

**Technická univerzita v Košiciach  
Fakulta elektrotechniky a informatiky**

**Aplikácia neurónových sietí pre detekciu  
atmosférických spršok v dátach z  
experimentu EUSO-SPB**

**Diplomová práca**

**2019**

**Bc. Peter Szakács**

**Technická univerzita v Košiciach  
Fakulta elektrotechniky a informatiky**

**Aplikácia neurónových sietí pre detekciu  
atmosférických spršok v dátach z  
experimentu EUSO-SPB**

**Diplomová práca**

Študijný program: Informatika  
Študijný odbor: 9.2.1 Informatika  
Školiace pracovisko: Katedra počítačov a informatiky (KPI)  
Školiteľ: doc. Ing. Ján Genči, PhD.  
Konzultant: Ing. Michal Vrábel

**Košice 2019**

**Bc. Peter Szakács**

Názov práce: Aplikácia neurónových sietí pre detekciu atmosférických spršok v dátach z experimentu EUSO-SPB

Pracovisko: Katedra počítačov a informatiky, Technická univerzita v Košiciach

Autor: Bc. Peter Szakács

Školiteľ: doc. Ing. Ján Genči, PhD.

Konzultant: Ing. Michal Vrábel

Dátum: 26. 4. 2019

Kľúčové slová: experiment EUSO-SPB, konvolučné neurónové siete, kozmické žiarenie vysokých energií (UHECR)

**Abstrakt:** V tejto práci skúmame potenciál trénovaných modelov založených na konvolučných neurónových sieťach pre spracovanie obrazu ako spôsobu klasifikácie dát z vedeckého experimentu EUSO-SPB určeného na štúdium extenzívnych atmosférických spršok vyvolaných časticami kozmického žiarenia vysokých energií. Skúmaná klasifikačná schéma je binárna, pričom cieľom je vybrať zo zaznamenaných dát tie, ktoré potenciálne obašujú stopu spršky. Pre tréning siete sme použili dáta generované 2 rôznymi spôsobmi. Následne sme porovnali senzitivitu a špecificitu modelov trénovaných osobitne na jednom type dát a na zmesi oboch typov dát pri rozpoznávaní spršok. Naše výsledky naznačujú, že jeden z typov generovaných dát produkoval celkovo lepšie modely z hľadiska sledovaných metrík. Po výbere modelov vykazujúcich najlepšie metriky sme ich aplikovali na dáta z experimentu a našli sme vyše 200 potenciálne zaujímavých javov v dátach. Naše výsledky dodatočne indikujú, že pri vhodne zvolenej sade trénovacích dát je možné použiť postupy založené na strojovom učení ako spôsob spracovania obrovského množstva dát produkovaných experimentom EUSO-SPB a iným jemu podobným.

Thesis title: Application of neural networks for air shower detection in data from EUSO-SPB experiment

Department: Department of Computers and Informatics, Technical University of Košice

Author: Bc. Peter Szakács

Supervisor: doc. Ing. Ján Genči, PhD.

Tutor: Ing. Michal Vrábel

Date: 26. 4. 2019

Keywords: EUSO-SPB, convolutional neural networks, UHECR

Abstract: In this work, we explore the potential of trained convolutional neural networks for image processing as a way to classify data from the EUSO-SPB science experiment to study ultra-high energy cosmic ray-induced extensive atmospheric air showers. The classification scheme explored is binary, where the goal is to select those recorded events which might potentially contain the track of a shower. For training, we use data generated using 2 different methods and compare the sensitivity and specificity of the models trained on both types of data separately and on a mix of them with regards to shower detection. Our results indicate that using one type of generated data yielded better overall results. After selecting the models with the best metrics, we applied them for classifying the data recorded during the experiment and found over 200 potential events of interest. Our results also indicate that, assuming a proper set of training data, using machine learning-based approaches is a possible way to process the large amounts of data produced by this experiment and others similar to it.

## ZADANIE DIPLOMOVEJ PRÁCE

Študijný odbor: **Informatika**  
Študijný program: **Informatika**

Názov práce:

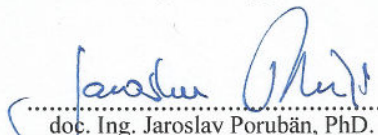
**Aplikácia neurónových sietí pre detekciu atmosferických spŕšok v dátach z experimentu EUSO-SPB**  
Application of neural networks for air shower detection in data from EUSO-SPB experiment

Študent: **Bc. Peter Szakács**  
Školiteľ: **doc. Ing. Ján Genči, PhD.**  
Školiace pracovisko: **Katedra počítačov a informatiky**  
Konzultant práce: **Ing. Michal Vrábel**  
Pracovisko konzultanta: **Katedra počítačov a informatiky**

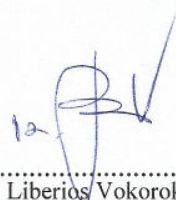
Pokyny na vypracovanie diplomovej práce:

1. Oboznámiť sa s princípom experimentu EUSO-SPB a spúšťacím algoritmom prvej úrovne.
2. Vypracovať stručný prieskum aktuálneho stavu problematiky použitia konvolučných neurónových sietí pre klasifikáciu dát podobných dátam z experimentu EUSO-SPB.
3. Vytvoriť výpočtovo nenáročnú metódu simulácie atmosferickej spŕšky pre účely tréningu neurónovej siete.
4. Navrhnuť a implementovať model konvulčnej neurónovej siete pre klasifikáciu dát aspoň do dvoch tried - šum a atmosférická spŕška.
5. Použitím navrhutej konvulčnej neurónovej siete klasifikovať dáta v rôznych poskytnutých formátoch.
6. Vizualizovať váhy a aktivácie vrstiev netrénovaného modelu konvulčnej neurónovej siete.
7. Vyjadriť úspešnosť klasifikácie používaných modelov konvulčných neurónových sietí aj v závislosti od parametrov simulovanej atmosferickej spŕšky.
8. Vypracovať dokumentáciu podľa pokynov vedúceho práce.

Jazyk, v ktorom sa práca vypracuje: slovenský  
Termín pre odovzdanie práce: 26.04.2019  
Dátum zadania diplomovej práce: 31.10.2018

  
doc. Ing. Jaroslav Porubán, PhD.  
vedúci garantujúceho pracoviska



  
prof. Ing. Liberios Vokorokos, PhD.  
dekan fakulty

### **Čestné vyhlásenie**

Vyhlasujem, že som záverečnú prácu vypracoval(a) samostatne s použitím uvedenej odbornej literatúry.

Košice, 26.4.2019

.....

*Vlastnoručný podpis*

## **Podakovanie**

Na tomto mieste by som rád poďakoval svojmu vedúcemu práce za jeho čas a odborné vedenie počas riešenia mojej záverečnej práce.

Rovnako by som sa rád poďakoval svojim rodičom a priateľom za ich podporu a povzbudzovanie počas celého môjho štúdia.

# Obsah

---

<b>Motivácia</b>	<b>1</b>
<b>1 Formulácia úlohy</b>	<b>3</b>
<b>2 EUSO-SPB</b>	<b>4</b>
2.1 Motivácia a koncept pozorovania . . . . .	4
2.2 Štruktúra detektora . . . . .	6
2.3 Spúšťače . . . . .	7
2.3.1 Spúšťač prvej úrovne . . . . .	8
2.4 Priebeh misie a zozbierané dáta . . . . .	9
2.5 Softvérový rámec ESAF . . . . .	11
2.5.1 Rekonštrukcia spršky - existujúce algoritmy . . . . .	13
<b>3 Strojové učenie v časticovej fyzike</b>	<b>14</b>
3.1 Konvolučné siete . . . . .	15
3.1.1 Konvolučné vrstvy . . . . .	16
3.1.2 Vzorkovacie vrstvy . . . . .	18
3.2 Konvolučné siete v časticovej fyzike . . . . .	18
3.2.1 Návrhy pre pozemné detektory kozmického žiarenia . . . . .	19
3.2.2 Detektor NOvA . . . . .	21
<b>4 Implementácia riešenia</b>	<b>23</b>
4.1 Formát dát . . . . .	24
4.2 Dátové sady . . . . .	25
4.3 Práca s dátovými sadami . . . . .	27
4.3.1 Zlučovanie sád . . . . .	27
4.3.2 Vizualizácia položiek sady . . . . .	28



---

4.4	Spracovanie simulovaných dát a dát z misie . . . . .	29
4.4.1	Charakteristiky vytvorenej sady . . . . .	30
4.4.2	Režimy spracovania zoznamu paketov . . . . .	31
4.5	Syntetické dáta . . . . .	32
4.5.1	Parametre spŕšky a pozadia . . . . .	33
4.5.2	Algoritmus generovania stopy spŕšky . . . . .	34
4.5.3	Simulácia chybových EC modulov . . . . .	35
4.6	Architektúry sietí . . . . .	37
4.6.1	Výber architektúr . . . . .	38
<b>5</b>	<b>Overenie riešenia</b>	<b>41</b>
5.1	Sledované metriky . . . . .	41
5.2	Dáta pre tréning modelov . . . . .	42
5.2.1	Simulované dáta . . . . .	42
5.2.2	Syntetické dáta . . . . .	43
5.2.3	Zmiešané dáta . . . . .	43
5.3	Dáta pre vyhodnotenie modelov . . . . .	44
<b>6</b>	<b>Dosiahnuté výsledky</b>	<b>46</b>
6.1	Modely tréňované na simulovaných dátach . . . . .	47
6.2	Modely tréňované na syntetických dátach . . . . .	49
6.3	Modely tréňované na zmiešaných dátach . . . . .	52
6.4	Vizualizácie máp príznakov skrytých vrstiev modelu . . . . .	56
6.5	Vyhodnotenie dát z misie . . . . .	57
6.6	Interpretácia výsledkov . . . . .	60
<b>7</b>	<b>Záver</b>	<b>62</b>
	<b>Literatúra</b>	<b>64</b>
	<b>Zoznam skratiek</b>	<b>69</b>
	<b>Zoznam príloh</b>	<b>70</b>

# Zoznam obrázkov

---

2.1	Plocha zemského povrchu snímaná detektorom JEM-EUSO pri kolmom pohľade dolu (nadir mód, tyrkysová) a pri odklone od kolmice o uhol $20^\circ$ (biela) a $30^\circ$ (oranžová). Obrázok prevzatý od Adams et al., 2015. . . . .	6
2.2	Hierarchia modulov tvoriacich hardvér detektora EUSO-SPB aj JEM-EUSO. Obrázok prevzatý od MacCarone et al., 2011. . . . .	7
2.3	Horný rad: frekvencia aktivácie spúšťača FLT počas misie EUSO-SPB. Parametre $P$ a $R$ odpovedajú hodnotám $N_{pst}$ a $N_{ctd}$ použitým v logike spúšťača. Dolný rad: priemerná intenzita pixelov v zaznamenaných paketoch snímok. Hodnoty $P = 4$ a $R = 2$ odpovedajú parametrom pre misiu JEM-EUSO. Obrázok prevzatý od M. E. Bertaina et al., 2017 . . . . .	10
2.4	Vývoj výšky letu balóna počas misie EUSO-SPB. Hrubou zelenou sú označené obdobia, počas ktorých prebiehal zber dát. Obrázok prevzatý od Wiencke et al., 2017 . . . . .	11
2.5	Parametre atmosférickej spfšky. Obrázok prevzatý od Vrábel, 2017	12
3.1	Príklad typickej architektúry konvolučnej neurónovej siete použitej pre klasifikáciu obrázkov. Konvolučná časť siete je tvorená vrstvami označenými ako Conv (extrakcia príznakov) a Pool (zmenšenie vstupu). Plne prepojená časť je tvorená plne prepojenými vrstvami označenými ako FC (z angl. Fully connected) a Softmax. Obrázok prevzatý od Dertat, 2017. . . . .	16
3.2	Príklad konvolučnej operácie aplikovanej na vstupný obraz s tromi farebnými kanálmi. Obrázok prevzatý od Machine Learning Guru, 2017. . . . .	17

3.3	Snímky zachytávajúce stopy zanechané tromi klasifikovanými typmi častíc v pokuse s hmlovou komorou. Obrázok prevzatý od Lincoln, 2017. . . . .	19
4.1	Štruktúra paketu ako aj projekcie z pohľadu určujúcich osí. Snímky sú so zvýraznením kontrastu a vo falošnej farbe. . . . .	24
4.2	Vizualizácia spršky zo syntetických dát. . . . .	29
4.3	Priebeh funkcie pre generovanie hodnôt signálu simulovanej spršky pre hodnotu <code>shower_max</code> rovnú 15 a stopu s dĺžkou 30 pixelov. Modré štvorce predstavujú zaokrúhlenú celočíselnú hodnotu funkcie pre danú súradnicu čiar. Vpravo je vidno stopu tejto spršky v xy projekcii paketu s nulovým pozadím a vyznačenou počiatočnou súradnicou . . . . .	35
4.4	Stopa generovanej spršky z obrázka 4.3 po vyhladení pri pohľade z xy projekcie na pozadí bez šumu a so šumom. . . . .	36
4.5	Stopa generovanej spršky z obrázka 4.4 v xy projekcii po simulovanom vyradení 4 EC modulov a xy projekcia paketu z misie EUSO-SPB (25.4.2017) s rovnakým počtom nefunkčných EC modulov. . . .	36
6.1	Vývoj celkovej úspešnosti predpovedí na sade simulovaných dát po 30 epochách tréningu pre siete <code>github_net3b</code> (oranžová) a <code>triple_net</code> (modrá). . . . .	47
6.2	Senzitivita natrénovaného modelu ako funkcie skutočnej energie simulovanej primárnej častice v MeV. . . . .	48
6.3	Senzitivita natrénovaného modelu ako funkcie azimutálneho uhla $\phi$ v radiánoch. . . . .	48
6.4	Senzitivita natrénovaného modelu ako funkcie zenitového uhla $\theta$ v radiánoch. . . . .	49
6.5	Vývoj celkovej úspešnosti predpovedí na sade syntetických dát po 30 epochách tréningu pre siete <code>github_net3b</code> (oranžová) a <code>triple_net</code> (modrá). . . . .	50
6.6	Senzitivita natrénovaného modelu ako funkcia parametra <code>shower_max</code> . 50	
6.7	Senzitivita natrénovaného modelu ako funkcia parametra <code>bg_lambda</code> . 51	
6.8	Senzitivita natrénovaného modelu ako funkcia parametra <code>duration</code> . 51	
6.9	Senzitivita natrénovaného modelu ako funkcia parametra <code>track_length</code> . 52	

---

6.10	Senzitivita modelov trénovaných na zmiešaných sadách dát oproti modelom trénovaným len na simulovaných dátach ako funkcia skutočnej energie simulovanej primárnej častice v MeV. . . . .	54
6.11	Senzitivita modelov trénovaných na zmiešaných sadách dát oproti modelom trénovaným len na simulovaných dátach ako funkcia zenitového uhla $\theta$ v radiánoch. . . . .	55
6.12	Senzitivita modelov trénovaných na zmiešaných sadách dát oproti modelom trénovaným len na simulovaných dátach ako funkcia azimutálneho uhla $\phi$ v radiánoch. . . . .	56
6.13	Porovnanie aktivácií prvej konvolučnej vrstvy spracúvajúcej projekciu XY pri modeloch založených na architektúre <code>triple_net</code> . . . . .	58
6.14	Dáta z misie z 28.4.2017 označené modelmi ako spíška ukazujúce náhle rozptýlené zjasnenie meracej plochy detektora. . . . .	59
6.15	Dáta z misie z 28.4.2017 označené modelmi ako spíška ukazujúce nárazovité zjasnenie pixelu resp. viacerých pixelov a postupné zoslabnutie intenzity. . . . .	60

# Zoznam tabuliek

---

4.1	Architektúry sietí použitých pri pokuse. Zvýraznené sú odchýlky oproti základnému modelu s poradovým číslom 1. Vysvetlivky použitých skratiek sú v tabuľke 4.2. Hodnota "n/a"indikuje absenciu vrstvy oproti základnému modelu . . . . .	39
4.2	Vysvetlenie skratiek použitých pri popise vrstiev siete. . . . .	39
4.3	Úspešnosť jednotlivých sietí po 11 epochách tréningu . . . . .	40
5.1	Parametre syntetických dát použitých pri tréningu aj vyhodnotení .	44
5.2	Typy a parametre dát použitých pri vyhodnotení modelov . . . . .	45
6.1	Súhrnné vyhodnotenie metrík modelov. Stĺpce SP (skutočné pozitíva) a FN (falošné negatíva) označujú položky sady obsahujúce simulovanú spŕšku klasifikované modelom do správnej resp. nesprávnej triedy. Stĺpce SN (skutočné negatíva) a FP (falošné pozitíva) označujú položky sady obsahujúce šum klasifikované modelom do správnej resp. nesprávnej triedy. . . . .	46
6.2	Vyhodnotenie špecificity modelov . . . . .	59

# Úvod

---

Misia EUSO-SPB (Scotti et al., 2017), aktívna medzi 25.4. až 6.5.2017, realizovala pozorovanie atmosférických javov v oblasti ultrafialového (UV) žiarenia. Primárnym cieľom bolo pozorovanie extenzívnych atmosférických spršok (EAS, z angl. extensive air shower) vyvolaných preletom častíc kozmického žiarenia vysokých energií (UHECR, z angl. ultra-high energy cosmic rays) atmosférou. Misia bola pôvodne plánovaná na okolo 100 dní, no pre technické problémy bola predčasne ukončená po 12 dňoch zrútením do oceánu. Väčšinu dát zaznamenaných počas misie sa však podarilo zachrániť a sú naďalej predmetom vyhodnocovania.

Táto misia bola súčasťou série misií zastrešovaných medzinárodnou kolaboráciou JEM-EUSO. Fenu et al., 2017, s. 7 predpokladajú, že reálna šanca výskytu EAS v týchto dátach je veľmi nízka, okolo 0.8-1.5 EAS počas celého trvania misie. Tieto dáta sú však cenným zdrojom informácií pre doladenie dodatočných plánovaných misií s rovnakým cieľom napr. EUSO-SPB2 (Adams Jr. et al., 2017), Mini-EUSO (Ricci, 2018), ako aj kolaboráciou dlhodobo plánovaný orbitálny detektor JEM-EUSO (Adams et al., 2015) s vysokým rozlíšením. Navyše môžu poslúžiť aj ako užitočný nástroj pri vyhodnocovaní účinnosti algoritmov na detekciu spršok a rekonštrukciu ich parametrov. Motiváciou pre spracovanie dát z týchto misií algoritmami, na rozdiel od manuálneho spracovania je primárne množstvo generovaných dát a aj rýchlosť ich generovania príslušnými detektormi.

Pre analýzu dát v kontexte plánovanej finálnej konfigurácie v podobe detektora JEM-EUSO bolo navrhnutých a odskúšaných množstvo algoritmov použiteľných s menšími úpravami aj pri dátach tejto misie. Nie sú však medzi nimi modely natréňované pomocou metód strojového učenia. Keďže dáta z týchto misií sa dajú interpretovať ako sekvencie snímok, sú v tomto kontexte primárnym kandidátom konvulčné neurónové siete. Tie v posledných rokoch zaznamenali výrazný úspech v oblastiach spracovania dát v podobe snímok alebo videa, najmä vďaka

dostupnosti väčšieho množstva dát na ich tréning a výkonného hardvéru v podobe grafických kariet. Ich existujúce, prípadne navrhované využitia v kontexte iných fyzikálnych experimentov podobného charakteru (Aurisano et al., 2016; Lukas Holch et al., 2017; Erdmann et al., 2017; Nieto et al., 2017) sú dodatočnou motiváciou pre ich odskúšanie aj v kontexte spracovania dát z misie EUSO-SPB.

# 1 Formulácia úlohy

---

Primárnym cieľom tejto práce je overenie vhodnosti modelov založených na konvolučných neurónových sieťach trénovaných metódami strojového učenia pre klasifikáciu dát z misíí zastrešovaných kolaboráciou JEM-EUSO za účelom vyhľadania dát obsahujúcich stopy sprššok v nich. V práci zároveň skúmame možnosti trénovania modelov pomocou rôznych typov simulovaných dát. Pri natrénovaných modeloch sa bude skúmať na týchto sadách simulovaných dát počet správne klasifikovaných záznamov obsahujúcich sprššku na základe hodnôt atribútov sprššky, prípadne pozadia. Zároveň bude skúmaný počet správne klasifikovaných záznamov obsahujúcich iba čistý šum. Okrem týchto je cieľom práce aj aplikácia takehoto modelu alebo modelov na dáta z misie EUSO-SPB a vyhodnotenie dát označených ako spršška. Takýto model alebo modely by mali byť schopné kompenzovať prípadné dočasné výpadky častí meracej plochy detektora, prípadne iné elektronické poruchy.



## 2 EUSO-SPB

---

Misia EUSO-SPB (Scotti et al., 2017) bola súčasťou série skúšobných aj reálnych misií zastrešovaných kolaboráciou JEM-EUSO (Marcelli et al., 2019) s podobným konceptom pozorovania aj vedeckými cieľmi. Primárnym cieľom je zber štatisticky relevantných dát o kozmickom žiarení vysokých energií (UHECR) pomocou pozorovania extenzívnych atmosférických spršok (EAS) vyvolaných interakciou častíc tohto žiarenia s atmosférou. Misia pozostávala z detektora zaveseného na vysokotlakovom stratosférickom balóne (SPB, z angl. Super-pressure balloon) a otočeného pre pozorovanie kolmo nadol. Balón bol vypustený z Nového Zélandu 25.4.2017 pre misiu plánovanú na 100 dní pozorovaní nad južnou časťou Tichého oceánu. Pre technické problémy misia skončila predčasne po 12 dňoch zrušením do oceánu. Dáta boli zaznamenané na palube detektora a periodicky sťahované, vďaka čomu sa ich väčšinu podarilo stiahnuť a sú naďalej predmetom vyhodnocovania.

### 2.1 Motivácia a koncept pozorovania

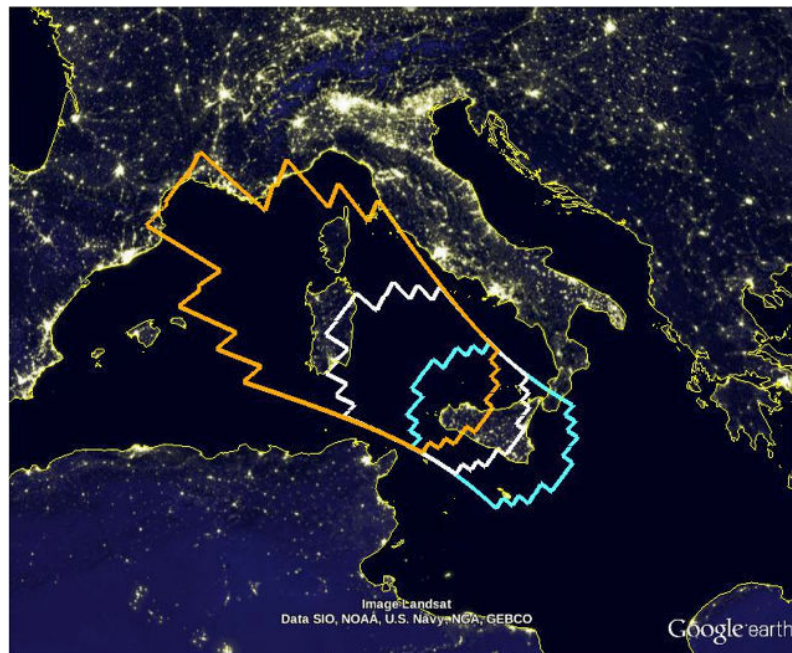
Motiváciou vzniku misie EUSO-SPB a podobných misií sú nedostatky existujúcich spôsobov zachytenia týchto spršok v podobe pozemných sietí detektorov ako je Observatórium Pierre Augera (Aab et al., 2015). Pri nich sa pomocou siete pozemných detektorov zachytávajú na zemský povrch dopadajúce častice žiarenia produkované pri EAS. Z informácií o čase dopadu týchto produktov EAS na jednotlivé detektory siete ako aj nameraných hodnôt energie v nich sa dá rekonštruovať profil EAS ako aj parametre primárnej častice. Takéto metódy sú však obmedzené z hľadiska rozlohy povrchu, ktorú môžu detektory pokrývať. V dôsledku znižujúcej sa frekvencie výskytu týchto javov na jednotku plochy so zvyšujúcou sa energiou primárnej častice je množstvo takýmto spôsobom zachytených javov

nepostačujúce pre ich podrobný výskum. Kvôli tomu bol už dlhšiu dobu považovaný presun pozorovania na detektory na orbite ako ďalší logický krok (Benson et al., 1981).

Základným konceptom týchto misií je použitie detektorov umiestnených vo veľkých výškach nad zemským povrchom, prípadne až na orbite, a otočených smerom k povrchu buď kolmo nadol alebo pod uhlom. Pri interakcii častíc produkovaných EAS s atmosférou vzniká ultrafialové (UV) žiarenie zachytiteľné detektorom umiestneným nad objemom atmosféry, v ktorom dochádza k rozvoju EAS. Detektor zaznamenáva úroveň tohto žiarenia počas fixných časových intervalov. Záznam má podobu snímku, v ktorom jeden pixel reprezentuje intenzitu dopadajúceho žiarenia počas tohto intervalu. Palubná elektronika detektora má v sebe implementovaný hierarchický mechanizmus spúšťačov, ktoré monitorujú potenciálne vzory odpovedajúce EAS. Pokiaľ je v zozbieraných dátach spúšťačmi identifikovaná potenciálna prítomnosť EAS, je uložená sekvencia záznamov s dĺžkou 128 snímok, označovaná ako paket, zachytávajúca postupný rozvoj EAS. Z týchto dát je následne možné procesom rekonštrukcie odhadnúť parametre EAS ako aj primárnej častice ako jej energia, uhol dopadu a podobne.

Dlhodobým cieľom kolaborácie JEM-EUSO je realizácia rovnomenného detektora na orbite s vysokým rozlíšením pre pokrytie čo najväčšej plochy a tým aj množstva zachytených EAS. Tento má byť umiestnený na Medzinárodnej vesmírnej stanici (ISS) alebo samostatne obiehať Zem na nízkej obežnej dráhe. Na obrázku 2.1 je vidieť plochu zachytiteľnú takýmto detektorom umiestneným na stanici ISS pri kolmom a naklonenom pohľade nadol. Pri kolmom pohľade je rozloha územia pokrytého týmto detektorom okolo  $2 \times 10^5 \text{ km}^2$ . Pre porovnanie uvádzame, že celková rozloha povrchu pokrytého detektorom Pierre Augera v Argentíne je okolo  $3 \times 10^3 \text{ km}^2$ , čiže takmer o dva rády menšia.

Medzi výhody misie na orbite patrí aj možnosť relatívne rovnomerného pokrytia takmer celého povrchu planéty použitím inklinovanej obežnej dráhy, akú používa aj stanica ISS. Takto je možné zmerať aj odchýlky vo frekvencii výskytu EAS v závislosti od geografickej polohy, čo v kombinácii s rekonštruovanými atribútmi slúži ako indikátor regiónov oblohy s vyšším výskytom ich zdrojov. Na druhej strane sa stávajú problémom pri takomto type pozorovania zdroje UV žiarenia na pozadí, ktoré nie sú zo spíšky ako sú meteory a svetlá miest, ktoré musia byť spravidla odfiltrované už v rámci detektora.



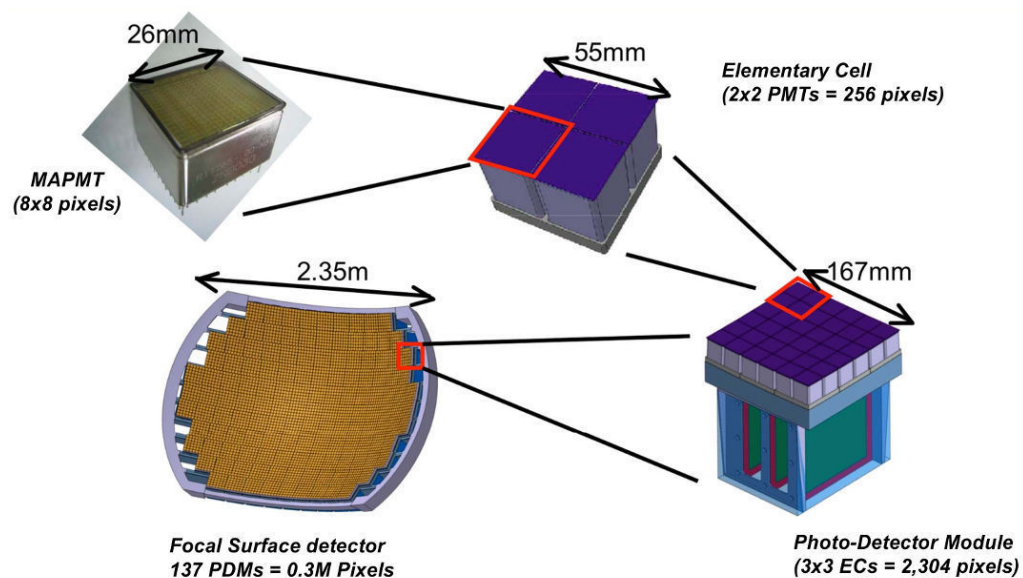
Obr. 2.1: Plocha zemského povrchu snímaná detektorom JEM-EUSO pri kolmom pohľade dolu (nadir mód, tyrkysová) a pri odklone od kolmice o uhol  $20^\circ$  (biela) a  $30^\circ$  (oranžová). Obrázok prevzatý od Adams et al., 2015.

## 2.2 Štruktúra detektora

Detektor misie EUSO-SPB obsahuje základné komponenty plánované aj pre misiu JEM-EUSO. Pre zachytenie žiarenia produkovaného EAS je použitý snímací povrch tvorený multianódovými fotonásobičmi (MAPMT, z angl. Multi-anode photomultiplier tube) s rozlíšením  $8 \times 8$  pixelov citlivými aj na veľmi nízke úrovne UV žiarenia. Žiarenie dopadajúce na ne vyvoláva pomocou fotoelektrického javu veľmi slabé elektrické signály, ktoré tento prístroj zosilňuje a vytvára tak merateľný signál. Vďaka tomu dokážu počítať aj jednotlivé fotóny, ktoré na ne dopadajú. Celkový signál v jednom pixeli nameraný počas fixnej časovej jednotky známej ako GTU, z angl. Gate time unit je následne prevedený do digitálnej podoby, reprezentujúcej hodnotu daného pixelu. Táto hodnota predstavuje 8-bitové celé číslo bez znamienka. Pri misii ako aj plánovanom detektore JEM-EUSO odpovedá jedno GTU  $2,5 \mu s$

Čítanie dát prebieha na úrovni skupín týchto modulov známych ako elementárne bunky (EC, z angl. Elementary cell) tvorených štyrmi MAPMT modulmi v konfigurácii  $2 \times 2$ . 9 EC modulov je zoskupených v konfigurácii  $3 \times 3$  do modulu

označovaného ako fotodetekčný modul (PDM, z angl. Photo-detector module). Ten obsahuje okrem samotných fotonásobičov aj základnú dosku s riadiacou elektronikou a implementáciou spúšťača prvej úrovne. Celý snímací povrch je umiestnený za sadou Fresnelových šošoviek pre sústredenie svetla zo zorného uhla  $\pm 11^\circ$ . Oproti plánovanému detektoru JEM-EUSO bol pri misii použitý len 1 PDM modul. Hierarchiu komponentov detektora aj s ich predpokladaným usporiadaním v plánovanom detektore JEM-EUSO ilustruje obrázok 2.2. Podrobnejší popis palubnej elektroniky poskytujú Osteria et al., 2017.



Obr. 2.2: Hierarchia modulov tvoriacich hardvér detektora EUSO-SPB aj JEM-EUSO. Obrázok prevzatý od Maccarone et al., 2011.

## 2.3 Spúšťače

Pri konfigurácii plánovaného detektora JEM-EUSO uvedenej v časti 2.2 je predpokladaný čistý objem dát produkovaný celou snímacou plochou okolo 1Tbps. Práca v prostredí stanice ISS reprezentuje výrazné obmedzenia na dostupnú šírku pásma vyhradenej pre telemetriu dát s riadiacim strediskom na Zemi ako aj pre maximálny výkon, ktorý môže byť dodaný detektoru. Kvôli tomu je už v rámci detektora potrebná rozsiahla redukcia tohto objemu dát pomocou predbežného vyhodnotenia pravdepodobnosti prítomnosti spŕšky v nich. Toto sa docieľuje pomocou spúšťačov (z angl. triggers). Každý z nich predstavuje algoritmus s na-

staviteľnými parametrami hľadajúci vzory resp. ukazovatele prítomnosti spršky. Rozoznávajú sa dve úrovne spúšťačov:

- Spúšťač prvej úrovne, z angl. First-level trigger (FLT) (Abdellaoui et al., 2017) je implementovaný v riadiacej doske PDM modulu.
- Spúšťač druhej úrovne, z angl. Second-level trigger (SLT) (Bayer et al., 2013) je implementovaný v komponente označovanom ako riadiaca doska klastra (CCB, z angl. cluster control board), ktorá je prepojená s riadiacou doskou PDM modulu.

Riadiace dosky palubnej elektroniky a aj mechanizmus spúšťačov v nich implementovaný bol prítomný aj v misii EUSO-SPB, okrem iného aj kvôli podobným obmedzeniam v danej misii a možnosti otestovania týchto komponentov. V prípade misie JEM-EUSO má spúšťač SLT spracúvať dáta až ôsmich PDM modulov pripojených na príslušnú dosku CCB. Misia EUSO-SPB niesla na palube zmenšenú verziu dosky CCB, ktorá bola použitá aj v predošlej misii EUSO-Balloon (Ballmoos, 2016) pripojenú len k jednému PDM modulu. V oboch misiách však pracuje spúšťač SLT rovnakým spôsobom.

Po aktivácii signálu spúšťača FLT sú doskou CCB načítané dáta z pripojených PDM modulov. Signál spršky v sekvencii snímok vytvára priamku tvorenú pixelmi so zvýšenými hodnotami oproti pozadiu. Preto algoritmus vykonáva nad zozbieranými dátami fitovanie priamky pomocou súčtu hodnôt pixelov v snímkoch pozdĺž preddefinovanej množiny smerov okolo možných počiatkových pozícií stopy spršky. Pokiaľ súčet hodnôt v niektorom z nich presahuje istú prahovú hodnotu, je aktivovaný signál spúšťača SLT a dáta sú posunuté pre uloženie a ďalšie spracovanie.

### 2.3.1 Spúšťač prvej úrovne

Podrobnejšie fungovanie spúšťača FLT, vrátane blokovej schémy jeho implementácie v hardvéri riadiacej dosky opisujú Abdellaoui et al., 2017. Je aplikovaný priebežne na dáta zozbierané modulom PDM. Pracuje na úrovni EC modulov. Hľadaným vzorom je lokálne zvýšený signál oproti predtým vypočítaným prahovým hodnotám pozadia pretrvávajúci v čase. Snímací povrch každého MAPMT modulu tvoriaceho príslušný EC modul, je rozdelený na prekrývajúce sa skupiny pi-

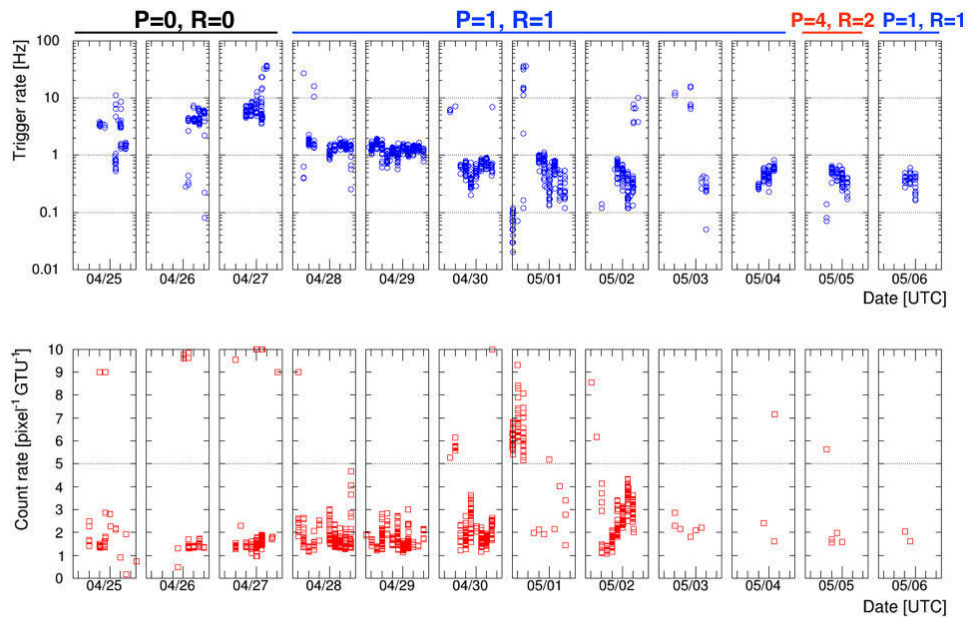
xelov o veľkosti 3x3. Skupiny tvorené pixelmi dvoch susediacich MAPMT modulov nie sú platné. Algoritmus sa aktivuje pokiaľ je zaznamenané lokálne zvýšenie signálu v niektorom z pixelov nad prahovú hodnotu  $n_{pix}^{thr}$  a postupuje nasledovne:

1. V každej skupine, ktorej je pixel súčasťou, sa počas nasledujúcich  $N_{pst}$  GTU sleduje prítomnosť pixelov s hodnotami nad prahovou hodnotou  $n_{pix}^{thr}$ . Zároveň sa sčítavajú hodnoty týchto jasnejších pixelov v skupine, od ktorých je ale odčítaná prahová hodnota pozadia.
2. Pokiaľ počet GTU, počas ktorých je v rámci skupiny minimálne jeden pixel nad prahovou hodnotou  $n_{pix}^{thr}$ , je väčší alebo rovný nastavenej hodnote  $N_{ctd}$  a zároveň súčet presahuje prahovú hodnotu  $n_{pix}^{thr}$  skupiny, je aktivovaný signál pre ďalšie vyhodnotenie.
3. Po prvej aktivácii signálu je sledovaný počas ďalšieho rozsahu  $N_{GTU}$  GTU počet opätovných aktivácií daného signálu v rámci EC modulu. Pre každé GTU s aktívnym signálom sa počet aktivácií inkrementuje o hodnotu 1.
4. Pokiaľ počet aktivácií nepresahuje stanovenú hodnotu  $N_{GTU}^{thr}$  GTU, je aktivovaný signál spúšťača FLT a čaká sa na vyžiadanie dát spúšťačom SLT.

Aktívny signál trvajúci viac ako  $N_{GTU}^{thr}$  GTU je typický pre javy nesúvisiace s EAS ako sú meteory alebo blesky. Hodnoty  $N_{GTU}$  a  $N_{GTU}^{thr}$  boli stanovené na 73 a 72 GTU. Pri misii bolo použitých niekoľko rôznych nastavení parametrov  $N_{pst}$  a  $N_{ctd}$ , ako ukazuje obrázok 2.3. M. E. Bertaina et al., 2017, od ktorých je obrázok 2.3 prevzatý, analyzovali výkon spúšťača prvej úrovne v tejto misii. Ich záver bol, že funguje v rámci obmedzení kladených na misiu JEM-EUSO, hoci nižšia frekvencia aktivácie by bola vhodná kvôli obmedzeniam telemetrie, ktoré boli prítomné aj pri misii EUSO-SPB. Podotýkajú, že časy, keď stúpla frekvencia aktivácie sa zrejme dajú vysvetliť šumom v elektronike. Wiencke et al., 2017 uvádzajú vo svojej práci ako možný zdroj aktivácií spúšťačov aj zasiahnutie samotného detektora nízkoenergetickými časticami kozmického žiarenia.

## 2.4 Priebeh misie a zozbierané dáta

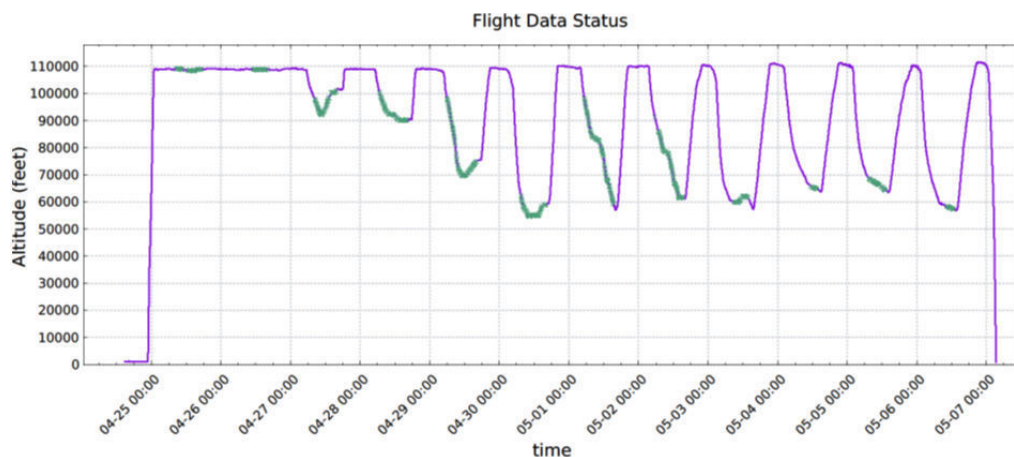
Podrobnejší priebeh misie ako aj prvotný pohľad na dáta možno nájsť v práci od Wiencke et al., 2017. Balón nesúci detektor bol vypustený z novozélandského



Obr. 2.3: Horný rad: frekvencia aktivácie spúšťača FLT počas misie EUSO-SPB. Parametre  $P$  a  $R$  odpovedajú hodnotám  $N_{pst}$  a  $N_{ctd}$  použitým v logike spúšťača. Dolný rad: priemerná intenzita pixelov v zaznamenaných paketoch snímkov. Hodnoty  $P = 4$  a  $R = 2$  odpovedajú parametrom pre misiu JEM-EUSO. Obrázok prevzatý od M. E. Bertaina et al., 2017

mesta Wanaka 25.4.2017. Na tretí deň po vypustení misie sa ukázalo, že z balóna zrejme unikol plyn, ktorý ho nadnášal. Výsledkom bolo, že namiesto plánovanej stabilnej výšky 30 km nad morom balón striedavo stúpol a klesol počas dňa podľa okolitej teploty, niekedy aj s rozdielom až 14 km. Vývoj výšky letu možno vidieť na obrázku 2.4. Misia vykonávala pozorovania iba počas nocí bez mesačného svitu, pričom počas väčšiny misie toto bolo v obdobiach, kedy zostúpil balón do nižšej výšky letu oproti plánu.

Dáta uložené detektorom po prepustení mechanizmom spínačov sú sekvencie snímkov reprezentujúcich pozorovanie s dĺžkou 128 GTU, ďalej označované ako pakety. Každý snímok paketu obsahuje 8-bitové bezznamienkové celočíselné hodnoty. Celkove boli počas misie uložených na lokálnych pevných diskoch nesených inštrumentom pakety s celkovým trvaním asi 40 hodín. Z toho 30 hodín, alebo 75% dát bolo stiahnutých pred zrútením balóna do oceánu. Celkove tieto pakety



Obr. 2.4: Vývoj výšky letu balóna počas misie EUSO-SPB. Hrubou zelenou sú označené obdobia, počas ktorých prebiehal zber dát. Obrázok prevzatý od Wiencke et al., 2017

predstavujú okolo 60 GB dát.

Ako Fenu et al., 2017 uvádzajú, predpokladané množstvo spŕšok zachytiteľných počas výsledného trvania tejto misie bol na úrovni 0.8-1.5 spŕšok. Keďže je nepomerne viac dát ako by tento predpoklad čakal, je vidieť, že spúšťače samotné nie sú samy osebe postačujúce pre definitívne rozlíšenie prítomnosti spŕšok v dátach a je potrebné ich doplniť o analýzu dát aj mimo nich.

## 2.5 Softvérový rámec ESAF

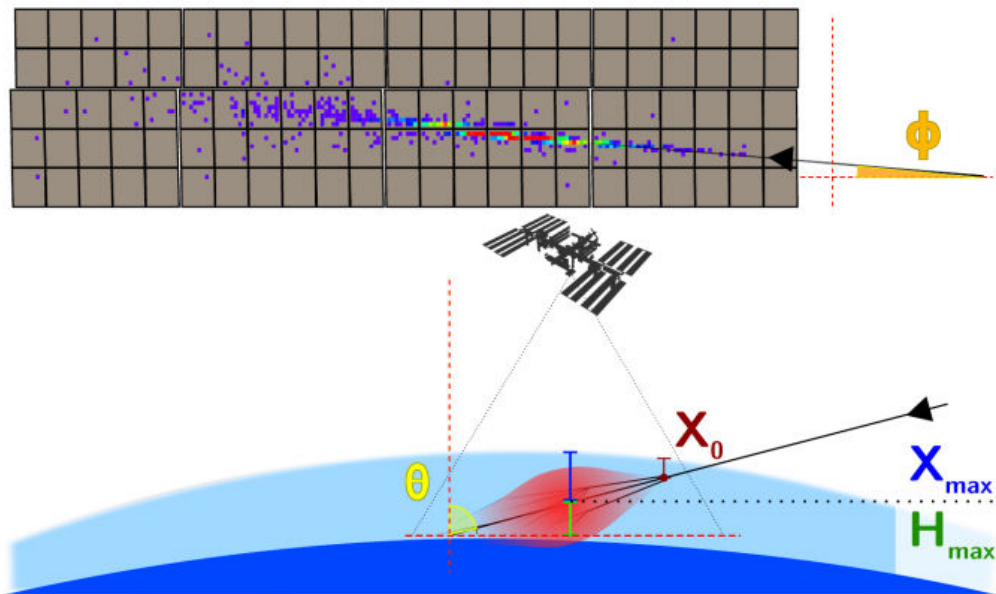
Pre účely simulácie misií kolaborácie JEM-EUSO a analýzy ich očakávaných vlastností pri plnení ich vedeckých cieľov bol vyvinutý softvérový rámec EUSO pre simuláciu a analýzu (ESAF, z angl. EUSO simulation and analysis framework). Rámec ESAF je koncepčne rozdelený na dva hlavné moduly: simulačný (Simu) a rekonštrukčný (Reco).

Simulačný modul je zameraný na vytváranie dát metódami schopnými simulovať fyzikálne procesy sprevádzajúce rozvoj atmosférickej časticovej spŕšky. Tieto procesy zahŕňajú interakcie častíc spŕšky s atmosférou a generovanie fotónov, ich prechod atmosférou k detektoru, simuláciu samotného hardvéru teleskopu, interakciu fotónov s modulmi detektora ako aj aplikáciu mechanizmu spúšťačov na príslušné dáta. Pre jednotlivé fázy je možné použiť viacero vzájomne zame-



niteľných modulov prípadne importovať údaje z externých modulov. Dáta vytvorené týmito modulmi odzrkadľujú fyzikálne reálie danej fázy vývoja spršky s viac alebo menej presnou schopnosťou ich aproximácie a často s tým spojenými rozdielmi aj vo výpočtovej náročnosti.

Rekonštrukčný modul číta dáta vytvorené simulačným modulom a skúša z nich rekonštruovať parametre spršky ako aj primárnej častice v niekoľkých fázach. Úlohou je rekonštrukcia nasledujúcich atribútov, graficky znázornených na obrázku 2.5:



Obr. 2.5: Parametre atmosférickej spršky. Obrázok prevzatý od Vrábel, 2017

- Smer príchodu, určený pomocou zenitového ( $\theta$ ) a azimutálneho uhla ( $\phi$ ).
- Hĺbka atmosféry meraná od jej vrchu v bode stopy s maximálnou produkciou sekundárnych častíc ( $X_{max}$ ).
- Výška tohto bodu nad povrchom ( $H_{max}$ ).
- Energia primárnej častice ako aj potenciálne jej typ.

Prvá fáza, nazývaná tiež rozpoznávanie vzorov (z angl. Pattern recognition), je o segmentácii obrazu. Vyžaduje sa pri nej oddeliť signál spršky od pozadia selekciou pixelov patriacich stope spršky v sekvencii snímok z detektora. Ďalšou fázou je uhlová rekonštrukcia spršky, kde použitý modul sa pokúša z tejto čistej

stopy a údajov o časovej následnosti snímok vypočítať aj smer vývoja spršky určený zenitovým ( $\theta$ ) a azimutálnym ( $\phi$ ) uhlom. Následne iný modul z takto rekonštruovaných dát sa pokúša odvodiť aj bod maximálnej intenzity vo vývoji spršky a jeho výšku nad povrchom a následne aj energiu primárnej častice. Kvôli sekvencnému vykonávaniu týchto fáz rekonštrukcie je potrebné povedať, že výber a nastavenie parametrov algoritmu skoršej fázy vie často silno ovplyvniť aj kvalitu rekonštrukčného procesu v ďalších fázach.

Oba moduly zdieľajú pre svoju činnosť iba informácie o atmosfére ako aj o optike teleskopu, nakoľko aj v reálnych podmienkach misie sa predpokladá, že jediný dodatočný parameter k dispozícii pre rekonštrukčné algoritmy popri dátach z detektora budú iba informácie o atmosfére zachytené monitorovacím systémom teleskopu. Takéto rozdelenie je potrebné pre zaistenie dôkladnosti testov. V rámci nie je presne stanovená fáza klasifikácie konkrétneho paketu a overenie, Táto úloha je ale najviac podobná fáze segmentácie, keďže uhlová rekonštrukcia pracuje s extrahovaným signálom.

### 2.5.1 Rekonštrukcia spršky - existujúce algoritmy

Ako bolo spomenuté vyššie, rekonštrukcia spršky sa skladá z niekoľkých fáz, preto je potrebné povedať, že jednotlivé algoritmy resp. moduly tu spomenuté sa dajú kategorizovať aj podľa fázy, kde sú použité. Vo fáze segmentácie sú to techniky PWISE (Guzman et al., 2013) a LTTPreClustering (M. Bertaina et al., 2014, s. 24-25). Primárne algoritmy pre fázu uhlovej rekonštrukcie sú založené na redukovani problému určenia smeru spršky priestorovo do dvoch rozmerov pomocou konceptu označovaného v angl. literatúre ako Track detector plane (TDP). Ide o myslenú rovinu, ktorá obsahuje samotnú stopu spršky ako aj detektor na orbite. Merrik, 2014 podrobne popisuje tento koncept ako aj samotné algoritmy. Pre poslednú fázu rekonštrukcie je v súčasnosti dostupný len jeden modul PmtToShowerReco (Fenu, 2013).

## 3 Strojové učenie v časticovej fyzike

---

Pokusy o využitie metód strojového učenia v časticovej fyzike nie sú novým trendom (Handzel et al., 1993). Ich motiváciou je pomocou reálnych alebo simulovaných dát natrénovať model, ktorý dokáže reprezentovať fyzikálnu interakciu a analyzovať dáta z experimentov automaticky a bez potreby ručného vytvorenia takéhoto modelu pomocou programového kódu. V časticovej fyzike boli tieto metódy už skôr minimálne skúšané ak nie použité napr. v podobe viacvrstvových perceptrónov (MLP, z angl. Multi-layer perceptron) (Handzel et al., 1993) prípadne klasifikátorov SVM s dátami z Tevatronu vo Fermilabe (Vaiciulis, 2002) alebo experimentu OPAL vykonávaným vo švajčiarskom CERNe (Vannerem et al., 1999).

Jestvuje ale aj široké spektrum meracích prístrojov v časticovej fyzike, ktorých výstupom je v podstate obraz fyzikálnej interakcie v experimentoch. Medzi ne patria napr. vzorkovacie kalorimetre so scintilátormi, časové projekčné komory s tekutým argónom alebo na vode založené detektory Čerenkovovho žiarenia, ktoré zaznamenávajú drobné množstvá energie naprieč objemom detektora. Tieto predstavujú materiál, ktorý je možné analyzovať pomocou prostriedkov počítačového videnia.

Základným typom neurónovej siete je práve MLP. Je tvorený minimálne dvoma plne prepojenými vrstvami neurónov tzn., že každý neurón je pripojený na každý výstup predošlej vrstvy. Hoci MLP sa dajú aplikovať aj na obrazové dáta, majú jednu principiálnu nevýhodu, ktorou je vysoký počet nezávisle trénovaných parametrov pri učení. Obrazové dáta aj v prípade napr. jedinej snímky veľkosti  $100 \times 100px$  bez farebných kanálov predstavuje 10000 vstupov pre prvú vrstvu. Keď uvažíme prvú vrstvu tvorenú iba 128 neurónmi znamená táto skutočnosť, že sieť má až takmer 1 280 000 nezávisle nastaviteľných váh, inými slovami, ide už o veľmi komplexný model. Pridanie farebných kanálov v obraze a zvýšenie jeho rozlíšenie

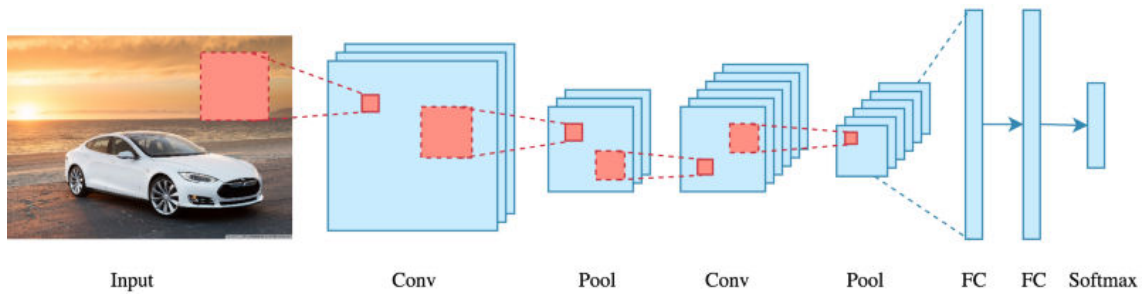
iba zvyšujú komplexnosť modelu.

Komplexné modely sú náchylnejšie na preučenie (z angl. *overfitting*), ktoré sa prejavuje veľmi zlou schopnosťou natrénovaného modelu reprezentovať dáta mimo trénovacej sady. Na zabránenie preučenia sa vyžaduje vyšší počet trénovacích dát. Toto je možné dosiahnuť augmentáciou trénovacej sady napr. rotáciou alebo orezaním a zväčšením snímok a ich pridaním do trénovacej sady ako nové položky. Väčšia sada však znamená aj dlhší čas tréningu. Dajú sa použiť aj špeciálne techniky ako *dropout* (Srivastava et al., 2014) alebo regularizácia pomocou normy L2 (Laarhoven, 2017). Ale ani pri kombinácii týchto metód nie je vhodné ani praktické na snímky s vysokým rozlíšením aplikovať modely založené na architektúre MLP. Naopak, veľmi dobré výsledky boli dosiahnuté v poslednej dobe pomocou konvolučných neurónových sietí.

### 3.1 Konvolučné siete

Základná štruktúra konvolučných neurónových sietí je inšpirovaná štruktúrou neurónov v zrakových centrách mozgu. Typická architektúra konvulčnej siete používanej pre klasifikáciu snímok do vopred danej množiny disjunktných tried má dve základné časti: konvulčnú a plne prepojenú. Príklad takejto architektúry je možné vidieť na obrázku 3.1. Konvulčnú časť tvoria vrstvy spracúvajúce trojrozmerné vstupy a vytvárajúce trojrozmerné výstupy. Tieto slúžia na extrakciu relevantných príznakov z obrazových dát od jednoduchých geometrických tvarov a farebných gradientov až po objekty v obraze ako napr. koleso auta, oči, tvár a pod. Zároveň môžu vrstvy tejto časti znižovať rozmer pôvodného vstupu z hľadiska jeho výšky a šírky alebo aplikovať rôzne normalizačné techniky na ich vstup.

Plne prepojená časť siete je spravidla tvorená plne prepojenými vrstvami a odpovedá klasickému MLP. Veľká väčšina parametrov, ktoré sa sieť učí, sa nachádza medzi prvou plne prepojenou vrstvou a predchádzajúcou vrstvou. V závislosti od úlohy majú tieto plne prepojené vrstvy rôzne štruktúry, no pri úlohách klasifikácie snímok sa na výstupe používajú plne prepojené vrstvy s počtom neurónov rovným počtu tried. Aktivačnou funkciou poslednej vrstvy v takejto úlohe je spravidla *softmax*. Jej výstupom sú pravdepodobnosti jednotlivých tried, pričom vektor týchto hodnôt, ktorý je výstupom poslednej vrstvy dáva pri sčítaní hodnotu 1.

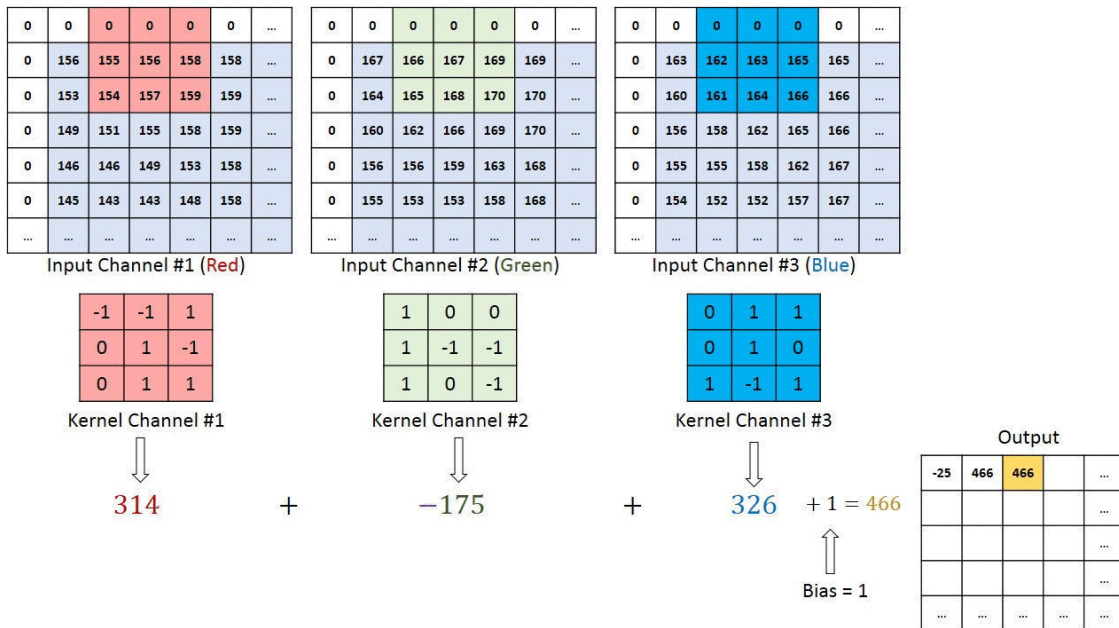


Obr. 3.1: Príklad typickej architektúry konvolučnej neurónovej siete použitej pre klasifikáciu obrázkov. Konvolučná časť siete je tvorená vrstvami označenými ako Conv (extrakcia príznakov) a Pool (zmenšenie vstupu). Plne prepojená časť je tvorená plne prepojenými vrstvami označenými ako FC (z angl. Fully connected) a Softmax. Obrázok prevzatý od Dertat, 2017.

### 3.1.1 Konvolučné vrstvy

Základom konvolučnej časti je rovnomenný typ neurónovej vrstvy. Na rozdiel od plne prepojených vrstiev tradičného MLP, pri konvolučných vrstvách nemajú neuróny prepojenie na celý vstup ale len na jeho malú časť označovanú ako receptívne pole (z angl. receptive field). Ak je vstupom konvolučnej vrstvy trojrozmerná matica s rozmermi  $w$  (šírka),  $h$  (výška) a  $d$  (hĺbka), potom receptívne pole reprezentuje úsek obrazu v podobe matice s rozmermi  $f_w$ ,  $f_h$  a  $d$ . Sada váh aplikovaná na takúto časť vstupu sa nazýva filter prípadne kernel a jeho úlohou je hľadať v tomto vstupe príslušný príznak. Možno si všimnúť, že hoci je aplikovaný len na časť vstupu z pohľadu výšky a šírky, je aplikovaný na celú hĺbku vstupu  $d$ . Okolo výšky a šírky vstupu je možné pred aplikáciou konvolúcie vytvoriť akúsi výplň alebo „čalúnenie“ (z angl. padding) pomocou napr. nulových hodnôt. Koncept výplne ako aj výsledok operácie konvolúcie vykonanú každým filtrom ilustruje obrázok 3.2.

Ako vidno z obrázka 3.2, základná operácia vykonaná touto vrstvou je súčet súčinov prvkov na rovnakých pozíciách v príslušnom receptívnom poli a sade váh. Nad výsledným súčinom je aplikovaná aktivačná funkcia neurónov danej vrstvy. Pre zlepšenie rozlišovacej schopnosti sa často používajú nelineárne aktivačné funkcie, pričom v súčasnosti prevažuje v tomto smere ReLU (z angl. Rectified linear unit). Tá výstup z príslušného filtra nastaví na nulu, ak je záporný, inak ho ponechá nezmenený.



Obr. 3.2: Príklad konvolučnej operácie aplikovanej na vstupný obraz s tromi farebnými kanálmi. Obrázok prevzatý od Machine Learning Guru, 2017.

Rovnaká sada váh je aplikovaná naprieč celou výškou a šírkou vstupu vrátane prípadnej výplne okolo. Výsledkom aplikácie jediného filtra nad vstupom je dvojrozmerné pole hodnôt s rozmermi  $w_2, h_2$ . Tento výstup sa nazýva mapa príznakov (z angl. feature map), pričom hodnoty v ňom reprezentujú mieru prítomnosti príznaku hľadaného príslušným filtrom vo vstupe. Konvolučná vrstva má spravidla viac než jeden filter, každý hľadajúci iné príznaky. Preto je výstupom konvolučnej vrstvy trojrozmerná matica s rozmermi  $w_2, h_2$  a  $n_f$ , tvorená mapami príznakov, ktorej hĺbka  $n_f$  odpovedá počtu filtrov vrstvy. Rozmery  $w_2$  a  $h_2$  závisia od veľkosti vstupu, filtrov a nastavení hodnôt posunov podľa vzorcov 3.1 (Machine Learning Guru, 2017), kde  $p_x$  a  $p_y$  reprezentujú rozdiel medzi šírkou resp. výškou vstupu s výplňou a bez výplne.

$$w_2 = \frac{w - f_w + 2 * p_x}{s_x} + 1$$

$$h_2 = \frac{h - f_h + 2 * p_y}{s_y} + 1$$
(3.1)

Každú mapu príznakov je možné vizualizovať ako snímok bez farebných kanálov, často vo falošnej farbe. Mapy príznakov vytvárané hlbšími vrstvami nemožno priamo interpretovať, keďže nie je vopred známe, aký príznak reprezentuje daná

mapa. Zeiler et al., 2013 však vyvinuli spôsob interpretácie pomocou techniky označovanej ako dekonvolúcia. Táto mapuje hodnoty mapy príznakov na regióny vstupu, ktoré spôsobili zvýšené hodnoty v danej mape. Kým na najnižšej vrstve priamo pripojenej na vstup sú extrahovanými príznakmi jednoduché textúry ako sú hrany a farebné vzory, na vyšších vrstvách už sieť hľadá komplexnejšie vzory ako napr. zaoblené tvary, psie uši, a napokon aj samotného psa.

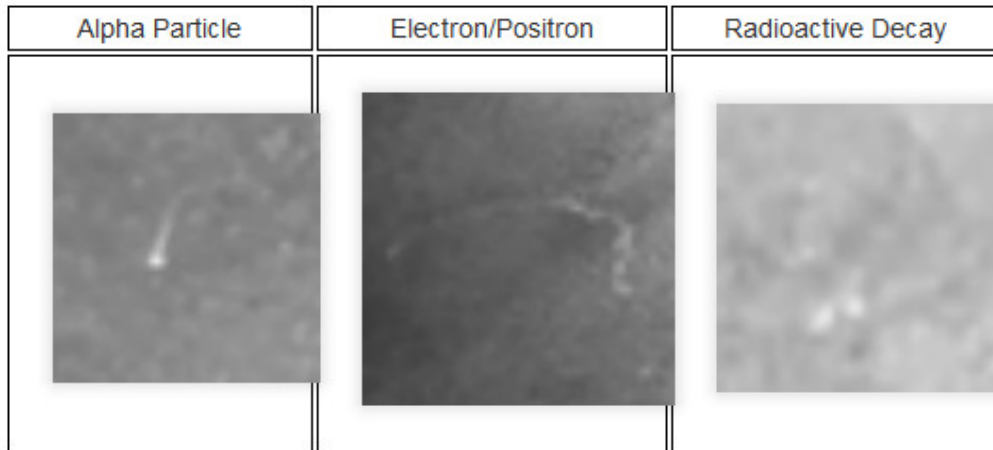
### 3.1.2 Vzorkovacie vrstvy

Okrem konvolučných sa používajú aj tzv. vzorkovacie (z angl. pooling) vrstvy, ktoré slúžia na explicitné zmenšenie veľkosti vstupu v horizontálnom a vertikálnom smere pri zachovaní hĺbky aplikovaním funkcie s jedinou výstupnou hodnotou na neprekrývajúce sa regióny vstupu s rozmermi  $h_p \times w_p$  pre každý kanál pozdĺž hĺbky  $d$ . Myšlienkou na ich použitie je, že absolútna pozícia niektorého príznaku nie je natoľko dôležitá ako jeho relatívna pozícia voči iným príznakom pre vyhľadávanie komplexnejších štruktúr. Napr. pozícia očí vo vstupnom obraze nie je natoľko dôležitá ako ich relatívna pozícia voči nosu alebo ústam pri hľadaní tváre v snímku. Na druhej strane, príliš časté využitie týchto vrstiev môže viesť k strate relevantných informácií. Najviac používaný typ vzorkovania je vzorkovanie maximálnych hodnôt (z angl. max-pooling) kde sa jednoducho vyberie maximálna hodnota z tohto regiónu.

## 3.2 Konvolučné siete v časticovej fyzike

Konvolučné siete ako prostriedok dátovej analýzy v časticovej fyzike boli skúmané pri viacerých projektoch. Jedným z jednoduchších príkladov, ktorý sa však podobá v mnohých aspektoch našej úlohe, a preto sme na ňom založili aj naše riešenie, je model pre klasifikáciu častíc, ktorý navrhol Lincoln, 2017. Úloha, ktorú riešil, bola klasifikácia častíc do troch rôznych tried na základe tvaru stôp nimi zanechaných v monochromatických snímkoch z hmlovej komory. Snímky boli manuálne extrahované z videozáznamu interakcií častíc v hmlovej komore, zmenšené na veľkosť  $50 \times 50$  pixelov a označené ako stopy produkované alfa časticami, elektrónmi prípadne pozitronmi a rádioaktívnym rozpadom. Príklad snímky s jednotlivými typmi stôp je možné vidieť na obrázku 3.3 Dodatočne boli aplikované

techniky pre augmentáciu sady ako napr. zrkadlové prevrátenie, rotácia a rôzne ďalšie transformácie pre zväčšenie sady.



Obr. 3.3: Snímky zachytávajúce stopy zanechané troma klasifikovanými typmi častíc v pokuse s hmlovou komorou. Obrázok prevzatý od Lincoln, 2017.

Tento experiment nebol vykonávaný v rámci formálneho výskumu, no podobné rozlíšenie snímok ako snímky v dátach z misie EUSO-SPB ( $48 \times 48$  pixelov) a podobný počet sieťou klasifikovaných tried ako pri našom probléme nám vyzdvihol tento príklad ako možného kandidáta na riešenie. Autor podotkol, že aj po augmentácii bolo v dátovej sade len niekoľko stoviek snímok, čo je menej než ideálne pre tréning. Autorom skúšané siete mali dve konvolučné vrstvy, každá nasledovaná vrstvou pre vzorkovanie maximálnych hodnôt, a plne prepojenú časť tvorenú 2 skrytými plne prepojenými vrstvami. Okrem nich trénoval aj model založený na architektúre MLP. Všetky modely boli trénované na 85% dát príslušnej sady, kým zvyšných 15% bolo vyhradených ako validačná sada pre vyhodnotenie modelu. Najlepšia úspešnosť predpovedí konvolučného modelu bola 62%, čo autor pripisuje nedostatku trénovacích dát.

### 3.2.1 Návrhy pre pozemné detektory kozmického žiarenia

Nieto et al., 2017 sa pokúšali využiť konvolučné siete na vytvorenie klasifikátora dát z plánovaného observatória Čerenkovove pole teleskopov (z angl. Cherenkov telescope array, CTA). Observatórium je zamerané na hľadanie atmosférických spŕšok vyvolaných gama žiarením zachytávaním vydávaného Čerenkovovho žiarenia. Autormi navrhovaný klasifikátor mal byť schopný rozlíšiť v simulovaných



dátach z jediného detektora spršky vyvolané gama žiarením od spršok vyvolaných hadrónovým kozmickým žiarením. Pri svojich pokusoch testovali dve architektúry konvolučných sietí: Inception V3 (Szegedy; Vanhoucke et al., 2015) a ResNet (He et al., 2015). Obe predstavovali modifikované verzie pôvodných architektúr GoogLeNet (Szegedy; Liu et al., 2015) a ResNet (Xiangyu Zhang et al., 2015) a na ich tvorbe sa minimálne sčasti zúčastnili aj pôvodní autori pôvodných architektúr.

Namiesto pôvodnej výstupnej vrstvy použili binárny klasifikátor a zároveň pred plným tréningom urobili niekoľko skúšobných behov pre výber najlepšieho optimalizačného algoritmu pre spätnú propagáciu pre ich úlohu. Simulované dáta pozostávali z vyše 4 miliónov simulovaných snímok a boli rozdelené na tréningovú, validačnú a testovaciu sadu v pomere 80% ku 10% ku 10%. Validáčna sada slúžila na priebežnú kontrolu pri tréningu a monitorovanie náznakov preučenia. Testovacia sada slúžila na overenie finálneho natrénovaného modelu. Pri ich pokusoch vyšli pre obe siete približne podobné výsledky, hoci úspešnosť klasifikácie bola vyššia o niekoľko desiatín percenta pri modeloch založených na architektúre Inception V3 pri nižších aj vyšších energiách spršky. Výsledky zároveň ukazovali autormi očakávaný nárast presnosti pri zvyšovaní energie spršky, ktorá sa prejavuje jasnejšími stopami v obrazoch a tým aj vyšším kontrastom. Kým pre nižšie úrovne energie bola presnosť na úrovni okolo 81%, pre najintenzívnejšie spršky stúpila presnosť až na 91%.

Erdmann et al., 2017 použili konvolučné siete pri rekonštrukcii spršok vyvolaných kozmickým žiarením v simulovaných dátach z pozemných detektorov. Riešili rovnakú úlohu ako algoritmy pre rekonštrukciu implementované v softvérovom rámci ESAF, ale v dátach z pozemných polí detektorov ako je napr. Observatórium Pierre Augera. Ich sieť využívala podobne ako ResNet "skratky" medzi konvolučnými vrstvami, a prijímala dva vstupy. Prvý predstavoval obraz poľa detektorov, kde každý "pixel" obsahoval dve hodnoty: celkovú energiu z častíc dopadajúcich na konkrétny detektor a čas dopadu prvých častíc. Druhý vstup reprezentoval časový priebeh vývoja spršky a v podstate by sa dal charakterizovať ako krátka sekvencia obrazov, čiže video. Kým prvý vstup bol ihneď použiteľný sieťou, druhý vstup bol najprv spracovaný niekoľkými 3D konvolučnými vrstvami, ktoré fungujú rovnako ako bežné konvolučné vrstvy ale majú rozmer navyše a následne výstup z nich bol upravený do vhodného tvaru a zlúčený s prvým vstupom pred

postúpením zvyšku siete. Tento na výstupe dával tri hodnoty: celkovú energiu, uhol dopadu ako aj výšku bodu s najväčšou produkciou energie ( $X_{max}$ ).

Pri prvých dvoch atribútoch autori skúšali aj výkon siete pri absencii druhého prúdu dát a minimálne v prípade rekonštrukcie uhla dopadu nezaznamenali púť výrazný pokles presnosti. Pri rekonštrukcii energie bolo ale viditeľný výrazné zlepšenie presnosti pri dostupnosti oboch dátových prúdov. Pri rekonštrukcii hodnoty  $X_{max}$  sú relevantné informácie uložené v druhom prúde, preto ho nie je možné vynechať. Pri všetkých troch hodnotách zaznamenali hodnoty rozlíšenia a citlivosti siete, ktoré boli veľmi uspokojujúce.

### 3.2.2 Detektor NOvA

Aurisano et al., 2016 vo svojej práci popisujú svoj pokus s vytvorením klasifikátora založenom na spomenutom GoogLeNet (Szegedy; Liu et al., 2015) pre klasifikáciu výstupov z detektora neutrín NOvA v americkom Fermilabe. Základným prvkom oboch detektorov v experimente je vzorkovací kalorimeter, ktorý sa skladá zo za sebou poukladaných rovín tvorených stĺpcami naplnenými scintilačnou kvapalinou v ktorých je kus neizolovaného optického vlákna ktoré zachytáva vyžarovanie energie z kvapaliny v prítomnosti časticovej spŕšky. V rámci roviny sú stĺpce usporiadané buď horizontálne alebo vertikálne, pričom toto usporiadanie sa striedavo mení od jedného ku druhému koncu detektora. Pri interakcii neutrín s detektorom vznikajú podľa typu neutrína rôzne typy sekundárnych častíc ako aj žiarenie rôznej úrovne, ktoré vytvárajú charakteristické vzory alebo stopy pri pohľade zhora resp. z boku na snímané dáta detektora. Práve rozlíšenie typu interakcie a na základe toho aj primárnej častice z týchto vzorov bol primárny cieľ výskumu.

Pre reprezentáciu dát boli použité výrezy zaznamenaných dát pri pohľade zhora a z boku s pevnou šírkou a výškou vycentované na oblasť pravdepodobnej primárnej interakcie častice s detektorom. Ako uviedli autori, na rozdiel od tradičných úloh pre konvolučné siete, kde vstupom je len jeden obrázok s viacerými farebnými kanálmi, v tomto prípade ide o dve rôzne pohľady na rovnakú interakciu. Súčasne s tým sa pre kódovanie jednotlivých pixelov reprezentujúcich hodnoty zaznamenatej energie v danej časti detektora využili 8-bitové celé čísla bez znamienka namiesto 32 bitových čísel v pohyblivej rádovej čiarky, kvôli úspore úložného priestoru aj pamäťovej náročnosti, pričom autori nezaznamenali žiaden výrazný posuv v presnosti reprezentácie množstva energie.

Kvôli týmto skutočnostiam uskutočnili niekoľko úprav oproti pôvodnému GoogLeNet vo svojej verzii tejto siete nazývanej CVN. Prvou z nich bolo vytvorenie dvoch vstupných prúdov, kde každý z nich spracúva príslušný pohľad zhora alebo z boku. Taktiež keďže snímky majú jednoduchšiu štruktúru ako plne farebné obrázky použité pri zvyčajných úlohách využívajúcich konvolučné siete, využitie extra hlbkej architektúry v podobe GoogLeNet neprináša žiadne zlepšenie v presnosti, preto bola hĺbka siete skrátaná. Výstupy z oboch konvolučných prúdov sú následne zlúčené pomocou jedného Inception modulu, na ktorý následne nadväzuje pooling vrstva vyberajúca priemerné hodnoty z regiónov s definovanou veľkosťou a za ňou klasifikátor softmax.

Treba povedať, že hoci striktne sú definované 4 typy interakcií, kde posledný nie je použiteľný pri identifikácii typu neutrína, prvé 3 z nich sa dajú rozdeliť aj na 4 podtypy, čo vyúsťuje až do 13 tried pre klasifikáciu. Autori sa tu rozhodli pri vyhodnotení zosumarizovať pravdepodobnosti pre jednotlivé podtypy z klasifikátora typu softmaxu ako výslednú pravdepodobnosť rozpoznávania siete. Hoci sieť podľa nich sa ukazovala byť sľubná aj pri jemnozrnnejšom rozpoznávaní, bolo pre začiatok vhodnejšie sa zamerať na hrubozrnnejšiu identifikáciu. Pri tréningu a vyhodnotení boli použité simulované dáta v podobe 4,7 milióna snímok rozdelené v pomere 80% ku 20% na tréningové a testovacie dáta. Táto množina dodatočne bola upravená napr. pridaním Gaussovho šumu so štandardnou odchýlkou 1% aby viac odpovedala skutočným zaznamenaným dátam.

Pri svojich pokusoch zaznamenali autori, že neurónová sieť nebola horšia ako existujúci identifikačný algoritmus pri interakciách muónových neutrín, ktoré sa pomerne ľahko odlišujú od ostatných tried. Na druhej strane, pri interakciách elektrónových neutrín zaznamenali výrazný pokrok v presnosti oproti existujúcemu algoritmu až o 14%. Základný poznatok, ktorý autori odvodili, je že konvolučné siete môžu efektívne pracovať aj s obrazmi nie prirodzeného pôvodu ako sú výstupy zo vzorkovacieho kalorimetra.

## 4 Implementácia riešenia

---

Pre použitie metód strojového učenia sú dáta potrebné pre samotný tréning modelu, pričom táto sada dát nesmie byť identická so sadou dát použitých na vyhodnotenie natrénovaných modelov. Pri prvotných pokusoch bola použitá sada dát vytvorená v rámci spracovania dát metódou využívajúcou extrakciu príznakov (Vrábel, 2019). Táto naniesla simulovanú spříšku vytvorenú v rámci ESAF na vybrané vzorky pozadia paketov z misie. V začiatkoch sme narazili na komplikácie, kedy ani jedna z použitých sietí nebola pri tréningu schopná rozpoznávať hľadané vzory v dátach. Keďže sa ukázali aj chyby v samotnej metóde generovania týchto simulovaných dát, boli nami skúšané siete v tejto fáze tréňované iba pomocou jednoduchých dát generovaných vlastným spôsobom.

Po doladení architektúr sme tento alternatívny spôsob generovania dát, podrobnejšie opísaný v časti 4.5, ponechali, čiastočne rozvinuli a používali ho naďalej, okrem iného aj kvôli relatívne nízkej náročnosti výpočtu pri ich vytváraní v porovnaní s rámcom ESAF. Siete natrénované na týchto dátach navyše vykazovali pri predbežnom vyhodnotení dát z letu vcelku uspokojivé výsledky, kvôli čomu sme sa rozhodli preskúmať vhodnosť takýchto dát pre tréning sietí samy osebe prípadne augmentované simulovanými dátami. V súhrne boli pri tréningu aj vyhodnotení použité nasledujúce typy dát:

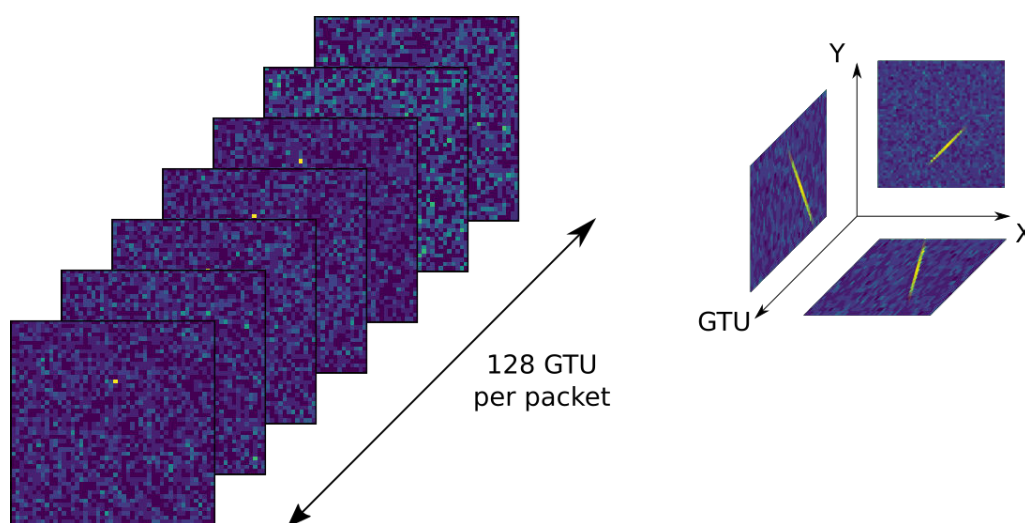
1. Simulované dáta vytvorené pomocou rámca ESAF, kde pozadie je použité z vybraných dát z misie, ďalej len **simulované** dáta.
2. Simulované dáta vytvárané vlastnou metódou, kde pozadie je vytvorené pomocou Poissonovej distribúcie hodnôt v pakete, ďalej označované pre odlíšenie ako **syntetické**, kvôli umelej povahe pozadia.

V rámci riešenia sme dodatočne vyvinuli sadu nástrojov všeobecnejšieho použitia pre prácu s oboma typmi dát ako aj pre ich použitie pri tréningu a vyhod-

notení modelov. Taktiež boli vyvinuté dve podporné nástroje pre vyhodnotenie klasifikačných metrík modelov tréovaných na oboch typoch dát. Všetky nástroje boli implementované ako skripty s rozhraním v príkazovom riadku. Dodatočne pri ich návrhu zavážila myšlienka, aby využívali výstup jeden druhého na vykonanie svojej úlohy.

## 4.1 Formát dát

Špecifickou charakteristikou nami zvoleného postupu je okrem iného aj použitý formát dát posúvaných pri tréningu aj vyhodnotení modelov. Kým väčšina algoritmov pre rekonštrukciu spŕšky pracujú v podstate priamo so surovými sekvenciami snímok, čiže paketmi, u nás boli pakety prevedené do formátu jednoduchých snímok reprezentujúcich projekcie maximálnych hodnôt paketu pri pohľade z rôznych osí. Tento koncept je ilustrovaný na obrázku 4.1.



Obr. 4.1: Štruktúra paketu ako aj projekcie z pohľadu určujúcich osí. Snímky sú so zvýraznením kontrastu a vo falošnej farbe.

Hoci takáto reprezentácia dát nemusí byť vhodná pre rekonštrukciu parametrov spŕšky, pre klasifikáciu bola zhodnotená ako postačujúca, keďže rozpoznanie prítomnosti spŕšky je primárne otázkou hľadania pretrvávajúceho lokálne zvýšeného signálu pohybujúceho sa v čase a majúceho isté charakteristiky. Tento formát navyše dovoľuje používať tradičné konvolučné siete pre spracovanie obrazových dát namiesto 3D konvolučných sietí, ktoré sú výpočtovo náročnejšie na tréning.

Inou charakteristikou je navyše aj skutočnosť, že sa nepoužíva celý paket ale pevne daný počet snímok z neho, hoci nemusia byť z rovnakého rozsahu GTU v rámci paketu. Dôvodom je, že v reálnych podmienkach sa môžu použiť aj informácie zo spúšťačov na presnejšie zameranie rozsahu snímok, kde sa najpravdepodobnejšie nachádza príslušná sprška. Okrem toho súvisí výber iba časti snímok paketu aj s povahou niektorých vzoriek šumu použitých pri sade dát vytvorenej rámcom ESAF, podrobnejšie opísanej v časti 5.2.1. Pri dátach pre tréning aj overenie sme použili výrez 20 snímok z paketu. V nasledujúcom texte je použitý takýto názov jednotlivých typov projekcií:

- **projekcia XY:** Projekcia maximálnych hodnôt snímok paketu v čase. Rozmer je rovnaký ako rozmer snímky.
- **projekcia GTUX:** Projekcia maximálnych hodnôt pixelov pozdĺž osi Y. Rozmer je určený šírkou snímky a počtom použitých snímok paketu.
- **projekcia GTUY:** Projekcia maximálnych hodnôt pixelov pozdĺž osi X. Rozmer je určený výškou snímky a počtom použitých snímok paketu.

## 4.2 Dátové sady

Simulované dáta aj dáta z misie sú ukladané v podobe veľkého množstva súborov v súborovom systéme. Toto nie je veľmi vhodné pre ich rýchle načítanie pre tréning prípadne vyhodnotenie siete. Preto boli prevedené do vlastného formátu implementovaného pre jednoduchosť použitia a manipulácie s dátami v sade. Tento formát sme použili aj pre syntetické dáta a používajú ho všetky nami implementované nástroje pre prácu buď so samotnými sadami alebo s modelmi. Každú dátovú sadu charakterizujú 3 zložky s rovnakým počtom položiek, ktoré v sebe obaľuje:

- **dáta:** Samotné položky vytvárané zo snímok paketu.
- **očakávané výstupy:** Očakávaný klasifikačný výstup neurónovej siete pre dátovú položku na rovnakej pozícii.
- **metadáta:** Dodatočné informácie o dátovej položke na rovnakej pozícii.

Každá dátová položka sady je sama o sebe kolekciou obsahujúcou kombináciu jedného zo 4 platných definovaných typov, ktorými sú paket a projekcie maximálnych hodnôt z neho vytvorených. Nie všetky typy musia byť prítomné v rámci sady no pokiaľ áno, tento typ musí byť prítomný v každej položke sady. Typy dátových položiek sú pevne dané pri vytvorení sady a musí v nej byť aspoň jeden zo štyroch platných typov. Typy dátových položiek sú navyše vytvárané z výrezu s rovnakým počtom snímok. Tvar tohto výrezu v podobe trojice hodnôt prirodzených čísiel: (počet snímok, výška snímky, šírka snímky) je popri typoch položiek jedným z atribútov uložených spolu so sadou.

Metadáta danej položky nemajú pevne stanovené atribúty ani ich typy. Navyše možno dynamicky pridávať nové atribúty pre danú položku aj po vytvorení sady. Nie sú použité pri tréningu siete, ale primárne pri vyhodnotení pre účely analýzy prediktívnej výkonnosti siete v závislosti od hodnôt atribútov v metadátach. Každá dátová sada má navyše tieto vlastnosti:

1. Každá dátová sada sprístupňuje dáta, očakávané výstupy aj metadáta pomocou číselných indexov začínajúcich od hodnoty 0.
2. Každá dátová sada je identifikovaná iba menom, ktoré tvorí prefix názvov súborov, ktoré predstavujú túto sadu po uložení v súborovom systéme.
3. Všetky súbory sady sú uložené v rovnakom adresári.
4. Používateľské dáta súborov sady nezahŕňajú cestu v súborovom systéme, kde sa súbory nachádzajú. Premiestnenie sady do iného adresára prípadne kopírovanie sady je možné vykonať relevantnými príkazmi podporovanými operačným systémom, aplikovanými na všetky súbory sady.
5. Používateľské dáta súborov sady taktiež nezahŕňajú názov sady. Preto jej premenovanie je iba záležitosťou premenovania príslušných súborov, presnejšie prefixov ich názvov.
6. Sada má definované typy dátových položiek, ktoré sú vždy prítomné v uloženej podobe. Typy sady sú tvorené ľubovoľnou kombináciou jedného zo štyroch typov, z ktorých minimálne jeden musí byť prítomný: projekcie XY, GTUX, GTUY a paket, z ktorého sú vytvárané.

7. Sada môže byť načítaná do pamäte len s podmnožinou zadaných typov, čo znamená nižšiu pamäťovú záťaž, zvlášť pri veľkých sadách.

## 4.3 Práca s dátovými sadami

Pre prácu s dátovými sadami bola vytvorená sada nástrojov v podobe skriptov. Hlavný účel, okrem vytvárania sád, je aj ich kombinovanie a ďalšie úpravy sád, kvôli vytvoreniu finálnych sád použitých pri overení. Tieto sú:

- Skript pre prevod simulovaných dát aj dát z misie do dátovej sady.
- Skript pre vytvorenie sady syntetických dát.
- Skript pre vizualizáciu položiek sady.
- Skript pre zlúčenie resp. konkatenáciu sád do jednej.
- Skript pre náhodné premiešanie sady.
- Skript pre odčlenenie súvislého rozsahu položiek sady do novej sady.
- Skript pre zobrazenie distribúcie hodnôt metadát v sade.

Skripty pre vytváranie dátových sád ich prevodom alebo generovaním sú popísané v častiach 4.4 resp. 4.5. Motiváciou pre vytvorenie skriptu pre premiešanie sady bolo, že usporiadanie položiek rovnakej triedy v sadách vytváraných spracovaním simulovaných alebo generovaním syntetických dát, nebolo náhodné kvôli spôsobu ich vytvárania. Vizualizácia položiek bola použitá aj pri vývoji spôsobu generovania dát. Zobrazenie distribúcie hodnôt metadát bolo použité pre overenie, že sady majú približne rovnaké zastúpenie prvkov so sledovanými atribútmi pri vyhodnotení senzitivity klasifikácie správok, bližšie popísanej v časti 5.1.

### 4.3.1 Zlučovanie sád

Výsledkom zlučovania je dátová sada obsahujúca počet položiek rovný súčtu počtov položiek jednotlivých sád. Skript zachováva poradie položiek z jednotlivých sád, pričom do výslednej sady sú pridávané položky v tom poradí, v akom boli tieto sady posunuté skriptu na príkazovom riadku. Toto správanie je modelované na základe príkazu `cat`, známeho z viacerých operačných systémov typu



UNIX. Zároveň skript vytvorí nový atribút v metadátach vytváranej sady s názvom `orig_dataset`, ktorého hodnotou je názov pôvodnej dátovej sady, ktorej bola položka súčasťou. Tento atribút však pridá iba v prípade ak v pôvodnej sade nebol tento atribút metadát prítomný. Platí, že iba sady, ktoré majú rovnaké typy prvkov vytvorené z rovnako veľkých rozsahov snímok v pakete môžu byť zlúčené.

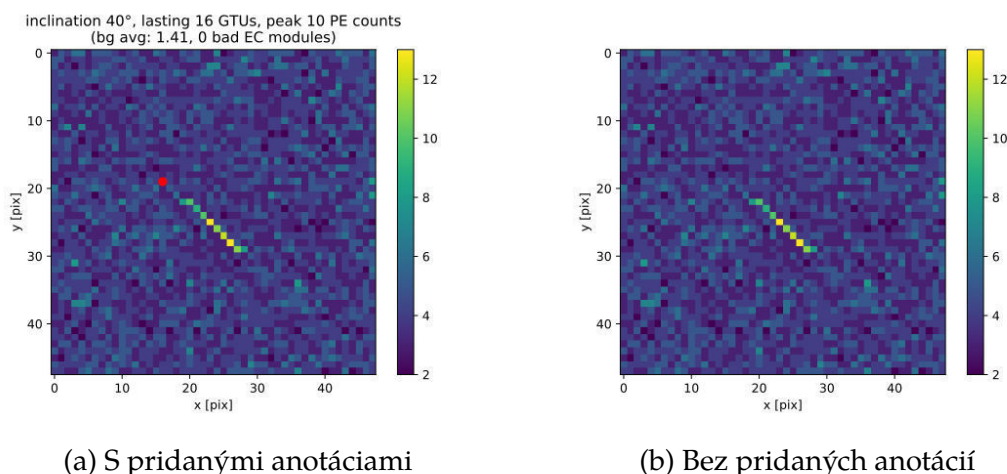
Motiváciou pre vytvorenie skriptu bola potreba postupnej kompozície dátových sád tvorených simulovanými dátami, keďže príslušný skript iba priraduje pevne stanovený očakávaný výstup pre každú položku vytvorené z posunutého zoznamu paketov. Podrobnejšie je jeho fungovanie opísané v časti 4.4.1 Tento skript taktiež slúži pre augmentáciu sady syntetických dát simulovanými pre vytvorenie sád zmiešaných dát pre tréning modelov 5.2.3.

### 4.3.2 Vizualizácia položiek sady

Vizualizácia položiek prebieha prevedením dát z projekcií na jednoduché snímky vo formáte SVG. Aktuálne ale nepodporuje vizualizáciu samotných paketov, iba projekcií z nich vytvorených v rámci sady. Dodatočne umožňuje vizualizovať len časť položiek sady v súvislom rozsahu od počiatočného indexu po prvý index za týmto rozsahom. Všetky vizualizované položky sú uložené do používateľom zadaného výstupného adresára, pričom je v ňom vytvorený podadresár `img/`, ktorý ich obsahuje. Položky sú rozdelené v adresári `img/` podľa typov do vlastných podadresárov nasledujúcim spôsobom:

- `img/yx` Pre vizualizované projekcie XY.
- `img/gtux` Pre vizualizované projekcie GTUX.
- `img/gtuy` Pre vizualizované projekcie GTUY.
- `img/raw` Pre vizualizované pakety.

Pokiaľ podadresár `img/` vo výstupnom adresári už existuje, skript skončí vypísaním chybovej hlášky a nezmení existujúci podadresár. Toto správanie je možné zmeniť špecifikovaním prepínača na príkazovom riadku. V takom prípade budú prepísané aj prípadné podadresáre s položky aj ich obsah, pokiaľ majú rovnaké mená. Pri názvoch súborov vizualizovaných položiek je použitý formát `frame-<index>`, kde `<index>` reprezentuje číselný index položky v sade.



Obr. 4.2: Vizualizácia spršky zo syntetických dát.

Skriptu je dodatočne možné špecifikovať aj definovaný typ dát v sade (syntetické, simulované, dáta z misie). Pre každý z týchto 3 typov sú definované povinné atribúty metadát položiek aj s ich typom. Informácie z metadát budú doplnené vo vytvorených obrázkoch ako anotácie v podobe nadpisov, a v prípade syntetických dát je vyznačená počiatočná pozícia spršky. Typ sa špecifikuje pomocou jedného z trojice prepínačov, pričom nie je možné použiť viac než jeden z prepínačov. Doplnenie metadát do nadpisu je však nepovinné a pokiaľ nie je žiaden z prepínačov použitý, položky sú vizualizované bez nadpisu. Príklad vizualizovaných projekcie XY syntetických dát s pridanými anotáciami a bez nich.

Pri použití dát z misie alebo simulovaných dát, pokiaľ je medzi metadátami atribút `event_id`, reprezentujúci ich identifikátor v databáze (podrobnosti v časti 4.4), je tento použitý ako nadpis, inak je použitá cesta k súboru a číselný index paketu v ňom, z ktorého bola vytvorená položka sady. V druhom prípade však takýto nadpis môže byť ťažko čitateľný, lebo absolútna cesta k súborom s paketmi je veľmi dlhá, rovnako aj názov súborov.

## 4.4 Spracovanie simulovaných dát a dát z misie

Dáta z misie ako aj simulované dáta vytvárané metódou využívajúcou rámec ESAF sú uložené v podobe množstva súborov obsahujúcich pakety. V oboch prípadoch sú snímky všetkých paketov uložené za sebou a pakety nie sú v súboroch viditeľne oddelené. Na priradenie snímku paketu je potrebné poznať počet snímok v jed-

nom pakete (128 snímok) a vydeliť číselný index snímku v súbore týmto číslom, pričom výsledkom je číselný index paketu. Každý paket navyše má aj jedinečný identifikátor známy ako `event_id` a na základe neho sú o ňom vedené záznamy v centrálnej databáze.

Simulované dáta sú ukladané v podobe súborov formátu knižnice NumPy (Oliphant, 2006) obsahujúcich dva pakety, kde prvý je pozadie na ktoré sa zakresľuje stopa spršky a druhý je rovnaký paket ale už so zakreslenou stopou. Rovnako aj dáta z misie sú uložené v podobe súborov obsahujúcich variabilný počet paketov. Tieto sú vo formáte definovanom knižnicou ROOT (Brun et al., 1997) od agentúry CERN. Okrem samotného záznamu signálu v čase obsahujú aj dodatočné informácie ako napr. GPS súradnice v čase zachytenia, konfigurácia detektora, informácie spúšťača prvej úrovne a pod.

- V prvom prípade stačí na extrakciu knižnica NumPy, ktorú už používame v našich nástrojoch.
- V druhom prípade bol použitý dodaný iterátor nad špecifickou štruktúrou súboru (Vrábel, 2019)

Keďže v rámci skriptu, ktorý tieto dáta spracúva sme nechceli zakomponovať aj samotné podmienky výberu paketov, rozhodli sme sa, že skript bude pracovať s vopred vybraným zoznamom paketov v podobe TSV súboru. Každý záznam v tomto zozname musí obsahovať aspoň stĺpec obsahujúci absolútnu cestu k súboru s paketmi. Každý záznam môže taktiež obsahovať dodatočné stĺpce, ktoré sú dokonca potrebné pre jeden z režimov spracovania. V zozname môžu byť prítomné súbory vo formáte NPY aj ROOT, pričom sa spôsob extrakcie paketov z oboch vyberá na základe prípony názvu súboru. Poradie položiek vo vytvorenej sade odpovedá poradiu paketov v spracovanom zozname.

#### 4.4.1 Charakteristiky vytvorenej sady

Povinným argumentom skriptu je aj očakávaný výstup modelu, ktorý sa nastaví rovnaký pre každú vytvorenú položku v dátovej sade. Preto sa sady dát určené pre tréning, kde sa vyžaduje vyrovnaný počet položiek so sprškou a s čistým šumom, vytvárajú z týchto typov dát spravidla tak, že sa spracuje niekoľko zoznamov a pre každý sa vytvorí sada dát z paketov v ňom obsiahnutých. Následne sa

pomocou skriptu pre zlúčenie dátových sád tieto čiastkové sady spoja do jednej a poprípade sa ešte pomocou druhého skriptu položky vo vytvorenej sade premiešajú.

Metadáta vytváranej sady dát vždy obsahujú tieto atribúty:

- `source_file_acquisition_full` Absolútna cestu k použitému súboru s paketmi.
- `packet_id` Číselný index paketu v súbore.
- `start_gtu` Index prvého použitého snímku z paketu.
- `end_gtu` Index snímku za posledným použitým snímkom z paketu.

Skript dodatočne akceptuje užívateľom zadaný zoznam mien atribútov prítomných vo vstupnom zozname, ktoré majú byť pridané k metadátam príslušnej položky. Každá položka v sebe pomocou povinných metadát stále obsahuje referenciu na originálny paket z ktorého bola vytvorená minimálne pomocou kombinácie cesty k súboru a číselného indexu paketu. Taktiež je možné používateľom zahrnúť explicitne v metadátach identifikátor paketu v databáze (atribút `event_id`), pokiaľ je prítomný vo vstupnom zozname, čo môže urýchliť spätné dohľadanie informácií o zdrojovom pakete v prípade potreby.

#### 4.4.2 Režimy spracovania zoznamu paketov

Skript je schopný pracovať v jednom z troch režimov, ktoré udávajú ako sa každý záznam spracuje.

- **default** Predvolený režim. Index paketu v súbore a rozsah snímok, ktoré sa z neho použijú je zadaný používateľom.
- **gtupack** Z angl. „GTU in packet“. Index paketu a rozsah snímok je daný dodatočnými stĺpcami v spracovanom TSV súbore.
- **allpack** Z angl. „all packets“. Zoznam paketov sa interpretuje iba ako zoznam súborov s paketmi, pričom sa extrahujú všetky pakety v rámci súboru. Rozsah snímok je zadaný používateľom.

V režime `gtupack` je potrebné, aby v rámci TSV súboru, ktorý skript spracúva, boli prítomné aj stĺpce `packet_id` a `gtu_in_packet`. Prvý atribút je index paketu v súbore a druhý predstavuje index snímku v rámci tohto paketu, kde bol prvýkrát aktivovaný spúšťač FLT. Používateľ zadáva v tomto režime iba počet snímok bezprostredne pred a za týmto snímkom, ktoré chce zahrnúť pri spracovaní, predvolené nastavenie je 4 snímky pred a 15 snímok za týmto snímkom. Tento režim taktiež zahŕňa ošetrovanie prípadov, kedy vypočítaný rozsah snímok presahuje hranice daného paketu. V takom prípade sa jednoducho vybrané okno snímok posúva dopredu prípadne dozadu, kým nie je opäť v rozsahu snímok paketu. Táto vlastnosť sa však dá vypnúť, ak nie je žiadaná; kód v takom prípade iba vyhodí výnimku a ukončí spracovanie.

V predvolenom režime ani v režime `allpack` nie je potrebné aby spracovaný zoznam mal hore uvedené stĺpce. Keďže sa však na každý záznam aplikuje rovnaká logika spracovania, či už ide o výber jedného paketu alebo všetkých paketov zo súboru, je potrebné na rozdiel od režimu `gtupack` aby zoznam nemal viacej záznamov s rovnakým zdrojovým súborom paketov. V opačnom prípade bude viacero položiek v sade vytvorených z rovnakého paketu, čo nie je považované za žiaduce. V oboch režimoch sa rozsah snímok stanovuje pomocou číselných indexov prvého a posledného snímku v rámci paketu, ktoré majú byť vybrané.

Predvolený režim aj režim `allpack` boli implementované ako prvé a používané prevažne pre vytvorenie sád dát iba zo simulovaných dát vytvorených rámcom ESAF resp. iba dát z misie bez potreby selektovať záznamy. V súčasnosti sa preferuje skôr režim `gtupack`, ktorý výraznejšie využíva informácie o pakete uložené v databáze.

## 4.5 Syntetické dáta

Okrem dát generovaných pomocou rámca ESAF sme vytvorili aj vlastný o niečo menej výpočtovo náročný algoritmus generovania dát. Tento vytvára v cykle pakety snímok, kde pixel snímky má rozsah hodnôt rovnaký ako výstup PDM modulu, čiže 8-bitové celé číslo bez znamienka. Výška, šírka aj počet snímok paketu sú nastaviteľné používateľom, čomu následne odpovedá aj rôzna veľkosť projekcií, pokiaľ sú vytvárané. Po vytvorení paketu je doň zakresľovaná stopa simulovanej spŕšky generovaná vlastným algoritmom a voliteľne sa potom simuluje aj efekt

chybových EC modulov. Vytvorený paket je namieste konvertovaný na položku sady dát a je k nemu priradený príslušný očakávaný klasifikačný výstup ako aj metadáta. Vďaka konverzii namieste sa pri vytváraní sady dát obsahujúcej iba projekcie takto zamedzuje zbytočnej pamäťovej záťaži množstvom paketov, ktoré na konci aj tak nie sú ukladané.

#### 4.5.1 Parametre spříšky a pozadia

Algoritmus prijíma ako argumenty niekoľko parametrov, každý v podobe dvojice hodnôt, ktoré určujú vlastnosti pozadia ako aj generovanej stopy spříšky. Tieto dvojice predstavujú krajné body uzavretého intervalu určujúceho rozsah použitých hodnôt parametra pre celú generovanú sadu dát. Pre každú konkrétnu položku sady sú skutočné použité hodnoty náhodne generované z tohto intervalu počas vytvárania paketu. Pokiaľ obe hodnoty pre daný parameter sú si rovné, čiže krajné body intervalu sa rovnajú, bude pre každú položku použitá konštantná hodnota tohto parametra. Skutočná hodnota použitá pre danú položku je ukladaná ako metadáta k vytvorenej položke. Plný zoznam parametrov podliehajúci tomuto pravidlu a ich význam je nasledujúci, pričom treba povedať, že okrem parametra `bg_lambda` sú všetky celočíselného typu:

- `bg_lambda`: Stredná hodnota pixelov pozadia paketu. Odpovedá parametru  $\lambda$  v Poissonovej distribúcii.
- `num_bad_EC`s: Maximálny počet EC modulov v rámci položky, ktoré majú byť simulované ako chybové.
- `start_gtu`: Prvé GTU resp. snímok paketu na ktorom sa má zakresľovať stopa spříšky.
- `start_y`: Prvý pixel pozdĺž výšky paketu, na ktorý sa má zakresľovať stopa spříšky.
- `start_x`: Prvý pixel pozdĺž šírky paketu, na ktorý sa má zakresľovať stopa spříšky.
- `shower_max`: Maximálna intenzita spříšky meraná relatívne voči nulovému pozadiu.

- *duration*: Trvanie spršky vyjadrené počtom GTU resp. snímok paketu na ktorom je sprška.
- *track\_length*: Dĺžka stopy spršky pri pohľade z xy projekcie paketu.

#### 4.5.2 Algoritmus generovania stopy spršky

Po vytvorení paketu s pozadím je vygenerovaná počiatočná pozícia stopy simulovanej spršky ako bod v 3D priestore reprezentovaný trojicou hodnôt *start\_gtu*, *start\_y* a *start\_x*. Pomocou hodnôt *duration* a *track\_length* je odvodený aj koncový bod stopy spršky pomocou trigonometrie v 3D priestore podľa vzorca 4.1, kde  $\alpha$  predstavuje uhol stopy spršky v xy projekcii.

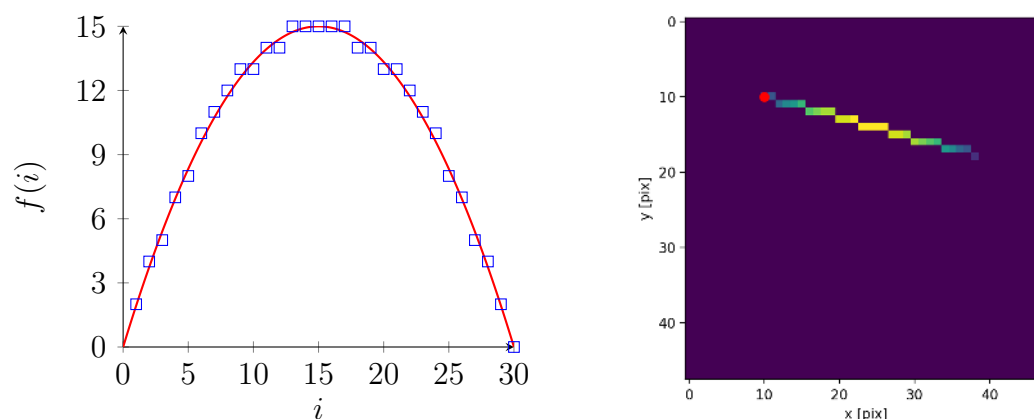
$$\begin{aligned}gtu &= start\_gtu + duration, \\y &= start\_y + round(\cos(\alpha) * track\_length), \\x &= start\_x + round(\sin(\alpha) * track\_length)\end{aligned}\tag{4.1}$$

V prípade, že vytvorená čiara sa nezmestí do rozsahov súradníc určených rozmermi paketu, sú orezané tie súradnice, ktoré sú mimo týchto rozsahov. Pre odvodenie súradníc medzi začiatkom a koncom stopy je použitý Bressenhamov algoritmus kreslenia čiary v 3D priestore. Následne sú vygenerované aj hodnoty pozdĺž tejto čiary odpovedajúce počtu fotoelektrónov reprezentujúcich čistý signál spršky. Pre vytvorenie hodnôt je použitý vzorec 4.2.

$$v_i = round(shower\_max \cdot (-\left(\frac{2 \cdot i}{len} - 1\right)^2) + 1), \forall i > 0 \wedge i < len\tag{4.2}$$

Vo vzorci  $v_i$  predstavuje hodnotu signálu stopy pre  $i$ -tú súradnicu od počiatočnej súradnice stopy spršky, *len* je počet súradníc orezanej stopy, čiže dĺžka čiary v pixeloch. Príklad priebehu hodnôt generovaných týmto vzorcom je vidieť na grafe na obrázku 4.3.

Hodnoty stopy sú zapísané do paketu s nulovými pixelmi a je na ne aplikovaný Gaussov filter v 3D priestore, čím sa dosahuje efekt vyhladenej mierne realistickejšej stopy. Filter má pevne nastavenú hodnotu štandardnej odchýlky  $\sigma$  rovnú 0,7. Pixely, kde boli zapísané pôvodné hodnoty sa označujú ako centrálna os spršky a nenulové hodnoty okolo nej vytvorené aplikáciou filtra ako okolie. Hodnoty okolia sú však väčšinou menšie ako 1 a pri pretypovaní na celé čísla by sa jednoducho stratili, preto sú vynásobené konštantou 1000, čím sa dosiahne rozsah hodnôt



Obr. 4.3: Priebeh funkcie pre generovanie hodnôt signálu simulovanej spŕšky pre hodnotu `shower_max` rovnú 15 a stopu s dĺžkou 30 pixelov. Modré štvorce predstavujú zaokrúhlenú celočíselnú hodnotu funkcie pre danú súradnicu čiary. Vpravo je vidno stopu tejto spŕšky v xy projekcii paketu s nulovým pozadím a vyznačenou počiatočnou súradnicou

medzi 0 a okolitými hodnotami pri centrálnej osi stopy. Centrálna os ostáva nezmenená. Konštanty použité pri výpočte boli odvodené experimentálne berúc do úvahy maximálny rozsah hodnôt v pixeli.

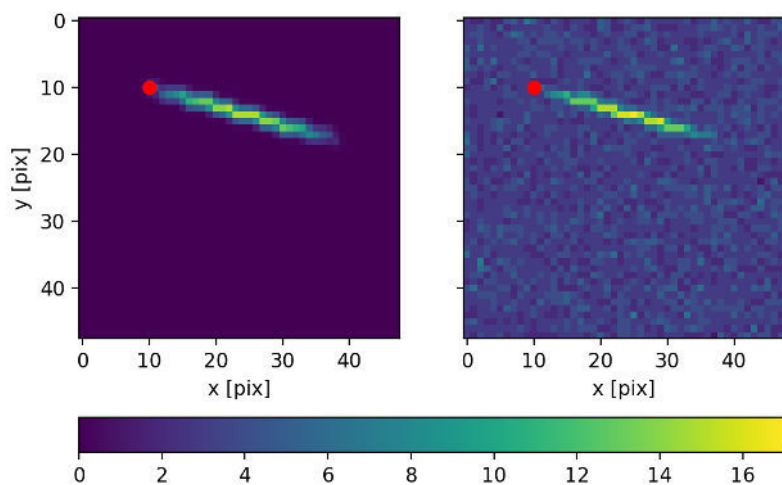
Stopa spŕšky v obrázku 4.3 po aplikovaní vyhladzovania je zobrazená na obrázku 4.4 na pozadí bez šumu aj so šumom s hodnotou `bg_lambda` rovnou 1.

### 4.5.3 Simulácia chybových EC modulov

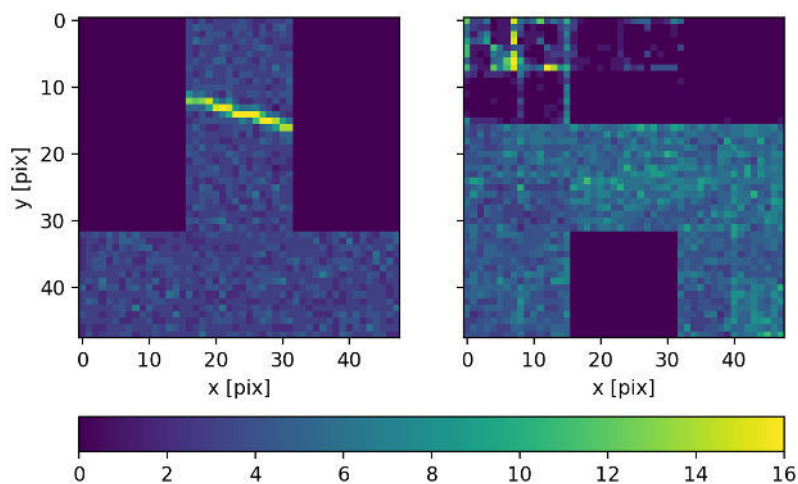
Voliteľne je možné pri generovaní položiek nastaviť aj efekt simulujúci vyradenie niektorých EC modulov, ktorý je možné vidieť aj pri dátach z letu. Pre realizáciu efektu sa vyberú náhodné EC moduly podľa ich číselného označenia, vynechajúc iba ten EC modul, v ktorom súčet hodnôt zo signálu spŕšky je maximálny. Následne sa pre vybrané EC moduly nastavia hodnoty pixelov v nich obsiahnutých v každom snímku paketu.

Aktuálne je simulácia nefunkčných EC modulov limitovaná len na vynulovanie príslušných pixelov. Toto je najmenej realistická časť v tejto metóde vytvárania simulovaných dát, keďže v dátach z letu boli viditeľné aj EC moduly, u ktorých sa chyby prejavovali aj výrazným jasom v príslušnej časti detektora, ako možno vidieť na projekciách na obrázku 4.5.





Obr. 4.4: Stopa generovanej spŕšky z obrázka 4.3 po vyhladení pri pohľade z xy projekcie na pozadí bez šumu a so šumom.



Obr. 4.5: Stopa generovanej spŕšky z obrázka 4.4 v xy projekcii po simulovanom vyradení 4 EC modulov a xy projekcia paketu z misie EUSO-SPB (25.4.2017) s rovnakým počtom nefunkčných EC modulov.

## 4.6 Architektúry sietí

Pri sieťach bolo vyskúšaných niekoľko rôznych architektúr, ktoré sa dajú klasifikovať do 2 všeobecnejších kategórií:

- konvolučné siete, spracúvajúce iba projekciu XY.
- konvolučné siete, spracúvajúce všetky 3 typy projekcií

Konvolučné siete iba s jedným vstupom prevažne slúžili na overenie konceptu, či je neurónová sieť vôbec schopná rozpoznať hľadaný vzor v dátach. Pri vývoji sme začali jednoduchými architektúrami obsahujúcimi iba 2 konvolučné vrstvy, každá nasledovaná max-pooling vrstvou. Po konvulčnej časti siete nasledovali 2 plne prepojené vrstvy a klasifikátor softmax. Motiváciou pre vytvorenie siete s 3 vstupmi bola prítomnosť paketov v reálnych dátach, kde je viditeľné dočasné zosilnenie signálu v jednom snímku resp. GTU, ktorý sa pri pohľade v xy projekcii javí ako stopa spŕšky. Keďže však takýto jav nie je viditeľný vo viac ako 1 snímku alebo 2 snímkoch, myšlienkou je, že sieť, ktorá má aj informácie o časovom priebehu napr. aj v podobe časových projekcií by mala byť lepšie schopná filtrovať takéto falošné pozitíva a teda vykazovať lepšiu celkovú úspešnosť klasifikácie.

Architektúry boli inšpirované prevažne 2 existujúcimi architektúrami. Prvou je architektúra, ktorú použil Lincoln, 2017, podrobnejšie opísaná v časti 3.2. Druhou je príklad konvulčnej siete od Ballard et al., 2016, používaný pre klasifikačnú úlohu MNIST. Ide o úlohu, pri ktorej sa rozpoznávajú ručne nakreslené číslice od 0 do 9 v monochromatických snímkoch s rozlíšením  $28 \times 28$  pixelov. Autori túto sieť implementovali ako príklad použitia v príručke k softvérovému rámcu TFlearn, ktorý sme použili. Všetky architektúry sietí odvodené od tej, ktorú vyvinul Lincoln, 2017, sú ďalej označované len ako siete `GitHub_net`. Architektúry odvodené od architektúry, ktorú vyvinuli Ballard et al., 2016, sú ďalej označované len ako siete MNIST

Po prvotných experimentoch podrobnejšie popísaných časti 4.6.1, boli siete MNIST vylúčené z ďalších pokusov, lebo vykazovali veľmi slabé výsledky, akonáhle sa v paketoch a teda aj v projekciách začal vyskytovať šum na pozadí. Toto dodatočne ovplyvnilo aj návrh siete s 3 vstupmi.

### 4.6.1 Výber architektúr

Prvotné návrhy sietí spracúvajúce iba projekcie XY využívali výstupnú vrstvu indikujúcu prítomnosť spŕšky pomocou jedinej výstupnej hodnoty reprezentujúcej pravdepodobnosť prítomnosti časticovej spŕšky v dátach. To znamená hodnota 1 na výstupe znamenala jej prítomnosť a hodnota 0 čistý šum. Pri takejto reprezentácii výstupu však vykazovali siete veľmi slabé výsledky s približne 50% úspešnosťou predpovedí aj pri dátach s nulovým pozadím a jasne viditeľnými čiarami bez ohľadu na počet epoch tréningu. Zlepšenie presnosti sa nám napokon podarilo dosiahnuť zmenou tejto schémy na vektor dvoch hodnôt reprezentujúcich prítomnosť spŕšky resp. šumu v dátach. Inými slovami, šum sa považoval odteraz za vlastnú triedu v klasifikácii. Po úprave všetky testované siete vykazovali takmer 100% úspešnosť pri klasifikácii dát s nulovými hodnotami pozadia. Následne sme vykonali aj pokus s rozpoznávaním čiar rôznej intenzity pri zavedení šumu pozadia do dát.

Cieľom pokusu bola selekcia vhodných architektúr resp. hyperparametrov siete z existujúcich návrhov. Pri experimentoch sa primárne pozeralo na to, ktorá architektúra vykazuje najlepšiu úspešnosť pri tréningu ako aj priebežnom testovaní a zopakovanie pokusu so znižovanou intenzitou stopy spŕšky v xy projekcii. Tieto pokusy mali v sebe niekoľko obmedzení, spomedzi nich aj skutočnosť, že náš generátor dát v danom čase bol schopný vytvoriť iba xy projekcie priamo bez vytvárania paketu ako prvého kroku a bol schopný zakresľovať iba vyhladené čiary konštantnej intenzity. Napriek tomu nám dali spätnú väzbu o tom, ktoré architektúry neboli vhodné na klasifikáciu snímok z letu. Pokus mal tieto parametre:

- Boli použité dve sady dát, obe s 20000 položkami, pričom prvá mala nulové pozadie a druhá pozadie určené Poissonovou distribúciou s hodnotou  $bg\_lambda$  rovnou 10.
- Dáta boli dynamicky rozdelené pomocou použitého softvérového rámca na tréningové a testovacie v pomere 9:1 (čiže 18000:2000 snímkov).
- Doba tréningu bola stanovená na 11 epoch.

Pri pokuse boli použité architektúry odvodené od siete Ballarda a Damiana Ballard et al., 2016, ďalej označované len ako siete MNIST ako aj od pôvodného Lincol-

novho príkladu Lincoln, 2017, ďalej označované len ako siete `GitHub_net`, všetky uvedené v tabuľke 4.1.

MNIST1	MNIST2	MNIST3	GitHub1	GitHub2	GitHub3
C(32,3x3,1x1)	C(64,3x3,1x1)	C(32,3x3,1x1)	C(64,3x3,3x3)	C(64,3x3,1x1)	C(64,3x3,1x1)
MP(2x2,2x2)	MP(2x2,2x2)	MP(2x2,2x2)	MP(2x2,2x2)	MP(2x2,2x2)	MP(2x2,2x2)
LRN	LRN	n/a	n/a	n/a	LRN
C(64,3x3,1x1)	C(64,3x3,1x1)	C(64,3x3,1x1)	C(64,3x3,3x3)	C(64,3x3,1x1)	C(64,3x3,1x1)
MP(2x2,2x2)	MP(2x2,2x2)	MP(2x2,2x2)	MP(2x2,2x2)	MP(2x2,2x2)	MP(2x2,2x2)
LRN	LRN	n/a	dropout(0,3)	dropout(0,3)	LRN
n/a	n/a	n/a	flatten	flatten	n/a
FC(128, tanh)	FC(256, relu)	FC(128, tanh)	FC(128, relu)	FC(128, relu)	FC(128, relu)
dropout(0,8)	dropout(0,8)	n/a	dropout(0,5)	dropout(0,5)	dropout(0,5)
FC(256, tanh)	FC(256, relu)	FC(256, tanh)	FC(50, relu)	FC(50, relu)	FC(50, relu)
dropout(0,8)	dropout(0,8)	n/a	n/a	dropout(0,5)	dropout(0,5)
FC(2, softmax)	FC(2, softmax)	FC(2, softmax)	FC(2, softmax)	FC(2, softmax)	FC(2, softmax)

Tabuľka 4.1: Architektúry sietí použitých pri pokuse. Zvýraznené sú odchýlky oproti základnému modelu s poradovým číslom 1. Vysvetlivky použitých skratiek sú v tabuľke 4.2. Hodnota "n/a" aindikuje absenciu vrstvy oproti základnému modelu

C(N, WxH, SWxSH)	Konvolučná vrstva s počtom filtrov N, každý so šírkou W a výškou H a posunom SW v horizontálnom a SH vo vertikálnom smere.
FC(N, A)	Plne prepojená vrstva s počtom neurónov N a aktivačnou funkciou A.
MP(WxH, SWxSH)	Vzorkovacia vrstva maximálnych hodnôt aplikovaná na regióny vstupu so šírkou W a výškou H a s posunom okna SW v horizontálnom a SH vo vertikálnom smere.
LRN	Vrstva pre normalizáciu lokálnej odozvy.
dropout(P)	Vrstva na prepustenie vstupu predošlej vrstvy s pravdepodobnosťou P, použitá iba pri tréningu.

Tabuľka 4.2: Vysvetlenie skratiek použitých pri popise vrstiev siete.

Sumárne výsledky experimentu sú uvedené v tabuľke 4.3. Ako vidno siete MNIST vykazovali pri dátach bez šumu, čiže s nulovým pozadím úspešnosť skoro

100% no pri dátach so šumom s hodnotou  $bg\_lambda$  rovnou 1 klesla na úroveň 50%. Na druhej strane, siete `GitHub_net` vykazovali veľmi vysokú úspešnosť vyše 90% pri oboch množinách dát.

Názov siete	Dáta bez šumu		Dáta so šumom	
	Úspešnosť pri tréningu	Úspešnosť pri testovaní	Úspešnosť pri tréningu	Úspešnosť pri testovaní
MNIST1	98,72%	100,00%	52,51%	52,05%
MNIST2	99,50%	100,00%	50,68%	49,30%
MNIST3	100,00%	100,00%	48,10%	50,85%
GitHub1	98,37%	97,90%	93,14%	96,60%
GitHub2	100,00%	100,00%	99,43%	99,95%
GitHub3	100,00%	100,00%	99,55%	99,95%

Tabuľka 4.3: Úspešnosť jednotlivých sietí po 11 epochách tréningu

Vyskúšali sme vytvoriť ešte jeden variant, `GitHub3b`, ktorého štruktúra konvolučnej časti siete presne kopíruje štruktúru konvolučnej časti siete `MNIST1`, pričom úspešnosť sa ustálila na rovnakej úrovni ako pri sieti `GitHub3`. Jediná iná vlastnosť, v ktorej sa siete odvodene od príkladu `Lincolna` líšia je štruktúra plne prepojených vrstiev, kde druhá vrstva bližšie ku výstupnej má pri `MNIST` sieťach vyšší resp. rovnaký počet neurónov oproti prvej vrstve kým pri sieťach `GitHub_net` je vidieť nižší počet neurónov.

Výsledok pokusov okrem iného ovplyvnil aj návrh architektúry siete s 3 vstupmi, ďalej označovanú ako `triple_net`. Pri nej sme pre každý vstup použili konfiguráciu konvolučných vrstiev rovnakú ako v prípade `GitHub3` resp. `MNIST1`, kým štruktúra plne prepojených vrstiev kopíruje siete `GitHub_net`. Výstupy konvolučnej časti siete pre všetky 3 posunuté projekcie sú následne pred posunutím prvej plne prepojenej vrstve transformované na jednorozmerný vektor hodnôt a spojené do jediného vektora v poradí: výstup z projekcie `XY`, `GTUX` a `GTUY`.

Štruktúra plne prepojených vrstiev nebola menená pri súčasnom zväčšení celkovej veľkosti vstupu z konvolučnej časti siete a tomu odpovedajúcemu nárastu parametrov. Táto sieť však vykazovala rovnakú úspešnosť pri tréningu ako siete `GitHub_net` iba so vstupom v podobe projekcie `XY`. V konečnom dôsledku boli pre testovanie zvolené siete `GitHub3b` a `triple_net`.

## 5 Overenie riešenia

---

Počas tréningu modelov sa ako overenie najprv sledoval vývoj celkovej úspešnosti klasifikácií na validačnej podmnožine trénovacej sady dát. Modely boli trénované osobitne na sadách tvorených simulovanými, syntetickými ako aj zmesi oboch typov dát. Doba tréningu bola stanovená na 30 epoch, čiže 30 iterácií cez celú trénovaciu sadu.

Po dokončení tréningu boli natrénované modely vyhodnotené podľa metrick klasifikácie opísaných v časti 5.1 na jednej alebo viacerých vyhodnocovacích sadách. Taktiež boli vizualizované výstupy skrytých vrstiev modelov v konvolučnej časti, čiže mapy príznakov ako možný indikátor vysvetlenia prípadných rozdielov medzi modelmi pri sledovaných metrikách. Napokon boli vybrané modely vykazujúce aspoň 90% celkovú úspešnosť aplikované na dáta z misie a boli vybrané tie dáta, ktoré každý z modelov označil, že obsahujú stopu EAS a zhodnotené.

### 5.1 Sledované metriky

Pri overení sa sledovali primárne 2 metriky binárnej klasifikácie:

- **Senzitivita**, definovaná ako pomer úspešných klasifikácií položky ako spŕšky ku celkovému počtu položiek obsahujúcich spŕšku.
- **Špecificita**, definovaná ako pomer úspešných klasifikácií položky ako šumu ku celkovému počtu položiek so šumom v sade

Vysoká senzitivita znamená, že je model schopný identifikovať veľkú väčšinu spŕšok v posunutej vzorke, kým vysoká špecificita značí nízku úroveň falošných pozitív. Je preto dobré, aby model vykazoval vysokú úroveň oboch atribútov, no nižšia senzitivita sa toleruje menej ako nižšia špecificita.

Pri vyhodnotení senzitivity sme dodatočne overili túto metriku v závislosti od hodnôt atribútov charakterizujúcich položky v sade dát ako trvanie spŕšky v počte snímok paketu, hodnotu uhlov  $\theta$  a  $\phi$  a energie primárnej častice. Keďže pri šume neexistujú tieto atribúty resp. nemajú žiaden zmysel, meriame počet falošných pozitív a aj špecificitu pre celú podmnožinu položiek sady, kde očakávaný výstup je šum.

Okrem týchto metrík sme nechali vizualizovať aj aktivácie skrytých vrstiev modelu ako aj filtrov konvolučných vrstiev pre účely porovnania naučených filtrov pri modeloch trénovaných na oboch typoch dát.

## 5.2 Dáta pre tréning modelov

Pre tréning modelov boli použité osobitné sady obsahujúce len simulované alebo syntetické dáta. Avšak modely trénované len na syntetických dátach vykazovali neuspokojivé výsledky pri vyhodnotení ich špecificity na validačnej podmnožine simulovaných dát. Kvôli tejto skutočnosti sme preskúmali aj vplyv augmentácie sady syntetických dát postupne zvyšovaným množstvom simulovaných dát na merané metriky modelu, zvlášť špecificitu.

### 5.2.1 Simulované dáta

Pre tréning na simulovaných dátach boli dáta vytvorené rámcom ESAF doplnené o vzorky čistého šumu, ktoré pozostávajú z dát z misie, ktoré boli buď manuálne alebo automatizovane pomocou iných metód vyhodnotené ako šum s veľmi vysokou pravdepodobnosťou (Vrábel, 2019). Pre výslednú sadu dát je však potrebné, aby bol približne rovnaký počet vzoriek so simulovanou spŕškou aj šumom, pričom tieto vzorky čistého šumu dokopy tvorí iba asi 14000 paketov, čo obmedzuje množstvo simulovaných spŕšok, ktoré je možné vytvoriť. Ako riešenie bola použitá technika, kde sa vytvorila rozsiahla vzorka čistého šumu s niečo menej než 100000 paketmi, vytvorená tak, že sa vybralo po 30 súvislých snímok z paketov z misie a následne sa táto sekvencia zopakovala pre vytvorenie paketu so 128 snímkami.

Treba povedať, že toto riešenie nie je úplne ideálne, keďže takýto vzor sa v reálnych dátach z misie nenachádza, no bolo potrebné, aby pozadie v tejto sade dát malo charakteristiky typické pre dáta z letu. Toto obmedzenie sme obišli tak, že

namiesto spracovania celého paketu sa v sade ponecháva iba výrez snímok z neho, presnejšie 20 snímok. Vo výslednej sade dát sme použili 33000 položiek so simulovanou spŕškou a zo spomenutej množiny bolo teda použitých 19000 vzoriek šumu pre doplnenie zvyšných 14000 originálnych paketov z letu. Sada bola niekoľkokrát po vytvorení premiešaná aby sa zaistilo približne rovnomerné rozdelenie položiek so spŕškou a čistým šumom. Pre validačnú podmnožinu pri tréningu bolo použitých prvých 15000 položiek z celkovej sady, čo reprezentuje okolo 22% položiek sady.

### 5.2.2 Syntetické dáta

V prípade syntetických dát (popísané v časti 4.5, kde je pozadie generované pomocou Poissonovej distribúcie neexistuje limitovaný vopred existujúci počet vzoriek so šumom. Tu bolo skôr potrebné vytvoriť dostatočne rozsiahlu množinu dát pre zachytenie dostatočného množstva exemplárov reprezentujúcich čo najväčšie množstvo kombinácií hodnôt maximálnej intenzity spŕšky a pozadia, uhla orientácie stopy spŕšky a pod. V dôsledku toho sme pri tréningu použili sadu obsahujúcu až 400000 položiek, pričom pre validačnú podmnožinu bolo použitých prvých 40000 položiek, čiže 10% zo sady. Podobne ako predošlá sada aj táto sada tvorená syntetickými dátami bola premiešaná pred použitím pre zaistenie vyrovnaného počtu položiek so spŕškou a šumom vo validačnej sade.

Význam parametrov použitých pri generovaní syntetických dát je vysvetlený v časti 4.5.1. Syntetické dáta použité pri tréningu mali charakteristiky uvedené v tabuľke 5.1.

### 5.2.3 Zmiešané dáta

Ako bolo spomenuté v časti 5.2, modely tréňované len na syntetických dátach nevykazovali uspokojivú špecifitu pri vyhodnotení na validačnej podmnožine simulovaných dát. Primárnym účelom zmiešanej sady bolo preskúmať vhodnosť augmentácie sady syntetických dát postupne zvyšovaným množstvom simulovaných dát na merané metriky klasifikácie vykazované modelmi natréňovanými na takto zmiešaných dátach.

Pre vytvorenie tejto sady bolo najprv odčlenených prvých 15000 snímok zo sady simulovaných dát, čo je validačná podmnožina príslušnej sady, aby sa zais-



Parameter	Minimum	Maximum
bg_lambda	0,6	2,6
num_bad_ECcs	0	4
start_gtu	0	7
start_y	2	45
start_x	2	45
shower_max	3	13
duration	7	40
track_length	3	40

Tabuľka 5.1: Parametre syntetických dát použitých pri tréningu aj vyhodnotení

tilo, že nebude existovať priesek medzi simulovanými dátami použitými pre tréning a vyhodnotenie. Následne sa zo zvyšných dát sady vybralo prvých 15000, 30000 a napokon 45000 položiek a primiešali sa do skopírovanej pôvodnej sady syntetických dát. Následne sa takto vytvorené sady niekoľkokrát premiešali, aby sa zaistilo rovnomerné rozdelenie oboch typov dát v sadách. Pre validačnú podmnožinu počas tréningu bolo použitých prvých 10% položiek, čo odpovedá 41500, 43000 a 44500 položkám, resp.

### 5.3 Dáta pre vyhodnotenie modelov

Pri sieťach trénovaných len na jednom type dát prebiehalo overenie senzitivity a špecificity modelov na validačnej podmnožine príslušnej trénovacej sady dát. Táto je počas tréningu používaná na priebežné vyhodnocovanie úspešnosti predpovedí modelu, pričom dáta z tejto podmnožiny nie sú použité na nastavovanie nových hodnôt váh počas tréningu. Dodatočne sme siete natrénované na syntetických dátach overili aj na rovnakej validačnej podmnožine sady simulovaných dát použitej pri tréningu príslušných modelov.

Pre modely trénované na zmiešaných sadách dát bola pre vyhodnotenie použitá validačná podmnožina simulovaných dát. Toto bolo primárne pre možnosť priameho porovnania metrík týchto modelov s modelmi trénovanými len na syntetických alebo simulovaných dátach. V sumáre boli všetky overované modely vyhodnotené na sadách dát s parametrami uvedenými v tabuľke 5.2.

Typ dát trénovacej sady	Typ dát vyhodnocovacej sady	Vzorky so šumom vo vyhodnocovacej sade	Vzorky so simulovanou spŕškou vo vyhodnocovacej sade
simulované	simulované	7467	7533
syntetické	syntetické	20042	7533
syntetické	simulované	7467	7533
zmiešané (15000 simu.)	simulované	7467	7533
zmiešané (30000 simu.)	simulované	7467	7533
zmiešané (45000 simu.)	simulované	7467	7533

Tabuľka 5.2: Typy a parametre dát použitých pri vyhodnotení modelov

## 6 Dosiiahnuté výsledky

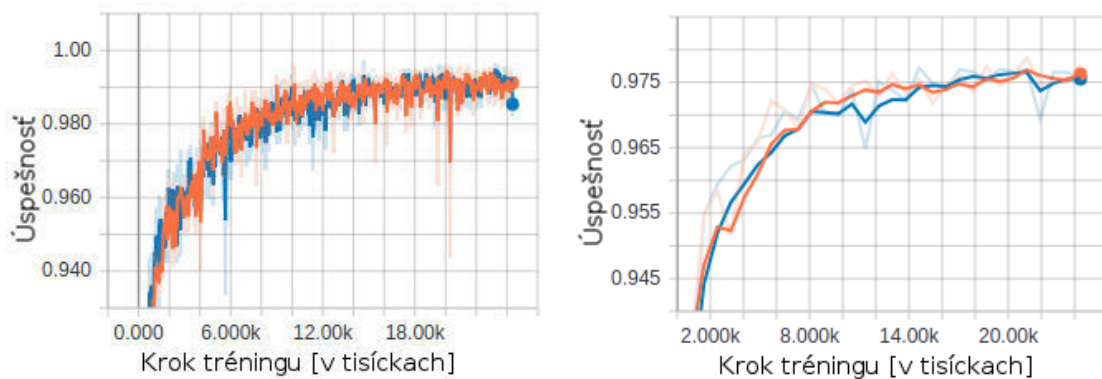
Nižšie uvádzame výsledky testovania spôsobom uvedeným v časti 5. Pri krivkách vývoja úspešnosti počas tréningu je aplikované isté vyhladzovanie, tzn. znižovanie extrémnych výkyvov v úspešnosti, aby bol lepšie viditeľný širší trend vývoja tejto metriky počas tréningu. Skutočné hodnoty sú naznačené slabšími odtieňmi farby v grafe. Súhrnné výsledky sledovaných metrík uvádzame v tabuľke 6.1.

Typ tréningových dát			Typ vyhodnocovacích dát			
Architektúra	SP	FN	SN	FP	Špecificita	Senzitivita
simulované			simulované			
github_net3b	7308	225	7357	110	98,53%	97,01%
triple_net	7199	334	7433	34	99,54%	95,57%
syntetické			syntetické			
github_net3b	14349	5609	19281	761	96,2%	71,9%
triple_net	14612	5346	19729	313	98,44%	73,2%
syntetické			simulované			
github_net3b	7460	73	232	7235	3,11%	99,03%
triple_net	7232	301	2473	4994	33,12%	96%
zmiešané (15000 simu.)			simulované			
github_net3b	6547	986	7340	127	98,3%	86,91%
triple_net	6819	714	6898	569	92,38%	90,52%
zmiešané (30000 simu.)			simulované			
github_net3b	6938	595	7303	164	97,8%	92,10%
triple_net	7021	512	7286	181	97,58%	93,20%
zmiešané (45000 simu.)			simulované			
github_net3b	7059	474	7298	169	97,74%	93,71%
triple_net	7067	466	7390	77	98,97%	93,81%

Tabuľka 6.1: Súhrnné vyhodnotenie metrík modelov. Stĺpce SP (skutočné pozitíva) a FN (falošné negatíva) označujú položky sady obsahujúce simulovanú spršku klasifikované modelom do správnej resp. nesprávnej triedy. Stĺpce SN (skutočné negatíva) a FP (falošné pozitíva) označujú položky sady obsahujúce šum klasifikované modelom do správnej resp. nesprávnej triedy.

## 6.1 Modely trénované na simulovaných dátach

Priebeh tréningu po je znázornený na obrázku 6.1. Možno vidieť, že približne v polovici tréningu sa úspešnosť predpovedí pri oboch modeloch ustálila okolo 97% na validačnej sade a 99% na trénovacej sade. Výsledky vyhodnotení senzitivity na validačnej podmnožine sú uvedené pre tri atribúty: skutočná energia primárnej častice a uhly  $\theta$  a  $\phi$ .



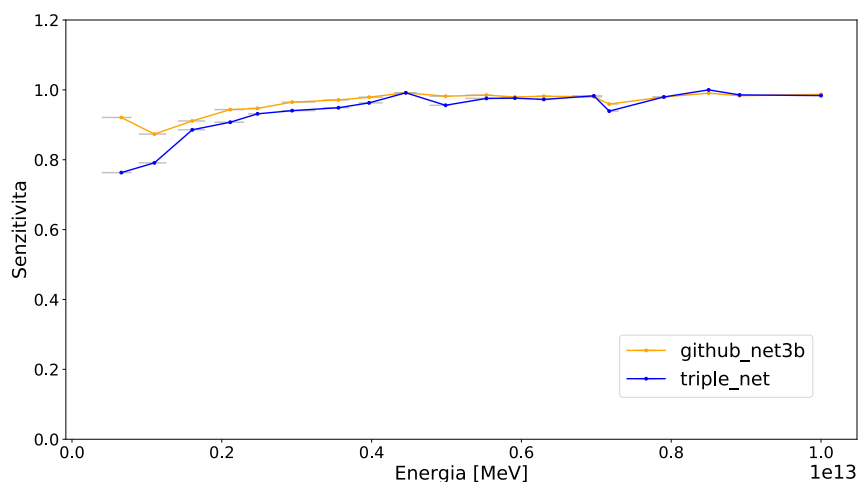
(a) Úspešnosť na trénovacej podmnožine (b) Úspešnosť na validačnej podmnožine

Obr. 6.1: Vývoj celkovej úspešnosti predpovedí na sade simulovaných dát po 30 epochách tréningu pre siete `github_net3b` (oranžová) a `triple_net` (modrá).

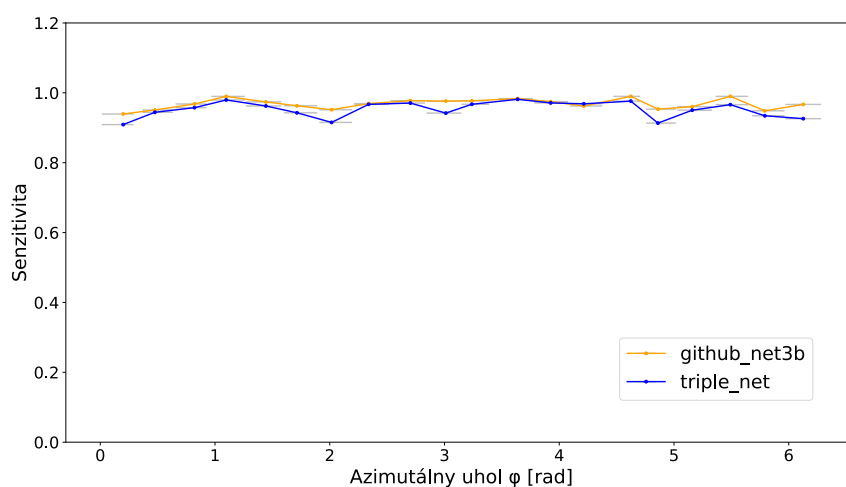
Vyhodnotenie senzitivity podľa energie primárnej častice na grafe na obrázku 6.2 ukázalo istý pokles senzitivity pre nižšie hodnoty energie. Na druhej strane, v rozsahu energií menej než  $0,1 \times 10^{13}$  MeV je oveľa menej vzoriek než v prípade rovnako širokého rozsahu vyšších energií, preto je odhad senzitivity menej istý pri týchto hodnotách. Okrem občasných fluktuácií je vidieť istý slabší pokles prediktívnej schopnosti so znižujúcou sa energiou. V rámci hodnôt atribútu však aktuálne nevidno bod, v ktorom by ktorákoľvek z testovaných sietí prestala mať stále nadpolovičnú prediktívnu schopnosť.

Na grafe vyhodnotenia podľa azimutálneho uhla  $\phi$  na obrázku 6.3 je možné pozorovať fluktuácie, ktoré môžu súvisieť s výkyvmi v počte vzoriek s danou hodnotou atribútu, ale inak si siete zachovávajú relatívne konštantnú prediktívnu schopnosť.

Podobne ako v prípade uhla  $\phi$  je možné podobný vzor pozorovať aj v prípade možných hodnôt zenitového uhla  $\theta$  na obrázku 6.4. Najmenej vzoriek v rámci sady



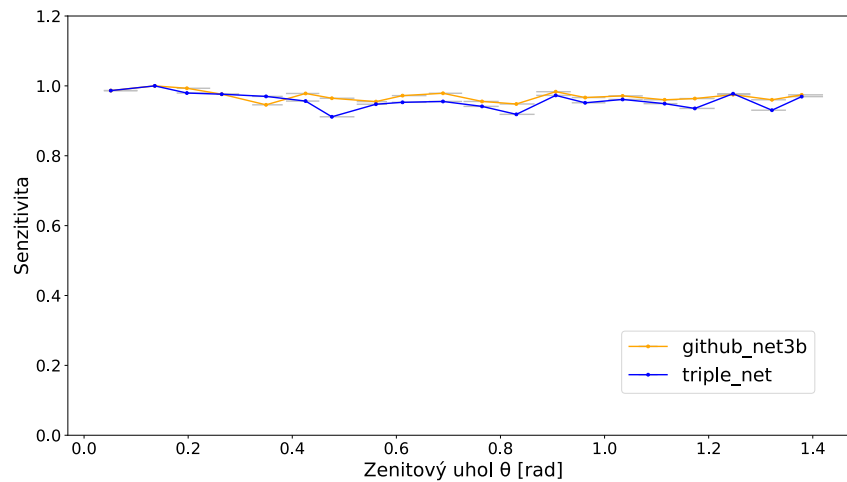
Obr. 6.2: Senzitivita natrénovaného modelu ako funkcie skutočnej energie simulovanej primárnej častice v MeV.



Obr. 6.3: Senzitivita natrénovaného modelu ako funkcie azimutálneho uhla  $\phi$  v radiánoch.

dát je pre krajné hodnoty tohto uhla, no v prípade validačnej podmnožiny je relatívne nižší počet vzoriek je aj v prípade hodnôt uhla theta rovných 0,8 a 1,3 radiánom. Ani jedna zo sietí nevykazovala nižšiu schopnosť predikcie v závislosti od hodnoty uhlov.

Ako vidno z tabuľky 6.1, obe modely vykazovali veľmi vysoký počet správnych klasifikácií šumu do správnej triedy a z toho vyplývajúci nízky počet faloš-



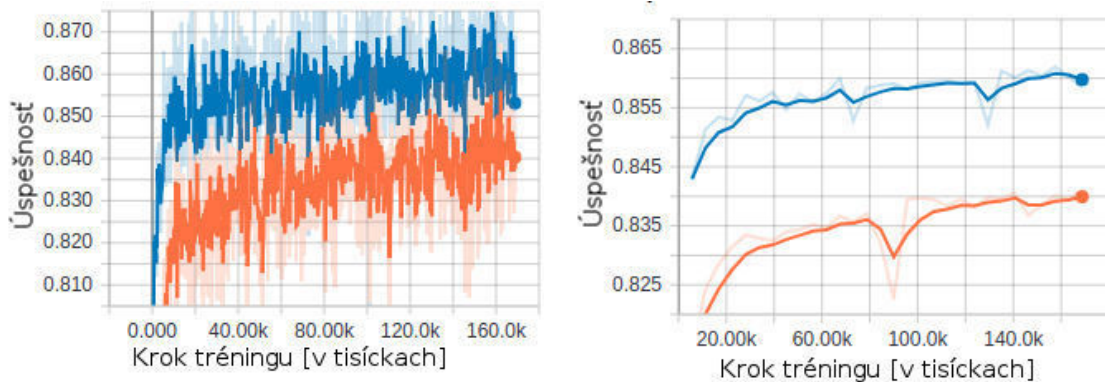
Obr. 6.4: Senzitivita natrénovaného modelu ako funkcie zenitového uhla  $\theta$  v radiánoch.

ných pozitív. Aby sme to kvantifikovali, zaznamenali sme, že z celkového počtu 7467 položiek v sade obsahujúcich šum bolo sieťou `github_net3b` klasifikovaných nesprávne iba 110 položiek. V prípade `triple_net` to bolo ešte menej, presne 34 položiek nesprávne klasifikovaných ako spŕšky.

## 6.2 Modely tréované na syntetických dátach

Priebeh tréningu znázornený na obrázku 6.5. Po istej dobe sa pri modeli s architektúrou `github_net3b` ustálila úspešnosť predpovedí na úrovni okolo 83%, kým pri architektúre `triple_net` na úrovni okolo 85%. Tento rozdiel úspešnosti oproti modelom trévaným len na simulovaných dátach nebol úplne neočakávaný, keďže sada syntetických dát bola zámerné vytvorená tak, aby v nej boli zahrnuté aj veľmi slabo viditeľné spŕšky s maximálnou hodnotou stopy oproti pozadiu na úrovni 3. Pri vyhodnotení senzitivity na validačnej podmnožine sa práve pri týchto hodnotách parametra `shower_max` ukazuje najvýraznejší prepád počtu úspešne klasifikovaných pozitív, ako vidno na obrázku 6.6.

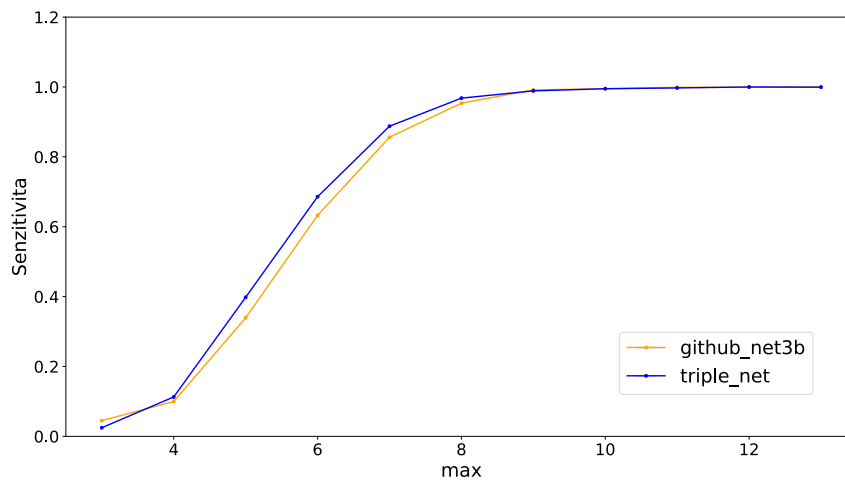
Výsledky vyhodnotení senzitivity na validačnej podmnožine sú uvedené pre 4 atribúty: `shower_max`, `bg_lambda`, `duration` a `track_length`. Pri vyhodnotení senzitivity podľa maxima spŕšky na obrázku 6.6 je vidieť najvýraznejší prepád správnej klasifikačnej presnosti pri hodnotách príslušného parametra menších než 7,



(a) Úspešnosť na trénovacej podmnožine (b) Úspešnosť na validačnej podmnožine

Obr. 6.5: Vývoj celkovej úspešnosti predpovedí na sade syntetických dát po 30 epochách tréningu pre siete `github_net3b` (oranžová) a `triple_net` (modrá).

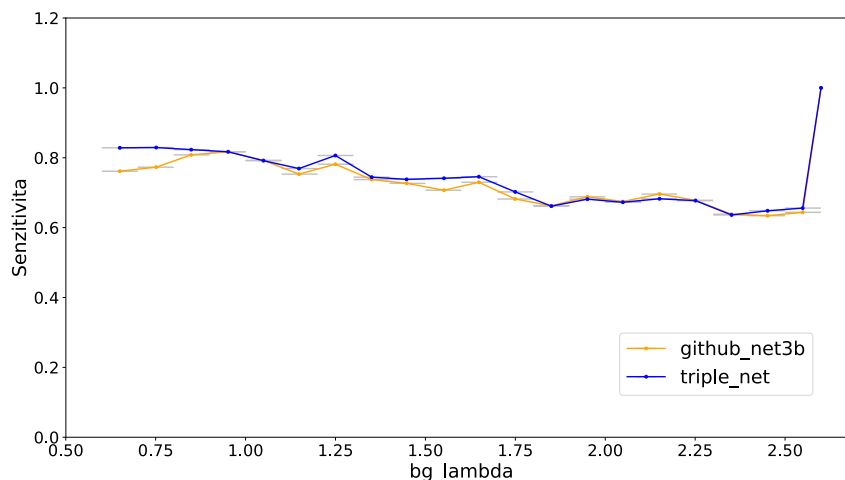
pričom pre hodnoty menej než 6 je vidieť dokonca prevahu falošných negatív.



Obr. 6.6: Senzitivita natrénovaného modelu ako funkcia parametra `shower_max`.

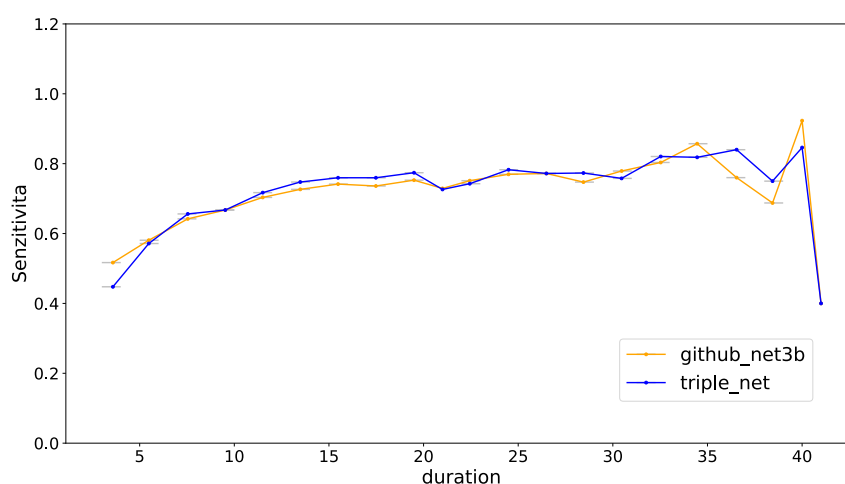
Pri intenzite pozadia je možné vidieť mierny pokles schopnosti správne klasifikovať spŕšky so zvyšujúcou sa intenzitou ako vidno na obrázku 6.7. Náhlý skok pri krajne vysokej hodnote tohto parametra je v tomto prípade zrejme dôsledkom nedostatku položiek pre danú úroveň pozadia; presnejšie 938 položiek pre hodnotu 2,6 v porovnaní s okolo 2000 položkami pre ostatné hodnoty odstupňované v intervaloch po 0,1.

Vyhodnotenie z pohľadu trvania spŕšky v počte snímok (obrázok 6.8) je zaujím-



Obr. 6.7: Senzitivita natrénovaného modelu ako funkcia parametra bg\_lambda.

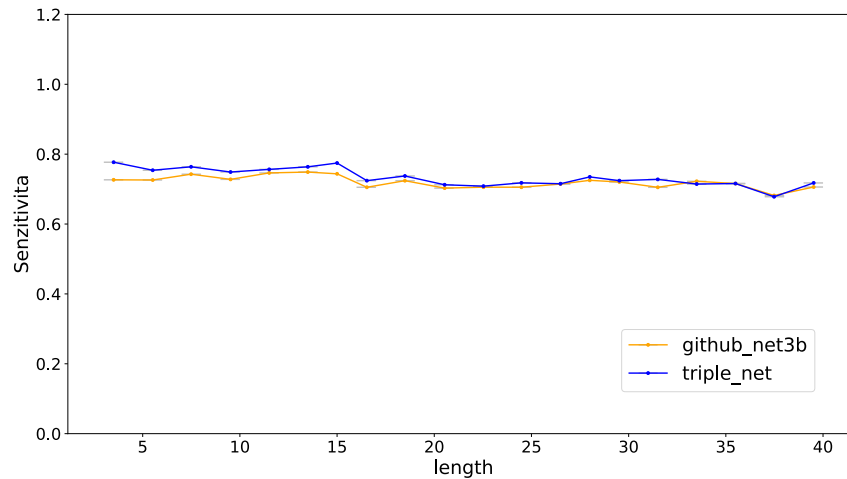
mavé aj z toho pohľadu, že zámerne boli vybrané hodnoty tohto parametra, ktoré presahujú hranice použitého výrezu 20 snímok paketu. Dôsledkom tejto voľby je, že stopa spŕšky bude viditeľná len z počiatkovej časti, kedy je rozdiel oproti pozadiu menší a teda stopa nemusí v dôsledku toho nikdy dosiahnuť intenzitu rovnajúcu sa shower\_max. Preto by sme očakávali, že rozpoznávanie spŕšok bude horšie pri takýchto vyšších hodnotách príslušného parametra, ale prekvapivo pozorujeme pokles špecificity len pri veľmi krátko trvajúcich spŕškach.



Obr. 6.8: Senzitivita natrénovaného modelu ako funkcia parametra duration.



Pri zdanlivej dĺžke stopy (obrázok 6.9) sme zaznamenali relatívne konštantnú senzitivitu pri jednotlivých hodnotách, ktoré naznačujú, že model je schopný pracovať podľa očakávaní aj s relatívne pomaly sa pohybujúcimi sprškami s pohľadu detektora.



Obr. 6.9: Senzitivita natrénovaného modelu ako funkcia parametra `track_length`.

Všetky modely vykazovali na validačnej podmnožine simulovaných dát približne rovnaké krivky senzitivity ako modely tréované len na simulovaných dátach. Rozdiely sa však primárne ukázali v počte falošných pozitív, ktoré názorne ukazuje tabuľka 6.1. Pri sieťach tréovaných len na simulovaných dátach bol počet falošných pozitív pri vyhodnotení na validačnej podmnožine týchto dát na úrovni maximálne 1,5% z celkového počtu položiek so šumom. Pri sieťach tréovaných na syntetických dátach však počet falošných pozitív stúpol veľmi výrazne, pri modeli siete `github_net3b` dokonca na úroveň okolo 98%, čo je neprimerane vysoko.

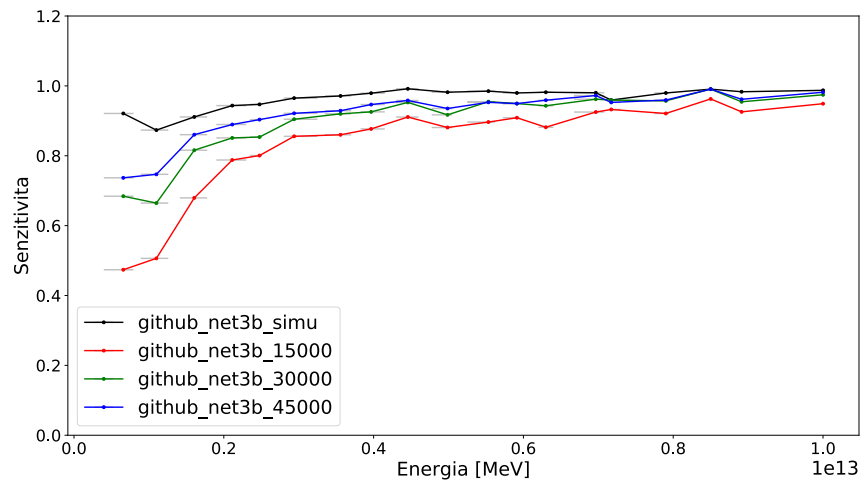
### 6.3 Modely tréované na zmiešaných dátach

Pri modeloch tréovaných na zmiešaných dátach sme pozorovali, že úspešnosť klasifikácie na vlastnej validačnej podmnožine pri tréningu sa ustálila na rovnakú prípadne vyššiu hodnotu ako modely tréované len na syntetických dátach. Dôležitejšie však bolo výrazné zlepšenie špecificity na modeloch takto tréovaných.

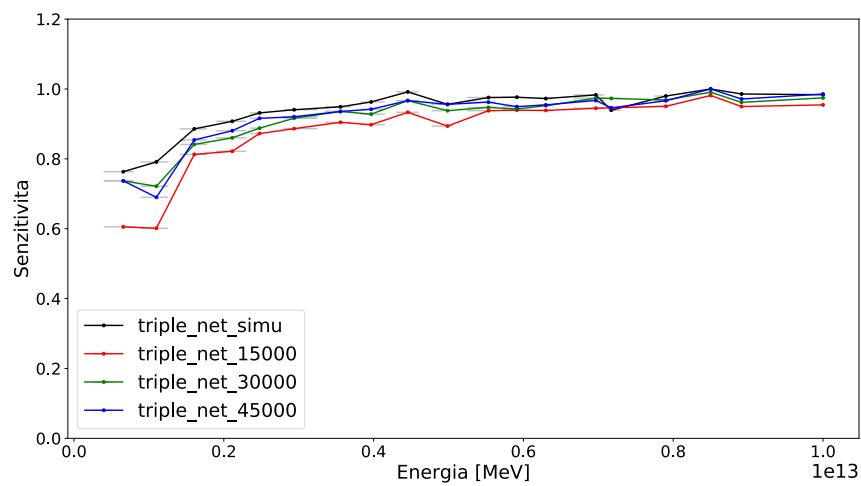
Ako vyplýva z tabuľky 6.1, pridaním len 15000 položiek simulovaných dát ku syntetickým sme na validačnej podmnožine simulovaných dát dosiahli výrazný pokles počtu falošných pozitív.

Zároveň vidieť, že pridaním ďalších simulovaných dát sme dosahovali postupne lepšie úrovne špecificity, hoci nie také dobré ako pri tréningu iba na simulovaných dátach. Modely založené na architektúre `github_net3b` mali nižší počet falošných pozitív ako modely založené na architektúre `triple_net`. Toto však trvalo iba dovtedy, kým nebola skúšaná sada augmentovaná so 45000 položkami simulovaných dát, kde model založený na architektúre `triple_net` vykazoval lepšie výsledky.

Krivky senzitivity podľa hodnôt energie (obrázok 6.10) a uhlov  $\theta$  (obrázok 6.11) a  $\phi$  (obrázok 6.11) boli porovnávané aj medzi modelmi trénovanými s rôznym podielom simulovaných dát v trénovacej sade. Taktiež sú krivky porovnané s krivkou vykazovanou modelom s rovnakou architektúrou no trénovanom iba na simulovaných dátach. Pri všetkých modeloch je vidieť postupné zvyšovanie senzitivity s nárastom podielu simulovaných dát. Zároveň je vidieť, že sieť `triple_net` vykazovala pri všetkých zmiešaných sadách krivku, ktorá je oveľa bližšie ku krivke vykazovanej ekvivalentným modelom, no trénovanom len na simulovaných dátach.

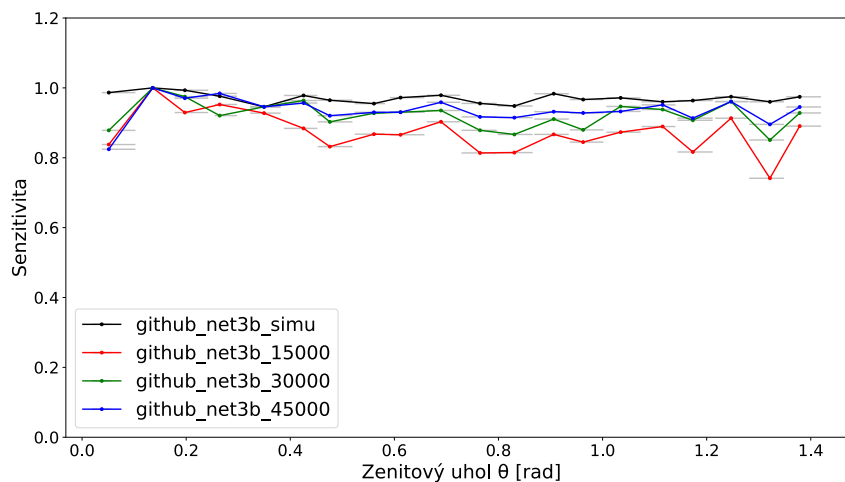


(a) github\_net3b

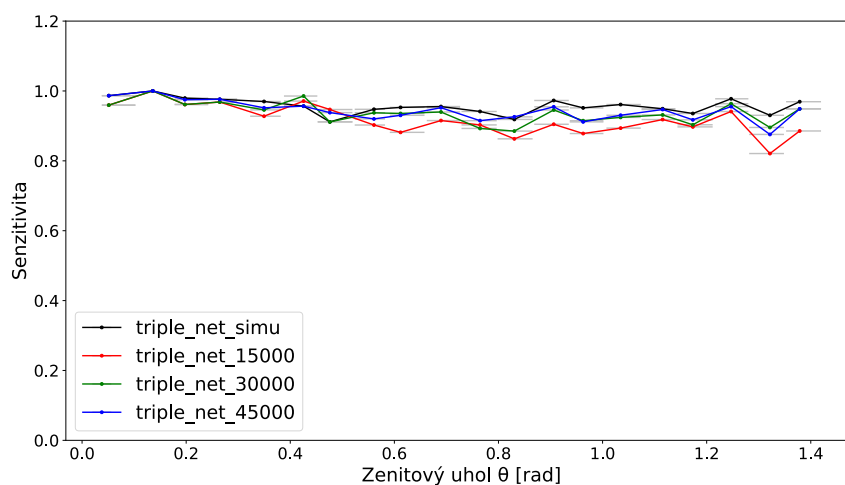


(b) triple\_net

Obr. 6.10: Senzitivita modelov trénovaných na zmiešaných sadách dát oproti modelom trénovaným len na simulovaných dátach ako funkcia skutočnej energie simulovanej primárnej častice v MeV.

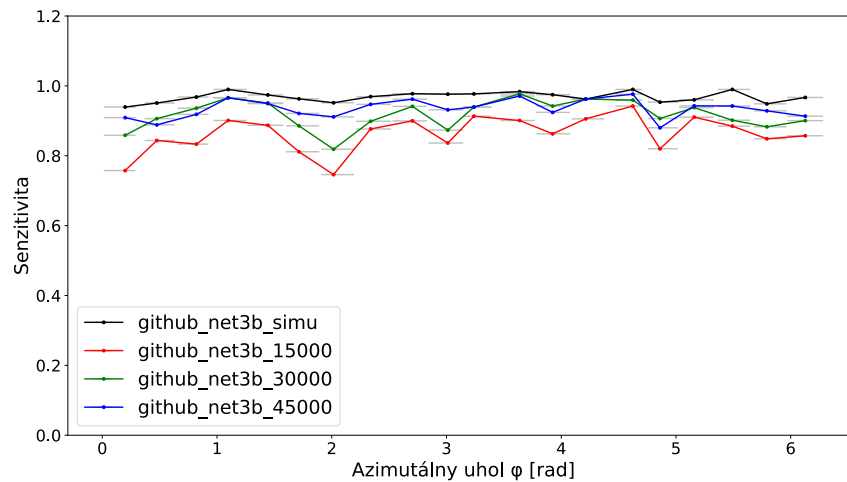


(a) github\_net3b

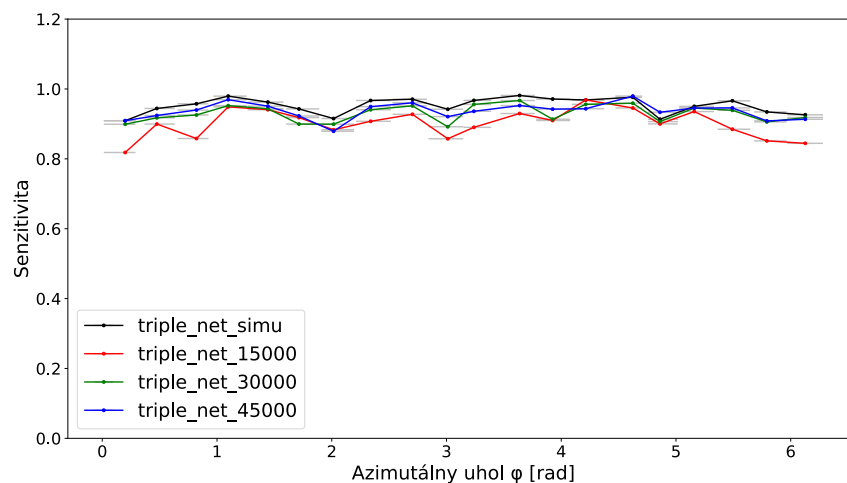


(b) triple\_net

Obr. 6.11: Senzitivita modelov tréovaných na zmiešaných sadách dát oproti modelom tréovaným len na simulovaných dátach ako funkcia zenitového uhla  $\theta$  v radiánoch.



(a) github\_net3b



(b) triple\_net

Obr. 6.12: Senzitivita modelov trénovaných na zmiešaných sadách dát oproti modelom trénovaným len na simulovaných dátach ako funkcia azimutálneho uhla  $\phi$  v radiánoch.

## 6.4 Vizualizácie máp príznakov skrytých vrstiev modelu

Pri tejto úlohe sme vizualizovali výstup prvej konvolučnej vrstvy pri všetkých modeloch. Jednou zo zaujímavostí, ktoré sme si všimli, je skutočnosť, že pri modeloch

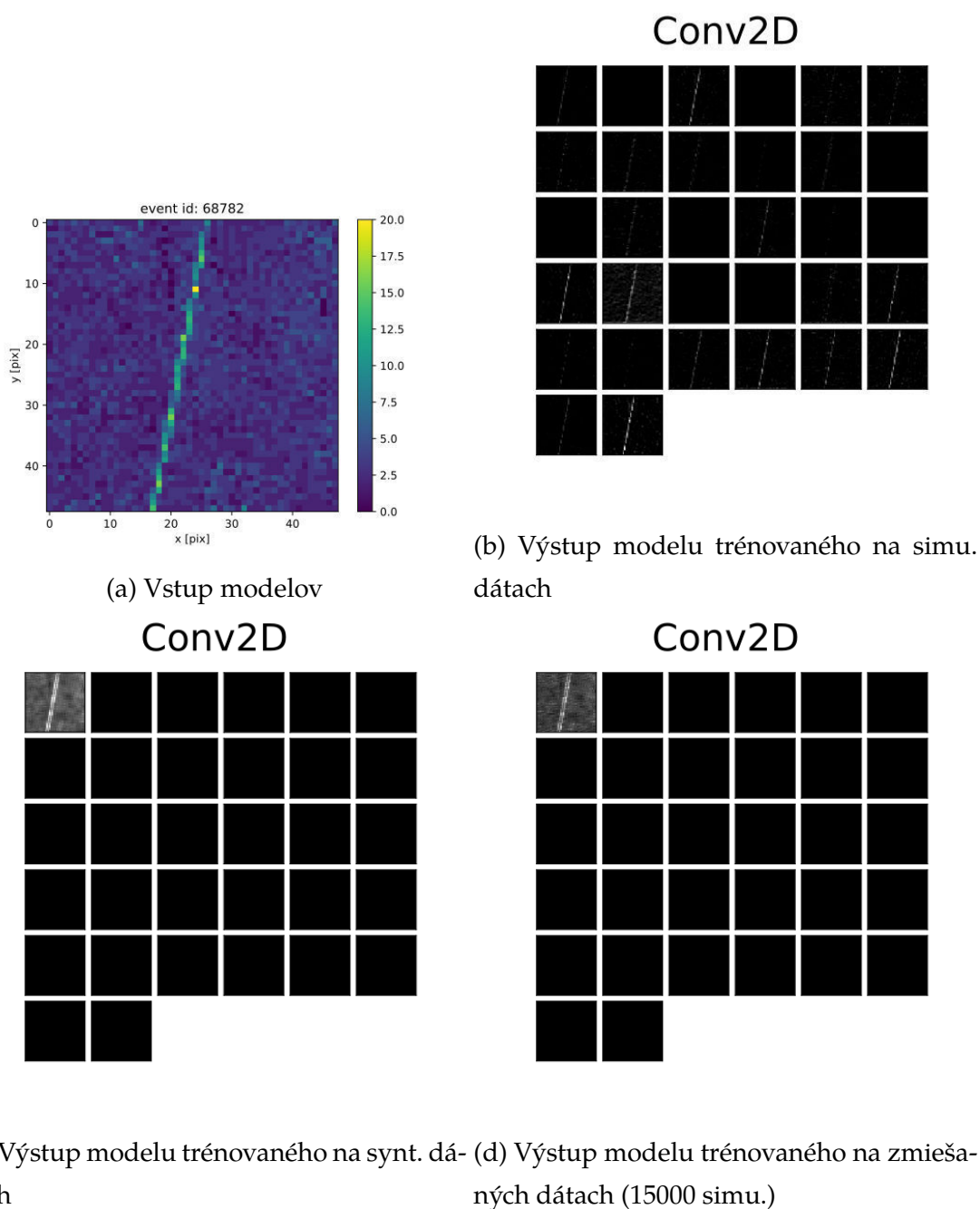
trénovaných len na syntetických dátach bol pri vybranej vzorke 20 položiek obsahujúcich šum aj spŕšku aktívny len výstup jedného filtra v konvolučnej časti siete pre spracovanie projekcie XY. Pri ostatných modeloch bolo aktívnych minimálne niekoľko filtrov, ako vidno pri porovnaní na obrázku 6.13. Dodatočne si možno všimnúť, že tento filter vykazuje oveľa menší kontrast medzi stále viditeľnou stopou spŕšky a pozadím, než pri zvyšných modeloch. Takýto nízky kontrast sme si všimli aj pri aktiváciách modelov trénovaných len na syntetických dátach, keď sme im posunuli vzorku obsahujúcu tento typ dát.

Pri výstupoch častí spracúvajúcich časové projekcie, ktoré sú prítomné len pri modeloch založených na architektúre `triple_net` sme zaznamenali o niečo vyššie počty aktívnych filtrov ako aj lepší kontrast medzi stopou spŕšky a pozadím, no stále menej ako pri modeli trénovanom len na simulovaných dátach. Toto naznačuje, že pri modeloch trénovaných na oboch typoch dát sa modely spracúvajúce aj časové projekcie naučili používať informácie v nich na rozpoznávanie a klasifikáciu, čo môže čiastočne vysvetliť približne desaťnásobný rozdiel v počte falošných pozitív medzi oboma modelmi trénovanými len na syntetických dátach v tabuľke 6.1.

## 6.5 Vyhodnotenie dát z misie

Pre vyhodnotenie dát z misie sme použili všetky modely okrem modelov trénovaných na syntetických dátach, kvôli neuspokojivým výsledkom pri ich vyhodnotení na sade simulovaných dát. Výsledky klasifikácie sú v tabuľke 6.2. Podľa očakávania modely trénované na najnižšom množstve simulovaných dát vyhodnotili oveľa viac položiek ako spŕšky. Okrem modelov trénovaných na zmesi s 15000 simulovanými dátami je vidieť nižší počet dát klasifikovaných ako spŕška pri modeloch založených na architektúre `triple_net`, než pri tých založených na architektúre `github_net3b`.

Po aplikovaní modelov na dáta sme vybrali podmnožinu tých dát z letu, ktoré každý model klasifikoval ako spŕšku a nechali sme tieto položky vizualizovať. Pri tejto analýze sme identifikovali 209 položiek, ktoré každý z modelov predpovedal, že obsahujú spŕšku. Z nich až 139 bolo zaznamenaných v rovnaký deň (28.4.2017) a v približne rovnakom čase a pochádzali z rovnakého zdrojového súboru. Keď sme si prešli príslušný zdrojový súbor, vyšlo nám, že dáta všetkých paketov v ňom



Obr. 6.13: Porovnanie aktivácií prvej konvolučnej vrstvy spracúvajúcej projekciu XY pri modeloch založených na architektúre `triple_net`.

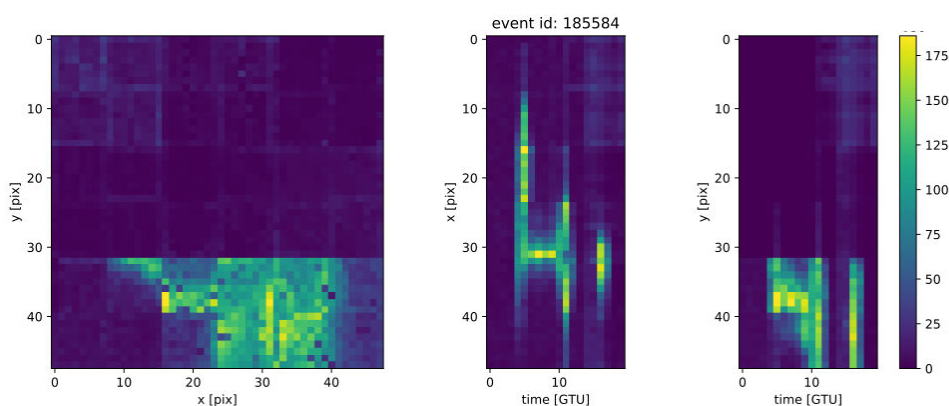
sú identické. Nie je nám celkom jasné ako k tomuto mohlo dôjsť. Týchto 139 položiek sme ignorovali ako možný dôsledok chyby pri sťahovaní alebo ukladaní.

Pri ostatných položkách klasifikovaných všetkými modelmi ako sprška sme pozorovali najčastejšie dve špecifické kategórie úkazov. Prvou je rozptýlené zjas-

Architektúra	Trénovacie dáta	Počet spŕšok
github_net3b	simulované	2200
triple_net	simulované	628
github_net3b	zmiešané (15000 simu.)	18763
triple_net	zmiešané (15000 simu.)	19315
github_net3b	zmiešané (30000 simu.)	8484
triple_net	zmiešané (30000 simu.)	4331
github_net3b	zmiešané (45000 simu.)	5299
triple_net	zmiešané (45000 simu.)	3290

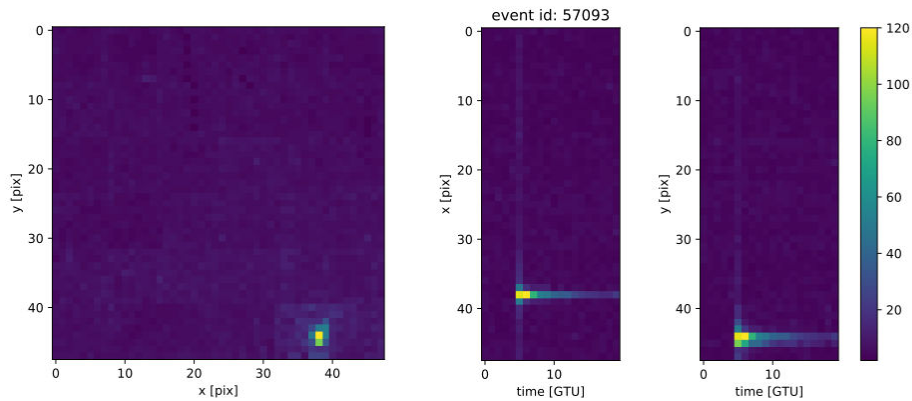
Tabuľka 6.2: Vyhodnotenie špecificity modelov

nenie veľkej časti snímacej plochy detektora, ako možno vidieť na obrázku 6.14. Druhou boli dáta, kde vidno výrazné zvýšenie intenzity v jednom pixeli alebo skupine združených pixelov, ktoré časom mizne. Takýto príklad je vidieť na obrázku 6.15. Ďalej sme zistili, že 45 týchto položiek tvoria dáta zaznamenané 28.4. a 29.4.2017. Pri pohľade na profil výšky letu počas misie na obrázku 2.4 si možno všimnúť, že počas týchto dvoch dní sa počas pozorovania nachádzal balón iba o 6 km nižšie oproti plánovanej výške 30 km nad povrchom. Z ďalších dní letu, kedy balón pri pozorovaní zostúpil aj do polovičnej výšky oproti plánovanej, bolo iba 6 položiek označených všetkými modelmi, že obsahujú spŕšku.



Obr. 6.14: Dáta z misie z 28.4.2017 označené modelmi ako spŕška ukazujúce náhle rozptýlené zjasnenie meracej plochy detektora.





Obr. 6.15: Dáta z misie z 28.4.2017 označené modelmi ako spíška ukazujúce nárazovité zjasnenie pixelu resp. viacerých pixelov a postupné zoslabnutie intenzity.

## 6.6 Interpretácia výsledkov

Výsledky naznačujú, že konvolučné siete pracujúce iba s projekciami sú pomerne perspektívnym spôsobom dátovej analýzy v misii EUSO-SPB a pri doladení a prispôbení potenciálne aj pri iných misiách. Zároveň sa ukazuje istá nižšia úspešnosť predikcií pri modeloch založených na architektúre spracúvajúcej len jednu projekciu oproti modelom spracúvajúcim všetky tri projekcie pri dostatočnom množstve vhodných tréningových dát. Toto pripisujeme skutočnosti, že druhý typ architektúr má prístup aj k informáciám o časovej následnosti snímok paketu.

Hoci rozdiel je relatívne malý, pri analýze dostatočne veľkého množstva dát môže reprezentovať rádovo aj stovky alebo tisíce nesprávne klasifikovaných položiek. S tým sú zároveň ale spojené vyššie náklady pri tréningu architektúr spracúvajúcich viaceré projekcie v podobe väčšieho počtu tréningových dát, keďže počet silno závisí od množstva nastaviteľných parametrov modelu. Zaujímavou myšlienkou môže byť v tomto ohľade sieť spracúvajúca iba jednu alebo obe projekcie v čase, keďže takáto sieť by mala menej parametrov než sieť s tromi projekciami a zároveň prístup k informáciám o čase.

Zároveň sa v aktuálnej podobe ukazuje náš spôsob generovania dát pre tréning ako nevhodný pre tréning modelov pre spracovanie dát z misie. Naša interpretácia je, že spôsob generovania týchto dát zrejme nedokáže plne reprezentovať ar-

tefakty šumu pozadia je lokálne zvýšenie signálu počas jedného GTU. Na druhej strane, takéto schopnosti by bolo možné zakomponovať do nášho generátora dát, no je otázne či zvýšená komplexnosť výpočtu nespôsobí nakoniec, že výsledný spôsob generovania dát bude rovnako náročný ako rámec ESAF.

Modely trénované na zmesi syntetických dát so simulovanými boli schopné zlepšiť celkovú úspešnosť klasifikácií na validačnej sade simulovaných dát pomerne rýchlo. Je však otázne do akej miery sú syntetické dáta vhodné na augmentáciu trénovacích dát. Prvé modely približujúce sa v celkovej úspešnosti tých trénovaných len na simulovaných dátach potrebovali prítomnosť minimálne 30000 simulovaných dát, čo už samo osebe môže predstavovať dostatočne veľkú tréningovú sadu minimálne pre sieť so vstupom len jedenej projekcie.

## 7 Záver

---

V tejto práci sme sa zamerali na perspektívu využitia metód strojového učenia, špecificky konvolučných neurónových sietí pre spracovanie obrazových dát, pre klasifikáciu dát z misie EUSO-SPB za účelom vyhľadania potenciálnych dát obsahujúcich stopu extenzívnej atmosférickej spŕšky. Použili sme dáta generované dvoma rozličnými spôsobmi za účelom tréningu modelov a natrénovali sme modely osobitne na jednom z typov dát a oboch typoch dát. Natrénované modely sme vyhodnotili pomocou metrík senzitivity na prítomnosť stopy spŕšky v dátach validačnej sady v závislosti od hodnôt špecifických atribútov dát obsahujúcich spŕšku. Taktiež sme vyhodnotili na príslušnej validačnej sade aj pomer správne klasifikovaných negatív, čiže položiek so šumom.

Zohľadňujúc výsledky vyhodnotenia sme aplikovali 8 natrénovaných modelov vykazujúcich aspoň 90% špecificitu na vzorke simulovaných dát a pomocou nich vyhodnotili dáta z misie. Identifikovali sme takto vyše 200 položiek dát, ktoré každý z modelov vyhodnotil ako potenciálne zaujímavé. Vzhľadom na krátke trvanie misie a odhadovaný počet zaznamenaných spŕšok počas nej, nie je vôbec isté, že vo vybraných dátach je spŕška vôbec prítomná. Spoločné črty vybranej vzorky dát ale naznačujú, že takéto modely majú perspektívu minimálne na dodatočnú redukciu alebo predvýber zozbieraných dát v kombinácii s manuálnym vyhodnotením tých dát označených ako spŕška. Tiež naznačujú možnú aplikáciu v takejto úlohe aj pri dátach z iných misií zastrešovaných kolaboráciou JEM-EUSO.

Taktiež nám z našich výsledkov vyplýva, že pre tréning takýchto modelov v kontexte spracovania dát tejto misie sú potrebné dáta s realistickým pozadím, ktoré naša metóda používajúca pre jeho simuláciu iba Poissonovu distribúciu hodnôt pixelov, nebola schopná dosiahnuť. Na druhej strane nám tiež vyplýva, že aj dáta z tejto misie môžu slúžiť ako pozadie pre trénovacie dáta so simulovanými spŕškami. Hoci niektoré artefakty prítomné v dátach z misie naša metóda gene-

rovania dát nevie simulovať, bolo by možné takúto simuláciu implementovať. Výhody takto upraveného spôsobu generovania dát sú však otáznе, zvlášť pokiaľ takýto generátor by v konečnom dôsledku bol rovnako náročný výpočtovo prípadne aj konceptuálne ako existujúci softvérový rámec ESAF, používaný na rovnaký účel.

Naša práca sa primárne zameriavala na klasifikáciu dát. Myšlienkou bolo zredukovať množstvo dát z misie vylúčením tých označených ako šum a zvyšné dáta posunúť automatizovaným algoritmom pre rekonštrukciu parametrov spŕšky, prípadne manuálnym metódam klasifikácie. Uviedli sme však medzi príbuznými výskumami aj Erdmann et al., 2017, ktorí sa zamerali na využitie strojového učenia pre rekonštrukciu dodatočných príznakov EAS v dátach pozemných detektorov. Z tohto pohľadu by bolo zaujímavé overiť vhodnosť metód strojového učenia aj pri rekonštrukcii parametrov spŕšky a primárnej častice.

Dodatočne, prítomnosť časovej dimenzie sa zdá byť dôležitá pre odlišenie falošných pozitív, no zároveň architektúry prijímajúce všetky tri projekcie cez separátne sady konvolučných vrstiev sú zároveň náročnejšie na množstvo tréningových dát keďže majú aj vyšší počet parametrov. Možným kompromisom by mohla byť architektúra spracúvajúca buď len jednu alebo obe projekcie v čase, teda úplné vynechanie projekcií XY zo vstupných dát.

Zámerne sme počas našej práce vynechali typy 3D konvolučné siete ako aj iné typy neurónových sietí určené na spracovanie videa. Keďže dáta z misie reprezentujú vo svojej podstate krátke sekvencie snímok, jedným z možných smerov výskumu v tejto oblasti môže byť aj preskúmanie ich vhodnosti pri klasifikácii dát buď z tejto misie alebo prípadne aj z iných misií podobného charakteru. Inou možnosťou výskumu je aj rozšírenie počtu klasifikačných tried pre rozlíšenie viacerých typov šumu, ako je používané pri existujúcich pokusoch o manuálnu klasifikáciu dát z misie.

# Literatúra

---

- AAB, A. et al., 2015. The Pierre Auger Cosmic Ray Observatory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. Roč. 798, s. 172–213. ISSN 0168-9002. Dostupné z DOI: <https://doi.org/10.1016/j.nima.2015.06.058>.
- ABDELLAOUI, G. et al., 2017. Cosmic ray oriented performance studies for the JEM-EUSO first level trigger. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. Roč. 866, s. 150–163. Dostupné z DOI: [10.1016/j.nima.2017.05.043](https://doi.org/10.1016/j.nima.2017.05.043).
- ADAMS JR., James H. et al., 2017. White paper on EUSO-SPB2. Dostupné z eprint: [arXiv:1703.04513v2](https://arxiv.org/abs/1703.04513v2).
- ADAMS, J. H. et al., 2015. The JEM-EUSO mission: An introduction. *Experimental Astronomy*. Roč. 40, č. 1, s. 3–17. Dostupné z DOI: [10.1007/s10686-015-9482-x](https://doi.org/10.1007/s10686-015-9482-x).
- AURISANO, A. et al., 2016. A Convolutional Neural Network Neutrino Event Classifier [2016 JINST 11 P09001]. Dostupné z DOI: [10.1088/1748-0221/11/09/P09001](https://doi.org/10.1088/1748-0221/11/09/P09001).
- BALLARD, W.; DAMIEN, A., 2016. *Convolutional Neural Network for MNIST dataset classification task* [online] [cit. 2018-06-01]. Dostupné z: [https://github.com/tflearn/tflearn/blob/master/examples/images/convnet\\_mnist.py](https://github.com/tflearn/tflearn/blob/master/examples/images/convnet_mnist.py).
- BALLMOOS, Peter von, 2016. The EUSO-BALLOON mission. In: *The EUSO-BALLOON mission. Proceedings of The 34th International Cosmic Ray Conference — PoS(ICRC2015)*. Sissa Medialab. Dostupné z DOI: [10.22323/1.236.0322](https://doi.org/10.22323/1.236.0322).

- BAYER, J. et al., 2013. Second level trigger and Cluster Control Board for the JEM-EUSO mission. In: *Second level trigger and Cluster Control Board for the JEM-EUSO mission. The JEM-EUSO Mission: Contributions to the ICRC 2013*, s. 99–102. Dostupné z eprint: arXiv:1307.7071.
- BENSON, R.; LINSLEY, J., 1981. Satellite observation of cosmic ray air showers. *International Cosmic Ray Conference*. Roč. 8, s. 145–148.
- BERTAINA, Mario E.; BAYER, Joerg; FENU, Francesco; MIGNONE, Marco; MIYAMOTO, Hiroko; SHINOZAKI, Kenji; CUMMINGS, Austin; ESER, Johannes B.; JUNG, Aera, 2017. The trigger logic of EUSO-SPB and its performance. In: *The trigger logic of EUSO-SPB and its performance. Proceedings of 35th International Cosmic Ray Conference — PoS(ICRC2017)*. Sissa Medialab. Dostupné z DOI: 10.22323/1.301.0443.
- BERTAINA, M. et al., 2014. Performance and air-shower reconstruction techniques for the JEM-EUSO mission. *Advances in Space Research*. Roč. 53, č. 10, s. 1515–1535. Dostupné z DOI: 10.1016/j.asr.2014.02.018.
- BRUN, Rene; RADEMAKERS, Fons, 1997. ROOT — An object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. Roč. 389, č. 1, s. 81–86. ISSN 0168-9002. Dostupné z DOI: [https://doi.org/10.1016/S0168-9002\(97\)00048-X](https://doi.org/10.1016/S0168-9002(97)00048-X). New Computing Techniques in Physics Research V.
- DETTAT, Arden, 2017. Applied Deep Learning - Part 4: Convolutional Neural Networks. *Towards Data Science* [online] [cit. 2019-04-20]. Dostupné z: <https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2>.
- ERDMANN, M. et al., 2017. A Deep Learning-based Reconstruction of Cosmic Ray-induced Air Showers. *Astroparticle Physics*. ISSN 0927-6505. Dostupné z DOI: 10.1016/j.astropartphys.2017.10.006.
- FENU, Francesco, 2013. *A Simulation Study of the JEM-EUSO Mission for the Detection of Ultra-High Energy Cosmic Rays*. Dizertačná práca. Eberhard Karls Universität Tübingen.

- FENU, Francesco; BERTAINA, Mario E.; BORTONE, Alberto; CUMMINGS, Austin; SAKAKI, Naoto; VENEZIANI, Andrea; CAMBURSANO, Simone, 2017. Expected number of Extensive Air Showers observable by EUSO-SPB. In: *Expected number of Extensive Air Showers observable by EUSO-SPB. Proceedings of 35th International Cosmic Ray Conference — PoS(ICRC2017)*. Sissa Medialab. Dostupné z DOI: 10.22323/1.301.0426.
- GUZMAN, A; MERNIK, T; SANTANGELO, A; MEDINA-TANCO, G; BERTAINA, M; SHINOZAKI, K; FENU, F; GORGI, A, 2013. The Peak and Window Searching Technique for the EUSO Simulation and Analysis Framework: Impact on the Angular Reconstruction of EAS. *Journal of Physics: Conference Series*. Roč. 409, s. 012104. Dostupné z DOI: 10.1088/1742-6596/409/1/012104.
- HANDZEL, Amir A.; GROSSMAN, T.; DOMANY, E.; TAREM, S.; DUCHOVNI, E., 1993. A neural network classifier in experimental particle physics. *International Journal of Neural Systems*. Roč. 04, č. 02, s. 95–108. Dostupné z DOI: 10.1142/s0129065793000109.
- HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian, 2015. Deep Residual Learning for Image Recognition. *CoRR*. Roč. abs/1512.03385. Dostupné z arXiv: 1512.03385.
- LAARHOVEN, Twan van, 2017. L2 Regularization versus Batch and Weight Normalization. *CoRR*. Roč. abs/1706.05350. Dostupné z arXiv: 1706.05350.
- LINCOLN, Nicholas, 2017. *Neural Network Classification of Cosmic Particles* [online] [cit. 2019-04-20]. Dostupné z: <https://www.linkedin.com/pulse/neural-net-classification-cosmic-particles-nicholas-lincoln>.
- LUKAS HOLCH, T. et al., 2017. Probing Convolutional Neural Networks for Event Reconstruction in  $\gamma$ -Ray Astronomy with Cherenkov Telescopes. *ArXiv e-prints*. Dostupné z arXiv: 1711.06298.
- MACHINE LEARNING GURU, 2017. *Understanding Convolutional Layers in Convolutional Neural Networks (CNNs)* [online] [cit. 2017-11-18]. Dostupné z: [https://web.archive.org/web/20171118012106/http://machinelearningguru.com/computer\\_vision/basics/convolution/convolution\\_layer.html](https://web.archive.org/web/20171118012106/http://machinelearningguru.com/computer_vision/basics/convolution/convolution_layer.html).
- MACCARONE, M. Concetta; KAJINO, F, 2011. *The JEM-EUSO Purple Book: Report on the Phase A Study 2010, December 2010*.

- MARCELLI, L. et al. *JEM-EUSO Experiment* [online] [cit. 2019-04-20]. Dostupné z: <http://jem-euso.roma2.infn.it>.
- MERNIK, Thomas, 2014. *The Expected Angular Resolution of the JEM-EUSO Mission*. Dizertačná práca. Eberhard Karls Universität Tübingen.
- NIETO, D.; BRILL, A.; KIM, B.; HUMENSKY, T. B.; CHERENKOV TELESCOPE ARRAY, for the, 2017. Exploring deep learning as an event classification method for the Cherenkov Telescope Array [PoS(2017)809]. Dostupné z eprint: arXiv:1709.05889.
- OLIPHANT, Travis, 2006. *NumPy: A guide to NumPy* [USA: Trelgol Publishing].
- OSTERIA, Giuseppe; PERFETTO, Francesco; SCOTTI, Valentina; CAFAGNA, Francesco; AND, Claudio Fornaro, 2017. The Data Processor System of EUSO-SPB. In: *The Data Processor System of EUSO-SPB. Proceedings of 35th International Cosmic Ray Conference — PoS(ICRC2017)*. Sissa Medialab. Dostupné z DOI: 10.22323/1.301.0461.
- RICCI, Marco, 2018. Mini-EUSO: A Precursor Mission on the International Space Station for the Observation of Atmosphere and Earth in the UV Light. In: *Mini-EUSO: A Precursor Mission on the International Space Station for the Observation of Atmosphere and Earth in the UV Light. Proceedings of 2016 International Conference on Ultra-High Energy Cosmic Rays (UHECR2016)*. Journal of the Physical Society of Japan. Dostupné z DOI: 10.7566/jpscp.19.011040.
- SCOTTI, Valentina; OSTERIA, Giuseppe, 2017. The EUSO-SPB mission. In: *The EUSO-SPB mission. Proceedings of The European Physical Society Conference on High Energy Physics — PoS(EPS-HEP2017)*. Sissa Medialab. Dostupné z DOI: 10.22323/1.314.0024.
- SRIVASTAVA, Nitish; HINTON, Geoffrey; KRIZHEVSKY, Alex; SUTSKEVER, Ilya; SALAKHUTDINOV, Ruslan, 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* Roč. 15, č. 1, s. 1929–1958. ISSN 1532-4435. Dostupné tiež z: <http://dl.acm.org/citation.cfm?id=2627435.2670313>.
- SZEGEDY, Christian; LIU, Wei; JIA, Yangqing; SERMANET, Pierre; REED, Scott; ANGUELOV, Dragomir; ERHAN, Dumitru; VANHOUCHE, Vincent; RABINOVICH, Andrew, 2015. Going Deeper with Convolutions. In: *Going Deeper*



- with Convolutions. Computer Vision and Pattern Recognition (CVPR)*. Dostupné tiež z: <http://arxiv.org/abs/1409.4842>.
- SZEGEDY, Christian; VANHOUCHE, Vincent; IOFFE, Sergey; SHLENS, Jonathon; WOJNA, Zbigniew, 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR*. Roč. abs/1512.00567. Dostupné z arXiv: 1512.00567.
- VAICIULIS, A., 2002. Support Vector Machines in Analysis of Top Quark Production [Nucl.Instrum.Meth. A502 (2003) 492-494]. Dostupné z DOI: 10.1016/S0168-9002(03)00479-0.
- VANNEREM, P.; MUELLER, K. -R.; SCHOELKOPF, B.; SMOLA, A.; SOLDNER-REMBOLD, S., 1999. *Classifying LEP Data with Support Vector Algorithms*. Dostupné z eprint: arXiv:hep-ex/9905027.
- VRÁBEL, Michal, 2017. *Pattern Recognition in the JEM-EUSO Experiment*. Písomná práca ku dizertačnej skúške.
- VRÁBEL, Michal, 2019. An Approach for Air Shower Pattern Recognition in EUSO-SPB Data. In: *An Approach for Air Shower Pattern Recognition in EUSO-SPB Data. Proceedings of 18th Scientific Conference of Young Researchers* [online]. Fakulta elektrotechiky a informatiky v Košiciach, Technická univerzita v Košiciach, s. 30–31 [cit. 2019-04-25]. ISBN 978-80-553-3273-4. Dostupné z: [http://scyr.kpi.fei.tuke.sk/wp-content/scyr-files/winners/SCYR\\_2019\\_Proceedings.pdf](http://scyr.kpi.fei.tuke.sk/wp-content/scyr-files/winners/SCYR_2019_Proceedings.pdf).
- WIENCKE, Lawrence; OLINTO, Angela, 2017. EUSO-SPB1 Mission and Science. In: *EUSO-SPB1 Mission and Science. Proceedings of 35th International Cosmic Ray Conference — PoS(ICRC2017)*. Sissa Medialab. Dostupné z DOI: 10.22323/1.301.1097.
- XIANGYU ZHANG, Kaiming He and; REN, Shaoqing; SUN, Jian, 2015. Deep Residual Learning for Image Recognition. *CoRR*. Roč. abs/1512.03385. Dostupné z arXiv: 1512.03385.
- ZEILER, Matthew D.; FERGUS, Rob, 2013. Visualizing and Understanding Convolutional Networks. *CoRR*. Roč. abs/1311.2901. Dostupné z arXiv: 1311.2901.

# Zoznam skratiek

---

**CCB** Doska pre riadenie klastra, z angl. Cluster control board.

**EAS** Extenzívna atmosférická spŕška, z angl. Extensive air shower.

**EC** Elementárna bunka, z angl. Elementary cell.

**ESAF** EUSO softvérový rámec pre simuláciu a analýzu, z angl. EUSO simulation and analysis framework.

**FLT** Spúšťač prvej úrovne, z angl. First-level trigger.

**GTU** Základná časová jednotka, z angl. Gate time unit; hodnota  $2.5\mu s$ .

**ISS** Medzinárodná vesmírna stanica, z angl. International space station.

**MAPMT** Multi-anódový fotonásobič, z angl. Multi-anode photomultiplier tube.

**MLP** Viacvrstvový perceptrón z angl. Multi-layer perceptron.

**PDM** Fotodetekčný modul, z angl. Photo-detector module.

**SLT** Spúšťač druhej úrovne, z angl. Second-level trigger.

**SPB** Vysokotlakový balón, z angl. Super-pressure balloon.

**UHECR** Kozmické žiarenie vysokých energií, z angl. Ultra-high energy cosmic ray.

**UV** Ultrafialové, z angl. Ultra-violet.

# Zoznam príloh

---

**Príloha A** Používateľská príručka

**Príloha B** Systémová príručka

**Príloha C** CD médium – zdrojové kódy projektu, záverečná práca bez príručiek, používateľská príručka a systémová príručka v elektronickej podobe.