Klasifikácia žánrov hudby pomocou konvolučných neurónových sietí

Čorbová, Gereg, Košarič

***Abstract*—V tejto práci sme predstavili prístup ku klasifikácii hudobných žánrov pomocou Mel-spektrogramov a konvolučných neurónových sietí (CNN). Súčasný výskum zdôrazňuje výhody CNN modelov pri spracovaní časovo-frekvenčných rysov v spektrogramoch, ktoré sú kľúčové pre identifikáciu hudobných žánrov. CNN efektívne extrahujú nízkoúrovňové aj vysokoúrovňové vzory, čím prekonávajú tradičné metódy strojového učenia a výrazne zlepšujú presnosť klasifikácie. Pomocou nášho navrhnutého prístupu, ktorý využíva prenesené učenie (transfer learning) pomocou modelu ResNet-50, sme dosiahli presnosť 72,93% s balanced accuracy 0.7293 a stratou 1.7436, čo naznačuje potenciál pre využitie v aplikáciách a systémoch pre automatizované odporúčanie hudby. Tento prístup tiež prekonáva presnosť 58,33%, ktorú sme dosiahli pomocou modelu ResNet-18 a presnosť 63.07% ktorú sme dosiahli pomocou 1D-CNN-RNN modelu.**

***Index Terms*—Klasifikácia hudobných žánrov, konvolučné neurónové siete, Mel-spektrogramy, hlboké učenie.**

I. Úvod

POPULÁRNYM prístupom ku klasifikácii žánrov hudby je využitie spektrogramov a hlbokých neurónových sietí. Jeho cieľom je zlepšiť automatizovanú klasifikáciu a odporúčanie hudby na základe zvukových dát. V prehľade literatúry pokrývame rôzne moderné prístupy využívané na identifikáciu hudobných žánrov, pričom zdôrazníme výhody a nevýhody jednotlivých prístupov. CNN sa osvedčili pri analýze komplexných vzorov v rôznych typoch médií vrátane obrazu a zvuku, čím preukázali svoju schopnosť identifikovať nízkoúrovňové aj vysokoúrovňové črty v hudobných nahrávkach. Cieľom tejto práce je preskúmať existujúce prístupy klasifikácie žánrov a na základe nich implementovať hlbokú neurónovú sieť na klasifikáciu žánrov hudby.

II. Existujúce prístupy

V práci [1] autori navrhli systém na klasifikáciu hudobných žánrov a odporúčanie hudby pomocou CNN, ktoré efektívne spracúvajú spektrogramy zvukových dát. Trénovanie a testovanie prebehlo na datasete GTZAN. Tento dataset pozostáva z 1000 zvukových stôp rozdelených do 10 žánrov, pričom každá skladba má dĺžku 30 sekúnd. Tento dataset je štandardne používaný pri hodnotení modelov na klasifikáciu hudobných žánrov. Autori na základe získaných výsledkov odporúčajú CNN pre aplikácie na odporúčanie hudby, ktoré sa prispôsobujú preferenciám používateľa.

Kombináciou CNN a rekurentných neurónových sietí (RNN) vznikajú CRNN, ktoré spájajú lokálnu extrakciu vlastností pomocou CNN s globálnou časovou sumarizáciou cez RNN, čím zlepšujú výkon modelu pri klasifikácii. V článku [16] v CRNN sa po vrstvách CNN používajú GRU (Gated Recurrent Units) na agregáciu časových vzorov, čo umožňuje efektívne modelovanie časových vzťahov s nižšou spotrebou pamäte, čo bolo preukázané pri testovaní na veľkých datasetoch, ako je Million Song Dataset.

V práci [2] autori klasifikujú hudobné žánre pomocou CNN a spektrogramov ako vstupných dát. Navrhovaný prístup zahŕňa predspracovanie, extrakciu rysov a klasifikáciu. Tréning na datasete GTZAN s použitím augmentácie dát (horizontálne otočenie a náhodné vystrihovanie) prekonal existujúce metódy, vrátane CRNN čím sa potvrdilo, že CNN s použitím správnych augmentácií dokáže výrazne zlepšiť presnosť klasifikácie hudobných žánrov.

V práci [4] autori porovnávajú rôzne prístupy ku klasifikácii hudobných žánrov, vrátane CNN, RNN, k-Nearest Neighbors (kNN), Naive Bayes a Support Vector Machines (SVM) na datasete GTZAN. Rozdelením zvukových vzoriek na menšie úseky zvýšili veľkosť datasetu. Najlepšie výsledky dosiahli metódy SVM a CNN, zatiaľ čo RNN-LSTM, kNN a Naive Bayes Classifier vykázali nižšiu presnosť. Z výsledkov vyplývavhodnosť CNN a SVM algoritmov pre klasifikáciu hudobných žánrov.

V práci [13] autori klasifikujú indické hudobné žánre pomocou CNN na spektrogramoch. Použitý bol dataset Indian Music Dataset (IMD2) obsahujúci 7000 skladieb z piatich žánrov. Rozdelením skladieb na päťsekundové segmenty zvýšili množstvo vzoriek v datasete. Experimenty počas trénovania pozostávali z porovnania presnosti spektrogramov skladajúcich sa iba z vokálnej zložky a spektrogramov obsahujúcou aj inštrumentálnu zložku.

V práci [14] autori klasifikujú hudobné žánre pomocou CNN na spektrogramoch, ktoré poskytujú časovo-frekvenčné informácie o skladbách. Použitý MK2 dataset obsahuje skladby z rôznych žánrov (Pop, Rock, Hip-Hop, Classical, Jazz, EDM, Soul), pričom každý žáner má aspoň 200 vzoriek vo formáte WAV. Počas trénovania boli použité modely vlastnej CNN siete, VGG16, ResNet34 a CRNN.

V práci [3] autori navrhli prístup ku klasifikácii hudobných žánrov kombinujúci zvukové obrazové rysy (auditory image features) s tradičnými akustickými a spektrálnymi rysmi, inšpirovaný ľudským auditívnym systémom. Model bol

testovaný na datasetoch GTZAN, GTZAN-NEW, ISMIR2004 a Homburg. Výsledky ukázali, že kombinácia týchto rysov značne zlepšuje presnosť klasifikácie. Použitie tejto kombinovanej metódy sa ukázalo ako efektívnejšie než tradičné metódy.

V práci [5] autori porovnali klasické modely strojového učenia (SVM, Random Forests, XGBoost) a CNN na klasifikáciu hudobných žánrov pomocou datasetu GTZAN. Pre CNN boli ako vstup použité Mel-spektrogramy, zatiaľ čo tradičné modely využívali rysy ako MFCC a spektrálny roll-off. Výsledky ukazujú, že CNN siete sú efektívnejšie pri spracovaní spektrogramov a lepšie zachytávajú komplexné zvukové rysy oproti testovaným metódam strojového učenia.

V práci [9] autori porovnávali tradičné modely strojového učenia (kNN, SVM, Logistická regresia) s CNN na klasifikáciu hudobných žánrov z datasetu GTZAN. Štúdia skúmala rôzne dĺžky vstupov – trojsekundové a tridsaťsekundové úseky. Experimenty preukázali vlastnosti bežných metód strojového učenia dosiahnuť vyššiu presnosť oproti CNN pri použití trojsekundových hudobných vstupoch.

V práci [12] autori porovnali CNN, CRNN a Logistickú regresiu (LR) na klasifikáciu hudobných žánrov pomocou FMA datasetu (variant fma_small) obsahujúceho 8000 skladieb z ôsmich žánrov. CNN, využívajúca spektrogramy ako vstup, dosiahla vyššiu presnosťv porovnaní s CRNN a LR. Výsledky ukazujú, že CNN je efektívnejšia než tradičné modely pri spracovaní zvukových stôp pomocou spektrogramov.

V práci [6] autori skúmajú klasifikáciu hudobných žánrov pomocou CNN a CRNN s využitím transfer learning a fine-tuningu. Implementovali multiframe prístup, ktorý analyzuje skladby v kratších časových oknách na lepšie zachytenie časovo-frekvenčných rysov. Tréning a testovanie prebehlo na datasete GTZAN s Mel-spektrogramami ako vstupom. Pri experimentoch dosiahol CRNN model vyššiu presnosť oproti použitému základnému CNN modelu.

V práci [15] autori používajú kombinovaný prístup CNN a RNN, konkrétne s použitím LSTM sietí. Tento prístup využíva CNN na extrakciu priestorových rysov zo spektrogramov a LSTM na zachytenie časových závislostí, čo je výhodné pri spracovaní sekvenčných dát, ako je hudba. Autori navrhli model, ktorý kombinuje CNN s vrstvami max-poolingu pre extrakciu priestorových rysov a LSTM vrstvami, ktoré pomáhajú spracovať časové vzory v hudobných skladbách. Autori zároveň využili dataset GTZAN. Pri spracovaní dát CNN extrahuje priestorové rysy zo spektrogramov, zatiaľ čo LSTM zachytáva časové závislosti. Výsledky experimentov naznačujú, že CNN dokáže efektívne identifikovať žánrové rysy a LSTM pridáva robustnosť v spracovaní časových vzorov.

V práci [7] autori klasifikujú hudobné žánre pomocou CRNN modelov (CNN-GRU a CNN-LSTM) na datasete GTZAN s použitím MFCC ako vstupných dát. Počas experimentov bola zistená skutočnosť, že CNN-LSTM model dokáže lepšie zachytiť časové a fekvenčné vzorky zvuku oproti CNN-GRU modelu.

V práci [8] autori skúmajú možnosť klasifikácie hudobných žánrov pomocou vizualizovaných spektrogramov a modelu YOLOv4, ktorý je bežne používaný na úlohy detekcie objektov. Tento prístup využíva Mel-frekvenčné spektrá, ktoré sú spracované pomocou MFCC, a konvertuje ich na vizuálne reprezentácie vhodné pre klasifikáciu prostredníctvom YOLOv4. Na prácu bol použitý dataset GTZAN, kde z každého zvukového súboru bol vytvorený Mel-spektrogram pomocou funkcií jazyku MATLAB. Výsledky modelu YOLOv4 boli zozbierané počas 10 experimentov. Tieto výsledky dokazujú, že použitie spektrogramov v kombinácii s YOLOv4 poskytuje vysokú presnosť pri klasifikácii hudobných žánrov.

V práci [10] autori aplikujú prístup založený na kNN na klasifikáciu hudobných žánrov. Na extrakciu rysov použili MFCC, ktoré sú často využívané pre ich schopnosť reprezentovať časovo-frekvenčné rysy zvuku. MFCC zjednodušuje komplexné zvukové signály na špecifické rysy, ktoré môžu byť následne analyzované pomocou kNN. Autori v práci využívajú dataset GTZAN. Získaný výsledok bol porovnaný s inými modelmi, pričom kNN poskytol lepšiu výkonnosť v porovnaní s niektorými tradičnými prístupmi, ako sú logistická regresia a Gaussovský model.

V práci [11] autori skúmajú dva prístupy ku klasifikácii hudobných žánrov: použitie modelu LSTM a hybridného modelu LSTM+SVM, pričom oba prístupy využívajú časovo frekvenčné rysy hudby, ktoré extrahujú pomocou MFCC. Pre tréning a testovanie modelov bol použitý GTZAN dataset, ktorý bol predspracovaný odstránením tichých segmentov, po ktorom boli audio dáta rozdelené na menšie úseky s dĺžkou deviatich sekúnd, čo zvýšilo množstvo vzoriek v trénovacích dátach. LSTM model dosiahol dosiahol na validačných dátach nižšiu presnosť v porovnaní s LSTM+SVM hybridným modelom, čo naznačuje, že tento hybridný prístup poskytuje lepšie výsledky než samotný LSTM model.

V práci [17] autori použili normalizované spektrálne modulové spektrum (NCMS) ako vstup pre hlbokú neurónovú sieť a ukázali, že tento prístup poskytuje presnosť porovnateľnú alebo lepšiu ako tradičné metódy využívajúce spektrálne črty. Pri testovaní na datasete GTZAN dosiahli pri použití samotných časových čŕt najnižšiu presnosť, ktorá sa zvýšila použitím samotných spektrálnych čŕt. Kombináciou temporálnych a spektrálnych čŕt sa presnosť zvýšila ešte viac, čo potvrdilo, že obidva typy čŕt sa navzájom dopĺňajú a poskytujú komplexnejšiu reprezentáciu hudobných dát.

III. Zhrnutie výsledkov existujúcich prístupov

Prehľad literatúry a doterajší výskum ukázal, že CNN modely dosahujú vyššiu presnosť v porovnaní s tradičnými

metódami strojového učenia, ako sú SVM či kNN. Tieto modely sú schopné lepšie zachytávať špecifické časovo-frekvenčné vzory charakteristické pre jednotlivé hudobné žánre, vďaka čomu poskytujú robustnejšie výsledky. Prehľad literatúry zároveň ukázal možnosť využitia hybridných prístupov CNN a RNN sietí, avšak poukázal aj na neistú úspešnosť pri použití týchto modelov.

Schopnosti CNN sietí zachytávať špecifické časovo-frekvenčné vzory robia z týchto sietí ideálnou voľbou pre automatizované systémy klasifikácie hudby a odporúčacie systémy. Vďaka schopnosti efektívne analyzovať a rozpoznávať zložité vzory v spektrogramoch môžu CNN modely presne identifikovať žánrové charakteristiky a prispôsobiť sa rôznorodosti hudobných štýlov. Tento prístup predstavuje spoľahlivé riešenie, ktoré môže výrazne zlepšiť výkon aplikácií odporúčania hudby a automatickej klasifikácie žánrov.

IV. Navrhovaný prístup

Na základe prehľadu literatúry a analýzy dostupných prístupov sme ako primárny postup zvolili CNN. Rozhodli sme sa otestovať použitie architektúr ResNet, konkrétne upravených predtrénovaných modelov ResNet-18 a ResNet-50, aby sme využili transfer learning. Táto voľba vychádza z osvedčenej schopnosti CNN efektívne spracovávať obrazové dáta, v našom prípade reprezentácie zvukových dát s pomocou Mel-spektrogramov. Mel-spektrogramy poskytujú časovo-frekvenčné informácie o hudobných nahrávkach, ktoré sú kľúčové pre identifikáciu žánrov. Pre prácu sme zvolili dataset GTZAN [18], podobne ako v mnohých iných prácach. Okrem toho sme implementovali aj model 1D CNN-RNN, ktorý kombinuje silu CNN a RNN s predpokladom, že model bude schopný lepšie zachytávať časové vzory v hudbe. Tento model je ďalším krokom v pokuse o zlepšenie presnosti klasifikácie.

Na optimalizáciu hyperparametrov sme sa pokúsili využiť framework Optuna, ktorý by nám umožnil jednoduchšie a automatizované testovanie rôznorodých konfigurácií modelov a tak by mohol prispieť k dosiahnutiu lepších výsledkov pri optimalizácii učenia.

V. Popis datasetu GTZAN

Dataset GTZAN je verejne dostupný dataset určený na klasifikáciu hudobných žánrov, ktorý obsahuje 1000 hudobných skladieb rozdelených do 10 rôznych žánrov. Každá skladba má dĺžku 30 sekúnd a je uložená vo formáte WAV, s vzorkovacou frekvenciou 22,05 kHz. Pre našu úlohu klasifikácie hudobných žánrov tento dataset poskytuje ideálny základ, pretože obsahuje rôznorodé hudobné žánre, ktoré sú dobre vyvážené.

Všetky skladby sú zaradené do desiatich rôznych žánrov, pričom každý žáner obsahuje 100 skladieb. Žánre zahrnuté v datasete sú: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae a rock. Tento vyvážený prístup zaručuje rovnaký počet skladieb pre každý žáner, čo je výhodné pre trénovanie modelov strojového učenia, ktoré sa učia na rovnocenných dátach pre každú triedu.

GTZAN je veľmi populárny v oblasti výskumu hudobnej klasifikácie, a to vďaka svojej štruktúre a rôznorodosti hudobných žánrov, ktoré umožňujú testovanie rôznych techník strojového učenia, ako sú CNN, RNN a ďalšie moderné metódy. Tento dataset je široko používaný na hodnotenie výkonu modelov pri úlohách, ako je rozpoznávanie žánrov a automatizované odporúčanie hudby.

Pre naše experimenty sme sa rozhodli použiť tento dataset, aby sme otestovali výkon našich modelov, ako sú ResNet-18, ResNet-50, a 1D CNN-RNN, pri úlohe klasifikácie hudobných žánrov. GTZAN poskytol dostatočne širokú škálu žánrov, čo nám umožnilo analyzovať a porovnávať rôzne modely a optimalizovať ich výkonnosť pri riešení tejto úlohy.

VI. Predspracovanie dát

Predspracovanie dát je kľúčovým krokom v procese trénovania modelu, pretože kvalita vstupných dát má zásadný vplyv na výkon výsledného klasifikátora. V našom prístupe sme vykonali nasledovné kroky predspracovania:

Zvukové nahrávky z datasetu boli spracované pomocou knižnice Librosa. Každá zvuková stopa bola normalizovaná na segmenty s dĺžkou 30 sekúnd. Z týchto segmentov boli následne generované Mel-spektrogramy, ktoré poskytujú vizuálnu reprezentáciu frekvenčného spektra v čase. Mel-spektrogramy sú vhodné pre spracovanie s pomocou CNN sietí, nakoľko premieňajú zvukový signál na obrazové dáta.

K predspracovaniu dát patrí aj implementácia triedy DynamicMusicDataset, ktorá načítava audio súbory zo zvoleného adresára, rozdeľuje ich na segmenty, pre ktoré sú aplikované zvukové augmentácie ako pridanie Gaussovho šumu *(parametre min_amplitude=0.001, max_amplitude=0.005 a p=0.3)*, zmenu rýchlosti prehrávania *(parametre min_rate=0.95, max_rate=1.05, leave_length_unchanged=True a p=0.3)* a posun tónu *(parametre min_semitones=-1, max_semitones=1 a p=0.3)*, ktoré pomáhajú zlešiť schopnosť generalizácie modelu na základe pridania miernej variability pre dáta. Takto predspracované zvukové dáta sú naśledne použité na generovanie Mel-spektrogramov. Tieto spektrogramy sú následne transformované do formátu vhodného pre vstup do neurónovej siete pomocou transformácií torch.ToTensor(), torch.Resize() a torch.Normalize() v súlades predspracovaním potrebným pre využitie sietí ResNet.

Na zabezpečenie flexibilného spracovania a zladenia rôznych datasetov v rámci experimentu používame vlastné dátové triedy. CustomConcatDataset umožňuje zlúčiť viaceré dátové sady a zabezpečuje, že označenia tried sú konzistentné medzi rôznymi datasetmi. Na to, aby sme efektívne zvládli rozdelenie dát na trénovacie, validačné a testovacie množiny, používame CustomSubset, ktorý umožňuje definovať vlastné indexy pre každý podmnožinu dát získanú pomocou stratifikovaného rozdelenia dát.

Všetky dáta sú potom rozdelené na testovaciu a holdout množinu v pomere 20:80, pričom holdout množina je následne rozdelená na 5 častí pomocou 5 násobnej stratifikovanej krížovej validácie.

VII. Použitý model

Modely ResNet-18 a ResNet-50 sú predtrénované hlboké neurónové siete s 18 a 50 vrstvami, ktoré využívajú tzv.

reziduálne bloky na efektívne učenie aj veľmi hlbokých modelov. Tieto modely patria medzi štandardné architektúry CNN využívané pre obrazovú klasifikáciu. Tieto modely boli upravené pre nami riešený problém.

Prvá konvolučná vrstva bola prispôsobená pre vstup rôznych rozmerov (224x224, 448x448,...) farebných spektrogramov s tromi farebnými kanálmi. Model ResNet-18 využíval vstup s rozmermi 224x224, zatiaľčo ResNet-50 využíva primárne vstupy 448x448 pixelov.

Okrem toho sme implementovali aj 1D CNN-RNN model, ktorý kombinuje výhody CNN a RNN. Tento model sa skladá z vlastne navrhnutej CNN siete a GRU siete, a je špeciálne navrhnutý na spracovanie časových vzorcov v hudobných dátach a je schopný zlepšiť výsledky klasifikácie v porovnaní s klasickými CNN modelmi, ktoré sa zameriavajú len na extrakciu statických vizuálnych čŕt. Model zahŕňa 1D konvolučné vrstvy na extrakciu lokálnych charakteristík amplitúdy a frekvencie v danom časovom úseku a GRU vrstvy na modelovanie časovej závislosti medzi rôznymi časovými úsekmi.

Rôzne konfigurácie modelov môžete vidieť v nasledujúcich tabuľkách.

TABLE I
KONFIGURÁCIE CNN MODELOV

Experiment	Model	FC	Dropout
1	ResNet-18	∅	∅
2	ResNet-50	128, 256	0.2
3	ResNet-50	84, 336	0.1023
4	ResNet-50	73, 499	0.4994
5	ResNet-50	128	0.2

TABLE II
KONFIGURÁCIE 1D-CNN-RNN MODELOV

Experiment	FC	CNN počet	RNN počet vrstiev	RNN skrytá vrstva
6	256, 128	256, 256, 128, 128	1	1024
7	338	178	2	294
8	73, 449	384, 319, 107,107, 59	3	359
9	256, 128	256, 256, 128, 128	1	1024

VIII. TRÉNOVANIE

Trénovanie modelu prebiehalo počas 10 000 epoch pomocou optimalizátora Adam pri ResNet-18 a s počiatočnou rýchlosťou učenia nastavenou na 0.0003 a AdamW pri ResNet-50 s počiatočnou rýchlosťou učenia nastavenou na 0.0001 a weight decay 0.0001. Proces trénovania zahŕňal stratifikovanú krížovú validáciu s K-prekrytiami (Stratified K-Fold cross-validation) s piatimi prekrytiami, čo zabezpečilo robustné hodnotenie modelu. Pre adaptívne znižovanie rýchlosti učenia sme využili CosineAnnealingLR scheduler, ReduceLROnPlateu scheduler (alebo žiadny), ktorý dynamicky upravoval rýchlosť učenia počas trénovania. Výkon modelu bol monitorovaný pomocou validačnej straty, pričom bol implementovaný mechanizmus early stopping, ktorý ukončil trénovanie, ak sa validačná strata nezlepšila počas dvadsiatich po sebe idúcich epoch. Ako chybovú funkciu sme použili CrossEntropyLoss, ktorá sa bežne používa pri klasifikácii do viacerých tried.

Okrem týchto metód sme implementovali aj optimalizáciu hyperparametrov pomocou frameworku Optuna. Tento framework umožňuje teoreticky testovať rôzne konfigurácie modelu (napr. počet vrstiev, veľkosti vrstiev, rýchlosť učenia) a nájsť tie, ktoré najlepšie vyhovujú dátam a riešenému problému.

Rôzne konfigurácie experimentov môžete vidieť v nasledujúcej tabuľke.

TABLE III
KONFIGURÁCIE PREBIEHANÝCH EXPERIMENTOV

Experiment	Učiaci parameter	Veľkosť obrazu	Hudobné augmentácie	Optuna
1	0.0003	224 x 224	n	n
2	0.0001	448 x 448	n	n
3	0.0007	372 x 372	n	a
4	0.0022	332 x 332	a	a
5	0.0001	448 x 448	a	n
6	0.0001	448 x 448	n	n
7	0.0014	276 x 276	n	a
8	0.0022	332 x 332	a	a
9	0.0001	448 x 448	a	n

IX. Výsledky

Z experimentov sme vyvodili nasledovné závery: Model ResNet-18 neposkytuje dostatočnú komplexitu pre zachytenie vlastností hudby v podobe Mel-spektrogramov. Vlastný 1D CNN-RNN model poskytol mierne zlepšenie oproti modelu ResNet-18. Najlepšie výsledky dosiahol model ResNet-50. Pre všetky modely bola dosiahnutá najvyššia presnosť s dvomi alebo tromi skrytými vrstvami, o veľkostiach medzi 128 a 256 neurónov. Použitie zvukových augmentácií mierne zlepšilo schopnosti siete generalizovať svoje znalosti. Využitie frameworku Optuna v tomto konkrétnom prípade neprinieslo žiadané zlepšenia. Výsledky najlepších iterácií jednotlivých experimentov môžete vidieť v nasledovnej tabuľke.

TABLE IV
Dosiahnuté výsledky jednotlivých experimentov

Experiment	Balanced Accuracy	Strata
1	0.5833 ± 0.02	1.2198 ± 0.0738
2	0.7213 ± 0.0346	1.743 ± 0.0333
3	0.4107 ± 0.0893	2.0432 ± 0.0892
4	0.2593 ± 0.036	2.1779 ± 0.0404
5	**0.7293 ± 0.0293**	**1.7436 ± 0.0271**
6	0.6147 ± 0.0153	1.8547 ± 0.0107
7	0.2913 ± 0.1013	2.1579 ± 0.1116
8	0.1 ± 0.0	2.3026 ± 0.0
9	**0.6307 ± 0.0374**	**1.8339 ± 0.0386**

X. Záver

V tejto práci sme predstavili prístup ku klasifikácii hudobných žánrov pomocou CNN, konkrétne modelov ResNet-18 a ResNet-50, ktoré boli trénované na spektrogramoch hudobných nahrávok. Cieľom bolo preskúmať efektívnosť týchto modelov pri klasifikácii hudobných žánrov.

Výsledky experimentov ukázali, že model ResNet-50 dosiahol najlepšiu hodnotu balanced accuracy 72,93%, pričom ResNet-18 vykázal mierne nižšie hodnoty s presnosťou 58,33%. Model 1D CNN-RNN dosiahol hodnotu balanced accuracy 63,07%, čo naznačuje, že tento model má potenciál zlepšiť výsledok v tejto úlohe.

Implementácia stratifikovanej krížovej validácie s 5-prekrytiami poskytla stabilnejšie a konzistentnejšie hodnotenie modelov, zatiaľ čo early stopping pomohol optimalizovať dĺžku trénovania a zabrániť preučeniu. Použitie zvukových augmentácií prispelo k lepšej schopnosti sietí generalizovať získané znalosti. Všetky trénované modely do určitej miery preukázali schopnosť extrahovať nízkoúrovňové aj vysokoúrovňové črty zo Mel-spektrogramov.

Na základe dosiahnutých výsledkov môžeme konštatovať, že navrhovaný prístup s využitím modelov ResNet-18 a ResNet-50 predstavuje efektívne riešenie pre automatizovanú klasifikáciu hudobných žánrov. Zároveň však chceme podotknúť, že experimentovanie s rôznymi druhmi CNN a RNN sietí by mohlo viesť k zvýšeniu dosiahnutej presnosti trénovaných modelov a zlepšiť výsledky v tejto úlohe.

References

[1] J. Dias, H. Deshmukh, V. Pillai, and A. Shah, *Music Genre Classification & Recommendation System using CNN*, SSRN Electronic Journal, 2022. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4111849

[2] A. K. Athulya and S. Sindhu, *Deep Learning Based Music Genre Classification Using Spectrogram*, 2nd International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), 2021. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3883911

[3] X. Cai and H. Zhang, *Music Genre Classification Based on Auditory Image, Spectral and Acoustic Features*, Multimedia Systems, vol. 28, pp. 779–791, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s00530-021-00886-3

[4] G. Chettiar and K. S., *Music Genre Classification Techniques*, Conference Paper, Vellore Institute of Technology, India, November 2021. [Online]. Available: https://www.researchgate.net/publication/356377974_Music_Genre_Classification_Techniques

[5] M. Shah, L. Gohil, N. Pujara, T. Vyas, K. Mangaroliya, and S. Degadwala, *Music Genre Classification using Deep Learning*, Proceedings of the Sixth International Conference on Computing Methodologies and Communication (ICCMC), pp. 974-978, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9753953

[6] Mounika K. S., Deyaradevi S., Swetha K., and V. Vanitha, *Music Genre Classification Using Deep Learning*, 2021 International Conference on Advances in Electrical, Electronics, Communication, Computing and Automation (ICAECA), pp. 1-6, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9675685

[7] N. Srivastava, S. Ruhil, and G. Kaushal, *Music Genre Classification using Convolutional Recurrent Neural Networks*, 2022 IEEE 6th Conference on Information and Communication Technology (CICT), pp. 1-5, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9997961

[8] Y.-H. Cheng, P.-C. Chang, and C.-N. Kuo, *Music Genre Classification Based on Visualized Spectrogram*, 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), pp. 217-221, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9778067

[9] N. Ndou, R. Ajoodha, and A. Jadhav, *Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches*, 2021 International IoT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1-6, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9422487

[10] J. K. Bhatia, R. D. Singh, and S. Kumar, *Music Genre Classification*, 2021 5th International Conference on Information Systems and Computer Networks (ISCON), pp. 1-5, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9702303

[11] D. S. and B. G. Prasad, *Music Classification based on Genre using LSTM*, Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 985-991, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9182850

[12] S. Chillara, K. A. S., S. A. Neginhal, S. Haldia, and V. K. S., *Music Genre Classification using Machine Learning Algorithms: A Comparison*, International Research Journal of Engineering and Technology (IRJET), vol. 6, no. 5, pp. 851–858, May 2019. [Online]. Available: https://www.irjet.net/archives/V6/i5/IRJET-V6I5174.pdf

[13] P. Kalapatapu, S. Gupta, A. Sharma, J. L. Sravani, and A. Malapati, *Genre Classification using Spectrograms as Input to CNN on Indian Music*, 2021 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), pp. 12-15, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9574060

[14] M. K. Abbas, M. Mudassir, K. Gupta, and R. Jain, *Classification of Musical Genres Using Audio Spectrograms*, 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 286-291, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10541572

[15] A. A. Khamees, H. D. Hejazi, M. Alshurideh, and S. A. Salloum, *Classifying Audio Music Genres Using CNN and RNN*, in Advanced Machine Learning Technologies and Applications. AMLTA 2021. Advances in Intelligent Systems and Computing, vol. 1339, A. E. Hassanien, K. C. Chang, and T. Mincong, Eds., Springer, Cham, pp. 315–323, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-69717-4_31

[16] Choi, K. et al. (2017). Convolutional recurrent neural networks for music classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2392–2396.*

[17] Il-Young, J., Kyogu. L. (2016). Learning Temporal Features Using a Deep Neural Network and its Application to Music Genre Classification. *Ismir, 434–440.*

[18] Tzanetakis, G., Essl, G., Cook, P. (2001). Automatic Musical Genre Classification of Audio Signals. [Online]. The International Society for Music Information Retrieval. Available: http://ismir2001.ismir.net/pdf/tzanetakis.pdf