

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky

**Predikcia geomagnetických búrok
pomocou hlbokého učenia**

Diplomová práca

2023

Bc. Veronika Motúzová

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky

**Predikcia geomagnetických búrok
pomocou hlbokého učenia**

Diplomová práca

Študijný program: Hospodárska informatika
Študijný odbor: Informatika
Školiace pracovisko: Katedra kybernetiky a umelej inteligencie (KKUI)
Školiteľ: doc. Ing. Peter Butka, PhD.
Konzultanti: Ing. Viera Maslej Krešňáková, PhD.
 RNDr. Šimon Mackovjak, PhD.

Košice 2023

Bc. Veronika Motúzová

Abstrakt v SJ

Diplomová práca sa venuje téme z oblasti dátovej analytiky, a to predikcii geomagnetických búrok pomocou metód hlbokého učenia. Geomagnetické búrky sú jedným z najvýraznejších nebezpečných prejavov kozmického počasia. Vznikajú ako dôsledok narušenia magnetosféry Zeme energetickými časticami, ktoré unikajú zo Slnka. Cieľom diplomovej práce je navrhnúť model, založený na strojovom učení s využitím hlbokých neurónových sietí, ktorý by bol schopný poskytovať predikcie indexov kozmického počasia s niekoľko hodinovým predstihom. Hlavnou metódou využívanou v tejto práci je rekurentná neurónová sieť. Modely rekurentných neurónových sietí sú implementované v programovacom jazyku Python a evaluované pomocou štandardných metrík pre vyhodnocovanie klasifikačných modelov.

Klúčové slová

predikcia časových radov, predikcia geomagnetických búrok, hlboké učenie, rekurentná neurónová sieť, DST index, kozmické počasie, OMNI dataset

Abstrakt v AJ

The diploma thesis deals with a topic from the field of data analytics, namely the prediction of geomagnetic storms using deep learning methods. Geomagnetic storms are one of the most prominent and dangerous manifestations of space weather. They arise as a result of the disruption of the Earth's magnetosphere by energetic particles escaping from the Sun. The aim of the thesis is to design a model based on machine learning using deep neural networks, which would be able to provide predictions of space weather indices several hours in advance. The main method used in this work is a recurrent neural network. Models of recurrent neural networks are implemented in the Python programming language and evaluated using standard metrics for evaluating classification models.

Klúčové slová v AJ

Time series prediction, geomagnetic storm prediction, deep learning, recurrent neural network, DST index, space weather, OMNI dataset

68259

TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY
Katedra kybernetiky a umelej inteligencie

**ZADANIE
DIPLOMOVEJ PRÁCE**

Študijný odbor: **Informatika**
Študijný program: **Hospodárska informatika**

Názov práce:

Predikcia geomagnetických búrok pomocou hlbokého učenia

Prediction of geomagnetic storms by deep learning

Študent: **Bc. Veronika Motúzová**
Školiteľ: **doc. Ing. Peter Butka, PhD.**
Školiace pracovisko: **Katedra kybernetiky a umelej inteligencie**
Konzultant práce: **RNDr. Šimon Mackovjak, PhD., Ing. Viera Maslej
Krešňáková, PhD.**
Pracovisko konzultanta: **Ústav experimentálnej fyziky SAV**

Pokyny na vypracovanie diplomovej práce:

1. Podať teoretický prehľad problematiky metód hlbokého učenia, s dôrazom na vhodné metódy pre predikciu geomagnetických búrok z časového radu dát.
2. Získať a predspracovať dátá pre predikciu geomagnetických búrok.
3. Navrhnuť, realizovať a vyhodnotiť experimenty s predikciou geomagnetických búrok pomocou hlbokého učenia na pripravenej množine dát.
4. Vypracovať dokumentáciu podľa pokynov katedry a vedúceho práce.

Jazyk, v ktorom sa práca vypracuje: slovenský
Termín pre odovzdanie práce: 21.04.2023
Dátum zadania diplomovej práce: 31.10.2022




prof. Ing. Liberios Vokorokos, PhD.
dekan fakulty

Čestné vyhlásenie

Vyhlasujem, že som diplomovú prácu vypracovala samostatne s použitím uvedenej odbornej literatúry.

Košice 21. 4. 2023

.....

Vlastnoručný podpis

Podakovanie

Ďakujem vedúcemu mojej diplomovej práce doc. Ing. Petrovi Butkovi, PhD. a mojim konzultantom pani Ing. Viere Maslej Krešňákovej, PhD., RNDr. Šimonovi Mackovjakovi, PhD. a Ing. Lenke Kališkovej za cenné rady, pripomienky a čas, ktorý venovali dohľadu nad mojou prácou. Zároveň ďakujem aj priateľovi, rodine a mojim najbližším za podporu počas celého môjho štúdia.

Obsah

Úvod	1
1 Hlboké učenie pre predikciu časových radov	2
1.1 Úvod do strojového učenia	2
1.2 Úvod do hlbokého učenia a neurónových sietí	4
1.2.1 Definícia neurónovej siete	6
1.2.2 Topológia neurónovej siete	7
1.2.3 Aktivačné funkcie	8
1.2.4 Chybové funkcie	12
1.2.5 Optimalizácia neurónových sietí	15
1.2.6 Regularizácia neurónových sietí	17
1.3 Rekurentná neurónová siet	19
1.4 Predikcia časových radov	22
1.5 Metriky pre výhodnocovanie modelov	24
1.6 Použité technológie	29
2 Predikcia geomagnetickej búrky	32
2.1 Geomagnetická búrka	32
2.2 Analýza súčasného stavu	37
3 Praktická časť	40
3.1 Pochopenie cieľa	41
3.2 Pochopenie dát	42
3.3 Príprava dát	46
3.4 Modelovanie	48
3.5 Vyhodnotenie	60
3.6 Nasadenie	63
Zoznam príloh	69

Zoznam obrázkov

1–1	Vzťah medzi AI, ML a DL	2
1–2	Delenie, metódy a algoritmy ML	5
1–3	Rozdiel medzi ML a DL	6
1–4	Neurónová sieť v ľudskom mozgu	7
1–5	Typy vrstiev v neurónovej sieti	7
1–6	Dopredná a rekurentná sieť	8
1–7	Graf Sigmoid funkcie	10
1–8	Graf ReLU funkcie	10
1–9	Graf <i>Tanh</i> funkcie	11
1–10	Graf MSE chybovej funkcie	13
1–11	Porovnanie grafov MSE (modrá) a MAE (červená) krivka	14
1–12	Plne prepojená NN s vyznačeným aktivačným tokom signálu a v opačnom smere spätné šírenie chýb	15
1–13	Podučený, optimálny a preučený model	17
1–14	Efekt Dropout funkcie	18
1–15	ROC krivka a plocha AUC pod ňou	28
2–1	Vplyv výronu koronálnej hmoty na zemskú magnetosféru	32
2–2	DST index počas priebehu búrky	35
2–3	Príklad nameraného DST indexu a úrovne intenzity geomagnetickej búrky	35
2–4	Výstup z aplikácie <i>DstLive</i>	39
3–1	Hodnoty hlavného atribútu DST v čase	42
3–2	Chýbajúce hodnoty v datasete	43
3–3	Heatmapa korelácií atribútov s DST	44
3–4	Hodnoty atribútov DST, F10_index a BZ_GSM v čase	45
3–5	Rozdelenie dátovej množiny	47
3–6	Architektúra RNN s veľkosťou vstupov na jednotlivých vrstvách .	49

3–7	Sumarizácia modelu	49
3–8	Vyhodnotenie TPR, TNR, PPV a NPV experimentu 1	51
3–9	Vyhodnotenie ACC a MCC experimentu 1	52
3–10	Predikované a skutočné hodnoty DST indexu experimentu 1	53
3–11	Predikované a skutočné hodnoty DST+6 experimentu 2	56
3–12	Predikované a skutočné hodnoty DST indexu experimentu 3	59
3–13	Predikcie oboch modelov v porovnaní so skutočnými hodnotami . .	61

Zoznam tabuliek

1–1	Kontingenčná tabuľka	25
2–1	Úrovne intenzity geomagnetických búrok a ich vplyvy	36
2–2	Výsledky výskumu Gruet et al. (2018)	38
3–1	Štatistické informácie o cieľovom atribúte DST index	42
3–2	Kontingenčná tabuľka modelu s časovým oknom 6 hodín	52
3–3	Výsledky metrík modelu s časovým oknom 6 hodín	52
3–4	Vizualizácie metrík z experimentu 2	55
3–5	Kontingenčná tabuľka modelu predikujúceho DST+6	55
3–6	Výsledky metrík modelu predikujúceho DST+6	56
3–7	Vizualizácie metrík z experimentu 3	58
3–8	Kontingenčná tabuľka modelu predikujúceho DST+6 pomocou pridaných atribútov	59
3–9	Výsledky metrík modelu predikujúceho DST+6 pomocou pridaných atribútov	59
3–10	Sumarizácia výsledkov pre predikciu so 6 hodinovým predstihom	60
3–11	Výsledky Gruet et al. (2018) v porovnaní s vlastnými výsledkami	62
3–12	Kontingenčná tabuľka prispôsobenej klasifikácie do 3 tried	62
3–13	Výsledky metrík prispôsobenej klasifikácie do 3 tried	63

Zoznam symbolov a skratiek

ACC z angl. *accuracy*, úspešnosť klasifikácie

Adam z angl. *Adaptive Moment Estimation*, optimalizačná funkcia

AI z angl. *Artificial Intelligence*, umelá inteligencia

ARIMA z angl. *Autoregressive Integrated Moving Average*, autoregresívna metóda

AUC z angl. *Area Under the Curve*, plocha pod ROC krivkou

BCE z angl. *Binary Crossentropy*, binárna krížová entropia

CCE z angl. *Categorical Crossentropy*, kategorická krížová entropia

CME z angl. *Coronal Mass Ejection*, výrony koronálnej hmoty zo Slnka

CNN z angl. *Convolutional Neural Network*, konvolučná neurónová sieť

DL z angl. *Deep Learning*, hlboké učenie

DST index z angl. *Disturbance Storm Time index*, časový index rušenia búrk

FDR z angl. *False Discovery Rate*, miera falošných objavov

FN z angl. *False Negative*, falošne negatívne

FNR z angl. *False Negative Rate*, miera falošnej negativity

FP z angl. *False Positive*, falošne pozitívne

FPR z angl. *False Positive Rate*, miera falošnej pozitivity

GPUs z angl. *Graphics processing units*, grafické procesorové jednotky

GRU z angl. *Gated Recurrent Unit*, typ rekurentnej neurónovej siete

GSM z angl. *Geocentric Solar Magnetospheric*, súradnicový systém

LSM z angl. *Liquid State Machine*, typ rekurentnej neurónovej siete

LSTM z angl. *Long Short-Term Memory*, typ rekurentnej neurónovej siete

MAE z angl. *Mean Absolute Error*, stredná absolútна chyba

MCC z angl. *Matthews Correlation Coefficient*, korelačný koeficient

ML z angl. *Machine Learning*, strojové učenie

MSE z angl. *Mean Squared Error*, stredná kvadratická chyba

NLP z angl. *Natural Language Processing*, spracovanie prirodzeného jazyka

NN z angl. *Neural Network*, neurónová sieť

NPV z angl. *Negative Predictive Value*, negatívna prediktívna hodnota

OMNI z angl. *Outer Magnetospheric Multiscale Mission Integrated Science dataset*

PPV z angl. *Positive Predictive Value*, pozitívna prediktívna hodnota

ReLU z angl. *Rectified Linear Unit*, retifikovaná lineárna jednotka

RMSprop z angl. *Root Mean Square Propagation*, optimalizačná funkcia

RNN z angl. *Recurrent Neural Network*, rekurentná neurónová siet

ROC z angl. *Receiver Operating Characteristic*, názov krivky

SAV Slovenská akadémia vied

SGD z angl. *Stochastic Gradient Descent*, stochastická gradientová metóda

TN z angl. *True Negative*, pravdivo negatívne

TNR z angl. *True Negative Rate*, miera pravdivej negativity

TP z angl. *True Positive*, pravdivo pozitívne

TPR z angl. *True Positive Rate*, miera pravdivej pozitivity

Úvod

Geomagnetické búrky sú prírodné javy, ktoré majú výrazný negatívny vplyv na mnohé aspekty nášho života. Tieto búrky vznikajú v dôsledku zmien v magnetickom poli Zeme a môžu ovplyvniť fungovanie technologických systémov, ako sú elektroenergetické siete, komunikačné siete alebo satelitné navigačné systémy. Preto je klúčové mať k dispozícii spolahlivé nástroje na predpovedanie týchto búrok s dostatočným predstihom, aby sa mohli prijať opatrenia na minimalizáciu ich negatívneho vplyvu.

V posledných rokoch sa metódy hlbokého učenia stávajú čoraz populárnejšími, a to nielen v oblasti predikcie geomagnetických búrok. Hlavnou metódou používanou v tejto práci bude rekurentná neurónová sieť, ktorá prišla ako požiadavka spolu s téhou práce z Ústavu experimentálnej fyziky Slovenskej akadémie vied, v.v.i. V tejto práci bude našim cieľom navrhnúť, natrénovať a vyhodnotiť model rekurentnej neurónovej siete, ktorý bude schopný predikovať výskyt geomagnetických búrok s čo najvyššou presnosťou a niekoľko hodinovým časovým predstihom na základe datasetu časového radu historických hodnôt indexov kozmického počasia.

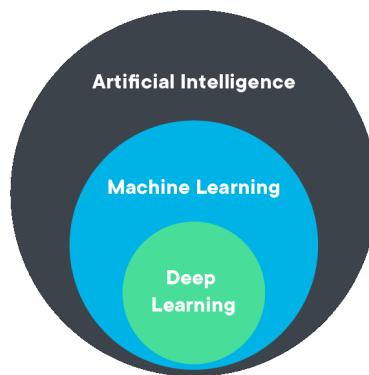
V Kapitole 1 sa zaoberáme úvodom do strojového a hlbokého učenia, teóriou neurónových sietí. Nájdeme tu stručné vysvetlenia potrebných pojmov, ktoré ďalej využívame v praktickej časti. Kapitola obsahuje napríklad definíciu neurónovej siete, jej topológiu, jej aktivačné, chybové a optimalizačné funkcie, a napokon regularizáciu neurónových sietí. Hlbšie je vysvetlený aj princíp rekurentnej neurónovej siete, ktorú v práci využívame. Popísané sú aj metriky vyhodnocovania úspešnosti predikčných a klasifikačných modelov, aby sme porozumeli dosiahnutým výsledkom. Kapitola je ukončená popisom použitých technológií. V Kapitole 2 sa bližšie pozrieme na kozmické počasie a jeho prejavy, konkrétnie geomagnetické búrky. Vysvetlíme si príčiny ich vzniku a možné následky ich vplyvu podľa intenzity. Pozrieme sa na analýzu súčasného stavu problematiky, čiže ako boli riešené práce s podobnou téhou a aké výsledky dosiahli. V Kapitole 3 sa nachádza praktická časť s vyhodnotenými experimentami.

1 Hlboké učenie pre predikciu časových radov

Táto kapitola poskytne stručný teoretický náhľad do oblasti strojového učenia a odvetvia hlbokého učenia. Priblížená je tu zároveň aj problematika predikcie časových radov, sú tu vysvetlené potrebné metriky na vyhodnocovanie predikčných modelov, ktoré budeme ďalej používať v praktickej časti. V závere kapitoly sú spomenuté aj použité technológie.

1.1 Úvod do strojového učenia

Pojem strojové učenie (angl. *Machine Learning*, skr. **ML**) združuje všetky metódy, ktoré automatizujú proces samostatného učenia výpočtových strojov. Ide o odvetvie umelej inteligencie (angl. *Artificial Intelligence*, skr. **AI**). AI sa zaoberá vývojom algoritmov a techník pre stroje. Stroje by následne mali byť schopné rozpoznávať dôležité príznaky a predpovedať na základe zadaných dát. Hlboké učenie (angl. *Deep Learning*, skr. **DL**) je pokročilá metóda ML, ktorá využíva algoritmy zvané neurónové siete. Vzájomný vzťah medzi AI, ML a DL zobrazuje diagram na Obrázku 1 – 1¹.



Obrázok 1 – 1: Vzťah medzi AI, ML a DL

Cieľom ML je využiť spomínané algoritmy a techniky na riešenie komplexných problémov a úloh v praxi. ML má široké využitie v rôznych oblastiach. Riešená

¹Zdroj Obrázka 1 – 1: www.flatironschool.com/blog/deep-learning-vs-machine-learning/

problematika môže byť napríklad spracovanie a rozpoznávanie obrazu, rozpoznanie textu a reči, či predikovanie hodnôt v oblasti finančníctva, ekonómie alebo aj v medicíne. ML sa čoraz viac uplatňuje aj v priemyselnej a výskumnej oblasti, stáva sa tak neoddeliteľnou súčasťou moderného sveta.

Bežné algoritmy strojového učenia sú založené na matematických a štatistických metódach, napríklad regresia, klasifikácia, zhľukovanie, rozhodovacie stromy a neurónové siete. Tieto metódy umožňujú výpočtovým strojom nájsť v dátach skryté vzťahy a súvislosti, ktoré sú pre ľudský mozog len ťažko rozpoznateľné. ML je učený proces, ktorý sa obvykle skladá z nasledovných krokov: získanie a pochopenie dát, príprava dát, návrh a trénovanie modelu, vyhodnotenie modelu a nasadenie v praxi.

Podľa Machová (2002) delíme strojové učenie z hľadiska viacerých aspektov nasledovne:

1. vo všeobecnosti sa delí na:

- induktívne - z informácií na najnižšej úrovni všeobecnosti indukuje, respektíve generuje všeobecnejšiu znalosť,
- deduktívne - naopak zo znalosti na vysokej úrovni všeobecnosti dedukuje znalosť menej všeobecnú a zložitú, ktorá je lepšie prispôsobená na riešenie konkrétneho druhu problémov,

2. podľa stupňa kontroly:

- kontrolované - disponuje spätnou väzbou o úspešnosti učenia,
- nekontrolované - neexistuje spätná väzba o vhodnosti výkonu,

3. na základe spôsobu, akým sa získavajú trénovacie príklady:

- úlohy typu online - získavajú trénovacie príklady jednotlivo a postupne,
- úlohy typu offline - disponujú všetkými trénovacími príkladmi súčasne,

4. podľa postupu v učení:

- inkrementálne - je učenie pomocou algoritmu, ktorý spracováva jeden trénovací príklad za druhým, po každom príklade poskytuje algoritmus riešenie, zodpovedá online učiacim úlohám,
 - neinkrementálne - je učenie pomocou algoritmu, ktorý spracováva celú množinu trénovacích príkladov odrazu, zodpovedá offline učiacim úlohám.
- (Machová, 2002)

Diagram na Obrázku 1–2 nám prehľadne zobrazuje delenie strojového učenia v praxi a ku každému typu učenia sú zobrazené aj príklady algoritmov, ktoré sa najčastejšie využívajú. V spodnej časti diagramu je odvetvie hlbokého učenia so svojim rozdelením, na ktoré sa v nasledujúcej podkapitole pozrieme detailnejšie.

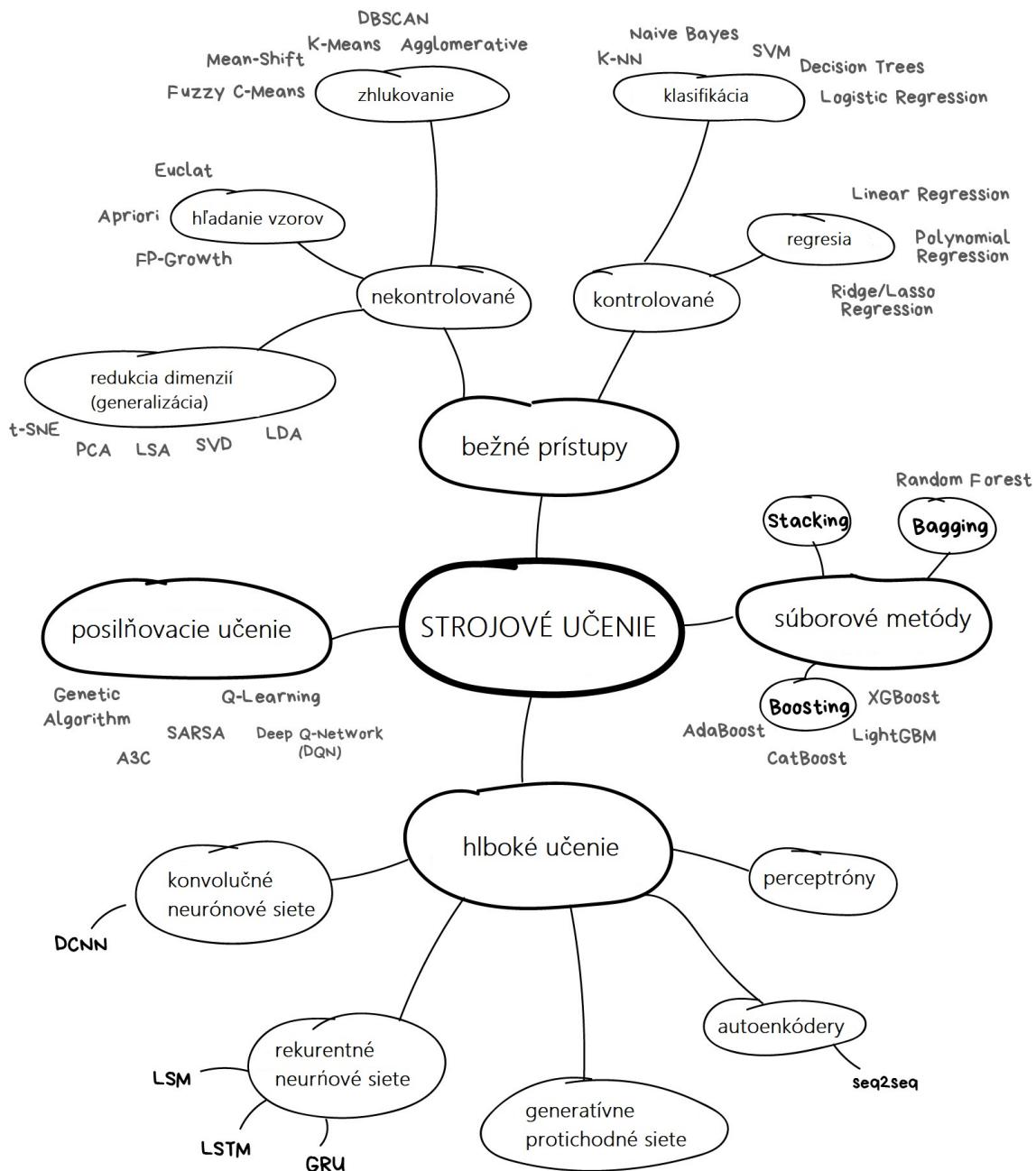
1.2 Úvod do hlbokého učenia a neurónových sietí

DL je pokročilá metóda strojového učenia, ktorá využíva algoritmy zvané neurónové siete. ML sa zameriava na trénovanie algoritmov, ktoré môžu analyzovať dátá, identifikovať vzory a naučiť sa predikovať budúce výstupy na základe histórie vstupných dát. To znamená, že ML sa zameriava na učenie sa z jednoduchších, plochých dát a modelov.

Na druhej strane, DL je technika strojového učenia, ktorá sa zameriava na trénovanie neurónových sietí s viacerými vrstvami, ktoré vyžadujú hlbšie porozumenie vstupných dát. DL umožňuje strojom naučiť sa z abstraktnejších, hierarchických dát a modelov.

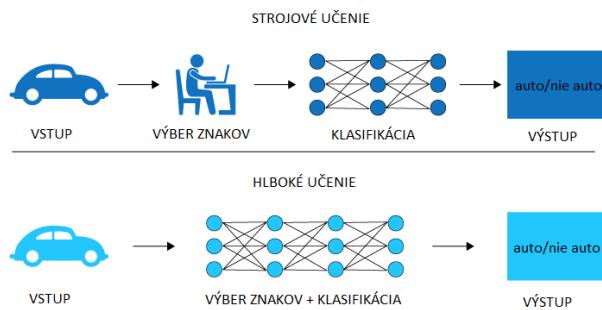
Podstatou DL je aj fakt, že stroj by mal byť schopný sám rozpoznať príznaky dôležité pre jeho učenie, narozdiel od iných metód ML, kde je potrebné príznaky na učenie vybrať manuálne - človekom.

V podstate teda DL je špecifická technika ML, ktorá využíva neurónové siete s viacerými vrstvami na riešenie komplexnejších úloh.



Obrázok 1 – 2: Delenie, metódy a algoritmy ML

Tento fakt tvorí hlavný rozdiel medzi ML a DL, znázorňuje ho aj Obrázok 1 – 3. DL je stále aktívnym výskumným odborom, kde sa neustále objavujú zlepšenia a inovácie.



Obrázok 1 – 3: Rozdiel medzi ML a DL

1.2.1 Definícia neurónovej siete

DL využíva neurónové siete, čo sú matematické modely pozostávajúce z vrstiev prepojených neurónov. Každý neurón prijíma vstupné dátá, tie sú vážené a spracované pomocou aktivačnej funkcie, čím vytvára výstupné dátá, ktoré sa odosielajú neurónom v ďalšej vrstve. Neurónová sieť (angl. *Neural network*, skr. **NN**) je schopná rozpoznať príznaky v dátach, na ich základe vie klasifikovať alebo predikovať dátá podobne ako ľudský mozog. Ľudský mozog sa skladá z biologických sietí nervových buniek – neurónov, viď Obrázok 1 – 4².

NN môžu mať rôzne architektúry a využívať sa na riešenie rôznych úloh. Využitie NN je široké, vie sa naučiť rozpoznať obrazové, zvukové i textové dátá. Trénovanie NN je výpočtovo náročné a vyžaduje množstvo vstupných dát a výkonných výpočtových zdrojov. Podľa Sinčák and Andrejková (1996) definícia NN znie: „Neurónová sieť je masívne paralelný procesor, ktorý má sklon k uchovávaniu experimentálnych znalostí a ich ďalšieho využívania. Napodobňuje ľudský mozog v dvoch aspektoch: poznatky sú zbierané v NN počas učenia a medzineurónové spojenia (synaptické váhy) sú využívané na ukladanie znalostí.“

²Zdroj Obrázka 1 – 4: https://mb.cision.com/Public/11572/2497645/b7e3494b8cc0b4af_800x800ar.jpg

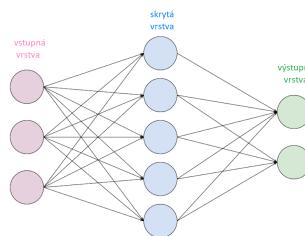


Obrázok 1 – 4: Neurónová sieť v ľudskom mozgu

1.2.2 Topológia neurónovej siete

Topológia, inak tvar alebo štruktúra, NN by podľa Sinčák and Andrejková (1996) mala byť popísateľná ľubovoľným orientovaným grafom, kde by vrcholy predstavovali neuróny a hrany synapsie medzi nimi. Topológia NN je najlepšie pochopiteľná na pravidelnej viacvrstvovej štruktúre, ktorá je zložená z nasledovných typov vrstiev, viď. Obrázok 1 – 5 :

- **vstupná vrstva** (angl. *input layer*), ktorej neuróny dostávajú vstup len z vonkajšieho sveta a výstup obvykle pokračuje k ďalším neurónom,
- **skryté vrstvy** (angl. *hidden layers*), v ktorých neuróny dostávajú vstup z ostatných neurónov alebo aj z externého sveta cez prahové prepojenia a ich výstupy pokračujú ďalej do NN,
- **výstupná vrstva** (angl. *output layer*), je podobná ako skrytá vrstva, iba že výstup z tejto vrstvy vyúsťuje do externého sveta.



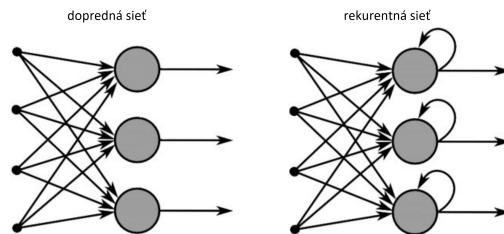
Obrázok 1 – 5: Typy vrstiev v neurónovej sieti

Typy neurónov analogicky rozlišujeme podľa toho, v ktorej vrstve sa nachádzajú na vstupné, skryté a výstupné neuróny (Sinčák and Andrejková, 1996).

Podľa smeru šírenia signálu rozlišujeme 2 základné typy NN:

- **dopredné NN** (angl. *Feed-Forward NN*), v ktorých sa signál šíri len jedným smerom,
- **rekurentné NN** (angl. *Recurrent NN*), kde sa signál šíri zložitejšie – oboma smermi, preto je ľahšie rozlišovať vstupnú a výstupnú vrstvu.

Rozdiel medzi týmito dvoma typmi je očividný na Obrázku 1 – 6. Dokonca aj v rámci jednotlivých neurónov sa signál môže šíriť viacerými spôsobmi, a to buď **synchrónne**, **blok-synchrónne**, **sekvenčne** alebo **asynchónne**. Zamyslenie nad vyššie uvedenými skutočnosťami ohľadom štruktúry NN viedie k uvedomieniu si zložitosti fungovania NN (Sinčák and Andrejková, 1996).



Obrázok 1 – 6: Dopredná a rekurentná siet

1.2.3 Aktivačné funkcie

Podľa Sharma (2017) sú aktivačné funkcie dôležitou súčasťou neurónových sietí, pretože určujú, ako budú neuróny reagovať na vstupné signály a ako sa prenesie táto informácia do ďalších vrstiev siete.

Aktivačné funkcie v neurónových sietach slúžia na transformáciu vstupného signálu na výstupný v rámci každej vrstvy. Signál do vrstvy neurónov vstúpi, transformuje sa pomocou aktivačnej funkcie a spracovaný postupuje do nasledujúcej vrstvy.

Neurónová sieť by bez aktivačných funkcií fungovala ako lineárny regresný model s obmedzeným výkonom. Aktivačná funkcia teda zabezpečuje jej nelineárnosť,

čo umožňuje NN sa učiť na komplikovaných typoch dát, ako sú obrázky, videá, zvuk, reč a text. NN sa považuje za univerzálny approximátor funkcií, vie vyriešiť aj najzložitejšiu nelineárnu funkciu.

Dôležitou vlastnosťou aktivačnej funkcie je, že musí byť differencovateľná, aby sme mohli spätnými krokmi optimalizovať NN za účelom zvýšenia jej presnosti. Podľa typu úlohy je potrebné zvoliť vhodné aktivačné funkcie pre jednotlivé vrstvy.

Existuje niekoľko typov aktivačných funkcií, ktoré sa používajú v neurónových sietach. Medzi najbežnejšie patria Sigmoid, rektifikovaná lineárna jednotka (angl. *Rectified Linear Unit*, skr. **ReLU**), hyperbolický tangens (skr. **Tanh**), Softmax, funkcia binárneho kroku, lineárna funkcia a parametrizovaná ReLU. Každá aktivačná funkcia má svoje vlastnosti a výhody, ktoré určujú jej použitie v konkrétnom kontexte, v závislosti od úlohy a architektúry siete.

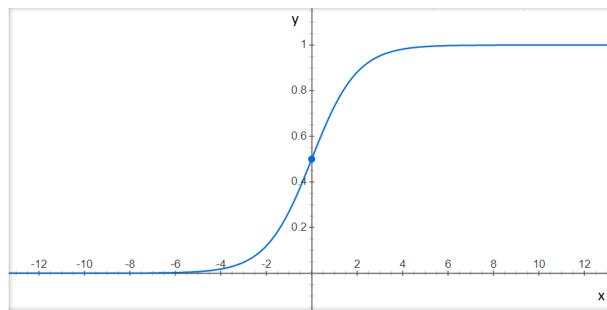
Sigmoid

Ide o najbežnejšie používanú nelineárnu aktivačnú funkciu. Používa sa často na klasifikáciu binárnych dát. Funkcia Sigmoid transformuje hodnoty na rozsah $\langle 0, 1 \rangle$. Hodnota 0,5 sa považuje za hraničnú.

Sigmoid je neustále differencovateľná a spojitá funkcia v tvare písmena S (Sharma, 2017). Ako príklad sa najčastejšie uvádzá logistická funkcia. Matematický predpis tejto logistickej funkcie je daný:

$$y = \frac{1}{1 + e^{-x}}, \quad (1.1)$$

a jej krivka je zobrazená na Obrázku 1–7.



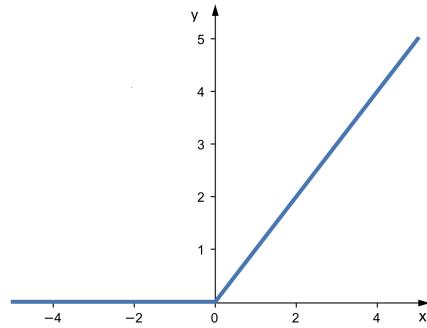
Obrázok 1 – 7: Graf Sigmoid funkcie

Rektifikovaná lineárna jednotka

ReLU je momentálne svetovo najpoužívanejšou aktivačnou funkciou, pretože sa používa takmer vo všetkých konvolučných neurónových sietach. Vstup, ktorý je menší ako 0, sa nastaví na 0, zatiaľ čo vstup väčší ako 0 sa ponechá bez zmeny. Vo väčšine prípadov funguje oveľa lepšie ako ostatné aktivačné funkcie. ReLU je opísaná rovnicou:

$$y = \max(0, x). \quad (1.2)$$

Priebeh funkcie popisuje graf na Obrázku 1 – 8, z ktorého vidíme, že záporné hodnoty premennej x nadobúdajú vždy nulovú funkčnú hodnotu. Výhodou ReLU je, že derivácia na intervale $(0, \infty)$ je 1. Preto ReLU dobre slúži na spätné vylepšovanie jednotlivých váh neurónov (Sharma, 2017). Využíva sa najmä v skrytých vrstvách NN.

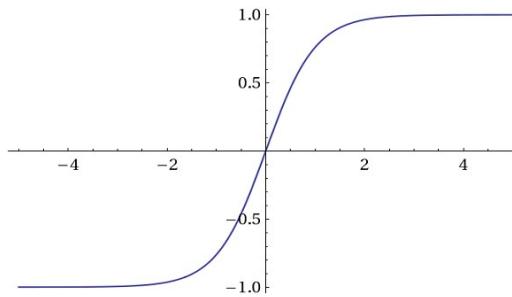


Obrázok 1 – 8: Graf ReLU funkcie

Hyperbolický tangens

Aktivačná funkcia Tanh sa podobne ako ReLU používa v skrytých vrstvách NN. Má podobný tvar ako funkcia Sigmoid – v tvare písmena S. Používa sa hlavne pri klasifikácii dát, ktoré majú viac ako dve triedy. Graf tejto funkcie vidíme na Obrázku 1 – 9³. Nadobúda hodnoty v rozsahu $\langle -1, 1 \rangle$ (Sharma, 2017). Je popísaná rovnicou:

$$y = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1. \quad (1.3)$$



Obrázok 1 – 9: Graf *Tanh* funkcie

Softmax

Aktivačná funkcia Softmax je zložená z viacerých funkcií typu Sigmoid. Vďaka tomu poskytuje nie len binárnu klasifikáciu ale aj klasifikáciu do viacerých tried. Funkcia nadobúda, rovnako ako Sigmoid, hodnoty v rozmedzí $\langle 0, 1 \rangle$. Tieto hodnoty predstavujú pravdepodobnosť zaradenia do danej triedy. Objekt nakoniec zaradí do triedy, ktorá má najväčšiu pravdepodobnosť. Matematický zápis funkcie znie:

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^K e^{x_j}}, \quad (1.4)$$

pre $i = 1, \dots, K$, pričom K predstavuje počet tried klasifikácie. Funkcia Softmax sa používa vo výstupnej vrstve NN a počet neurónov v tejto vrstve zodpovedá počtu tried klasifikácie (Sharma, 2017).

³Zdroj Obrázka 1 – 9: www.quora.com/What-is-the-range-of-tanh-activation-function

1.2.4 Chybové funkcie

Chybová funkcia je kritérium, ktoré určuje, ako dobre sa neurónová sieť naučila danú úlohu. Zároveň je dôležitým komponentom NN. Cieľom trénovania neurónovej siete je minimalizovať hodnotu chybovej funkcie. Existuje mnoho rôznych chybových funkcií, ktoré sa používajú v závislosti od typu úlohy, ktorú siet rieši. Čím menšiu hodnotu nadobúda chybová funkcia, tým presnejší je model NN a naopak.

Podľa Janocha and Czarnecki (2017) niektoré z najbežnejších chybových funkcií v neurónových sieťach sú:

- stredná kvadratická chyba (angl. *Mean Squared Error*, skr. **MSE**),
- krížová entropia (angl. *Cross-Entropy*),
- stredná absolútна chyba (angl. *Mean Absolute Error*, skr. **MAE**).

Ďalšie typy chybových funkcií sú napr. binárna krížová entropia (skr. BCE), kategorická krížová entropia (skr. CCE), logistická chybová funkcia, Huberova chyba (angl. *Huber loss*), chyba log-cosh a iné (Janocha and Czarnecki, 2017). Existuje mnoho ďalších chybových funkcií, ktoré sa používajú v neurónových sieťach v závislosti od typu úlohy a výstupu siete. Výber správnej chybovej funkcie môže výrazne ovplyvniť výkonnosť siete. V praktickej časti práce je využívaná MSE a MAE.

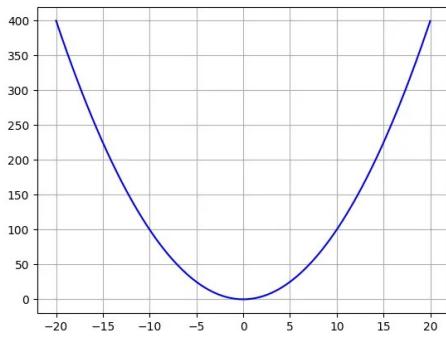
Stredná kvadratická chyba

Stredná kvadratická chyba je inak nazývaná aj priemerná kvadratická odchýlka. Táto funkcia sa používa na regresné úlohy a vypočíta rozdiel medzi skutočnými a predpokladanými hodnotami. Patrí medzi najbežnejšie chybové funkcie.

MSE sa vypočíta ako súčet kvadratických rozdielov medzi skutočnými a predpokladanými hodnotami. MSE nikdy nebude záporná, pretože chyby vždy umocňujeme na druhú. Graf tejto chybovej funkcie sa nachádza na Obrázku 1–10. Matematická definícia tejto funkcie je:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (1.5)$$

kde N je veľkosť testovacej množiny, y_i je skutočná hodnota, \hat{y}_i je predpokladaná hodnota (Janocha and Czarnecki, 2017).



Obrázok 1 – 10: Graf MSE chybovej funkcie

Stredná absolútна chyba

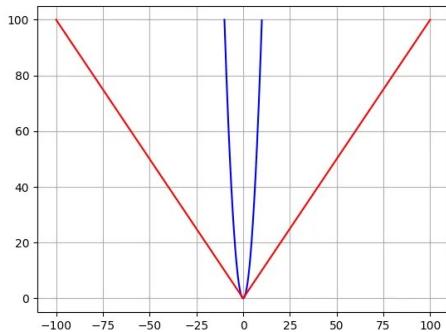
Stredná absolúttna chyba alebo stredná absolúttna odchýlka, táto funkcia sa používa na regresné úlohy a vypočíta rozdiel medzi skutočnými a predpokladanými hodnotami. MAE sa vypočíta ako súčet absolútnych hodnôt rozdielov medzi skutočnými a predpokladanými hodnotami podľa vzorca :

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|, \quad (1.6)$$

kde n je veľkosť testovacej množiny, y_j je skutočná hodnota, \hat{y}_j je predpokladaná hodnota (Janocha and Czarnecki, 2017). Na Obrázku 1 – 11 je zobrazené porovnanie grafov MSE (modrá krivka) a MAE (červená krivka).

Algoritmus spätného šírenia chýb v neurónových sieťach

Algoritmus spätného šírenia chýb (angl. *Backpropagation algorithm*) je podľa Wyt-hoff (1993) jeden z najdôležitejších algoritmov v tréningu neurónových sietí. Jeho cieľom je minimalizovať chybu predikcie siete tým, že sa upravia váhy medzi neurónmi. Tento algoritmus je najzákladnejším princípom v NN. Vystihuje podstatu



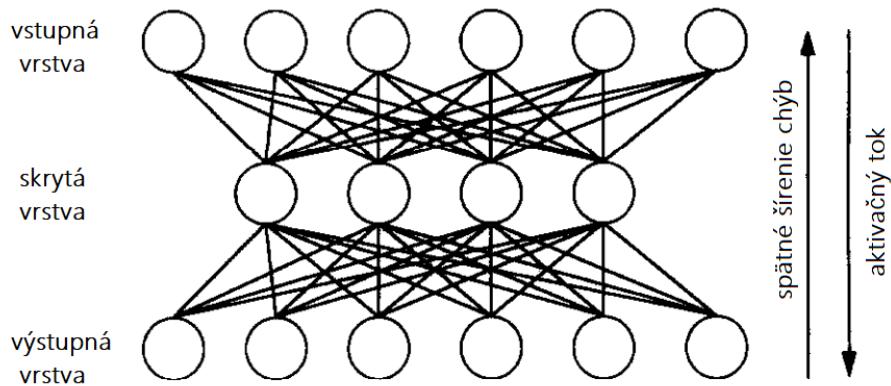
Obrázok 1 – 11: Porovnanie grafov MSE (modrá) a MAE (červená) krivka

učenia NN. Počas trénovania sa upravujú váhy a prahy medzi jednotlivými neurónmi, aby sa minimalizovala chybosť siete a zlepšila sa jej schopnosť vytvárať presné výstupy. Algoritmus sa používa na efektívne trénovanie neurónovej siete pomocou takzvaného reťazového pravidla. V krátkosti vysvetlené, po každom prechode signálu sieťou sa vykoná spätné šírenie chýb a úprava váh jednotlivých neurónov. Pri spätnom šírení chýb hrá dôležitú úlohu práve zvolená chybová funkcia NN.

Algoritmus spätného šírenia chýb pracuje nasledovne:

1. Siet spracuje dátá na vstupe a vykoná sa dopredný prechod, kde sa každému neurónu priradí vážený vstup a vypočíta sa jeho aktivácia.
2. Vypočíta sa chyba predikcie siete pomocou vybranej chybovej funkcie. Táto chyba je následne použitá na určenie toho, ako sa majú upraviť váhy siete.
3. Algoritmus vykonáva spätný pohyb cez siet a vypočíta sa gradient chybovej funkcie vzhľadom k váham siete.
4. Tento gradient sa následne použije na aktualizáciu váh siete. Akým spôsobom sa váhy upravia závisí od aktuálneho algoritmu optimalizácie.
5. Tento proces sa opakuje pre každý vstup v trénovacej množine a zvyčajne sa opakuje niekoľkokrát, kým sa dosiahne požadovaná presnosť siete.

Hlavnou výhodou algoritmu spätného šírenia chýb je, že umožňuje sieti učiť sa na základe skúseností. Tým, že váhy siete sa upravujú na základe chyby predikcie, sa sieť môže adaptovať na nové vstupy a zlepšovať svoju presnosť. Na Obrázku 1 – 12 vidíme plne prepojenú NN s vyznačeným aktivačným tokom signálu a v opačnom smere spätné šírenie chýb.



Obrázok 1 – 12: Plne prepojená NN s vyznačeným aktivačným tokom signálu a v opačnom smere spätné šírenie chýb

Je však potrebné dávať pozor na takzvané pretrénovanie siete, keď si sieť presne zapamäta trénovacie dátá a stráca schopnosť reagovať správne na nové vstupy. Preto je dôležité použiť techniky ako regularizácia či Dropout na zabránenie pretrénovania siete, ktorým sa budeme ďalej venovať v tejto kapitole.

1.2.5 Optimalizácia neurónových sietí

Pojem samotnej optimalizácie evokuje snahu sa zlepšovať, byť efektívnejším a ideálnejším. S vlastnosťou matematického modelovania prichádza ruka v ruke aj možnosť optimalizácie. Pre modelovanie NN je dôležitá nielen vhodná voľba aktivačnej a chybovej funkcie, ale aj voľba optimalizačnej funkcie. Optimalizácia neurónových sietí sa zaobera nájdením najlepších nastavení pre parametre siete, ktoré minimalizujú chybu predikcie a zlepšujú presnosť siete. K tomuto účelu sa používajú rôzne optimalizačné algoritmy – funkcie, ktoré sa snažia nájsť optimálne váhy siete.

Poznáme viacero takýchto optimalizačných funkcií, napr. gradientová metóda, stochastická gradientová metóda (angl. *Stochastic gradient descent*, skr. **SGD**), adaptívny algoritmus optimalizácie miery učenia (angl. *Adaptive moment optimizer*, skr. **Adam**), RMSprop, Nadam, Adamax, Adagrad, Adadelta a iné.

Optimalizačné funkcie sa často používajú v kombinácii s rôznymi inými technikami ako regularizácia, *Dropout*, alebo *Batch* normalizácia, ktoré pomáhajú zabrániť pretrénovaniu siete a zlepšujú jej schopnosť generalizácie na nové vstupy.

Stochastická gradientová metóda

Stochastická gradientová metóda alebo metóda stochastického gradientového zo-stupu je variácia gradientovej metódy, ktorá pracuje s náhodnými podmnožinami trénovacích dát a aktualizuje váhy siete na základe chyby len pre tieto náhodné podmnožiny (Keskar and Socher, 2017).

Tento prístup zvyčajne vedie k rýchlejšiemu učeniu siete, pretože aktualizácia váh sa vykonáva častejšie. Podľa Keskar and Socher (2017) sa zistilo, že adaptívne optimalizačné metódy ako *Adam*, *Adagrad* alebo *RMSprop* zle zovšeobecňujú v porovnaní s SGD. Tieto metódy majú tendenciu dosahovať dobré výsledky v počiatocnej časti tréningu, ale v neskôrších fázach tréningu ich SGD prekonáva.

Adam

Adam je podľa Kingma and Ba (2014) moderný optimalizačný algoritmus, ktorý kombinuje výhody SGD a gradientovej metódy. Tento algoritmus využíva adaptívne krokovanie a adaptívne nastavovanie momentu pre každý parameter siete. Opiera sa o gradient prvého rádu.

Funkciu je možné priamo implementovať, je efektívna na výpočty a nenáročná na pamäť. Slúži na efektívne aktualizovanie váh neurónov v NN. Adam sa ukázal ako veľmi účinný v optimalizácii hlbokých neurónových sietí.

RMSprop

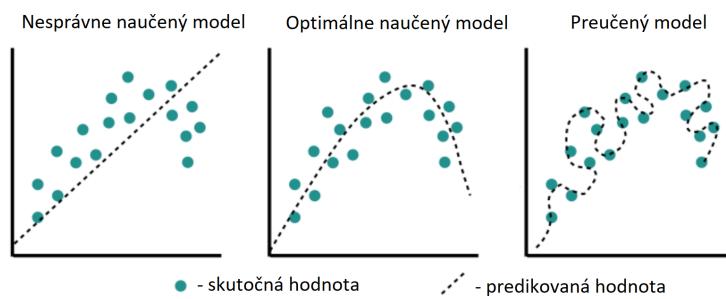
Funkcia *Root Mean Square Propagation* (skr. **RMSprop**) je optimalizačný algoritmus, ktorý adaptívne mení veľkosť kroku učenia pre každý parameter siete v závislosti od toho, ako sa menia gradienty týchto parametrov.

Po funkciu Adam je druhým najpoužívanejším optimalizačným algoritmom. Radí sa k metódam stochastickej optimalizácie a rovnako ako Adam sa opiera o gradient prvého rádu.

K normalizácii gradientu mu slúži klízavý priemer štvorcových gradientov. RMSprop nepovažuje rýchlosť adaptívneho učenia (angl. *adaptive learning rate*) za hyperparameter, teda rýchlosť učenia sa mení v čase (Hinton et al., 2012).

1.2.6 Regularizácia neurónových sietí

Regularizácia je technika používaná v neurónových sietach s cieľom zabrániť pretrénovaniu siete a zlepšiť jej schopnosť generalizovať sa na nové vstupy. Pretrénovanie (angl. *overfitting*) je stav, keď siet sa naučí zapamätať trénovacie dátá a nedarí sa jej dobre predikovať na nových dátach – testovacích. Regularizácia NN sa pomáha vyhnúť preučeniu. Na Obrázku 1 – 13 môžeme vidieť príklad na porovnanie predikcií a skutočných hodnôt podučeného, optimálneho a preučeného modelu. Podučený model je nedostatočne naučený model, ktorý predikuje hodnoty veľmi nepresne. Našim cieľom je zlatý stred – optimálny model.



Obrázok 1 – 13: Podučený, optimálny a preučený model

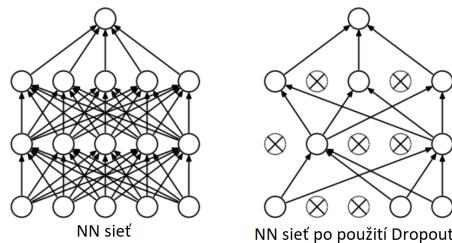
Poznáme viacero regularizačných techník napríklad takzvaná L1 regularizácia (*Lasso*), L2 regularizácia (Tikhonovova regularizácia), *Dropout*, *Early Stopping*, *Batch* normalizácia, augmentácia – zväčšenie trénovacej množiny dát rôznymi spôsobmi, najmä pri obrazových dátach je to posunutie, otočenie, zmena veľkosti či zmena farieb). V našej práci sme využili regularizáciu *Dropout* a *Early Stopping*.

Lasso – L1 a Tikhonovova– L2 regularizácia

Techniky L1 a L2 regularizácie pridávajú do chybovej funkcie neurónovej siete dodatočný parameter, ktorý znižuje hodnoty váh siete. L1 regularizácia pridáva absolútnu hodnotu váh do chybovej funkcie, zatiaľ čo L2 regularizácia pridáva ich druhú mocninu. Tým sa zabezpečí, aby váhy neurónovej siete nevystúpali do veľkých hodnôt.

Dropout

Funkcia *Dropout* má už v názve sloveso *drop*, čo v slovenskom preklade znamená pustiť, vyhodiť. Táto technika spočíva v náhodnom vynechávaní niektorých neurónov a ich synapsí zo siete počas trénovalia, ako môžeme vidieť na Obrázku 1 – 14.



Obrázok 1 – 14: Efekt Dropout funkcie

Tým sa zabezpečí, že neurónová sieť sa naučí rozpoznávať príznaky aj bez niektorých neurónov, čím sa zlepší jej schopnosť generalizácie na nové vstupy (Srivastava et al., 2014). Pri použití funkcie *Dropout* je potrebné určiť tzv. *dropout rate*, čo je podľa Srivastava et al. (2014) najčastejšie číslo od 0,1 - 0,8 a predstavuje percentuálny počet neurónov, ktoré sa zo siete vynechajú. Ak je *dropout rate* = 0,15; z NN

sa vynechá 15% zo všetkých neurónov. V našej práci využívame len nižší stupeň *dropout rate* 10%.

Early Stopping

Funkcia angl. *Early Stopping*, v preklade včasné zastavenie, spočíva v sledovaní vývoja chyby neurónovej siete na validačnej množine počas trénovania a zastavení trénovania, keď sa chyba začne zvyšovať. Tým sa zabezpečí, aby sieť nebola trénovaná príliš dlho a podporí jej schopnosť generalizácie. Ide o jednoduchú, ale účinnú metódu. Chyba na validačných dátach sa sleduje po každej epoche trénovania, a ak sa po určitom počte epoch (tzv. *patience*) nezlepší, trénovanie sa zastaví. V našom prípade funkcia *Early Stopping* bude sledovať chybu MAE na validačnej množine a výrazne skráti čas trénovania.

Batch normalizácia

Batch normalizácia spočíva v normalizácii vstupov a výstupov z každej vrstvy neurónovej siete počas trénovania. Táto normalizácia zabezpečuje, aby výstupy z každej vrstvy mali priemernú hodnotu blízkú 0 a smerodajnú odchýlku blízkú 1. Tým sa zabezpečí, že výstupy z každej vrstvy sú v rovnakom rozsahu. Vedie to k zlepšeniu stability a zvýšeniu efektivity trénovania. *Batch* normalizácia má tiež pozitívny vplyv na schopnosť generalizácie NN na nové vstupy a môže zlepšiť jej výkon na testovacích dátach.

1.3 Rekurentná neurónová sieť

Rekurentná neurónová sieť (z angl. *Recurrent Neural Network*, skr. RNN) je špeciálnym typom neurónovej siete, ktorá má schopnosť spracovávať a uchovávať informácie o predošlých udalostiach v sekvencii, čo ju robí vhodnou pre prácu so sekvenčnými dátami, ako sú napríklad texty, zvuky, reč alebo časové rady.

Hlavný rozdiel medzi rekurentnými a bežnými doprednými neurónovými sieťami

podľa Medsker and Jain (2001) spočíva v tom, že rekurentné siete obsahujú spätné väzby (takzvané *feedback loops*), ktoré umožňujú prenášať informácie z predoších krokov v sekvencii do aktuálneho kroku. Týmto spôsobom môže sieť uchovávať pamäť o predoších stavoch a využívať ich pri rozhodovaní o aktuálnom stave. Rozdiel v topológii siete medzi doprednými a rekurentnými sietami sme už videli na Obrázku 1–6.

RNN sú často využívané na riešenie nasledovných typov úloh:

- Predikcia textu – generovanie nového textu na základe predoších slov alebo správ, preklad textu z jedného jazyka do druhého.
- Rozpoznávanie reči – rozpoznávanie slov a fráz zo zvukových dát.
- Časové rady – predikcia budúcich hodnôt v časových radoch, ako napríklad cenové pohyby na finančných trhoch.

Rekurentné siete majú rôzne architektúry a varianty, ako napríklad jednoduché rekurentné siete (*Simple RNN*), *Long Short-Term Memory* siete (skr. **LSTM**), *Liquid state machine* (skr. **LSM**) a *Gated Recurrent Unit* siete (skr. **GRU**). Každá z týchto variánt má svoje vlastné vlastnosti a vhodnosť pre rôzne úlohy.

Vďaka schopnosti uchovávať informácie o predoších stavoch v sekvencii, sú rekurentné siete dôležitým nástrojom v oblasti strojového učenia a hlbokého učenia, a môžu byť použité na riešenie mnohých úloh, ktoré vyžadujú spracovanie sekvenčného typu dát.

Liquid State Machine

Liquid State Machine je typ rekurentnej neurónovej siete. LSM siete pozostávajú z troch hlavných častí: *liquid*, *readout* vrstvy a trénovacej fázy. Vrstva *readout* je plne prepojená s *liquid* vrstvou a zodpovedá za mapovanie stavových premenných na výstup siete. V trénovacej fáze sa *liquid* inicializuje a následne vstupné vzory

prejdú cez *liquid*, aby sa vytvorili stavové premenné. Tieto premenné sa potom použijú na trénovanie *readout* vrstvy, ktorá sa pripojí k liquidu.

Výhodou LSM siete je, že vytvára komplexné a dynamické reprezentácie vstupov, ktoré sa môžu meniť v čase. Taktiež má veľkú paralelnosť, čo umožňuje spracovať veľké množstvá vstupov naraz. Ďalšou výhodou je, že LSM siete nie sú také náročné na trénovanie ako bežné RNN siete, pretože sa sústredia na učenie lineárnych mapovaní, ktoré sú jednoduchšie ako učenie nelineárnych funkcií.

Long Short-Term Memory

Long Short-Term Memory je špeciálny typ RNN navrhnutý na efektívne riešenie problému takzvaného miznúceho gradientu (angl. *vanishing gradient*) a explozívneho gradientu (angl. *exploding gradient*), ktoré môžu nastať pri trénovaní bežnej RNN. LSTM siete sú navrhnuté tak, aby boli schopné uchovávať dlhodobé závislosti a pamäť v sekvenciách, ktoré sú nevyhnutné pre mnohé úlohy v oblasti ML (Yu et al., 2019).

Hlavnou vlastnosťou LSTM siete podľa Yu et al. (2019) sú takzvané brány (angl. *gates*), ktoré umožňujú regulovať tok informácií v sieti a rozhodovať o tom, ktoré informácie sa majú preniesť ďalej a ktoré sa majú zabudnúť.

LSTM má tri hlavné typy brán:

- Brána zabudnutia (angl. *forget gate*): rozhoduje, ktoré informácie pamäti majú byť zabudnuté a ktoré sa majú udržať.
- Brána vstupu (angl. *input gate*): rozhoduje, ktoré nové informácie sa majú uložiť do pamäte a aký majú význam.
- Brána výstupu (angl. *output gate*): rozhoduje, ktoré informácie z pamäte sa majú použiť na výstup siete.

Tieto brány sú implementované pomocou lineárnych vrstiev s aktivačnými funkciami *Sigmoid* a *Tanh*, funkcie určujú, ktoré hodnoty majú byť zapnuté a ktoré vypnuté.

Takýmto spôsobom môže LSTM siet efektívne uchovávať informácie o predošlých stavoch a rozhodovať o tom, aké informácie sú dôležité pre aktuálny stav.

LSTM siete sú v súčasnosti výrazne používané v oblasti spracovania prirodzeného jazyka (angl. *Natural language processing*, skr. **NLP**) a rozpoznávania reči. Vďaka ich schopnosti uchovávať dlhodobú pamäť sú vhodné pre úlohy, ako napríklad predikcia slov v texte alebo preklad medzi jazykmi. LSTM siete môžu byť trénované pomocou algoritmu spätného šírenia chyby, podobne ako bežné dopredné neurónové siete (Yu et al., 2019).

Gated Recurrent Unit

GRU je typ RNN. GRU siete majú dve brány. *Reset* bránu a *update* bránu. *Reset* brána určuje, kolko z predošlého stavu by malo byť zabudnuté, zatiaľ čo *update* brána určuje, kolko informácií by malo byť uložených v novom stave. Výhodou GRU sietí je, že majú menej parametrov ako bežné RNN siete a sú schopné učiť sa dlhšie závislosti medzi vstupnými hodnotami (Medsker and Jain, 2001).

1.4 Predikcia časových radov

Predikcia časových radov je oblasťou ML, ktorá sa zaobrá predikciou budúcich hodnôt časových radov na základe historických dát. Časové rady (angl. *Time Series*) sú postupnosti dát, ktoré sú usporiadane v časovom poradí, napríklad denné teploty, mesačné tržby alebo hodinové záznamy srdcového tepu.

Praktické využitie predikcie časových radov môžeme uplatniť v rôznych odvetviach, napr. ekonomickej predpovede, kde ide najčastejšie o predikciu cien akcií, predpovede hospodárskej situácie či predikciu inflácie. V predpovedi počasia, teploty, zrážok, vetra a ďalších meteorologických javov. Vo výrobe a plánovaní, napríklad predikcia dopytu po produktoch, optimalizácia zásobovania a plánovanie výroby. V oblasti zdravotníctva na predikciu vývoja ochorení a monitorovanie pacientov.

V predikcii časových radov sa používajú rôzne metódy ML. Tieto metódy sa líšia

vo svojom prístupe k predikcii časových radov a môžu byť použité na rôzne typy dát a úloh. Spomenieme si niektoré najpoužívanejšie metódy:

- ARIMA (angl. *Autoregressive Integrated Moving Average*) – metóda založená na autoregresii a pohyblivom priemere.
- Exponenciálny vyrovnávací model (angl. *Exponential Smoothing*) – metóda pracuje na základe expozície minulých hodnôt, využíva sa na predpovedanie krátkodobých trendov a nemá schopnosť zachytia sezónnosti.
- LSTM (angl. *Long Short-Term Memory*) – rekurentná neurónová sieť, môže zachytia komplexné závislosti a interakcie medzi rôznymi faktormi, ktoré ovplyvňujú časové rady.
- Prophet – určená na predpovedanie časových radov s komplexnými sezónnymi vzormi a trendmi, je založená na Bayesovskej dekompozícii a môže byť veľmi účinná.
- WaveNet – využíva hlboké neurónové siete a generatívne modelovanie. Je schopná zachytia aj extrémne vzorce a môže byť veľmi účinná pri predikcii časových radov v oblastiach, kde sa vyskytujú zložité a nepredvídateľné vzorce.

Predikcia časových radov môže byť veľmi ťažkou úlohou, pretože tieto rady môžu byť ovplyvnené mnohými faktormi, ktoré nemusia byť zrejmé alebo sú zložité na modelovanie. Niektoré z problémov, ktoré sa riešia pri predikcii časových radov, sú:

- Trendy – časové rady zvyčajne majú tendenciu postupne rást alebo klesať, čo má vplyv na predpovede. Tieto trendy sú buď lineárne, alebo nelineárne a bývajú ovplyvnené rôznymi faktormi, ako sú ekonomicke zmeny alebo sezónne faktory.
- Sezónnosť – mnoho časových radov má sezónne výkyvy, ktoré sa opakujú v určitom časovom období. Sezónne faktory môžu byť ročné, mesačné alebo týždenné.

- Náhodné vplyvy – časové rady môžu byť ovplyvnené náhodnými vplyvmi, ako sú napríklad udalosti, ktoré nemajú žiadny súvis s bežným trendom alebo sezónnosťou.
- Nejasné vzťahy medzi dátami – časové rady môžu mať medzi atribútami nejasné vzťahy.
- Zmeny v dátach – časové rady môžu byť ovplyvnené zmenami v dátach, ktoré môžu byť spôsobené zmenami v procese, v ktorom sa dátá generujú. Tieto zmeny sú buď náhle, alebo postupné.

Všetky tieto problémy komplikujú modelovanie a majú vplyv na presnosť predpovedí. Riešenie týchto problémov môže byť tažké, ale výberom vhodnej metódy podľa typu úlohy sa vieme dopracovať k vysoko presným výsledkom. Spomínanú presnosť určujeme prostredníctvom takzvaných štandardných metrík pre vyhodnocovanie klasifikačných modelov, ktorým sa budeme venovať v nasledujúcej podkapitole.

1.5 Metriky pre vyhodnocovanie modelov

Pri klasifikácii v strojovom učení sa úspešnosť modelov vyjadruje číselne pomocou určitých metrík. To platí nie len pre NN, ale aj ostatné algoritmy ML. Tieto metriky sa udávajú v rozpätí $\langle 0, 1 \rangle$ alebo v percentách. Pre konkrétny príklad klasifikácie do dvoch tried, sa budú metriky úspešnosti počítať pomocou nasledovných 4 hodnôt:

- Pravdivo negatívne (angl. *True Negative*, skr. **TN**) – označuje počet správne identifikovaných negatívnych prípadov,
- Pravdivo pozitívne (angl. *True Positive*, skr. **TP**) – označuje počet správne identifikovaných pozitívnych prípadov,
- Falošne negatívne (angl. *False Negative*, skr. **FN**) – označuje počet nesprávne identifikovaných negatívnych prípadov,

- Falošne pozitívne (angl. *False Positive*, skr. **FP**) – označuje počet nesprávne identifikovaných pozitívnych prípadov.

Tieto 4 hodnoty sa zapisujú do kontingenčnej tabuľky (angl. *Confusion Matrix*), viď Tabuľka 1–1.

Tabuľka 1–1: Kontingenčná tabuľka

		predikované hodnoty	
		0	1
skutočné hodnoty	0	TN	FP
	1	FN	TP

Úspešnosť

Úspešnosť klasifikácie (angl. *accuracy*, skr. **ACC**) je najbežnejšou metrikou a určuje sa ako podiel správne identifikovaných prípadov ku všetkým prípadom. Pre úspešnosť teda platí nasledujúci vzťah:

$$\text{úspešnosť} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FN} + \text{TP} + \text{FP}} \quad (1.7)$$

Návratnosť

Návratnosť (angl. *recall*) alebo inak citlivosť (angl. *sensitivity*) sa týka iba pozitívnych prípadov dátovej vzorky, a je to podiel správne určených pozitívnych prípadov a všetkých pozitívnych prípadov. Návratnosť sa inak nazýva aj miera pravdivej pozitivity (angl. *True Positive Rate*, skr. **TPR**). Návratnosť vypočítame podľa vzťahu:

$$\text{návratnosť} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1.8)$$

Špecifickosť

Špecifickosť (angl. *specificity*) na rozdiel od návratnosti si všíma negatívne prípady, vypočíta sa ako návratnosť, avšak pre negatívne prípady. Špecifickosť sa inak na-

zýva aj miera pravdivej negativity (angl. *True Negative Rate*, skr. **TNR**). Vzorec pre výpočet špecifickosti:

$$\text{špecifickosť} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (1.9)$$

Presnosť

Presnosť (angl. *precision*) sa určuje pre každú triedu osobitne, teda v našom prípade pre pozitívne a negatívne prípady. Ide o podiel správne identifikovaných položiek v danej triede a počet položiek, ktoré model zaradil do danej triedy. Vzťahy pre výpočet presnosti v oboch triedach sú:

$$\text{presnosť}_1 = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{presnosť}_0 = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (1.10)$$

F1-skóre

F1-skóre (angl. *F1-score*) je harmonickým priemerom presnosti a návratnosti pre danú triedu. Vzťah pre výpočet F1-skóre je:

$$\text{F}_1 \text{ skóre} = 2 * \frac{\text{presnosť} * \text{návratnosť}}{\text{presnosť} + \text{návratnosť}} \quad (1.11)$$

Miera falošnej pozitivity

Miera falošnej pozitivity (angl. *False Positive Rate*, skr. **FPR**) je miera nesprávne klasifikovaných negatívnych príkladov ku všetkým negatívnym prípadom. Platí pre ňu vzťah:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (1.12)$$

Miera falošnej negativity

Miera falošnej negativity (angl. *False Negative Rate*, skr. **FNR**) je miera nesprávne klasifikovaných pozitívnych príkladov ku všetkým pozitívnym prípadom. Výpočet:

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \quad (1.13)$$

Pozitívna prediktívna hodnota

Pozitívna prediktívna hodnota (angl. *Positive Predictive Value*, skr. **PPV**) je pravdepodobnosť, že klasifikačný model správne určí pozitívnu triedu. PPV sa vypočíta ako podiel počtu pravdivo pozitívnych prípadov k celkovému počtu prípadov, ktoré boli klasifikované ako pozitívne. Ide o už vyššie spomínanú presnosť, v tomto prípade pozitívnej triedy, platí rovnaký vzorec 1.10 ako pre presnosť_1 .

Negatívna prediktívna hodnota

Negatívna prediktívna hodnota (angl. *Negative Predictive Value*, skr. **NPV**) je pravdepodobnosť, že klasifikačný model správne určí negatívnu triedu. NPV sa vypočíta ako podiel počtu pravdivo negatívnych prípadov k celkovému počtu prípadov, ktoré boli klasifikované ako negatívne. Ide o už vyššie spomínanú presnosť, v tomto prípade negatívnej triedy, platí rovnaký vzorec 1.10 ako pre presnosť_0 .

Miera falošných objavov

Miera falošných objavov (angl. *False Discovery Rate*, skr. **FDR**) je podiel falošne pozitívnych prípadov k celkovému počtu prípadov, ktoré boli klasifikované ako pozitívne. FDR sa používa na vyhodnotenie, aká veľká časť pozitívnych výsledkov je falošne pozitívna. Platí vzťah:

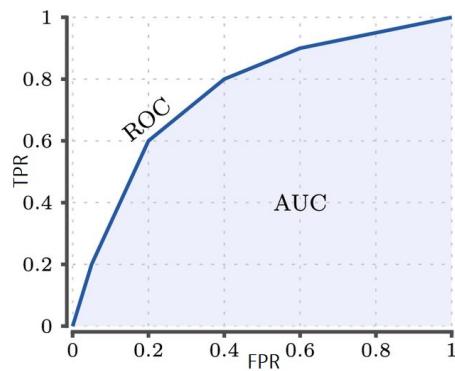
$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}} \quad (1.14)$$

Krivka prevádzkovej charakteristiky prijímača a plocha pod krivkou

Krivka prevádzkovej charakteristiky prijímača (angl. *Receiver Operating Characteristic*, skr. **ROC**) a plocha pod krivkou (angl. *Area Under the Curve*, skr. **AUC**) sú dva úzko súvisiace pojmy. ROC krivka je grafickým nástrojom, ktorý určuje, ako dobre model rozlišuje medzi dvoma triedami. ROC krivka zobrazuje závislosť medzi TPR a FPR modelu pre rôzne prahové hodnoty. ROC krivka je významná aj preto,

že jej plocha pod krivkou AUC poskytuje celkovú mieru úspešnosti klasifikačného modelu, vidľ. Obrázok 1 – 15⁴.

Výpočet AUC spočíva v určení plochy pod ROC krivkou. Hodnota AUC sa pohybuje v rozsahu od 0 do 1, kde hodnota 1 znamená dokonalú klasifikáciu a hodnota 0,5 znamená, že klasifikátor má náhodné rozhodnutie. V praxi sa AUC používa na porovnávanie výkonu klasifikačných modelov a na výber najlepšieho modelu. AUC poskytuje jednoduchý a intuitívny ukazovateľ na mieru úspešnosti modelu, ktorý je nezávislý na priemernej hodnote tried, a teda je vhodný aj pre nerovnomerné distribúcie tried.



Obrázok 1 – 15: ROC krivka a plocha AUC pod ňou

Priemerná presnosť oboch tried

Priemerná presnosť (angl. *Balanced Accuracy*) je metrika používaná na vyhodnocovanie výkonnosti klasifikačného modelu v prípade, že dátá sú nevyvážené. Nevyvážené dátá sú také, kde počet príkladov v jednej triede je výrazne väčší ako v druhej triede.

Úspešnosť môže byť v takomto prípade neinformatívna, pretože model môže mať vysokú presnosť klasifikácie pre dominantnú triedu, a zároveň veľmi nízku presnosť pre menej početnú triedu. Ak sú dátá vyvážené, čiže ak počet príkladov v oboch triedach je približne rovnaký, potom je *Balanced Accuracy* rovnaká ako úspešnosť.

⁴Zdroj Obrázka 1 – 15: www.researchgate.net/publication/357044864_Maintaining_AUC_and_H-measure_over_time

Balanced Accuracy vyjadruje priemernú presnosť pre obe triedy, kde každá trieda má rovnaký význam. Vypočíta sa ako priemer TPR a TNR podľa nasledujúceho vzťahu:

$$\text{balanced accuracy} = \frac{1}{2} \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right) \quad (1.15)$$

Korelačný koeficient medzi skutočnými a predpokladanými triedami

Tento korelačný koeficient (angl. *Matthews Correlation Coefficient*, skr. **MCC**) môže nadobúdať hodnoty od -1 do 1 , pričom hodnota 1 znamená dokonalú koreláciu medzi predikovanými a skutočnými triedami, hodnota -1 znamená úplnú opačnú koreláciu a hodnota 0 znamená náhodnú predikciu.

MCC je výhodná pre nevyvážené dátové sady, pretože berie do úvahy všetky hodnoty kontingenčnej tabuľky a nie len správne klasifikované príklady. Ďalšou výhodou je, že je vhodná pre klasifikátory s nízkou presnosťou, pretože môže ukázať, že aj klasifikátor s nízkou presnosťou môže mať vysokú koreláciu s predikovanými triedami. Pre výpočet MCC používame nasledujúci vzťah:

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (1.16)$$

1.6 Použité technológie

V tejto podkapitole sa budeme venovať technológiám, ktoré priamo využívame v praktickej časti tejto práce: programovací jazyk Python, kľúčové knižnice pre DL - Keras, Tensorflow, knižnica pre interaktívnu vizualizáciu Bokeh, okrajovo spomenieme aj pomocné knižnice, ktoré zjednodušujú prístup k dátam z vesmírnych misií Heliopy a Aidapy.

Programovací jazyk Python

Python je vysokoúrovňový interpretovaný objektovo orientovaný *open-source* programovací jazyk, ktorý bol vyvinutý v 90. rokoch Guidom van Rossumom. Jazyk

bol navrhnutý s cieľom byť jednoduchý, prehľadný a intuitívny, s dôrazom na čitateľnosť kódu. Má široké použitie v rôznych oblastiach, ako sú webové aplikácie, databázové správy, vedecké výpočty, ML, AI a iné. Má obrovskú knižnicu modulov a nástrojov, ktoré umožňujú vývojárom rýchlo a efektívne vytvárať a testovať svoje aplikácie. Python je tiež známy svojimi silnými komunitami vývojárov a používateľov, ktoré pravidelne prispievajú k jeho rastu a rozvoju. Najvýznamnejšie pythonovské knižnice v oblasti ML podľa Costa (2020) sú: **Tensorflow**, **Keras**, **PyTorch**, **Scikit-learn**, **Pandas**, **Numpy**, **NLTK**, **Theano**, **MXNet** a **Spark MLlib**.

Tensorflow

Tensorflow je jedna z najlepších dostupných knižníc pre ML v Pythone. Ide o rýchlu, flexibilnú a škálovateľnú *open-source* knižnicu. Je poskytovaná spoločnosťou Google a umožňuje ľahké vytváranie modelov ML. Oblasti, v ktorých **Tensorflow** vyniká sú napríklad práca s hlbokými neurónovými sieťami, rozpoznávanie obrázkov, textu a reči, ale aj riešenie parciálnych diferenciálnych rovníc (Costa, 2020).

Keras

Keras je populárna open-source knižnica, ktorá sa zaoberá najmä neurónovými sieťami. Je navrhnutá spoločnosťou Google. V súčasnosti je **Keras** rozšírením knižnice **Tensorflow**. **Keras** obsahuje hlavné stavebné prvky a nástroje NN: vrstvy NN, aktyvačné, chybové a optimalizačné funkcie, Batch normalizáciu, **Dropout** a metriky využívacie hodnotovanie (Costa, 2020).

Bokeh

Bokeh je *open-source* knižnica pre vizualizáciu dát pomocou jazyka Python. Umožňuje tvorbu interaktívnych a estetických grafov, diagramov a vizualizácií. **Bokeh** podporuje rôzne typy vizualizácií, vrátane čiarových grafov, stĺpcových grafov, histogramov, kruhových grafov, heatmaps a mnohých ďalších. Jednou z výhod **Bokeh**

je možnosť interaktivity. **Bokeh** umožňuje užívateľom interagovať s vizualizáciou pomocou rôznych funkcií, ako sú zväčšovanie, priblíženie, pohyb, výber.

Heliopy

Heliopy je knižnica v jazyku Python určená na prácu s dátami z heliosféry, ktorá zahŕňa dátá z misií ako napríklad *ACE*, *WIND*, *STEREO* alebo *Parker Solar Probe*. Knižnica obsahuje funkcie na načítanie, spracovanie a analýzu dát z týchto misií, ako aj na generovanie vizualizácií. **Heliopy** sa zameriava najmä na fyziku slnečnej koróny, heliosféry a interakcie slnečného vetra s planétami a kozmickým prostredím. Knižnica je otvoreným zdrojom, a je k dispozícii na stránke GitHub.

Aidapy

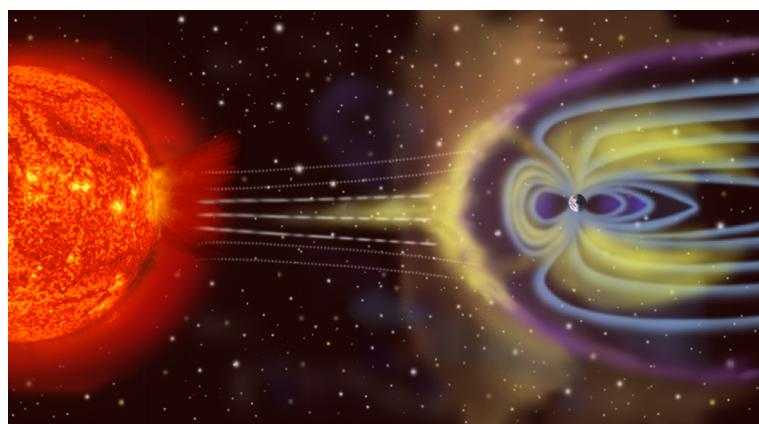
Knižnica **Aidapy** zjednodušuje prístup k dátam z kozmických misií, k simuláciám vesmírnej fyziky a k pokročilým štatistickým nástrojom. Knižnicu v práci využívame hlavne na zjednodušený a rýchly prístup k skúmaným dátam.

2 Predikcia geomagnetickej búrky

Táto kapitola sa venuje doménovej časti diplomovej práce. Vysvetlíme si základné pojmy ohľadom geomagnetických búrok, príčiny ich vzniku, mieru intenzity, priebeh a ich rôzne negatívne vplyvy. Pozrieme sa aj na analýzu súčasného stavu, kde spomenieme niektoré podobné výskumy a porovnáme ich výsledky.

2.1 Geomagnetická búrka

Geomagnetická búrka je nebezpečným prejavom kozmického počasia a príčinou jej vzniku sú najmä výrony koronálnej hmoty zo Slnka (angl. *Coronal Mass Ejection*, skr. **CME**), ktoré vylúčia smerom k Zemi energeticky nabité častice, a tie narušia zemskú magnetosféru, vid Ď Obrázok 2–1⁵. Takáto geomagnetická búrka môže mať rôzne efekty, ako napríklad zvýšenie radiácie v atmosfére Zeme, výpadok sateлитov, energetických sietí a transformátorov, ale zapríčinuje aj vznik polárnej žiary. Spomínané efekty však závisia od intenzity geomagnetickej búrky (Astafyeva et al., 2014).



Obrázok 2 – 1: Vplyv výronu koronálnej hmoty na zemskú magnetosféru

⁵Zdroj Obrázka 2 – 1: (Leone, 2023)

Príčiny vzniku geomagnetickej búrky

Geomagnetickú búrku môžme popísat ako náhle zosilnenie a zmenu zloženia magnetického poľa Zeme. Je spôsobená vplyvom slnečného vetra. Slnečný vietor je tok nabitej častice a magnetického poľa zo Slnka do slnečnej sústavy. Keď sa tieto nabité častice dostanú k Zemi, interagujú so zemským magnetickým poľom, čím ho narušia.

Geomagnetická búrka podľa Gonzalez et al. (1994) vzniká interakciou slnečného vetra s magnetickým poľom Zeme. Keď slnečný vietor dorazí do blízkosti Zeme, interakcia jeho nabitej častice a magnetického poľa Zeme spôsobí prúdenie elektrických nábojov v horných vrstvách atmosféry. Tento prúd sa nazýva "aurorálny prúd" a vzniká v oblasti nad magnetickým pólem Zeme. Pri silnej interakcii slnečného vetra a magnetického poľa Zeme sa aurorálny prúd výrazne zosilní a môže sa objaviť aj v oblastiach, kde by sa bežne nenachádzal.

Aurorálny prúd spôsobuje v hornej atmosfére ohrev a excitáciu atómov a molekúl plynu, čo vedie k emisii svetla a tvorbe polárnej žiary. Zároveň sa vytvárajú turbulentné elektrické a magnetické polia, ktoré spôsobujú fluktuácie a zmeny v magnetickom poli Zeme (Gonzalez et al., 1994).

Miera geomagnetickej aktivity v blízkosti Zeme

Miera geomagnetickej aktivity v blízkosti Zeme (angl. *Disturbance storm time index*, skr. **DST** index) je jedným z indexov týkajúcich sa kozmického počasia. Pre našu prácu je veľmi dôležitým, nakoľko je hlavným indikátorom geomagnetických búrok. Mernou jednotkou je nanotesla (nT). Jeho hodnoty sa pohybujú v rozmedzí od -600 do +100 nT. Jeho hodnota sa vypočítava zo štyroch hodnôt magnetického poľa získaných z meraní na štyroch rôznych observatóriách v blízkosti rovníka. Podľa Lethy et al. (2018) DST index meria intenzitu globálne symetrického rovníkového elektrojetu (prstencového prúdu), ktorý je spôsobený CME. Na základe hodnoty DST indexu vieme povedať, či geomagnetická búrka nastala alebo nie. Ak jeho hodnota klesne na hodnotu -20 nT a menej, hovoríme o geomagnetickej búrke.

Na základe hodnoty DST indexu sa určuje aj intenzita prípadnej geomagnetickej búrky. DST index je súčasťou datasetu *Outer Magnetospheric Multiscale Mission Integrated Science*, skr. **OMNI**⁶, spolu s ďalšími indexami, ktoré ovplyvňujú vesmírne počasie. DST index je meraný v hodinových intervaloch už od 1. januára 1963 (Papitashvili, n.d.).

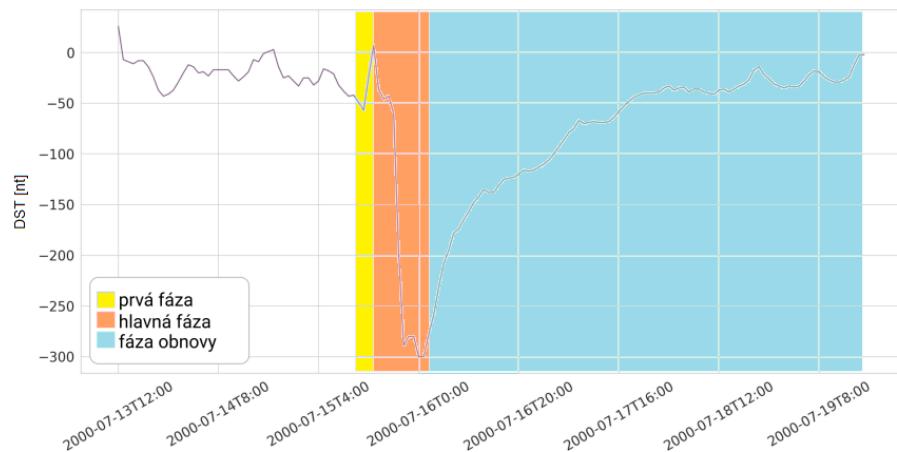
Priebeh geomagnetickej búrky

Geomagnetická búrka má rôzne trvanie, niektoré búrky trvajú niekoľko hodín, u iných rátame dni. Podľa Gonzalez et al. (1994) je počas veľkých geomagnetických búrok možné zaznamenať 3 fázy búrky:

1. Prvá fáza – mierny nárast DST indexu, na Obrázku 2 – 2 je fáza v žltom pásme.
2. Hlavná fáza – náhly pokles pod hladinu -20nT, následne je hladina chvíľu stála, na Obrázku 2 – 2 oranžové pásma.
3. Fáza obnovy – posledná fáza, magnetické pole sa zotavuje, hladina DST indexu pomaly stúpa na svoju pokojovú hodnotu, fáza na Obrázku 2 – 2 v modrom pásme.

Vizualizácia zobrazuje priebeh búrky v rozpätí niekoľkých dní z roku 2000, kedy búrka dosiahla intenzitu -300nT.

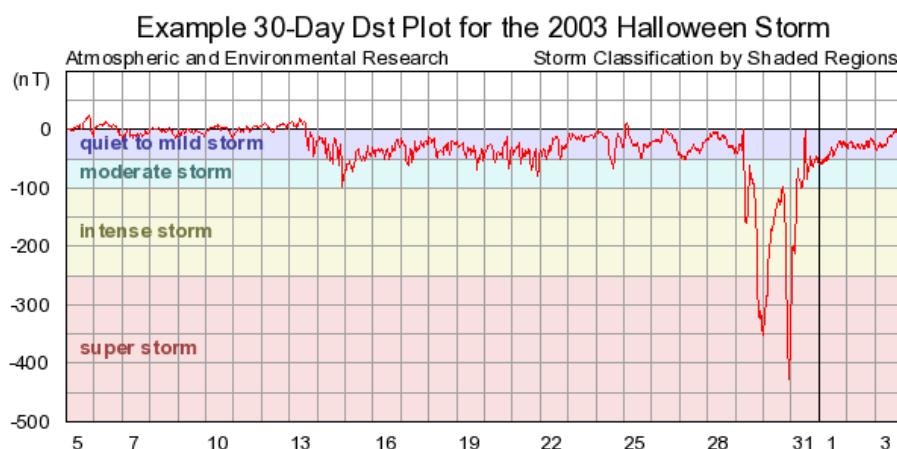
⁶Referencia na hl. zdroj datasetu: <https://omniweb.gsfc.nasa.gov/>



Obrázok 2 – 2: DST index počas priebehu búrky

Intenzita búrky a jej vplyvy

Geofyzici používajú pre rozdelenie intenzity geomagnetických búrok podľa Gonzalez et al. (1994) na rôzne úrovne – takzvané G-kategórie. Nižšie v Tabuľke 2 – 1 uvádzame zjednodušený popis týchto kategórií, a príklady ich vplyvov. Obrázok 2 – 3⁷ taktiež zobrazuje pásmá intenzity búrok farebne.



Obrázok 2 – 3: Príklad nameraného DST indexu a úrovne intenzity geomagnetickej búrky

⁷Zdroj Obrázka 2 – 3: www.aer.com/science-research/space/space-weather/space-weather-index/

Tabuľka 2 – 1: Úrovne intenzity geomagnetických búrok a ich vplyvy

označenie	intenzita	rozpätie DST (nT)	vplyvy
G1	slabá	$\langle -20, -30 \rangle$	minimálny vplyv na životné prostredie a technológie
G2	stredne silná	$\langle -30, -50 \rangle$	polárna žiara, rušenia signálov, výpadky elektrického prúdu
G3	silná	$\langle -50, -100 \rangle$	závažnejšie výpadky prúdu, výskyt nežiadúcich prúdov vo vedeniach, výskyt polárnej žiary
G4	veľmi silná	$\langle -100, -250 \rangle$	rozsiahle výpadky prúdu, poškodenie družíc, výskyt polárnej žiary v nižších zemepisných šírkach
G5	extrémne silná – superbúrka	-250 a menej	katastrofické poruchy v celom elektroenergetickom systéme, vypnutie transformátorov a vedení, nezvratné poškodenie družíc, polárna žiara aj v blízkosti rovníka

Z Lakhina and Tsurutani (2016) sa dozvedáme, že najsilnejšia geomagnetická búrka, ktorá bola zaznamenala pomocou DST indexu, sa udiala 13. marca 1989. Búrka mala DST index -589 nT a bola spôsobená výronom koronálnej hmoty zo Slnka. Výron spôsobil silný prúd nabitéh častíc, ktoré zasiahli Zem a vyvolali búrku, ktorá mala výrazné dopady na elektroenergetické siete.

Pred začiatkom merania DST indexu najsilnejšia zaznamenaná geomagnetická

búrka bola takzvaná Carringtonova búrka, ktorá zasiahla Zem 1. a 2. septembra 1859.

Následky tejto búrky boli v tom čase obmedzené hlavne na telegrafické siete, kde zaznamenali výboje a iskry, ktoré poškodili telegrafické zariadenia a niektoré dokonca zapálili papier. Okrem toho bola búrka doprevádzaná polárnou žiarou, ktorá bola viditeľná až v Karibiku a Afrike (Lakhina and Tsurutani, 2016).

Dnes by následky takejto búrky boli oveľa závažnejšie, pretože naše moderné technológie sú oveľa citlivejšie na vplyvy geomagnetických búrok.

2.2 Analýza súčasného stavu

Príbužným témam sa už doposiaľ venovalo viacero článkov, prác, výskumov a experimentov. Preto je dôležité zanalyzovať súčasný stav danej problematiky predikcie geomagnetických búrok z čo najaktuálnejších zdrojov.

Medzi prvé výskumy danej problematiky patrí japonský výskum *Prediction of the geomagnetic storm associated Dst index using an artificial neural network algorithm* publikovaný v roku 1999 (Kugblenu et al., 1999). Autori sa zameriavalí na geomagnatické búrky ako eventy s dovtedy najväčšou intenzitou, a dátovo analyzovali príčiny ich vzniku.

V článku *Operational forecasts of the geomagnetic Dst index* (Lundstedt et al., 2002) sa dozvedáme, že ďalšie výskumy podobného typu prebehli v roku 2002, kde bol odprezentovaný model neurónovej siete predikujúci real-time hodnoty DST indexu, čiže bez časového predstihu. Bola použitá takzvaná Elmanova architektúra neurónovej siete.

O 2 roky neskôr, slovenský výskum *Fuzzy neural networks in the prediction of geomagnetic storms* (Andrejková et al., 2004), sa opäť pokúšal o predikciu geomagnatických búrok pomocou neurónových sietí. Cieľom tohto výskumu bolo predikovať geomagnetickú búrku s predstihom 4 hodín. Vo výskume boli použité dátá namenaných indexov geomagnetických búrok z rokov 1982-1984, 1989, 1990 a 1992. Vo vý-

skume bola použitá jednoduchá neurónová sieť a výsledky nie sú interpretované štandardnými metrikami. Pre predstavu, podarilo sa im zachytiť najviac 29 z 37, 28 z 52 a 57 z 89 búrok v rôznych datasetoch za rôznych nastavení parametrov.

Existuje viacero starších podobných výskumov, žiadne však neudáva presnosť predikcie. Z novších publikácií je úspešná napríklad *Multiple-Hour-Ahead Forecast of the Dst Index Using a Combination of Long Short-Term Memory Neural Network and Gaussian Process*⁸ (Gruet et al., 2018). V tomto článku je predstavený model na predpovedanie geomagnetického indexu Dst 1 - 6 hodín dopredu pomocou LSTM a Gaussovho procesu. Výsledky sú interpretované pomocou TPR a FPR, v závislosti od intervalu DST indexu. Najlepší výsledok dosiahli v predikcii 1 h dopredu, najslabší pre 6 h dopredu. Výsledky sú zhrnuté v Tabuľke 2 – 2.

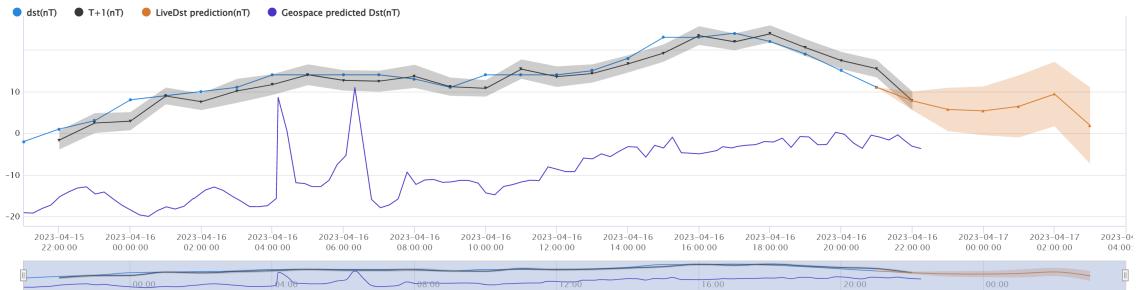
Tabuľka 2 – 2: Výsledky výskumu Gruet et al. (2018)

trieda	DST	pri predikcii			
		o 1 hod vopred		o 6 hod vopred	
		TPR	FPR	TPR	FPR
1	DST<-250	0.969	$6.40 \cdot 10^{-4}$	0.50	$8.34 \cdot 10^{-3}$
2	-250<DST<-50	0.927	0.0719	0.859	0.167
3	DST>-50	0.929	0.0705	0.834	0.130

Neskôr čínsky výskum z roku 2020 *Prediction of the Dst Index with Bagging Ensemble-learning Algorithm* (Xu et al., 2020), využíva *Bagging ensemble-learning* algoritmus, ktorý kombinuje 3 algoritmy — *the artificial neural network, support vector regression* a LSTM. Model bol testovaný na dátach z roku 2016, od 12. do 17. októbra. Najviac presná predpoveď geom. búrky je predpoveď s 1 hodinovým predstihom, a má výsledok RMSE 3.7327 nT, korelačný koeficient 0.9928 a presnosť predikcie intervalu 96.69 %.

⁸Dostupná online: <https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018SW001898>

Za zmienku jednoznačne stojí aj webová aplikácia *DstLive*⁹. Ide o model pre predpoved DST indexu v reálnom čase, vycháza z výskumu Hu et al. (2022). *DstLive* predpovedá DST 1 až 6 hodín dopredu, s pridruženými pravdepodobnosťami. Ako vyzerá spomínaná online predpoveď vidíme na Obrázku 2 – 4.



Obrázok 2 – 4: Výstup z aplikácie *DstLive*

Táto práca pokračuje v úsilí vyššie spomínaných výskumov a pokúša sa dosiahnuť porovnatelne dobrý klasifikačný model ako celosvetovo najlepšie modely. Vlastná metodika modelu sa využije pre výskumné účely na Ústave experimentálnej fyziky SAV, v.v.i., odkiaľ prišla požiadavka a námet na prácu. K práci vznikne aj vedecký článok v anglickom jazyku, ktorý sa pokúsime v blízkej dobe publikovať.

⁹Dostupná online na: <https://swx-trec.com/dst/>

3 Praktická časť

Praktická časť diplomovej práce sa zaobrá využitím poznatkov, postupov a techník spomínaných v predošlých kapitolách. Používame štandardný postup dolovania v dátach a celý proces pozostáva z nasledujúcich fáz:

1. Pochopenie cieľa – pochopenie reálneho problému, cieľov a požiadaviek a transformácia na konkrétnu definíciu úlohy dolovania.
2. Získanie dát – zber dát, prístup k dátam, uloženie dát do súboru.
3. Pochopenie dát – fáza zahŕňa oboznámenie sa s dátami a vzťahmi medzi nimi, zistenie ich kvality a charakteru.
4. Príprava dát – vytvorenie množiny dát pre modelovanie. Patrí sem výber atribútov, odvodenie nových atribútov, ako aj ich transformácia a čistenie dát.
5. Modelovanie – modelovanie RNN s LSTM, nastavenie a vyladenie parametrov na optimálne hodnoty, súvisí s fázou prípravy dát a vyžaduje priebežné vyhodnocovanie.
6. Vyhodnotenie – v tejto etape sú klasifikačné modely evaluované pomocou štandardných metrík pre vyhodnocovanie klasifikačných a predikčných modelov, úzko súvisí s fázou modelovania a na základe výsledkov sa priebežne vylepšuje existujúci klasifikačný model.
7. Nasadenie – odovzdanie výsledku práce a zdrojových kódov na výskumné účely Ústavu experimentálnej fyziky Slovenskej akadémie vied, v.v.i., ktorá sa aktívne podieľala na vzniku tejto práce.

Zdrojové kódy a dátá k jednotlivým fázam ako aj celému výskumu sú zverejnené v online repozitári¹⁰. Zdrojové kódy sú vo formáte Jupyter Notebook, dátá a predikcie v súboroch sú vo formáte .csv, natrénované klasifikačné modely vo formáte .hdf5.

¹⁰Dostupné online na: https://github.com/motuziik/DP_Veronika_Motuzova

Trénovanie modelov realizované online s povolením KKUI TUKE¹¹ na GPUs Tesla K40c, Quadro RTX4000 a Quadro P4000.

3.1 Pochopenie cieľa

Cieľom je predpovedať blížiacu sa geomagnetickú búrku, čiže musíme vo vhodnom časovom predstihu rozhodnúť, či búrka nastane alebo nie. Ide teda o binárnu klasifikačnú úlohu. Neskôr môžeme vyhodnotiť aj aká silná búrka sa blíži. Bolo by potrebné klasifikovať DST index do viacerých tried podľa intenzity búrok vysvetlenej v časti 2.1 Intenzita búrky a jej vplyvy.

Získanie dát

OMNI dataset je dostupný cez portál¹². V programovacom jazyku Python však existujú knižnice pre pohodlný prístup k dátam – Aidapy a Heliopy. Využívame teda prístup cez spomínané knižnice. Pred prístupom k dátam je potrebné nainštalovať všetky predispozície a zvoliť si časové obdobie, v ktorom boli skúmané dátá zaznamenané. Pre čo najlepšie výsledky chceme pracovať so všetkými dostupnými dátami, preto volíme obdobie od 1. januára 1963 až po 14. máj 2022 (v danom momente najaktuálnejší dostupný záznam). Celkovo dostávame k dispozícii 520 443 záznamov s hodinovým intervalom a ukladáme ich do súboru surových dát `omni_full.csv`. Zdrojový kód tejto fázy s kompletným postupom sa nachádza v súbore `1_ulozenie_a_stiahnutie_dat.ipynb`, dostupné online¹³.

¹¹Prístup umožňuje CHI: <https://datalab.kkui.fei.tuke.sk/gpu>

¹²Zdroj dát: <https://omniweb.gsfc.nasa.gov/>

¹³Získanie dát: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/1_ziskanie_dat

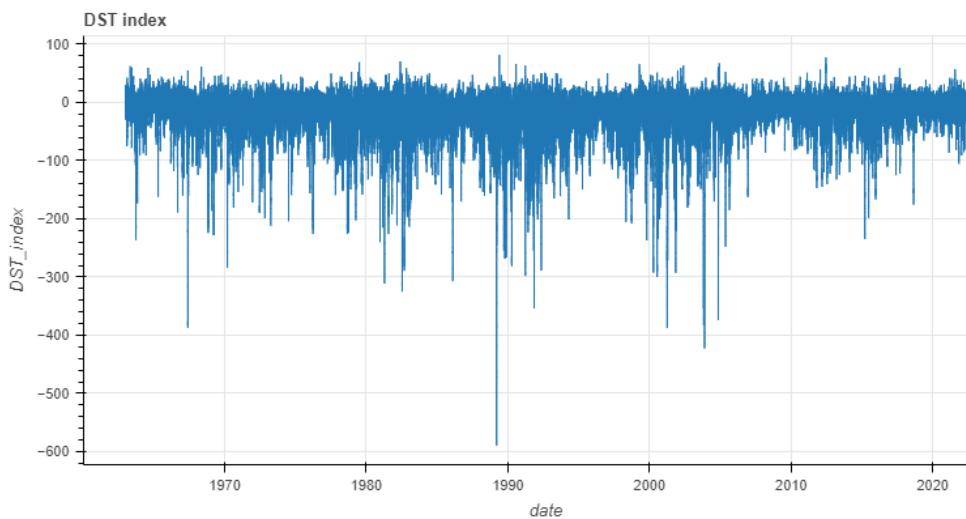
3.2 Pochopenie dát

Fáza pochopenia dát pozostávala z niekoľkých krokov, ktoré si bližšie popíšeme. Najprv bolo potrebné zistiť tvar a rozsah dát – zistili sme počet záznamov a atribútov. K dispozícii máme 520 443 záznamov (riadkov) a 55 atribútov (stĺpcov).

Nasledujúcim krokom bolo určenie cieľového atribútu – atribút, ktorý určuje, či geomagnetická búrka nastala alebo nie, vyjadruje DST index. V datasete je nazvaný jednoducho DST. Našou úlohou je tento atribút predikovať a na základe neho klasifikovať blížiaci stav búrky s predstihom niekoľkých hodín. Prahová hodnota pre výskyt geomagnetickej búrky je $-20nT$. DST index menší ako prahová hodnota bude predstavovať stav geomagnetickej búrky (klasifikovaný ako 1), v opačnom prípade ide o normálny stav (klasifikovaný ako 0). Štatistické informácie o atribúte sú zhrnuté v Tabuľke 3 – 1. Priebeh hodnôt hlavného atribútu DST vidíme na Obrázku 3 – 1 v podobe vizualizácie Bokeh.

Tabuľka 3 – 1: Štatistické informácie o cieľovom atribúte DST index

priemer	štandardná odchýlka	min	25%	50%	75%	max
-14.13	22.03	-589.00	-22.00	-10.00	-1.00	81.00



Obrázok 3 – 1: Hodnoty hlavného atribútu DST v čase

Ďalej je dôležité zistiť prítomnosť a počet chýbajúcich hodnôt v dátach pre každý atribút – vid' Obrázok 3–2. Väčšina atribútov má viac než 100 000 chýbajúcich hodnôt, chýbajúce hodnoty sa vyskytujú najmä pri starších záznamoch spred 80-tých rokov. Náš hlavný atribút DST index však chýbajúce hodnoty neobsahuje.

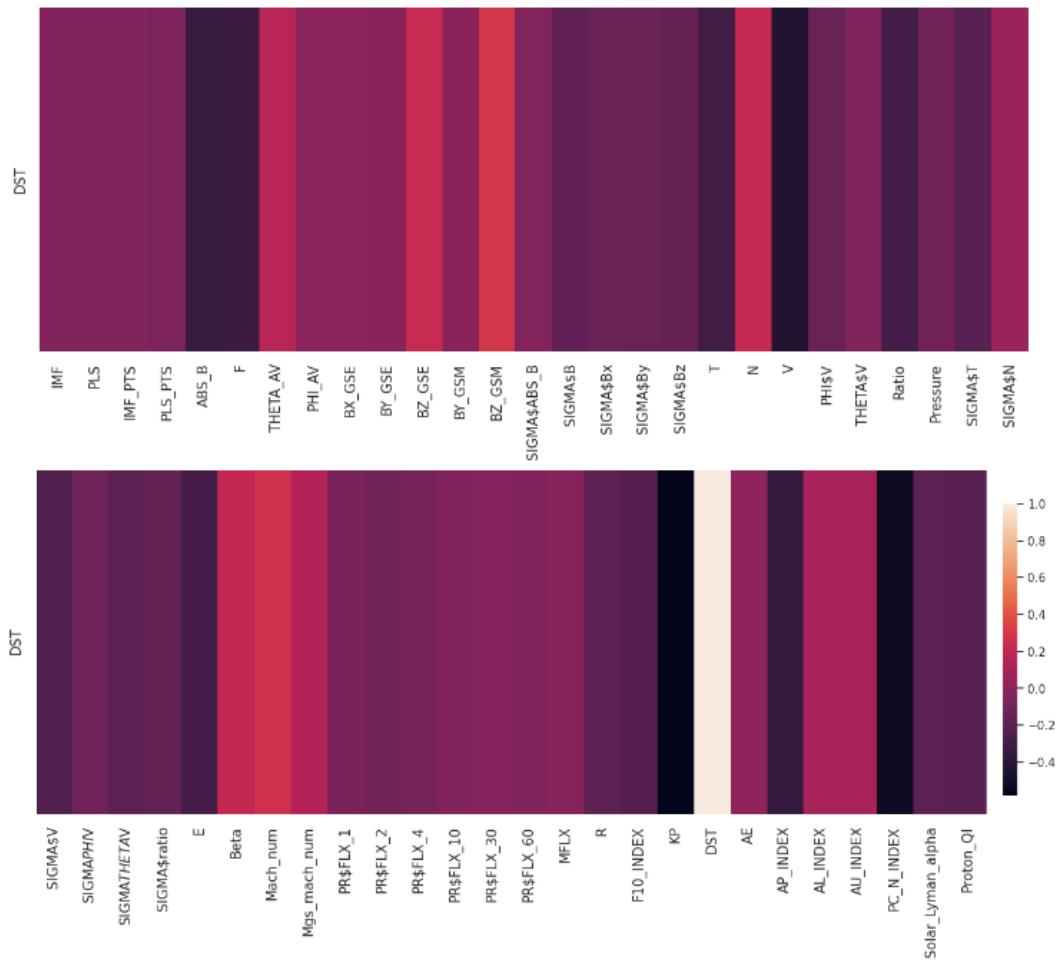
time1	0	BZ_GSM	124093	SIGMA\$N	165563	PR\$FLX_60	155455
Rot\$	0	SIGMA\$ABS_B	141086	SIGMA\$V	159753	MFLX	0
IMF	0	SIGMA\$B	124223	SIGMA\$PHI\$V	174407	R	0
PLS	0	SIGMA\$Bx	124622	SIGMA\$THETA\$V	208894	F10_INDEX	523
IMF PTS	0	SIGMA\$By	124609	SIGMA\$ratio	228207	KP	0
PLS PTS	0	SIGMA\$Bz	124606	E	151606	DST	0
ABS_B	124067	T	161020	Beta	176713	AE	0
F	124067	N	144145	Mach_num	160867	AP_INDEX	0
THETA_AV	124067	V	127822	Mgs_mach_num	176708	AL_INDEX	0
PHI_AV	124067	PHI\$V	159323	PR\$FLX_1	286726	AU_INDEX	0
BX_GSE	124067	THETA\$V	194992	PR\$FLX_2	323149	PC_N_INDEX	117482
BY_GSE	124067	Ratio	228207	PR\$FLX_4	323150	Solar_Lyman_alpha	41
BZ_GSE	124067	Pressure	144153	PR\$FLX_10	155408	Proton_QI	160870
BY_GSM	124093	SIGMA\$T	164786	PR\$FLX_30	155441		

Obrázok 3–2: Chýbajúce hodnoty v datasete

Nasledovala korelačná analýza medzi cielovým atribútom a ostatnými atribútmi – zistenie závislostí medzi atribútmi. Ak atribút výraznejšie koreluje, môže byť užitočný pre predikciu. Silu korelácie sme zistovali pomocou indikátora – Pearsonov korelačný koeficient.

Ak sa hodnoty korelačného koeficientu pohybujú okolo 0, atribúty navzájom nie sú závislé. Čím bližšie je hodnota k -1 alebo 1, sila korelácie rastie. Najvýraznejšia pozitívna korelácia sa zistila s atribútmi: BZ_GSM (0.29), Mach_num (0.26), BZ_GSE (0.22); negatívna korelácia s atribútmi: KP (-0.58), PC_N_INDEX (-0.53), V (-0.44). Korelácie vizualizuje heatmapa na Obrázku 3–3. Čím svetlejsia je farba, tým je vyššia pozitívna korelácia DST s daným atribútom. Farba približujúca sa k čiernej hlási výraznejšiu negatívnu koreláciu. Výsledky korelačnej analýzy zavážia aj pri výbere vhodných atribútov pre predikciu.

Výber atribútov pre predikciu – na základe korelačnej analýzy, počte chýbajúcich hodnôt a odporúčaní konzultantov, sme sa rozhodli predikciu urobiť najprv na základe samotného DST indexu a neskôr pridať atribúty F10_index a BZ_GSM. Aby sme mohli porovnať, či pridané atribúty zlepšia predikciu.

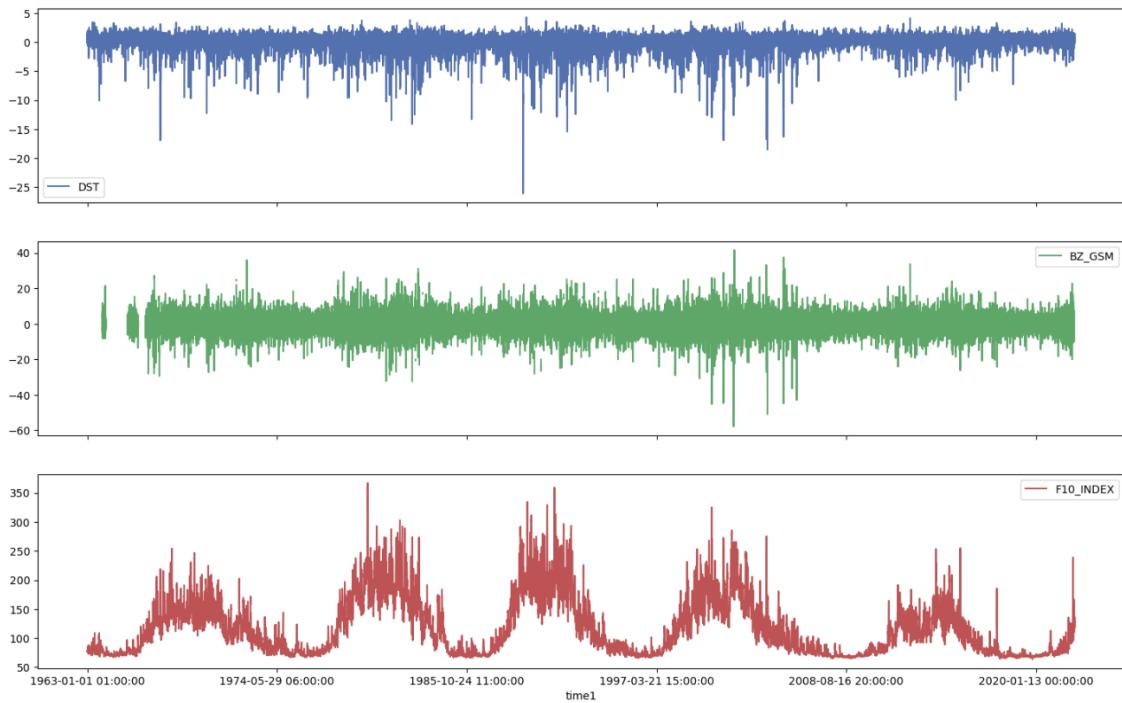


Obrázok 3 – 3: Heatmapa korelácií atribútov s DST

Vizualizácia hodnôt zvolených atribútov v čase – ukážka priebehu hodnôt atribútov DST, F10_index a BZ_GSM na Obrázku 3–4. Všimnime si, že atribút F10_INDEX má výraznú sezónnosť.

F10_INDEX je číslo, ktoré sa používa na meranie slnečnej rádiovej aktivity vo frekvenčnom pásme 10.7 cm. V časových radoch F10 indexu môžeme vidieť pravidelné sezónne vzory, ktoré sa prejavujú v cykloch slnečnej aktivity. Tieto cykly sa opakujú v intervaloch približne 11 rokov, pričom maximálna slnečná aktivita nastáva v polovici cyklu a minimálna aktivita na konci a na začiatku cyklu. Všimnime si tiež, že v časových úsekokoch, kde je vysoká hladina F10 indexu, je vyšší výskyt geomagnetických búrok. V sezónnosti všetkých 3 atribútov vidíme súvislost.

BZ_GSM označuje zložku medziplanetárneho magnetického poľa v smeroch sever-juh v súradnicovom systéme **GSM** (angl. *Geocentric Solar Magnetospheric coordinate system*). Rovnako ako DST je vyjadrený v jednotkách nanotesla.



Obrázok 3 – 4: Hodnoty atribútov DST, F10_index a BZ_GSM v čase

Stručný popis vybraných atribútov podľa OMNIweb portálu¹⁴:

- time1 – časová pečiatka v sekundách od 1. januára 1963 00:00:00 UTC,
- BZ_GSM – zložka magnetického poľa v smere Z v GSM súradničiach,
- F10_INDEX – solárny rádiový tok,
- DST – index magnetických búrok, meria zmenu v magnetickom poli Zeme.

Celá fáza pochopenia dát s kompletnými vizualizáciami a korelačnými analýzami sa nachádza v súbore s názvom `2_data_introduction.ipynb`, dostupné online¹⁵.

¹⁴Dostupné online: https://omniweb.gsfc.nasa.gov/html/ow_data.html

¹⁵Pochopenie dát: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/2_pochopenie_dat

3.3 Príprava dát

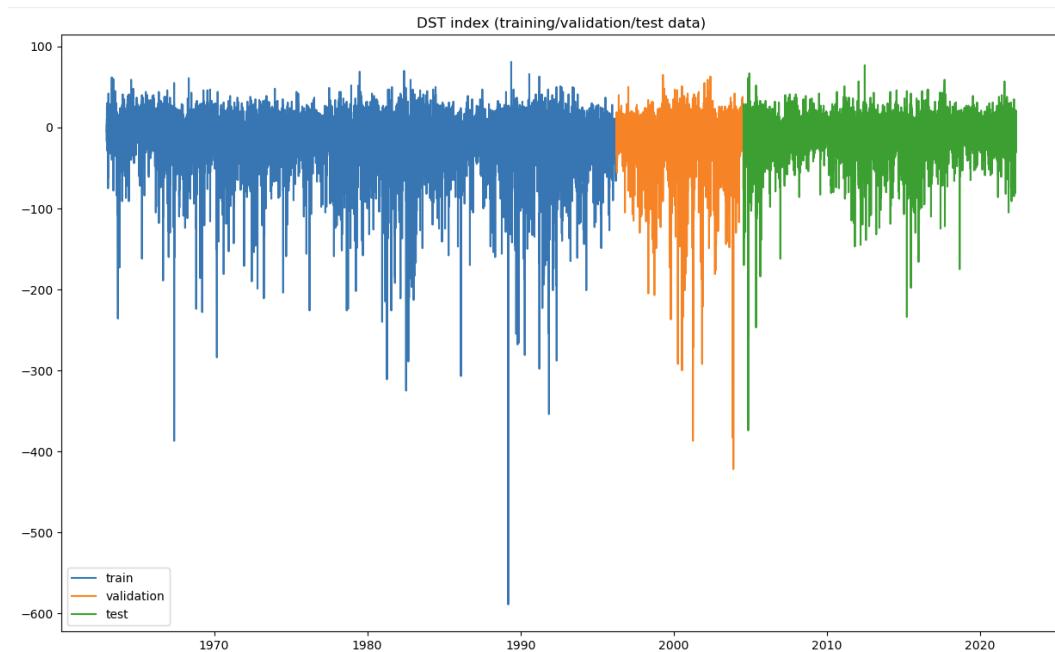
Fáza prípravy dát pozostáva z niekoľkých krokov potrebných pre ďalšiu fázu modelovania a vyhodnotenia. Najskôr sme začali predikovať DST index a na základe neho klasifikovať výskyt geomagnetických búrkov iba pomocou historických hodnôt samotného DST indexu, ktorý neobsahuje žiadne chýbajúce hodnoty. Pre potreby predikcie blížiacej sa geomagnetickej búrky si musíme odvodiť nové atribúty posunutého DST indexu v čase, napríklad nový atribút DST+1 predstavuje budúcu hodnotu DST indexu o 1 hodinu vopred. Analogicky si vytvárame nové atribúty DST+2, DST+3, DST+4, DST+6, DST+8, DST+12, DST+24 a DST+48.

Dataset s novými atribútmi si pre potreby trénovania a vyhodnocovania modelu musíme rozdeliť na trénovaciu a testovaciu množinu. Dôležité je, aby testovacia množina bola pri všetkých experimentoch rovnaká, zabezpečí sa tak porovnatelnosť výsledkov každého natrénovaného modelu. Do testovacej množiny si preto odberieme 30% zo všetkých dát – z konca dátovej množiny, čiže v testovacej množine máme najaktuálnejšie záznamy (približne od roku 2005 po súčasnosť) a pre trénovacie účely nám ostali staršie záznamy (1963 po 2005).

V testovacej množine sme po konzultácii nechali iba atribúty time1, BZ_GSM, F10_INDEX, DST, DST+1, DST+2, DST+3, DST+4, DST+6, DST+8, DST+12, DST+24 a DST+48. Tieto atribúty sme sa rozhodli ďalej využívať na predikciu DST indexu na základe výsledkov korelačných analýz a odporúčaní Ústavu experimentálnej fyziky Slovenskej akadémie vied, v.v.i. Posledným krokom bolo odstránenie chýbajúcich hodnôt a uloženie jednotnej testovacej množiny do súboru `test_omni.csv`. Vznikla teda testovacia množina s 13 atribútmi a 155 259 záznamami.

Trénovaciu množinu operatívne upravujeme pre potreby každého trénovania modelu podľa toho, ktoré atribúty práve využívame, je uložená v súbore `train_omni.csv`. Obsahuje 364 310 záznamov. Pred trénovaním sa z trénovacej množiny operatívne odčlení 20% do validačnej množiny, ktorá slúži na priebežné vyhodnocovanie modelu počas jeho učenia. Na Obrázku 3–5 vidíme farebné rozdelenie hodnôt DST indexu

na trénovaciu množinu – modrou, validačnú – oranžovou a testovaciu – zelenou farbou.



Obrázok 3 – 5: Rozdelenie dátovej množiny

Všetky súbory z fázy prípravy dát s kompletnými postupmi, ako aj rozdelený dataset sa nachádzajú v priečinku s názvom `3_priprava_dat`, dostupné online¹⁶.

¹⁶Príprava dát: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/3_priprava_dat

3.4 Modelovanie

Fáza modelovania zahŕňa návrh, trénovanie a priebežné vyhodnocovanie modelu RNN s LSTM vrstvou. Navrhli sme jednoduchú RNN:

- s LSTM vrstvou,
- lineárnu aktivačnou funkciou,
- chybovou funciou MSE,
- model ukladáme na základe MAE,
- používame optimalizačnú funkciu Adam,
- proti preučeniu modelu používame Dropout 10% a Early Stopping.

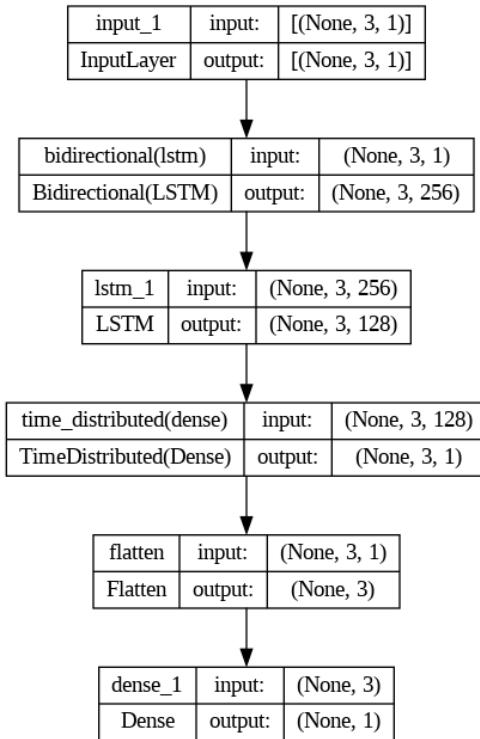
Obrázok 3–6 zobrazuje architektúru modelu spolu s veľkosťou vstupov na jednotlivých vrstvách a celkovú sumarizáciu modelu vidíme na Obrázku 3–7.

Modelovanie a priebežné vyhodnocovanie prebieha v troch na seba nadvádzajúcich kolách experimentov, a všetky súbory z fázy modelovania s kompletnými postupmi, priebežným vyhodnotením predikcie a metrík sa nachádzajú v priečinku s názvom **4_modelovanie**, dostupné online¹⁷.

Tri kolá experimentov boli zamerané na zistenie:

- 1. kolo – aké veľké časové okno zvoliť pre predikciu.
- 2. kolo – s akým najväčším časovým predstihom vieme stále kvalitne predikovať geomagnetickú búrku.
- 3. kolo – aký vplyv má pridanie ďalších atribútov na kvalitu predikcie.

¹⁷Modelovanie: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/4_modelovanie



Obrázok 3 – 6: Architektúra RNN s veľkosťou vstupov na jednotlivých vrstvách

Layer (type)	Output Shape	Param #
<hr/>		
input_1 (InputLayer)	[(None, 3, 1)]	0
bidirectional (Bidirectional)	(None, 3, 256)	133120
lstm_1 (LSTM)	(None, 3, 128)	197120
time_distributed (TimeDistributed)	(None, 3, 1)	129
flatten (Flatten)	(None, 3)	0
dense_1 (Dense)	(None, 1)	4
<hr/>		
Total params: 330,373		
Trainable params: 330,373		
Non-trainable params: 0		
<hr/>		

Obrázok 3 – 7: Sumarizácia modelu

Hľadanie najvhodnejšej šírky časového okna

V prvom kole experimentov sme predikovali súčasnú hodnotu DST indexu a na základe neho klasifikovali výskyt geomagnetickej búrky tak, že sme použili iba DST index a jeho historické hodnoty s rôznou šírkou časového okna.

Model predikoval DST index postupne na základe jeho 3, 6, 8, 12, 24 a 48 predchádzajúcich nameraných hodnôt. Následne sme porovnali výsledky klasifikácie a výbrali najvhodnejšiu šírku časového okna. Na Obrázkoch 3–8 a 3–9 vidíme priebeh metrík TPR, TNR, PPV, NPV, ACC a MCC. Vysvetlenie týchto metrík nájdeme v teoretickej časti práce 1.5. Dynamika vizualizácie nám napovie, ako závisí kvalita predikcie od šírky časového okna. Každý bod na krivkách predstavuje nový model. Jediná vec, v ktorej sa modely líšia je práve šírka časového okna. Obidva grafy majú na osi x použitú šírku časového okna a na osi y hodnotu danej metriky. Pri značkách vidíme presné hodnoty metrík zaokrúhlené na 4 desatinné miesta. Hraničná hodnota pre záchytenie búrky je nastavená na -20 nT.

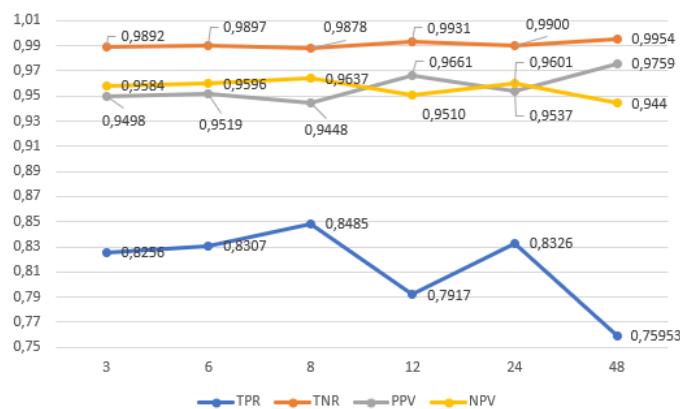
Zistili sme, že nie je vhodné použiť krátke 3-hodinové okno a ani príliš dlhé 12, 24 a 48-hodinové okno. Tieto časové okná dosiahli horšie výsledky TPR, ACC a MCC. Po zvážení výsledkov experimentu sme sa rozhodovali medzi 6 a 8 hodinovým časovým oknom. Úspešnosť klasifikácie 8 hodinového časového okna bola súčasťou o 0.002 vyššia, no nakoniec sme zvolili časové okno 6 hodín, pretože pri predikcii časových radov býva použitie menšieho časového okna efektívnejšie. Tu je niekoľko dôvodov:

- Aktuálnejšie informácie: použitie menšieho časového okna zahŕňa aktuálnejšie informácie do predikcie. Je to užitočné, keď sa časový rad mení rýchlo a zahrnutie starších dát by skreslilo predikciu.
- Citlivosť na rýchle zmeny: časový rad bude obsahovať menej dát, ale bude mať väčšiu citlivosť na rýchle zmeny.
- Zabránenie preučeniu: použitie menšieho časového okna môže pomôcť zabrániť preučeniu.

- Kratší výpočtový čas: menšie časové okno znamená menej dát na spracovanie a trénovanie sa urýchli.

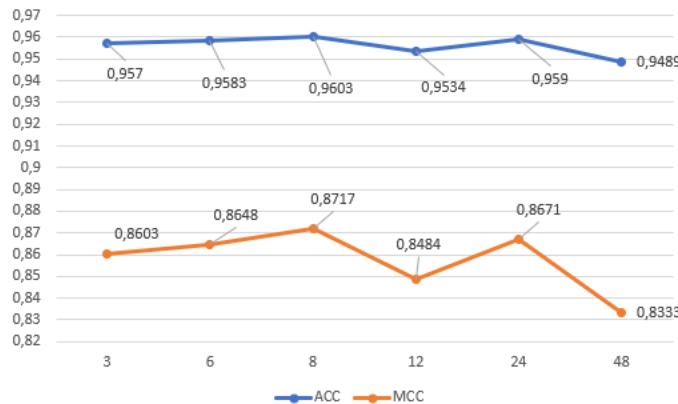
Podrobnejšie výsledky klasifikácie modelu s časovým oknom 6 hodín vidíme v kontingenčnej Tabuľke 3–2. Tabuľka 3–3 zobrazuje výsledky ďalších metrík, ako presnosť, návratnosť, F1-skóre a úspešnosť klasifikácie. Na Obrázku 3–10 vidíme graf predikcií a skutočných hodnôt. Môžeme si všimnúť, že predikcie vyzerajú veľmi úspešne, avšak náš model nadhodnocuje negatívne extrémne hodnoty a zároveň podhodnocuje pozitívne extrémne hodnoty. Tento problém môžeme neskôr pri klasifikácii vyriešiť posunutím prahovej hodnoty klasifikácie.

Všetky súbory z tohto kola experimentov, natréované modely a uložené .csv predikcie sa nachádzajú v podpriečinku s názvom **Experiment_1**, dostupné online¹⁸.



Obrázok 3–8: Vyhodnotenie TPR, TNR, PPV a NPV experimentu 1

¹⁸Experiment 1: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/4_modelovanie/Experiment_1



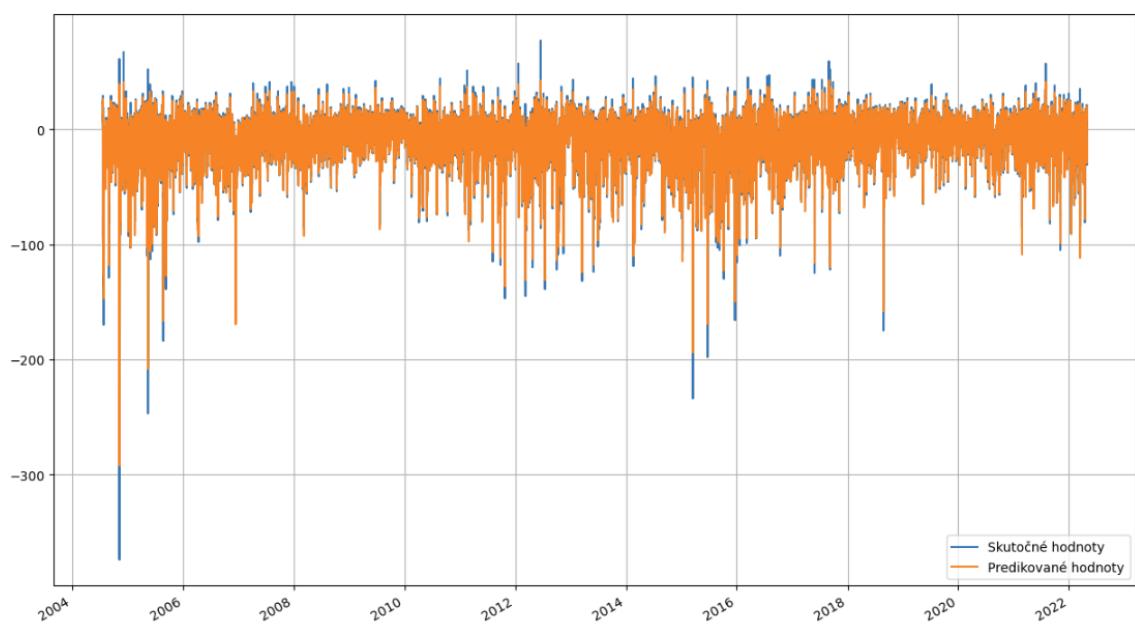
Obrázok 3 – 9: Vyhodnotenie ACC a MCC experimentu 1

Tabuľka 3 – 2: Kontingenčná tabuľka modelu s časovým oknom 6 hodín

		predikované hodnoty		
		0	1	
skutočné hodnoty	0	123 350	1 283	
	1	5 182	25 438	

Tabuľka 3 – 3: Výsledky metrík modelu s časovým oknom 6 hodín

	presnosť	návratnosť	F1-skóre	dátové položky
0	0.96	0.99	0.97	124 633
1	0.95	0.83	0.89	30 620
úspešnosť			0.96	155 253
macro avg	0.96	0.91	0.93	155 253
weighted avg	0.96	0.96	0.96	155 253



Obrázok 3 – 10: Predikované a skutočné hodnoty DST indexu experimentu 1
(pomocou 6-tich predchádzajúcich hodinových záznamov pred predikovanou hodnotou)

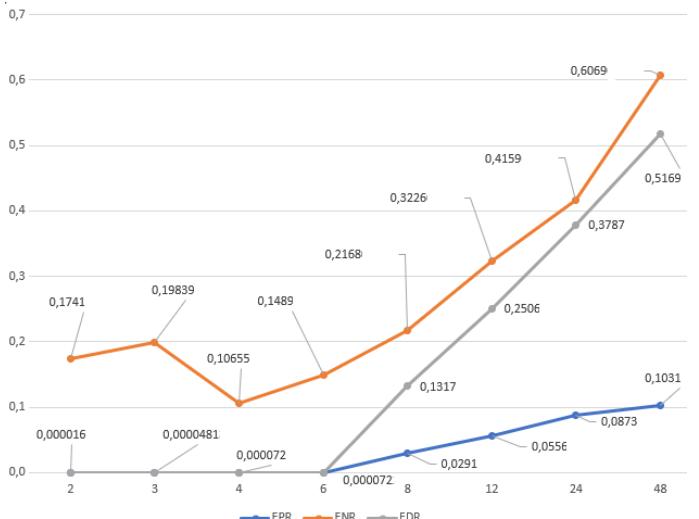
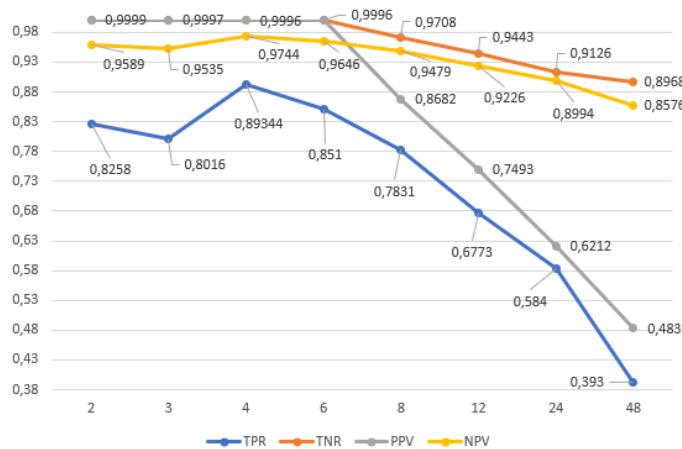
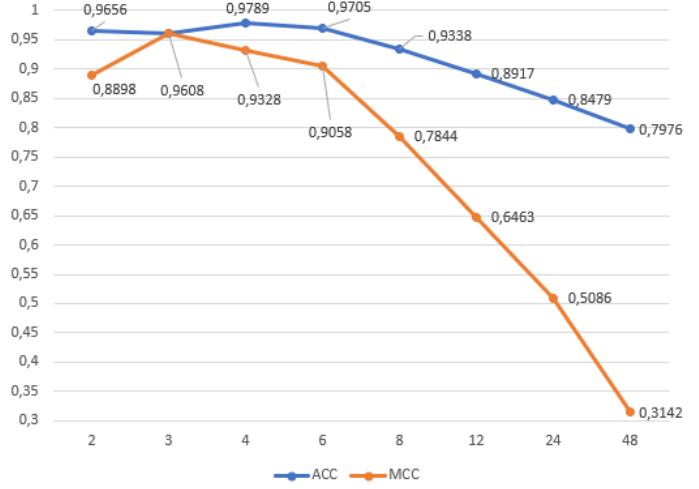
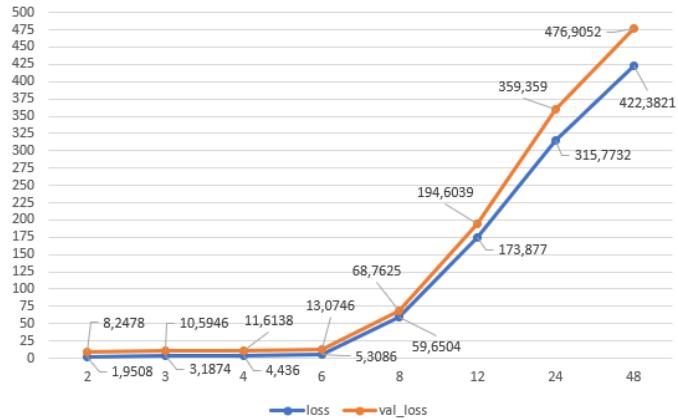
Predikcia rôzne vzdialených budúcich hodnôt

Druhé kolo experimentov sa venuje predikcie budúcich hodnôt DST indexu s predstihom 2, 3, 4, 6, 8, 12, 24 a 48 hodín. Využijeme poznatok z predchádzajúceho kola experimentov a veľkosť časového okna nastavíme na 6 hodín, to znamená, že budúca hodnota sa predikuje na základe predchádzajúcich 6 nameraných hodnôt. Predikované atribúty sú postupne DST+2, DST+3, DST+4, DST+6, DST+8, DST+12, DST+24 a DST+48. Každý natrénovaný model sme vyhodnotili pomocou metrík.

Na paneloch Obrázku 3 – 4 vidíme priebeh metrík loss – MSE, val_loss – val_MSE, TPR, TNR, PPV, NPV, FPR, FNR, FDR, ACC a MCC. Všetky tieto grafy majú na osi x časový predstih predikcie a na osi y hodnotu danej metriky. Hraničná hodnota pre zachytenie búrky je nastavená na -20 nT.

Z vizualizácií vieme vyčítať, aká kvalitná je predikcia a kolko hodín dopredu vieme určiť DST index, než sa výsledky výrazne pokazia. Každý bod na krvkách predstavuje nový model. Jediná vec, v ktorej sa modely líšia je vzdialenosť predikovaného atribútu od súčasnej hodnoty v čase dopredu. Napríklad predikcia DST+24 nám povie, aký DST index bude o 1 deň. Všimnime si, že každá jedna metrika zaznamenala výrazný zlom a rapídne zhoršila výsledky a tým aj kvalitu predikcie, ak sme sa snažili predpovedať hodnotu DST indexu vzdialenejšiu ako 6 hodín dopredu. Z toho vyplýva, že predstih 6 hodín je najväčší, pri ktorom si vieme zachovať kvalitu predikcie. Kedže cieľom práce je predikovať geomagnetickú búrku s čo najväčším časovým predstihom, považujme časový predstih 6 hodín zatial za najúspešnejší. Prejdeme preto na ďalšie kolo experimentov a zistíme, ako model zareaguje na pridanie ďalších vstupných atribútov do predikcie.

Podrobnejšie výsledky klasifikácie tohto modelu vidíme v kontingenčnej Tabuľke 3 – 5. Tabuľka 3 – 6 zobrazuje výsledky ďalších metrík, ako presnosť, návratnosť, F1-skóre a úspešnosť klasifikácie. Na Obrázku 3 – 11 vidíme graf predikcií a skutočných hodnôt tohto modelu. Všetky súbory z tohto kola experimentov, natrénované modely a uložené predikcie sa nachádzajú v podpriečinku s názvom **Experiment_2**.

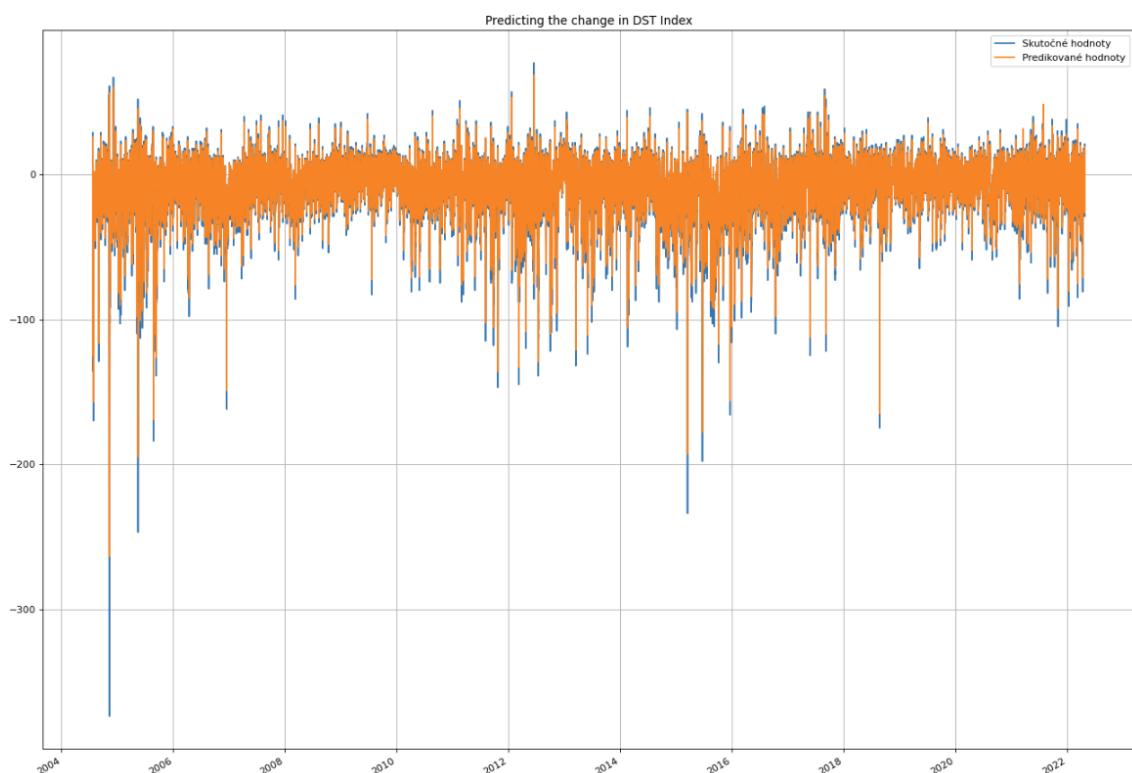
Tabuľka 3 – 4: Vizualizácie metrík z experimentu 2**Tabuľka 3 – 5:** Kontingenčná tabuľka modelu predikujúceho DST+6

(pomocou 6-tich predchádzajúcich hodinových záznamov pred predikovanou hodnotou)

		predikované hodnoty	
		0	1
skutočné hodnoty	0	124 633	9
	1	4 581	26 030

Tabuľka 3 – 6: Výsledky metrík modelu predikujúceho DST+6
 (pomocou 6-tich predchádzajúcich hodinových záznamov pred predikovanou hodnotou)

	presnosť	návratnosť	F1-skóre	dátové položky
0	0.96	1.00	0.98	124 642
1	1.00	0.85	0.92	30 611
úspešnosť			0.97	155 253
macro avg	0.98	0.93	0.95	155 253
weighted avg	0.97	0.97	0.97	155 253



Obrázok 3 – 11: Predikované a skutočné hodnoty DST+6 experimentu 2
 (pomocou 6-tich predchádzajúcich hodinových záznamov pred predikovanou hodnotou)

Pridanie atribútov pre predikciu rôzne vzdialenosťí budúcich hodnôt DST indexu s časovým oknom 6 hodín

V treťom kole experimentov sme sa rozhodli k DST indexu pridať atribúty F10_INDEX a BZ_GSM. Na predikciu budúcich hodnôt DST indexu teda používame 3 atribúty. Predikované atribúty sú DST+2, DST+3, DST+4, DST+6, DST+8, DST+12, DST+24 a DST+48. Šírka časového okna je 6 hodín. Sledujeme, ako sa zmení kvalita predikcie s pridanými atribútmi.

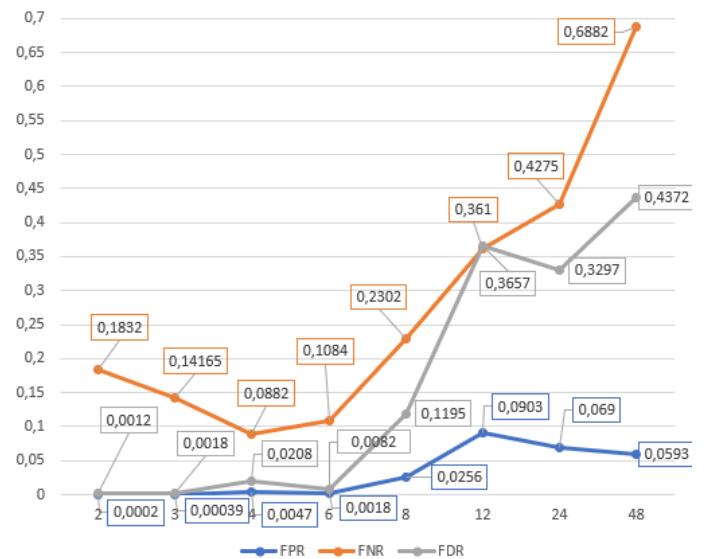
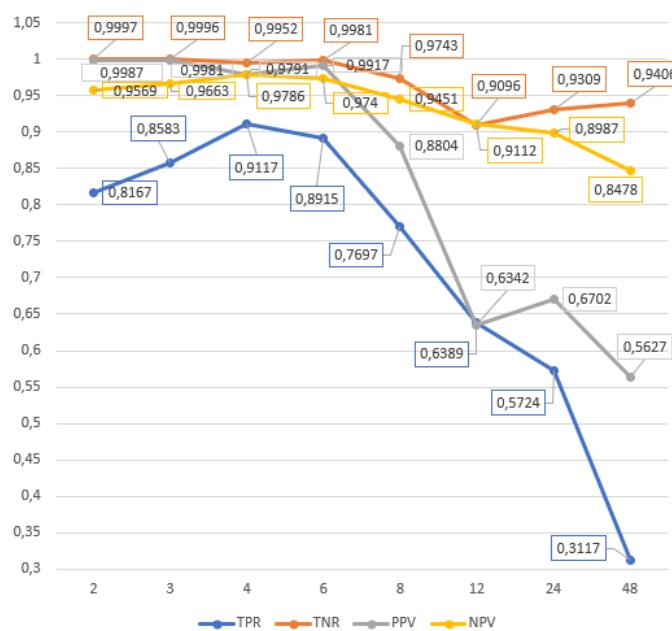
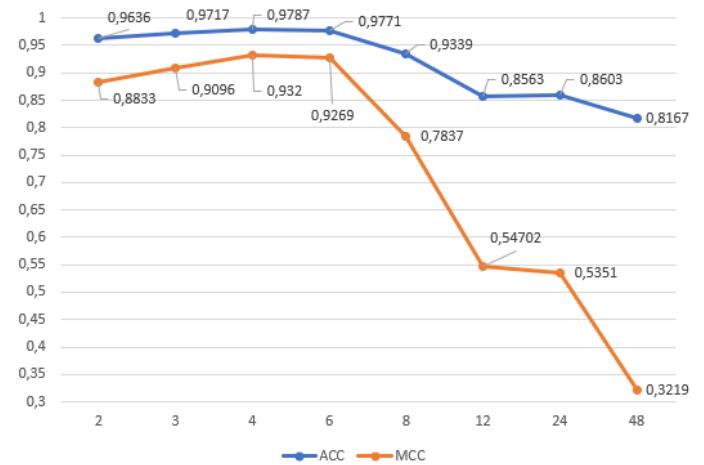
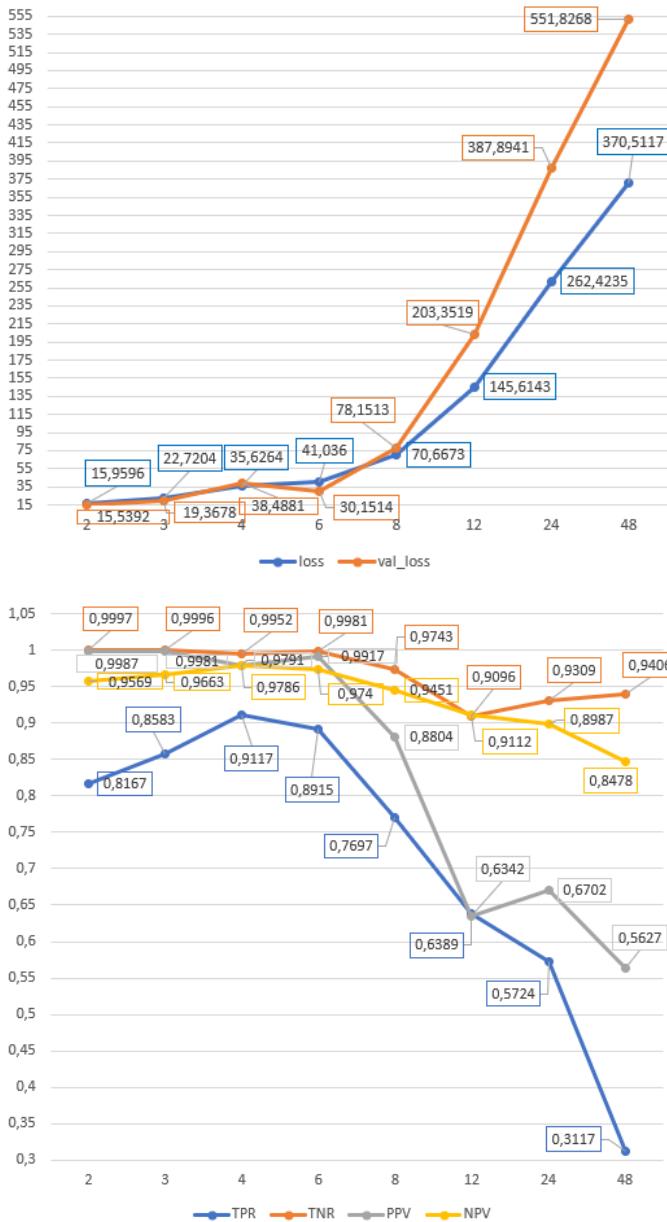
Opäť sme každý natrénovaný model vyhodnotili pomocou metrík. Vyhodnotenia vidíme na paneloch Obrázku 3 – 7. Analogicky, ako pri predchádzajúcim experimente porovnávame metriky loss – MSE, val_loss – val_MSE, TPR, TNR, PPV, NPV, FPR, FNR, FDR, ACC a MCC. Všetky tieto grafy majú na osi x časový predstih predikcie a na osi y hodnotu danej metriky. Pri značkách vidíme presné hodnoty metrík zaokrúhlené na 4 desatinné miesta. Hraničná hodnota pre zachytenie búrky je nastavená na -20 nT.

Najväčší časový predstih, pri ktorom sa nám podarilo zachovať kvalitu predikcie je model predikujúci DST s predstihom 6 hodín. Oproti experimentu 2 sme zaznamenali zlepšenie.

Podrobnejšie výsledky klasifikácie tohto modelu vidíme v kontingenčnej Tabuľke 3 – 8. Tabuľka 3 – 9 zobrazuje výsledky ďalších metrík, ako presnosť, návratnosť, F1-skóre a úspešnosť klasifikácie. Na Obrázku 3 – 12 vidíme graf predikcií a skutočných hodnôt tohto modelu.

Všetky súbory z tohto kola experimentov, natrénované modely a uložené .csv predikcie sa nachádzajú v podpriečinku s názvom **Experiment_3**, dostupné online¹⁹.

¹⁹Experiment 3: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/4_modelovanie/Experiment_3

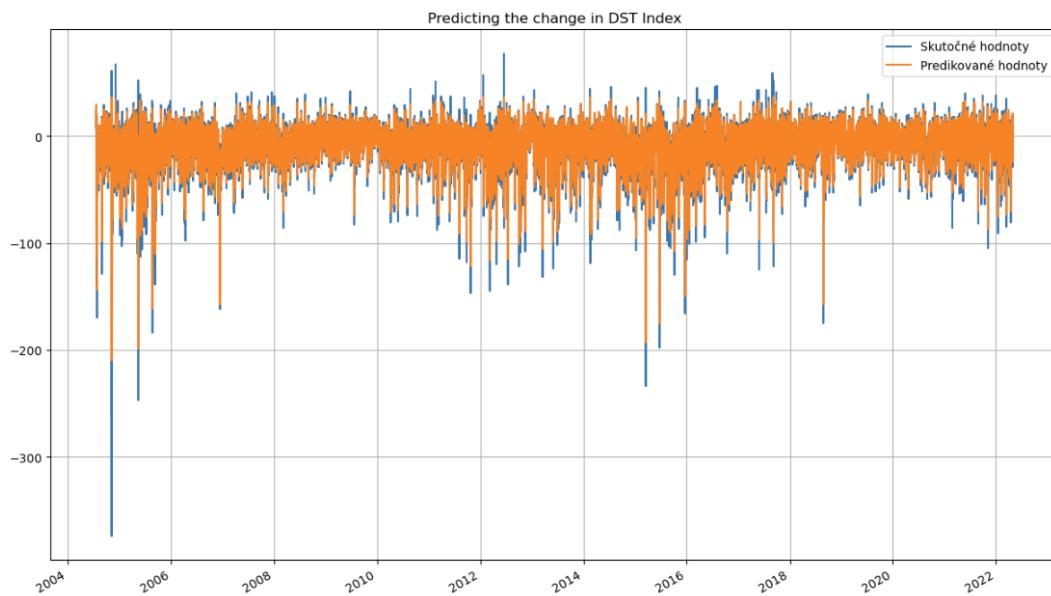
Tabuľka 3 – 7: Vizualizácie metrík z experimentu 3

Tabuľka 3 – 8: Kontingenčná tabuľka modelu predikujúceho DST+6 pomocou pridaných atribútov

		predikované hodnoty	
		0	1
skutočné hodnoty	0	124 416	226
	1	3 321	27 290

Tabuľka 3 – 9: Výsledky metrík modelu predikujúceho DST+6 pomocou pridaných atribútov

	presnosť	návratnosť	F1-skóre	dátové položky
0	0.97	1.00	0.99	124 642
1	0.99	0.89	0.94	30 611
úspešnosť			0.98	155 253
macro avg	0.98	0.94	0.96	155 253
weighted avg	0.98	0.98	0.98	155 253



Obrázok 3 – 12: Predikované a skutočné hodnoty DST indexu experimentu 3
(pomocou pridaných atribútov)

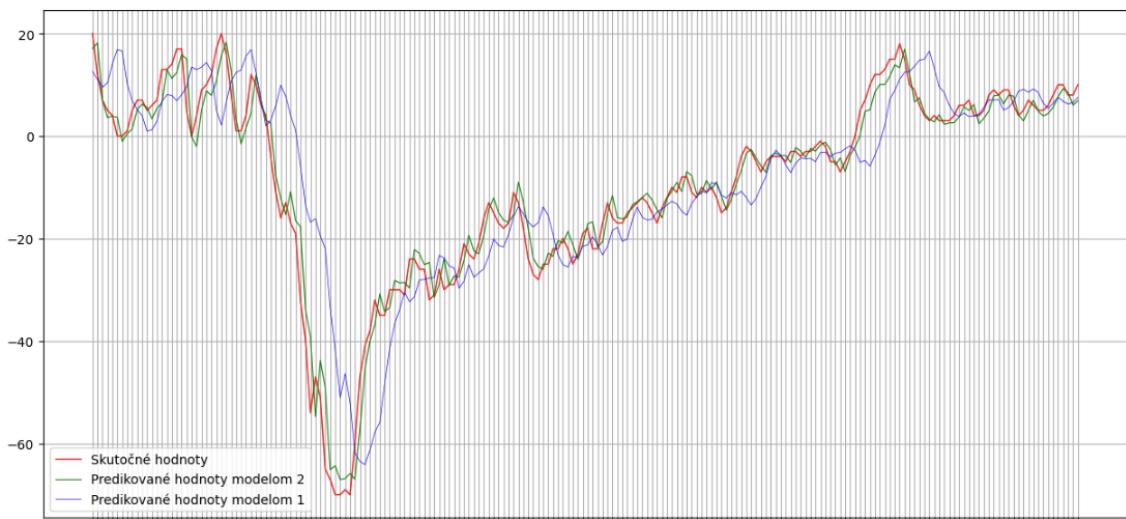
3.5 Vyhodnotenie

V tejto fáze si zhrnieme výsledky práce. Tabuľka 3 – 10 sumarizuje finálne výsledky dvoch najdôležitejších modelov. Model 1 – model z experimentu 2, model je založený iba na hodnotách DST indexu, predikuje DST index s časovým predstihom 6 hodín. Model 2 – model z experimentu 3, predikciu určuje na základe 3 atribútov: DST index, F10_INDEX a BZ_GSM. Pridaním atribútov sa model vylepšil vo viacerých metrikách. Úspešnosť stúpla o 1%, návratnosť geomagnetických búrok stúpla o 4%, F1-skóre búrok stúpla o 2% a rovanko aj Matthewsov korelačný koeficient.

Zlepšenie pozorujeme aj na Obrázku 3 – 13, kde je priblížený graf predikcií. Skutočné hodnoty zobrazuje krivka červenej farby. Krivka zelenej farby zobrazuje hodnoty, ktoré predikoval Model 2. Modrá farba reprezentuje Model 1. Vidíme, že zelená krivka sa podstatne viac blíži k skutočným hodnotám ako modrá. Aj tento fakt potvrdzuje vyššiu úspešnosť Modelu 2.

Tabuľka 3 – 10: Sumarizácia výsledkov pre predikciu so 6 hodinovým predstihom

metrika		Model 1	Model 2
úspešnosť		0.97	0.98
presnosť	0	0.96	0.97
	1	1.00	0.99
návratnosť	0	1.00	1.00
	1	0.85	0.89
F1-skóre	0	0.98	0.99
	1	0.92	0.94
korelačný koeficient		0.91	0.93



Obrázok 3 – 13: Predikcie oboch modelov v porovnaní so skutočnými hodnotami

Pre potreby porovnania s výskumom Gruet et al. (2018) sme klasifikáciu upravili. Použili sme klasifikáciu do 3 tried namiesto binárnej klasifikácie. Testovacie množiny týchto 2 výskumov sú odlišné, toto porovnanie má iba informatívny charakter, či sa nám podarilo priblížiť k výsledkom úspešných modelov.

Pri klasifikácii sme rozdelili predikcie do 3 tried. Pre vylepšenie klasifikácie sme použili metódu posúvania prahov pre predikcie, ktorá dobre funguje najmä, ak má model tendenciu podhodnocovať alebo nadhodnocovať skutočnú hodnotu. Upravené prahové hodnoty sme použili iba pri našich vygenerovaných predikciách. Prahouvé hodnoty pre testovaciu množinu ostali nezmenené.

Pôvodné hraničné hodnoty pre DST index sú -250 pre zachytenie superbúrky a -50 pre zachytenie silnejšej búrky. Hodnoty nad -50 sú považované za normálne. Prahy -250 sme kvôli nadhodnocovaniu modelu posunuli na -180 a hodnotu -50 na -40. Tieto hodnoty sme určili na základe grafu predikcií, kde model extrémne hodnoty kvôli nízkemu výskytu nadhodnocuje. Čím sú hodnoty bližšie k 0, tým je nadhodnotenie nižšie, pretože ich výskyt je častejší.

Údaje v Tabuľke 3 – 11 porovnávajú výskum s našimi výsledkami. Tieto výsledky ukazujú zlepšenie vo všetkých triedach a oboch metrikách oproti výskumu Gruet

et al. (2018) (Table 3, strana 10-11, zvýraznené). Bližšie výsledky prispôsobenej klasifikácie interpretuje kontingenčná Tabuľka 3 – 12 a Tabuľka 3 – 13. Zdrojový kód fázy vyhodnotenia sa nachádza v priečinku s názvom `5_vyhodnotenie`, dostupné online²⁰.

Tabuľka 3 – 11: Výsledky Gruet et al. (2018) v porovnaní s vlastnými výsledkami

		Gruet et al. (2018)				vlastné výsledky	
trieda	DST	o 1 hod vopred		o 6 hod vopred		o 6 hod vopred	
		TPR	FPR	TPR	FPR	TPR	FPR
1	DST<-250	0.969	$6.40 \cdot 10^{-4}$	0.50	$8.34 \cdot 10^{-3}$	1.00	$1.09 \cdot 10^{-4}$
2	-250<DST<-50	0.927	0.0719	0.859	0.167	0.9935	0.0122
3	DST>-50	0.929	0.0705	0.834	0.130	0.9878	0.0015

Tabuľka 3 – 12: Kontingenčná tabuľka prispôsobenej klasifikácie do 3 tried

		predikované hodnoty		
		1	2	3
skutočné hodnoty	1	11	0	0
	2	17	3 358	5
	3	0	1 856	150 006

²⁰Vyhodnotenie: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/5_vyhodnotenie

Tabuľka 3 – 13: Výsledky metrík prispôsobenej klasifikácie do 3 tried

	presnosť	návratnosť	F1-skóre	dátové položky
1	0.39	1.00	0.58	11
2	0.64	0.99	0.78	3 380
3	1.00	0.99	0.99	151 862
úspešnosť			0.99	155 253
macro avg	0.68	0.99	0.78	155 253
weighted avg	0.99	0.99	0.99	155 253

3.6 Nasadenie

Vo fáze nasadenia sme vytvorili jednoduché používateľské rozhranie vo formáte Jupyter Notebook. Používateľ si môže vybrať, či chce použiť model predikujúci iba na základe DST indexu alebo model vychádzajúci z 3 atribútov: DST, F10_INDEX a BZ_GSM.

Používateľ musí zadať vstupné hodnoty a načítať natrénovaný model zo súboru a na výstupe dostane jedno číslo – predikciu DST indexu očakávaného o 6 hodín. Celá práca a jej výstup poslúžia na výskumné účely na Ústave experimentálnej fyziky Slovenskej akadémie vied, v.v.i.

Súbor z fázy nasadenia sa nachádza v priečinku s názvom `6_nasadenie`, dostupné online²¹.

²¹Nasadenie: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/6_nasadenie

Záver

Táto diplomová práca sa zaoberala predikciou geomagnetických búrok pomocou hlbokého učenia. Práca vznikla v spolupráci s Ústavom experimentálnej fyziky Slovenskej akadémie vied, v.v.i. a poslúži na ich výskumné účely. Motiváciou pre predikciu geomagnetických búrok je minimalizácia ich negatívneho vplyvu na človeka, životné prostredie a moderné technológie. Hlboké učenie ako odvetvie sa neustále a veľmi rýchlo vyvíja. Hlboké neurónové siete sú veľmi dobrým a úspešným nástrojom na rôzne typy predikcie. Hlavnou metódou hlbokého učenia použitou v tejto práci bola rekurentná neurónová sieť.

Teoretická časť práce sa zaoberala stručným úvodom do strojového učenia a podrobnejšie do jeho odvetvia – hlbokého učenia. Obsahuje všetky teoretické základy, ktoré sú potrebné na porozumenie práci ako celku. Doménová časť práce sa zaoberala samotnými geomagnetickými búrkami. Stručne popisuje príčiny ich vzniku, vplyvy, prejavy a stupne intenzity. Poskytuje prehľad výskumov, ktoré sa venovali podobnej téme. Praktická časť podrobne popisuje celý proces výskumu od fázy pochopenia ciela, cez prípravu dát, modelovanie, vyhodnotenie až po nasadenie. Celý proces zaznamenávajú pomocné vizualizácie a reporty, ktoré sprehľadňujú jednotlivé chronologické fázy.

Cieľom diplomovej práce bolo navrhnúť model rekurentnej neurónovej siete, ktorý by bol schopný poskytovať predikcie indexov kozmického počasia s niekoľko hodinovým predstihom. Ciel bol splnený a natrénovali sme viacero modelov. Výstupom práce sú 2 najúspešnejšie predikčné modely.

Prvý model predikuje DST index s časovým predstihom 6 hodín iba na základe jeho historických hodnôt a používa časové okno 6 hodín. Tento model poskytuje úspešnosť klasifikácie 97%. Jeho presnosť pre negatívne prípady je 96% a pre pozitívne 100%. Návratnosť negatívnych prípadov 100%, pozitívnych 85%. F1-skóre pre negatívne prípady je 98%, pre pozitívne 92%. Matthewsov korelačný koeficient je 91%.

Druhý model predkuje index DST s rovnakým časovým predstihom 6 hodín a rovnakým 6-hodinovým časovým oknom, avšak k vstupným atribútom pribudli F10_INDEX a BZ_GSM. Úspešnosť klasifikácie tohto modelu je 98%. Jeho presnosť pre negatívne prípady je 97% a pre pozitívne 99%. Návratnosť negatívnych prípadov 100%, pozitívnych 89%. F1-skóre pre negatívne prípady je 99%, pre pozitívne 94%. Matthewsov korelačný koeficient je 93%.

K práci vznikne aj článok s cieľom publikácie v blízkej budúcnosti. V budúcnosti by práca mohla pokračovať vývojom a optimalizáciou iných architektúr rekurentných neurónových sietí, prípadne by sa na predikciu mohla použiť iná kombinácia vstupných atribútov. Užitočná by určite bola aj optimalizácia hyperparametrov siete. Dátová množina, ktorej sa výskum venuje, sa s plynúcim časom neustále zväčšuje, čo bude viest k zlepšeniu predikcie v budúcych výskumoch.

Literatúra

Andrejková, G., Tóth, H. and Kudela, K. (2004). Fuzzy neural networks in prediction of geomagnetic storms.

Astafyeva, E., Yasyukevich, Y., Maksikov, A. and Zhivetiev, I. (2014). Geomagnetic storms, super-storms, and their impacts on gps-based navigation systems, *Space Weather* **12**(7): 508–525.

Costa, C. D. (2020). Best python libraries for machine learning and deep learning.

URL: <https://towardsdatascience.com/best-python-libraries-for-machine-learning-and-deep-learning-b0bd40c7e8c>

Gonzalez, W., Joselyn, J.-A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. and Vasyliunas, V. (1994). What is a geomagnetic storm?, *Journal of Geophysical Research: Space Physics* **99**(A4): 5771–5792.

Gruet, M. A., Chandorkar, M., Sicard, A. and Camporeale, E. (2018). Multiple-hour-ahead forecast of the dst index using a combination of long short-term memory neural network and gaussian process, *Space Weather* **16**(11): 1882–1896.

Hinton, G., Srivastava, N. and Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent, *Cited on* **14**(8).

Hu, A., Camporeale, E. and Swiger, B. (2022). Multi-hour ahead dst index prediction using multi-fidelity boosted neural networks, *arXiv preprint arXiv:2209.12571*.

Janocha, K. and Czarnecki, W. M. (2017). On loss functions for deep neural networks in classification, *arXiv preprint arXiv:1702.05659*.

Keskar, N. S. and Socher, R. (2017). Improving generalization performance by switching from adam to sgd, *arXiv preprint arXiv:1712.07628*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*.

Kugblenu, S., Taguchi, S. and Okuzawa, T. (1999). Prediction of the geomagnetic storm associated dst index using an artificial neural network algorithm, *Earth Planets Space*.

Lakhina, G. S. and Tsurutani, B. T. (2016). Geomagnetic storms: historical perspective to modern view, *Geoscience Letters* **3**: 1–11.

Leone, D. (2023). Nasa mulls participation in sino-european small space mission.

URL: <https://spacenews.com/nasa-mulls-participation-in-sino-european-small-space-mission/>

Lethy, A., El-Eraki, M. A., Samy, A. and Deebees, H. A. (2018). Prediction of the dst index and analysis of its dependence on solar wind parameters using neural network.

URL: <https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018SW001863>

Lundstedt, H., Gleisner, H. and Wintoft, P. (2002). Operational forecasts of the geomagneticdstindex, *Geophysical Research Letters* **29**(24).

Machová, K. (2002). *Machine Learning. Principles and Algorithms.*, ELFA s.r.o.

URL: <https://kristina.machova.website.tuke.sk/pdf/SU4.pdf>

Medsker, L. R. and Jain, L. (2001). Recurrent neural networks, *Design and Applications* **5**: 64–67.

Papitashvili, N. (n.d.). Spdf - about omniweb data.

URL: https://omniweb.gsfc.nasa.gov/html/ow_data.html

Sharma, S. (2017). Activation functions in neural networks.

Sinčák, P. and Andrejková, G. (1996). Neurónové siete inžiniersky prístup (1. diel), *Elfa: Kosice*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research*.

Wythoff, B. J. (1993). Backpropagation neural networks: a tutorial, *Chemometrics and Intelligent Laboratory Systems* **18**(2): 115–155.

Xu, S. B., Huang, S. Y., Yuan, Z. G., Deng, X. H. and Jiang, K. (2020). Prediction of the dst index with bagging ensemble-learning algorithm, *The Astrophysical Journal Supplement Series* .

Yu, Y., Si, X., Hu, C. and Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures, *Neural computation* **31**(7): 1235–1270.

Zoznam príloh

Príloha A DVD médium – záverečná práca v elektronickej podobe, príručky v elektronickej podobe a zdrojové kódy.

Príloha B Používateľská príručka

Príloha C Systémová príručka

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie

Predikcia geomagnetických búrok pomocou hlbokého učenia

Diplomová práca

Používateľská príručka

Vedúci diplomovej práce:

doc. Ing. Peter Butka, PhD.

Autor:

Bc. Veronika Motúzová

Konzultanti diplomovej práce:

Ing. Viera Maslej Krešňáková, PhD. a RNDr. Šimon Mackovjak, PhD.

Košice 2023

Obsah

1	Funkcia programu	1
2	Súpis obsahu dodávky	1
3	Inštalácia programu	1
3.1	Požiadavky na technické prostriedky	1
3.2	Požiadavky na programové prostriedky	2
3.3	Vlastná inštalácia	2
3.4	Popis štruktúry programu	2
3.5	Popis správ pre systémového programátora	2
4	Popis vstupných, výstupných a pracovných súborov	3

1 Funkcia programu

Tento program slúži na predikciu geomagnetických búrok, pomocou metód hlbokého učenia a to konkrétnie rekurentnej neurónovej siete. Predkuje sa atribút DST index, na základe ktorého klasifikujeme stav na geomagnetickú búrku (1) alebo normálny stav(0).

2 Súpis obsahu dodávky

Priložené DVD obsahuje:

- zdrojové kódy - DP_Motuzova.zip,
- systémovú príručku v elektronickej forme - Veronika_Motuzova_SP.pdf,
- používateľskú príručku v elektronickej forme - Veronika_Motuzova_PP.pdf

3 Inštalácia programu

Tento program nie je potrebné inštalovať. Je potrebné rozbalíť si súbor DP_Motuzova.zip z priloženého DVD, repozitár je dostupný aj online¹. Súbory .ipynb otvoríme v programe JupyterLab, alebo v inom ktorý podporuje .ipynb súbory. Pred spustením je nutné nainštalovať v prostredí JupyterLab (alebo inom) potrebné balíky. Zoznam balíkov je v súbore requirements.txt, ktorý sa taktiež nachádza vo vyššie spomínanom .zip súbore.

3.1 Požiadavky na technické prostriedky

Trénovanie modelov bolo realizované online s povolením KKUI TUKE² na severoch Tesla K40c, Quadro RTX4000 a Quadro P4000. Na používanie či spustenie programu nie sú potrebné špeciálne technické požiadavky.

¹https://github.com/motuziik/DP_Veronika_Motuzova

²Prístup umožňuje CHI: <https://datalab.kkui.fei.tuke.sk/gpu>

3.2 Požiadavky na programové prostriedky

Na spustenie programu je potrebné mať nainštalované prostredie JupyterLab, alebo iné prostredie, ktoré podporuje súbory typu .ipynb. Zároveň je potrebné nainštalovať balíky, ktorých zoznam je v súbore requirements.txt.

3.3 Vlastná inštalácia

Pri spustení je potrebné rozbaliť súbor DP_Motuzova.zip do pracovného adresára. Ďalšia inštalácia nie je potrebná.

3.4 Popis štruktúry programu

Zdrojový kód s používateľským rozhraním sa nachádza v priečinku s názvom 6_nasadenie, dostupné online³, v súbore user.ipynb. Ide o súbor typu Jupyter Notebook, ktorý je charakteristický rozdelením kódu do blokov. Každý blok predstavuje čiastkovú funkcionality celej úlohy. Obsah blokov je bližšie popísaný v systémovej príručke v časti Popis algoritmov a údajových štruktúr, globálnych premenných.

3.5 Popis správ pre systémového programátora

Ak má blok zdrojového kódu nejaký očakávaný výstup, napr. výpis tvaru dát, typ dát, obrázok architektúry, sumarizácia modelu, vyhodnotenie úspešnosti alebo chyby, vyhodnotenie metrík či kontingenčnú tabuľku, tento výstup sa po spustení zobrazí v konzole pod blokom zdrojového kódu.

³Nasadenie:https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/6_nasadenie

4 Popis vstupných, výstupných a pracovných súborov

Súbor DP_Motuzova.zip obsahuje aj dátové súbory, ktoré predstavujú vstup do hlavného programu, a to omni_full.csv, test_omni.csv a train_omni.csv. Výstupom programu je naučená neuronová sieť, ktorá je schopná predikovať DST index a detektovať tak stav blížiacej sa búrky. Takáto neurónová sieť je dostupná v priečinku s názvom 6_nasadenie , dostupné online⁴, v súbore 6_6_pridane_atr.hdf5 a 6_dozadu_6_dopredu.hdf5.

⁴Nasadenie: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/6_nasadenie

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie

Predikcia geomagnetických búrok pomocou hlbokého učenia

Diplomová práca

Systémová príručka

Vedúci diplomovej práce:

doc. Ing. Peter Butka, PhD.

Autor:

Bc. Veronika Motúzová

Konzultanti diplomovej práce:

Ing. Viera Maslej Krešňáková, PhD. a RNDr. Šimon Mackovjak, PhD.

Košice 2023

Obsah

1 Funkcia programu	1
2 Analýza riešenia	1
3 Popis programu	1
3.1 Popis riešenia	2
3.2 Popis algoritmov a údajových štruktúr, globálnych premenných . . .	2
3.3 Zoznam zdrojových textov	4
4 Zhodnotenie riešenia	4

1 Funkcia programu

Program bol vytvorený na predikciu geomagnetických búrok, pomocou metód hlbokého učenia, a to konkrétnie rekurentnej neurónovej siete. Predikuje sa atribút DST index, na základe ktorého klasifikujeme stav na geomagnetickú búrku (1) alebo normálny stav(0).

Program bol vyvíjaný v programovacom jazyku Python, čo je v súčasnosti jeden z popredných jazykov využívaných v oblasti hlbokého učenia. Program načíta potrebné knižnice na prácu, prijme vstup od používateľa, upraví dátu na požadovaný tvar, načíta natrénovaný model, pomocou ktorého určí hodnotu DST indexu s predstihom 6 hodín, ktorá sa vypíše používateľovi do konzoly.

2 Analýza riešenia

Hlavnou úlohou pri vývoji riešenia bolo navrhnuť RNN sieť, ktorá bude dosahovať čo najvyššiu úspešnosť predikcie. Model RNN bol navrhnutý a optimalizovaný v priebehu troch kôl experimentov.

3 Popis programu

Program je zložený z viacerých častí. Hlavná časť pre používateľa je v súbore typu Jupyter Notebook, nachádza sa v priečinku `6_nasadenie`, dostupné online¹, v súbore `user.ipynb`. Je potrebné spustiť ho v programe podporujúcom tento typ súboru, napr. JupyterLab. Zdrojový kód je písaný v blokoch, ktoré na seba logicky naväzujú. Pre správne fungovanie je potrebné spustiť bloky v poradí, v akom sú napísané. Dôležité sú aj súbory `6_6_pridane_attr.hdf5` a `6_dozadu_6_dopredu.hdf5`. Tieto súbory obsahujú natrénovaný model, ktorý sa v súbore `user.ipynb` načíta. Tieto súbory sú uložené v spoločnom priečinku.

¹Nasadenie:https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/6_nasadenie

3.1 Popis riešenia

Celý technický postup návrhu modelov je v priečinku s názvom `4_modelovanie`, dostupné online².

3.2 Popis algoritmov a údajových štruktúr, globálnych premenných

Hlavný skript je tvorený z malých blokov, ktorých funkciunalitu si teraz popíšeme:

- Blok 1 : importuje všetky potrebné knižnice alebo jednotlivé konkretne funkcie z nich.

```
from tensorflow import keras
from keras.preprocessing.sequence import TimeseriesGenerator
from keras.models import Model
import numpy as np
```

- Blok 2 : načítanie vstupných dát od používateľa pre prvý model.

```
DST = [-10,13,16,-20,-25,-40,-30]
# používateľ zadá sekvenciu aspoň 7 historických hodnôt DST indexu
```

- Blok 3 : úprava vstupných dát.

```
generator = TimeseriesGenerator(DST, [0,0,0,0,0,0,0], length=6, batch_size=256)
```

- Blok 4 : načítanie natrénovaného predikčného modelu zo súboru.

```
model = keras.models.load_model('6_dozadu_6_dopredu.hdf5')
```

²Modelovanie: https://github.com/motuziik/DP_Veronika_Motuzova/tree/main/4_modelovanie

- Blok 5 : predikcia hodnoty DST indexu s predstihom 6 hodín.

```
DST_6 = model.predict(generator)
```

- Blok 6 : výpis výslednej predikovanej hodnoty.

```
print(DST_6)
```

- Blok 7 : načítanie vstupov od používateľa pre druhý model, spojenie vstupov do jedného poľa.

```
DST = np.array([-10, 13, 16, -20, -25, -40, -30])
# používateľ zadá sekvenciu aspoň 7 historických hodnôt všetkých 3 atribútov
F10_INDEX = np.array([177.69, 177.69, 178.39, 178.39, 178.39, 178.39, 178.39])
BZ_GSM = np.array([1.79, 2.29, 1.60, 1.29, 1.89, 3.20, 0.40])

inputs = np.concatenate((DST, F10_INDEX, BZ_GSM), axis=1)
```

- Blok 8 : kontrolný výpis vstupu.

```
print(inputs)
```

- Blok 9 : úprava vstupu pre potreby modelu.

```
generator = TimeseriesGenerator(inputs, [0,0,0,0,0,0], length=6, batch_size=256)
```

- Blok 10 : načítanie natrénovaného predikčného modelu zo súboru.

```
model = keras.models.load_model('6_6_pridane_attr.hdf5')
```

- Blok 11 : predikcia hodnoty DST indexu s predstihom 6 hodín.

```
DST_6 = model.predict(generator)
```

- Blok 12 : výpis výslednej predikovanej hodnoty.

```
print(DST_6)
```

3.3 Zoznam zdrojových textov

Priložené DVD obsahuje súbor `DP_Motuzova.zip`, v ktorom sú všetky zdrojové kódy zatriedené do pod-adresárov jednotlivých fáz. Súbor je dostupný aj online³. V každom pod-adresári sú zdrojové kódy a potrebné dátové súbory pre každú fazu, v `requirements.txt` sú spísané balíky, ktoré potrebujeme nainštalovať pred spustením.

4 Zhodnotenie riešenia

Riešenie danej úlohy bolo splnené. Úspešne sme natrénovali rekurentnú neurónovú sieť, ktorá je schopná predikovať DST index a detektovať tak stav blížiacej sa búrky. K práci vznikne aj článok s cieľom publikácie v blízkej budúcnosti. V budúcnosti by práca mohla pokračovať vývojom a optimalizáciou iných architektúr rekurentných neurónových sietí, prípadne by sa na predikciu mohla použiť iná kombinácia vstupných atribútov. Užitočná by určite bola aj optimalizácia hyperparametrov siete. Dátová množina, ktorej sa výskum venuje, sa s plynúcim časom neustále zväčšuje, čo bude viest k zlepšeniu predikcie v budúcich výskumoch. Práca vznikla v spolupráci s Ústavom experimentálnej fyziky Slovenskej akadémie vied, v.v.i. a poslúži na ich výskumné účely.

³https://github.com/motuziik/DP_Veronika_Motuzova