Another Approach for Understanding the Relationship Between Noise, Spectral Bias, and Performance in an Audio Classifier

Dumitriu Marian
Universitatea Babes-Bolyai, Cluj-Napoca, Romania

Abstract

Lucrarea investigheaza cum zgomotul afecteaza comportamentul si performanta unui clasificator audio bazat pe retele neuronale convolutionale. Se propune o metoda numita Spectral Bias Score (SBS), care arata cat de mult modelul se bazeaza pe anumite frecvente din sunet. Printr-o serie de experimente controlate, in care se adauga diferite tipuri de zgomot si se modifica nivelul semnalului, se observa cum schimbari in SBS se leaga de modificari concrete ale performantei. Scopul principal este un protocol experimental care ajuta la intelegerea modului in care zgomotul afecteaza modelul si la gasirea de solutii pentru a-l face mai robust.

AMS: 68T07, 68T10;
ACM: Computing methodologies → Machine learning → Neural networks;
Applied computing → Sound and music computing.

1 Introduction

Obiectivul acestei lucrari este investigarea daca si in ce masura clasificatoarele audio bazate pe retele neuronale convolutionale prezinta un *bias spectral*, adica o dependenta disproportionala de anumite benzi de frecventa, si cum aceasta afecteaza robustetea la zgomot. Accentul se pune pe masurare experimentala si interpretare a semnalelor interne ale modelului (gradient, saliency) mai mult decat pe optimizarea performantelor.

2 Related Work

In literatura de specialitate, clasificarea audio automata este dominata de arhitecturi profunde si complexe, precum ResNet, EfficientNet sau Transformers adaptati pentru spectrograme [16]. Majoritatea eforturilor de cercetare se concentreaza pe cresterea acuratetei globale pe seturi de date standard (ESC-50, UrbanSound8k) prin tehnici de augmentare a datelor, cum ar fi SpecAugment, menite sa previna overfitting-ul [18].

Cu toate acestea, exista o lacuna in intelegerea modului in care aceste modele percep structura spectrala a sunetului. Desi metode precum Grad-CAM ofera o vizualizare a zonelor de interes, ele nu cuantifica dependenta modelului de anumite benzi de frecventa. Mai mult, studiile existente compara adesea performanta doar in conditii de zgomot alb (white noise), ignorand distorsiunile spectrale structurate care apar in lumea reala (ex: efectul de telefon - low-high pass, sau sunetul infundat - low pass).

Abordarea aceasta se distinge prin utilizarea unei arhitecturi CNN clasice, simple, ca punct de analiza al clasificarii sunetelor. In loc sa incercam sa obtinem "state-of-the-art" accuracy cu modele complexe, folosim acest model baseline pentru a studia corelatia dintre metrica propusa, Spectral Bias Score (SBS), si schimbarea performantei sub efectul diferitelor filtre acustice. Ne propunem sa demonstram ca modelele neuronale tind sa dezvolte "preferinte" spectrale care nu coincid neaparat cu perceptia umana.

Literatura relevanta acopera: pipeline-uri CNN pe spectrograme, augmentare cu zgomot, mecanisme de atentie time–frequency, metode de denoising si discutii teoretice despre spectral bias. Mai jos se prezinta sumarul a cinci lucrari selectate, utile pentru studiu.

2.1 Noise robust sound event classification with convolutional neural network [1]

Lucrarea propune o metoda de crestere a robustetii la zgomot a clasificatoarelor audio prin utilizarea CNN-urilor pe spectrograme procesate. Se descriu detalii privind conversia semnalului in spectrograma, reducerea dimensiunii imaginilor spectrale, arhitectura CNN si strategii de preprocesare. Pentru studiul prezent, aceasta lucrare ofera validarea alegerii reprezentarii pe spectrograma si un punct de comparatie pentru tehnici simple de augmentare/denoising.

2.2 Abnormal respiratory sounds classification using deep CNN through artificial noise addition [2]

Lucrarea exploreaza augmentarea prin adaugare de zgomot artificial pentru a imbunatati robustețea la zgomot in clasificarea sunetelor respiratorii. Sunt prezentate detalii despre preprocesare, extragerea de trasaturi (inclusiv utilizarea FFT si spectrogramelor) si efectele augmentarii asupra performantelor. In contextul proiectului, aceasta lucrare inspira designul ablațiilor: comparatia antrenare pe date curate versus antrenare cu augmentare.

2.3 A CNN sound classification mechanism using data augmentation [3]

Lucrarea extinde tehnici de augmentare prin aplicarea de filtre direct pe spectrograme Mel, ceea ce creste diversitatea datelor de antrenament. Metoda propusa ofera idei pentru testarea robustetii modelului la distorsiuni spectrale specifice, nu doar la zgomot alb. Aceasta literatura sustine testele de band-masking si experimentele cu filtre aplicate pe spectrograme.

2.4 Environmental sound classification using temporal-frequency attention based convolutional neural network [4]

Lucrarea introduce un mecanism de atentie time–frequency care permite modelului sa se concentreze pe regiunile cele mai informative din spectrograma. Implementarea si vizualizarile de atentie sunt relevante ca metodologie pentru a obtine explicatii interne (Grad-CAM, attention maps) si pentru a compara modul in care atentie vs. gradient reflecta dependenta spectrala.

2.5 Noise Suppression in Audio Classification [5]

Lucrarea prezinta o abordare de preprocesare prin suprimare a zgomotului (denoising) inainte de clasificare, folosind modele dedicate (ex.: U-Net). Sunt analizate efectele denoising-ului asupra acuratetii si asupra calitatii semnalului rezultat. Acest punct este util pentru ablatia privind mitigarea: compararea clasificatorului pe date curate, zgomotoase si „curatate".

3 Methodology

3.1 Research hypothesis

Zgomotul care afecteaza anumite benzi de frecventa modifica dependenta clasificatorului de acele benzi si aceasta modificare poate fi cuantificata prin Spectral Bias Score (SBS). Cresterea SBS in anumite conditii trebuie sa coreleze cu degradarea performantei cand eliminam acele benzi. Mai multe detalii despre modul de a analiza relatia experimental se gaseste la capitolele urmatoare.

3.2 Definitions and diagnostics

- **SNR (signal-to-noise ratio)**: raportul in decibeli intre puterea semnalului si puterea zgomotului.
- **Band-masking**: atenuarea sau eliminarea unei benzi de frecventa la testare.
- **Input gradient**: g, vectorul de derivate ale pierderii fata de fiecare esantion; indica sensibilitatea deciziei la perturbatii.
- **Spectral Bias Score (SBS)**: definim SBS ca o masura a sensibilitatii locale a retelei. Fie L functia de loss si S spectrograma de intrare. Calculam harta de salienta $G = |\nabla_S L|$. SBS pentru o banda de frecventa B este raportul dintre gradientii acumulati in acea banda si gradientul total:

$$SBS(B) = \frac{\sum_{f \in B} \sum_t G_{t,f}}{\sum_{f,t} G_{t,f}}$$

  Aceasta metrica ne permite sa prezicem de dinainte cat de mult va suferi modelul daca o anumita portiune din spectru este eliminata. Daca gradientul ar fi definit in domeniul temporal in loc de spectrograma, se calculeaza aplicand Fast Fourier Transform pe g, obtinandu-se energia per frecventa.

4 Data and preprocessing

Se foloseste un dataset de sunete scurte, ESC-50 (alternativ UrbanSound8K); pentru prototip acceptam un subset (20 de clase, 800 exemple). Fisierele se resampleaza la 44.1 kHz si se normalizeaza. Reprezentarea principala propusa este Mel-spectrograma, insa se va compara si o varianta cu raw-waveform (1D) pentru a investiga diferentele de comportament spectral si sensibilitate la zgomot. Aceasta comparatie face parte din contributia practica: se va arata daca si cum reprezentarea afecteaza SBS si robusteatea in ambele cazuri.

5 Model and training settings

Pentru a asigura o analiza potrivita a bias-ului spectral, s-a dezvoltat un cadru experimental care izoleaza variabilele de frecventa. S-a optat pentru o arhitectura CNN standard (baseline), evitand mecanismele de "self-attention" sau blocurile reziduale complexe care ar putea masca dependentele spectrale primare.

5.1 Arhitectura Modelului si Datele

Modelul utilizat este un CNN secvential clasic, compus din 4 blocuri de convolutie (Conv2D), urmate de straturi de MaxPolling si BatchNormalization. Clasificarea finala se realizeaza printr-un strat Dense cu activare Softmax. Intrarea in retea este reprezentata de Log-Mel Spectrograme (128 benzi Mel $\times$ 431 cadre), extrase din ferestre audio de 5 secunde.

5.2 Protocolul de Perturbare Spectrala

Pentru a valida robustetea modelului la alterarea caracteristicilor acustice, am generat 5 variatii ale setului de date, simuland diferite efecte audio:

- **Clean (D_{clean})**: Semnalul original, referinta.
- **White Noise (D_{wn})**: Adaugare de zgomot aleator, afectand uniform tot spectrul.
- **Low Pass (D_{lp})**: Filtru Butterworth (cutoff 8250 Hz). Simuleaza un sunet infundat.
- **High Pass (D_{hp})**: Filtru Butterworth (cutoff 2000 Hz). Simuleaza un sunet ascutit/subtire.

- **Low-High Pass (D_{lh})**: Eliminarea benzii 3500-7500 Hz.

6 Experiments

Experimentele se concentreaza pe masurare: ce degradeaza, ce imbunatateste si cum se reflecta aceste schimbari in diagnosticele spectrale.

Am rulat experimentele pe un subset initial de date pentru a valida metodologia. Modelul baseline CNN a atins o acuratete de **97.62%** pe datele curate (D_{clean}), demonstrand ca arhitectura este suficienta pentru task-ul de baza. Totusi, comportamentul sau sub diferite efecte audio a scos la iveala limitari interesante.

6.1 Rezultate si Impactul Efectelor Audio

Analiza SBS a indicat ca modelul isi concentreaza peste 70% din "atentie" (gradienti) in zona de frecvente joase (<4000 Hz). Testarea pe seturile filtrate a confirmat aceasta dependenta:

Tip Efect (Dataset)	**Acuratete**	**Observatii Acustice**
Clean	97.62%	Referinta
Low Pass (<8.25k Hz)	93.00%	Sunet infundat
Low-High Pass (fara 3.5k-7.5k)	77.25%	Lipsa "culorii" sunetului
High Pass (>2k Hz)	48.88%	Sunet mai subtire, gol

Table 1: Performanta modelului sub diferite distorsiuni spectrale

6.2 Interpretarea Rezultatelor si Comparatia cu Perceptia Umana

Rezultatele experimentale arata diferente clare intre modul in care modelul si oamenii percep sunetul. Analiza filtrelor, corelata cu scorurile SBS, explica mecanismele prin care CNN-ul ia decizii.

6.2.1 Energie versus Detalii

Diferenta dintre performanta pe setul **Low Pass** (93.00%) si cea pe setul **High Pass** (48.88%) indica faptul ca modelul cauta "scurtaturi". Pentru urechea umana, frecventele inalte sunt esentiale pentru claritate si detalii. Totusi, modelul a functionat bine doar cu frecventele joase. Asta sugereaza

ca reteaua nu asculta "obiectul" sonor, ci detecteaza doar zonele cu energie mare (in acest caz bas, ritm). Cand aceste zone sunt eliminate (High Pass), modelul nu se mai descurca, chiar daca sunetul ramane inteligibil pentru om.

6.3 Calibrarea Metricii SBS

In experimentul **Low-High Pass** (eliminarea benzii 4k-8k Hz), modelul a inregistrat o scadere a performantei de 19.4%. Aceasta valoare este remarcabil de apropiata de predictia SBS, care atribuia acestei benzi o importanta de aproximativ 14% din totalul energiei gradientilor. Aceasta proximitate valorica (14% estimat vs 19% real) valideaza ipoteza noastra: SBS este o metrica *calibrata*. Diferenta mica de 5 puncte procentuale poate fi atribuita pierderii coerentei semnalului, insa ordinul de marime confirma ca SBS poate cuantifica corect, a-priori, cat de mult se bazeaza reteaua pe o anumita portiune din spectru.

6.4 Analiza pe Clase de Sunete

Un aspect crucial este comportamentul diferentiat in functie de natura sursei sonore. Desi acuratetea globala scade sub filtre, analiza matricei de confuzie arata ca anumite clase sunt reziliente:

- **Sunetele de joasa frecventa** (ex: motoare, tunete, impact mecanic) si-au pastrat o rata mare de detectie in experimentul Low-Pass, deoarece componenta lor spectrala principala a ramas intacta.
- **Sunetele de inalta frecventa** (ex: ciripit de pasari, sticla sparta) inregistreaza o scadere abrupta a performantei in acelasi scenariu, fiind adesea confundate cu zgomotul de fundal.

Aceste rezultate sugereaza ca "bias-ul spectral" nu provine doar din modul de antrenare, ci reflecta si particularitatile fizice ale claselor. Modelul identifica corect zonele cu energie predominanta, dar nu reuseste sa invete suficient de bine trasaturile secundare care ar asigura robustete in conditii de distorsiune.

6.5 Confirmarea Vizuala prin Grad-CAM

Hartile Grad-CAM sustin aceste observatii. Cand modelul clasifica corect, Grad-CAM arata activare clara pe zonele cu SBS ridicat. In schimb, cand modelul greseste pe datele filtrate, vedem o activare imprastiata in spectru. Aceasta activare haotica indica faptul ca, odata ce frecventele sale dominante

sunt alterate, reteaua nu poate reorienta eficient atentia pe frecventele ramase. Modelul nu are mecanisme interne de adaptare sau de completare a informatiei lipsa.

7 Metrics and analysis

Rezultatele obtinute confirma validitatea scorului SBS ca predictor al performantei. Corelatia stransa intre importanta spectrala calculata teoretic si scaderea reala a acuratetei demonstreaza utilitatea metodei. Se masoara accuracy, precision, recall si F1, impreuna cu SBS pentru a evalua dependenta de frecventa. Se vizualizeaza:

- cum scade acuratetea la zgomot crescut;
- cum se modifica SBS;
- zonele de frecventa importante pentru model.

Scopul este de a observa cum afecteaza zgomotul comportamentul spectral al clasificatorului si daca metodele propuse pot reduce acest efect.

8 Metrics and Analysis

Rezultatele obtinute confirma validitatea scorului SBS ca predictor al performantei. Corelatia stransa intre importanta spectrala calculata teoretic si scaderea reala a acuratetei demonstreaza utilitatea metodei. Evaluarea s-a realizat pe setul de testare, utilizand metricile: SBS, Accuracy, F1-Score, Precision si Recall pe clase si matricea de confuzie.

8.1 Performanta Generala

Modelul a obtinut o acuratete de baza ridicata pe setul curat, insa degradarea sub zgomot variaza semnificativ in functie de tipul filtrului aplicat. Tabelul 1 prezinta o comparatie detaliata a metricilor principale.

Un aspect notabil este diferenta dintre *White Noise* si *High Pass*. Desi zgomotul alb afecteaza tot spectrul, modelul pastreaza o acuratete decenta (88.5%), deoarece informatia din frecventele joase ramane partial accesibila. In schimb, eliminarea completa a joaselor (High Pass) duce la prabusirea modelului (48.88%), confirmand dependenta critica de aceasta zona spectrala.

8.2 Analiza Detaliata pe Clase

Pentru a intelege nuantele, am analizat raportul de clasificare. Performanta nu este uniforma; unele sunete sunt mult mai robuste decat altele. Analiza F1-Score (2) pe clase specifice demonstreaza ca bias-ul spectral al modelului se manifesta diferit in functie de structura fizica a sunetului.

Clasa	**F1 (Cl)**	**F1 (LP)**	**F1 (HP)**	**F1 (LH)**	**F1 (WN)**
Helicopter	0.976	**0.879**	0.000	0.816	0.000
Sea Waves	0.962	0.812	0.000	**0.049**	0.306
Rain	0.952	0.810	0.120	**0.048**	0.133
Clapping	0.937	0.791	0.314	0.545	0.000

Table 2: Comparatie F1-Score pe clase pentru diferite conditii spectrale.

Aceasta analiza confirma ipoteza "Shortcut Learning": clasele care se definesc prin ritm si anvelopa de joasa frecventa (elicopter) rezista filtrarii, in timp ce clasele definite prin timbru si armonice inalte (foc, drujba) devin de nerecunoscut pentru model.

1. Asimetria de Dependenta: Clasa "Helicopter" (dependenta de frecventele joase) ilustreaza bias-ul fundamental al modelului:

- Cand eliminam frecventele inalte (**Low Pass**), F1-Score ramane ridicat (0.879).
- Cand eliminam frecventele joase (**High Pass**), F1-Score scade la 0.000 (esec total).

Acest contrast masiv confirma ca informatia din banda de inalta frecventa (care este eliminata in testul Low Pass) este aproape irelevanta pentru decizie, in timp ce eliminarea benzii de joasa frecventa duce la colaps. Aceasta valideaza SBS, demonstrand ca modelul nu are o reprezentare redundanta si se bazeaza exclusiv pe acea banda critica (Low Frequencies).

2. Dependenta de Context (Ex: Rain / Sea Waves): In cazul claselor texturale, metrica SBS a indicat o importanta scazuta ($\sim$ 14%) pentru banda medie (3.5k–7.5k Hz). Totusi, testul Low-High Pass a aratat o prabusire a performantelor (F1 $\approx$ 0.05).

- Diagnosticul SBS: Acest contrast arata o limita a SBS. Metrica masoara contributia directa a unei benzi, dar nu surprinde rolul acesteia de zona de legatura intre benzile joase si cele inalte. Zgomotul aplicat

in banda medie nu doar elimina continut, ci intrerupe si continuitatea dintre aceste regiuni.
- Solutie pentru robustete: Augmentarea cu zgomot nu este suficienta. O solutie este mascarea complementara, prin care modelul este fortat ca in unele iteratii de antrenare sa clasifice pe baza benzii medii. Acest lucru creste importanta SBS a acestei benzi si introduce redundanta utila.

3. Fragilitatea Sunetelor Scurte (Ex: Clapping): Pentru sunetele scurte, analiza SBS arata o distributie a gradientului pe tot spectrul, ceea ce ar sugera robustete. Cu toate acestea, performanta scade puternic in orice scenariu de filtrare (LP: 0.791, HP: 0.314).

- Diagnosticul SBS: Desi informatia este distribuita, modelul nu trateaza benzile ca fiind independente. Pentru un tranzient scurt, are nevoie de tot spectrul simultan pentru a reconstrui evenimentul. Zgomotul partial afecteaza forma temporala, iar arhitectura CNN nu poate compensa lipsa unor portiuni din spectru.
- Solutie pentru robustete: In acest caz solutia este in domeniul timpului. Augmentarile temporale (deplasari, modificari de durata) sau arhitecturile care pastreaza rezolutia temporala pe toate benzile pot mentine coerenta necesara pentru detectarea tranzientelor.

Ca remarca, analiza pe clase valideaza ipoteza **"Shortcut Learning"**: modelul identifica corect zona dominanta de energie, dar nu reuseste sa invete trasaturi redundante care ar permite recunoasterea sunetului in cazul in care banda principala este compromisa.

9 Reproducibility and organization

Codul va fi organizat in scripturi. Setarile experimentelor se pastreaza. Un notebook Jupyter va reproduce figurile principale.

10 Limitations

Limitari anticipate: hardware-ul pentru a rula testarile, dimensiunea si diversitatea dataset-ului; estimari zgomotoase ale gradientului pe probe individuale; posibile asimetrii (ex.: nivel mediu de amplitudine) ce necesita controale.

11 Contribution

Aceasta lucrare aduce trei contributii majore:

1. **Metoda de Diagnostic (White-Box):** Introducerea SBS ca instrument standardizat pentru cuantificarea dependentei spectrale, permitand evaluarea riscului de esec al unui model inainte de implementare.
2. **Validarea Empirica a Bias-ului Spectral:** Demonstrarea faptului ca acuratetea (o metrica Black-Box) este direct corelata cu SBS (o metrica White-Box), confirmand ca modelele CNN sunt predispuse la "Shortcut Learning" bazat pe energia frecventelor.
3. **Directii de Augmentare:** Propunerea strategiei *SBS-Guided Augmentation*. In loc de mascarea aleatorie (SpecAugment), se sugereaza mascarea tintita a benzilor cu SBS maxim in timpul antrenarii. Acest lucru ar forta modelul sa invete sa foloseasca tot spectrul, sporind robustetea.

12 Conclusions

S-a prezentat un plan experimental pentru cuantificarea interactiunii dintre zgomot, comportamentul spectral si performanta in clasificatoare audio bazate pe retele neuronale convolutionale. Rezultatele pot ghida practici de antrenare, preprocesare sau postprocesare pentru aplicatii si clasificatoare de audio robuste.

References

[1] A. Özer and E. Can, “Noise robust sound event classification with convolutional neural network,” in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, Izmir, Turkey, 2018, pp. 1-4. doi:10.1109/SIU.2018.8404557

[2] S. Zulfiqar, F. H. Gilani, S. A. R. Rizvi, I. Zafar, and G. G. Ali, “Abnormal respiratory sounds classification using deep CNN through artificial noise addition,” *Journal of Medical and Biological Engineering*, vol. 41, pp. 581–591, 2021. doi:10.1007/s40846-021-00637-6

[3] I. Chu, M. Park, J. H. Lee, and J. W. Shin, “A CNN Sound Classification Mechanism Using Data Augmentation,” *IEEE Signal Processing Letters*, vol. 30, pp. 692-696, 2023. doi:10.1109/LSP.2023.3283733

[4] J. Ma, S. Zhang, Z. Niu, and J. Liu, “Environmental sound classification using temporal-frequency attention based convolutional neural network,” *Scientific Reports*, vol. 11, no. 1, p. 14782, 2021. doi:10.1038/s41598-021-94235-8

[5] S. Leonardsson, V. Bergkvist, "Noise Suppression in Audio Classification", Lund University, 2025 lup.lub.lu.se/student-papers/search/publication/9200623

[6] S. P. Dubagunta, A. T. L. Ho, V. K. T. Truong, W. L. T. Oo, and A. W. H. Khong, “Speech Denoising for Robust Audio Classification,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 671-675. doi:10.1109/ICASSP40776.2020.9054452

[7] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, “On the Spectral Bias of Neural Networks,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, vol. 97, pp. 5301-5310. URL: proceedings.mlr.press/v97/rahaman19a.html

[8] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification,” in *Proceedings of the 23rd ACM International Conference on Multimedia (MM ’15)*, 2015, pp. 1015–1018. doi:10.1145/2733373.2806390

[9] S. Hershey et al., “CNN Architectures for Large-Scale Audio Classification,” in *2017 IEEE International Conference on Acous-*

tics, Speech and Signal Processing (ICASSP), 2017, pp. 131–135. doi:10.1109/ICASSP.2017.7952132

[10] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279-283, 2017. (Also available as arXiv:1608.04363) doi:10.1109/LSP.2017.2657381

[11] Y. Tokozume and T. Harada, "Learning from Between-class Examples for Deep Sound Recognition," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018. URL: arxiv.org/abs/1711.10282

[12] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access*, vol. 11, pp. 106620-106653, 2023. doi:10.1109/ACCESS.2023.3319207

[13] J. Ekanayake, S. Shanthakumar, S. Shakila, and S. Pathirana, "Environmental Sound Classification Using Deep Learning," *Uwa Wellassa University Technical Report*, 2021. URL: lib.uwu.ac.lk/handle/123456789/5431

[14] X. Zhu, K. Xu, D. Wang, L. Zhang, B. Li, and Y. Peng, "Environmental Sound Classification Based on Multi-temporal Resolution Convolutional Neural Network Combining with Multi-level Features," *arXiv preprint arXiv:1805.09752*, 2018. URL: arxiv.org/abs/1805.09752

[15] D. T. N. Le, C. T. H. Le, T. H. Truong, and H. T. T. Nguyen, "A review: Deep learning for environmental sound classification on embedded devices," *Artificial Intelligence Review*, vol. 56, pp. 11451–11494, 2023. doi:10.1007/s10462-023-10443-4

[16] Y. Xu, Q. Duan, W. Yang, H. Zhu, E. Yang, and H. Chen, "Comparison of pre-trained CNNs for audio classification using transfer learning," *Future Generation Computer Systems*, vol. 138, pp. 21-33, 2023. doi:10.1016/j.future.2022.08.010

[17] J. Li, W. Dai, F. Metze, and S. Das, "A Comparison of Deep Learning Methods for Environmental Sound Detection," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 116-120. doi:10.1109/ICASSP48485.2024.10447225

[18] S. Latif, R. Qayyum, M. Usman, and J. Qadir, "A Survey on Deep Learning for Audio Signal Processing," *arXiv preprint arXiv:2105.03027*, 2021. URL: arxiv.org/abs/2105.03027

[19] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proceedings of the 22nd ACM International Conference on Multimedia (MM '14)*, 2014, pp. 1041–1044. doi:10.1145/2647868.2655045

[20] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 921-928. doi:10.1109/SLT.2018.8639585