom naboju naelektrisanih čestica koje se proizvedu u svakom *pp* sudaru.

Sistem unutrašnjeg detektora obezbeđuje merenje položaja čestica u oblasti  $|\eta| < 2.5$  [12], što je omogućeno kombinujući informacije iz tri poddetektorska dela. Cilindrični centralni deo se nalazi oko same cevi sa snopom i pokriva  $|\eta| < 1.5$ . End-cap delovi se sastoje iz diskova postavljenih normalno na pravac snopa, pri čemu pokrivaju oblast  $1.5 < |\eta| < 2.5$ .

#### 3.1.2 Kalorimetri

ATLAS-ov kalorimetarski sistem se nalazi između unutrašnjeg detektora i mionskog sistema. Sastoji se iz:

1. elektromagnetnog kalorimetra i

#### 2. hadronskog kalorimetra.

**Elektromagnetni kalorimetar.** Podeljen je na *barrel* i *end-cap* deo. Po svojoj konstrukciji može se okarakterisati kao tečni *sampling* kalorimetar, sa tečnim argonom kao aktivnom sredinom i olovom kao apsorberom (LAr). Neke od glavnih odlika ovakvih kalorimetara su visoka uniformnost odgovora, jednostavna kalibracija i visoka radijaciona izdržljivost. Što se tiče geometrije ovog kalorimetra, slojevi aktivne sredine apsorbera su raspoređeni tako da formiraju oblik harmonike, čime je smanjeno vreme sakupljanja signala što je jedan od kritičnih parametara za rad u okruženju LHC-a. Takođe poseduje visoku granularnost.

Elektromagnetni kalorimetar meri energiju i položaj elektromagnetnih pljuskova unutar oblasti od  $|\eta| < 3.2$  [12]. Takođe je podeljen na *barrel*, koji pokriva oblast sa pseudorapiditetom od  $|\eta| < 1.475$ , i dva *end-cap* dela koji pokrivaju  $1.375 < |\eta| < 3.2$ . Oblast preklapanja između pokrivenosti ovih delova je  $1.37 < |\eta| < 1.52$  i ona se zapravo ne smatra oblašću visoko preciznog merenja.



Slika 3.3: Skica elektromagnetnog kalorimetra u ATLAS detektoru na kojoj su prikazani svi segmenti u vertikalnoj i horizonatlnoj ravni u oblasti  $\eta = 0$  [12]

Hadronski kalorimetar. Sastoji se iz tri dela - *barrel*, *end-cap* i *forward* regiona. Poslednja dva, kao i elektromagnetni kalorimetri, spadaju u *sampling* kalorimetre sa tečnim argonom kao aktivnom sredinom ali sa gvožđem kao apsorberom. *Barrel* oblast predstavlja scintilacioni kalorimetar sa gvozdenim pločama (TileCal), koje su orijentisane paralelno osi. Time se postiže longitudinalna i transverzalna segmentacija i omogućava dobra energetska rezolucija hadronskog kalorimetra. Dimenzije hadronskog kalorimetra su mnogo veće od dimenzija elektromagnetnog kako ne bi došlo do proboja hadronskih pljuskova u mionski sistem.

### 3.1.3 Mionski spektrometar

ATLAS-ov mionski sistem [13] je dizajniran tako da je u mogućnosti da obezbedi merenja impulsa miona potpuno nezavisno od unutrašnjeg detektora. Sastoji se iz dve oblasti - *barrel* i *end-cap*. Princip rada ovog dela detektora se zasniva na tome što superprovodni toroidni magneti sa vazdušnim jezgrom stvaraju jako magnetno polje, u kome mioni skreću, čime se omogućava merenje njihovog impulsa. Komponente čiji je to zadatak su MDT (*Monitored Drift Tubes*) i CSC (*Cathode Strip Chambers*) komorama), pri čemu CS komore imaju veću granularnost jer se nalaze bliže tački interakcije (Slika 3.4). Pored ovih merenja, ovaj sistem vrši nezavisno i trigerovanje događaja sa mionima pa se u te svrhe koriste sledeće komponente: TGC (*Thin Gap Chambers*) i RPC (*Resistive Plate Chambers*).



Slika 3.4: Trodimenzionalni prikaz ATLAS-ovog mionskog sistema [13]

### 3.1.4 Triger

S obzirom na ogromnu količinu podataka koja se proizvede prilikom sudara u LHC-u, od čega ne tako malo broj događaja se mora odbaciti, potreban je sistem koji će na osnovu određenih kriterijuma birati podatke koji će biti trajno upisani i potom korišćeni za analizu. Za opisanu selekciju događaja je zadužen ATLAS-ov triger i DAQ (*Data Acquisition*) sistem [14]. On tu selekciju događaja vrši u tri nivoa, pri čemu se poslednja dva mogu grupisati u HLT (*High Level Trigger*):

• L1 - početnu frekvencu događaja od 40 MHz svodi na 100 kHz;

- L2 sa 100 kHz na 1 kHz;
- EH sa 1 kHz na  $\sim$  100 događaja trajno zapisanih svake sekunde.

Neophodno je da vreme potrebno L1 trigeru za procesiranje bude što kraće zbog zahtevnog LHC okruženja i ono stoga iznosi 2.5  $\mu$ s. Upravo toliko vremena se informacija zadržava u *pipeline* memoriji koju čine integrisana elektronska kola, te se iz tog razloga L1 naziva hardverskim trigerom. Događaji ostaju tu sve dok ne budu ili odbačeni ili prihvaćeni od strane L2 nivoa trigera. Na kraju, poslednji korak u selekciji vrši EF nivo, koji koristi i informacije o kalibraciji, poravnanju detektorskih sistema i mapi magnetnog polja.

## **3.2** Identifikacija fotona na eksperimentu ATLAS

#### 3.2.1 Rekonstrukcija fotona

Interakcije fotona i elektrona sa elektromagnetnim kalorimetrom dovode do nastanka veoma sličnih elektromagnetnih pljuskova, čime se na tim mestima ostavlja velika količina energije ali raspoređene u strogo određeni broj susednih kalorimetrijskih ćelija [12]. Kako su njihove signature u EM kalorimetru skoro identične, potrebno je detaljno definisati način rekonstrukcije elektrona i fotona. Rekonstrukcija kandidata za elektron obuhvata poseban algoritam koji traga za njihovim putanjama i povećava efikasnost rekonstrukcije za kandidate niske energije. Rekonstrukcija konvertovanih i nekonvertovanih fotona se u fazi Run 2 vrši na sličan način kao i u fazi Run 1.

#### 3.2.2 Identifikacija fotona

Identifikacija fotona na ATLAS eksperimentu se uglavnom oslanja na primenu pravougaonih *cut*-ova koristeći kalorimetrijske varijable, čime se postiže dobra separacija promptnih fotona i pogrešnih signatura od nepromptnih fotona koji potiču iz raspada neutralnih hadrona u džetovima i QCD džetova koji ostavljaju veliki deo svoje energije u elektromagnetnom kalorimetru. Ove varijable su prikazane u Tabeli 3.1, kao i na Slici 3.5, sa odgovarajućim definicijama kojim je prikazano na koji način karakterišu bočni i uzdužni tok elektromagnetnog pljuska u elektromagnetnom kalorimetru (EMC), kao i deo pljuska koji odlazi na proboj hadrona u hadronski kalorimetar (HCAL). Promptni fotoni obično proizvode uže energijske "naslage" u EMC-u i imaju manji proboj u HCAL, u poređenju sa pozadinskim fotonima iz džetova.

Pored ovoga, postoje i dva određena seta kriterijuma, odnosno *cut*-ova, *loose* i *tight*, koji su specifično definisani tako da odgovaraju podacima dobijenim u pp sudarima na  $\sqrt{s} = 13$  TeV tokom 2015. i 2016. godine [12]. Iako je isti set varijabli bio korišćen i tokom Run 1 faze, selekcioni *cut*-ovi su podešeni tako da umanje zavisnost identifikacione efikasnosti od *pile-up*-a<sup>2</sup>,

<sup>&</sup>lt;sup>2</sup>Primena statistike je krucijalna za detekciju retkih događaja. Da bi se stvorila takva mogućnost, potrebno je dostići visoku stopu sudara, npr. visokom luminoznošću. LHC je osmišljen tako da dostigne luminoznost od  $10^{34}$  cm<sup>-2</sup>s<sup>-1</sup> [15], čime se prikuplja velika količina podataka svake godine. To zahteva skladištenje velikog broja protona u svaki *bunch*, kao i optimizaciju optičkih karakteristika snopa. Kao posledica ova dva zahteva, dolazi do nekoliko istovremenih *pp* sudara prilikom jednog ukrštanja snopova. Upravo ovaj efekat se naziva *pile up*.

čime bi se postigli bolji rezultati s obzirom na nešto grublje uslove u Run 2 fazi. Ovo je imalo za posledicu manje stroge selekcione kriterijume za konvertovane fotone, gde dolazi do znatnog uticaja velikog broja interakcija po ukrštanju snopova na šire elektromagnetne pljuskove.



Slika 3.5: Shematska reprezentacija varijabli korišćenih pri identifikaciji fotona [12]. Veličina  $E_C^{S_N}$  predstavlja elektromagnetnu energiju sakupljenu u N-tom longitudinalnom sloju elektromagnetnog kalorimetra u klasteru sa svojstvima obeleženim sa C, kojim se označava broj i/ili karakteristike obeleženih ćelija.  $E_i$  je energija u *i*-toj ćeliji, dok je  $\eta_i$  pseudorapiditet centra date ćelije.

Loose selekcioni kriterijumi se zasnivaju na oblicima pljuska u drugom sloju elektromagnetnog kalorimetra, kao i na energiji deponovanoj u hadronskom kalorimetru [12]. Tight selekcioni kriterijumi doprinose informacijama iz fino segmentisanog strip sloja kalorimetra, i posebno su optimizovani za nekonvertovane i konvertovane fotone, jer se uzima u obzir i to da drugi daju generalno širi bočni profil pljuska. Pragovi ovih kriterijuma se razlikuju u 7 intervala psedorapiditeta  $|\eta|$  rekonstruisanog fotona (0.0 - 0.6, 0.6 - 0.8, 0.8 - 1.15, 1.15 - 1.37, 1.52 - 1.81, 1.81 - 2.01, 2.01 - 2.37), pri čemu je razlog za tu izdeljenost pre svega geometrija samog kalorimetra, ali isto tako i različiti efekti na oblike pljuskova koje može imati materijal ispod kalorimetra.

Raspodele datih varijabli za promptne i pozadinske fotone su takođe pod uticajem i dodatnih efekata, kao što je *pile-up* [12]. Rezultat toga je prisustvo aktivnosti niske energije  $E_T$  u detektoru, uključujući ostatke energije u EM kalorimetru. Što je veći broj superponiranih ppdogađaja,  $\mu$ , to će se oblik fotonskog pljuska više proširiti usled dodatne energije deponovane u kalorimetru. Na samom kraju, time se smanjuje efikasnost identifikacije za veće vrednosti  $\mu$ .

Kategorija	Definicija	Naziv	loose	tight
Akseptansa	$ \eta <2.37,$ pri čemu je oblast $1.37\leq  \eta <1.52$ isključena	-	$\checkmark$	~
Proboj hadrona u HCAL	Odnos $E_T$ u prvom sampling sloju hadronskog kalorimetra i $E_T$ elektro- magnetnog klastera (u oblasti $ \eta  < 0.8$ ili $ \eta  > 1.52$ )	$R_{had_1}$	~	~
	Odnos $E_T$ u hadronskom kalorimetru i $E_T$ elektromagnetnog klastera (u oblasti $0.8 <  \eta  < 1.37$ )	$R_{had}$	$\checkmark$	~
Srednji sloj u EM kalorimetru	Odnos energije u $3 \times 7$ $(\eta \times \phi)$ ćelijama i energije u $7 \times 7$ ćelijama koje se nalaze oko fotonskog klastera	$R_{\eta}$	$\checkmark$	~
	Poprečna širina pljuska, $\sqrt{(\Sigma E_i \eta_i^2)/(\Sigma E_i) - ((\Sigma E_i \eta_i)/(\Sigma E_i))^2},$ gde je $E_i$ energija a $\eta_i$ pseudorapiditet za datu ćeliju $i$ , pri čemu je suma računata unutar prozora $3 \times 5$ ćelija	$w_{\eta 2}$	~	~
	Odnos energije u $3 \times 3$ $(\eta \times \phi)$ ćelijama i energije u $3 \times 7$ ćelijama koje se nalaze oko fotonskog klastera	$R_{\phi}$		~
EM <i>strip</i> layer	Poprečna širina pljuska, $\sqrt{(\Sigma E_i(i-i_{max})^2)/(\Sigma E_i)}$ , gde <i>i</i> broji sve trake u prozoru od $3 \times 2$ $(\eta \times \phi)$ trake, a $i_{max}$ predstavlja indeks trake sa najvećom energijom izračunatom iz tri trake oko trake sa najvećom deponovanom energijom	$w_{s3}$		~
	Poprečna širina pljuska, $\sqrt{(\Sigma E_i(i-i_{max})^2)/(\Sigma E_i)}$ , gde <i>i</i> broji sve trake u prozoru od 20 × 2 $(\eta \times \phi)$ trake, a $i_{max}$ predstavlja indeks trake sa najvećom izmerenom energijom u <i>strip</i> sloju	w <sub>s tot</sub>		~
	Energija izvan kora tri centralne trake ali unutar sedam traka, podeljena sa en- ergijom iz tri centralne trake	$f_{side}$		~
	Razlika između energije koja odgovara drugom maksimumu u <i>strip</i> sloju i en- ergije rekonstruisane u traci sa mini- malnom vrednošću pronađenom u inter- valu između prvog i drugog maksimuma	$\Delta E_s$		~
	Odnos energijske razlike između mak- simalne deponovane energije i energije deponovane u drugom maksimumu u klasteru prema sumi ovih energija	$E_{ratio}$		~
	Odnos energije u prvom sloju prema ukupnoj energiji u EM klasteru	$f_1$		~

Tabela 3.1: Varijable korišćene za looseitightidentifikaciju fotona $\left[ 12\right]$ 

# Glava 4 Rezultati i diskusija

U ovom radu, korišćena su tri uzorka dobijena Monte Karlo simulacijama na ATLAS detektoru:

- $h \to Za;$
- $h \to Z\gamma;$
- $\pi^0 \to \gamma \gamma$ .

Razlog koji leži iza izbora ovakvih uzoraka leži u pronalasku pravilnosti, ili pak nepravilnosti, koje se mogu javiti u slučaju pojave aksiona koji se manifestuje raspadom na dva fotona. Čest je slučaj da se nastala dva fotona detektuju kao jedan, usled čega se upravo takva situacija i posmatra, te uzima za *signal*. Iz preostala dva uzorka takođe se izdvaja slučaj sa samo jednim rekonstruisanim fotonom, pri čemu dobijamo potrebni *background*. Kao što će se videti kasnije, *signal* i *background* su dve osnovne klase u problemu na kome se vrši treniranje i učenje veštačke neuronske mreže, kao i algoritama mašinskog učenja.

## 4.1 Korišćeni Monte Karlo uzorci

Za svaki od navedenih Monte Karlo uzoraka, date su osnovne informacije koje se tiču njihove produkcije, kao i pregled atributa selektovanih fotona, odabranih za trening neuronske mreže. To su transverzalni impuls  $(p_T)$ , pseudorapiditet  $(\eta)$ , azimutalni ugao  $(\phi)$ , kao i *shower shape* varijable iz Tabele 3.1. Pored toga, opisani su glavni kriterijumi korišćeni prilikom analize i odabira željenih događaja.

#### 4.1.1 Uzorak $h \rightarrow Za$

Uzorak koji predstavlja signal u ovom istraživanju opisuje proces  $gg \to h \to Za \to Z\gamma\gamma$ . Simulacija je vršena za proton-proton sudare na energiji  $\sqrt{s} = 13$  TeV i potiče iz 2016. godine. Dobijena je pomoću standardnog alata za generisanje visokoenergijskih sudara - *PYTHIA 8*, kao i programskog okvira *POWHEG* koji obezbeđuje određene kalkulacije u programima koji proizvode Monte Karlo simulacije. Za masu aksiona uzeta je vrednost od  $m_A = 0.5$  GeV, dok ukupan broj događaja iznosi 40 000.

Selekcioni kriterijumi su sprovedeni na nekoliko nivoa i oni su sledeći:

- 1. Svi događaji koji sadrže par *truth* fotona koji potiče od aksiona;
- 2. Svi događaji osim onih u kojima se javlja bilo koja druga čestica u konusu veličine  $\Delta R < 0.1$  oko aksiona ili već selektovanih *truth* fotona;
- 3. Svi događaji u kojima za date *truth* fotone postoje odgovarajući rekonstruisani (*reco*) fotoni u konusu  $\Delta R < 0.1$  oko jednog od *truth* fotona iz obeleženog para, i to posebno:
  - (a) 1 reco foton;
  - (b) 2 reco fotona.

Slučaj koji nas posebno zanima jeste 3 (a) selekcioni nivo, koji daje 8 311 događaja. Oni predstavljaju *signal* korišćen za učenje veštačke neuronske mreže.



Slika 4.1: Nenormirana raspodela transverzalnog impulsa  $p_T$ 

Slika 4.2: Normirana raspodela transverzalnog impulsa  $p_T$ 











Slika 4.5: Nenormirana raspodela azimutalnog ugla $\phi$ 





Slika 4.7: Shower shape varijabla  $R_{had}$ 



Slika 4.8: Shower shape varijabla  $R_{had_1}$ 



Slika 4.9: Shower shape varijabla  $R_{\eta}$ 



Slika 4.10: Shower shape varijabla  $w_{\eta 2}$ 



Slika 4.11: Shower shape varijabla  $R_\phi$ 

Slika 4.12: Shower shape varijabla  $w_{stot}$ 



Slika 4.13: Shower shape varijabla  $f_{side}$ 



Slika 4.15: Shower shape varijabla  $E_{ratio}$ 



Slika 4.14: Shower shape varijabla  $\Delta E$ 



Slika 4.16: Shower shape varijabla  $f_1$ 

#### 4.1.2 Uzorak $h \rightarrow Z\gamma$

Jedan od uzoraka koji sačinjava *background* deo podataka opisuje proces  $gg \to h \to Z\gamma$ . Monte Karlo simulacija je takođe vršena za proton-proton sudare na energiji  $\sqrt{s} = 13$  TeV, a softverski alati korišćeni za generisanje jesu *PYTHIA 8, POWHEG* i *EVTGEN*. Ukupno je generisano 130 000 događaja.

Nivoi selekcionih kriterijuma su sledeći:

- 1. Svi događaji koji sadrže truth fotone koji potiču od Higsovog bozona;
- 2. Svi događaji osim onih u kojima se javlja foton koji ne potiče od Higsovog bozona u konusu veličine  $\Delta R < 0.1$  oko već selektovanih truth fotona;
- 3. Svi događaji u kojima za date truth fotone postoje odgovarajući reco fotoni u konusu  $\Delta R < 0.1$  oko obeleženog truth fotona, i to posebno za slučaj 1 reco fotona.

Ovde se za *background* opet uzimaju samo događaji koji su prošli 3. selekcioni nivo, i njih ima ukupno 80 146.



Slika 4.19: Nenormirana raspodela pseudorapiditeta  $\eta$ 

Slika 4.20: Normirana raspodela pseudorapiditeta  $\eta$ 



Slika 4.21: Nenormirana raspodela $\phi$ 

Slika 4.22: Normirana raspodela $\phi$ 



Slika 4.23: Shower shape varijabla  $R_{had}$ 



Slika 4.24: Shower shape varijabla  $R_{had_1}$ 



Slika 4.25: Shower shape varijabla  $R_\eta$ 

Slika 4.26: Shower shape varijabla  $w_{\eta 2}$ 



Slika 4.27: Shower shape varijabla  $R_\phi$ 

Slika 4.28: Shower shape varijabla  $w_{s tot}$ 



Slika 4.29: Shower shape varijabla  $f_{side}$ 



Slika 4.30: Shower shape varijabla  $\Delta E$ 



Slika 4.31: Shower shape varijabla  $E_{ratio}$ 



Slika 4.32: Shower shape varijabla  $f_1$ 

### 4.1.3 Uzorak $\pi^0 \rightarrow \gamma \gamma$

Drugi uzorak koji čini *background* deo podataka opisuje proces  $\pi^0 \to \gamma\gamma$ . Monte Karlo simulacija je takođe vršena za proton-proton sudare na energiji  $\sqrt{s} = 13$  TeV, pomoću softverskog alata Particle Gun - u svojoj najjednostavnijoj formi, on daje označene čestice (pripisan im je odgovarajući *pdg id*) sa njenim relativističkim kvadrivektorom . Ukupno je generisano 1 997 000 događaja.

Nivoi selekcionih kriterijuma su:

- 1. Svi događaji koji sadrže *truth* fotone koji potiču od piona;
- 2. Svi događaji osim onih u kojima se javlja foton koji ne potiče od piona u konusu veličine  $\Delta R < 0.1$  oko već selektovanih truth fotona;
- 3. Svi događaji u kojima za date *truth* fotone postoje odgovarajući *reco* fotoni u konusu  $\Delta R < 0.1$  oko jednog od *truth* fotona iz obeleženog para, i to posebno:
  - (a) 1 reco foton;
  - (b) 2 reco fotona.

Ovde se još jednom za *background* uzimaju samo događaji koji su prošli 3 (a) selekcioni nivo, i njih ima ukupno 394 821.



Slika 4.33: Nenormirana raspodela transverzalnog impulsa  $p_T$ 

Slika 4.34: Normirana raspodela transverzalnog impulsa  $p_T$ 



Slika 4.37: Nenormirana raspodela $\phi$ 



Slika 4.39: Shower shape varijabla  $R_{had}$ 

Slika 4.38: Normirana raspodela $\phi$ 



Slika 4.40: Shower shape varijabla  $R_{had_1}$ 



Slika 4.41: Shower shape varijabla  $R_\eta$ 

Slika 4.42: Shower shape varijabla  $w_{\eta 2}$ 



Slika 4.43: Shower shape varijabla  $R_\phi$ 



Slika 4.44: Shower shape varijabla  $w_{s tot}$ 



Slika 4.45: Shower shape varijabla  $f_{side}$ 





Slika 4.47: Shower shape varijabla  $E_{ratio}$ 

Slika 4.48: Shower shape varijabla  $f_1$ 

## 4.2 Implementacija veštačke neuronske mreže

Prvi metod korišćen za identifikaciju fotona koji potiču od aksiona jeste veštačka neuronska mreža. Pre njene implementacije, neophodno je pripremiti podatke, kao i izvršiti njihovu analizu. Nakon toga sledi priprema skupova podataka (*dataset*-ova), a zatim treniranje neuronske mreže.

#### 4.2.1 Priprema i analiza podataka iz uzoraka

Na samom početku potrebno je izvršiti pripremu uzoraka: ukupno je odabrano 16 621 događaja, od kojih jedna polovina (8 311) pripada signalu, a druga *background*-u - 4155 potiče od  $h \rightarrow Z\gamma$  uzorka, a 4155 od  $\pi^0 \rightarrow \gamma\gamma$  uzorka. Čitav proces pripreme i analize podataka pre implementacije neuronske mreže je rađen pomoću *Pandas* biblioteke, koja predstavlja *Python* biblioteku za analizu podataka.

Nakon što su podaci promešani, učitani su u vidu DataFrame-a, sa kojim je tabelarni prikaz podataka najlakši, kao i rad sa atributima. Ukupno postoji 14 kolona, pri čemu svaka odgovara jednom od 13 atributa, dok je u poslednjoj koloni naznačeno da li data instanca pripada signalu ili *background*-u. Svaki red odgovara jednoj instanci, odnosno jednom fotonu.

Za uspešnost učenja neuronske mreže često je potrebno pažljivo odabrati atribute, i odstraniti one koji možda nisu neophodni jer bi inače samo dodatno opterećivali mrežu. Veoma čest način na koji se njihova važnost može ispitati jeste pomoću korelacione matrice. Ovim putem se računa korelacija između svih atributa, i konkretno ovde je upotrebljena *Pearson* korelacija [23], čiji koeficijent  $\rho$  se računa kao:

$$\rho = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}.$$
(4.1)

(X, Y) predstavlja par varijabli između kojih se korelacija računa,  $\sigma_X$  i  $\sigma_Y$  standardne devijacije, dok je  $\operatorname{cov}(X, Y)$  kovarijansa kojom se opisuje jačine veze između datih varijabli. Kovarijansa se može predstaviti pomoću srednje vrednosti  $(\mu)$  i očekivane vrednosti (E):

$$cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)].$$
(4.2)



Slika 4.49: Matrica korelacije; može se primetiti da su najveći koeficijenti korelacije između dve različite varijable  $\rho(R_{had}, R_{had_1}) = 0.703390$ ,  $\rho(f_1, E_{ratio}) = 0.603118$  i  $\rho(f_1, w_{\eta 2}) = 0.570811$ .

Sledeći korak jeste priprema skupa podataka, sačinjenog iz 16 621 fotona (background + signal), pri čemu svaki foton ima po 13 atributa. Ispod su grafički prikazani dobijeni rezultati koji pokazuju razliku između background i signal fotona.



Slika 4.50: Raspodela transverzalnog impulsa<br/>  $$p_T$$ 



Slika 4.51: Raspodela pseudorapiditeta  $\eta$ 



Slika 4.52: Raspodela azimutalnog ugla $\phi$ 



Slika 4.53: Shower shape varijabla  $R_{had}$ 



Slika 4.55: Shower shape varijabla  $R_\eta$ 



Slika 4.54: Shower shape varijabla  $R_{had_1}$ 



Slika 4.56: Shower shape varijabla  $w_{\eta 2}$ 



Slika 4.57: Shower shape varijabla  $R_{\phi}$ 



Slika 4.59: Shower shape varijabla  $f_{side}$ 



Slika 4.61: Shower shape varijabla  $E_{ratio}$ 



Slika 4.58: Shower shape varijabla  $w_{s tot}$ 



Slika 4.60: Shower shape varijabla  $\Delta E$ 



Slika 4.62: Shower shape varijabla  $f_1$ 

#### 4.2.2 Podela podataka na trening, validacioni i test skup podataka

Kako bi moglo da se vrši treniranje mreže, potrebno je obezbediti odgovarajući skup podataka za sam proces učenja, odnosno treniranja, kao i za testiranje uspešnosti klasifikacije neuronske mreže. Međutim, često se uvodi i validacioni skup podataka, što je i ovde slučaj, koji služi za evaluaciju neuronske mreže u procesu treninga, nakon koje se vraćamo na trening i vršimo fino podešavanje i unosimo odgovarajuće izmene kako bismo poboljšali performans mreže.

Sva tri skupa podataka su standardizovana funkcijom StandardScaler(), koja srednju vrednost približava 0, a standardnu devijaciju 1. Ovo se vrši za svaki atribut ponaosob, kako bi njihove vrednosti što više ličile na normalnu raspodelu, za koju se pokazalo da je mašinsko učenje uspešnije. Standardizacija skupa podataka je uobičajen zahtev za mnoge estimatore u mašinskom učenju, koji se mogu ponašati loše u slučaju kada se ona ne primeni, odnosno ukoliko srednja vrednost nije 0 a standardna devijacija 1.Nakon toga sledi primena metode fit\_transform(), koja fituje podatke a potom ih transformiše tako što stare vrednosti zamenjuje novim.[24]

#### 4.2.3 Treniranje modela

Ključni korak je naravno proces treniranja modela veštačke neuronske mreže. Model koji je odabran u ovom radu jeste *Sequential* model definisan u *Keras* biblioteci. Ovaj model predstavlja linearno slagane slojeve, koji se dodaju metodom add(Dense()), gde Dense() predstavlja gusto povezani sloj neuronske mreže (svaki neuron u njemu je povezan sa svakim neuronom u sloju pre i posle), koji obuhvata operaciju: output = activation(dot(input, kernel) + bias), gde su sve uvedeni parametri definišu prilikom samog uvođenja Dense sloja. Pored ulaznog i izlaznog sloja, dodata su još dva skrivena Dense sloja. Za funkciju koja određuje način inicijalizacije težina je odabrana he\_ normal funkcija, koja se zasniva na normalnoj raspodeli. Kao funkcije aktivacije korišćena je ReLu funkcija, dok je samo za poslednji *output* sloj korišćena sigmoidna funkcija.

Kao što je ranije pomenuto, i ovde je rađeno sa tri (pod)skupa podataka: trening, validacionim i test *dataset*-om. Sam proces učenja, odnosno treninga je vršen na trening *dataset*-u, dok je svaka provera uspešnosti rađena pomoću validacionog skupa podataka. Često se na osnovu dobijene tačnosti prilikom validacije može utvrditi da li je došlo do *overfitting*-a - pojave u kojoj je neuronska mreža previše dobro istrenirana. Tačnije, ona je naučila da rešava zadati problem samo za trening *dataset*. Problem *overfitting*-a se rešava regularizacijom, koja se može primeniti na tri različita načina:

- L1 i L2 regularizacija ovim se uvode dodatni članovi u *cost* funkciju koji umanjuju doprinos članova višeg reda i tako pomažu u održavanju jednostavnosti rešenja;
- Kombinacija modela neophodno je da dati modeli, kao i arhitekture njhovih neuronskih mreža, budu različiti i potpuno nezavisni;
- Dropout tehnika kombinacija prethoodne dve. Praktično, prilikom njene primene privremeno se gasi određeni deo neurona u neuronskoj mreži, koji se biraju nasumično u svakom epoch-u<sup>1</sup>.

<sup>&</sup>lt;sup>1</sup>Jedan *epoch* predstavlja slučaj u kome čitav skup podataka prođe samo jednom kroz neuronsku mrežu.[25]

U ovom trenutku se takođe procenjuje koji metod za regularizaciju je najbolje primeniti, i u ovom radu odabrana je Dropout() funkcija iz *Keras*-a. Za parametar koji definiše broj ugašenih neurona odabrana je vrednost 0.5, što znači da prilikom jednog *epocha*-a radi 50% neurona.

#### 4.2.4 Zaključni plotovi

Nakon validacije, dobija se da je tačnost performanse mreže 89.38%, dok je gubitak 0.2891. Za test *dataset* dobija se tačnost<sup>2</sup> 89.59%, a gubitak 0.2961. Prilikom analize dobijenih rezultata, dobijena su tri plota prikazana ispod.



Slika 4.63: Potiskivanje fona (eng. background rejection) u zavisnosti od efikasnosti signala



Slika 4.64: Gubitak za svaki epoch

 $<sup>^2 {\</sup>rm odnos}$ broja tačno predviđenih rezultata i ukupnog broja uzoraka



Slika 4.65: Tačnost za svaki epoch

## 4.3 Primena metoda mašinskog učenja

Pored veštačke neuronske mreže, na istom setu podataka su primenjeni i neki od opisanih algoritama mašinskog učenja. Jedan od glavnih mera korišćenih u proceni uspešnosti datih klasifikatora jeste matrica konfuzije, čija definicija je prikazana Tabelom 4.1.

Tabela 4.1: Matrica konfuzije

	Actual True	Actual False
Predicted True	True Positive (TP)	False Negative (FN)
Predicted False	False Positive (FP)	True Negative (TN)

Pored matrice konfuzije, koriste se i sledeći parametri:

- Tačnost broj instanci uspešno klasifikovanih: (TP + TN)/N, gde je N ukupan instanci u skupu podataka, odnosno N = TP + FP + FN + TN;
- Preciznost: TP/(TP + FP);
- Odziv<sup>3</sup>: TP/(TP + FN);
- F1 mera: (2 \* preciznost \* odziv )/(preciznost + odziv).

 $^{3}$ eng. *recall* 





Slika 4.66: Matrica konfuzije za algoritam knajbližih suseda; na osnovu dijagonalnih članova, vidi se da je tačno predviđeno 2 995 od ukupno 3 325 instanci.



Slika 4.68: Matrica konfuzije za Random Forest algoritam; uspešno je predviđeno 2 982 od ukupno 3 325 instanci.



Slika 4.67: Matrica konfuzije za Decision Tree algoritam; tačno je predviđeno 2 963 od ukupno 3 325 instanci.



Slika 4.69: Matrica konfuzije za Support Vector Machine algoritam, takođe, prema elementima na dijagonali matrice zabune dobija se da je tačno predviđeno 3 023 od ukupno 3 325 instanci.

Kada su svi modeli ispitani, procenjene su njihove tačnosti i date u tabeli ispod.

	Preciznost [%]	Odziv [%]	F1 mera [%]	Tačnost [%]
Veštačka neuronska mreža	93.74	52.27	67.11	89.59
KNN	90	90	90	90
Decision Tree	90	88	89	89
Random Forest	90	89	90	90
SVM	92	90	91	91

Tabela 4.2: Mer	e uspešnosti	klasifikatora
-----------------	--------------	---------------

## Zaključak

Nakon izvršene uvodne analize, pripreme podataka, kao i treniranja različitih algoritama mašinskog učenja na njima, dobijaju se rezultati čiji pregled je dat u Tabeli 4.2.

Ono na šta je potrebno skrenuti pažnju jeste da je prilikom računanja navedenih parametara u slučaju neuronske mreže, za vrednost praga na osnovu kog se procenjuje da li je klasifikacija izvršena dobro ili ne, uzeto 0.5. To znači da ukoliko je mreža predvidela vrednost 0.7 u intervalu [0, 1], pri čemu je 0 *background* a 1 signal, data instanca će biti računata kao *True Predicted*. To dalje znači da postoji čitav niz mogućih vrednosti za dati prag, od kojih neke nude moguće bolje vrednosti za F1 meru za ovaj konkretni slučaj. Međutim, ovde je uzeta samo jedna vrednost, kako bi se ručno mogle izračunati mere uspešnosti klasifikatora i tako uporedile sa ostalim algoritmima mašinskog učenja.

Što se tiče poslednje kolone, odnosno tačnosti, za slučaj neuronske mreže ona je uzeta prilikom evaluacije modela isprobanog na *test* skupu podataka, i nije računata ručno kao preciznost, odziv i F1 mera neuronske mreže.

Na osnovu dobijenih rezultata, možemo da zaključimo da neuronska mreža postiže najveću preciznost (93.74%). Ovakav rezultat se mogao i očekivati, s obzirom na to da je takav model dosta složeniji od ostalih algoritama mašinskog učenja. Takođe, njegova ogromna prednost u odnosu na ostale testirane modele jeste pre svega prilagodljivost. Naime, veoma jednostavno se može dodavati još slojeva, kao i podešavati broj neurona, funkcije aktivacije i metode regularizacije. Za uspešnost neuronske mreže ne postoji univerzalno rešenje, već se njena arhitektura bira prema samom skupu podataka sa kojima se radi, kao i na osnovu mnogo pokušaja i naknadnih podešavanja. Zato se kao cilj ne postavlja samo postizanje najveće tačnosti, već pripremanje modela za nove podatke i strukture unutar njih, čime se on osposobljava za generalizaciju i optimalniji proces učenja.

## Bibliografija

- [1] K. Albertsson et al. Machine Learning in High Energy Physics Community White Paper. arXiv:1807.02876 (2019)
- [2] V. Oberški. The strong CP problem and axions. Ljubljana University (2017)
- [3] C. A. Baker et al. An Improved Experimental Limit on the Electric Dipole Moment of the Neutron. arXiv:hep-ex/0602020 (2006)
- [4] D. Mrđa, I. Bikit. Osnove fizike čestica i nuklearne fizike. Prirodno-matematički fakultet, Univerzitet u Novom Sadu (2016)
- [5] Large Hadron Collider. Preuzeto sa: https://home.cern/topics/large-hadron-collider
- [6] CERN's Accelerator Complex. https://cds.cern.ch/record/1621583
- [7] A. Thamm, T. Melia, P. Charitos. (19.9.2018) Axion-like particle searches at the LHC. Preuzeto sa: https://ep-news.web.cern.ch
- [8] M. Bauer, M. Neubert, A. Thamm. Collider Probes of Axion-Like Particles. arXiv:1708.00443 (2017)
- [9] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. JINST, 3:S08003, 2008.
- [10] About the ATLAS Experiment. Preuzeto sa: https://atlas.cern/discover/about
- [11] The Inner Detector. Preuzeto sa: https://atlas.cern/discover/detector/inner-detector
- [12] The ATLAS Collaboration. Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run 2 data collected in 2015 and 2016. (CERN-EP-2018-216), 2018.
- [13] S. Jakobsen. Commissioning of the Absolute Luminosity For ATLAS detector at the LHC. Niels Bohr Institute, University of Copenhagen, 2013.

- [14] Trigger and Data Acquisition System.Preuzeto sa: https://atlas.cern/discover/detector/trigger-daq
- [15] G. Soyez. Pileup mitigation at the LHC. Institut de Physique Théorique, CEA Saclay, CNRS UMR 3681, 2018.
- [16] D. Cielen, A. D. B. Meysman, M. Ali. Introducing Data Science. Manning Publications, 2016.
- [17] P. Fernandez. (6.12.2018) The Difference between Artificial Intelligence, Machine Learning, and Deep Learning.
   Preuzeto sa: https://medium.com
- [18] T. M. Mitchell. Machine Learning. McGraw-Hill, 2016.
- [19] T. Borovicka, M. Jirina Jr, P. Kordik, M. Jirina. (18.6.2012) Selecting Representative Data Sets. Preuzeto sa: https://www.intechopen.com/books/advances-in-data-mining-knowledgediscovery-and-applications
- [20] M. Jain. (17.1.2019) The Basics of Neural Networks. Preuzeto sa: https://medium.com
- [21] S. Gavran. Veštačke neuronske mreže u istraživanju podataka: Pregled i primena. Univerzitet u Beogradu, 2016.
- [22] A. Sharma. (30.3.2017) Understanding Activation Functions in Neural Networks. Preuzeto sa: https://medium.com
- [23] Kent State University. SPSS Tutorials: Pearson Correlation. Preuzeto sa: https://libguides.library.kent.edu/SPSS/PearsonCorr
- [24] Documentation of scikit-learn 0.21.3. Preuzeto sa: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
- [25] S. Sharma. (23.9.2017) Epoch vs Batch Size vs Iterations. Preuzeto sa: https://medium.com

## Biografija



Olivera Vujinović rođena je 3. decembra 1995. godine u Novom Sadu. Završila je osnovnu školu "Miroslav Antić" kao đak generacije i gimnaziju "Svetozar Marković" u Novom Sadu sa Vukovom diplomom. 2018. godine završila je osnovne akademske studije fizike - istraživački smer, na Prirodno-Matematičkom fakultetu u Novom Sadu. Iste godine je upisala i master studije na Katedri za Nuklearnu fiziku. Dobitnik je stipendije Ministarstva prosvete, kao i Fonda za mlade talente Republike Srbije - Dositeja. Bila je učesnik u dvomesečnom CERNovom letnjem programu namenjenom studentima zainteresovanim za oblast fizike elementarnih čestica. Deo je saradnje koju ima Institut za Fiziku u Beogradu sa grupom na Univerzitetu Johannes Gutenberg u Majncu, a koja se realizovala u vidu projekta iz kog je proistekao i master rad.

#### UNIVERZITET U NOVOM SADU PRIRODNO - MATEMATIČKI FAKULTET KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj: RBR Identifikacioni broj: IBR Tip dokumentacije: Monografska dokumentacija TDTip zapisa: Tekstualni štampani materijal TΖ Vrsta rada: Master rad VR Autor: Olivera Vujinović AU Mentor: prof. dr Jovana Nikolov, dr Nenad Vranješ MN Naslov rada: Potraga za aksionima na eksperimentu ATLAS korišćenjem metoda mašinskog učenja NR Jezik publikacije: srpski (latinica)  $\mathbf{JP}$ Jezik izvoda: srpski/engleski JI Zemlja publikovanja: Republika Srbija  $\mathbf{ZP}$ *Uže geografsko područje*: Vojvodina UGP Godina: 2019. GO Izdavač: Autorski reprint  $\mathbf{IZ}$ Mesto i adresa: Prirodno-matematički fakultet, Trg Dositeja Obradovića 4, Novi Sad  $\mathbf{M}\mathbf{A}$ Fizički opis rada: 6 poglavlja/60 strana/25 referenci/89 slike FO

Naučna oblast: Fizika

### NO

Naučna disciplina: Fizika visokih energija

### ND

 $Predmetna \ odrednica/\ ključne\ reči:$ ATLAS eksperiment, Aksioni, Fizika visokih energija ${\bf PO}$ 

UDK

 $\check{C}\!uva\ se$ : Biblioteka departmana za fiziku, PMF-a u Novom Sadu

ČU

Važna napomena: nema

## VN

*Izvod*: Zadatak eksperimentalne fizike visokih energija svodi se na ispitivanje Standardnog modela merenjima visoke preciznosti, kao i na potragu za novim česticama, što predstavlja fiziku izvan Standardnog modela. Obe stavke zahtevaju identifikaciju retkih signala, što će biti znatno otežano povećanim *pile-up*-om usled dodatnih sudarajućih protona na HL-LHC. Svrha ovog rada jeste upravo ispitivanje uspešnosti različitih algoritama mašinskog učenja u potrazi za aksionima kao novim česticama, pri čemu je centralni deo rada zasnovan na modelu veštačke neuronske mreže i njegovoj uspešnosti.

## ΙZ

Datum prihvatanja teme od NN veća: Septembar, 2018.

## DP

*Datum odbrane*: 30.9.2019.

#### DO

Članovi komisije:

### KO

Predsednik:	dr Dušan Mrđa, redovni profesor
član:	dr Danijela Boberić Krstićev, vanredni profesor
član:	dr Jovana Nikolov, vanredni profesor, mentor
član:	dr Nenad Vranješ, viši naučni saradnik, mentor

#### UNIVERSITY OF NOVI SAD FACULTY OF SCIENCE AND MATHEMATICS KEY WORDS DOCUMENTATION

Accession number: ANO Identification number: INO Document type: Monograph publication DT Type of record: Textual printed material  $\mathbf{TR}$ Content code: Final paper CC Author: Olivera Vujinović AU Mentor: doc. dr Jovana Nikolov, dr Nenad Vranješ MN *Title*: Search for the axion-like particles in ATLAS experiment with application of machine learning methods  $\mathbf{TI}$ Language of text: srpski (latinica) LTLanguage of abstract: English LA Contry of publication: Republika Srbija CP Locality of publication: Vojvodina  $\mathbf{LP}$ Publication year: 2019. PY Publisher: Author's reprint PU Publication place: Faculty of Science and Mathematics, Trg Dositeja Obradovića 4, Novi Sad PP Physical description: 6 chapters/60 pages/25 references/89 pictures PD Scientific field: Physics SF Scientific discipline: High Energy Physics SDSubject/Key words: ATLAS Experiment, Axions, High Energy Physics SKW UDK

 $Holding\ data:$ Library of Department of Physics, Tr<br/>g Dositeja Obradovića 4 ${\bf HD}$ 

*Note*: none

#### Ν

*Abstract*: The experimental high-energy physics has two main goals - probing the Standard Model with precision measurements and searching for new particles which are associated with physics beyond the Standard Model. Both objectives demand the identification of rare signals in immense backgrounds. This will become a very serious challenge due to the higher rate of pile-up collisions from additional protons at the HL-LHC. The aim of this thesis is to evaluate the accuracy of each of machine learning algorithms, while the main part of the thesis is focused on an artifical neural network model and its performance.

### AB

Accepted by the Scientific Board: September, 2019.

### ASB

Defended on: 30.9.2019.

#### DO

Thesis defend board:

### DB

President: dr Dušan Mrđa, Full Professor

- Member: dr Danijela Boberić Krstićev, Assistant Professor
- Member: dr Jovana Nikolov, Assistant Professor, mentor
- Member: dr Nenad Vranješ, Senior Research Associate, mentor