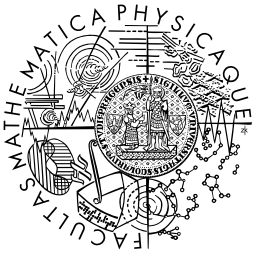


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Samuel Amrich

Vývoj, implementace a testování algoritmů pro měření elektromagnetických signálů detekovaných na palubě stratosférického balónu nad bouřkovými oblastmi

Katedra fyziky povrchů a plazmatu

Vedoucí bakalářské práce: prof. RNDr. O. Santolík, Dr.
Studijní program: Fyzika (B0533A110001)
Studijní obor: FP (0533RA110001)

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Ďakujem vedúcemu bakalárskej práce prof. RNDr. Ondřejovi Santolíkovi, Dr. za jeho cenné vedenie, rady a pripomienky počas celej doby písania práce. A v neposlednom rade aj za samotnú možnosť pracovať na danej téme. Ďalej moje poďakovanie patrí aj konzultantke práce Ing. Ivane Kolmašovej, Ph.D. za jej neochvejnú a vždy dostupnú výpomoc. Nerád by som som zabudol na poďakovanie svojím rodičom. Mojej mame ktorá ma vždy podporovala a môjmu otcovi ktorý ma naučil mnoho do života. A za dlhoročnú spoluprácu a veľké množstvo získaných znalostí patrí moje poďakovanie aj doc. RNDr. Rudolfovi Gálisovi, PhD. a RNDr. Šimonovi Mackovjakovi, PhD.

Název práce: Vývoj, implementace a testování algoritmů pro měření elektromagnetických signálů detekovaných na palubě stratosférického balónu nad bouřkovými oblastmi

Autor: Samuel Amrich

Katedra: Katedra fyziky povrchů a plazmatu

Vedoucí bakalářské práce: prof. RNDr. O. Santolík, Dr., Katedra fyziky povrchů a plazmatu

Abstrakt: V tejto práci sa zaoberáme využitím postupov strojového učenia na skúmanie rádiových záznamov bleskových výbojov. Cieľom bolo vytvoriť algoritmus, ktorý by dokázal autonómne detegovať a klasifikovať rôzne typy bleskových výbojov alebo ich skupín alebo ich vývojových častí. Na tento účel sme otestovali rôzne metódy klasického strojového učenia ako aj hlbokých neurónových sieti. Všetky tieto modely boli trénované iteratívnym postupom na archívnych dátach. Výsledky práce ukazujú, že je možné efektívne využiť metódy strojového učenia na detekciu a klasifikáciu za účelom nasadenia na palube stratosférického balónu projektu STRATELEC. V budúcnosti by naše výsledky mali byť nasadené na palube stratosférického balóna a mohli by byť využité na zlepšenie pochopenia pochodov v pri tvorbe bleskového výboja v búrkovom oblaku.

Klíčová slova: bleskové výboje, šírenie elektromagnetických vĺn, strojové učenie

Title: Development, implementation and testing of algorithms for measurements of electromagnetic signals detected onboard a stratospheric balloon above thunderstorms

Author: Samuel Amrich

Department: Department of Surface and Plasma Science

Supervisor: prof. RNDr. O. Santolík, Dr., Department of Surface and Plasma Science

Abstract: In this work, we focus on the application of machine learning techniques to study radio recordings of lightning discharges. Our goal was to develop an algorithm that could autonomously detect and classify different types of lightning discharges, as well as their groups or developmental stages. To this end, we tested various methods of classical machine learning as well as deep neural networks. All of these models were trained iteratively on archival data. The results of our work show that it is possible to effectively use machine learning methods for detection and classification purposes, with the aim of deploying them on board the STRATELEC stratospheric balloon project. In the future, our findings should be deployed on board the stratospheric balloon and could be used to improve understanding of the processes involved in creating lightning discharges in a thundercloud.

Keywords: lightning discharges, radio wave propagation, machine learning

Obsah

Úvod

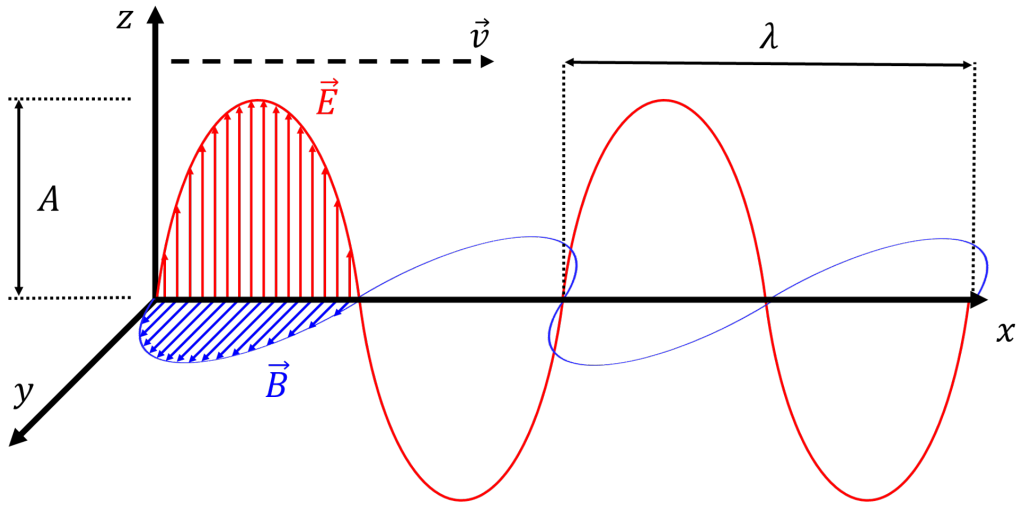
Počas búrok dochádza v búrkových oblakoch k celej škále javov, ktoré vedú k vzniku bleskového výboja. Ako tieto vnútrooblakové javy, tak aj bleskový výboj samotný vyžarujú elektromagnetické vlny. Tie sa šíria do celého priestoru od miesta vzniku a môžu byť zachytené rádiovými prijímačmi na pozemných staniciach, balónoch alebo dokonca družiciach. Povaha týchto elektromagnetických vĺn je odzrkadlením povahy udalosti ktorá danú vlnu vyvolala. Pričom hlbším štúdiom týchto vĺn sa vieme dozvedieť nie len o príslušnej kategórií udalosti ale aj jej hlbších vlastnostiach.

V prípade balónových alebo družicových meraní je prenosová kapacita značne obmedzená. Preto vzniká nutnosť efektívnych algoritmov pre výber a predspracovanie nameraných dát. K tomuto účelu je možné využiť programovateľné hradlové polia. Čo je prípad aj projektu STRATELEC (STRatéole-2 ATmospheric ELECtricity), Ktorý bude vypustený počas poslednej balónovej kampane projektu Stratéole-2 s plánovaným štartom v roku 2025. Celkovo sa má projekt zamerať na štúdium dynamiky na rozhraní medzi troposférou a stratosférou. Jednotlivé body ktoré má STRATELEC naplniť sú nasledujúce. Dokumentovanie elektrického stavu atmosféry a vysoko energetickej radiácie na mieste jeho vzniku za účelom lepšieho pochopenia a zlepšenia modelovania búrok. Ďalej to je identifikácia špičkových technológií ktoré bude možné využiť v nasledujúcich balónových letoch. Balón bude technicky zaštiťovať francúzska vesmírna agentúra CNES a prístrojové vybavenie je kolaboráciou dvoch francúzskych laboratórií a Akademie věd České republiky. Konkrétne Laboratoire aerologie v Toulouse a Laboratoire de Physique et de Chimie de lEnvironnement et de lEspace v Orleans a oddělení kosmické fyziky Ústavu fyziky atmosféry. STRATEOLE bude vybavený širokopásmovým prijímačom rádiových vĺn a detektorom gamma žiarenia XStorm. Komunikácia so Zemou bude sprostredkovaná pomocou satelitu Irídium. Celková dĺžka misie sa počíta na niekoľko mesiacov.

V nasledujúcej práci sme využívali dáta zo širokospektrálnych meracích prístrojov umiestnených na niekoľkých Európskych stanovištiach Ústavu fyziky atmosféry AV ČR. Tie využívajú tienené slučkové antény so vzorkovacou frekvenciou 200 MHz. Konkrétne sme využívali merania zo štyroch stanovíšť, Milešovka (ML), Dlouhá louka (DL), Lomnický Štít (LS) a Krupka (KR). Tieto dáta následne analyzujeme, a z nich nadobudnuté znalosti následne využijeme pre vytvorenie algoritmu využívajúceho strojové učenie, ktorý bude schopný rozpoznať prítomnosť a typ bleskového výboja na časti záznamu z antény.

1. Elektromagnetické vlny

Elektromagnetické vlny sú rovinné vlny tvorené osciláciou vektorov intenzít elektrického $\vec{E}$ a magnetického $\vec{H}$ poľa. Vo vákuu sa šíria rýchlosťou svetla. Charakterizovať sa dajú pomocou vlnovej dĺžky λ alebo pomocou frekvencie ν. Elektrická a magnetická intenzita sú navzájom prepojené a spolu s ďalšími dvoma veličinami a to s elektrickou $\vec{D}$ a magnetickou $\vec{B}$ indukciou. Rovnice ktoré tieto všetky štyri parametre prepájajú sa nazývajú Maxwellove rovnice (z knihy Bedřich Sedlák (2012)) (viď obrázok 1.1).



Obrázok. 1.1: Schematické znázornenie elektromagnetického vlnenia vo vákuu. Modrým je označený vektor magnetickej intenzity $\vec{B}$ a červeným je označený vektor elektrickej intenzity $\vec{E}$. Vektorom $\vec{v}$ je označený smer šírenia vlny, λ označuje vlnovú dĺžku vlnenia a A je amplitúda vlnenia.

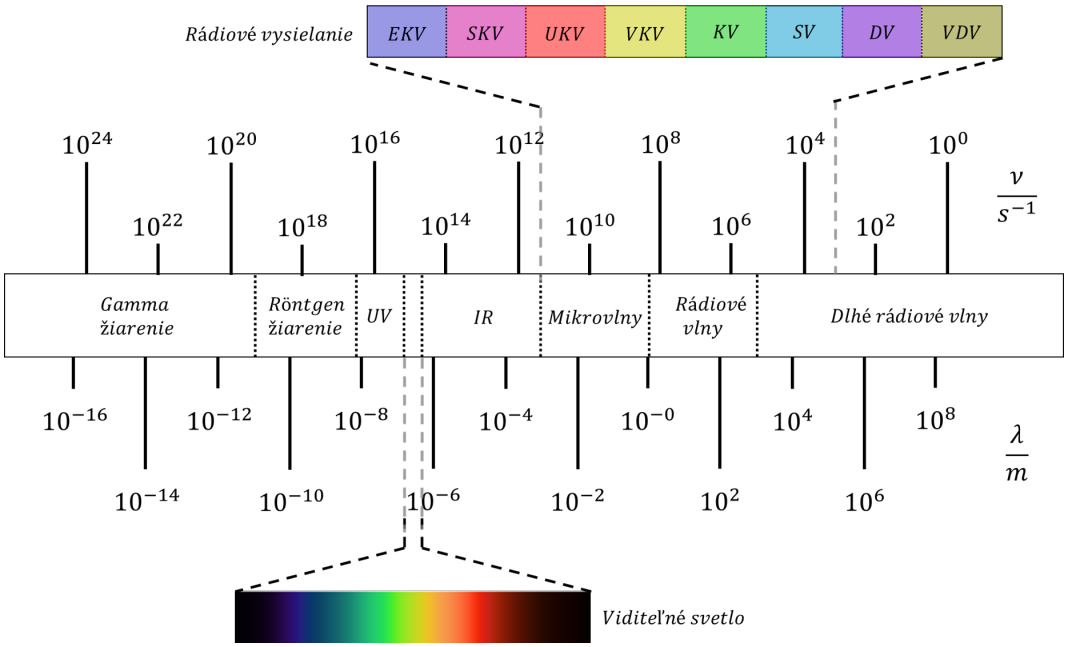
1.1 Popis

V modernej fyzike sa popisuje prenos elektromagnetického vlnenia po kvantách. Pričom energia takéhoto kvanta je priamo úmerná frekvencií daného vlnenia. A práve podľa tejto energie kvanta, a teda frekvencie, sa elektromagnetické vlnenie vydeľuje na jednotlivé triedy v celom spektre (podľa knihy Malý (2020)):

- Rádiové vlny
- Mikrovlny
- Infračervené žiarenie
- Viditeľné svetlo
- Ultrafialové žiarenie
- Röntgenové žiarenie

- Gamma žiarenie

Elektromagnetické vlnenie je schopné prenášať energiu a teda aj informáciu. V našom výskume sa budeme ďalej zaoberať hlavne rádiovými vlnami (viď obrázok 1.2).



Obrázok. 1.2: Znázornenie celého spektra elektromagnetického vlnenia s dôrazom na viditeľné svetlo a rádiové vlny. Kde EKV znamená extrémne krátke vlny; SKV super krátke vlny; UKV ultra krátke vlny; VKV veľmi krátke vlny; KV krátke vlny; SV stredné vlny; DV dlhé vlny; VDV veľmi dlhé vlny.

1.2 Zdroje

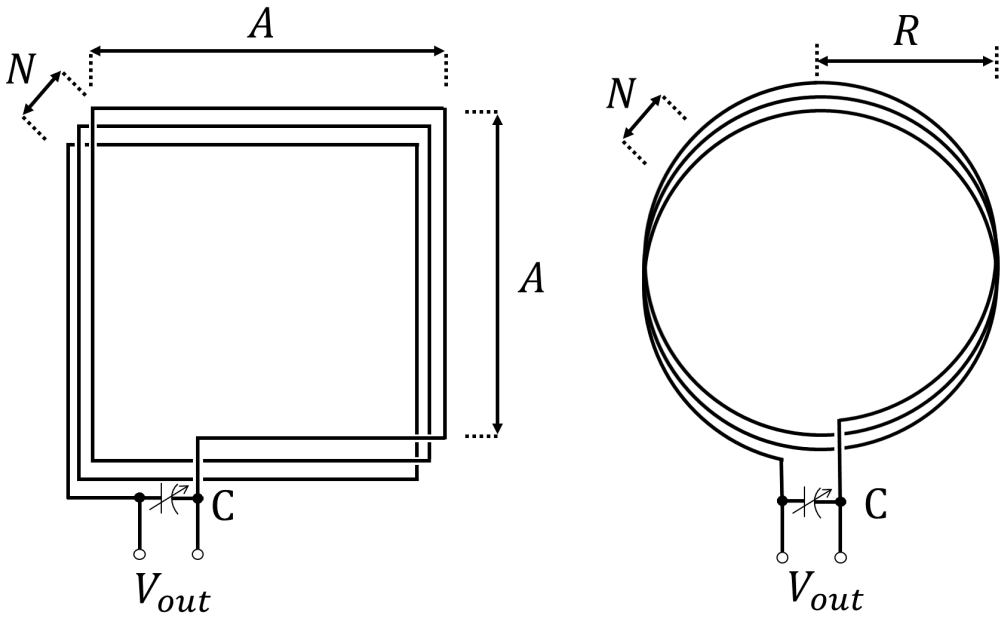
Základné rozdelenie ktoré môžeme prisúdiť zdrojom elektromagnetického žiarenia sú zdroje prirodzené a zdroje umelé. Medzi nimi neexistuje žiaden fundamentálny rozdiel a líšia sa len v intencií. Umelými zdrojmi sa nebudeme ďalej zaoberať. Prirodzenými zdrojmi sú každý náboj ktorý sa pohybuje so zrýchlením (podľa knihy Feynman a kol. (1965)). A preto je zdrojom rádiových vĺn aj bleskový výboj. Keďže ale spektrálna charakteristika, doba trvania a ďalšie vlastnosti vlnenia závisia od zdroja a každý bleskový výboj je charakterizovaný iným presunom náboja. Tak aj uvoľnené vlnenie nesie charakteristiku ktorá v sebe obsahuje vlastnosti výboja.

1.3 Spôsoby merania

Spôsob merania závisí od druhu elektromagnetického žiarenia. V našom prípade nás zaujímajú rádiové vlny a teda sa zameriame na to ako merať práve takéto vlnenie. Rádiové vlny sú všeobecne prijímané pomocou antén. Anténa je najčastejšie kovová konštrukcia v ktorej pri prechode elektromagnetickej vlny dochádza k oscilácií elektrónov. Táto oscilácia je následne zosilnená a zaznamenaná.

Najčastejšie sa stretávame s jedným z dvoch druhov antén a to s dipólovou (meranie elektrickej zložky žiarenia) alebo so slučkovou anténou (meranie magnetickej zložky žiarenia). V našom prípade sa používa práve slučková anténa.

Slučková anténa (podľa knihy League (1982)) pozostáva z vodiča ktorý je kontinuálne namotaný (viď obrázok 1.3). Výstup z antény je úmerný časovej derivácií magnetickej indukcie v smere kolmom k ploche antény. Výstup je samozrejme analógový a pre rozumné ďalšie nakladenie s ním je potrebná jeho digitalizácia. Tá sa vykonáva meraním napätia na výstupe z antény pomocou analógovo-digitálneho prevodníku (ADC) v osciloskope.



Obrázok. 1.3: Schéma slučkovej antény v dvoch najčastejších vyhotoveniach. Naľavo je štvorcová varianta, napravo je kruhová varianta. N je počet namotaní, A a R je šírka, resp. polomer antény. Anténa generuje napätie V_{out}.

2. Bleskové výboje, ich príčina, prejav a klasifikácia

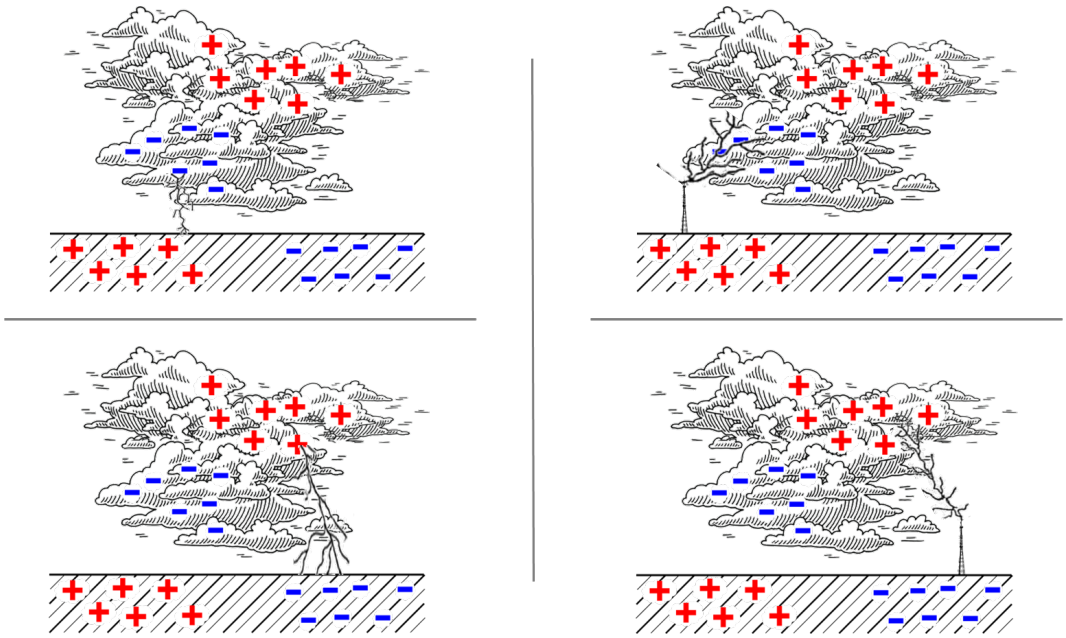
Typicky je elektrický výboj v atmosfére spôsobený nehomogenitou rozloženia náboja v búrkovom oblaku a následným rýchlym presunom tohto náboja za tvorby a v rámci ionizovaného plazmového kanálu. Tie sú sprevádzané emisiou rádiového signálu. Ten sa môže generovať v rozdielnej dobe elektrického výboja a vlastnosti tohto rádiového signálu môžu niesť informáciu o výboji ktorý ho vyvolal. Môžeme takto pozorovať blesky Oblak-Zem (Cloud-to-ground (CG)), Zem-Oblak (Ground-to-cloud (GC)), alebo blesk Oblak-Oblak (Intracloud (IC)). Takto emitovaný rádiový signál je možné detegovať na veľké vzdialenosti pomocou, nato prispôsobených, prijímacích antén.

2.1 Bleskový výboj

Blesk alebo správnejšie bleskový výboj je atmosférický jav. Ten sa vyskytuje v búrkových oblakoch. Tam je totiž najväčšia pravdepodobnosť nehomogenity rozloženia náboja v oblaku alebo v rozdiely náboja medzi oblakom a zemou. Samotný výboj je následne rýchly presun tohto náboja medzi oblasťami s rozdielnym elektrickým nábojom. Tento presun ma niekoľko fáz. Takýto bleskový výboj trvá zlomok sekundy a je ľuďmi pozorovateľný. Najzreteľnejší je pruh svetla (laický označovaný ako blesk) a v krátkom slede za tým nasleduje zreteľný zvuk charakteru výbuchu (laicky označovaný ako hrom). Samotným pruhom svetla ktorý pozorujeme je tzv. bleskový kanál, čo je súvislá krivka tvorená plazmou ktorá ako veľmi dobrý vodič umožňuje výmenu elektrického náboja (viď obrázok 2.1). Bleskový výboj ako rýchly presun náboja v plazmovom kanály sa prejavuje typický jasným svetelným zábleskom. Pričom prúdová vlna sa pohybuje rýchlosťou tretiny až polovice rýchlosti svetla. Výbojový prúd blesku môže dosahovať až stovky kiloampérov a teplota bleskového kanálu sa môže vyšplhať až na 30 000 °C. To ale nie je jediný prejav. Takýto rýchly presun elektrického náboja taktiež generuje žiarenie v oblasti rádiových vĺn.

2.2 Pôvod vzniku blesku

Primárna otázka je čo spôsobuje nerovnomerné rozloženie náboja v búrkovom oblaku (ďalej len oblaku) alebo v oblaku voči zemi. Toto nerovnomerné rozloženie nastane z počiatočného homogénneho stavu presunom náboja alebo jeho nehomogénnou tvorbou v rámci oblaku. Táto tvorba a presun sa v súčasnosti vysvetľujú dvojicou mechanizmov (Rakov a Uman (2007)). Prvým je mechanizmus konvekcie. Hlavným princípom je, že nadbytočný náboj je do oblaku dodávaný externými zdrojmi ako je napríklad radón, nachádzajúci sa v pôde alebo blízko nad zemou, alebo kozmické žiarenie nad oblakom. Následne je vzostupným prúdením teplého vzduchu unášaný pozitívny náboj od zeme na vrchol oblaku (viď článok Williams a Stanfill (2002)). Tento prenos je zabezpečovaný rôznorodými časticami ktoré sa v oblaku nachádzajú, a ktoré nazývame hydrometeory (viď článok Mansell a kol.



Obrázok. 2.1: Znázornenie štyroch základných typov výboja. (ľavý horný) CG záporný výboj. (pravý horný) GC záporný výboj. (ľavý dolný) CG pozitívny výboj. (pravý dolný) GC pozitívny výboj.

(2005) a Liu a Chandrasekar (2000)). Medzi ne patria vodné kvapky, prechladené vodné kvapky, snehové a ľadové krupky ako aj krúpy. Všetky tieto hydrometeory sa nabíjajú nábojom mechanizmami spomenutými vyššie. Ťažšie hydrometeory padajú zatiaľ čo ľahšie sú vynášané vyššie. Práve ľahšie sa nabíjajú kladne a tento svoj náboj vynášajú do vyšších vrstiev. Zatiaľčo tie ťažšie, záporne nabité, klesajú. Druhým je mechanizmus "Graupel-ice"(viď článok Ziegler a kol. (1991) a Dye a kol. (1986)). V tomto prípade je zdroj náboja priamo v oblaku a predáva sa zrážkami ľadových kryštálikov a krúp. Pri týchto zrážkach dochádza k separácií nábojov a ich prenosu. Pri nižších teplotách sú ťažšie krúpy, padajúce dole, nabíjané záporne a ľadové kryštáliky, stúpajúce hore, nabíjané kladne. Pri vyšších teplotách sú polarity nábojov vymenené. Skupiny hydrometeorov nesúci rovnaký náboj sa týmto mechanizmom zoskupia do nábojových vrstiev. Pokiaľ elektrické pole medzi nábojovými vrstvami dosiahne určitú hodnotu, tak sa naštartuje proces vzniku bleskového výboja, ktorý v časti prípadov končí ako obrovský výboj medzi oblakom a zemou (viď článok Kostinskiy a kol. (2020)).

2.3 Časový vývoj blesku

Typická sekvencia vývoja negatívneho Oblak-Zem bleskového výboja (CG-) tak ako je popísaná v knihe Rakov a Uman (2007) je nasledujúca (viď obrázok 2.2).

Celý proces začína, typicky v búrkovom oblaku, kde sa vytvoria tri vrstvy náboja. Na vrchu je obvykle oblasť kladného náboja. Uprostred je oblasť záporného náboja. A na spodnom okraji je oblasť slabého pozitívneho náboja.

Nasleduje počiatočný elektrický prieraz. Jednotný názor na presný priebeh tejto etapy nie je, ale všeobecne sa domnieva, že môže ísť o výboj, ktorý premos-

ťuje hlavné oblasti záporného a nižšieho kladného náboja. Typický trvá desiatky mikrosekúnd. Predstavuje hlavného strojcu podmienok pre vytvorenie stupňovitého vedúceho výboja (viď článok Kolmašová a kol. (2014), Wu a kol. (2016) a Liu a kol. (2022)).

Stupňovitý vedúci výboj je negatívne nabitá plazma ktorá sa posúva k zemskému povrchu v sérií diskrétnych krokov. Stupňovitý vedúci výboj slúži na vytvorenie vodivého plazmového kanálu medzi nábojom v oblaku a zemou. Typické trvanie sú jednotky až desiatky milisekúnd.

V náväznosti na priblíženie stupňovitého vedúceho výboja k zemskému povrchu sa začne tvoriť do vrchu mieriaci výboj začínajúci na zemskom povrchu (alebo vodivom objekte spojeného so zemou). Jeho tvorba a približovanie k stupňovitému vedúcemu výboju sa označuje ako proces spájania.

Finálnym spojením dole a hore smerujúcich výbojov je spätný výboj. Ten cestuje vertikálne smerom nahor po záporne nabitej vodivej línií. Pričom rýchlosť tohto pohybu môže dosiahnuť až polovicu rýchlosti svetla. Zaujímavá je ale závislosť, že s rastúcou výškou rýchlosť klesá. Po ukončení spätného výboja môže celý vývoj blesku skončiť. A takýto blesk následne voláme jednovýbojový (single-stroke).

V skutočnosti ale častejšie nastane, že po konci spätného výboja sa uskutočnia vnútro-oblakové prerovnania náboja po ktorých nasleduje spojitý vedúci výboj výboj a eventuálne ďalšie spätné výboje.

2.4 Klasifikácia etáp bleskového výboja

Pre potreby možnej identifikácie pomocou postupov strojového učenia bolo potrebné jednotlivé výboje alebo ich skupiny a ich vývojové časti rozdeliť do kategórií (viď článok Zhu a kol. (2021) a Lu a kol. (2012)), ktoré majú v časovo frekvenčných spektrogramoch rozdielne charakteristiky. Tieto charakteristiky vo vyžarovaní elektromagnetického vlnenia sú konkrétne napísané nižšie kde sa následne budeme venovať ich podrobnejšiemu popisu. Časť popisov vychádza z knihy Cooray (2016).

- Iniciačná fáza CG (Breakdown pulse CG)
- Iniciačná fáza IC (Breakdown pulse IC)
- Skupina mikrosekudnových pulzov (Microsecond pulses)
- Úzka bipolárna udalosť (Narrow bipolar event (NBE))
- IC aktivita (IC activity)
- Spätný výboj (Return stroke)

2.4.1 Iniciačná fáza CG

Na záznamoch sa Iniciačná fáza CG prejavuje ako pulzy od seba vzdialené desiatky až malé stovky mikrosekúnd (viď článok Wu a kol. (2013), Kolmašová a kol. (2020) a Kolmašová a kol. (2022)). Sú krátke, v trvaní desiatky mikrosekúnd široké. Vyskytujú sa často v skupinách s pravidelným rozostupom. Skupina zvykne byť dlhá okolo jednej milisekundy (viď obrázok 2.3).

2.4.2 Iniciačná fáza IC

Iniciačná fáza IC sa prejavuje na záznamoch ako malé, desiatky mikrosekúnd široké pulzy, od seba vzdialené stovky mikrosekúnd. Najčastejšie sú nepravidelne usporiadané. Skupiny môžu byť aj niekoľko milisekúnd dlhé (viď článok Wu a kol. (2015) a obrázok 2.4).

2.4.3 Skupina mikrosekundových pulzov

Na zázname bývajú pulzy mikrosekundy široké, od seba vzdialené niekoľko mikrosekúnd. Často sú veľmi pravidelné rozostupy medzi nimi v skupinke a skupinky sú až stovky mikrosekúnd dlhé (viď článok Kolmašová a Santolík (2013) a Shi a kol. (2020) a obrázok 2.5).

2.4.4 Úzka bipolárna udalosť

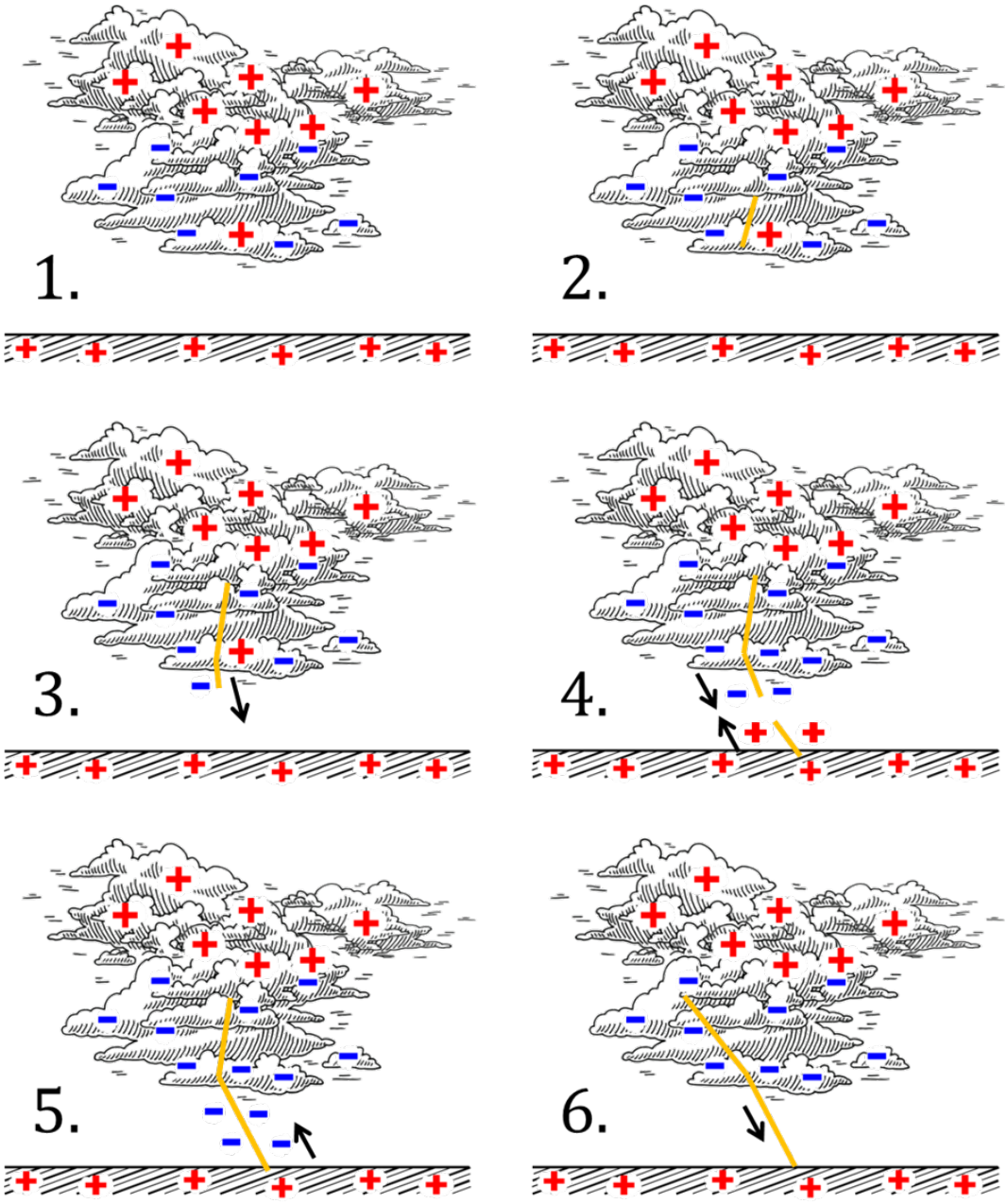
Na zázname sa prejavuje ako úzky bipolárny pulz (viď článok Nag a Rakov (2010)), často osamotený a mal by byť dobre rozlíšiteľný v dátach tým, že najviac zo všetkých žiari na vysokých frekvenciách, kde bude zaberať celé pásmo (viď obrázok 2.6).

2.4.5 IC aktivita

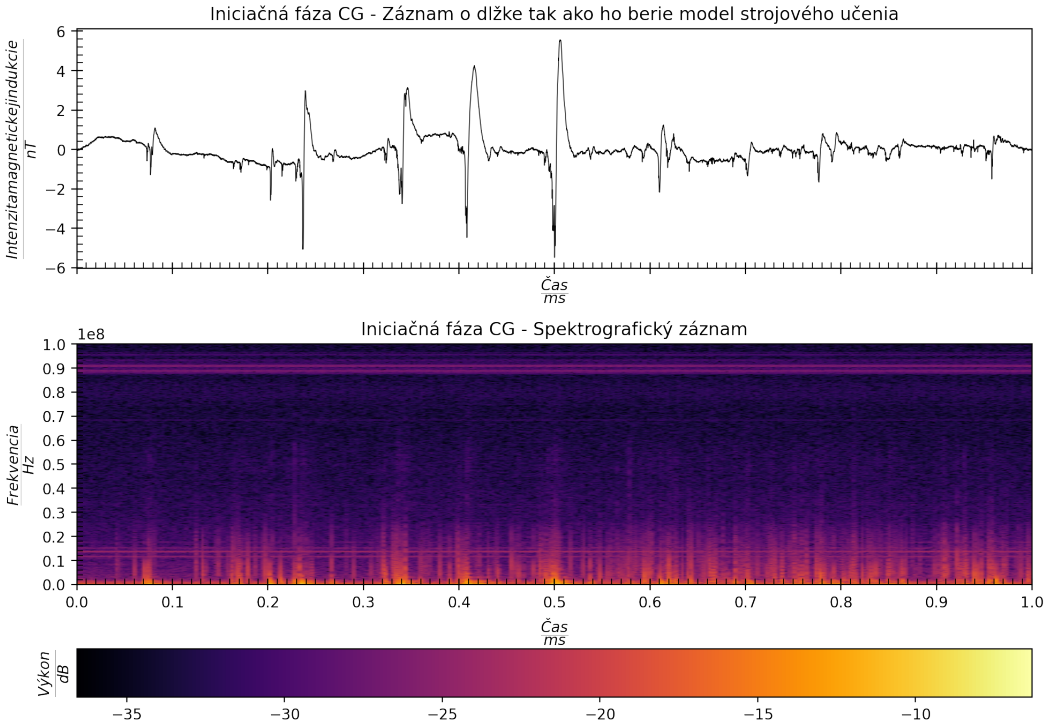
Prejavom na záznamoch sú bipolárne pulzy ktoré nie sú tak široké ako pri spätnom pulze, ale ani nie tak úzke ako pri úzkych bipolárnych udalostiach (NBE). Len málokedy sa vyskytujú ojedinele, ale býva ich iba zopár do milisekundy (viď obrázok 2.7).

2.4.6 Spätný výboj

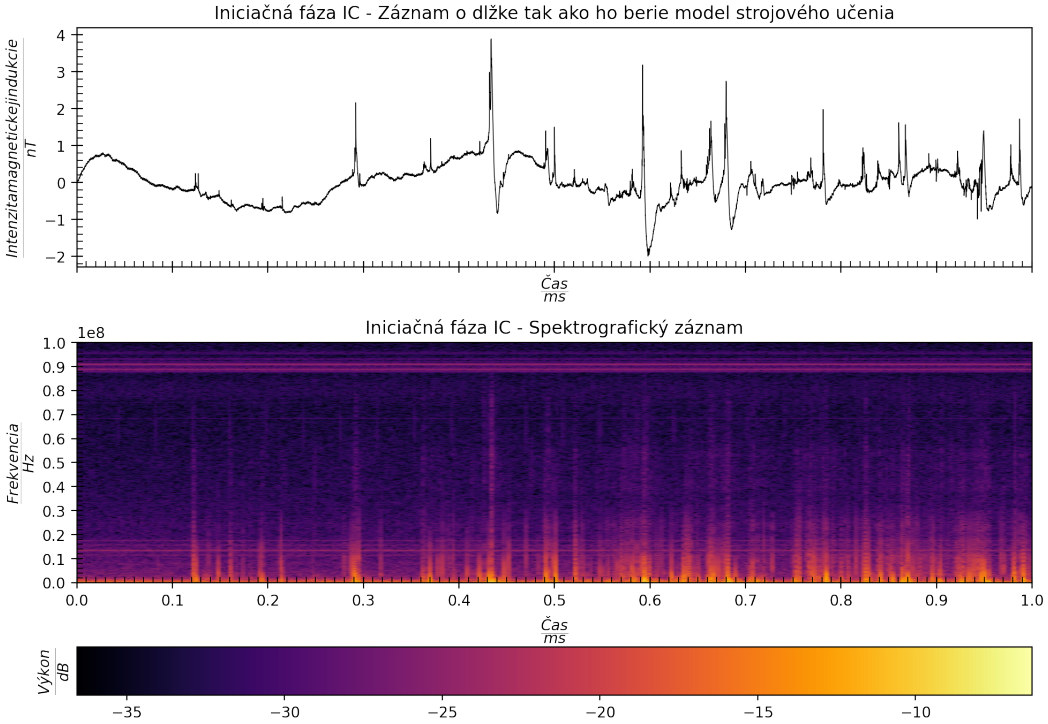
Na záznamoch sa prejavuje ako široký pulz ktorý začína prudkým nástupom do hlavného peaku a následne prestrelením do opačnej polarity (viď článok Nag a Rakov (2012) a Kašpar a kol. (2022)). Toto prestrelenie zvyčajne dosahuje iba polovičnú intenzitu oproti hlavnému peaku. Po prestrelení nasleduje pozvoľnejší, široký návrat do nuly. Pulz býva okolo 100 mikrosekúnd (μs) široký a najviac energie vyžiari na desiatkach kilohertzoch (viď obrázok 2.8).



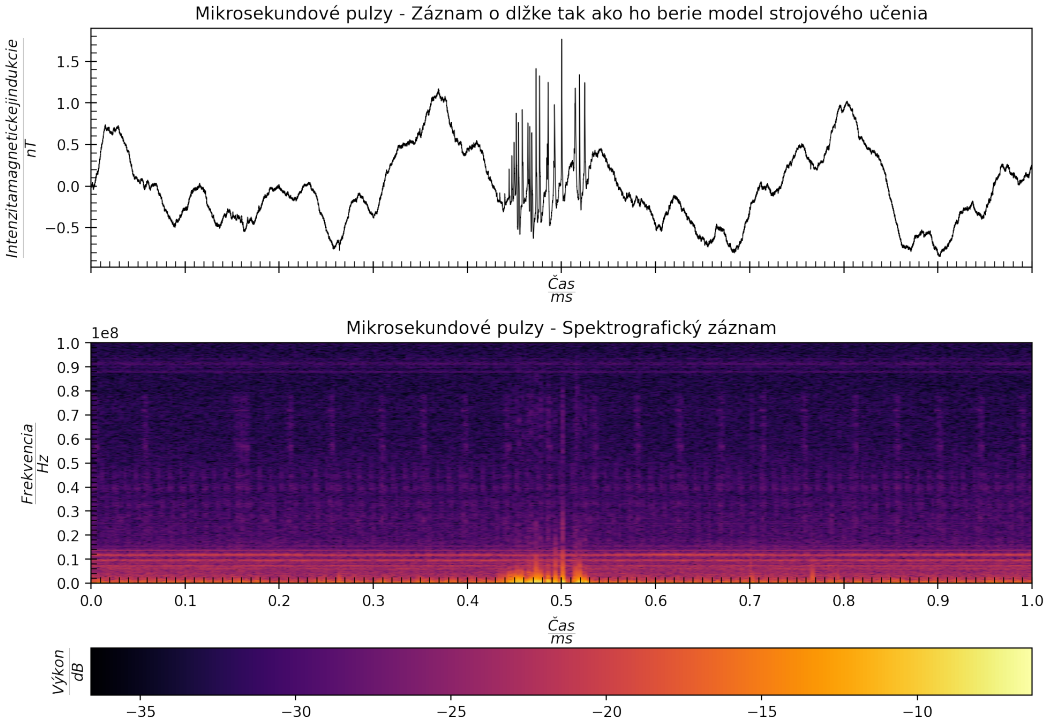
Obrázok. 2.2: Grafické znázornenie vývoja (CG-) blesku ako je popísaný v knihe Rakov a Uman (2007). Kde jednotlivé očíslovné časti predstavujú časový vývoj. 1. Znázorňuje rozloženie náboja v oblaku pred výbojom. 2. Znázorňuje počiatočný elektrický prieraz. 3. Znázorňuje skokový vedúci výboj. 4. Znázorňuje proces spájania. 5. Znázorňuje prvý spätný výboj. 6. Znázorňuje spojitý vedúci výboj , ktorý nasleduje po procese prerozdelenia náboja vo vnútri búrkového oblaku a po ktorom môže nasledovať druhý spätný výboj.



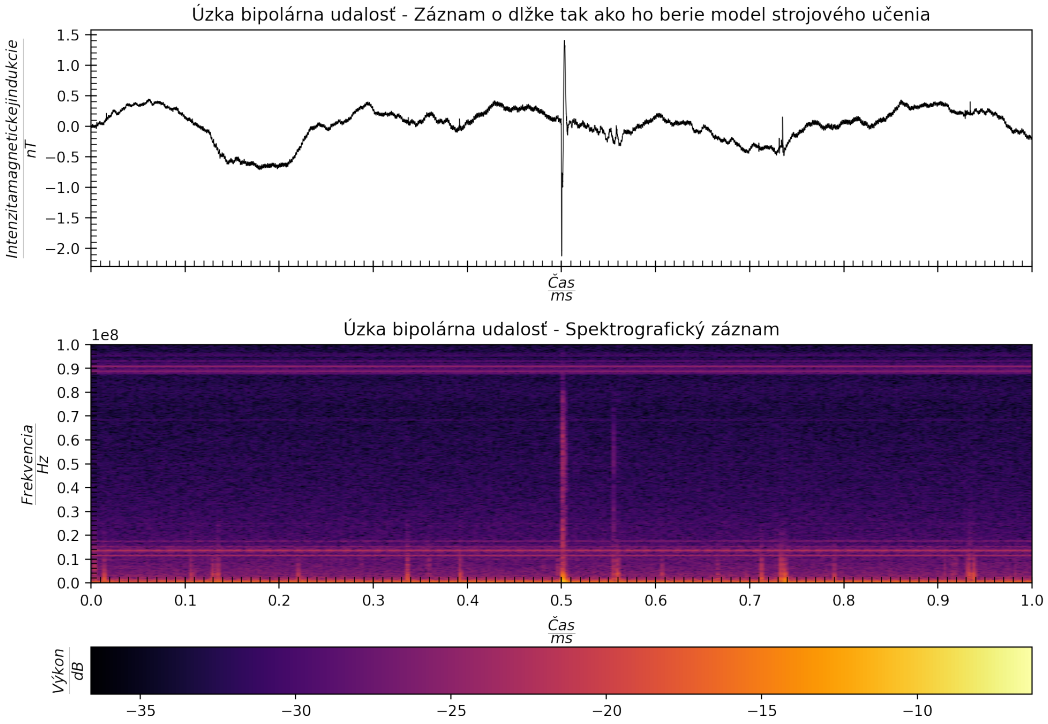
Obrázok. 2.3: Iniciačná fáza invertovaného IC blesku (rovnaká kategória ako Iniciačná fáza CG): Ukážka integrovanej vlnovej formy a prislúchajúceho spektrogramu. Pre stanicu ML, dátumu 02.07.2020 o 10.20.04.



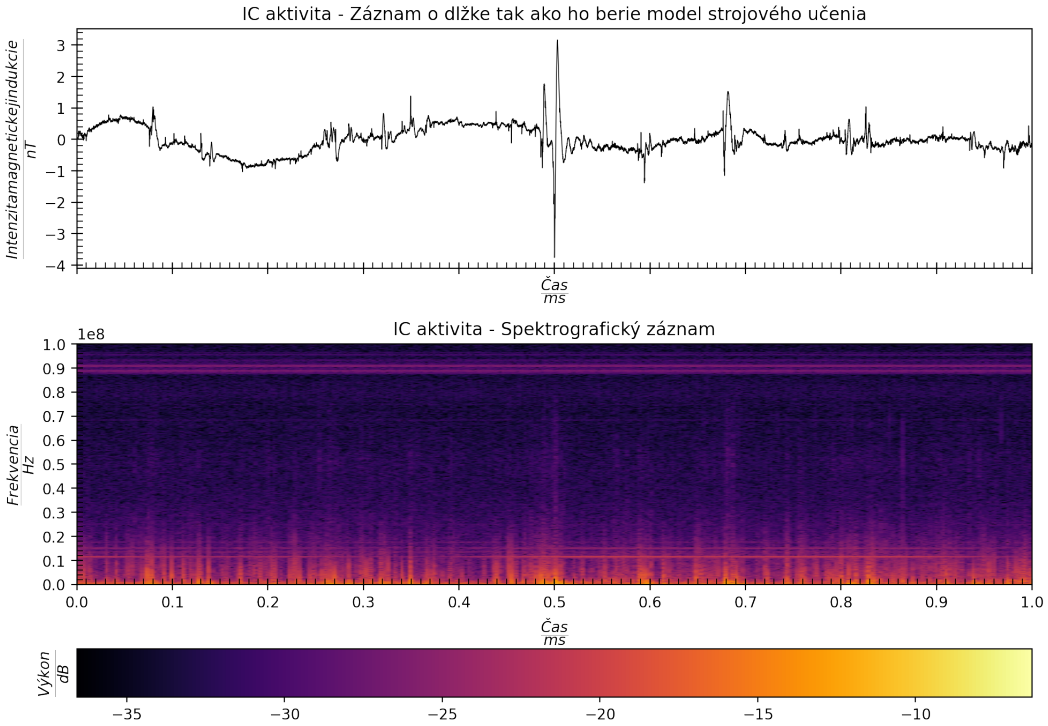
Obrázok. 2.4: Iniciačná fáza IC: Ukážka integrovanej vlnovej formy a prislúchajúceho spektrogramu. Pre stanicu ML, dátumu 02.07.2020 o 09.36.14.



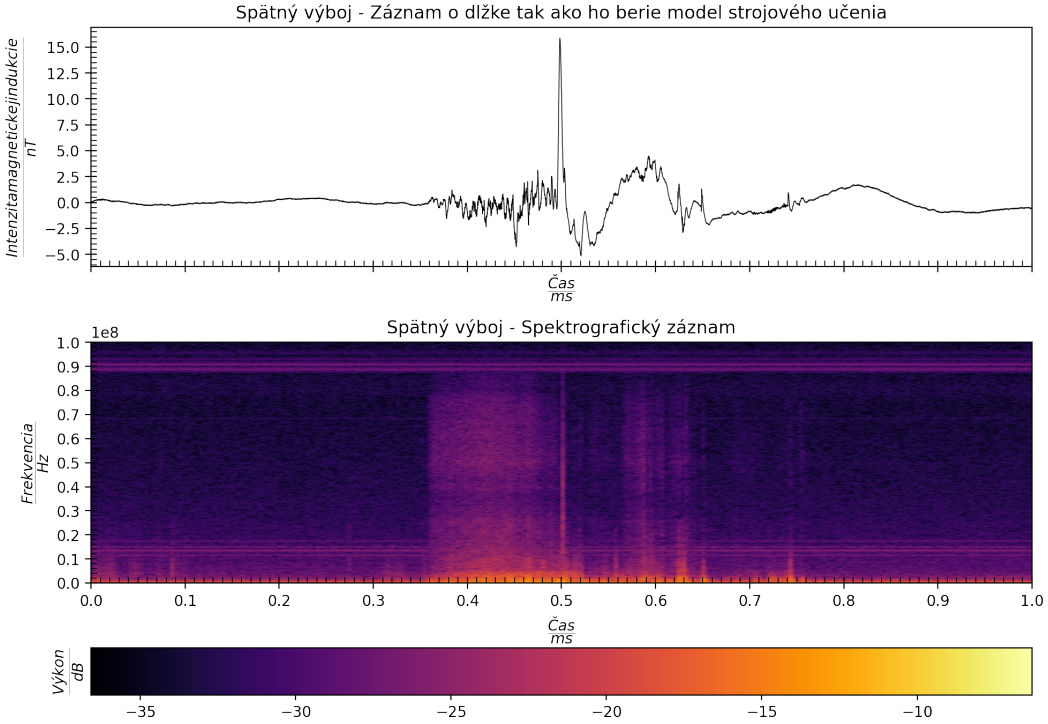
Obrázok. 2.5: Skupina mikrosekudnových pulzov: Ukážka integrovanej vlnovej formy a prislúchajúceho spektrogramu. Dátumu 18.06.2021 o 18.58.46.



Obrázok. 2.6: Úzka bipolárna udalosť: Ukážka integrovanej vlnovej formy a prislúchajúceho spektrogramu. Pre stanicu ML, dátumu 10.08.2020 o 10.51.03.



Obrázok. 2.7: IC aktivita: Ukážka integrovanej vlnovej formy a prislúchajúceho spektrogramu. Pre stanicu ML, dátumu 02.07.2020 o 10.32.57.



Obrázok. 2.8: Spätný výboj (s výrazným spojitým vedúcim výbojom pred vrcholom spätného výboja): Ukážka integrovanej vlnovej formy a prislúchajúceho spektrogramu. Pre stanicu ML, dátumu 10.08.2020 o 11.01.07.

3. Strojové učenie

3.1 Úvod

Strojové učenie (v angl. Machine learning) je podmnožina oboru zvaného umelá inteligencia (v angl. Artificial inteligence). Strojové učenie sa zaoberá možnosťou tvorenia algoritmov ktoré vykazujú schopnosť regresie, klasifikácie, segmentácie, grupovania, predikcie, asociovania a zmenšovania dimenzie problémov na základe učenia. Pod učením sa rozumie vystavenie algoritmu už známym dátam, teda dátam o ktorých náš algoritmus vie ich príslušnosť ku kategórií, oblasť záujmu, asociáciu alebo predikciu. Neplatí to ale všeobecne. Rozlišujeme totiž ešte podkategórie strojového učenia ktoré sú schopné pracovať aj bez učenia sa na známych dátach. Nebudeme sa im ale bližšie venovať pretože niesu potrebné pre našu prácu. Priradenie prislúchajúcej informácie k dátam sa väčšinou nazýva označovanie (v angl. labeling). Možnosť použitia strojového učenia pre podobný typ problémov už bol ukázaný v článku Maslej-Krešňáková a kol. (2021).

3.2 Klasické strojové učenie

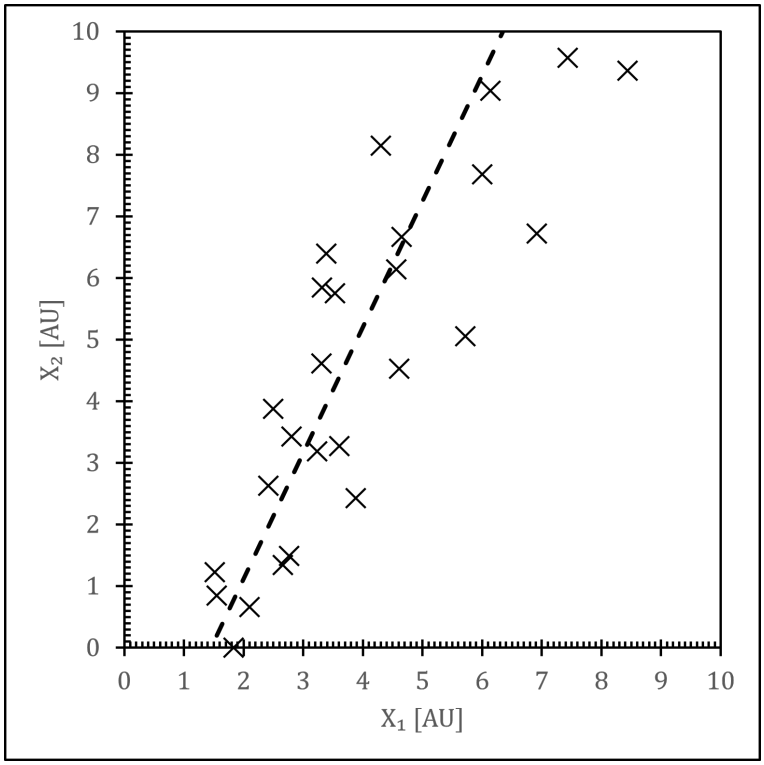
Časovo najstaršia podmnožina strojového učenia sa označuje ako klasické strojové učenie. Pod týmto súhrnným označením sa nachádza veľké množstvo rôznych prístupov k tomu akým spôsobom pracovať so známymi dátami a teda akým spôsobom učiť algoritmus. Pri určitých zjednodušeniach sa dá povedať, že sa klasicky strojové učenie dá deliť podľa troch faktorov. Tými sú prítomnosť učiteľa, oblasť využitia a napokon použitý algoritmus. Učenie má za úlohu nastaviť alebo teda nájsť ideálne parametre modelu ktorý následne bude najlepšie vystihovať merané dáta. Popri názve parametre modelu sa stretneme aj s tzv. hyperparametrami modelu. Čo sú čísla charakterizujúce architektúru alebo stavbu modelu, nie jeho naučené znalosti.

3.2.1 Lineárna regresia

Pod pojmom lineárne regresné modely sa rozumejú modely ktoré vyjadrujú závislosť závislej premennej od nezávislých premenných iba pomocou lineárnych funkcií. Tie sú charakterizované priamkami, resp. nadrovinami v mnohorozmerných priestoroch nezávislých parametrov (viď obrázok 3.1). V základe ide o veľmi jednoduchú ale do veľkej miery obmedzenú metódu. Výhodami je možnosť rýchleho učenia aj na veľkých dátových množinách (nad milióny vstupných dát). Nevýhodou ale je fundamentálna nemožnosť zachytenia nelineárnych procesov. Matematicky sa dá zapísať ako:

$$Y = a_0 + a_1 \cdot X + \epsilon. \tag{3.1}$$

Kde Y je závislá premenná (všeobecne vektor), a_0, a_1 sú koeficienty modelu, X sú vstupné nezávislé parametre a ϵ je náhodný šum pridaný do modelu. Dôvod pridania šumu a jeho nastavenie sú otázky ktorým sa budeme venovať ešte v časti venovanej učeniu modelov strojového učenia.



Obrázok. 3.1: Graf lineárnej regresie pre názorný príklad pre priemet do plochy dvoch parametrov X_1 a X_2.

3.2.2 Polynomiálna regresia

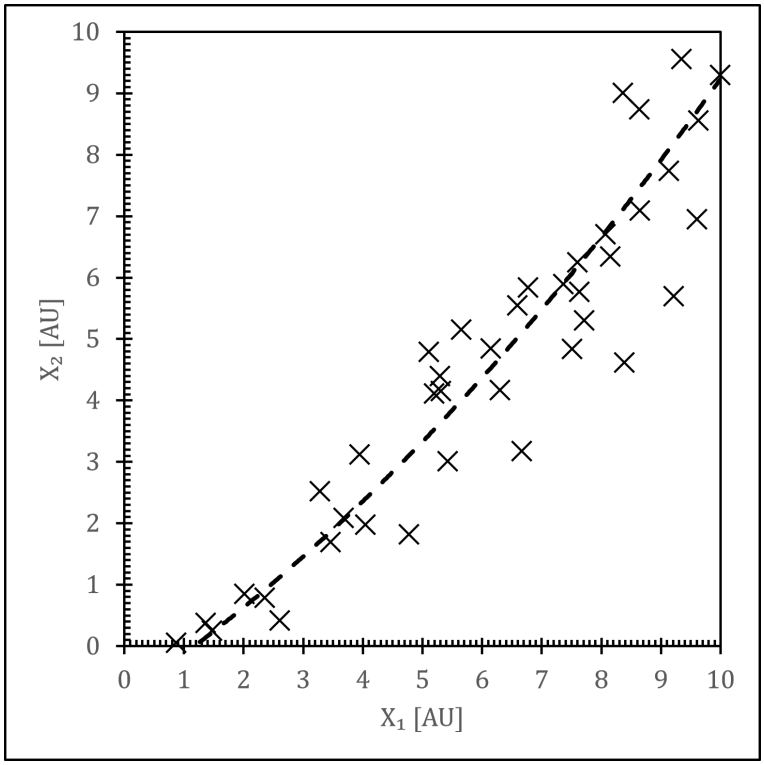
V nadväznosti na lineárnu regresiu a z tlaku na zachytenie nelineárnych javov sa prechádza od lineárnej k všeobecne polynomiálnej regresií. Tá modeluje závislosť závislej premennej od vstupných nezávislých ako polynóm stupňa n (viď obrázok 3.2). Takže výsledný modelom je polynomická krivka, resp. plocha. Výhodami je v dnešnej dobe takmer rovnako rýchla doba učenia ako pri lineárnych modeloch ako aj väčšie pole uplatnenia. Ako nevýhodou sa ale vynára určitá náchylnosť takýchto algoritmov preceňovať extrémne hodnoty v dátach ako aj tendencia pri vyšších n k presnému "naučeniu"vstupných dát. Namiesto žiadanejšieho vystihnutia trendu, jav známy ako over-fitting.

$$Y = a_0 + a_1 \cdot X^1 + a_2 \cdot X^2 + ... + a_n \cdot X^n + \epsilon. \tag{3.2}$$

Kde opäť Y je závislá premenná, $a_0, a_1, ..., a_n$ sú koeficienty modelu, X sú vstupné nezávislé parametre a ϵ je náhodný šum pridaný do modelu.

3.2.3 K-najbližších susedov

Samozrejme, v oblastiach kedy sa potýkame s nelineárnymi javmi a kde vstupné dáta majú tendenciu k široko rozptýleným hodnotám sa polynomiálna regresia nedá rozumne použiť. Častejšie sa siaha po iných prístupoch k riešeniu. Jedným z takých je aj metóda najbližších susedov, tiež známa pod anglickým označením K-nearest neighbours (KNN) (viď článok Fix a Hodges (1989)). Základným predpokladom je, že nemám žiadne parametre modelu ktoré nastavujem učením a



Obrázok. 3.2: Graf polynomiálnej regresie stupňa $n = 2$ pre názorný príklad pre priemet do plochy dvoch parametrov X_1 a X_2.

model je závislý iba od hyperparametrov. A testovací dátový vektor získa príslušnosť podľa príslušnosti najbližších susedov v priestore nezávislých premenných pri zadanej priestorovej metrike. Názorný príklad je zobrazený na obrázok 3.3. Matematicky sa tento model dá popísať aj ako:

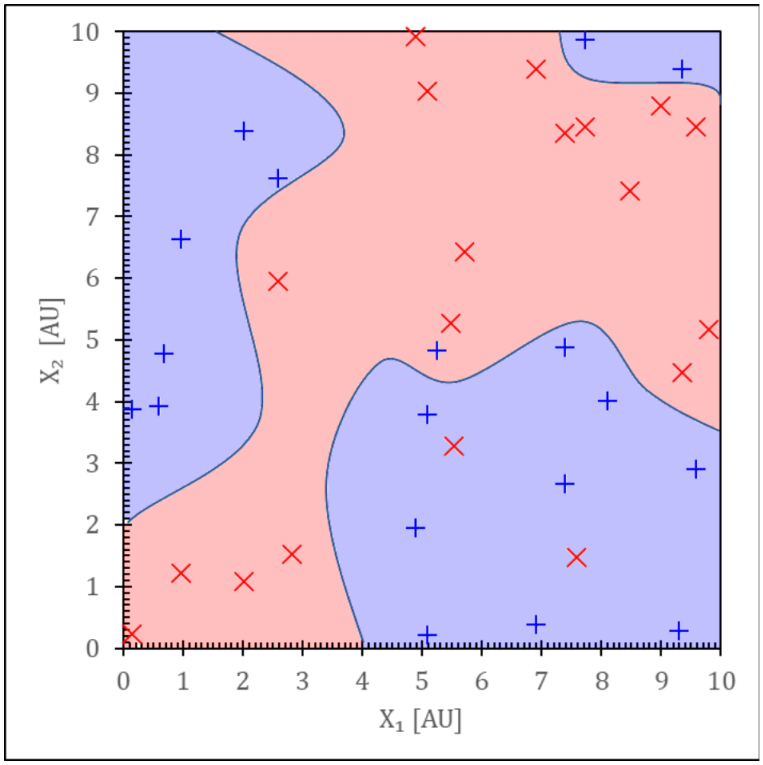
$$P(Y = y|X = x) = \frac{1}{K} \cdot \sum_{i \in A} I(Y^{(i)} = y). \tag{3.3}$$

Kde P je pravdepodobnosť príslušnosti k danej skupine, K je hyper-parameter modelu označovaný ako počet susedov. Výhodou tohto modelu je, že pri správnom nastavení hyperparametrov je schopný vystihnúť aj naozaj zložité zákonitosti. Problémom ale je, že s narastajúcim počtom vstupných dát značne narastá aj výpočetná náročnosť. Preto sa tento model v praxi používa iba do určitého nižšieho počtu vstupných dát.

3.2.4 Support vector machine (SVM)

Určitým naviazaním na lineárnu a polynomiálnu regresiu je metóda pomocných vektorov (viď článok Boser a kol. (1992)). Alebo známejšia pod svojím anglickým označením a skratkou ako Support vector machine (SVM). Hlavnou myšlienkou je nájdenie takej nadroviny (alebo eventuálne ľubovoľnej plochy) v priestore voľných parametrov ktorá najlepšie oddeľuje rozdielne triedy. takúto nadrovinu v lineárnom prípade matematicky popíšeme ako:

$$a_0 + a_1 \cdot X = 0. \tag{3.4}$$



Obrázok. 3.3: Graf aplikácie modelu K-najbližších susedov pre názorný príklad pre priemet do plochy dvoch parametrov X_1 a X_2. Modré a červené body patria do dvoch separátnych tried a farebná plocha reprezentuje oblasť každej danej triedy.

Takéto rozdelenie je najčastejšie myslené ako nadrovina od ktorej je najbližší bod každej kategórie čo najďalej (separabilný príklad) alebo do určitej vzdialenosti je bodov danej kategórie najmenej (neseprabilný prípad) (viď obrázok 3.4).

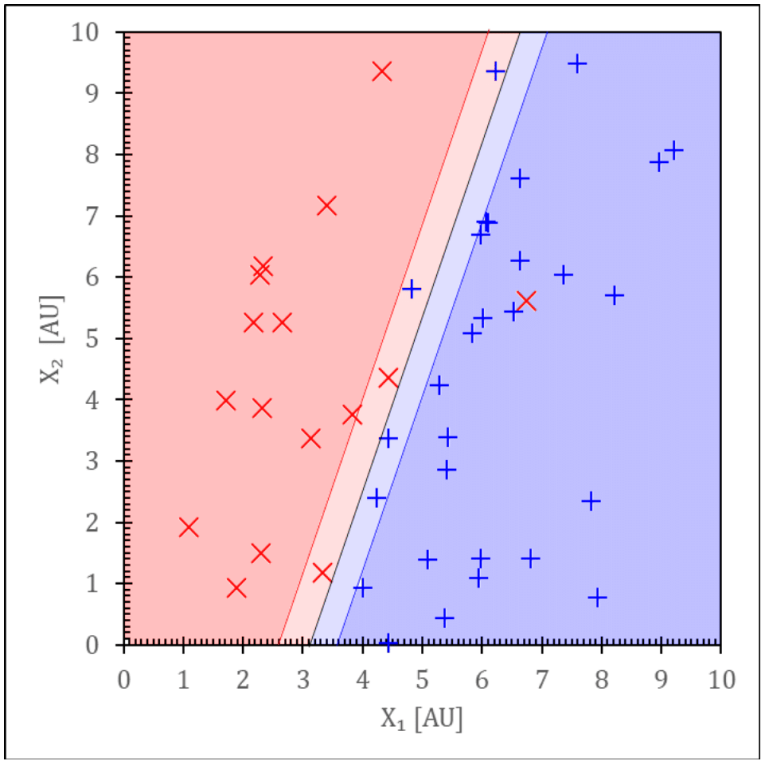
3.2.5 Rozhodovací strom

Opäť iným prístupom je skupina modelov ktoré sa označujú súhrne ako rozhodovacie stromy (Decision tree v angličtine). Základným princípom je rozdelenie analýzy na strom binárnych volieb (viď graf 3.5). Pričom učenie je nastavenie rozhodovacích kritérií tak aby sme čo najlepším spôsobom rozdelili v priestore nezávislých premenných (viď obrázok 3.7).

Veľkou výhodou je jednoduchosť a jasná interpretovateľnosť výsledkov. Ale prichádza to aj s určitým počtom nevýhod. Jedna z nich je, že pri väčších modeloch dochádza k prílišnej granulácií, teda k overfittingu a k celkovej nerobustnosti (tá sa dá pochopiť ako nechuť modelu k odhaleniu skrytých zákonitostí v dátach).

3.2.6 Random forest

V priamej náväznosti na rozhodovacie stromy a s nutnosťou vysporiadať sa s ich obmedzeniami sa objavila skupina modelov označovaných ako náhodný les (viď článok Breiman (2001)). Alebo známejšie ako Random forest z anglického jazyka. Myšlienka je jednoduchá, namiesto vytváranie jedného veľkého rozhodovacieho stromu sa vytvorí veľké množstvo (v praxi to sú aj desiatky tisíc) menších

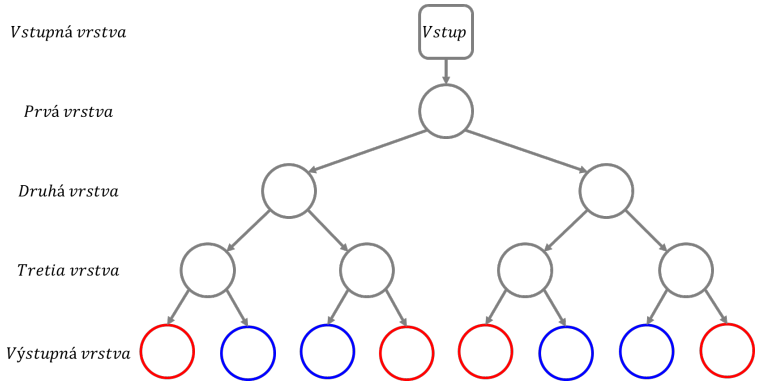


Obrázok. 3.4: Graf aplikácie modelu podporných vektorov susedov pre názorný príklad pre priemet do plochy dvoch parametrov X_1 a X_2. Modré a červené body patria do dvoch separátnych tried a farebná plocha reprezentuje oblasť každej danej triedy a odtieň určuje mieru istoty modelu.

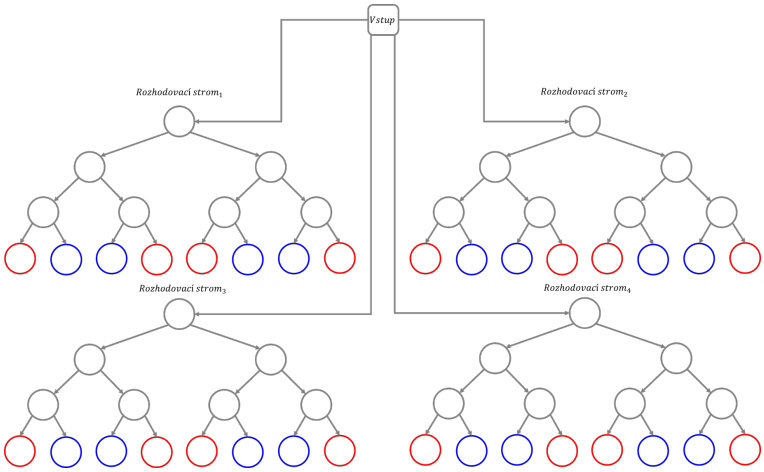
samostatných rozhodovacích stromov (viď obrázok 3.6) i keď vo výsledku graf rozdelenia vyzerá obdobne(viď obrázok 3.7). Každý z nich začína učenie na trochu inom subsete trénovacích dát (o tom ako sa vstupné dáta rozdeľujú sa viac rozpíšeme v kapitole o príprave dát 5.4) preto každý z jednotlivých stromov skončí s iným nastavením parametrov. Následne sa údaj ktorý chceme klasifikovať ukáže všetkým stromom a ako výsledok sa zoberie najčastejšie hodnota ktorú jednotlivé stromy vracali. Väčšina vlastností rozhodovacích stromov platí aj tu ale predtým jasná interpretovateľnosť je teraz ťažšia ale na druhú stranu random forest modely bývajú robustnejšie.

3.2.7 XGBoost

Random forest modely sa ukazovali ako jedny z najefektívnejších z klasických modelov strojového učenia. Preto sa aj pokračovalo v ich postupnej evolúcií. Jedným z výsledkov tejto snahy je skupina modelov označovaných ako gradient-boosted decision tree (GBDT). XGBoost je práve jeden z nich (viď článok Chen a kol. (2015)) Princíp je obdobný ako pri random forest ale pracuje sa na určité "fázy". Najprv dochádza k vytvoreniu základného random forest modelu ktorý sa natrénuje a otestuje. Dáta s ktorými má takýto model problém sa vyčlenia a vytvorí sa druhý (často krát ale už menší) model ktorý sa následne učí už iba na týchto vyčlenených dátach. To sa môže opakovať. Následne sa výsledný model vytvorí ako vyskladanie série takýchto modelov. Výsledný reprezentujúci graf sa



Obrázok. 3.5: Schéma názorného rozhodovacieho stromu hĺbky $n = 3$, ktorý zatrieďuje do dvoch tried (modrá, červená). (Zdroj: vlastná ilustrácia).



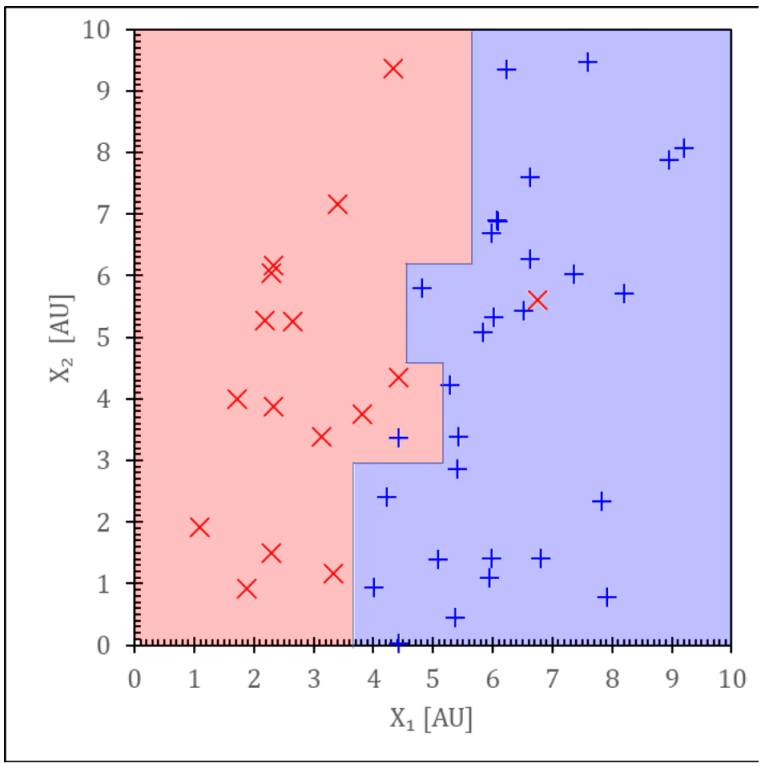
Obrázok. 3.6: Schéma názorného náhodného lesu s počtom stromov $n = 4$, ktorý zatrieďuje do dvoch tried (modrá, červená).

stále ale podobá práve rozhodovacím stromom (viď obrázok 3.7).

V súčasnej dobe sa XGBoost označuje ako jeden z najdokonalejších modelov z oblasti klasického strojového učenia. A v praxi už niekoľko rokov predstavuje podstatnú časť komplexných machine learning algoritmov. Výhodami je hlavne relatívne vysoká rýchlosť učenia aj na naozaj veľkých vstupných datasetoch a pritom vysoká účinnosť na širokej škále vstupných dát. Práve preto sa XGBoost v súčasnej dobe dá implementovať už na širokú škálu zariadení ako aj na FPGA zariadenia. Jenodu jasnou nevýhodou ale je tak ako pri samostatných random forest modeloch, že už je takmer nemožné reprezentovať význam parametrov v modeli.

3.3 Hlboké strojové učenie

Existuje ešte jedna značne veľká podskupina strojového učenia ktorá sa v poslednej dobe začína už vyčleňovať ako samostatná kategória a to je tzv. hlboké strojové učenie (v angl. deep learning). Táto oblasť sa zaoberá tvorbou algoritmov ktoré označujeme ako hlboké siete (vychádzajúc z knihy z práce S. C. Kleene a McCarthy (1956) a praktických znalostí z knihy Moroney (2020)). Tými sa



Obrázok. 3.7: Graf aplikácie modelu rozhodovacieho stromu/náhodného lesu/XGBoost pre názorný príklad pre priemet do plochy dvoch parametrov X_1 a X_2. Modré a červené body patria do dvoch separátnych tried a farebná plocha reprezentuje oblasť každej danej triedy.

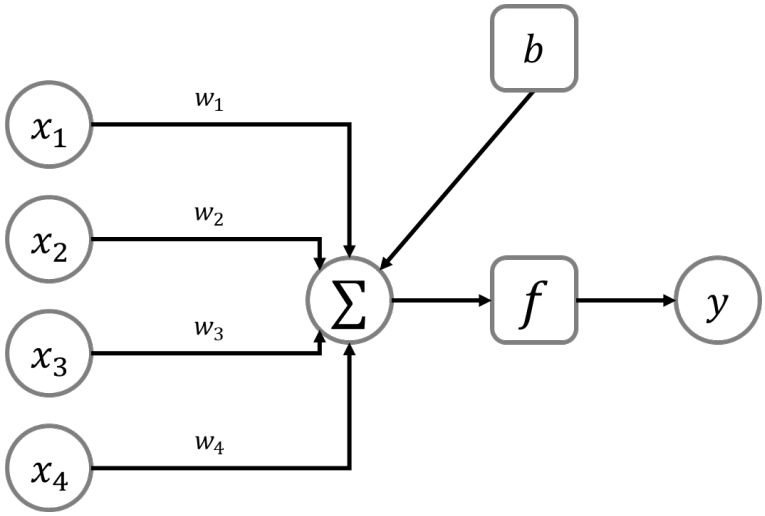
budeme ďalej zaoberať pretože môžu pri zložitých problémoch vykazovať lepšie výsledky no sú náročnejšie na návrh a učenie a zároveň sú výpočtovo náročnejšie čo predstavuje problém ak by sme ich chceli nasadiť na palube stratosférického balóna ktorý nedisponuje výpočtovým výkon. To je ale problém ktorý sa do určitej miery dá riešiť v súčasnosti množstvom postupov.

3.3.1 Neurónová a hlboká neurónová sieť

Asi najnákladnejší stavebný blok celého hlbokého strojového učenia sú neurónové a hlboké neurónové siete. Tie vychádzajú z inšpirácie v nie len ľudskom mozgu. Výrazy neurónová sieť a hlboká neurónová sieť budeme ďalej v práci používať ako zameniteľné pomenovania a budeme tým myslieť hlavne hlboké siete. Ako už názov napovedá, neurónová sieť je prepojená (buď plne alebo čiastočne) sieť jednotlivých neurónov (viď obrázok 3.9). Neurón je elementárna jednotka ktorá má mnoho vstupov ale iba jeden výstup (viď obrázok 3.8).

3.3.2 Konvolučné neurónové siete

Konvolučné neurónové siete (vychádzajúc z ich prvej zmienky v článku Fukushima (1980)) sú len kombináciou konvolučných a neurónových sieti (viď obrázok 3.10). Neurónovú časť sme už vysvetlili, preto sa zameráme na konvolučnú časť. Konvolúcia je matematická operácia (operátor) ktorá ako vstup berie dve

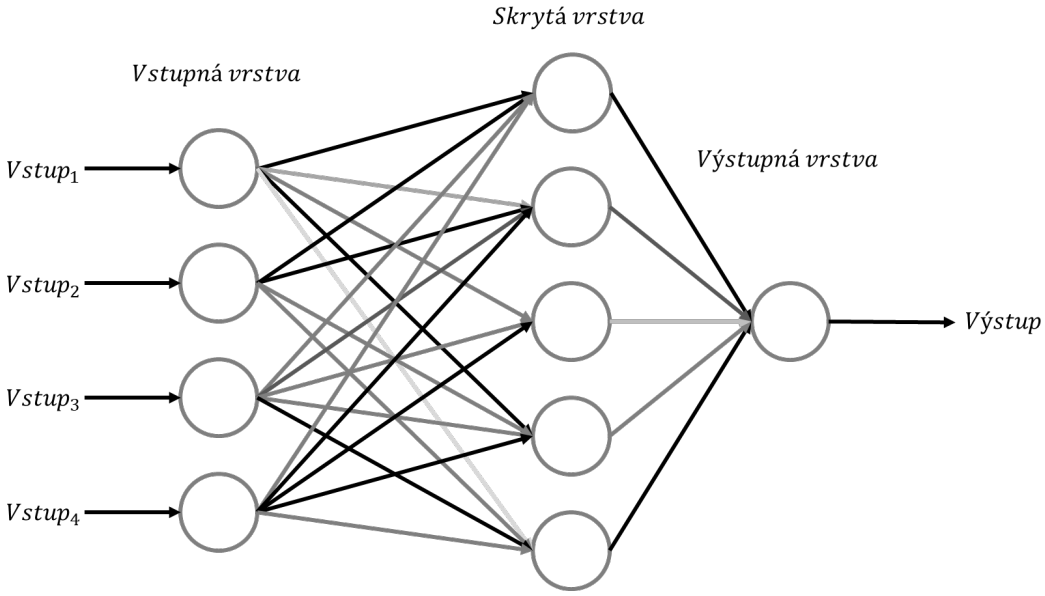


Obrázok. 3.8: Schéma principiálne ukazujúca architektúru a prácu jedného jednoduchého neurónu, tzv. perceptronu. $x_{1;...;4}$ sú vstupné údaje, $w_{1;...;4}$ sú váhy spojení, $\sum$ je jednoduchá sumácia dát krát váh. b je vstupný šum. f je aktivačná funkcia a y je výstupná hodnota.

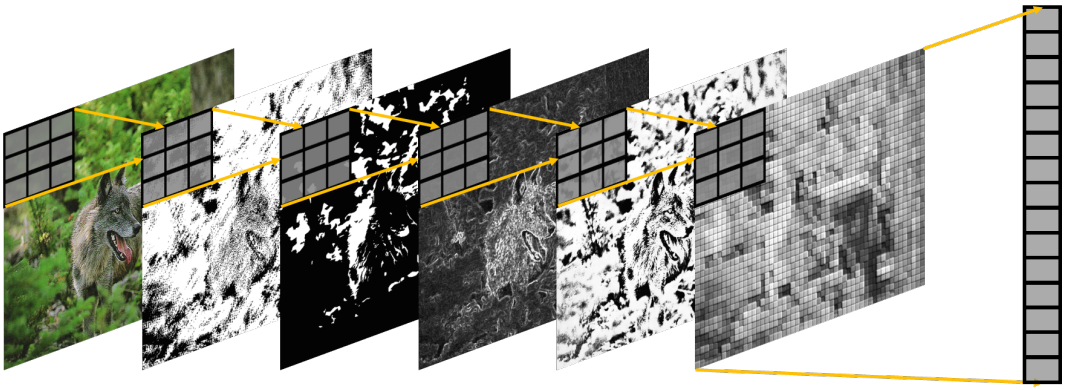
funkcie a vracia jednu. V počítačom svete sa ale najčastejšie stretneme s diskrétnou a dvoj-dimenzionálnou variantou konvolúcie. V takom prípade sa jedna z funkcií (bežne označovaná ako jadro konvolúcie) dá chápať ako tabuľka/konvolučná maska ktorá v prípade grafiky je preložená cez raster. Hodnoty v obrázku sa násobia s hodnotami v maske a výsledky sa sčítavajú. Konvolučná sieť je následne súsled takýchto konvolúcií. Pričom učenie konvolučnej siete je správne nastavenie číselných hodnôt v maske tak aby sme vo výsledku číselne vyjadrili prítomnosť a polohu dôležitých príznakov vo vstupných dátach. Výsledný vektor/ vektorizovaná matica je následne vstupom do neurónovej siete.

3.3.3 YOLOv5

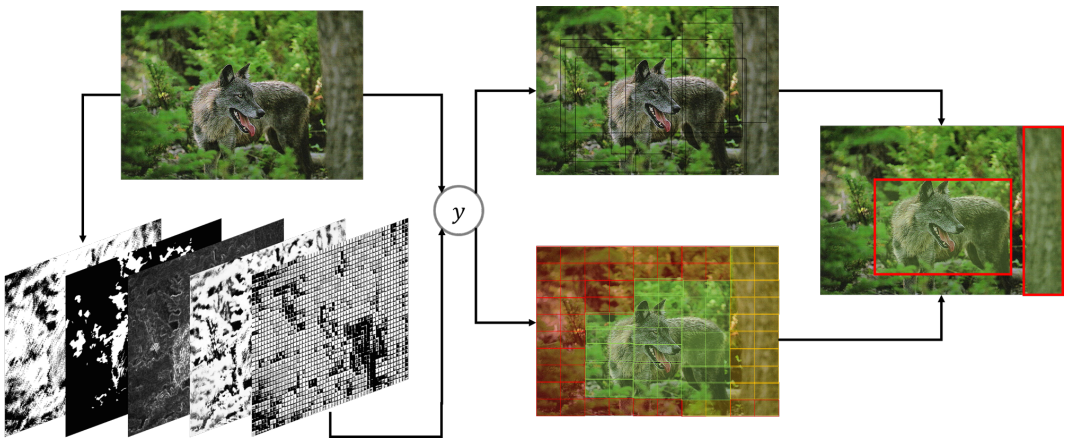
YOLOv5 je skupina modelov hlbokých neurónových konvolučných sieti založených na článku Redmon a kol. (2016), konkrétne už v piatej iterácií. Samotné označenie pochádza z anglického You only look once (YOLO). Základom je rozdelenie analýzy na dvojicu viac menej nezávislých analýz. Jedna sa zaoberá rozdelením obrazu do siete tak aby označila všetky separátne objekty. Druhá analýza už samotné takto vyčlenené objekty identifikuje (viď graf 3.11). Problematické je, že v súčasnej dobe je táto analýza tak zložitá, že sa používa výhradne ako "black box"metóda, teda bez možnosti interpretácie. Na druhú stranu YOLOv5 je extrémne populárne v praxi, pretože dovoľuje jednoducho a rýchlo tvoriť veľmi presné modely.



Obrázok. 3.9: Schéma jednoduchej neurónovej siete s jednou skrytou vrstvou. Odtieň šedej reprezentuje silu spojenia (veĺkosť váhy).



Obrázok. 3.10: Znázornenie práce diskrétnej konvolúcie na dvoj-dimenzionálnych dát, konkrétne obrázkov.



Obrázok. 3.11: Veľmi zjednodušené ukázanie princípu fungovanie modelu YOLOv5. Vstupný obrázok prechádza viacnásobnou konvolučnou vrstvou ktorá extrahuje všetky charakteristiky. Následne je analýza rozdelená na dva separátne algoritmy, jeden určuje boxy v ktorých sa pravdepodobne niečo nachádza, druhý určuje heatmapu pravdepodobnosti konkrétneho objektu na obrázku. Spojením týchto dvoch údajov získame pozadovanú segmentáciu obrázku.

4. Ciele práce

Z predchádzajúcich kapitol vyplývajú nasledujúce postupné ciele praktickej časti tejto bakalárskej práce:

1. Pripraviť software pre prehľadávanie niekoľkých rokov archívnych dát magnetických slučkových antén z meracích stanovíšť Ústavu fyziky atmosféry Milešovka, Lomnický Štít, Dlouhá louka a Krupka.
2. Dáta rozdeliť na spracovateľné úseky a po použití high-pass filtru numericky integrovať.
3. Vzniknuté záznamy priebehu magnetickej indukcie podrobiť frekvenčnej analýze a výsledné spektrogramy pripraviť pre použitie v strojovom učení.
4. Pripraviť množinu ručne označených dát pre nasledujúce kategórie javov:
   (a) Iniciačná fáza CG.
   (b) Iniciačná fáza IC.
   (c) IC aktivita.
   (d) Skupina mikrosekundových pulzov.
   (e) Úzka bipolárna udalosť.
   (f) Spätný výboj
5. Pripraviť metriky pre vyhodnotenie úspešnosti strojového učenia.
6. Pre rozpoznávanie kategórií uvedených v bode 4 použiť niekoľko rôznych metód strojového učenia a vyhodnotiť ich úspešnosť.

5. Príprava archívnych dát

Prvým krokom praktickej časti je prehľadávanie, rozbaľovanie a predpríprava archívnych dát z archívu Ústavu fyziky atmosféry AV ČR.

5.1 Využitý software a programovacie jazyky

Dáta sme pred-spracovali a analyzovali v programovacom jazyku Python z distribúcie anaconda verzie 3.9. Používané dodatočné knižnice sú: py7zr, os, glob, shutil, zipfile, matplotlib.pyplot, numpy, scipy.

5.2 Zdroj dát

Naším zdrojom bol archív dát z rokov 2018 až 2020. Dáta pochádzajú z viacerých miest Milešovka (ML), Lomnický Štít (LS), Dlouha louka (DL) a Krupka (KR). Merané boli pozemnými detektormi magnetického poľa. Takýmto detektorom je tienená slučková anténa. Tá je schopná operovať na frekvencií 5 kHz až 90 MHz. Zároveň má anténa integrovaný predzosilňovač s riadeným ziskom. Anténa je následne napojená na širokopásmový analyzátor ktorý vzorkuje na 200 MHz.

5.3 Prehľadávanie dát

Všetky dáta v archíve sme kompletne prešli pomocou rekurzívneho prehľadávania. Ukážka kódu je zobrazená nižšie. Takto sme schopní si v programe zaevidovať všetky dáta ktoré sa v archíve nachádzajú a jednoducho si následne rozbaliť dané meranie ktoré požadujeme. Každé meranie je separátny 7zip archív ktorý obsahuje jednak údaje z analyzátora a jednak sprievodný textový súbor s dodatočnými informáciami ako je napríklad čas merania, množstvo nameraných dát ale aj či sa jednalo o softvérový alebo hardvérový spúšťač. Typ spúšťaču nám hovorí o tom či dané meranie bolo zaznamenané iba ako priehľadové meranie (softvérový) alebo po tom ako meraná hodnota prekonala určité hranicu magnetického poľa (hardvérový). Výsledok tohoto prehľadávania je viacrozmerný slovník ktorý nám umožňuje overiť či požadovaný čas merania existuje a následne ho jednoducho vyhľadať v archíve.

```
def dict_all_sub(path):
    if os.path.isdir(path):
        temp = {}
        for name in os.listdir(path):
            temp[name] = dict_all_sub(os.path.join(path, name))
    else:
        temp = path
    return temp
```

5.4 Príprava dát

Rozbalené binárne dáta v priečinku sa načítajú do pamäte a konvertujú do celočíselných údajov. Tieto údaje sú séria čísel a metadát. Plánujeme aplikovať na dáta metódy strojového učenia, preto musíme rozdeliť celý priebeh záznamu na jednotlivé spracovateľné úseky dlhé 1 ms, čo pri vzorkovacej frekvencií 200 MHz predstavuje 200 000 meraní. Tie uložíme ako separátne textové súbory s názvom zodpovedajúcim času pozorovania. Názorným príkladom môže byť meranie (ml20200811hs_20200811_133533_5192852928). Kde (ML) znamená meranie zo stanice Milešovka. (20200811) znamená, že meranie bolo zaznamenané 11.08.2020. (133533) znamená, že meranie bolo zaznamenané o 13.35.33. Posledný údaj (5192852928) znamená, že daný 1 ms výsek je z údajov danom meraní s poradovým číslom 51928 až 52928. Všetky takéto záznamy sú o rovnakej časovej dĺžke 1 ms.

Prvou predprípravou dát je ich nízkopásmové orezanie na hladine 2 kHz. Nízkopásmové orezanie sa aplikuje aby po následnej integrácií signálu sa v ňom nenachádzalo lineárne zvyšovanie hodnoty spôsobené nízkofrekvenčným signálom. Hodnota 2 kHz sa vyberá ako hodnota kedy nenarúšame ešte samotné tvary následných integrovaných foriem ale už máme dostatočné odstránenie nízkych parazitných frekvencií.

5.5 Výber charakteristík

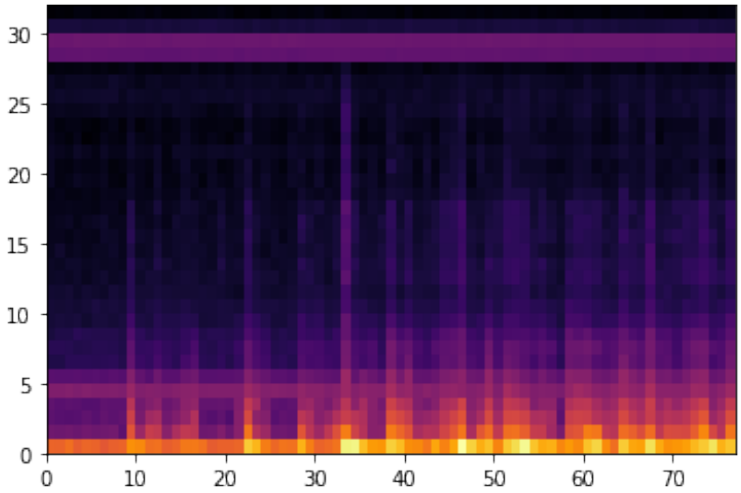
To čo vidí náš model strojového učenia je len vektor čísel ale mi si tieto informácie vieme zobraziť v ľubovoľnej etape. Základ z ktorého vychádzame je časovo-frekvenčný spektrogram. Ten tvoríme vždy z 200 000 súsledných údajov (1 ms). Parametrami Fourierovej transformácie sú šírka okna 1024 bodov, a prekryv 512 bodov. Čo reprezentuje 1ms času (viď druhú časť grafu 2.3).

Nasleduje samotná príprava dát na to aby ich bolo možné použiť v metódach strojového učenia. Keďže v klasickom prípade sa jedná o učenie s učiteľom je potrebné ručne extrahovať charakteristiky signálu vhodné pre model.

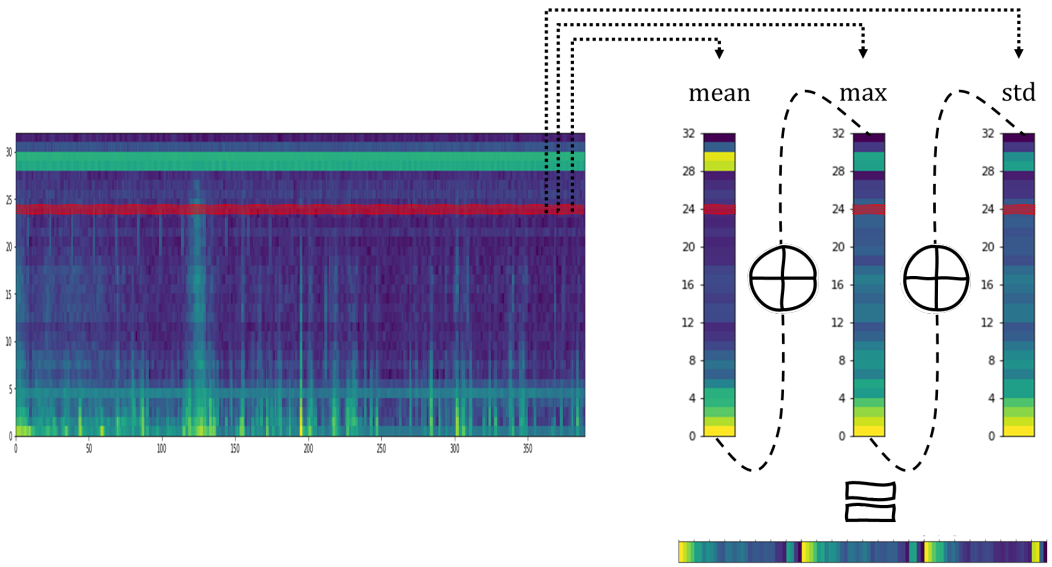
V prvom prípade sme vytvorený spektrogram rozdelili na 32 ekvidistantných frekvenčných intervalov a s tými sme ďalej pracovali ako s časovou radou údajov. Preto sme na tieto rady aplikovali 3 rôzne funkcie a to maximum, suma a smerodajná odchýlka. Takto získame trojicu vektorov o dĺžke 32 bodov. Z nich vytvoríme jeden veľký vstupný vektor o dĺžke 96 bodov pre náš model (viď graf 5.2).

V druhom prípade sme spektrogramu zmenšili rozlíšenie (rozpixelovali), čím sme vytvorili veľké celky (viď graf 5.1). Pričom ako vstupný vektor sme využili práve tento spektrogram rozvinutý do vektora.

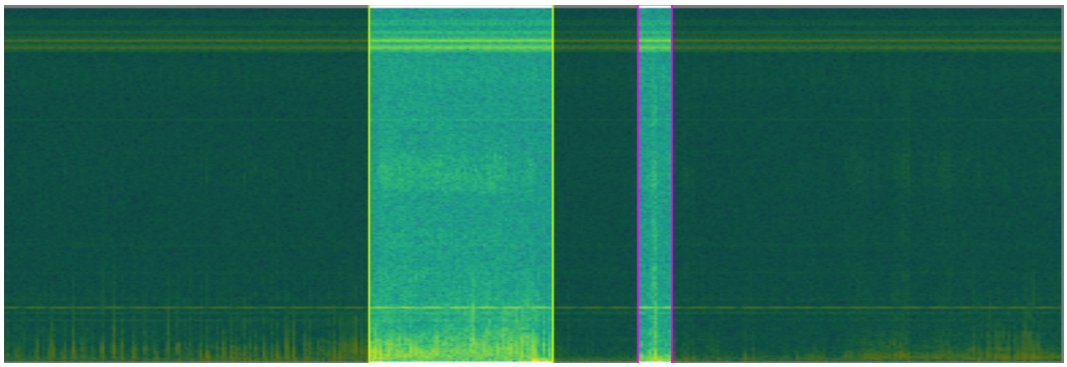
V prípade YOLOv5 si samotná sieť vyberá potrebné charakteristiky a teda sa jedná o učenie bez učiteľa. Ale zároveň je potrebné označiť konkrétnu polohu hľadaného úkazu (viď obrázok 5.3).



Obrázok. 5.1: Jedna z metód tvorby charakteristík pre metódy strojového učenia. V tomto prípade sa jedná o roxpixelovanie spektrogramu.



Obrázok. 5.2: Grafické znázornenie vytvorenia vstupného vektora pre metódy zdrojového učenia. Na spetrografický záznam sa aplikujú funkcie maximum, priemer, smerodajná odchýlka (anglické skratky *max*, *mean*, *std*). Takto vzniknú tri vektory ktoré spojíme do jedného výsledného ktorý je vstupom pre naše metódy strojového učenia.

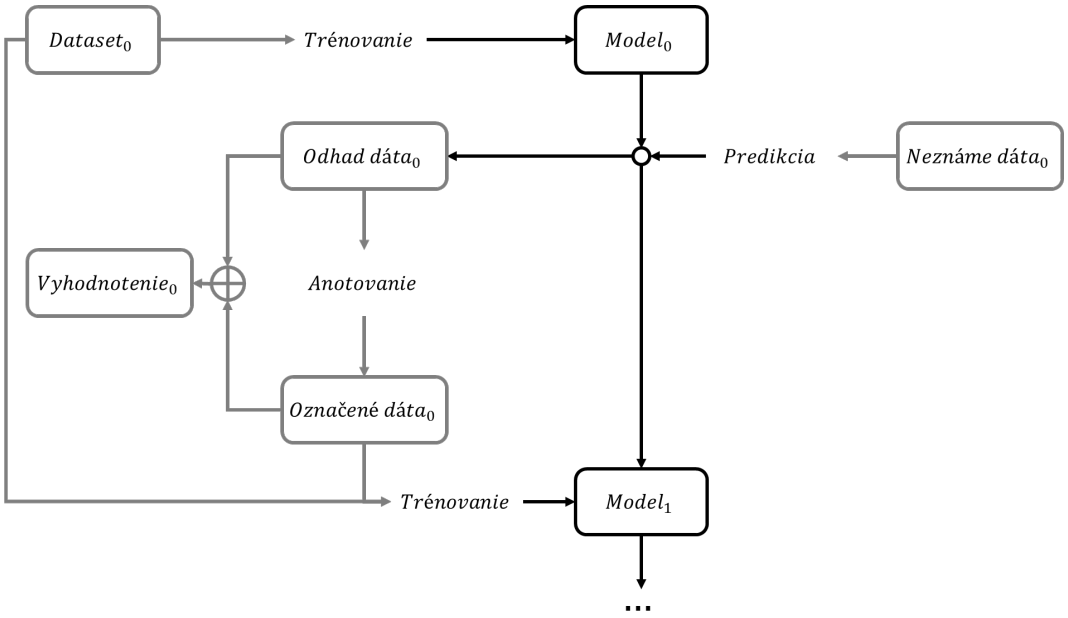


Obrázok. 5.3: Spektrogram ale ako čistý obrázok ktorý spolu s vyznačenými štvrocami slúži ako vstup pre metódy hlbokého učenia.

6. Analýza archívnych dát

6.1 Pipeline dát

Ručné označovanie dát pre analýzu je náročné ako na profesionalitu tak aj na čas. Preto sme pristúpili na iteratívny systém (viď obrázok 6.1). To znamená, že ako prvé sa anotuje malá množina meraní (ideálne sa vyberajú merania ktoré najlepšie reprezentujú dané kategórie). Táto malá množina sa použije na trénovanie počiatočného modelu. O ňom predpokladáme, že nebude dostatočný ale môžeme očakávať už jeho lepšiu schopnosť klasifikácie voči náhodnému rozdeľovaniu. Tento prvý model použijeme na anotovanie množiny, pre model neznámych, meraní. My následne urobíme anotáciu na tejto množine dát (urýchlenie spočíva v možnosti ponechania anotácie modelom ako aj v eradikácií falošne pozitívnych javov). Porovnaním anotácie modelom a našej vieme dokonalejšie vyhodnotiť úspešnosť modelu (máme väčšiu vzorku testovacích dát ktoré model pri trénovaní nevidel). Následne môžeme nami anotované merania pridať do trénovacieho datasetu. Model pretrénujeme na väčšej vzorke. čím dostaneme ideálne dokonalejší, presnejší model a celý proces predikcie, anotácie, vyhodnocovania a trénovania môžeme opakovať.

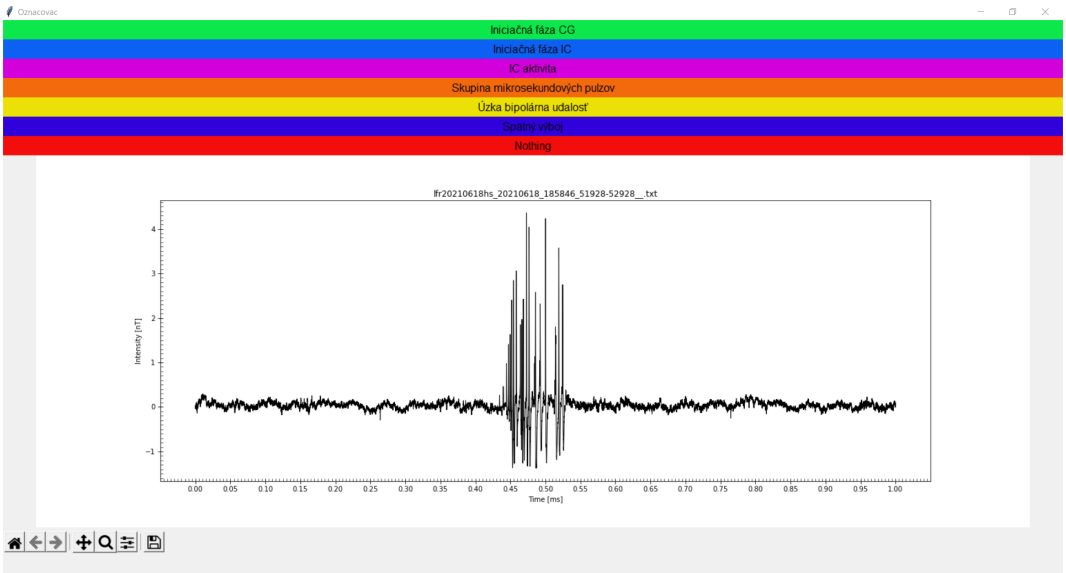


Obrázok. 6.1: Schematické znázornenie trénovania, validácie, testovania, anotácie a vyhodnocovania modelov a dát použitých v práci.

6.2 Označovanie dát

Pre uľahčenie anotovania dát sme potrebovali použiť nejaký nástroj ktorý by nám ukázal vlnovú formu merania (ideálne v interaktívnej forme) a dovoľoval by nám jednoducho označiť prislúchajúcu kategóriu a toto naše označenie uložiť. Žiaden z voľne dostupných nástrojov (napr. zooniverse) nám ale nevyhovoval. Vytvorili sme teda vlastný nástroj (viď obrázok 6.2). Ten práve fungoval spôsobom,

že do internej pamäte natiahol všetky merania z priečinku a urobil na nich potrebné predspracovanie. Následne nám zobrazoval jednu vlnovú formu za druhou a my sme mohli jednoduchým kliknutím na príslušné tlačidlo označiť príslušnosť do kategórie. Naše označenie sa priebežne ukladá do textového súboru, s ktorým následne pracujeme.



Obrázok. 6.2: Ukážka vlastného programu ktorý slúži na označovanie reálnej kategórie dát.

6.3 Použité metriky

Vždy je nutné charakterizovať účinnosť jednotlivých modelov a to ako dobre klasifikujú v komparácií s ostatnými. Preto sme pre jednotlivé modeli použili metriky uvedené nižšie:

- Koeficient determinácie (R2)
- Root-mean-square deviation (RMSE)

A pre vyhodnotenie úspešnosti modelu sme použili.

- Accuracy (ACC)
- Precision (PRC)

Kde R2 (matematicky zapísané aj ako R^2) je koeficient determinácie a je to jednoduchý, ale pritom veľmi efektívny nástroj. Je definovaný ako:

$$R^2 = 1 - \frac{S_{reg}}{S_{avg}}. \tag{6.1}$$

Kde S_{reg} je suma všetkých štvorcov odchýlky hodnoty od regresnej krivky a S_{avg} je suma všetkých štvorcov odchýlky hodnoty od priemeru hodnôt (viď článok

Renaud a Victoria-Feser (2010)). Nie je ale vhodné ho používať ako jediný rozhodovací parameter, pretože môže byť ľahko ovplyvniteľný malým množstvom chybných hodnôt.

Metrika RMSE je skratka pre strednú kvadratickú odchýlku (z anglického root-mean-square error). Vyjadruje smerodajnú odchýlku rezíduí, teda rozdielov meraných a predikovaných dát (viď článok Willmott a Matsuura (2005)). Matematicky zapísateľné ako:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum e_i^2}. \tag{6.2}$$

Kde e_i je rozdiel model-predikcia.

Metrika ACC (voľne preložiteľná ako presnosť) je v základe zadefinovaná pre binárnu klasifikáciu ale je možné definíciu rozšíriť aj na mnoho kategorickú klasifikáciu. Matematicky je ju možné zapísať ako:

$$Accuracy = \frac{\text{Počet správnych predikcií}}{\text{Počet všetkých predikcií}}. \tag{6.3}$$

Metrika PRC (voľne preložiteľná ako precíznosť) je podobne ako Accuracy zadefinovaná pôvodne na binárnu klasifikáciu nasledujúcim vzorcom:

$$Precision = \frac{\text{Počet správnych predikcií}}{\text{Počet všetkých kladných predikcií}}. \tag{6.4}$$

Pre mnoho kategorickú klasifikáciu je jedna z možnosti rozšírenia definície nasledujúca. Vypočítame Precision pre jednotlivé kategórie a výsledná hodnota je priemer Precision pre jednotlivé kategórie.

6.4 Použité modely

Z oblasti klasických modelov strojového učenia sme použili nasledujúce modely:

- Lineárna regresia
- K-najbližších susedov
- Rozhodovací strom
- Náhodný les
- Náshodný les boost
- Metóda pomocných vektorov
- XGBoost

6.5 Trénovací hardware a trénovanie

Krátkodobé trénovania modelov prebiehalo pravidelne počas vývoja ako nástroj kontroly zlepšovania modelov. V súčte boli použité štyri rôzne počítače (2 lokálne a 2 cloudové). Z počiatku bol celý výskum vykonávaný na lokálnom stroji s procesorom Intel® Core™ i7-9750H a s akceleračnou kartou Nvidia GeForce RTX 3060 Mobile (GA106 Ampere, 6 GB GDDR6). Na tomto stroji boli ku koncu aj vykonávané finálne učenia modelov strojového učenia. Záverečné analýzy boli ale z technických dôvodov dokončené na lokálnom stroji vybavený procesorom Intel® Core™ i3-1115G4 bez akceleračnej karty. Učenie modelov hlbokého učenia prebiehalo na vzdialenom stroji vybavený procesorom Intel® Xeon™ (1 jadro, 2 logické vlákna) a s akceleračnou kartou Nvidia K40 (GK110 Kepler, 12 GB GDDR5) alebo K80 (2 krát GK210 Kepler 2.0, 24 GB GDDR5) podľa aktuálnej inštancie učenia. Pričom informácie o dátovej množine použitej pri učení je v tabuľke 6.1.

Názov kat.	č. kat.	Množstvo dát	Z toho validačné
Iniciačná fáza CG	0	132	31
Iniciačná fáza IC	1	116	23
IC aktivita	2	346	69
Skupina mikrosekundových pulzov	3	103	22
Úzka bipolárna udalosť	4	163	40
Žiadna udalosť	5	176	34
Spätný výboj	6	281	45

Tabuľka 6.1: Tabuľka označenia a číslovania jednotlivých kategórií ako aj počtu meraní v danej kategórií celkovo a vo validačnej množine.

6.6 Dosiahnuté výsledky

6.6.1 Klasické strojové učenie

Optimalizáciou hyper parametrov modelov sme dosiahli najlepšie výsledky pre rozhodovacie stromy (príklad optimalizácie hyperparametrov na podmnožine je vidieť na grafe 6.3). Pričom výsledky pre zvyšné metódy sú zobrazené v grafe 6.4. Dôležité pre posúdenie úspešnosti modelu ale je konfúzna matica. Pre náš najlepší model je práve takáto matica vytvorená a zobrazená na obrázku 6.5. Z danej konfúznej matice vidíme, že náš model dokáže rozpoznať s dostatočnou presnosťou iniciačnú fázu CG a iniciačnú fázu IC. Problém nastáva pri rozpoznávaní IC aktivity, ktorú zamieňa za skupinu mikrosekundových pulzov alebo z časti aj za žiadnu udalosť. Na druhú stranu to či sa na zázname nachádzala žiadna udalosť vieme taktiež určiť s dostatočnou presnosťou. Miestom na zlepšenie je klasifikovanie spätného výboja, ktorý je identifikovaný zatiaľ častokrát ako úzka bipolárna udalosť alebo ako žiadna udalosť.

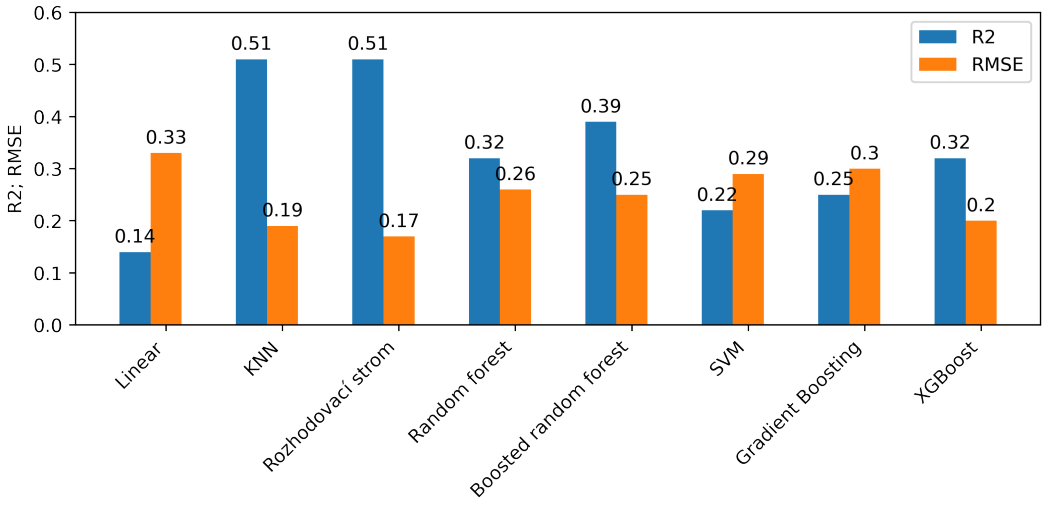
6.6.2 Hlboké strojové učenie

Hlboké strojové učenie nedosahovalo dostatočne zaujímavé výsledky aby sme sa s nimi museli ďalej zaoberať. Problematické bolo, že sa modely nikdy nedostali

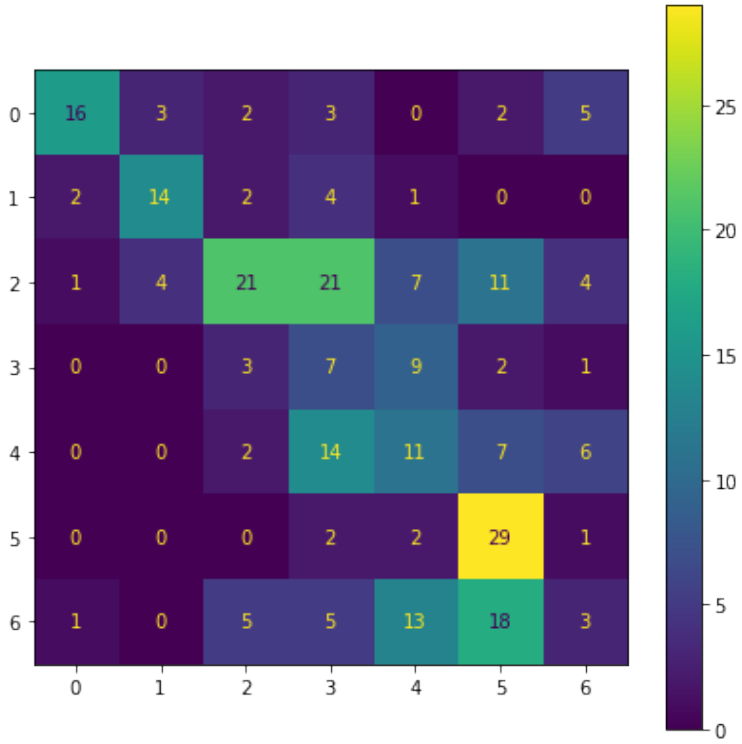
na validačnej množine nad úroveň náhodného označovania.

n\p	1	2	3	4	5	6	7	8	9	10
1	-0,12	-0,09	-0,33	-0,48	-0,39	-0,36	-0,43	-0,49	-0,64	-0,65
2	0,28	0,45	0,31	0,24	0,18	0,16	0,15	0,10	0,06	0,12
3	0,51	0,42	0,39	0,36	0,34	0,29	0,22	0,23	0,26	0,22
4	0,45	0,45	0,44	0,40	0,31	0,34	0,38	0,39	0,37	0,34
5	0,47	0,42	0,47	0,43	0,42	0,46	0,44	0,44	0,39	0,38
6	0,47	0,45	0,42	0,44	0,44	0,47	0,44	0,43	0,41	0,40
7	0,43	0,45	0,47	0,48	0,44	0,44	0,45	0,44	0,40	0,38
8	0,42	0,42	0,46	0,43	0,42	0,43	0,42	0,42	0,39	0,38
9	0,39	0,42	0,41	0,45	0,41	0,41	0,41	0,43	0,41	0,40
10	0,39	0,42	0,42	0,40	0,39	0,40	0,41	0,38	0,38	0,38

Obrázok. 6.3: Tabuľka hodnoty metriky R2 pre všetky rozumné nastavenia hyperparametrov modelu. V tomto prípade sa jedná o model KNN. Stĺpce (označené ako p) reprezentujú mocninu s ktorou sa ráta metrika v danom priestore. Riadky (označené ako n) reprezentujú koľko najbližších susedov sa berie do úvahy pri vážení daného merania.



Obrázok. 6.4: Stĺpcový graf úspešnosti jednotlivých modelov v jednotlivých metrikách. Je vidieť, že ako najúspešnejšie sa javí model rozhodovací strom a KNN.



Obrázok. 6.5: Konfúzna matica pre model rozhodovací strom. Vertikálna osa reprezentuje reálne kategórie validačného datasetu. Horizontálna osa reprezentuje predikované kategórie. Jednotlivé čísla reprezentujú počty meraní (počty podľa tabuľky 6.1).

Diskusia

V práci bolo potrebné ako prvé popísať vlastnosti elektromagnetických vĺn vyžarovaných elektrickými výbojmi v atmosfére. Nadobudnuté informácie nám pomohli s výberom vhodných charakteristík a predspracovania signálu v praktickej časti. Taktiež sme popísali fungovanie meracej antény ktorá sa používa na záznam signálu z bleskových výbojov.

V druhej kapitole teoretického úvodu sme zložili kompilát literatúry ktorá sa zaoberá témou bleskov, bleskových výbojov, ich klasifikácie a vývojových štádií. Čerpaním informácií zo širokého spektra článkov, kníh a inej literatúry sme sa snažili predísť chybným informáciám, výberovým skresleniam a zastaraním faktom. Práve takýto prístup považujeme za prínosný, pretože na rozdiel prác z oboru strojového učenia, ktoré pristupujú k problematike iba ako k práci so všeobecnými dátami sme sa snažili práve pochopiť fyzikálne princípy a tak lepšie nastaviť predspracovanie a výber charakteristík pre modely strojového učenia.

V tretej kapitole teoretického úvodu sa naša pozornosť zamerala na jednotlivé nástroje našej analýzy, strojové učenie. Teoretické zhrnutie princípu fungovania sme spísali na základe aktuálneho pohľadu na umelú inteligenciu a definíciu strojového učenia. Pre jednotlivé popisy použitých modelov sme sa snažili vychádzať z pôvodných zdrojov, teda článkov a kníh ktoré dané modely ako prvé navrhli. Neskôr v praktickej časti, pracujeme ďalej s pojmami metriky, ktoré taktiež spadajú aj do oboru strojového učenia. Tam sme vychádzali z matematických definícií pri popise a z dobrej praxe pri výbere vhodných metrík. V oboch prípadoch, výber modelov strojového učenia aj výber metrík, môže byť zaťažený výberovým efektom a našou subjektívnosťou. Proti tomu sme sa snažili bojovať hlavne odsledovaním hlavných moderných trendov v strojovom učení a prácou na všetkých modeloch v rovnakých podmienkach. Prvou praktickou výzvou práce bola snaha o extrakciu meraní z archívu. Dáta sú uložené v jednotlivých priečinkoch ktoré reprezentujú miesta merania, rok, mesiac a deň merania a jednotlivé súbory sú pomenované podľa konkrétneho času merania. Pričom samotný súbor je archív ktorý obsahuje ako meranie, tak aj metadáta (v ktorom je napr. informácia o tom, či sa jedná o reálny záznam nejakého úkazu alebo iba o rutinný pravidelný záznam). Problematická stránka bola práve prehľadávanie tohto archívu ak sme si chceli vyvolať konkrétne meranie alebo merania z konkrétneho rozpätia časov a miest. Z tohto dôvodu sme vytvorili vlastný program ktorý pracuje dvojfázovo. V prvej fáze si vytvorí mapu, rekurzívne zmapuje celý súborový systém. To nám dovoľuje následnú rýchlu orientáciu v dostupností meraní ako aj vo vyvolaní merania na základe intuitívneho zadávania času a miesta. Druhou fázou je práve rozbalenie a rozkúskovanie žiadaného merania na úseky s ktorými ďalej pracujeme. Jednou nevýhodou s ktorou sa stretávame je časová (súvisí s nie vysokou optimalizáciou kódu, zastaraním hardvérom, obmedzenou prenosovou rýchlosťou na vzdialený archív a s veľkosťou samotných meraní). Prvá fáza mapovania trvá približne desať minút a rozbalenie a rozkúskovanie jedného merania trvá približne nižšie minúty. Proti tomuto sme sa snažili bojovať simultánnym behom programu na všetkých dostupných logických jadrách procesora. To nám vo výsledku dovolilo sa dopracovať k pár tisíckam predpripravených meraní ktoré sme ďalej využili v práci.

Po nadobudnutí dostatočného množstva meraní bolo potrebné jednotlivé dáta označiť ručne podľa príslušnej kategórie do ktorej spadajú. Nami vytvorený program, navrhnutý pre tento konkrétny účel, dané označovanie významným spôsobom zrýchľoval a zjednodušoval. Hoc prvotný štart programu a jeho inicializačná fáza počas ktorej pripravoval grafy integrovaných foriem zabrala v priemere nižšie jednotky minút na 100 meraní. Tak následne bolo možné týchto sto meraní klasifikovať pod hodinu. To najme vďaka tomu, že nám náš program dovoľoval interaktívne si graf približovať a posúvať. Výber kategórie prebiehal stlačením prislúchajúceho tlačidla a naše označenie sa zapísalo do textového súboru s ktorým sme vedeli následne poľahky pracovať. Jedinou veľkou nevýhodou alebo skôr priestorom na zlepšenie je to, že každé meranie bolo brané ako jedinečné, pričom niekedy by sa nám hodila informácia o priebehu pred alebo za daným javom. Zobrazenia tejto informácie ale náš program nebol schopný. Druhou, menšou nevýhodou, ale nie tak programu ako celého označovania je, ak sa vyskytne dvojica rôznych udalostí v jednom zázname. V takomto prípade daný záznam priradíme do oboch kategórií duplicitne.

Nasledujúcim je samotná snaha o natrénovanie modelov strojového učenia. Jedným z problémov je prerozdelenie datasetu na trénovací a validačný subset. Keďže toto prerozdelenie nastáva náhodne, môže dôjsť k nerovnomernému rozloženiu jednotlivých kategórií udalosti do subsetov. Odstránenie tejto chyby je pretrénovávanie na vždy novo vygenerovaných subsetoch. Ďalším problémom môže byť, že naše trénovacie dáta majú mierne iné charakteristiky voči meraniam ktoré následne budú na balóne. Čo mže spôsobiť systematickú chybovosť modelu v praktickom nasadení.

Pre správnu implementáciu je potrebné štatisticky vyhodnocovať úspešnosť a sledovať zlepšovanie našich modelov. Na túto úlohu sme vytvorili program ktorý na základe dvojice textových súborov s pomenovaniami pridelených kategórií k meraniam vypočíta jednotlivé metriky. Táto dvojica súborov zodpovedá predikciám z modelu a našej klasifikácií. Je to len jednoduchý program ale dovoľuje nám efektívne sledovanie vývoja presnosti modelov cez jednotlivé iterácie.

Problematickým miestom našej práce môžu byť overfitting (zjednodušene sa jedná o stav kedy má náš model príliš veľký počet voľných parametrov a dochádza k naučeniu datasetu naspamäť namiesto toho aby model extrahoval hlbšie zákonitosti v dátach). To je ale najčastejšie problém metód hlbokého strojového učenia (vlastné hlboké konvolučné siete, modely YOLOv5). Tie ale vykazovali iné nevýhody (náročnosť na testovaciu vzorku, výpočetná náročnosť, fyzikálna neinterpretácia) ktoré ich spravili nevhodnými pre náš typ problému.

Záver

Teoreticky sme popísali bleskový výboj, štádia jeho vývoja v čase ako aj delenie ktoré je vhodné pre klasifikáciu metódami strojového učenia. Popísali sme pôvod vzniku blesku. Jednotlivé mechanizmy vzniku nehomogenity rozloženia elektrického náboja v búrkových oblakoch. Podrobnejšie sme rozobrali časový vývoj blesku. A záverom sme si jednotlivé kategórie zobrazili na 1 ms záznamoch spolu s ich prislúchajúcimi spektrografickými záznamami.

V našej práci využívame podmnožinu umelej inteligencie, ktorá sa nazýva strojové učenie. Preto sme popísali jednotlivé prístupy k strojovému učeniu. Popísali sme niekoľko bežne využívaných modelov. Ku každému sme uviedli vlastnosti, výhody a nevýhody. Na názorných príkladoch sme ukázali ako jednotlivé modely fungujú.

Prvým krokom praktickej časti bolo prehľadávanie archívu dát. Na túto prácu sme vytvorili vlastný program ktorý rekurzívnym spôsobom zmapuje súborovú štruktúru archívu a následne nám dovolí extrahovať meranie z ľubovoľnej stanice v ľubovoľnom čase.

Ná základe teoretického rozboru bleskového výboja sme sa rozhodli pre využitie spektrogramu daného merania ako zdroja charakteristík pre modely strojového učenia. Pričom ako najefektívnejší spôsob sa nám overila kombinácia rozpixelovania spektrogramu a zostrojenia štatistických momentov frekvenčných pásiem v spektrogramu.

Aby sme boli schopný efektívnej práce pri učení modelov strojového učenia, vytvorili sme program ktorý urýchľuje označovanie jednotlivých 1 ms záznamov do kategórií.

Pre potrebu určenia najvhodnejšieho modelu strojového učenia sme vytvorili program ktorý na nami označených dátach trénoval všetky vyššie vypísané modely strojového učenia. Za pomoci pretrénovaní na celom rozumnom priestore hyperparametrov sme vybrali kandidátov na najvhodnejšie modely.

Aby sme vedeli postupne sledovať zlepšovanie presnosti modelu vytvorili sme program ktorý nám spočíta jednotlivé štatistické metriky ná základe predpovedí modelu a našich označení dát.

Výsledkom snaženia bolo, že sme ako najvhodnejší model vybrali rozhodovací strom ktorý dosahoval hodnotu až $R^2 = 0.51$.

Do budúcna je možné za použitia už pripravených nástrojov iteratívnym spôsobom zlepšovať ešte ďalej presnosť modelu na klasifikáciu bleskových výbojov. Zároveň by do budúcna malo byť možné implementovať nami vytvorený model do programovateľných hradlových polí, ktoré sa využijú na palube stratosférického balóna.

Zoznam použitej literatúry

Bedřich Sedlák, I. Š. (2012). *Elektřina a magnetismus*. Karolinum, Praha. ISBN 978-80-246-2198-2.

Boser, B. E., Guyon, I. M. a Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. doi: 10.1145/130385.130401. URL https://doi.org/10.1145/130385.130401.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T. a kol. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, **1**(4), 1–4.

Cooray, V. (2016). *An Introduction to Lightning*. Springer Netherlands. ISBN 9789402401110. URL https://books.google.cz/books?id=AXQ_vgAACAAJ.

Dye, J. E., Jones, J. J., Winn, W. P., Cerni, T. A., Gardiner, B., Lamb, D., Pitter, R. L., Hallett, J. a Saunders, C. P. R. (1986). Early electrification and precipitation development in a small, isolated montana cumulonimbus. *Journal of Geophysical Research: Atmospheres*, **91**(D1), 1231–1247. doi: https://doi.org/10.1029/JD091iD01p01231. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JD091iD01p01231.

Feynman, R., Leighton, R., Sands, M. a Hafner, E. (1965). *The Feynman Lectures on Physics; Vol. I*, volume 33. AAPT.

Fix, E. a Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, **57**(3), 238–247. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403797.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, **36**(4), 193–202. ISSN 1432-0770. doi: 10.1007/BF00344251. URL https://doi.org/10.1007/BF00344251.

Kašpar, P., Kolmašová, I. a Santolík, O. (2022). Model of the first lightning return stroke using bidirectional leader concept. *Journal of Geophysical Research: Atmospheres*, **127**(24), e2022JD037459. doi: https://doi.org/10.1029/2022JD037459. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022JD037459. e2022JD037459 2022JD037459.

Kolmašová, I., Soula, S., Santolík, O., Farges, T., Bousquet, O., Diendorfer, G., Lán, R. a Uhlíř, L. (2022).

A frontal thunderstorm with several multi-cell lines found to produce energetic preliminary breakdown. *Journal of Geophysical Research: Atmospheres*, **127**(4), e2021JD035780. doi: https://doi.org/10.1029/2021JD035780. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JD035780. e2021JD035780 2021JD035780.

Kolmašová, I. a Santolík, O. (2013). Properties of unipolar magnetic field pulse trains generated by lightning discharges. *Geophysical Research Letters*, **40**(8), 1637–1641. doi: https://doi.org/10.1002/grl.50366. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50366.

Kolmašová, I., Santolík, O., Farges, T., Rison, W., Lán, R. a Uhlíř, L. (2014). Properties of the unusually short pulse sequences occurring prior to the first strokes of negative cloud-to-ground lightning flashes. *Geophysical Research Letters*, **41**(14), 5316–5324. doi: https://doi.org/10.1002/2014GL060913. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL060913.

Kolmašová, I., Santolík, O., Defer, E., Kašpar, P., Kolínská, A., Pedeboy, S. a Coquillat, S. (2020). Two propagation scenarios of isolated breakdown lightning processes in failed negative cloud-to-ground flashes. *Geophysical Research Letters*, **47**(23), e2020GL090593. doi: https://doi.org/10.1029/2020GL090593. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL090593. e2020GL090593 2020GL090593.

Kostinskiy, A. Y., Marshall, T. C. a Stolzenburg, M. (2020). The mechanism of the origin and development of lightning from initiating event to initial breakdown pulses (v.2). *Journal of Geophysical Research: Atmospheres*, **125**(22), e2020JD033191. doi: https://doi.org/10.1029/2020JD033191. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD033191. e2020JD033191 2020JD033191.

League, A. R. R. (1982). *The A. R. R. L. Antenna Book*. Number sv. 14 in Radio amateur's library. American Radio Relay League. ISBN 9780872594142. URL https://books.google.cz/books?id=6ohYAAAAYAAJ.

Liu, H. a Chandrasekar, V. (2000). Classification of hydrometeors based on polarimetric radar measurements: Development of fuzzy logic and neuro-fuzzy systems, and in situ verification. *Journal of Atmospheric and Oceanic Technology*, **17**(2), 140 – 164. doi: https://doi.org/10.1175/1520-0426(2000)017<0140:COHBOP>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atot/17/2/1520-0426_2000_017_0140_cohbop_2_0_co_2.xml.

Liu, N. Y., Scholten, O., Hare, B. M., Dwyer, J. R., Sterpka, C. F., Kolmašová, I. a Santolík, O. (2022). Lofar observations of lightning initial breakdown pulses. *Geophysical Research Letters*, **49**(6), e2022GL098073. doi: https://doi.org/10.1029/2022GL098073. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022GL098073. e2022GL098073 2022GL098073.

Lu, G., Cummer, S. A., Blakeslee, R. J., Weiss, S. a Beasley, W. H. (2012). Lightning morphology and impulse charge moment change of high peak current negative strokes. *Journal of Geophysical Research: Atmospheres*, **117** (D4). doi: https://doi.org/10.1029/2011JD016890. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JD016890.

Malý, P. (2020). *Optika*. Karolinum. ISBN 9788024622460. URL https://books.google.cz/books?id=sksEEAAAQBAJ.

Mansell, E. R., MacGorman, D. R., Ziegler, C. L. a Straka, J. M. (2005). Charge structure and lightning sensitivity in a simulated multicell thunderstorm. *Journal of Geophysical Research: Atmospheres*, **110**(D12). doi: https://doi.org/10.1029/2004JD005287. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JD005287.

Maslej-Krešňáková, V., Kundrát, A., Mackovjak, Š., Butka, P., Jaščur, S., Kolmašová, I. a Santolík, O. (2021). Automatic detection of atmospherics and tweek atmospherics in radio spectrograms based on a deep learning approach. *Earth and Space Science*, **8**(11), e2021EA002007. doi: https://doi.org/10.1029/2021EA002007. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021EA002007. e2021EA002007 2021EA002007.

Moroney, L. (2020). *AI and Machine Learning for Coders: A Programmer's Guide to Artificial Intelligence*. O'Reilly. ISBN 9781492078197. URL https://books.google.cz/books?id=462OzQEACAAJ.

Nag, A. a Rakov, V. A. (2010). Compact intracloud lightning discharges: 1. mechanism of electromagnetic radiation and modeling. *Journal of Geophysical Research: Atmospheres*, **115**(D20). doi: https://doi.org/10.1029/2010JD014235. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JD014235.

Nag, A. a Rakov, V. A. (2012). Positive lightning: An overview, new observations, and inferences. *Journal of Geophysical Research: Atmospheres*, **117** (D8). doi: https://doi.org/10.1029/2012JD017545. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JD017545.

Rakov, V. a Uman, M. (2007). *Lightning: Physics and Effects*. Cambridge University Press. ISBN 9781107268555. URL https://books.google.cz/books?id=Urz3CwAAQBAJ.

Redmon, J., Divvala, S., Girshick, R. a Farhadi, A. (2016). You only look once: Unified, real-time object detection.

Renaud, O. a Victoria-Feser, M.-P. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, **140**(7), 1852–1862. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2010.01.008. URL https://www.sciencedirect.com/science/article/pii/S0378375810000194.

S. C. Kleene, C. E. S. a McCarthy, J. (1956). *Representation of Events in Nerve Nets and Finite Automata*, pages 3–42. Princeton University Press, Princeton. ISBN 9781400882618. doi: doi:10.1515/9781400882618-002. URL https://doi.org/10.1515/9781400882618-002.

Shi, D., Wang, D., Wu, T. a Takagi, N. (2020). A comparison on the e-change pulses occurring in the bi-level polarity-opposite charge regions of the intracloud lightning flashes. *Journal of Geophysical Research: Atmospheres*, **125**(17), e2020JD032996. doi: https://doi.org/10.1029/2020JD032996. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD032996. e2020JD032996 2020JD032996.

Williams, E. a Stanfill, S. (2002). The physical origin of the land–ocean contrast in lightning activity. *Comptes Rendus Physique*, **3**(10), 1277–1292. ISSN 1631-0705. doi: https://doi.org/10.1016/S1631-0705(02)01407-X. URL https://www.sciencedirect.com/science/article/pii/S163107050201407X.

Willmott, C. J. a Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, **30**(1), 79–82.

Wu, B., Zhang, G., Wen, J., Zhang, T., Li, Y. a Wang, Y. (2016). Correlation analysis between initial preliminary breakdown process, the characteristic of radiation pulse, and the charge structure on the qinghai-tibetan plateau. *Journal of Geophysical Research: Atmospheres*, **121**(20), 12,434–12.459. doi: https://doi.org/10.1002/2016JD025281. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JD025281.

Wu, T., Takayanagi, Y., Funaki, T., Yoshida, S., Ushio, T., Kawasaki, Z.-I., Morimoto, T. a Shimizu, M. (2013). Preliminary breakdown pulses of cloud-to-ground lightning in winter thunderstorms in japan. *Journal of Atmospheric and Solar-Terrestrial Physics*, **102**, 91–98. ISSN 1364-6826. doi: https://doi.org/10.1016/j.jastp.2013.05.014. URL https://www.sciencedirect.com/science/article/pii/S1364682613001685.

Wu, T., Yoshida, S., Akiyama, Y., Stock, M., Ushio, T. a Kawasaki, Z. (2015). Preliminary breakdown of intracloud lightning: Initiation altitude, propagation speed, pulse train characteristics, and step length estimation. *Journal of Geophysical Research: Atmospheres*, **120**(18), 9071–9086. doi: https://doi.org/10.1002/2015JD023546. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JD023546.

Zhu, Y., Bitzer, P., Rakov, V. a Ding, Z. (2021). A machine-learning approach to classify cloud-to-ground and intracloud lightning. *Geophysical Research Letters*, **48**(1), e2020GL091148. doi: https://doi.org/10.1029/2020GL091148. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL091148. e2020GL091148 2020GL091148.

Ziegler, C. L., MacGorman, D. R., Dye, J. E. a Ray, P. S. (1991). A model evaluation of noninductive graupel-ice charging in the early electrification of a mountain thunderstorm. *Journal of Geophysical Research: Atmospheres*,

96(D7), 12833–12855. doi: https://doi.org/10.1029/91JD01246. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/91JD01246.

Zoznam obrázkov

Zoznam použitých skratiek

ML Milešovka

DL Dlouhá louka

LS Lomnický Štít

KR Krupka

CG Cloud-to-ground

GC Ground-to-cloud

IC Intracloud

NBE Narrow bipolar event

SVM Support vector machine

KNN K-nearest neighbours

YOLO You only look once

R2 Koeficient determinácie

RMSE Root-mean-square deviation

ACC Accuracy

PRC Precision