ŽILINSKÁ UNIVERZITA V ŽILINE

AUTOREFERÁT
DIZERTAČNEJ PRÁCE

Žilina apríl 2022

Ing. Miroslav Chochul

Žilinská univerzita v Žiline
Fakulta riadenia a informatiky

Ing. Miroslav Chochul

Autoreferát dizertačnej práce

KLASIFIKÁCIA ZVUKOV PROSTREDIA S VYUŽITÍM METÓD STROJOVÉHO UČENIA

na získanie akademického titulu **„philosophiae doctor“ (v skratke PhD.)**
v študijnom programe doktorandského štúdia
aplikovaná informatika

v študijnom odbore:
informatika

Žilina apríl 2022

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia **a Katedre technickej kybernetiky, Fakulte riadenia a informatiky Žilinskej univerzity v Žiline**

Predkladateľ: **Ing. Miroslav Chochul**
Katedra technickej kybernetiky
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Školiteľ: **doc. Ing. Peter Ševčík, PhD.**
Katedra technickej kybernetiky
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Oponent: **prof. Ing. Aleš Janota, PhD.**
Katedra riadiacich a informačných systémov
Fakulta elektrotechniky a informačných technológií
Žilinská univerzita v Žiline

Oponent: **Ing. Róbert Žalman, PhD.**
Vedúci oddelenia Smart Factory
Oddelenie Smart Factory
Asseco CEIT a.s.

Autoreferát bol rozoslaný dňa: .....

Obhajoba dizertačnej práce sa koná dňa ..... o ..... h. pred komisiou pre obhajobu dizertačnej práce schválenou pracovnou skupinou odborovej komisie v študijnom odbore **informatika v študijnom programe aplikovaná informatika**, vymenovanou dekanom Fakulty riadenia a informatiky Žilinskej univerzity v Žiline dňa .....

prof. Ing. Karol Matiaško, PhD.
predseda pracovnej skupiny odborovej komisie
v študijnom odbore **informatika**
v študijnom programe **aplikovaná informatika**

Fakulta riadenia a informatiky
Žilinská univerzita
Univerzitná 8215/1
010 26 Žilina

Abstrakt

Dizertačná práca sa zaoberá klasifikáciou environmentálnych zvukov, teda zvukov prostredia za pomoci metód strojového učenia. Klasifikačný model teda na základe akustického signálu predikuje druh zvuku. Teoretická časť prace je venovaná rozboru environmentálnych zvukov, ich pôvodu a spôsobu klasifikácie. Taktiež popisuje metódy strojového učenia, ich rozdelenie a využitie pre klasifikačné problémy. Ďalej sú tu popísané metódy extrakcie príznakov a druhy transformácie akustického signálu. Experimentálna časť práce je venovaná výberu a vývoju vhodnej architektúry klasifikačného modelu. Popisuje použité metódy predspracovania dát, ich augmentáciu a následnú extrakciu príznakov. Taktiež sa venuje popisu vývoju stratégie trénovania a vyhodnocovania klasifikačného modelu. Hlavným cieľom tejto práce bol návrh architektúry klasifikačného modelu, ktorý by mal nízku veľkosť, z čoho vyplýva nízky počet parametrov, aby bolo možné takýto model implementovať na zariadenia s obmedzenou výpočtovou silou. Pre porovnanie bol zvolený referenčný model, ktorým bol nami navrhnutý klasifikačný model porovnávaný. Z tohto porovnania vyplýva, že s využitím 0.65% veľkosti referenčného modelu, je možné dosiahnuť takmer rovnakú presnosť klasifikácie.

Kľúčové slová: klasifikácia environmentálnych zvukov, strojové učenie, konvolučná neurónová sieť, nízko-parametrická architektúra.

Abstract

The topic of this thesis is a classification of environmental sounds, i.e. non-human sounds, using machine-learning methods. The classification model, based on an acoustic signal, predicts a source of a sound. The theoretical part of the thesis is dedicated to the analysis of environmental sounds, their origin, and classification approaches. In addition, machine-learning methods, their taxonomy and their usage in classification tasks are described in this part as well. Next described are the feature extraction methods and types of acoustic signal transformations. The experimental part of the thesis is dedicated to the choice and development of the suitable architecture of the classification model. Next, are the description of data pre-processing methods, data augmentation and feature extraction. Furthermore, the development of training and evaluation strategies of the classification model are detailed. The main goal of this thesis was the development of a classification model architecture with a small size, which means low parameter count, to make it possible to implement this kind of model on devices with limited computational power. For evaluation, a reference model was chosen, against which our classification model was compared. From this comparison results that by using a 0.65% size of the reference model it is possible to achieve nearly similar classification accuracy.

Key words: environmental sound classification, machine learning, convolution neural network, low-parametric architecture.

Úvod

Predstavme si, že stojíme na ulici v meste. Zavrieme oči, čo počujeme? Pravdepodobne okoloidúce autá a autobusy, kroky ľudí, ktorí prechádzajú okolo, možno smiech alebo plač dieťaťa. Na základe nášho sluchu vieme získať množstvo informácií o našom prostredí. Pre väčšinu ľudí je schopnosť počúvať samozrejmá a prirodzená, avšak v prípade výpočtovej techniky sa jedná o náročnú úlohu a algoritmy strojového počúvania, ktoré automaticky rozpoznajú zvukové udalosti, dodnes zostávajú otvorený problém.

Klasifikácia environmentálnych zvukov, pomocou ktorej by bolo možné analyzovať a kategorizovať akustické emisie okolia, má viacero možných využití. Ako príklad sa ponúka monitorovanie hlukového znečistenia v mestách, nakoľko zvuk je dôležitým zdrojom informácií o mestskom živote. Ďalšie zaujímavé využitie je v oblasti bioakustiky, kde sú pomocou akustických emisií skúmané rôzne živočíchy či celé biodiverzity. Využitie je možne aj pre takzvaný akustický bezpečnostný systém, nakoľko mikrofóny sú všeobecne menšie a lacnejšie než kamery a sú odolné voči environmentálnym podmienkam ako sú hmla či zmena denného svetla a vďaka faktu, že zvuk prechádza cez prekážky, je možné implementovať takýto systém aj na monitorovanie väčšej oblasti ako sú lesy alebo polia. Zároveň je zaznamenávanie zvuku zvyčajne menej energeticky náročné.

Smerovanie dizertačnej práce je orientované do oblasti monitorovania chránenej oblasti za účelom signalizácie alebo v skratke akustický bezpečnostný systém. Našou motiváciou je systém pre ochranu lesov, pred nelegálnou ťažbou alebo nepovoleným vstupom motorových vozidiel do lesných oblastí, ktorý by mohol v budúcnosti vzniknúť. Nelegálna ťažba dreva je pretrvávajúci problém, v policajných štatistikách, bolo v prípade trestného činu Nelegálnej (pytliackej) ťažby dreva zistených 489 prípadov v roku 2020, v roku 2019 to bolo 618 prípadov krádeže dreva, či už v štátnych alebo v súkromných lesoch, ktoré riešila polícia Slovenskej republiky. Z hľadiska nelegálneho vstupu motorového vozidla, tieto prípadu sú posudzované ako trestný čin Porušovanie ochrany živočíchov a rastlín, takýchto prípadov bolo v roku 2020 zistených 96 a v roku 2019 bolo týchto prípadov 70. Preto usudzujeme, že takéto monitorovanie by malo zmysel a zvuk ako informačné médium je vhodnou voľbou. Šírenie zvuku je zväčša odolné voči prekážkam, má vysokú informačnú hodnotu a jeho zaznamenávanie je energeticky výhodnejšie ako v prípade obrazu. Z tohto dôvodu sme sa rozhodli analyzovať úlohu klasifikácie environmentálnych zvukov.

Väčšina prístupov riešenia úlohy klasifikácie environmentálnych zvukov, ktoré využívajú metódy strojového učenia, sú založené na hlbokom učení, z čoho vyplývajú veľmi vysoké výpočtové požiadavky. Tieto prístupy dosahujú dobré výsledky, pokiaľ sa jedná o presnosť rozpoznávania, avšak ich implementácia na zariadenia s nižšou výpočtovou silou je pre ich veľkosť zvyčajne problematická. V našej práci sme sa preto rozhodli venovať návrhu klasifikátora s nízkou veľkosťou, ktorý by bolo možné implementovať aj na zariadenia s nižšiu výpočtovou silou.

Na základe tohto sme si určili nasledujúce ciele dizertačnej práce:

- Analýza metód strojového učenia a metód spracovávania akustického signálu.
- Návrh klasifikačného modelu pre klasifikáciu environmentálnych zvukov s využitím poznatkov z vykonanej analýzy. Tento klasifikačný model by mal byť navrhnutý s dôrazom na nízku veľkosť, preto boli určené dva sekundárne ciele:
    - Architektúra klasifikačného modelu by mala pracovať s čo najnižším počtom parametrov.
    - Klasifikačný model by mal pracovať s čo najmenšou vzorkou akustického signálu, ideálne do jednej sekundy.
- Porovnanie navrhnutého klasifikačného modelu so zvoleným referenčným modelom.

1 Environmentálne zvuky

Vo všeobecnosti môžeme rozdeliť environmentálne zvuky na zvukové, resp. akustické udalosti, pri ktorých je zvuk produkovaný separátnymi fyzickými zdrojmi hluku, ako napríklad prechádzajúce auto, spev vtákov alebo kostolný zvon. Zvukové udalosti majú len jeden zdroj, avšak definícia, čo sa ráta za jeden zdroj zvuku je subjektívna, príkladom môže byť prechádzajúce auto, zvuky kolies na vozovke a hukot motoru, čiže abstraktnejší zdroj alebo jeho parciálne časti. Zvukové udalosti sú zvyčajne vhodne definované v krátkom časovom úseku. Oproti tomu zvukové, resp. akustické scény odkazujú na komplexný zvuk, ktorý je tvorený spojenými zvukmi viacerých zdrojov, zvyčajne z reálneho prostredia, ako napríklad zvuková scéna ulice môže obsahovať zvuky prechádzajúceho auta, zvuk krokov, komunikáciu ľudí a iné. Zvuková scéna dom môže byť zložená zo zvukov práčky, hudby z rádia a detského smiechu [1].

Úloha extrakcie informácií o akustickej udalosti a scény z audio signálu v prípade použitia techník strojového učenia spadá do kategórie strojového vnímania, konkrétne strojové počúvanie (z angl. machine hearing), ktoré je podľa Bello et al. [2] akustickým ekvivalentom k strojovému videniu (z angl. machine vision), teda tiež kombinuje techniky spracovávania signálov so strojovým učením a vytvára tak systém, ktorý je schopný extrahovať užitočné informácie zo zvukov. Zjednodušene môžeme strojové počúvanie popísať ako schopnosť identifikovať a rozlišovať zvuky prítomné v audio signáli, s cieľom dosiahnuť rozpoznávanie zvuku na ľudskej úrovni [3]. Medzi typické úlohy strojového počúvania patrí klasifikácia, ktorej cieľom je kategorizovať akustické nahrávky do preddefinovaných kategórií. Klasifikovať môže jednotlivé akustické udalosti alebo celé akustické scény. Ďalšia úloha je detekcia, pri ktorej je cieľ určiť čas, kedy je špecifikovaný zvuk alebo zvuky aktívny. Zo špecifickejších úloh je jedna napríklad o odhadovaní, či dve audio nahrávky pochádzajú z jednej akustickej scény.

Z ohľadom na pôvod zvukovej informácie, rozpoznávame niekoľko oblastí výskumu. Tou najrozšírenejšou oblasťou je výskum rozpoznávania ľudskej reči. Druhá oblasť výskumu pracuje s hudbou, ktorá sa nazýva „získavanie hudobných informácií“ (z angl. Music information retrieval). Ďalšia oblasť výskumu je analýza každodenných zvukov, ktoré nepatria do predchádzajúcich oblastí, čiže zvuky okrem ľudskej reči a hudby. Nájdeme však paralely medzi jednotlivými úlohami týchto oblastí, napríklad klasifikácia akustickej scény, kedy chceme priradiť jedno označenie ako „reštaurácia“ alebo „park“ je príbuzná s úlohou rozpoznávania rečníka a rozpoznávaním hudobného žánru. Obdobne, úloha označovania zvukov, ktorej cieľom je priradiť množinu označení k nahrávke, napríklad pomenovanie počuteľných objektov, je príbuzná rozpoznávaniu hudobných nástrojov v nahrávke. Úloha detekcie akustických udalostí, ktorej cieľ je identifikovanie zvukových udalostí v dobe ich vzniku, v rámci audio signálu, je úloha príbuzná automatickému rozpoznávaniu reči alebo automatického prepisu hudby [1]. Aj vďaka týmto podobnostiam sú techniky, ktoré boli vyvinuté pre jednu úlohu, prenášané do iných oblastí. Je však dôležite si uvedomovať rozdiely v signáloch jednotlivých oblastí ako napríklad, že hudobný signál je zložený zo zvukov hudobných nástrojov, ktoré boli navrhnuté, aby mali harmonickú štruktúru, každodenné zvuky túto vlastnosť nezdieľajú.

1.1 Organizácia environmentálnych zvukov

Z hľadiska bolo organizácie environmentálnych zvukov bolo navrhnutých viacero taxonómií, podľa charakteristík. Schafer navrhol klasifikačnú schému na základe fyzikálnych charakteristík [4], Delage navrhol klasifikáciu na základe stupňa ľudskej aktivity [5] alebo v kontexte bioakustiky Pijanowski et al. navrhli klasifikačnú schému, podľa pôvodu zvuku. Avšak tieto klasifikačné schémy neberú v úvahu štruktúru akustických udalostí s ohľadom na

rôznu mieru abstraktnosti [7]. Rôzne princípy kategorizácie môžu koexistovať, zvlášť s ohľadom na zdroj zvuku a činnosť produkujúcu zvuk. Salamon et al. navrhli taxonómiu mestských zvukov, ktorá do určitej miery zahŕňa činnosti produkujúce zvuk [8]. Na najvyššej úrovni sú štyri kategórie - človek, príroda, mechanické a hudba. Na nižších úrovniach sú potom kategórie zdrojov zvukov, ktoré sú dostatočne rozdelené, aby boli jednoznačné, čiže napríklad brzdy auta, motor auta alebo klaksón namiesto jednoducho auto.

2 Strojové učenie

Metódy strojového učenia (ML, z angl. Machine Learning) implementujú algoritmy a štatistické modely, pomocou ktorých efektívne vykonávajú úlohy, bez nutnosti explicitne naprogramovať inštrukcie, pomocou ktorých sa majú tieto úlohy vykonávať. Namiesto toho sa pomocou algoritmu učenia naučí, ako danú úlohu vykonávať na základe poskytnutých dát. Sila strojového učenia je v jeho schopnosti poskytovať generalizované riešenie prostredníctvom architektúry, ktorá reprezentuje komplexné vzťahy v dátach. Využitím týchto metód môžeme docieliť, že výpočtové procesy budú efektívnejšie, spoľahlivejšie a cenovo výhodnejšie. ML sa konvenčne rozdeľuje do kategórií na základe procesu učenia na: Učenie s učiteľom, Učenie bez učiteľa, Kombinované učenie s učiteľom a bez učiteľa a Učenie formou odmeňovania.

2.1 Metódy klasifikácie

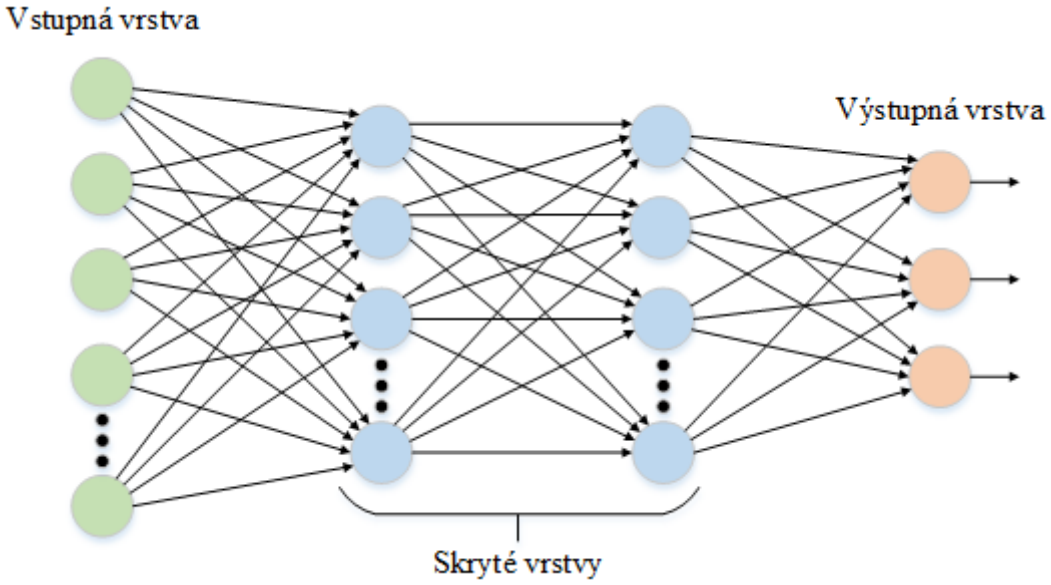
V obore strojového učenia je klasifikácia druh problému, pri ktorom je cieľom rozdeliť dáta do skupín na základe logického rozhodovania, čo najbližšie skutočnému rozdeleniu. Klasifikácia je teda úloha rozpoznávania vzorov. Klasifikátor je potom algoritmus, ktorý implementuje klasifikáciu. Hlavná úloha klasifikátora je identifikovať triedu, do ktorej patria nové pozorované vzorky. Tento algoritmus je vytváraný na základe trénovacej množiny dát, ktorá obsahuje vstupné vzorky aj ich výstupné triedy. Následne po tejto fáze trénovania by mal byť klasifikátor schopný kategorizovať do naučených tried aj vopred neznáme dáta. Typickým príkladom môže byť kategorizovanie obrázkov psov na základe ich rasy, lokalizácia objektov vo fotografii, určovanie, či je daný text pozitívny alebo negatívny v danej téme/kontexte alebo rozhodnutie z akustickej nahrávky, aký zvuk je prítomný.

2.2 Neurónové siete

Neurónové siete, alebo tiež známe aj ako umelé neurónové siete alebo simulované neurónové siete, je technika strojového učenia, ktorej štruktúra bola inšpirovaná ľudským mozgom. Napodobňuje spôsob, ktorým biologické neuróny signalizujú medzi sebou. Vďaka tomuto pozadiu sú neurónové siete vysvetľované na vyššej úrovni za pomoci neurobiologických termínov, ako neurón, axón a synapsie, ktoré ich spájajú [27], avšak ako už bolo spomenuté, napriek tomu, že neurónové siete boli inšpirované biologickým fungovaním mozgu, nie sú modelom mozgu.

2.2.1 Architektúra neurónovej siete

Architektúru neurónovej siete môžeme vo všeobecnosti popísať pomocou orientovaného grafu, kde vrcholy predstavujú neuróny a orientované hrany predstavujú synaptické prepojenia. Jednou z tradičných a pomerne dostatočne preskúmaných štruktúr je viacvrstvová štruktúra zobrazená na obrázku 1.



Obrázok 1 Viacvrstvová architektúra doprednej neurónovej siete

Ako je možné vidieť, vrstvy tejto štruktúry sú pomenované.
Rozpoznávame tri základné typy vrstiev[38]:

- Vstupná vrstva – neuróny tejto vrstvy prijímajú vstup z externého sveta a ich výstup je spracovávaný ďalšími neurónmi neurónovej siete.
- Skryté vrstvy – jedna alebo viacej vrstiev, ktoré sa nachádzajú medzi vstupnou a výstupnou vrstvou, ktorých neuróny prijímajú vstup z ostatných neurónov, alebo aj na základe prahového prepojenia z externého sveta. Ich výstup spracovávaný ďalšími neurónmi neurónovej siete.
- Výstupná vrstva – neuróny tejto vrstvy majú obdobnú funkciu ako u skrytých vrstiev, avšak ich výstup už nie je ďalej spracovávaný neurónovou sieťou a teda predstavuje odozvu neurónovej siete na daný vstup z externého sveta.

V tej najjednoduchšej forme má táto architektúra len vstupnú vrstvu, ktorá je priamo prepojená na výstupnú, nie však naopak. Takáto neurónová sieť je potom označovaná ako jednovrstvová sieť. Pridaním jednej alebo viacerých skrytých vrstiev neurónová sieť bude schopná zo vstupu extrahovať štatistiky vyššieho rádu [37].

2.3 Konvolučné neurónové siete

Konvolučné neurónové siete, alebo tiež známe aj ako konvolučné siete [33], sú špecializovaný typ neurónovej siete pre spracovávanie dát, ktoré majú známu mriežkovitú topológiu, napríklad časovú postupnosť dát môžeme chápať ako jednorozmernú mriežku, kedy berieme vzorky v pravidelných časových intervaloch alebo obrazové dáta, ktoré si môžeme predstaviť ako dvojrozmernú mriežku pixelov.

Ako názov konvolučná sieť napovedá, je v tejto neurónovej sieti implementovaná matematická operácia konvolúcie. V jednoduchosti je možne napísať, že konvolučné siete sú jednoducho neurónové siete, ktoré používajú konvolúciu namiesto skalárneho súčinu, aspoň v jednej zo svojich vrstiev [42]. V prípade tradičnej neurónovej siete je každý neurón v prvej skrytej vrstve prepojený na každú hodnotu vstupu. Tento prístup ale spôsobuje prudký nárast parametrov v prípade vysokodimenzionálneho vstupu, ak máme napríklad vstup obrázok s rozmermi 100x100x3 (100 pixelov široký, 100 pixelov vysoký a 3 farebné kanály), každý neurón prvej skrytej vrstvy, by mal 30000 trénovateľných parametrov. Naproti tomu konvolučné siete zavádzajú princíp, kedy je každý neurón prepojený len s malou časťou

vstupu (časť susedných položiek). Tento jav označujeme ako lokálna konektivita. Táto sa v konvolučných sieťach používa nielen vo vrstve prepojenej na vstupné dáta, ale aj v skrytých vrstvách a je propagovaná do celej siete. Typ neurónových sietí, ktoré tento princíp využívajú sú potom označované ako lokálne prepojené vrstvy [43]. Táto charakteristika poskytuje konvolučnej sieti dve zaujímavé vlastnosti [44]:

- Vzory, ktoré sa naučí sú invariantné – teda potom čo sa naučí určitý vzor napríklad v pravom dolnom rohu, môže ho konvolučné neurónová sieť rozpoznať kdekoľvek, napríklad v ľavom hornom rohu.
- Môžu sa naučiť priestorovú hierarchiu vzorov – Prvá konvolučná vrstva sa naučí malé lokálne vzory ako napríklad hrany. Druhá vrstva sa bude učiť zložitejšie vzory z prvej vrstvy a tak ďalej. Toto dovoľuje konvolučnej sieti sa efektívne naučiť stále komplexnejšie a abstraktnejšie vizuálne pojmy.

Druhým typickým rysom všeobecnej neurónovej siete je fakt, že každý neurón môže obsahovať unikátne synaptické váhy. Konvolučná sieť používa techniku zdieľania parametrov, čo znamená, že neuróny jednej vrstvy zdieľajú rovnaké hodnoty parametrov. Toto prináša výhodu v pamäťovej náročnosti, keďže počet parametrov, ktoré je nutné uchovať, je výrazne znížený. Zdieľanie parametrov vychádza z predpokladu, že každá vzorka vstupu obsahuje príznaky, ktoré sa v rámci nej opakujú. Príznak, ktorý rozpoznávame, je reprezentovaný práve množinou váh, ktorá je zdieľaná. Zároveň je zachovaná pozícia, kde bol daný príznak rozpoznaný, keďže konvolučné vrstvy majú na výstupe tzn. mapu príznakov. Intuitívne, jedna mapa príznakov rozpoznáva jeden príznak a mapuje ho na pozície vstupu. Je možné rozpoznávať viacero príznakov alebo vzorov v jednej vrstve. Vrstva potom obsahuje niekoľko množín váh a na výstupe je niekoľko máp príznakov, každá pre jeden vzor.

2.3.1 Konvolučná vrstva

Kľúčovým stavebným blokom konvolučných neurónových sietí sú konvolučné vrstvy. Synaptické váhy neurónov sa v kontexte konvolučných vrstiev nazývajú jadro alebo filter. Pre upresnenie jedna skupina váh sa označuje ako filter a jedna vrstva môže mať viacero takýchto filtrov. Počas prechodu vrstvou je realizovaných niekoľko konvolúcií, podľa počtu filtrov v danej vrstve. Hodnoty tohto filtra predstavujú, spolu s biasom, trénovateľné parametre konvolučnej vrstvy. Z ohľadom na rozmernosť vstupu rozdeľujeme konvolučné vrstvy na:

- 1D konvolučná vrstva – najjednoduchší typ, zvyčajne používaný pre sekvenčné množiny dát,
- 2D konvolučná vrstva – najčastejšie používaný typ v konvolučných sieťach, zvyčajne využívaný pre obrazové dáta,
- 3D konvolučná vrstva – tento typ vrstiev sa využíva pri detekcii udalosti vo videu alebo pri medicínskych 3D obrazoch.

Filter v týchto vrstvách má rovnakú dimenziu ako vstup, ale niekoľkonásobne menšiu veľkosť. Tento filter je posúvaný naprieč vstupným kanálom, zľava doprava a zhora dole. V každej pozícii sú hodnoty filtra vynásobené s aktuálne prekrytými hodnotami vstupu a následne sčítaná do jednej hodnoty. Kolektívny výsledok potom predstavuje mapu príznakov.

2.3.2 Podvzorkovanie

Okrem konvolučnej vrstvy ďalší prvok konvolučných neurónových sietí, ktorý ich robí efektívne, je postupné podvzorkovanie dát počas prechodu konvolučnými vrstvami. To núti model, aby sa naučil väčšie (s ohľadom na pôvodný vstupný priestor) a viac komplexnejšie príznaky (vzory vzorov) v neskorších vrstvách. Toto podvzorkovanie je možné vykonať tak,

že použijeme konvolučnú vrstvu a nastavíme veľkosť filtra a veľkosť kroku konvolúcie na rovnakú hodnotu [47].

Druhým spôsobom, ktorý sa využíva, je aplikácie tzv. vrstvy združovania (z angl. pooling layer). Tento druh vrstvy bol použitý už v LeNet5. Vrstva združovania má podobné nastavenia ako konvolučná vrstva, teda nastavenia veľkosti filtra a kroku, ktorý je vo väčšine prípadov volený tak, aby sa oblasti aplikácie neprekrývali. Každý kanál vstupnej mapy príznakov je spracovávaný samostatne. Z každej oblasti aplikácie je potom vybratá jedna hodnota, v závislosti od typu vrstvy, buď je to priemerná hodnota alebo častejšie využívaná maximálna hodnota.

3 Extrakcia príznakov

V prvej kapitole bol zvuk popísaný s ohľadom na jeho informačnú hodnotu. Z fyzikálneho hľadiska môžeme zvuk, resp. akustický signál definovať, ako usporiadaný kmitavý pohyb častíc prostredia, v ktorom sa zvuk šíri. Kmitanie častíc zdroja zvuku sa pomocou vzájomného pôsobenia prenáša na častice v okolí, ktoré sa tiež rozkmitajú, nedochádza však k presunu hmoty, len k presunu energie. Keďže dochádza pri prenose k určitému oneskoreniu, vzniká postupná vlna, ktorá sa šíri smerom od zdroja zvuku. Celý proces je v podstate mechanickým kmitaním pružného prostredia. Na základe frekvencie kmitania sa potom zvuky delia do troch pásiem: infrazvuk - pásmo 0,7 – 16 Hz, sú to zvuky pod hranicou počuteľnosti; počuteľné pásmo - pásmo 16 – 20 000 Hz, toto pásmo predstavuje zvuky, ktoré sú schopné vyvolať zvukový vnem; ultrazvuk - pásmo 20 – 50 kHz, to sú zvuky nad hranicou počuteľnosti. Skutočný rozsah počuteľného zvuku je subjektívny, avšak najhlasnejšie sú vnímané signály v oblasti 500 – 5 000 Hz [48], čiže v tejto oblasti je ľudské ucho najcitlivejšie, zároveň s narastajúcim vekom sa častokrát stráca citlivosť vo vyšších frekvenciách.

Keďže zvuk, ktorý môžeme premeniť z akustickej podoby na elektronický signál pomocou mikrofónu, ako mnoho iných signálov v prírode je spojitý v čase i úrovni a technické prostriedky pracujú s diskrétnymi hodnotami v diskrétnom čase, je nutné využiť procesu diskretizácie. Tento proces sa skladá z dvoch krokov [50][51]:

- Diskretizácia v čase – vzorkovanie – proces vzorkovania transformuje signál na časovo diskrétny signál. Prostredníctvom vzorkovania sa získavajú hodnoty časovo spojitého signálu v presne definovaných časových okamihoch. V prípade periodického vzorkovania je spojitý signál $x(t)$ nahradený postupnosťou vzoriek $x(nT)$, kde T predstavuje periódu vzorkovania.
- Diskretizácia v úrovni – kvantovanie – proces kvantovania transformuje signál spojitý v úrovni na signál diskrétny v úrovni. Každá hodnota signálu je nahradená hodnotou vybranou z konečnej množiny prípustných hodnôt. Počas tohto procesu vzniká tzv. kvantizačná chyba, čo je rozdiel medzi skutočnou a priradenou hodnotou.

Po procese vzorkovania a kvantovania dostávame signál diskrétny v čase i úrovni. Potom nasleduje proces kódovania, kde je každej diskrétnej hodnote číslicového signálu priradený istý kód, najčastejšie b-bitová binárna postupnosť.

3.1 Fourierová transformácia

Fourierová transformácia je jedným zo základných pilierov spracovávania signálov. S využitím tejto transformácie môžeme rozložiť signál s ohľadom na jeho harmonické (sínusové, resp. komplexne exponenciálne) komponenty, efektívne tak prevádza signál z časovej oblasti do frekvenčnej oblasti. Predstavuje tak efektívny nástroj na frekvenčnú

analýzu v oblasti spracovania signálov. Pomocou tejto transformácie vieme spracovávať signál, ktorý je aperiodický a môže byť spojitý alebo diskrétny v čase [51].
Pre výpočet Fourierovej transformácie bolo navrhnutých viacero algoritmov:

- Diskrétna Fourierová transformácia (DFT)
- Rýchla Fourierová transformácia (FFT)
- Krátkodobá Fourierová transformácia (STFT)
- Kĺzavá diskrétna Fourierová transformácia (SDFT)

3.1.2 Váhové funkcie

Algoritmus FFT predpokladá, že vstupný signál je periodický, teda že časová postupnosť s dĺžkou N sa cyklicky opakuje donekonečna. Ak frekvencia sínusového vstupného signálu nie je násobkom frekvenčného rozlíšenia f_r, tento predpoklad nie je pravdivý a FFT zaznamená diskontinuitu medzi poslednou a prvou vzorkou z dôvodu cyklického opakovania. Tieto umelé diskontinuity sa potom prejavia vo FFT ako vysokofrekvenčné komponenty, ktoré neboli prítomné v pôvodnom signáli. Výsledné spektrum teda nebude spektrum pôvodného signálu, ale jeho rozmazaná verzia, teda akoby energia jednej frekvencie presakovala to ostatných frekvencií. Tento fenomén je nazývaný presakovanie spektra (z angl. spectral leakage) [57].
Tento efekt je možné minimalizovať použitím váhovej funkcie, resp. váhového okna. Teda časová postupnosť je vynásobená váhovou funkciou pred aplikáciou FFT. Všetky váhové funkcie sa zhodujú v troch vlastnostiach:

- mimo oblasti ich definície nadobúdajú nulovú hodnotu,
- sú symetrické a na hranici symetrie, teda v strede nadobúdajú maximum,
- na začiatku aj na konci, nadobúdajú hodnoty blízke alebo rovné nule.

Na základe týchto vlastností je diskontinuita odstránená. Bolo definovaných niekoľko váhových funkcií, ktorých tvar väčšinou odráža kompromis medzi šírkou výsledného vrcholu vo frekvenčnej oblasti, presnosťou amplitúdy a pomerom zníženia presakovania spektra [52]. Inými slovami, široký hlavný lalok, zapríčiňujúci nepriaznivé frekvenčné rozlíšenie, je spojený s malou amplitúdou postranných lalokov, pri ktorých sa presakovanie znižuje, a naopak úzky hlavný lalok, umožňujúci presnejšie odčítanie frekvencie signálu, je spojený s väčšou amplitúdou postranných lalokov, keď je presakovanie spektra väčšie [58].

3.1.3 Krátkodobá Fourierová transformácia

S využitím krátkodobej Fourierovej transformácie môžeme sledovať spektrálne zmeny v čase. Princípom STFT, ako už bolo popísané, je segmentácia signálu do kratších úsekov pomocou váhovej funkcie a následne je na každý úsek aplikovaná FFT. Použité váhové okno a jeho veľkosť ovplyvňujú výsledné zobrazenie. STFT môžeme definovať ako [61]:

$$X(m,k) = \sum_{n=0}^{N-1} x(hm+n)w(n)e^{-2\pi i\frac{mn}{N}} \quad (1)$$

kde $m = 0, 1, 2, \dots N-1$, N označuje dĺžku váhového okna, k definuje poradové číslo aktuálneho segmentu a h je jeho posun. Dĺžka okna predstavuje kompromis medzi časovým a frekvenčným rozlíšením, nakoľko dĺžka okna ovplyvňuje frekvenčné rozlíšenie priamo úmerne a časové rozlíšenia nepriamo úmerne $\Delta t = N$. Výsledkom STFT sú komplexné čísla vyjadrujúce informáciu o fáze a amplitúde každého frekvenčného kroku. Amplitúdové spektrum potom môžeme vyjadriť ako absolútnu hodnotu výsledku (2), čím odstránime informáciu o fáze. Výkonové spektrum potom ako jeho druhú mocninu (3).

$$A(m) = |X(m)| \quad (2)$$

$$P(m) = |X(m)|^2 \quad (3)$$

4 Experimentálna časť

4.1 Datasety

Pre vytvorenie kvalifikátora zvukov prostredia je potrebná množina vstupných dát – dataset, ktorý pozostáva z akustických nahrávok rôznych zdrojov zvuku, resp. činností produkujúcich zvuk. Keďže naším cieľom je klasifikácia samostatných akustických udalostí, nie akustických scén, mali by tieto nahrávky obsahovať každá len jeden anotovaný zdroj zvuku. V rámci výskumu klasifikácie environmentálnych zvukov bolo zostavených niekoľko datasetov, ktoré túto podmienku spĺňajú, preto nebolo potrebné zostavovať pre potreby nášho výskumu vlastný. Z hľadiska množstva dostupných dát sa ponuka AudioSet [64], ktorý bol zostavený z 10 sekundových zostrihov zvuku z YouTube videí, pozostáva z 527 kategórií zvuku. Avšak v rámci týchto kategórií sú aj zvuky hudobných nástrojov a ľudskej reči, preto je pre naše potreby nevhodný. Ďalšia nevýhoda je, že nie sú k dispozícii čisté nahrávky, ale len vopred extrahované príznaky. Z ohľadom na veľkosť množiny nahrávok a kategórií sa ako ďalší ponuka FSD50K [65], ktorý inšpiráciu čerpal v AudioSet-e a prevzal z neho 200 kategórií zvuku, pomocou ktorých boli označené nahrávky prevzaté z projektu FreeSound [66]. Avšak aj tento dataset obsahuje nahrávky hudobných nástrojov a ľudskej reči. Tento dataset je už zostavený zo samotných nahrávok, avšak tieto môžu obsahovať viacero anotácií. Dataset, ktorý neobsahuje zvuky hudobných nástrojov, ani ľudskú reč, je UrbanSound8K [8], ktorý bol zostavený na základe sťažností na hluk v New Yorku medzi rokmi 2010 a 2014, z ktorých bolo určených 10 rozličných tried zvuku. Nahrávky boli potom taktiež ako FSD50K získané z projektu FreeSound. Dataset pozostáva z 8 732 audio nahrávok, každá s maximálnou dĺžkou štyri sekundy. Tento dataset je úzko spätý s výskumom zvukového znečistenia v mestskej časti. Pre potreby tejto práce, sme sa nakoniec rozhodli využiť dataset ESC-50 [26], ktorý bol zostavený K. Piczakom. Tento dataset je z ohľadom na veľkosť menší ako predchádzajúce, pozostáva z 2 000 nahrávok rozdelených do 50 tried, ktoré môžeme zhrnúť do piatich kategórií pôvodu. Každá nahrávka je dlhá päť sekúnd a má pôvod v projekte FreeSound. Rozmanitosť pôvodov zdrojov zvuku, ktoré nie sú viazané na konkrétnu oblasť, je dôvod prečo je tento dataset vhodný pre návrh všeobecného klasifikátora. V tabuľke 1 je možne vidieť stručné porovnanie vyššie spomenutých datasetov.

dataset	Počet nahrávok	Dĺžka nahrávky	Celková dĺžka	Počet tried	Zdroj nahrávok
AudioSet	2,1 M	10 s	5 731 h	527	Youtube
FSD50K	51 197	0,3 – 30 s	108 k	200	FreeSound
UrbanSound8K	8 732	≤ 4 s	8,8 h	10	FreeSound
ESC-50	2 000	5 s	2,8 h	50	FreeSound

Tabuľka 1 Všeobecné porovnanie datasetov

4.1.1 ESC-50

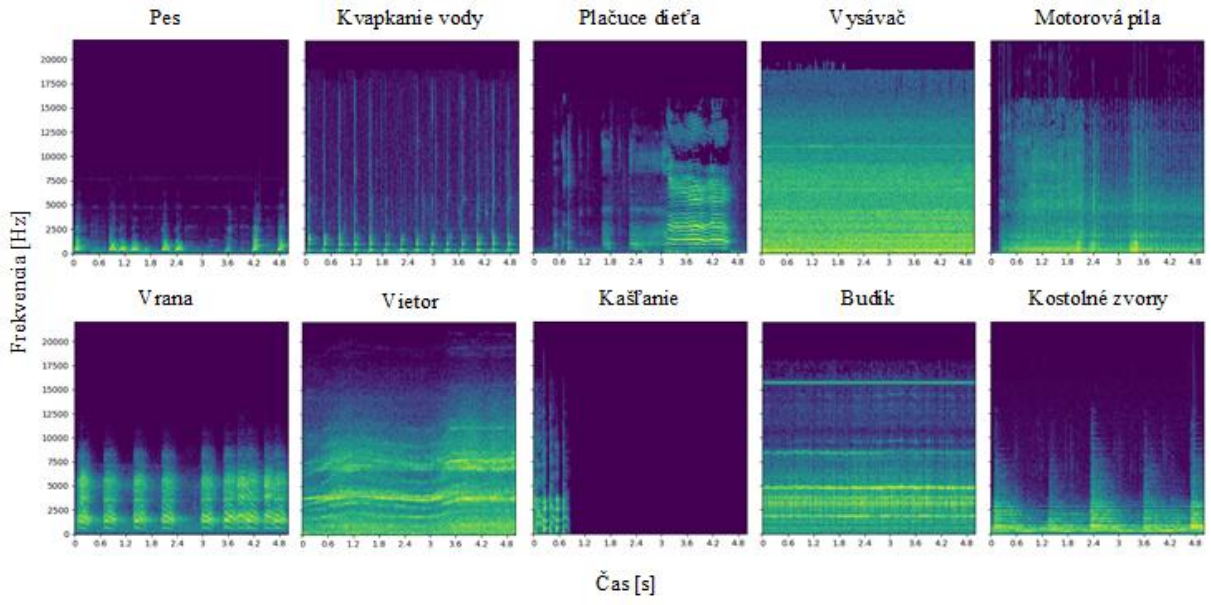
Ako už bolo spomenuté, dataset ESC-50 pozostáva z 2 000 anotovaných environmentálnych nahrávok samostatných akustických udalostí. Tieto nahrávky sú rovnomerne rozložené do päťdesiatich kategórií, 40 nahrávok na každú triedu. Tieto triedy sú rozložené do piatich voľne definovaných oblastí pôvodu, detailné zobrazenie tried je v tabuľke 2.

Ako je možné vidieť tento dataset obsahuje zvuky, ktoré sú veľmi bežné (smiech, štekot psa), niektoré celkom zreteľné (rozbíjanie skla, čistenie zubov) a niektoré, kde sú rozdiely jemnejšie (zvuk lietadla a helikoptéry). Uvedené zdroje zvuku sú veľmi heterogénne, čo sa týka dĺžky trvania či intenzity. Zároveň, ako môžeme vidieť na obrázku 2, zvuky majú vo frekvenčnej oblasti priebeh podobný šumu, ako napríklad vysávač a niektoré, ako napríklad budík, majú výrazné nosné signály, ktorých frekvencie môžeme odčítať. Toto môže

spôsobovať ťažkosti pre model strojového učenia, ktorý sa snaží naučiť zvuky, ktoré môžu byť odlišné nahrávku od nahrávky.

Zvieratá	**Zvuky prírody a zvuky vody**	**Ľudské nerečové zvuky**	**Interiérové/ domáce zvuky**	**Exteriérové/ mestské zvuky**
Pes	Dážď	Plačúce dieťa	Klopanie na dvere	Helikoptéra
Kohút	Morské vlny	Kýchnutie	Klikanie myšou	Motorová píla
Prasa	Praskanie ohňa	Tlieskanie	Písanie na klávesnici	Siréna
Krava	Cvrčky	Dýchanie	Dvere, vŕzganie dreva	Klaksón auta
Žaba	Čvirikajúce vtáky	Kašľanie	Otváranie konzervy	Motor
Mačka	Kvapkanie vody	Kroky	Práčka	Vlak
Sliepka	Vietor	Smiech	Vysávač	Kostolné zvony
Hmyz (lietajúci)	Nalievanie vody	Čistenie zubov	Budík	Lietadlo
Ovca	Spláchnutie toalety	Chrápanie	Tikanie hodín	Ohňostroj
Vrana	Búrka	Pitie, popíjanie	Rozbíjanie skla	Ručná píla

Tabuľka 2 Rozdelenie tried datasetu ESC-50



Obrázok 2 Príklady spektrogramov nahrávok v datasete ESC-50

Zároveň môžeme vidieť, že v niektorých prípadoch, aj keď dĺžka nahrávky je 5 sekúnd, užitočný zvuk je prítomný len určité percento tohto času. Keďže väčšina prístupov klasifikácie delí nahrávky na menšie rámce, ktoré potom klasifikuje podľa vopred volenej schémy, vyvstáva otázka slabého anotovania. Teda v prípade, že človek vychádza z predpokladu, že užitočný zvuk je prítomný po celé trvanie nahrávky a tá je rozdelená na rámce, ktoré zdedia anotáciu celej nahrávky, môžu niektoré rámce zdediť anotáciu, aj keď neobsahujú užitočný zvuk, čo môže mať za následok pokles presnosti rozpoznávania.

4.2 Metódy strojového učenia

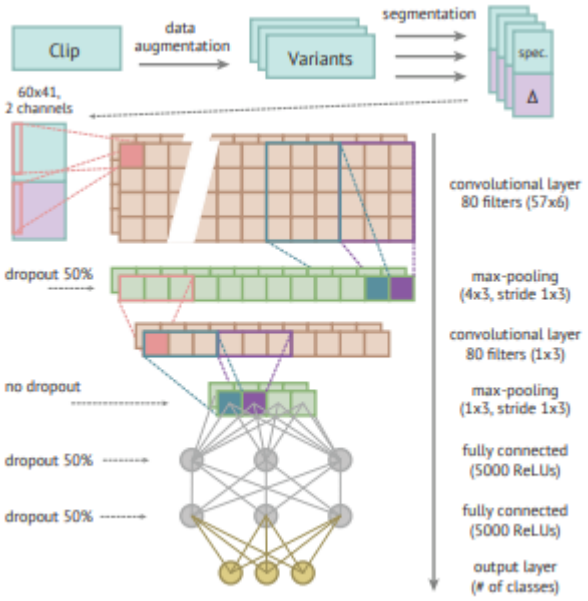
Ako už bolo spomenuté v predchádzajúcej časti, vybrali sme pre základ našej práce dataset ESC-50. Autor Karol Piczak, ktorý zostavil tento dataset environmentálnych zvukov, na ňom spravil niekoľko štúdií, ktoré boli zamerané na efektivitu klasifikačných metód strojového učenia [26][67]. Tieto metódy zahŕňali k-Najbližších susedov, ktorá dosahovala rozpoznávanie 32,2%, model najlepšie rozpoznával „kýchnutie“, „búrku“ a „otváranie plechovky“ a naopak problémové triedy zvuku boli „tikanie hodín“, „klaksón auta“ a „motor“, kde bolo rozpoznávanie takmer nulové. Ďalšia bola metóda podporných vektorov dosahovala rozpoznávanie 39,6%, model založený na tejto metóde najlepšie rozpoznával „búrku“ a „klopanie“ a problematické triedy boli „dvere, vŕzganie dreva“ „helikoptéra“ a „motor“. Poslednou z klasických metód klasifikácie, bola metóda náhodný les, ktorá dosiahla rozpoznávanie 44,3%, model rozpoznával najlepšie triedy „klopanie“ a „búrka“, naopak problém s rozpoznávaním mal pri triedach „tikanie hodín“, „klaksón auta“ a „kvapkanie vody“. V článku [67] potom popísal model, založený na umelých neurónových sieťach, konkrétne na konvolučných neurónových sieťach. Najlepší model potom dosiahol rozpoznávanie 64,5%. To nám ukazuje, že metóda strojového učenia založená na umelých neurónových sieťach dosahuje lepších výsledkov ako klasické klasifikátory, čo je ďalej podporené, že na GitHub-ovej[1] stránke tohto datasetu je uvedený rebríček modelov a presnosť ich rozpoznávania a väčšina týchto modelov je založená práve na umelých neurónových sieťach. Z toho dôvodu sme sa aj my rozhodli pre architektúru modelu založenú na neurónových sieťach, konkrétne na konvolučných neurónových sieťach. Model, ktorý navrhol K. Piczak, bol zvolený ako referenčný. Ako už bolo spomínané väčšina doterajších prístupov používa veľmi hlboké modely, ktorých presnosť rozpoznávania je vyššia ako v prípade Piczakovho modelu, rovnako je však vyššia aj ich veľkosť, v niektorých prípadoch dosahuje až 87 miliónov. Nakoľko naším cieľom práce je návrh architektúry s malou veľkosťou modelu, resp. s nízkym počtom parametrov, a tento typ prístupu nemá takú popularitu, rozhodli sme sa preto použiť Piczakov model ako referenčný, nakoľko sa jedná o prvotné riešenie použitého datasetu ESC-50.

Tento model využíval schému spracovávania environmentálnych zvukov na báze podrámcov. Teda schéma, kedy je každá nahrávka rozdelená do menších podrámcov, zvyčajne s nejakou úrovňou prekrytia a príznaky sú extrahované z každého rámca samostatne. Aby sa klasifikátor mohol trénovať, príznaky z jednotlivých podrámcov sú buď spojené do jedného veľkého vektora príznakov, alebo spriemerované tak, aby predstavovali jeden rámec. Druhá možnosť je nechať klasifikátor trénovať na každom podrámci a po spracovaní všetkých vykonať kolektívne zvolenie výslednej triedy pre celý rámec, na základe tried zvolených zo všetkých podrámcov, napríklad volenie podľa majority alebo volenie na základe pravdepodobnosti [68]. Mel spektrogramy, ktoré tento model požíva na extrakciu príznakov, boli rozdelené na 41 podrámcov s prekrytím 50% v prípade krátkeho variantu, alebo na 101 podrámcov s prekrytím 90% v prípade dlhého variantu. K tým boli pridané ich delta príznaky, ktoré sú počítane pomocou Savitsky-Golay filtrovania a vytvorili tak dvojkanálový vstup.

[1] https://github.com/karolpiczak/ESC-50

Predikcia triedy pre celú nahrávku, resp. celý rámec bola vykonaná buď na základe majority alebo na základe pravdepodobnosti predikovanej triedy z každého segmentu, čo v kombinácii s krátkym variantom poskytovalo najlepšie výsledky. Na obrázku 3 je možné vidieť architektúru modelu, ktorý navrhol K. Piczak. Táto sieť má v krátkom variante nasledujúce parametre:

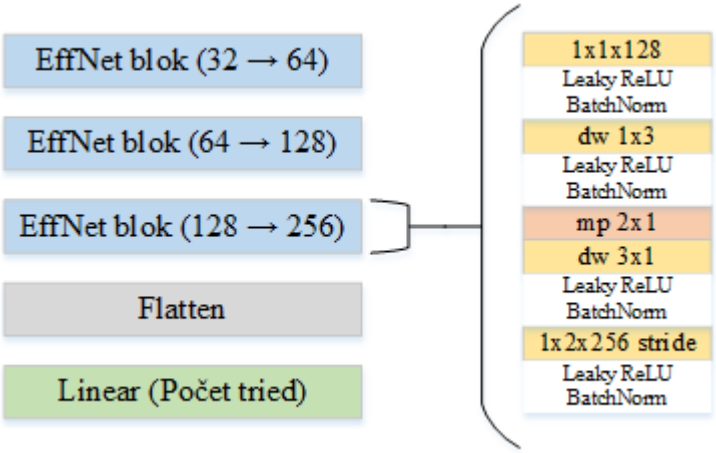
- Trénovateľných parametrov: 26 534 130
- Veľkosť parametrov: 106,14 MB
- Celkový počet násobenie-sčítanie operácií: 34,54 $*10^6$



Obrázok 3 Architektúra konvolučnej siete K. Piczak [67]

Keďže výsledky konvolučných neurónových sietí, dosahujú v oblasti rozpoznávania obrazu vynikajúce výsledky, rozhodli sme sa čerpať inšpiráciu práve v týchto architektúrach. Konkrétne sa teda jedná o architektúry, ktoré boli navrhnuté s ohľadom na výpočtovú efektivitu, určené pre mobilné zariadenia. Navrhnutých bolo viacero takýchto konvolučných sietí. My sme v rámci rozboru zvažovali troch kandidátov ShuffleNet [69], MobileNet [70] a EffNet [71]. V rámci úvahy sme zhodnotili, že architektúry ShuffleNetu a MobileNetu obsahujú komplexné bloky, v prípade ShuffleNetu je to miešanie kanálov, v prípade MobileNetu je to invertované residuálne prepojenie „bottleneck“ bloku, a tým sťažujú ich možnú úpravu pre potreby našej práce. Preto sme sa rozhodli adaptovať EffNet, nakoľko jeho pomerne jednoduchá architektúra ponúka priestor pre ďalšie úpravy a je to teda vhodná štartovacia sieť. Na obrázku 4 môžeme vidieť zobrazenie architektúry EffNet. Ako môžeme vidieť, táto sieť pozostáva z troch tzv. EffNet blokov. Jeden z týchto blokov je zobrazený detailne. Na začiatku tohto bloku je vykonaná bodová konvolúcia (angl. pointwise convolution), následne je použitý špeciálny druh konvolúcie, ktorý sa nazýva priestorovo oddeliteľná konvolúcia (angl. spatial separable convolution), ktorej princíp je nahradenie jednej operácie konvolúcie s veľkosťou jadra VxŠ, dvomi operáciami konvolúcie s veľkosťami jadra, najprv Vx1 a následne 1xŠ, pomocou čoho zníži počet parametrov. V prípade EffNet bloku bola medzi tieto dve konvolučné operácie vložená operácia združovania podľa maximálnej hodnoty, max-pool. Zároveň je za každú konvolučnú vrstvu pridaná kombinácia vrstva nelineárnej aktivačnej funkcie Leaky ReLU a vrstva dávkovej normalizácie, ktorá normalizuje vstup do vrstvy a dovolí tak využitie vyšších hodnôt parametrov učenia a zároveň zníži počet potrebných epoch pre trénovanie. Posledná vrstva tohto bloku je konvolučná vrstva s krokom konvolúcie rovnakým ako je veľkosť jadra, aby došlo k ďalšej redukcii dát. Po týchto troch blokoch nasleduje vrstva „Flatten“, ktorá

transformuje vstup na jednorozmerný výstupný vektor, ktorý je možné spracovať za pomoci plne-prepojenej vrstvy (Linear), ktorá vykonáva finálnu klasifikáciu.



Obrázok 4 Architektúra EffNet s detailným EffNet blokom. „dw“ znamená hĺbková konvolúcia (angl. depthwise convolution) a „mp“ znamená združovanie podľa maxima (z angl. max-pool)

4.3 Predspracovanie dát

Pri tvorbe tohto datasetu K. Pizcak rekonvertoval všetkých 2 000 nahrávok na jednotný formát:

- vzorkovacia frekvencia: 44 100 Hz,
- kanály: Mono,
- kodek: 16 bit PCM (S16 LE).

V rámci predspracovania dát sme sa rozhodli podvzorkovať tento akustický signál na vzorkovaciu frekvenciu 16 000 Hz. Hlavný dôvod, prečo sme sa rozhodli podvzorkovať tieto akustické signály, bola snaha o redukciu rozmernosti dát. V rámci našej práce využívame metódu pásmovo-obmedzenej sinc interpolácie, ktorá je implementovaná aj v knižnici librosa, aj PyTorch audio, ktorá je ideálna pre digitálny audio signál [89].

Z hľadiska procesu klasifikácie sme sa rozhodli pre schému používajúcu rámce, teda nahrávka je vopred rozdelená na rámce. Z týchto rámcov sú extrahované príznaky a tieto sú následne použité pre trénovanie alebo testovanie. Rozhodovanie klasifikátora o triede zvuku je vykonávané na každý rámec zvlášť, preto po sebe idúce rámce môžu patriť do rozličných tried. Nevýhoda tejto procesnej schémy je v tom, že niektoré zvuky sú krátkodobé, napríklad rozbitie skla a niektoré sú dlhodobé, napríklad búrka; preto je potrebné sa pri voľbe veľkosti rámca riadiť kompromisom. Ak je rámec príliš krátky, potom dlhodobé zmeny signálu nebudú zahrnuté počas extrakcie príznakov, naopak ak je rámec príliš dlhý, krátkodobé zvuky sa môžu stratiť [68].

Ďalší problém vyvstával, ako bolo spomenuté, s tým, že užitočný zvuk sa v nahrávke môže nachádzať len určité percento času. Prvá možnosť, ako tento problém vyriešiť je ignorovať ho, ale to by malo za následok slabo anotované dáta, resp. zle anotované dáta, nakoľko by rámec neobsahoval zvuk, ktorému je priradený, čo by malo za následok pokles presnosti rozpoznávania. Ďalšia metóda je orezať ticho alebo „prázdno“ a následne vytapetovať odrezaný čas užitočným zvukom tak, aby bola zachovaná dĺžka nahrávky 5 sekúnd, zároveň tak dataset zostane balancovaný . Túto metódu sme spočiatku využívali, ale neskôr bola vypustená, pretože neposkytovala signifikantné zlepšenie presnosti rozpoznávania; druhý dôvod bolo to, že táto metóda zavádza periodicitu tam, kde prirodzene nie je, napríklad rozbitie skla sa, požitím tejto metódy, stalo rozbitím niekoľkých skiel. Preto sme neskôr implementovali metódu, ktorá oreže ticho a ponechá len užitočný zvuk, čo má síce za následok, že dataset nie je plne balancovaný, ale inak nevytvára žiadnu nevýhodu.

Odstraňovanie ticha bolo vykonávané pomocou knižnice librosa[2]. Jedným z bežných problémov pri použití orezania ticha, ako metódy predspracovania pre celý dataset, je nastavenie hraničnej hodnoty v decibeloch, a teda hodnoty pod hranicou budú vyhodnocované ako ticho. Analyzované boli úrovne 10 dB, 20 dB, 40 dB, 50 dB a 80 dB s tým, že referenčná hodnota je maximálna hodnota výkonu signálu. Hranica orezávania ticha na úrovni 40 dB sa ukázala ako najvhodnejšia, nakoľko nedochádza k možnej strate užitočnej informácie ako pri 10 dB a 20 dB ale orezáva ticho dostatočne skoro, aby užitočný zvuk zostal dominantný.

4.4 Extrakcia príznakov

Na rozdiel od úloh klasifikácie obrazu, klasifikácia environmentálnych zvukov predpokladá využitie lokálne korelovaných jednorozmerných signálov, čiže vstup je natiahnutý pozdĺž jednej osi. Reprezentácia akustického signálu je odlišná od vizuálnych signálov, ako sú fotografie, ktoré majú lokálne korelácie v oboch priestorových dimenziách. Z toho dôvodu bolo navrhnutých niekoľko metód špecificky pre oblasť audio signálu. Jednou z oblastí týchto metód sú Vopred vypočítané časovo-frekvenčné reprezentácie a ich následne spracovanie pomocou 2D konvolučnej siete. Do tejto kategórie patrí aj Piczaková sieť, ktorá operuje so vstupom vo forme Mel spektrogramov. Základom väčšiny týchto metód je Krátkodobá Fourierová transformácia, a následne môže byť spektrum ďalej spravované, napríklad do podoby Mel spektrogramov, Mel-Frekvenčných kepstrálnych koeficientov, spektrálny kontrast, chromagram a ďalšie. Modely strojového učenia potom môžu využívať jednu z týchto metód, tzn. jedno-príznakový vstup alebo viacero týchto metód v kombinácii a vytvoriť tak viac-príznakový vstup, ako je napríklad Piczakov model, ktorý používa mel spektrogramy v kombinácii s delta príznakmi.

4.4.1 Metóda reformácie spektrogramu

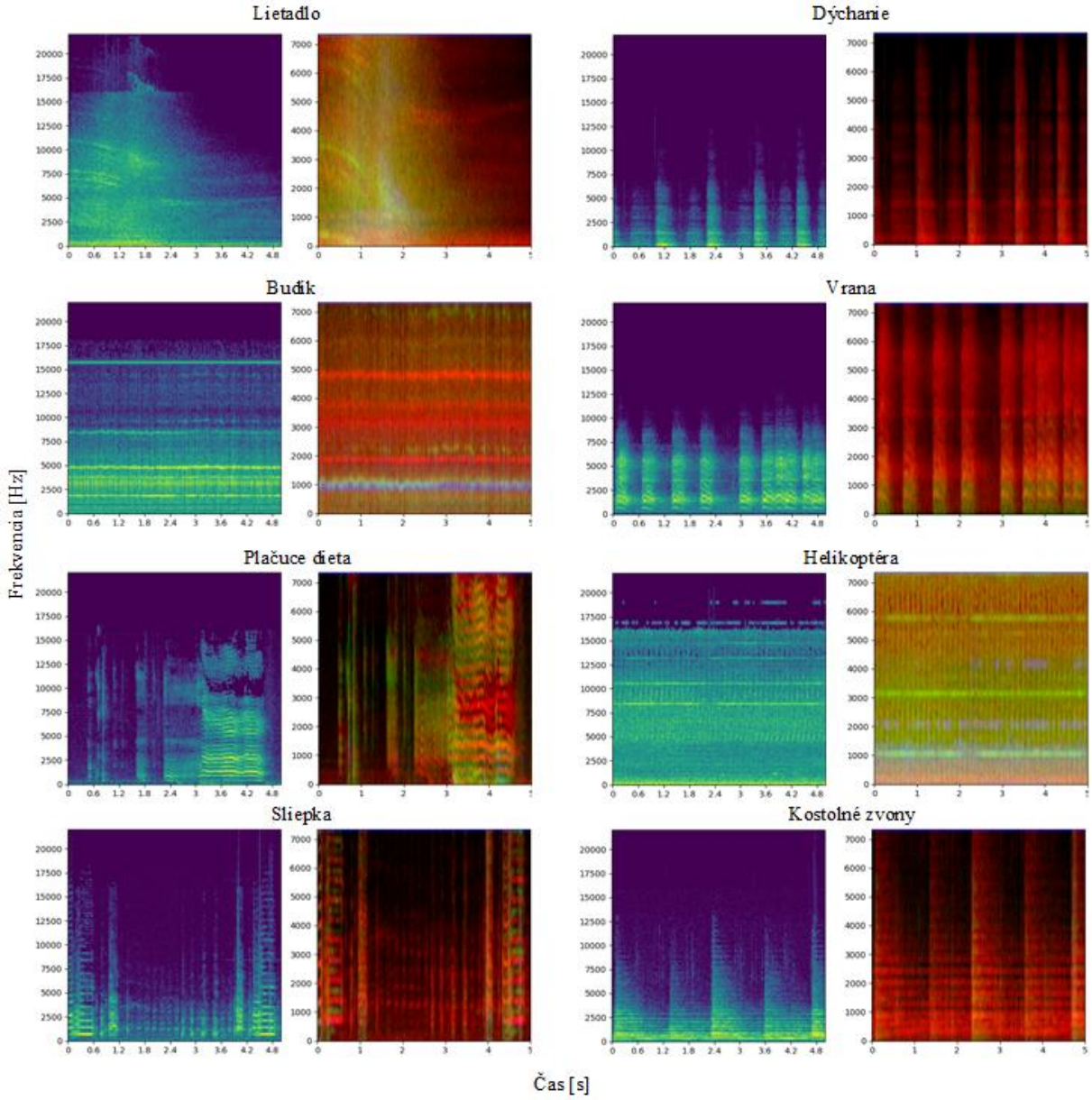
Ako je možné si povšimnúť, veľa z vyššie spomenutých metód má svoj pôvod v rozpoznávaní reči alebo hudby. My sme sa v našej práci rozhodli nepoužiť žiadnu z týchto metód. Tieto metódy síce ponúkajú dodatočnú redukciu rozmernosti vstupných dát, avšak prístupov založených na týchto metódach, ako tých čo využívajú Mel spektrogramy je už dostatok alebo ako v prípade Mel-Frekvenčných kepstrálnych koeficientov sa ukázalo, že neprodukujú príliš silné riešenia.

V rámci našej práce sme sa rozhodli využívať transformáciu vstupného signálu pomocou STFT, na základe čoho môžeme získať výkonový spektrogram, ktorý prevedieme do logaritmickej mierky.

Je vhodné poznamenať, že veľkosť konvolučnej neurónovej siete rastie spolu s veľkosťou vstupných dát. Zvyšuje sa počet jej parametrov, z toho vyplývajúca celková veľkosť, ako aj počet operácií násobenie-sčítanie. Tento nárast je zvlášť signifikantný v prípade plne- prepojenej vrstvy, pomocou ktorej je vykonávaná klasifikácia.
Ako metódu redukcie rozmernosti vstupných dát sme si zobrali za príklad rozpoznávanie farebného obrazu. V tomto prípade majú vstupné dáta tri kanály, teda jeden kanál pre každú farebná zložku formátu RGB. V rámci tejto metódy je vstupný jednokanálový spektrogram rozdelený pozdĺž frekvenčnej osi na tri frekvenčné pásma, ktoré sú potom mapované do troch kanálov. Spodné frekvenčné pásmo ako červená zložka farby, stredné frekvenčné pásmo ako zelená zložka farby a vrchné frekvenčné pásmo ako modrá zložka farby. Výsledné RGB zobrazenie je následne spracovávané konvolučnou neurónovou sieťou. Takto mapované vstupné dáta ponúkajú redukciu počtu parametrov konvolučnej neurónovej siete, jej celkovej veľkosti a počtu operácii násobenie-sčítanie na približne 35 % oproti prípadu, kedy sú ako vstupné dáta použité jednokanálové spektrogramy. Je nutné poznamenať, že v prípade tejto

[2] https://librosa.org/doc/latest/index.html

metódy nedochádza k zníženiu počtu hodnôt vo vstupných dátach; tato metóda je využitá pre redukciu rozmernosti konvolučnej siete ako takej. Na obrázku 5 je možné vidieť ukážky aplikácie tejto metódy- reformácie spektrogramu do RGB zobrazenia.



Obrázok 5 Príklady mapovania spektrogramov ako RGB zobrazenie

Následne bolo nutné tieto dáta pred vstupom do konvolučnej siete normalizovať, nakoľko tieto majú tendenciu dosahovať lepších výsledkov, ak sú vstupné dáta normalizované.

Všeobecne sa využívajú dva spôsoby normalizácie:

Min-max normalizácia, pri ktorej je výstup škálovaný do požadovaného rozsahu, bežne sa škáluje na rozsah <0,1> alebo <-1,1>, toto je vykonávané podľa vzorca:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}(b - a) + a \qquad (4)$$

kde x_{norm} je normalizovaný výstup, x je vstup, x_{min} je minimálna hodnota zo vstupu, x_{max} je maximálna hodnota vstupu a $\langle a, b \rangle$ sú hraničné hodnoty požadovaného rozsahu.

Druhou všeobecné používanou metódou normalizácie je tzv. z-skóre normalizácia, pomocou ktorej dosiahneme, že všetky vstupné dáta budú mať priemer rovný nule a smerodajnú odchýlku rovnú jednej. Táto normalizácia sa vykonáva podľa vzorca:

$$x_{norm} = \frac{x - \mu}{\sigma} \quad (5)$$

kde μ je priemerná hodnota a σ je smerodajná odchýlka.
V rámci úvodných testov sa ukázalo, že konvolučná sieť dosahuje lepšie výsledky v prípade z-skóre normalizácie ako min-max a preto bola táto metóda implementovaná do ďalších experimentov.

4.4.2 Experiment – Veľkosť analyzovaného rámca

Ako už bolo spomínané, konvolučná sieť rastie s veľkosťou vstupu, ktorý je priamo ovplyvnený veľkosťou analyzovaného rámca. Jedným z našich cieľov bolo, aby bol zvuk klasifikovaný na základe nahrávky do jednej sekundy. Konečná veľkosť rámca by však mala brať do úvahy krátkodobé aj dlhodobé zvuky. Preto sme sa rozhodli vykonať experimentálne meranie vplyvu veľkosti analyzačného rámca na presnosť EffNetu, ako nami adaptovanej siete.
V rámci tohto experimentu sme trénovali EffNet na rámcoch s čoraz väčšou veľkosťou od 250 ms po 1 sekundu s inkrementom po 10 ms a zaznamenávali sme presnosť rozpoznávania po natrénovaní. Pre každé nastavenie analyzačného rámca bol EffNet natrénovaný 10 krát. Je vhodné poznamenať, že čím menší je analyzačný rámec, na tým viacej rámcov je rozdelená jedna nahrávka a z toho vyplýva viacej dát pre trénovanie, čo môže mať za následok lepšie rozpoznávanie. Pre prácu s audio signálom bola využitá knižnica librosa.
Parametre extrakcie príznakov:

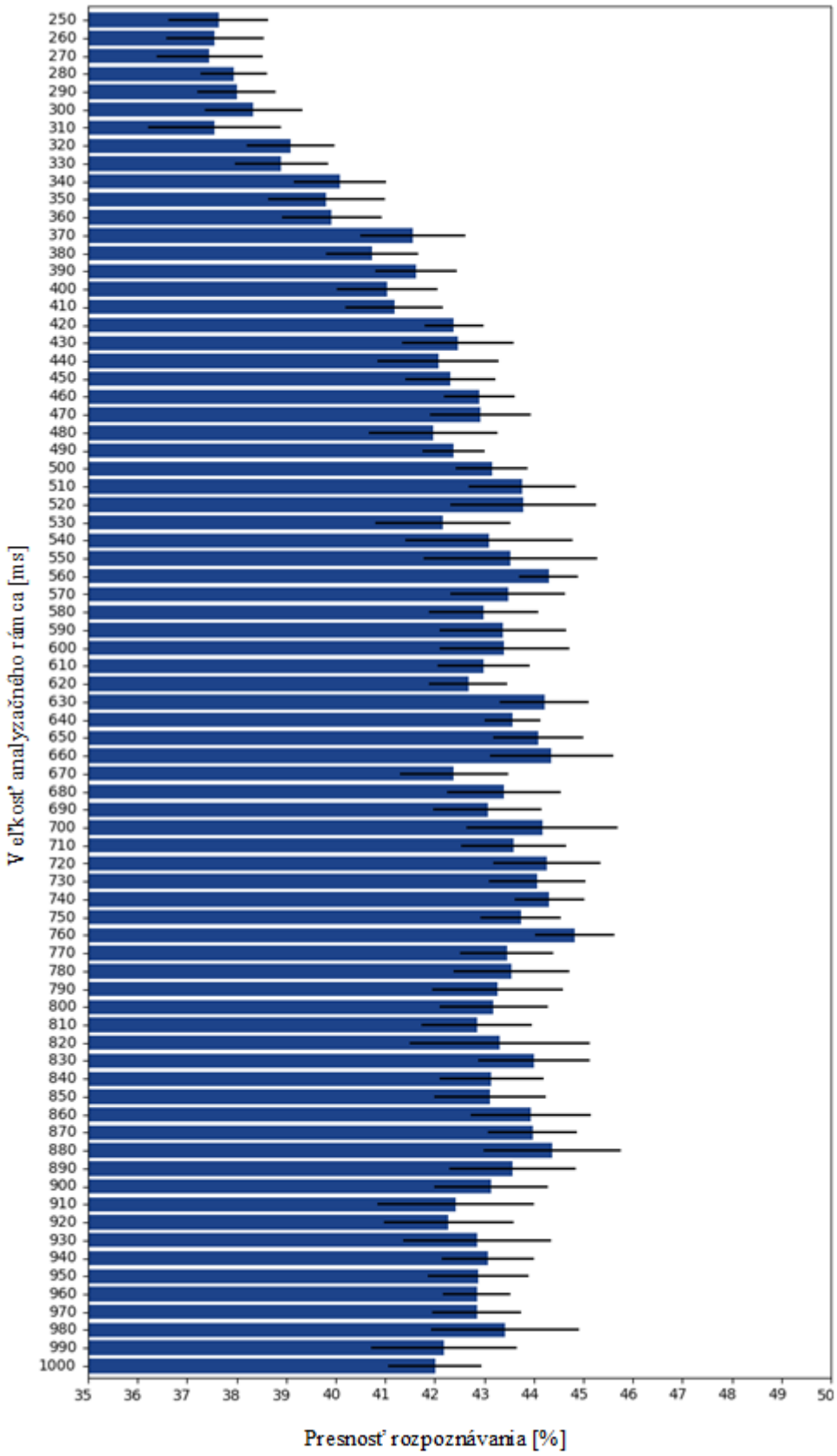
- podvzorkovanie na 16 000 Hz,
- prekrytie jednotlivých rámcov: 50%,
- dĺžka STFT: 512 vzoriek s prekrytím 50%,
- váhová funkcia: Hanning,
- ticho bolo vyplnené užitočným zvukom.

Pre implementáciu konvolučnej siete EffNet bol použitý framework PyTorch[3] a na trénovanie bola použitá knižnica Ignite[4]. Ako bolo spomenuté vyššie architektúra EffNetu používa pre finálnu klasifikáciu plne-prepojenú vrstvu, ktorej veľkosť sa mení s veľkosťou vstupu. To predstavovalo problém, nakoľko rôzne veľkosti analyzačného rámca vyžadovali rôzne nastavenia vstupu plne-prepojenej vrstvy. Riešenie pozostávalo z orezania EffNetu po treťom bloku, následne bol vykonaný jeden prechod sieťou, s použitím požadovaného rámca, a na základe výsledku bola vypočítaná potrebná veľkosť vstupu do plne-prepojenej vrstvy. Toto bolo vykonané po každej zmene veľkosti rámca.
Nastavenie parametrov trénovania:

- počet epoch: 15,
- veľkosť dávky: 16,
- optimalizátor: Stochastický gradientový zostup (z angl. Stochastic Gradient descent, SGD) s použitím Nestrovho momentu 0.9,
- parameter učenia (angl. learning rate): 0,005,
- stratová funkcia: Negatívna logaritmická vierohodnosť (z angl. Negative log likelihood loss, NLLLoss),
- zoslabovanie váh: 0,001.

Na obrázku 6 je možné vidieť graf predstavujúci vyhodnotenie tohto experimentu. Pre každú veľkosť analyzačného rámca bola vypočítaná priemerná presnosť rozpoznávania po 15 epochách trénovania, čierna čiara potom predstavuje smerodajnú odchýlku.

[3] https://pytorch.org/
[4] https://pytorch.org/ignite/index.html



Obrázok 6 Vyhodnotenie experimentu vplyvu veľkosti analyzačného rámca

Môžeme pozorovať, že aj keď rozdelenie nahrávok na menšie analyzačné rámce (250 – 310 ms) poskytuje viacej rámcov pre trénovanie, ich priemerná presnosť bola nízka. Najhoršiu priemernú presnosť rozpoznávania 37,43% dosahoval model pri použití veľkosti

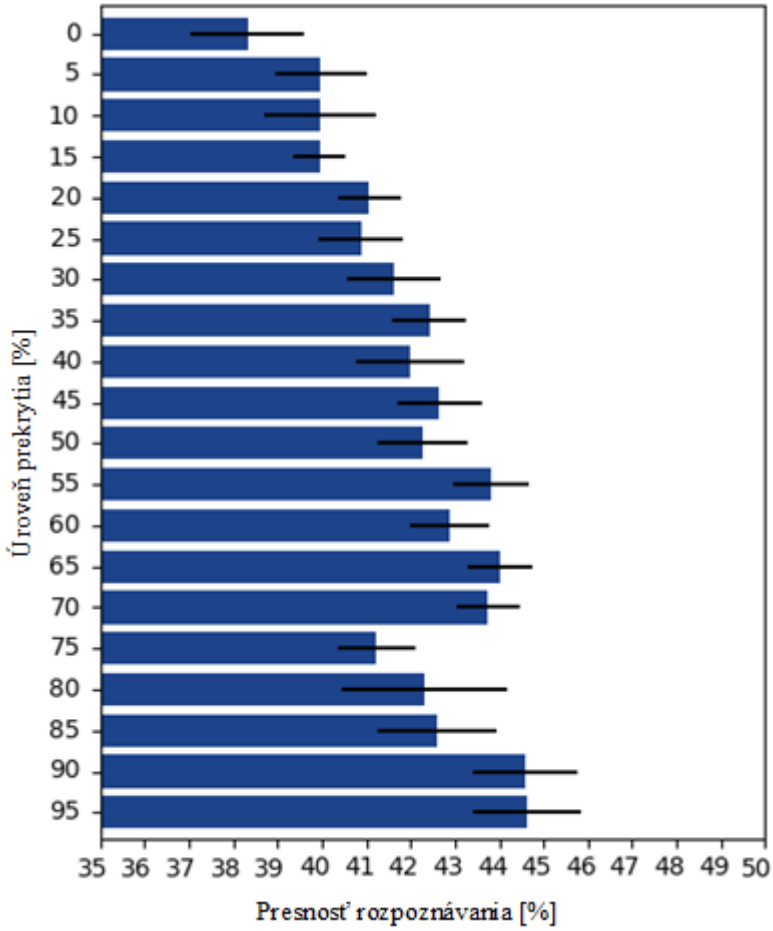
analyzačného rámca 270 ms. Najlepšia priemerná presnosť rozpoznávania bola dosiahnutá pri veľkosti okna 760 ms s priemernou presnosťou rozpoznávania 44,81%. Zároveň však toto nastavenie vykazovalo relatívne malú smerodajnú odchýlku, oproti nastaveniu, ktoré boli v presnosti za ním (550, 660, 880). Z týchto dôvodov sme sa rozhodli implementovať práve veľkosť analyzačného rámca 760 ms pre budúce trénovanie.

4.4.3 Experiment – Veľkosť prekrytia

V rámci tohto experimentu sme zisťovali, ako úroveň prekrytia ovplyvňuje presnosť rozpoznávania. Bola nastavená fixná veľkosť analyzačného rámca, následne bol EffNet natrénovaný 10 krát pre každé nastavenie úrovne prekrytia od 0 – 95% s inkrementom po 5%. Parametre extrakcie príznakov:

- podvzorkovanie na 16 000 Hz,
- veľkosť analyzačného rámca: 500 ms,
- dĺžka STFT: 512 vzoriek s prekrytím 50%,
- váhová funkcia: Hanning,
- ticho bolo vyplnené užitočným zvukom.

Nastavenie parametrov učenia konvolučnej siete bolo rovnaké ako v predchádzajúcom pokuse. Na obrázku 7 je možné vidieť graf predstavujúci vyhodnotenie tohto experimentu.



Obrázok 7 Vyhodnotenie experimentu úrovne prekrytia

Pre každé nastavenie bola vypočítaná priemerná presnosť rozpoznávania a smerodajná odchýlka, ako môžeme vidieť v grafe. Je vhodné poznamenať, že čím vyššie je prekrytie, tým viacej analyzačných rámcov je možné generovať z jednej nahrávky, avšak v prípade príliš vysokého prekrytia tieto rámce budú značne korelované. Keďže my klasifikujeme jednotlivé rámce samostatne a nie sekvencie menších rámcov, je vhodné, aby použitý dataset

vstupných dát nebol príliš korelovaný, preto sme vyradili nastavenia prekrytia 90% a 95%. V prípade, že by sme použili klasifikačnú schému rozdelenia na rámce, bolo by vhodné použitie týchto vyšších nastavení prekrytia. Rozhodli sme sa použiť nastavenie prekrytia 65% , nakoľko pri tomto nastavení dosahoval EffNet dobré výsledky a zároveň vzniknutý dataset analyzačných rámcov nebol príliš korelovaný.

4.4.4 Augmentácia dát

Počas analýzy postupov klasifikácie sme zistili, že implementovanie jednoduchých techník augmentácie dát môže mať za následok zlepšenie presnosti klasifikácie, resp. fungovať ako regulátor zabraňujúci preučeniu konvolučnej siete. Pomocou augmentácie dát je možné synteticky vygenerovať nové anotované dáta z už existujúcich, a tým tak efektívne rozšíriť trénovaciu množinu dát. Jednoduchú formu augmentácie dát je možné vykonať pomocou miernej modifikácie vstupných dát. Existuje viacero kategórií augmentácie dát, zväčša s ohľadom na signál, na ktorý sú tieto techniky aplikované, čiže na akustický signál, na vytvorené spektrogramy, miešanie viacerých nahrávok a ďalšie. My sme aplikovali modifikácie na čistý akustický signál pred transformáciou na vstupné reprezentácie dát, v našom prípade pred aplikovaním STFT.
Bežné techniky, ktoré sú aplikované priamo na akustický signál:

- Natiahnutie v čase – pomocou tejto techniky upravujeme temporálnu charakteristiku akustického signálu (zrýchlenie alebo spomalenie) bez úpravy jeho spektrálnej charakteristiky.
- Posun po časovej osi – táto technika spôsobí, že užitočný zvuk je posunutý po časovej osi bez zmien jeho temporálnej alebo spektrálnej charakteristiky. Táto technika je efektívne aplikovaná použitím prekrytia analyzačných rámcov.
- Posun výšky tónu – použitím tejto techniky zachovávame temporálnu charakteristiku signálu a manipulujeme jeho spektrálnu charakteristiku (zvýšenie alebo zníženie), je to v podstate opak natiahnutia v čase.

4.5 Zmeny trénovacieho procesu

Z rámci optimalizácie trénovania procesu sme experimentovali s viacerými nastaveniami.

Začiatočné hodnoty synaptických váh majú signifikantný efekt na proces trénovania. Tieto hodnoty by mali byť volené náhodne, ale zároveň podľa určených pravidiel, aby sa zabránilo stavom ako sú miznúci gradient alebo explodujúci gradient. Ak nastane niektorý z týchto stavov, gradient stratovej funkcie bude príliš malý alebo naopak veľký, aby spätná propagácia bola užitočná a konvergencia umelej neurónovej siete bude trvať dlho, resp. vôbec nenastane. Z tohto hľadiska bolo vytvorených niekoľko stratégií náhodného výberu hodnôt. Ako aktivačná funkcia je používaná Leaky ReLU, pri ktorej problém miznúceho gradientu nenastane, rovnako ako pri použití ReLU. Pôvodne sme pre všetky vrstvy využívali stratégiu, ktorú navrhol LeCun et al. [77], ktorá je v knižnici PyTorch predvolená. Pri nej sú náhodné hodnoty vyberané z rozloženia, ktoré má priemer nula a smerodajnú odchýlku

$$\sigma_w = \frac{1}{\sqrt{m}} \tag{6}$$

kde m je počet synaptických prepojení, ktoré vstupujú do uzla. Neskôr sme pre konvolučné vrstvy implementovali stratégiu inicializácie váh, ktorú navrhol He et al. [78], ktorá bola navrhnutá s ohľadom na použitie aktivačných funkcií ReLU, resp. Leaky ReLU, ktorá je vo svojej podstate modifikáciou stratégie LeCun. Rovnako ako pri LeCun, sú náhodné hodnoty vyberané z rozloženia, ktoré má priemer nula, avšak pre výpočet smerodajnej odchýlky je použitý nasledujúci vzorec:

$$\sigma_w = \frac{a}{\sqrt{m}} \tag{7}$$

kde m je počet synaptických prepojení, ktoré vstupujú, resp. vystupujú z uzla, v závislosti od použitého módu. Hodnota a je závislá od aktivačnej funkcie pre ReLU je to $\sqrt{2}$, v prípade nami používanej Leaky ReLU je to $\sqrt{\frac{2}{1+negativny\_sklon^2}}$. Bias je inicializovaný na nulu. V prípade Normalizačných vrstiev sme použili inicializáciu na konštantnú jednotku a bias na nulu, nakoľko pri tejto inicializácií mala stratová funkcia tendenciu klesať rýchlejšie.

Aby sme zabránili tomu, že hodnoty synaptických váh nadobudnú príliš vysokých hodnôt a taktiež, aby sme pomohli regulovať preučenie siete, využili sme techniku zoslabovania váh. V rámci tejto techniky pridávame malú penalizáciu, zvyčajne L2 norma (Eukleidivská norma) synaptických váh k stratovej funkcii. Hodnota zoslabovania váh bola nastavená na hodnotu $5 * 10^{-4}$.

Ďalším dôležitým parametrom, ktorý ma veľký vplyv na proces trénovania, je parameter učenia. Zistili sme že naša konvolučná sieť sa najlepšie trénuje s parametrom učenia $5 * 10^{-4}$. Zároveň sme sa rozhodli aplikovať dve techniky plánovania zmeny parametru učenia, ktoré napomáhajú trénovaniu a to je exponenciálne zoslabovanie parametru učenia. Pri tejto technike je po každej epoche trénovania parameter učenia vynásobený parametrom γ, v našom prípade sme zvolili hodnotu zoslabovania $\gamma = 0{,}985$. Druhá implementovaná technika je tzv. „ohrievanie“ (z angl. warm up) parametru učenia. Princíp „ohrievania“ je v tom, že je zvolená nízka hodnota parametru učenia a tá je následne zvyšovaná pokiaľ nedosiahne požadovanú hodnotu. Benefity tejto techniky boli demonštrované vo viacerých aplikáciách strojového učenia. Naša stratégia zmeny parametru učenia je teda nasledovná:

- parameter učenia je nastavený na $5 * 10^{-4}$,
- následne je lineárne „zohrievaný“ po 5 epoch, až kým nedosiahne hodnotu $5 * 10^{-3}$,
- po zvyšok trénovania je hodnota parametru učenia exponenciálne oslabovaná hodnotou $\gamma = 0{,}985$.

V rámci hľadania optimálneho nastavenia procesu trénovania sme otestovali nahradenie aktivačnej funkcie Leaky ReLu za ReLU, samozrejme s adekvátnou zmenou inicializácie váh korešpondujúcou s touto aktivačnou funkciou. Táto zmena však mala za následok pokles presnosti rozpoznávania, preto sme sa rozhodli ponechať aktivačnú funkciu Leaky ReLU. Ďalej sme testovali rôzne nastavenia úrovne negatívneho sklonu. Toto nastavenie nemalo signifikantný vplyv na presnosť rozpoznávania, preto sme úroveň negatívneho sklonu ponechali na hodnote 0,3.

Rozhodli sme sa otestovať optimalizátor Adam v základnom nastavení, ako ho definuje PyTorch, teda $\beta_1 = 0.9$ a $\beta_2 = 0.999$. Výsledná presnosť rozpoznávania sa príliš signifikantne nelíšila, preto sme sa aj naďalej rozhodli používať optimalizátor SGD s Nestrovím momentom.

4.6 Úpravy extrakcie príznakov

V rámci týchto zmien boli mierne upravené aj všeobecne nastavenia extrakcie príznakov. Bola implementovaná technika samotného orezávania ticha, bez tapetovania užitočným zvukom z dôvodov vyššie spomenutých.

Pri aplikácii STFT sme začali využívať váhovú funkciu Blackman-Harris, nakoľko poskytuje primeranú šírku pásma a veľmi nízke presakovanie spektra. V rámci zisťovania efektu dĺžky STFT, resp. zvoleného temporálneho rozlíšenia štúdie ukazujú, že tento parameter je viazaný na architektúru siete. Napríklad v štúdii [74] je ukázané, že AlexNet dosahuje lepšie výsledky pri temporálnom rozlíšení 30 ms a GoogLeNet zase pri temporálnom rozlíšení 40 ms. Na základe nášho testovania sme zistili, že naša konvolučná sieť dosahuje najlepších výsledkov pri temporálnom rozlíšení 45 ms, pri podvzorkovaní na 16 000 Hz. Zároveň sme dĺžku STFT začali počítať podľa vzorca $N = 2^a 3^b 5^c 7^d 11^e 13^f$. Uvedomujeme si, že funkcia STFT v knižnici librosa nie je optimalizovaná podľa tohto vzorca, avšak nepozorovali sme negatíva, naopak keďže podľa tohto vzorca je veľkosť váhového okna rovná dĺžke STFT, nemusíme dopĺňať váhové okno o nulové hodnoty a výsledok STFT je tak nižších rozmerov. Dĺžka STFT tak je 720 vzoriek namiesto 1024 ako by mala byť. Keďže sme použili váhovú funkciu Blackman-Harris, prekrytie váhových okien bolo nastavené na 66,1%. Na základe týchto nastavení sa nám veľkosť vstupného tenzora zmenila na 3x120x51.

Ďalej sme sa rozhodli otestovať online augmentáciu dát. Pre implementáciu online augmentácie dát bola nutná zmena stratégie predspracovania dát a extrakciu príznakov nasledovným spôsobom: fáza prípravy dát pozostávala len z podvzorkovania, orezania ticha a rozdelenia audio signálu na príslušné analyzačné okná. Počítanie spektrogramov a ich následná reformácia do RGB zobrazenia, bola vykonávaná priamo počas procesu učenia, nie vopred, ako tomu bolo pri offline augmentácii dát, kedy sú všetky tieto úkony vykonané ešte pred začiatkom procesu učenia. Vďaka tomu sme mohli implementovať aj online augmentáciu dát, ktorá bola vykonávaná s určitou pravdepodobnosťou. Teda každá vzorka mala určitú pravdepodobnosť, že na ňu bude aplikovaná niektorá z augmentácií. Myšlienka za týmto postupom je, aby sa dáta neustále menili, resp. aby sme model počas trénovania vystavili čo najväčšej rozmanitosti dát. Rozhodli sme sa aplikovať štyri druhy augmentácie dát, z ktorých bude použitá najviac jedna:

- posun výšky tónu,
- obrátenie časovej osi,
- zmena hlasitosti,
- pridanie farebného šumu.

Z hľadiska pravdepodobnosti bola určená pravdepodobnosť 50%, že bude daná vzorka augmentovaná; v prípade, že áno, pravdepodobnosť jednotlivých augmentácií bola rozložená rovnomerne. Je teda možné, že jedna vzorka bude počas jednej epochy augmentovaná a počas druhej už nie, čo by malo prispieť k robustnosti nášho modelu a následnému zlepšeniu presnosti rozpoznávania.

Naša hypotéza sa však nepotvrdila. Takto natrénovaná sieť dosahovala obdobných výsledkov ako v prípade využitia offline augmentácie dát. Zároveň určovanie pravdepodobnosti augmentácie dát pre každú vzorku samostatne predĺžilo proces trénovania. Z tohto dôvodu sme sa rozhodli ponechať offline augmentáciu dát a extrakciu príznakov.

4.7 Zmeny v architektúre EffNet

V tejto časti budú popísané ďalšie zmeny vykonané na tejto architektúre, aby sme docielili vyššieho rozpoznávania a ďalej znížili počet parametrov.

Pôvodný EffNet patril medzi tzv. konvenčné konvolučné siete, teda konvolučné siete, ktoré vykonávajú konvolúciu v nižších vrstvách a pre klasifikáciu sú výstupné mapy príznakov poslednej konvolučnej vrstvy vektorizované a následne spracované plne-

prepojenou vrstvou. Takáto štruktúra premosťuje konvolučnú štruktúru s tradičnými klasifikátormi na báze neurónovej siete. Avšak plne-prepojené vrstvy majú tendenciu sa preučiť, a zároveň sú „drahé“ z pohľadu parametrov, keďže jej veľkosť rastie spolu so vstupom. Preto sme sa rozhodli implementovať za poslednou konvolučnou vrstvou vrstvu globálneho združovania podľa priemeru (z angl. Global Average Pooling). Jednou z výhod tejto vrstvy je, že vynucuje zhody medzi mapami príznakov a kategóriami, čiže mapy príznakov môžu byť ľahšie interpretované ako mapy istoty kategórie [75]. Ďalšou výhodou je, že rapídne znižujú veľkosť plne-prepojenej vrstvy a zároveň túto veľkosť fixujú na jednu hodnotu, čím prestáva byť táto vrstva závislá od veľkosti vstupu.

Uvažovali sme o nahradení priestorovo oddeliteľnej konvolúcie iným typom bloku, napríklad hĺbkovo oddeliteľnou konvolúciou, nakoľko pri použití priestorovo oddeliteľnej konvolúcie je nutná špecifická symetria jadra, aby ho bolo možné rozdeliť, čo má za následok obmedzené množstvo jadier, ktoré je možné použiť. V prípade hĺbkovo oddeliteľnej konvolúcie takáto požiadavka nevzniká. Avšak priestorovo oddeliteľná konvolúcia dovoľuje využiť združovanie podľa maxima medzi jednotlivými vrstvami. To má za následok zníženie celkového počtu parametrov, čo je v súlade s našimi cieľmi, preto sme sa rozhodli ponechať tento typ konvolučného bloku.

Následne sme nahradili poslednú konvolučnú vrstvu v bloku za bodovú konvolučnú vrstvu a pridali sme vrstvu združovania podľa maxima 1x2 s príslušným krokom, čím sme ďalej redukovali počet trénovateľných parametrov. Avšak to malo za následok, že sieť stratila schopnosť sa doučovať. Preto sme sa rozhodli znovu reštrukturalizovať výpadové vrstvy, nakoľko ich hodnota pravdepodobnosti výpadu bola príliš vysoká pre tak malú sieť. Na základe experimentov sme určili, že po prvom bloku nie je výpadová vrstva vhodná vôbec. Dvojrozmerné výpadové vrstvy, ktoré nasledovali zvyšné bloky, mali znížený hodnotu pravdepodobnosti výpadu $p = 0{,}2$. Vo výpadovej vrstve pred plne-prepojenou vrstvou zostalo ponechané nastavenie pravdepodobnosti výpadu $p = 0{,}5$. Taktiež bol v rámci trénovania znížený parameter zoslabovania váh na hodnotu $1*10^{-4}$. Po týchto zmenách bola sieť znovu schopná sa doučiť.

Skúšali sme nahradiť vrstvu normalizácie dávky (angl. BatchNorm) za vrstvu renormalizácie dávky (angl. BatchRenorm) podľa článku [76], avšak neprinieslo to poznateľné zlepšenie presnosti rozpoznávania, preto sme sa rozhodli ponechať klasickú vrstvu normalizácie dávky, zároveň však využitie tejto vrstvy spôsobilo predĺženie procesu trénovania.

Ďalej sme sa rozhodli zmeniť počty kanálov jednotlivých blokov. Prvotne sme len pridali blok s počtom kanálov 512, resp. 1024, avšak tieto nastavenia neposkytli tak významné zlepšenie presnosti rozpoznávania, aby to ospravedlnilo zvýšenie počtu parametrov. Rozhodli sme sa preto reštrukturalizovať počty kanálov v celej konvolučnej sieti.

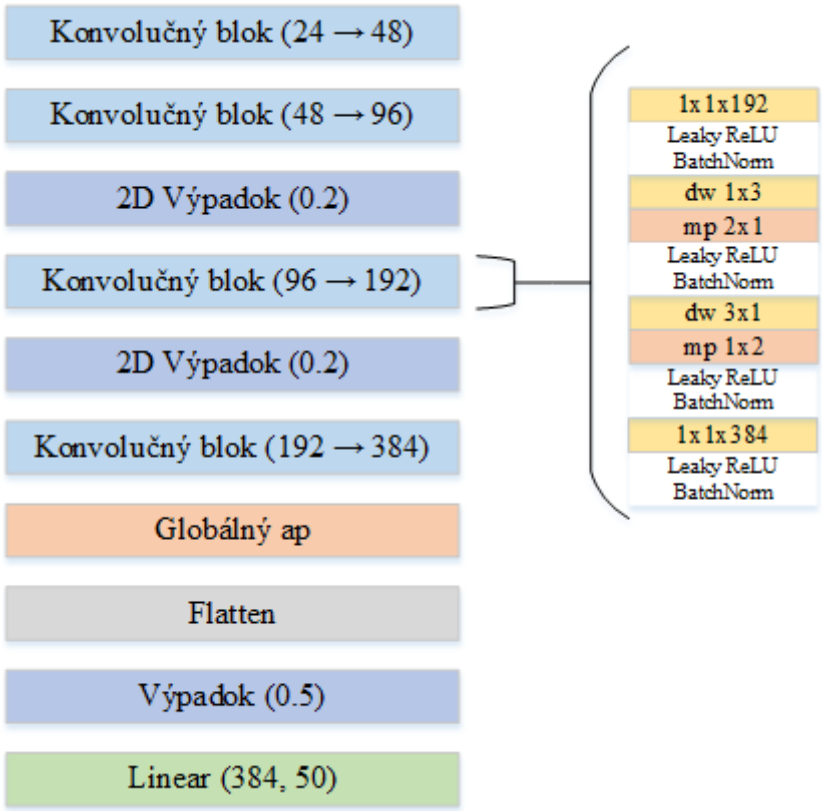
Konečná architektúra našej siete je zobrazená na obrázku 8.

Vykonané zmeny považujeme za natoľko významné, že túto sieť už nepovažujeme za EffNet, ale za novú konvolučnú neurónovú sieť.

V tejto konfigurácii má naša konvolučná sieť nasledujúce parametre:

- veľkosť vstupného tenzora: 3x120x51,
- trénovateľných parametrov: 171 386,
- veľkosť parametrov: 0,69 MB,
- celkový počet násobenie-sčítanie operácií: 18,37 $*10^6$.

Pri porovnaní s Piczakovým modelom zistíme, že naša konvolučná sieť operuje len s 0,65% trénovateľných parametrov oproti Piczakovmu modelu.



Obrázok 8 Architektúra našej konvolučnej neurónovej siete s detailom konvolučného bloku. „dw" znamená hĺbková konvolúcia, „mp" znamená združovanie podľa maxima a „ap" združovanie podľa priemeru.

4.8 Krížová validácia

Všetky doterajšie merania presnosti boli vykonané pri rozdelení datasetu na trénovacie, validačné a testovacie dáta. To však spôsobilo redukciu množiny dát pre trénovanie. Samotný autor tejto množiny dát zoradil nahrávky do piatich rovnomerných množín pre krížovú validáciu tak, aby nahrávky, ktoré pochádzajú z jedného zdrojového súboru boli vždy obsiahnuté v jednej množine [26]. Z tohto hľadiska sme sa rozhodli množinu dát nemiešať pred rozdelením na jednotlivé diely.

Preto sme sa aj mi rozhodli implementovať krížovú validáciu, nakoľko nám to ponúkne korektné porovnanie presnosti rozpoznávania s referenčným modelom. Zároveň je táto technika validácie vhodná pre menšie množiny dát, nakoľko táto je rozdelená v jednu dobu len na dve časti: trénovaciu a validačnú. Ďalšou výhodou je pomerne presný odhad klasifikačnej presnosti, ale za cenu časovej náročnosti, nakoľko treba model natrénovať viackrát.

Konkrétny typ krížovej validácie, ktorý využívame sa nazýva *k*-násobná krížová validácia, kde k je celé číslo, ktoré predstavuje počet dielov, na ktoré bude množina dát rozdelená. Typické hodnoty k sú päť a desať. Teda množina dát je rozdelená na k dielov, jeden z nich je delegovaný ako validačný a zvyšných $k-1$ sú delegované ako množina trénovacích dát, na ktorých je následne model natrénovaný. Výsledná presnosť rozpoznávania, resp. validačná presnosť je zaznamenaná a proces krížovej validácie pokračuje. Ďalší diel v poradí je delegovaný ako validačný a proces trénovania je znovu zahájený. Je vhodné poznamenať, že pred každým procesom trénovania je model nanovo inicializovaný. Tento proces je zopakovaný k krát, takže každý diel je delegovaný ako validačný práve jeden krát. Výsledkom *k*-násobnej krížovej validácie, označovaný tiež ako

krížovo-validačná presnosť, je priemerná hodnota presnosti rozpoznávania, vypočítaná z k validačných presností.

Pre určenie presnosti rozpoznávania našej siete, sme náš model natrénovali päťkrát pre každý diel a výslednú priemernú hodnotu pre každý diel sme použili pre výpočet výslednej krížovo-validačnej presnosti rozpoznávania. Výsledné priemerné presnosti rozpoznávania pre jednotlivé diely ako aj výslednú krížovo-validačnú presnosť je možné visieť v tabuľke 3.

Pridali sme taktiež metriku Top-5, ktorá predstavuje presnosť toho, koľkokrát je cieľové označenie v piatich najvyšších pravdepodobnostiach predikcie označenia našej siete. Nami doteraz používaná presnosť rozpoznávania je v podstate Top-1 metrika, ktorá predstavuje presnosť toho, koľkokrát je cieľové označenie najvyššia pravdepodobnosť predikcie označenia.

Validačný diel	Validačná presnosť Top-1	Smerodajná odchýlka Top-1	Validačná presnosť Top-5	Smerodajná odchýlka Top-5
Diel 1	59,87 %	0,66 %	84,83 %	0,26 %
Diel 2	59,46 %	0,75 %	85,36 %	0,69 %
Diel 3	62,56 %	0,81 %	85,62 %	0,36 %
Diel 4	66,06 %	0,67 %	88,70 %	0,24 %
Diel 5	59,46 %	0,71 %	85,52 %	0,82 %
Krížovo-validačná	61,48 %	2,66 %	86,01 %	1,47 %

Tabuľka 3 Výsledné presnosti rozpoznávania pri použití 5-násobnej krížovej validácie

4.9 Prenášané učenie

Keďže veľkosť nami používaného datasetu je relatívne malá, rozhodli sme sa využiť princíp tzv. prenášaného učenia (z angl. transfer learning). Hlavnou motiváciou pre implementáciu tejto metódy však bolo to, že vo viacerých štúdiách tento prístup napomáhal zvýšeniu presnosti rozpoznávania klasifikačného modelu.

V rámci našej práce induktívne prenášané učenie, keďže sme sa rozhodli využiť rozsiahlu množinu dát ImageNet, konkrétne ImageNet z roku 2012, ktorý sa sústreďoval na rozpoznávanie objektov. Táto obrovská množina dát pozostáva z viac než milióna trénovacích vzoriek rozdelených do 1 000 tried.

V rámci fázy predtrénovania sme trénovali náš klasifikačný model na tejto množine dát dvakrát a pre ďalšie spracovávanie sme použili ten najlepší model z nich. Toto obmedzené množstvo bolo preto, že jedno takého predtrénovanie trvalo vyše týždňa a to z dôvodu veľkosti množiny dát ako aj obmedzenej dostupnej výpočtovej sily. Zároveň sa však výsledné presnosti rozpoznávania príliš nelíšili a dosiahli približne 30% na testovacej množine. Uvedomujeme si, že táto hodnota nie je vysoká, ale ak vezmeme v úvahu nízky počet parametrov našej siete a fakt, že sa jedná o predtrénovanie, je tento výsledok dostatočný.

V rámci analýzy sme zistili, že existuje viacero prístupov k prenášanému učeniu. Tieto druhy zväčša závisia od veľkosti použitej množiny dát, ako aj od podobnosti úloh. Prístup, ktorý sa nám osvedčil najviac je tzn. dotrénovanie (z angl. fine-tune). Pri využití tohto prístupu je nanovo inicializovaná plne-prepojená vrstva a zvyšok modelu je inicializovaný pomocou predtrénovaného modelu, avšak žiadne parametre nie sú zmrazené, teda počas procesu trénovania si ponechajú schopnosť učiť sa. Takto inicializovaný model je následne trénovaný podľa zvolenej stratégie. Takto nastavený klasifikačný model sme znova trénovali pomocou využití 5 násobnej krížovej validácie 5 krát. Zistili sme, že klasifikačný model, ktorý bol inicializovaný týmto spôsobom dosahoval lepších výsledkov presnosti rozpoznávania ako v prípade náhodnej inicializácie, ktorú sme používali v predchádzajúcich testovaniach. Výslednú krížovo-validačnú presnosť rozpoznávania, výsledky pre jednotlivé validačné diely ako aj Top-5 presnosť je možné vidieť v tabuľke 4.

Validačný diel	Validačná presnosť Top-1	Smerodajná odchýlka Top-1	Validačná presnosť Top-5	Smerodajná odchýlka Top-5
Diel 1	59,47 %	0,15 %	86,10 %	0,20 %
Diel 2	61,22 %	0,51 %	85,94 %	0,73 %
Diel 3	63,18 %	0,68 %	86,65 %	0,17 %
Diel 4	67,96 %	0,34 %	88,97 %	0,22 %
Diel 5	59,20 %	0,52 %	84,70 %	0,32 %
Krížovo-validačná	62,21 %	3,25 %	86,47 %	1,45 %

Tabuľka 4 Výsledné presnosti rozpoznávania. Trénovanie s využitím prenášaného učenia – prístup dotrénovanie

Ako je možné si povšimnúť krížovo-validačná presnosť rozpoznávania modelu, ktorý bol inicializovaný pomocou hodnôt synaptických váh získaných z predtrénovania na množine dát ImageNet dosahuje vyšších hodnôt ako v prípade náhodnej inicializácie.

Ak teda porovnáme našu konvolučnú neurónovú sieť so zvolenou referenčnou sieťou, z hľadiska parametrov sa situácia nezmenila, teda naša sieť stále klasifikuje s použitím 0,65% veľkosti referenčného modelu, čo sa týka počtu trénovateľných parametrov a z toho vyplývajúca veľkosť modelu. Náš klasifikačný model dosahuje krízovo-validačnej presnosti rozpoznávania 62,21%, referenčný model dosahuje presnosť rozpoznávania 64,5%, ako teda môžeme vidieť náš klasifikačný model stráca 2,29% presnosti rozpoznávania oproti referenčnému. Avšak náš klasifikačný model je schopný klasifikovať na základe 760 milisekúnd akustického signálu, naproti tomu referenčný model vyžaduje 5 sekúnd.

Detailné porovnanie nášho klasifikačného modelu s referenčnou model K. Piczaka je možné vidieť v tabuľke 5.

	Referenčný (Piczakov) model	Náš klasifikačný model
Potrebná dĺžka akustického signálu	**5** s	**0.73** s
Veľkosť vstupného tenzora	$2 \times 60 \times 41$	$3 \times 120 \times 51$
Trénovateľných parametrov	**26 534 130**	**171 386**
Veľkosť parametrov	**106,14** MB	**0,69** MB
Celkový počet násobenie-sčítanie operácií	$34{,}54 \cdot 10^6$	$18{,}37 \cdot 10^6$
Presnosť rozpoznávania	**64,5 %**	**62,21 %**
Top-5 presnosť rozpoznávania	-	**86,47 %**

Tabuľka 5 Porovnanie klasifikačného modelu s referenčným

Záver

Táto práca sa zaoberá návrhom klasifikačného modelu pre klasifikáciu environmentálnych zvukov, tento model by mal byť založený na metódach strojového učenia. Hlavnou motiváciou, za vývojom tohto klasifikačného modelu, bolo jeho možné budúce využitie pre systém ochrany lesov a to pred nelegálnou ťažbou alebo nepovoleným vstupom motorových vozidiel do lesných oblastí, resp. ako základ akustického bezpečnostného systému.

Bolo nutné vykonať analýzu týchto metód s ohľadom na ich využiteľnosť pri riešení klasifikačných problémov. Dôraz bol kladený na nízku veľkosť modelu, aby v budúcnosti bola možná implementácia na zariadenie s obmedzenou výpočtovou silou. Z tohto hľadiska sa ukázalo ako najvhodnejší prístup využitie konvolučných neurónových sietí, ktorých využitie sa osvedčilo vo viacerých prístupoch.
Ďalej bola vykonaná analýza metód extrakcie príznakov s ohľadom na ich využiteľnosť spolu s konvolučnou neurónovou sieťou, ktorá vo väčšine prístupov predpokladá dvojrozmerné vstupné dáta. Z tohto dôvodu sme sa venovali transformáciám akustického signálu do časovo-frekvenčnej oblasti.

Ako inšpiráciu počas návrhu sme zobrali efektívne rozpoznávanie obrazu a náš klasifikačný model sme postavili na obdobných princípoch. Rozpoznávanie obrazu nás taktiež inšpirovalo k návrhu metódy reformácie spektrogramu, pomocou ktorej je jednokanálový spektrogram reformovaný na trojkanálové RGB zobrazenie. V rámci experimentov sme testovali vplyv dĺžky vstupného akustického signálu na presnosť klasifikácie a taktiež úroveň prekrytia medzi jednotlivými rámcami. Na základe týchto experimentov sme určili, že vhodná dĺžka akustického signálu, ktorý vstupuje do klasifikačného procesu je 0,76 sekúnd s prekrytím medzi jednotlivými rámcami 65%. Ďalej sme experimentovali s rôznymi druhmi augmentácie dát, ako aj so spôsobmi ich aplikácie (online vs. offline augmentácia). Na základe experimentov sme určili vhodnú stratégiu trénovania nášho klasifikačného modelu, ako aj jeho korektnú evaluáciu.

Bol zvolený referenčný model, proti ktorému sme porovnávali náš klasifikačný model. Náš model operuje s 0,65% veľkosti referenčného modelu. Nami navrhnutý klasifikačný model je schopný klasifikovať akustický signál s dĺžkou 0,76 sekúnd, oproti 5 sekundám, ktoré vyžaduje referenčný model. Avšak čo sa týka presnosti klasifikácie náš klasifikačný model stráca voči referenčnému modelu 2,29%.

Prínosom práce je najmä overenie vhodnosti použitia princípov efektívneho rozpoznávania obrazu pre návrh klasifikátora environmentálnych zvukov, samotný návrh klasifikačného modelu s nízkou veľkosťou, schopného klasifikácie na základe krátkeho akustického signálu. Ďalším prínosom je návrh metódy reformácie spektrokramu, pomocou ktorej sme redukovali veľkosť konvolučnej siete.

Aj keď je presnosť klasifikácie nášho modelu 62,21%, máme za to, že tento model je použiteľný ako základ akustického bezpečnostného systému, nakoľko aj keď priama presnosť rozpoznávania nie je vysoká, pri použití metriky Top-5 zistíme, že v najvyšších pravdepodobnostiach predikcie je cieľová skupina prítomna s presnosťou 86,47%, čo znamená, že je možná určitá kompenzácia presnosti.

Stručné zhrnutie, hlavným cieľom dizertačnej práce bol návrh klasifikačného modelu s nízkou veľkosťou, náš model má veľkosť 0,69 MB, tento bod teda rátame za splnený. Sekundárnym cieľom bolo, aby klasifikačný model pracoval s čo najmenšou vzorkou akustického signálu, ideálne do jednej sekundy; náš model klasifikuje na základe 0.76 sekundy, teda aj tento cieľ sme splnili. Záverečný cieľ bolo porovnanie nášho klasifikačného modelu s referenčným, výsledok tohto porovnania sa nachádza v tabuľke 5. Čím boli ciele dizertačnej práce splnené.

Referencie

[1] Cepoi, L., Donţu, N., Şalaru, V., & Şalaru, V. (2016). Removal of organic pollutants from wastewater by cyanobacteria. In *Cyanobacteria for bioremediation of wastewaters* (pp. 27-43). Springer, Cham.

[2] Bello, J. P., Silva, C., Nov, O., Dubois, R. L., Arora, A., Salamon, J., ... & Doraiswamy, H. (2019). Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, *62*(2), 68-77.

[3] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 776-780). IEEE.

[4] Schafer, R. M. (1977). Our Sonic Environment and the Tuning of the World: The Soundscape. *Vermont: Destiny Books Rochester*.

[5] Delage, B.: Paysage sonore urbain. Technical Report, Plan Construction, Paris (1979).

[6] Pijanowski, B. C., Farina, A., Gage, S. H., Dumyahn, S. L., & Krause, B. L. (2011). What is soundscape ecology? An introduction and overview of an emerging new science. *Landscape ecology*, *26*(9), 1213-1232.

[7] Guastavino, C. (2018). Everyday sound categorization. *Computational analysis of sound scenes and events*, 183-213.

[8] Salamon, J., Jacoby, C., & Bello, J. P. (2014, November). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 1041-1044).

[9] Morel, J., Marquis-Favre, C., Dubois, D., & Pierrette, M. (2012). Road traffic in urban areas: A perceptual and cognitive typology of pass-by noises. *Acta acustica united with acustica*, *98*(1), 166-178.

[10] Wold, E., Blum, T., Keislar, D., & Wheaten, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE multimedia*, *3*(3), 27-36.

[11] Wang, J. C., Wang, J. F., He, K. W., & Hsu, C. S. (2006, July). Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor. In *The 2006 IEEE international joint conference on neural network proceedings* (pp. 1731-1735). IEEE.

[12] Chu, S., Narayanan, S., Kuo, C. C. J., & Mataric, M. J. (2006, July). Where am I? Scene recognition for mobile robots using audio features. In *2006 IEEE International conference on multimedia and expo* (pp. 885-888). IEEE.

[13] Bello, J. P., Mydlarz, C., & Salamon, J. (2018). Sound analysis in smart cities. In *Computational Analysis of Sound Scenes and Events* (pp. 373-397). Springer, Cham.

[14] "Dublin City Noise web." [Online]. Dostupné: http://www.dublincitynoise.com.

[15] "Sound of New York (SONYC) web." [Online]. Dostupné: http://wp.nyu.edu/sonyc.

[16] Bello, J. P., Silva, C., Nov, O., Dubois, R. L., Arora, A., Salamon, J., ... & Doraiswamy, H. (2019). Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, *62*(2), 68-77.

[17] Farrés, J. C. (2015, June). Barcelona noise monitoring network. In *Proceedings of the Euronoise* (pp. 218-220).

[18] Laiolo, P. (2010). The emerging significance of bioacoustics in animal species conservation. *Biological conservation*, *143*(7), 1635-1645.

[19] Mporas, I., Ganchev, T., Kocsis, O., Fakotakis, N., Jahn, O., Riede, K., & Schuchmann, K. L. (2012, November). Automated acoustic classification of bird species from real-field recordings. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence* (Vol. 1, pp. 778-781). IEEE.

[20] Walters, C. L., Freeman, R., Collen, A., Dietz, C., Brock Fenton, M., Jones, G., ... & Jones, K. E. (2012). A continental-scale tool for acoustic identification of E uropean bats. *Journal of Applied Ecology*, *49*(5), 1064-1074.
[21] Stowell, D. (2018). Computational bioacoustic scene analysis. In *Computational analysis of sound scenes and events* (pp. 303-333). Springer, Cham.
[22] "Audio Analytic web" [Online]. Dostupné: https://www.audioanalytic.com/
[23] Krstulović, S. (2018). Audio event recognition in the smart home. *Computational Analysis of Sound Scenes and Events*, 335-371.
[24] Kumar, D. P., Amgoth, T., & Annavarapu, C. S. R. (2019). Machine learning algorithms for wireless sensor networks: A survey. *Information Fusion*, *49*, 1-25.
[25] Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw*, *19*(1-9), 2.
[26] Piczak, K. J. (2015, October). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1015-1018).
[27] Warr, K. (2019). *Strengthening deep neural networks: making AI less susceptible to adversarial trickery*. O'Reilly Media.
[28] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.
[29] Hebb, D. O. (1949). The first stage of perception: growth of the assembly. *The Organization of Behavior*, *4*, 60-78.
[30] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.
[31] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554-2558.
[32] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science.
[33] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541-551.
[34] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.
[35] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.
[36] Du, K. L., & Swamy, M. N. (2006). *Neural networks in a softcomputing framework* (Vol. 501). London: Springer.
[37] Simon, H. (2009). Neural Networks and Learning Machines. Third Edition /Simon Haykin.–the USA.
[38] Sinčák, P., & Andrejková, G. (1996). Neurónové siete Inžiniersky prístup (1. diel). *Elfa: Kosice*.
[39] Šíma J., Neruda R. (1996). Teoretické otázky neuronových sítí, 1. vyd. Praha: MATFYZPRESS, Dostupne: http://www2.cs.cas.cz/~sima/kniha.pdf
[40] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.
[41] Karpathy, A.: CS231n: Convolutional Neural Networks for Visual Recognition. 2017. Dostupné: http://cs231n.stanford.edu/
[42] Goodfellow, I.; Bengio, Y.; Courville, A.: Deep Learning. MIT Press, 2016, Dostupné: http://www.deeplearningbook.org.

[43] Le, Q. V. (2015). A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks. *Google Brain*, *20*, 1-20.
[44] Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.
[45] Dumoulin, V., & Visin, F. (2016). A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.
[46] Rao, D., & McMahan, B. (2019). *Natural language processing with PyTorch: build intelligent language applications using deep learning*. " O'Reilly Media, Inc.".
[47] Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.
[48] Ginsberg, J. H. (2018). *Acoustics: A Textbook for Engineers and Physicists* (Vol. 2, p. 698). New York: Springer.
[49] Nový. R. (2009). Hluk a Chveni. České vysoké učení technické v Praze Česká technika - nakladatelstvl ČVUT
[50] Miček, J., & Jurečka, M. (2013). Moderné prostriedky implementácie metód číslicového spracovania signálov 1. *Žilina: EDIS*.
[51] Proakis J.G., Manolakis D.G. (2007) Digital signal processing: principles, algorithms and applications, Prentice Hall, ISBN 0-13-187374-1.
[52] Heinzel, G., Rüdiger, A., & Schilling, R. (2002). Spectrum and spectral density estimation by the Discrete Fourier transform (DFT), including a comprehensive list of window functions and some new at-top windows.
[53] FFTW knižnica, Dostupná: http://www.fftw.org/
[54] Smith, S. W. (1999). The scientist and engineer's guide to digital signal processing. Second Edition
[55] Isaacson, E. (1989). Numerical Recipes in C: The Art of Scientific Computing (William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling); Numerical Recipes: Example Book (C)(William T. Vetterling, Saul A. Teukolsky, William H. Press, and Brian P. Flannery). *SIAM Review*, *31*(1), 142.
[56] Smith, J. O. (2007). *Mathematics of the discrete Fourier transform (DFT): with audio applications*. Julius Smith.
[57] Understanding FFTs and Windowing Dostupné: https://download.ni.com/evaluation/pxi/Understanding%20FFTs%20and%20Windowing.pdf
[58] Zimmermann, J. Spektrálna skladba segmentov rečového signálu.
[59] Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, *66*(1), 51-83.
[60] Smith, J.O. Spectral Audio Signal Processing, Dostupné: http://ccrma.stanford.edu/~jos/sasp/
[61] Dokumentácia Pytorch, Dostupné: https://pytorch.org/docs/
[62] Rabiner, L., & Schafer, R. (2010). *Theory and applications of digital speech processing*. Prentice Hall Press.
[63] Valero, X., & Alias, F. (2012). Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE Transactions on Multimedia*, *14*(6), 1684-1689.
[64] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 776-780). IEEE.
[65] Fonseca, E., Favory, X., Pons, J., Font, F., & Serra, X. (2020). FSD50k: an open dataset of human-labeled sound events. *arXiv preprint arXiv:2010.00475*.

[66] Font, F., Roma, G., & Serra, X. (2013, October). Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 411-412).
[67] Piczak, K. J. (2015, September). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)* (pp. 1-6). IEEE.
[68] Chachada, S., & Kuo, C. C. J. (2014). Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing*, *3*.
[69] Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 116-131).
[70] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).
[71] Freeman, I., Roese-Koerner, L., & Kummert, A. (2018, October). Effnet: An efficient structure for convolutional neural networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 6-10). IEEE.
[72] Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, *24*(3), 279-283.
[73] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.
[74] Boddapati, V., Petef, A., Rasmusson, J., & Lundberg, L. (2017). Classifying environmental sounds using image recognition networks. *Procedia computer science*, *112*, 2048-2056.
[75] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.
[76] Ioffe, S. (2017). Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. *Advances in neural information processing systems*, *30*.
[77] LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K. R. (2012). Efficient backprop. In Neural networks: Tricks of the trade (pp. 9-48). Springer, Berlin, Heidelberg.
[78] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).
[79] Jacobsen, E., & Lyons, R. (2003). The sliding DFT. IEEE Signal Processing Magazine, 20(2), 74-80.
[80] Jacobsen, E., & Lyons, R. (2004). An update to the sliding DFT. IEEE Signal Processing Magazine, 21(1), 110-111.
[81] Šarafín, P. (2014). Modul pre digitalizáciu a predspracovanie akustického signálu: diplomová práca. Žilina: UNIZA, 68s.
[82] Gabor, D. (1946). Theory of communication. Part 1: The analysis of information. Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering, 93(26), 429-441.
[83] Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. IEEE transactions on Computers, 100(1), 90-93.
[84] Salomon, D. (2004). Data compression: the complete reference. Springer Science & Business Media.
[85] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., ... & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear

and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences, 454(1971), 903-995.
[86] Karlovský V. (2016): Analýza variability slnečnej aktivity metódou EMD, Zborník referátov z 23. celoštátneho slnečného seminára, 1-8
[87] Dokumentácia PyTorch: normalizačná vrstva, dostupná https://pytorch.org/docs/stable/generated/torch.nn.BatchNorm2d.html
[88] Dokumentácia TensorFlow: normalizačná vrstva, dostupná https://www.tensorflow.org/api_docs/python/tf/keras/layers/BatchNormalization
[89] Smith, J.O. Digital Audio Resampling Home Page, http://www-ccrma.stanford.edu/~jos/resample/
[90] Dokumentácia PyTorch: prevzorkovania, dostupná https://pytorch.org/audio/stable/tutorials/audio_resampling_tutorial.html
[91] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

Zoznam publikácií

[1] *An overview of practices used in environmental sound classification*. M. Chochul and P. Ševčík. In: ICETA 2021 : 19th IEEE International Conference on Emerging eLearning Technologies and Applications : proceedings : 19th IEEE International Conference on Emerging eLearning Technologies and Applications : proceedings / zost. František Jakab. - 1. vyd. - Denver : Institute of Electrical and Electronics Engineers, 2021. - 441 s. [online, USB-key]. - ISBN 978-1-6654-2101-0. - s. 76-81

[2] *A survey of low power wide area network technologies*. M. Chochul and P. Ševčík. In: ICETA 2020 : 18th IEEE International conference on emerging elearning technologies and applications : Information and communication technologies in learning : proceedings : 18th IEEE International conference on emerging elearning technologies and applications : Information and communication technologies in learning : proceedings / zost. František Jakab. - 1. vyd. - Denver : Institute of Electrical and Electronics Engineers, 2020. - 789 s. [online]. - ISBN 978-0-7381-2366-0. - s. 1-5

[3] *Optical communication system for a robot in project Aeris*. M. Chochul and P. Ševčík. In: ICETA 2021 : 19th IEEE International Conference on Emerging eLearning Technologies and Applications : proceedings : 19th IEEE International Conference on Emerging eLearning Technologies and Applications : proceedings / zost. František Jakab. - 1. vyd. - Denver : Institute of Electrical and Electronics Engineers, 2021. - 441 s. [online, USB-key]. - ISBN 978-1-6654-2101-0. - s. [1-5]

[4] *Dynamic system parameter identification based on the acceleration data.* P. Šarafin, L. Formanek and M. Chochul. In: ICETA 2020 : 18th IEEE International conference on emerging elearning technologies and applications : Information and communication technologies in learning : proceedings : 18th IEEE International conference on emerging elearning technologies and applications : Information and communication technologies in learning : proceedings / zost. František Jakab. - 1. vyd. - Denver : Institute of Electrical and Electronics Engineers, 2020. - 789 s. [online]. - ISBN 978-0-7381-2366-0. - s. [1-4]

[5] *Prediction of temperature in WSN using artificial intelligence*. L. Formanek, M. Chochul and O. Karpiš. In: Sensors and electronic instrumentation advances : proceedings of the 5th international conference on sensors and electronic instrumentation advances : proceedings of the 5th international conference on sensors and electronic instrumentation advances / S. Y. Yurish. - 1. vyd. - Barcelona : IFSA Publishing, 2019. - ISBN 978-84-09-14413-6. - s. 126-129.

[6] *Compressed Sensing and Acoustic Analysis for Use in Localization Tasks*. V. Olešnaníková, O. Karpiš, P. Šarafín, L. Formanek, M. Chochul. In: Sensors and electronic instrumentation advances [electronic] : proceedings of the 5th international conference on sensors and electronic instrumentation advances. - 1. vyd. - Barcelona: IFSA Publishing, 2019. - ISBN 978-84-09-14413-6. - s. 333-338.
[7] *Forest fire detection and localization within WSN*. M. Chochul. In: Mathematics in science and technologies : proceedings of the MIST conference 2019. - [S.l.] : [s.n.]. - ISBN 9781794002180. - s. 28-32

Zoznam citácií

[1] *A survey of low power wide area network technologies*. M. Chochul and P. Ševčík. In: ICETA 2020 : 18th IEEE International conference on emerging elearning technologies and applications : Information and communication technologies in learning : proceedings : 18th IEEE International conference on emerging elearning technologies and applications : Information and communication technologies in learning : proceedings / zost. František Jakab. - 1. vyd. - Denver : Institute of Electrical and Electronics Engineers, 2020. - 789 s. [online]. - ISBN 978-0-7381-2366-0. - s. 1-5

- 2021 [01] UGWUANYI, S., PAUL, G., IRVINE, J. *Survey of iot for developing countries : performance analysis of lorawan and cellular nb-iot networks.* In: Electronics. ISSN 2079-9292, 2021, vol. 10, iss. 18, art. no. 2224, s. 1-30. SCOPUS; WoS